Provision Quality-of-Service Controlled Content Distribution in Vehicular Ad Hoc Networks

by

Hao Luan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

By equipping vehicles with the on-board wireless facility, the newly emerged vehicular networking targets to provision the broadband serves to vehicles. As such, a variety of novel and exciting applications can be provided to vehicular users to enhance their road safety and travel comfort, and finally raise a complete change to their on-road life. As the content distribution and media/video streaming, such as Youtube, Netflix, nowadays have become the most popular Internet applications, to enable the efficient content distribution and audio/video streaming services is thus of the paramount importance to the success of the vehicular networking. This, however, is fraught with fundamental challenges due to the distinguished natures of vehicular networking. On one hand, the vehicular communication is challenged by the spotty and volatile wireless connections caused by the *high mobility* of vehicles. This makes the download performance of connections very unstable and dramatically change over time, which directly threats to the on-top media applications. On the other hand, a vehicular network typically involves an extremely *large-scale* node population (e.g., hundreds or thousandths of vehicles in a region) with intense spatial and temporal variations across the network geometry at different times. This dictates any designs to be scalable and fully distributed which should not only be resilient to the network dynamics, but also provide the guaranteed quality-of-service (QoS) to users.

The purpose of this dissertation is to address the challenges of the vehicular networking imposed by its intrinsic dynamic and large-scale natures, and build the efficient, scalable and, more importantly, practical systems to enable the cost-effective and QoS guaranteed content distribution and media streaming services to vehicular users. Note that to effectively deliver the content from the remote Internet to in-motion vehicles, it typically involves three parts as: 1.) *an infrastructure grid of gateways* which behave as the data depots or injection points of Internet contents and services to vehicles, 2.) *protocol at gateways* which schedules the bandwidth resource at gateways and coordinates the parallel transmissions to different vehicles, and 3.) *the end-system control mechanism at receivers* which adapts the receiver's content download/playback strategy based on the available network throughput to provide users with the desired service experience. With above three parts in mind, the entire research work in this dissertation casts a systematic view to address each part in one topic with: 1.) design of large-scale cost-effective content distribution infrastructure, 2.) MAC (media access control) performance evaluation and channel time scheduling, and 3.) receiver adaptation and adaptive playout in dynamic download environment.

In specific, in the first topic, we propose a practical solution to form a large-scale and cost-effective content distribution infrastructure in the city. We argue that a large-scale infrastructure with the dedicated resources, including storage, computing and communication capacity, is necessary for the vehicular network to become an alternative of 3G/4G cellular network as the dominating approach of ubiquitous content distribution and data services to vehicles. On addressing this issue, we propose a fully distributed scheme to form a large-scale infrastructure by the contributions of individual entities in the city, such as grocery stores, movie theaters, etc. That is to say, the installation and maintenance costs are shared by many individuals. In this topic, we explain the design rationale on how to motivate individuals to contribute, and specify the detailed design of the system, which is embodied with distributed protocols and performance evaluation.

The second topic investigates on the MAC throughput performance of the vehicle-to-infrastructure (V2I) communications when vehicles drive through RSUs, namely drive-thru Internet. Note that with a large-scale population of fast-motion nodes contending the channel for transmissions, the MAC performance determines the achievable nodal throughput and is crucial to the on-top applications. In this topic, using a simple yet accurate Markovian model, we first show the impacts of mobility (characterized by node velocity and moving directions) on the nodal and system throughput performance, respectively. Based on this analysis, we then propose three enhancement schemes to timely adjust the MAC parameters in tune with the vehicle mobility to achieve the maximal the system throughput.

The last topic investigates on the end-system design to deliver the user desired media streaming services in the vehicular environment. In specific, the vehicular communications are notoriously known for the intermittent connectivity and dramatically varying throughput. Video streaming on top of vehicular networks therefore inevitably suffers from the severe network dynamics, resulting in the frequent jerkiness or even freezing video playback. To address this issue, an analytical model is first developed to unveil the impacts of network dynamics on the resultant video performance to users in terms of video start-up delay and smoothness of playback. Based on the analysis, the adaptive playout buffer mechanism is developed to adapt the video playback strategy at receivers towards the user-defined video quality. The proposals developed in the three topics are validated with the extensive and high fidelity simulations.

We believe that our analysis developed in the dissertation can provide insightful lights on understanding the fundamental performance of the vehicular content distribution networks from the aspects of session-level download performance in urban vehicular networks (topic 1), MAC throughput performance (topic 2), and user perceived media quality (topic 3). The protocols developed in the three topics, respectively, offer practical and efficient solutions to build and optimize the vehicular content distribution networks.

Acknowledgements

The past five years of my graduate and research life in Waterloo is truly the most unique, precious and awarding time of mine. Along this way, there are so many people who deserve to be thanked for their friendship, help and supports offered to me. Within this limited space, I cannot hope to thank them properly.

First and foremost, my deepest and sincerest gratitude goes to my supervisor Professor Xuemin (Sherman) Shen. It is his guidance, support, encouragement, patience and genuine expertise in the past five years that have made this dissertation possible. What I appreciate the most of Professor Shen is his great patience and understanding to students. His enthusiasm and dedication to his work, his students and his family are really inspiring to me.

I would also like to thank the honorable members of my thesis committee, Professor Jiangchuan Liu, Liang-liang Xie, Sagar Naik and Penfeng Li, for serving as my thesis readers. Their insightful comments have significantly affected the substance and presentation of my work.

Doing research and spending day and night in the lab can be boring and daunting. My fellow students have made my life at the University of Waterloo a colorful and enjoyable experience. I wish to thank Ning Lu, Xiaohui Liang, Xiaodong Lin, Miao Wang, Jian Qiao, Xiaoxia Zhang, Zhongmin Zheng, Kuan Zhang, Hongbin Liang, Mahdi Asefi, Yongkang Liu, Hao Liang, Kuan Zhang, Dr. Rongxing Lu, Dr. Yuanguo Bi, Dr. Zhiguo Shi, Dr. Yipin Sun, Dr. Huangguan Shan, Dr. Ping Wang, Dr. Wei Song, Dr. Bong Jun Choi, Dr. Kuang-Hao Liu, Dr. Fen Hou and many others for their continuous encouragement, selfless help and all the good times we spent together. In particular, I wish to especially thank Dr. Lin X. Cai, Dr. Ping Wang, Dr. Jiming Chen, Dr. Xinhua Ling, and Dr. Fan Bai for their inspiring discussions and invaluable suggestions on my research.

The thesis is dedicated to my father Huabin Luan and my mother Aiping Jiang. I would not be completing my studies if they did not teach me the value of hard work and dedication. The thesis is also dedicated to my girl friend Sanying Li. I owe them everything, and fear I cannot love them enough in return for that. Thanks to them all for their continuous and ever-caring support which made me always feel their presence so near to me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

The advent of advanced communication technologies in the past decades has ushered the era of information age. To date, the high-rate instant access to the media-rich information (e.g., online gaming and video on demands) and ubiquitous communications to each other (e.g., emails and social networks) have already become an integral part of our daily life. As reported in [2], the average Canadian spends 18 hours a week on the Internet, which has outpaced the time spent on watching TV.

Although vehicles nowadays represent the third place, after home and office, where a regular citizen spends more time daily, the high-rate and ubiquitous information access and communications in vehicles are still not available or very expensive in most regions of the world. As reported in [3], Americans spend 15 hours on average in a car each week. This is to say, in around 8.9% of a day, people would have very limited or no Internet access at all. The surge demand of ubiquitous access thus drives the vehicular communication ever more important.

Catering to the ever-increasing demand of vehicular communications, the vehicular networks (alternatively known as vehicular ad hoc networks (VANETs)) have recently been proposed. As shown in Figure 1.1, by equipping vehicles with the on-board wireless transceivers and computerized control modules, and meanwhile deploying the Internet gateways on the roadside (namely roadside unit (RSU)), the vehicles are able to communicate and connected as a network in two modes: the inter-vehicle communication (alternatively known as vehicle-to-vehicle (V2V) communication) for data message exchange among peer vehicles, and the vehicle-to-RSU (V2R) communication to provide the Internet access to vehicles with the assistance of RSU.

Over such a platform, three thrusts of applications can be provided to vehicles on the move. The first thrust is related to the road safety. By exchanging and propagating the road traffic information among vehicles using the V2V communications, drivers can be effectively notified on the traffic conditions and accordingly response timely to avoid the possible accidents and other extreme consequences. Figure 1.2 shows examples of the safety applications including the erratic lane changing, intersection collision warning, etc. The second thrust is as a key enabling technology of the Intelligent Transportation Systems (ITS). For example, using the vehicular networks, real-time road traffic information can be delivered to vehicles, and guide drivers to optimize their driving routes to avoid traffic jams and reduce the travel time. By installing wireless facility on the traffic lights and making them communicate with the driving through vehicles, adaptive traffic scheduling can be realized to enhance the road capacity with the smooth traffic flow. The third thrust is on the infotainment applications as a means to enhance the travel comfort of passengers. For example, through the V2V communications in proximity, passengers on the board can transfer files among each other, chatting online or playing games on the road. By connecting to the Internet through RSUs, travellers can surf the web, watch Youtube videos and access various media-rich applications just as they have at home. To summarize, the vehicular networks can make our travel more safe, pleasant and efficient.

Motivated by the significant commercial and social benefits brought by the vehicular networks, much standardization and research effort has been undertaken worldwide in the past decade to enable vehicular communications. Notably, the U.S. Federal Communications Commission (FCC) has allocated 75 MHz of spectrum in the 5.9 GHz frequency band exclusively used for the Dedicated Short Range Communications (DSRC) on vehicular networking. Over the DSRC spectrum, the IEEE 802.11 standard body has proposed a new amendment, IEEE 802.11p, to support the Wireless Access in Vehicular Environment (WAVE). Meanwhile, prominent automobile corporations in the worldwide have also lunched important projects to promote vehicular broadband communications. For instance, Mercedes-Benz proposes to deploy the “InfoFuel” stations along the roads to fuel on-road vehicles with the high throughput Internet access using the IEEE 802.11a radio [4]. General Motors launched the “OnStar” to provide the subscription-based communications and in-vehicle security on the move [5], and Toyota builds “Toyota Friend” to facilitate the social communications among owners of Toyota cars [6].

In this dissertation, we mainly focus on the third thrust of VANET applications. In particular, our goal is to develop the practical and efficient system and network protocols

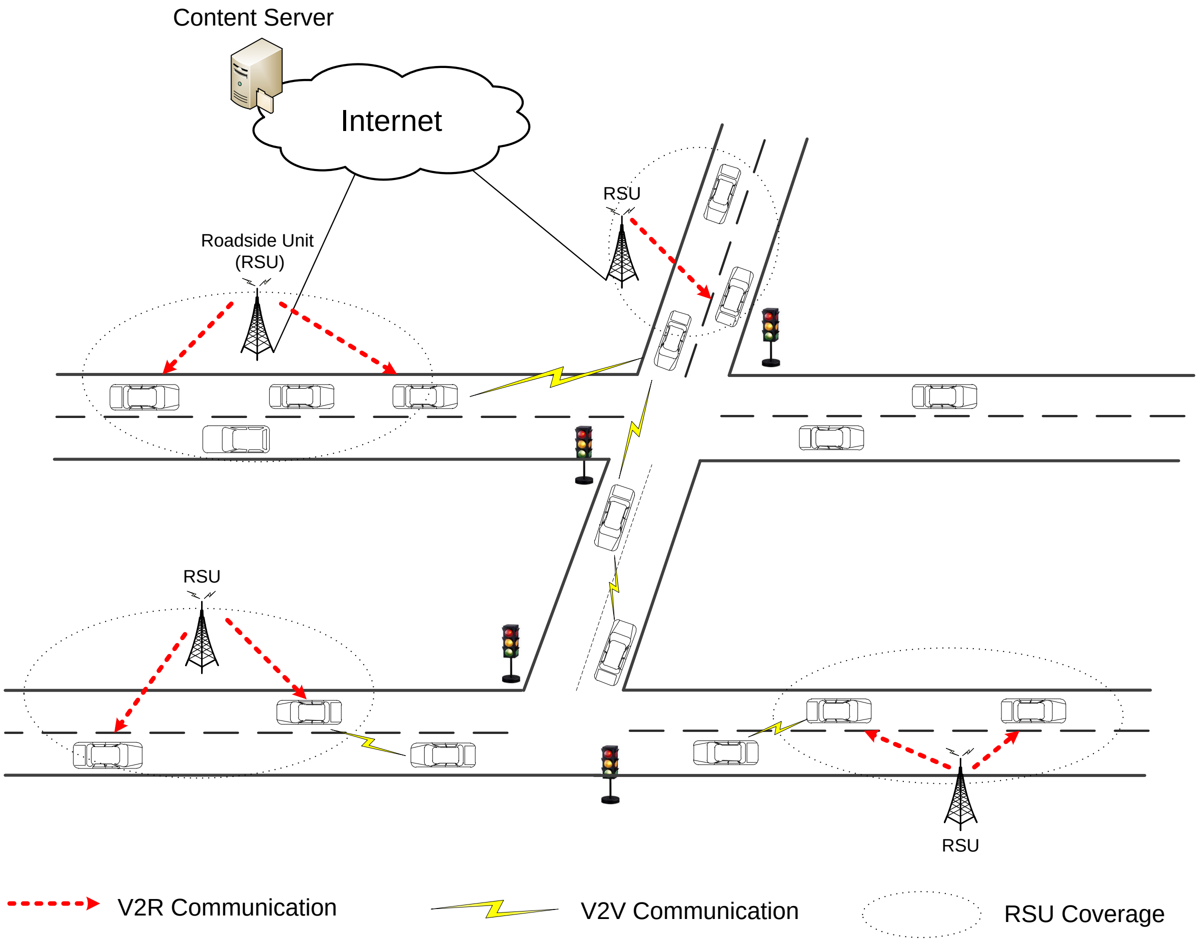
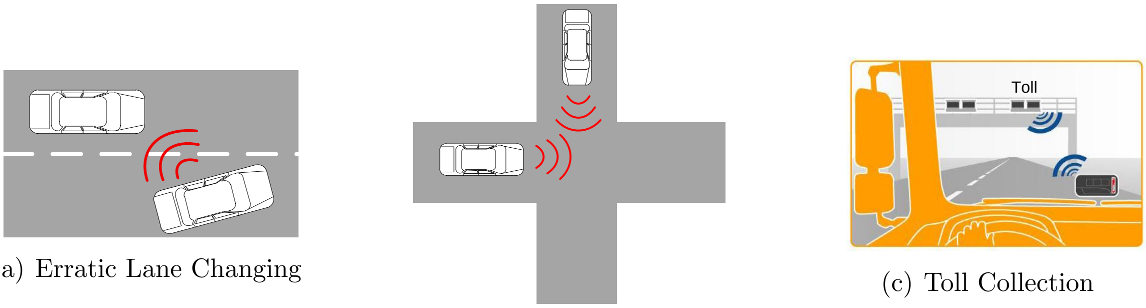


Figure 1.1: Vehicular communication on the road



(a) Erratic Lane Changing

(b) Intersection Collision Warning

(c) Toll Collection

Figure 1.2: Application examples of VANET

to provision the quality-of-service (QoS) guaranteed media content distribution services[1], such as live video streaming, content file distributions to vehicles on the move.

1.2 Research Challenges

Although with a bright future ahead, the way to efficient content distributions in VANETs is still fraught with fundamental engineering challenges. This mainly attributes to the following two distinguished features of VANETs.

1.2.1 High Node Mobility

In VANETs, each communication node is a fast-moving vehicle. As a result, the V2V and V2R communications are highly violative and susceptible to frequent interruptions with the transient contact durations among nodes. For example, [7] investigates on the download bandwidth of vehicles from the unplanned open residential WiFi access points in Boston. It is shown that the plethora IEEE 802.11b access points deployed in cities could provide vehicle nodes with the intermittent and short-lived connectivity, yet high throughput when the connectivity is available. The transient and intermittent download connectivity of vehicles inevitably renders significant impairments to the on-top content distribution applications, such as live and on-demand video streaming to vehicles.

[1] Rather than focusing on the packet level performance, the key of content distribution network design is to guarantee the session level performance of users, e.g., download delay of whole content. This will be elaborated in Chapter 2

Moreover, the dramatic changing connectivity and locations of vehicles lead to the dynamic network topology. This dictates any content distribution protocols to be resilient to the topology change and can self-heal quickly. Note that the mobility of vehicles are pertained to the road layout and presents specific patterns in different road environments. For example, the trajectory of vehicles on the highway is one-dimensional, and it is two-dimensional in the urban areas with dense street intersections. The heterogeneity and diversity of the vehicle mobility and topology patterns in VANETs thus indicate that there is no "one-for-all" solution to form the content distribution networks. Instead, to address the network dynamics and heterogeneities, the key is to *exploit the specific mobility patterns of vehicles in different deploying environments, and accordingly adapt the content distribution system and strategies to the particular topology patterns.*

1.2.2 Large Network Scale

The VANETs are typically of large-scale involving several hundreds or even thousands of nodes communicating at the same time, such as all vehicles on one highway section, in the downtown area of a city, or even in a region of multiple cities. Therefore, *the content distribution networks must be scalable to the network size.*

Moreover, due to the fast mobility of vehicles, the node density of the network presents apparent spatio-temporal variations. As reported in [8], by analyzing the real-world traces collected in Berkeley, California and Toronto, Canada, it is shown that highway traffic drastically changes over time (of different scales) across geographic locations. Also, it is intuitive that within a city, the vehicle density in the downtown area is typically greater than that in the outskirts. Even on the same road segment in a city, the vehicle density could fluctuate dramatically over time and is affected by the nearby traffic conditions. Therefore, *a practical design of the content distribution system needs to self-organize and be adaptive to the spatial- and time-varying road traffic.*

1.3 Our Approach and Contributions

In this dissertation, we report our solutions to provision QoS controlled vehicular content distribution services by addressing the aforementioned challenges. As shown in Figure 1.1, in a typical content distribution network, the content files are retrieved from the media servers residing in the remote Internet and then delivered to vehicles through the RSUs (or gateways/data depots/access points) along the trajectory of vehicles. Note that with random mobility, vehicles tend to have random connectivity to RSUs, and need to share the RSUs with many other vehicles simultaneously, the service quality perceived at the receiver is affected by multiple ingredients simultaneously, including the availability of RSUs, available resource at each RSU and the specific requirements of on-top applications. An integrated system design thus needs to address all the aspects. From this perspective, we generally divide the design of content distribution network into three research topics, i.e., infrastructure formation, RSU resource allocation and receiver adaption, and address each topic in one chapter as follows:

1.3.1 Practical Solution on Constructing Content Distribution Infrastructure

In the first topic, we propose a practical solution to build the large-scale infrastructure for the content distribution in the urban vehicular network. We claim that the dedicated infrastructure with reserved computing, capacity and buffer resource is a must to provide the QoS guaranteed content distribution services (e.g., news, advertisement broadcasting) to users in the urban. However, to form the large-scale infrastructure is a daunting task, if not impossible, due to the intensive investment cost and the dirty installation and maintenance work involved. In this thesis, we propose VTube, a cost-effective and large-scale infrastructure for content distribution to urban vehicular networks. In specific, VTube relies on the distributed low-cost storage buffers, namely roadside buffer (RSB), installed in the city facilities, such as stores, museums, cafeteria, etc., to provide content distribution services to vehicular users. Each RSB has a local buffer storage, and is equipped with a wireless transceiver which can communicate with the vehicles driving through its communication coverage and upload the stored files to vehicles. It is important to note that the RSBs spread in the city are installed, manipulated and maintained by distributed owners, such as grocery stores, movie theaters, etc., without the involvement of any central controller. This is enabled by two reasons. First, RSBs are low-cost devices which can be managed using wireless communications and are not connected to the Internet for any charges of bandwidth. Second, the RSBs bring benefit with the commercial usages such as advertisement distribution. RSBs across in the city are designed as an integrated whole to form VTube to serve the content distribution applications to vehicles. In this thesis, we unfold the design of VTube by first presenting the detailed design principles and practices of VTube. Given the content popularity and capacity of the buffer storage, we then develop a mathematical model to evaluate the mean download delay of contents for vehicular users. Using the estimated delay as an input, we formulate the content replication problem in RSBs as a stochastic programming problem to attain the mean system-wide minimum download delay. Finally, we propose a fully distributed random walk based algorithm to solve the optimization problem. Using extensive simulations, we demonstrate that VTube can minimize the download delay of users because of the exploitation of vehicle mobility and distributed buffer storage at different locations.

Contributions: The contributions of this research are two-fold. *First*, VTube is a very practical solution on the headache issue of vehicular infrastructure construction. It involves distributed entities to share the hardware cost, installation and acquaintance work of devices. Moreover, the formed infrastructure is extensible to serve different types of mobile users, such as smartphone, tablet users, etc., and behaves as a generic platform to support a variety of content distribution applications, such as news broadcast, road traffic monitor and information collector. *Second*, we propose the distributed algorithm for the content replications in RSBs. Thanks to the proposed algorithm, RSBs in VTube can be installed and managed distributedly and VTube can be additively formed with the newly deployed RSBs easily merged to the existing infrastructure.

1.3.2 MAC Performance Evaluation

In the second topic, we investigate on the MAC throughput performance of the V2R communications. In specific, motivated by the pervasive adoption and great success of WiFi, the newly proposed IEEE 802.11p standard adopts the contention-based IEEE 802.11e EDCA (Enhanced Distributed Channel Access) scheme as its MAC. Originally designed for the static indoor applications, the throughput performance of IEEE 802.11 MAC protocol in the outdoor vehicular environment is however unclear and arguable, especially in the scenario with a large number of fast-moving nodes transmitting simultaneously. In this thesis, we develop a mathematical model to evaluate the throughput performance of IEEE 802.11b DCF (Distributed Coordination Function) scheme in the highly mobile vehicular networks. The developed model is based on a three dimensional Markov chain model which incorporates the high node mobility with the modeling of DCF. By solving the Markov chain numerically, we show the impacts of vehicle mobility (characterized by node velocity and moving directions) on the resultant nodal and system throughput, respectively. Based on the developed model, we demonstrate that the throughput of DCF may reduce with the increasing node velocity. This is because that with the high mobility, distance between vehicles and RSUs change dramatically, resulting in the fast changing transmissions rates. However, the legacy DCF scheme cannot adapt quickly enough in tune with the changing transmission rates and hence fail to fully utilize the transient high throughput connectivity of vehicles. As a remedy to that, we propose three amendments to DCF to adaptively adjust the DCF parameters according to the node mobility. Using extensive simulations, we validate the accuracy of the developed analytical model and show the effectiveness of the proposed amendments.

Contributions: The contributions of this research are two-fold. *Firstly*, this work represents the first study in literature to model the impact of node mobility on the MAC performance of V2R communications. By showing the resultant system throughput performance at different vehicle traffic conditions, the work provides insightful lights on the real-world deployment of the V2R communications. Moreover, using the proposed analytical model, the optimal setting of DCF parameters can be determined. This enables the optimal and adaptive configurations of DCF according to the vehicle traffic flow. *Secondly*, this work proposes effective amendments to DCF to improve its performance in the outdoor dynamic environments.

1.3.3 Adaptive Video Playback

Networked video streaming, which represents the most advanced content distribution service over Internet, has achieved tremendous success in the past decade. However, to delivery the high-quality video services to vehicles is still a very, if not the most, challenging task. This is because that with the limited and intermittent connectivity of vehicles, the download rate tends to be very low and intensively fluctuating all the time which can hardly afford the strict QoS requirements of the on-top video applications. As an effort to address this issue, in the last topic of this thesis, we propose the optimal design of the video receiver at the user end to enable the user-oriented smooth video playback at vehicular networks. To be specific, we first established an analytical framework to investigate on the evolution of playout buffer at the receiver at dynamic networking environment. Notably, the playout buffer is deployed at the receiver to absorb the dynamic packet arrivals, and guarantee the smooth video playback. This is done by holding packets locally and meanwhile freezing the video playback for a short period until a threshold amount of packets is reached before the video playback initiates. With packets dynamically arriving to the playout buffer and then drawn for the application use by the upper layer video decoder, the playback can sustain as long as the playout buffer is non-empty. That is to say, the queue length of the playout buffer is an indicator of the smoothness of video playback. As such, by analyzing the queue length of the playout buffer, we evaluate the user perceived video quality in terms of video start-up delay and playback smoothness. In specific, we consider two cases of finite and infinite playout buffers and model them as the $G/G/1/N$ and $G/G/1/\infty$ queues, respective, with N denoting the buffer size. We then apply the diffusion approximation and derive the close-form expressions of the queue length distribution in the aforementioned two cases, respectively. Based on these results, we show the expressions of the media start-up delay and probability of media playback frozen, represented by the mean and variance of download and playback rates. Based on the developed analytical framework, we further propose the adaptive playout buffer management schemes to optimally manage the threshold of video playback towards the maximal user utility, according to different quality-of-experience (QoE) requirements of end users. The proposed framework is finally validated by extensive simulations.

Contributions: The primary contribution of this research is to establish a mathematical mapping framework between the network QoS metrics and user perceived performance, i.e., QoE. In specific, from the networks perspective, the resource allocation of video services is dimensioned by the QoS metrics such as throughput, packet loss rate, packet transmission delay and delay jitters, etc. Those network QoS metrics, however, are not directive to characterize the user perceived quality, which is referred to as QoE in literature. In order for the networked video streaming to achieve the optimal QoE, the foremost issue is therefore to establish a mapping between the QoE metrics and network Qos metrics. Our major research goal in this part is to provide a simple, accurate and directive mathematically translation between QoS and QoE. Using the developed framework, we then showcase how to use this translation to 1. design the receiver and playback strategies to achieve the user desired QoE provided the network throughput performance, and 2. design the network resource allocation protocols given the QoE, respectively.

1.4 Thesis Outline

The rest of this thesis is organized as follows: Chapter 2 introduce the background knowledge and basic concepts of VANETs. Chapter 3 presents a practical and cost-efficient solution on building the large-scale infrastructure for vehicular content distribution. Chapter 4 presents the model of the DCF MAC in the drive-thru Internet scenario and propose the enhancement mechanisms to improve the performance DCF in the highly mobile vehicular networks. Chapter 5 investigates on the impacts of the network dynamics on the video quality perceived by end users, and accordingly propose the optimal receiver adaption scheme to achieve the smooth video delivery in the vehicular networks. Finally, Chapter 6 concludes the thesis, and points out our future research directions.

Chapter 2

Background

This chapter first introduces the basic concepts and background knowledge of the emerging vehicular networks and the content distribution applications. After that, an overview and comparison of the existing works and technologies on the design and prototyping of vehicular content distribution networks are provided.

2.1 Enabling Technologies of VANETs

2.1.1 Key Components of Vehicular Communications

Since the first stream-powered vehicle was invented for human transportation in 1769, the vast and steady technological innovations over centuries have finally made vehicles evolve as an advanced, integrated and intelligent system on the road. As a result, users have acquired the persistently enhanced safety, comfort and convenience during the travel than ever before. The newly emerged vehicular networking technology further ushers in a new era of automobiles in which vehicles are enabled with wireless communications and connected as an integrated network on the wheel. Figure 2.1 shows the enabling components needed for vehicular communications based on the smart vehicle model described in [1]. It includes:

▷ Communication facility, also called on-board unit (OBU), to conduct wireless communications among vehicles and RSUs. The OBU operates on the DSRC radio which is specifically designed for short-range fast-motion vehicular communications. The details of DSRC will be discussed later in this chapter.

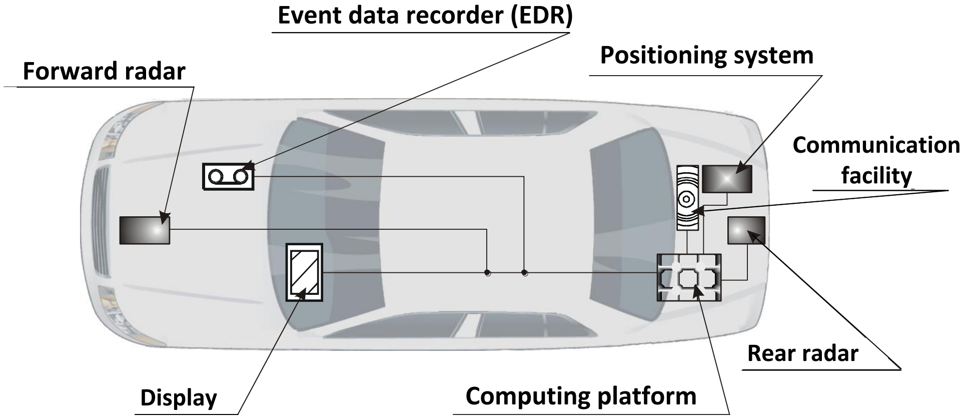


Figure 2.1: On-board facilities for vehicular communications [1]

- ▷ Positioning System, such as a GPS receiver which can timely and accurately identify the location information, such as latitude, longitude and velocity of vehicles.
- ▷ Sensors, which include the autonomous ranging sensors such as the front and rear radars, cameras, and on-board sensors such as steering wheel angle sensors and wheel speed sensors. The vehicular radar, such as the microwave, infrared and ultrasonic radar, plays an important role for road safety by detecting and tracking the non-communication vehicles or obstacles. The on-board sensors are important to assist the efficient communications. For example, they can be used for motion and trajectory prediction and accordingly provide valuable information to vehicular network formation and optimization.
- ▷ Computing, display and event data recorder facilities to enable on-board data processing, input/output and tracking to facilitate the vehicular communications.

Many of the above components, such as GPS, radars and on-board sensors, have already been mechanized and widely deployed in vehicular transportation[1]. In the near future, it can be envisioned that vehicles with the omni-capability of sensing, computing and communication will be pervasive. VANET will connect them as a ubiquitous network on the road.

[1] See "Technology on XTS, ATS Can Help Avoid Crashes", GM, 2012.3.27

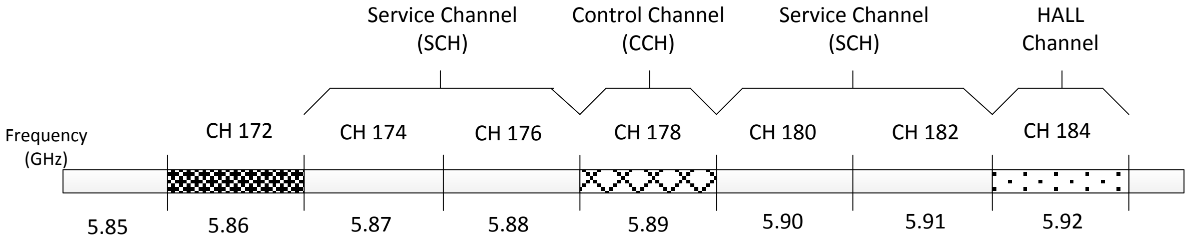


Figure 2.2: DSRC channel spectrum

2.1.2 Standardization of VANET

In 1999, the U.S. FCC allocated the 5.850 to 5.925 GHz frequency band for the ITS applications. The 75 MHz DSRC spectrum is further divided into seven channels as shown in Figure 2.2. Among the seven DSRC channels, the fourth channel (CH 178), namely control channel, is exclusively used for the purpose of control, coordination and safety message broadcasting. The first channel (CH 172) is unused and the last (high availability low latency) channel (CH 184) is left for future use. The other four channels, namely service channels, are for IP-based infotainment applications.

The operation of DSRC-based VANET communications is standardized by the WAVE (wireless access for vehicular environment) protocol suite [9] as shown in Figure 2.3, where the IEEE 802.11p standard specifies the MAC and physical layers and the IEEE 1609 protocol suite specifies the upper-layer operations, as

- ▷ IEEE 1609.1 (WAVE Resource Manager) that describes the key components of WAVE system architecture and defines data flow and resource management protocols,
- ▷ IEEE 1609.2 (Security Services) that covers methods for securing WAVE management and application messages,
- ▷ IEEE 1609.3 (Networking Services) that defines network and transport layer services, including addressing and routing, etc.,
- ▷ IEEE 1609.4 (Multi-channel Operations) that defines the operations of the wireless transceiver over the multi-channel DSRC radio,

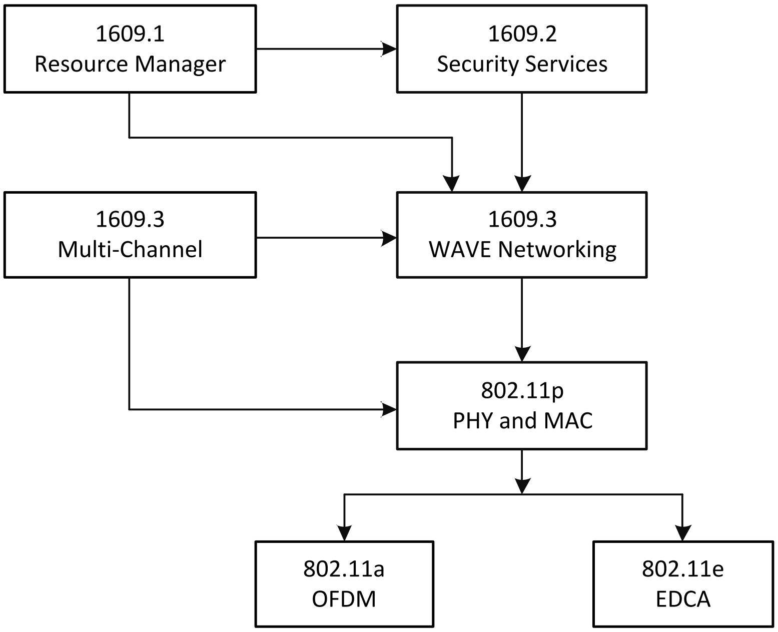


Figure 2.3: WAVE protocol suite

- ▷ IEEE 802.11p (MAC and Physical protocols) that specifies the MAC and physical layer operations of WAVE. The IEEE 802.11p protocol adopts the OFDM (Orthogonal Frequency Division Multiplexing) scheme in the physical layer which is similar to IEEE 802.11a. IEEE 802.11p targets a transmission range between 300m and 1km. It can provide the data transmission rates of $9, 12, 18, 24,$ and 27 Mbps to vehicles at the velocity of $0 - 60$ Km/hr, and $3, 4.5, 6, 9,$ and 12 Mbps at the vehicle velocity of $60 - 120$ Km/hr. In the MAC layer, the contention-based IEEE 802.11e EDCA (Enhanced Distributed Channel Access) scheme is adopted to support the QoS differentials of communications.

2.1.3 Real-World Evaluation of VANETs

As VANET is still in the infancy stage, a rich body of research has been developed to validate the effectiveness of wireless communications in the presence of high vehicular mobility.

Ott et al., [10] report the first measurement of the communication between a moving car with an external antenna and the roadside Wireless LAN access point (AP), namely drive-thru Internet. They show that using the off-the-shelf IEEE 802.11b device, a vehicle could maintain a connection to a roadside AP for around 600 meters and transfer 9MB of data at the velocity of 80km/h using either TCP or UDP. Moreover, it is observed that the data transmission rate in each drive-thru presents a bell-shape curve with the very low rates when the vehicle enters and exits the AP coverage and transient high rate when the vehicle is close to the AP. Accordingly, the communication session within each drive-thru can be divided into three phases: entry phase, production phase, and exit phase, each of around 200 meters in their experiments.

Bychkovsky et al., [7] have extended the experiments of drive-thru Internet in the metropolitan area. In the experiment, they make nine vehicles driven under the normal traffic conditions in the Boston metropolitan area, where vehicle is equipped with a wireless enabled computer and communicates with the "open" residential WiFi AP during the driving. Through the analysis of measurement data collected in over 290 "drive hours", they report that the vehicles can successfully establish the AP connection and transmit data at the velocity range from 0 to 60 km/h, with the mean duration of connections to be 13 seconds and the mean duration between connections to be 75 seconds. Therefore, this experiment verifies the feasibility of communications between the moving vehicle to roadside AP regardless the velocity in the urban. Moreover, they show that using the current IEEE 802.11 protocol, it takes several seconds to setup the connection, which is intolerably long and significantly reduces the effective communication time of the connection. As a remedy to that, they propose an IP address caching scheme which can effectively reduces the AP association delay by by-passing the DHCP.

2.2 Vehicular Content Distribution Networks

The term "content" refers to general digital files, such as text messages, audio/video files, etc., which can be broadly classified as *small-* and *medium-size contents* of several or tens of megabytes, such as music files, picture images and text messages, and *large volume contents*, such as video files of hundreds megabytes or several gigabytes. Based on its playback pattern at the receivers, the content distribution applications can be categorized into two groups as:

▷ *play-after-download*, in which the playback of the contents commences after the entire copy of the content file has been downloaded in the local cache

▷ *real-time streaming* (or play-when-downloading) where the content or media playback starts when the content is still downloading.

The play-after-download pattern can provide the best media playback quality to end users in terms of the visual quality and playback smoothness, as the intact content files have already been stored locally; nevertheless, it typically incurs the long start-up delay to users who has to wait for the file download. The real-time streaming shortens the annoying start-up delay with immediate media playback, but has the risk of frequent freezing of playback and severe video distortions due to the variable network delays and unpredictable packet losses.

The selection of the media playback pattern should not only cater to the specific content file size, but also conform to the available network throughput. For example, the playout of large volume contents should adopt the real-time streaming pattern in most scenarios to avoid the overly long start-up delays. Nevertheless, the medium-size contents, such as music files, should typically apply the download-then-play pattern for the guaranteed playback quality at the cost of modest start-up delay. However, it is important to note that the vehicular networks are highly dynamic with intermittent connectivity and intensively changing throughput. For example, as vehicles moving on the road with varying traffics and density of RSUs, the download throughput may change dynamically. As such, the content playback should adapt to the specific physical environments and download performance of vehicles along their trajectories. In Chapter 5 of the dissertation, we provide an analytical framework to guide the adaptive content playout with variable download throughput.

It is also important to note that unlike the traditional research which largely focuses on the packet-level performance, such as the packet losses and packet delays during the communication, the emphasis of content distribution network design is to guarantee the session-level performance and optimize the user's perceived performance subject to the application specific QoS or QoE requirements. For example, in the real-time video services, the design goal is to achieve the smooth video playback with modest start-up delays and high visual quality within the entire session of video playout. In the case of small content distributions, the goal is to reduce the download delay of the entire content, whereas the instantaneous transmission delays of individual packets are not that critical.

Moreover, in the context of vehicular networking, the content distribution network design should be tailored and adaptive to the specific deployment environments. Specifically, different deployment scenarios, such as city urban, rural, highways, have entirely different features in the road traffic and communication patterns. The approaches of content distribution in different scenarios should therefore explore the specific features in different deployment environment. For example, in the urban area, there typical exist far more infrastructure devices, such as cellular base stations, WiFi access points, than other districts, like the rural and highways. As such, the key is to explore the opportunistic high-rate V2R connections to achieve the QoS guaranteed and cost-effective content distribution services to vehicles. In the highway scenario, although the infrastructure is typically sparsely deployed in order to save the deployment cost, as vehicles are moving in the one-dimensional linear topology with the relatively static mobility pattern, it is more likely to form the long-lasting V2V connections among vehicles and accordingly enable cooperations among vehicles to collaboratively distribute the contents. However, with the diverse mobility behaviors (such as velocities and acceleration/decelations) in the highway, how to identify the long-lasting V2V connections and utilize the diversity of vehicle mobilities are the key to form effective cooperations for content distributions on the highway.

2.3 Related Works on Vehicular Content Distribution

Content dissemination in vehicular networks has recently attained considerable attention in a vast literature, ranging from file sharing [11, 12], data ferrying [13], sensing [14] to media streaming [15]. In what follows, we survey the recent works based on the architecture of the content distribution network from two perspective: infrastructure-less content distribution and infrastructure-based content distribution.

2.3.1 Infrastructure-less Content Distribution

Nandan et al. [11] introduce the first V2V content distribution protocol, namely SPAWN (swarming protocol for vehicular ad-hoc networks), to enable the cooperative content retrieval and sharing among vehicles. In SPAWN, a file is chopped into multiple pieces and swapped among vehicles in a BitTorrent (BT) style. Using the proximity-driven piece selection, SPAWN exploits the location information of vehicles for piece selection and download which can outperform the traditional rarest first scheme used in BT.

Targeting to the same application scenario, Li et al. propose CodeOn [16], which applies the network coding to facilitate the multicast content distribution on the highway. In CodeOn, the interest contents are first encoded using the symbol level network coding, which not only enhances the diversity of content blocks, but also makes the system robust to the severe packet loss due to the harsh wireless channels. Within this framework, according to the amount of useful content blocks each vehicle has, an optimal relay selection protocol is presented to maximize the downloading rate of the system. [17, 15, 18] also investigate on the network coding based vehicular content distribution network.

Zhang et al. propose Roadcast [19] in which two functions are provided to facilitate the efficient V2V content distribution. Firstly, a vector space model (VSM) based content search is enabled to help vehicles hunt for the interested files from the ocean of content files stored in the distributed vehicles. Secondly, a popularity-aware content dissemination protocol is proposed to minimize the integrated download delay of files. Specifically, assuming the fixed amount storage in the network, [20] shows that the minimum download delay of files is achieved when the number of file copies in the network is propositional to square-root of their popularity in the system steady state. In the light of [20], Roadcast assigns each file a cost value which is evaluated based on the popularity, size and estimated download delay. According to its cost value, files are evicted from the node buffer with different priorities, and finally the network is shown to converge to a stable state following the "square-root" law as in [20].

Zhang et al. propose V-PADA [21] a vehicle-platoon-aware scheme to utilize the clustering/platoon effect of vehicles on the highway for content dissemination. In specific, it is shown in [22] that vehicles on the highway tends to form clusters on the road. As such, vehicles within the same cluster can cooperative store contents and collaboratively distribute contents. Inspired by this idea, V-PADA deploys the stochastic time series analysis to predict the movement of vehicles in the same platoon and based on the mobility of vehicles to determine the content replications. As such, contents are optimally replicated at distributed vehicles in a stable platoon to minimize the download costs of vehicles in the platoon.

2.3.2 Infrastructure-based Content Distribution

Yuen et al. [23] investigate the non-cooperative content sharing in a mobile Infostation system[2] where a file is swapped among vehicles via social contacts: file is exchanged only when nodes upon the opportunistic contact have mutually interested contents.

Zhang et al. [25] develops a content scheduling algorithm at RSUs to provide different service priorities based on the data size and service deadline. In this case, given two requests with the same deadline, the vehicle which requests for a smaller amount of data will be served first. Given two requests asking for data with same size, the vehicle with an

[2] The concept of Infostation is early proposed in 1990s [24] which serves the same purpose as RSUs to provide low-cost short-range communications to vehicles.

earlier deadline will be served first. The algorithm is shown to outperform other scheduling algorithms, such as first-come-first-serve, first-deadline-first, and smallest data size first algorithms, to fully utilize the transient connection time of V2R communications.

Nandan et al. [26] describes AdTorrent to distribute the advertising contents pertaining to a local area. In AdTorrent, static wireless digital billboards are deployed on the roadside, continually pushing the advertising contents, e.g., hotel virtual tours, movie trailers, etc., to the driving through vehicles. Among the vehicles, the advertising contents are then swapped in a BitTorrent style similar to SPAWN.

Huang et al. [27] propose to deploy buffer storages in the city to enable the content retrieval of vehicles. Without involving the V2V communication, they seek the optimal content replications along the path of vehicles to maximize the delivery ratio of contents for each drive. On achieving this goal, it is assumed that the path information of vehicles are given as an input to a centralized system for a suboptimal solution.

Chapter 3

Engineering Distributed Infrastructure for Large-Scale Cost-Effective Content Dissemination over Urban Vehicular Networks

3.1 Introduction

Although the large-scale distributions of multimedia and user-generated contents (e.g., Youtube, MySpace, etc.) have already become the most popular Internet applications[1], low-cost ubiquitous content distribution services to mobile users on-board vehicles are however still a far dream to achieve. This not only attributes to the prohibitive bandwidth capacities required by the content distribution applications, but is also caused by the high bandwidth cost of traditional access networks. Notably, using the centralized cellular networks, not only the aggregate bandwidth per user is very limited with the need to serve a large group of users simultaneously[2], but also the incurred cost per user is high. To provide the low-cost broadband Internet access on the move, the city-wide WiFi is a plausible solution. According to a recent survey conducted by Devicescape [29], among more than 2700 respondents all over the world, 81% smartphone users prefer using WiFi over 3G for

[1] Youtube, the most popular video sharing site, features over 40 million videos and attracts around 20 million subscriptions per month.

[2] Customers of AT&T are reported to encounter dropped calls, spotty service, and delayed text/voice messages as iPhone overloads the network [28].

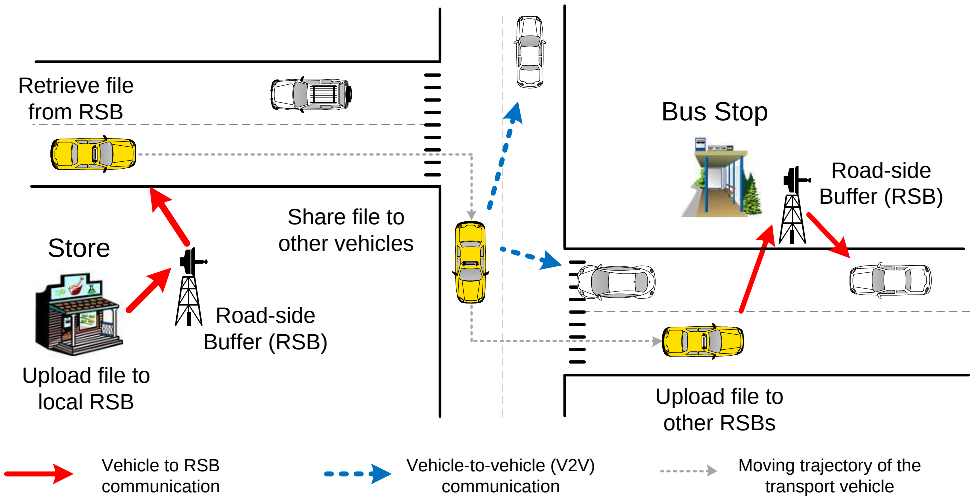


Figure 3.1: Content distribution through VTube

data services; 91% WiFi users expect WiFi while on the road and 79% respondents believe that WiFi should be free. However, the city-wide WiFi is still far to realize which is fraught with technical [30], societal and political issues [31]. To summarize, *a practical, efficient yet low-cost content distribution infrastructure is desirable to assist the large-scale content dissemination applications to mobile users.*

In this work, we investigate on providing the *low-cost large-scale* content distributions to vehicular users[3] in the urban area. In particular, we question on *how to construct a scalable content distribution network in the city which could bypass any excessive and long-term physical installations and investments and, more importantly, incurs the minimum monetary expense on bandwidth to individual users*? On addressing this issue, we propose VTube, a fully distributed large-scale infrastructure to facilitate the low-cost content disseminations. VTube is composed of a multitude of wireless buffer devices deployed on the roadside, namely roadside buffers (RSBs). Each RSB is equipped with a wireless transceiver, operating on the dedicated short-range communication (DSRC) radio, to communicate with the nearby vehicles using the vehicle-to-infrastructure communications [32]; it can selectively retrieve content files from the drive-thru vehicles and at the same time upload the cached files to vehicles upon their requests.

[3] In this work, we slightly abuse the terms by using vehicular users, vehicles and vehicle nodes interchangeably to denote the mobile users on-board vehicles.

Consider a motivating example as shown in Fig. 3.1 that a grocery store intends to distribute its recent flyers to citizens using VTube. To do so, the flyers are first uploaded to one or multiple RSBs nearby the store. The RSBs are then responsible to distribute the content files (flyers) to vehicles driving through the area and let the vehicles spread the flyers to other RSBs and vehicles across the city. In this process, the key issue is how to enable the collaborative caching among RSBs and vehicles to fully explore the buffer storage of RSBs and the mobility of vehicles to provide the best download performance to users.

Unlike the conventional centralized infrastructure (e.g., cellular base stations), a unique feature of VTube is that the infrastructure (i.e., RSBs) is deployed, owned and managed by distributed entities, such as hotels, movie theaters, schools, etc., for the purpose to distribute their own content files. For example, hotels may buy and deploy RSBs in their parking lot to distribute the room photos to vehicles on the road. The movie theaters can use their RSBs to periodically distribute the latest movie tailors to the public. In other words, the formation of VTube relies on the contributions of many individuals in the city with the shared investment cost and maintenance work. This is enabled due to the following three features of RSB:

- Cheap and simple to install: RSBs are cheap and light-weight devices composed of a wireless transceiver and small buffer. They can be configured and managed using the wireless communications (similar to the management of WLAN routers), which avoids the complex cabling works. As RSBs are deployed to distribute the local contents generated by their owners, they are not necessarily connected to Internet. As such, once deployed, RSBs incur no bandwidth costs to their owners.

- Easy to manage: the content distribution and buffer management of RSBs are purely self-organized which adapt to the time-varying network conditions [8] (e.g., the density of vehicle traffic, buffer availability) and are tailored to meet the local download demands. Therefore, except to regularly update the content files in their RSBs using wireless connections, the owners are not involved in any further operations.

- Profitable: The RSBs can bring commercial benefits to their owners by distributing the advertisements or other store information to the public. The RSBs can also be used for localization purpose to help users locate the stores[4].

The RSBs distributedly deployed in the city then cooperatively and collectively form VTube. By relying the fast-motion vehicles to transport contents among RSBs, VTube

[4] By connecting to a RSB, users in proximity can get aware of the store running the RSB, and locate the store through GPS and digital maps which can be buffered in the RSB.

enables the city-wide content distribution services to vehicles. In this work, we primarily make the following two contributions:

- ▷ We propose VTube, a new network architecture for low-cost content distribution in urban vehicular networks. The proposed architecture is formed by distributed infrastructure devices contributed by individual owners; with distributed entities united to construct VTube and sharing the investment and maintenance costs, VTube provides rich storage and bandwidth resource to users for media content distributions in the city. While we primarily target on vehicular users in this work, VTube can be easily extended to provide content distributions to a variety of other mobile users with different wireless portable devices, such as smartphones, tablets, laptops, etc.

- ▷ We embody VTube with a fully distributed content replication algorithm such that RSBs are enabled to intelligently and distributedly cache contents towards the maximal system objective. The proposed algorithm is based on the random walk algorithm and is scalable to different network size. Using extensive simulations, we show that the performance of the proposed algorithm approaches to the centralized global optimal content replication.

The remainder of this chapter is organized as follows: Section 5.2 describes related works, by positioning the original contributions of VTube. Section 3.3 presents the system model and formulate the VTube design as an optimization problem. Section 3.4 discusses on the solution of the problem and presents VTube protocol. The performance of VTube protocol is verified in Section 3.5, and Section 5.7 closes the chapter with summary.

3.2 Related Works

This section provides a brief survey on the related works and highlight our contributions in the light of existing literature.

The vehicular content distribution networks can be broadly categorized in two groups as: V2V-based systems in which the content distribution mainly relies on the collaborations among vehicles using the V2V communications, and the V2R-based systems which mainly exploit the capacity and opportunistic contact of infrastructure gateways on the roadside for content retrieval.

3.2.1 V2V-based System

Nandan et al., introduce the first V2V-based content distribution protocol, namely SPAWN (swarming protocol for vehicular ad-hoc networks) [11], to enable the cooperative content retrieval and sharing among vehicles. In SPAWN, a file is first chopped into multiple pieces and then swapped among vehicles in a BitTorrent (BT) style to facilitate the collaborative download. Moreover, SPAWN exploits the location information of vehicles and use the proximity-driven piece selection to schedule the piece transmission, which is shown to outperform the traditional rarest first scheme of BT used in the wireline peer-to-peer networks. Within the similar framework, Lee et. al., propose CodeTorrent [33, 34] which deploys the network coding to maximize the mutual differences of content pieces stored in the nearby vehicles, and accordingly reduce the search delay and coordinations of piece transmissions. Unlike SPAWN and CodeTorrent, in VTube we mainly focus on the design of the content distribution infrastructure and develop the distributed content replication protocols in the proposed infrastructure system. However, note that we consider both V2V and V2R communications, the proposed infrastructure system in VTube can work complementarily with the V2V system described in [11, 33, 34].

Li et al. propose CodeOn [16] for efficient content distribution over highways vehicular networks. In CodeOn, the content files are chopped into blocks and encoded using the symbol-level network coding which makes the system robust to the lossy wireless channels. Based on the amount of useful content blocks stored in distributed vehicle, an optimal relay selection protocol is developed to distributedly select the relay nodes and forward the content blocks over highways to the receivers. Similar to CodeOn, [15, 18, 17] also investigate on efficient content distributions over highways using network coding, and [21] develops a platoon-based content distribution protocol. In contrast to CodeOn, VTube considers the content distribution in urban areas. Unlike that on highways, the V2V connections in urban tend to be much more dynamic due to the diverse and complicated mobilities of vehicles, and moreover the V2V communications has much lower capacity and smaller coverage due to the intense interferences caused by the high node density and shadowing and fading effects caused by complex building environments. In this case, we argue that to explore the infrastructure is crucial. However, the approach to deploy distributed infrastructure as proposed in VTube can also be extended on the highway vehicular networks.

Acer et. al., propose [35] a V2V-based content distribution network to deliver the non-real time content information in the metropolitan area using the bus network. By exploring the stable bus schedule, predictable bus mobility and temporary storage at bus stops, [35] proposes a routing protocol which takes the randomness of rod traffic into consideration to deliver a single-copy file from the source to destination. Unlike [35], VTube targets to multicast a variety of files to a broad vehicles. However, it is interesting to combine the bus network with VTube by deploying the RSBs at bus stops and exploring the predictable bus mobility for content disseminations.

3.2.2 V2R-based System

In [25], Zhang et al., develop a scheduling algorithm at distributed RSUs to manage the V2R accesses of vehicles for service differential content distribution. Note that vehicles have transient and limited connection time to RSUs and diverse data volume to transmit. [25] provisions high transmission priority to vehicles with less data or having urgent deadlines. The algorithm is shown to outperform other scheduling algorithms, such as first-come-first-serve, first-deadline-first, and smallest data size first algorithms, to fully utilize the transient connection time of V2R communications.

Nandan et al. propose AdTorrent [36] to facilitate the distribution of advertising contents pertaining to a local area. In AdTorrent, static wireless digital billboards are deployed on the roadside which continually push the advertising contents, e.g., hotel virtual tours, movie trailers, etc., to the vehicles in proximity. Among vehicles, the advertising contents are then swapped in a BitTorrent style similar to SPAWN.

[25] focuses the content schedule and service provision at a single RSU. [36] investigates on the content distribution over a small region without considering the collaborative caching between wireless digital billboards and vehicles. In contrast, VTube targets to support the content distribution infrastructure over a large region to a large-scale node population. To do so, it is key to intelligently and fully utilize the buffer resource of the infrastructure.

Trullols-Cruces et. al., proposes [12] to explore opportunistic contacts and cooperative download among vehicles to enhance the content delivery rate. In specific, to distribute a file to the receiver, multiple relay vehicles are selected to carry the content file from roadside gateways to the in-motion receiver. This is enabled by the analysis of node mobility and road traffic. Similar to [35], [12] focuses on the single-copy file delivery whereas VTube focuses on multicasting files to a group of interested users.

Huang et al. propose to deploy buffer storages in the city to enable the content retrieval of vehicles [27]. Without involving the V2V communications, [27] seeks the optimal content replications along the path of vehicles to maximize the delivery ratio of contents for each drive. On achieving this goal, it is assumed that the path information of vehicles is available an input to a centralized system for a suboptimal solution. VTube differs from [27] in three aspects. Firstly, [27] assumes that the mobility trajectories of vehicles are given and determine the content replications accordingly. VTube, however, assumes the random mobility of vehicles given the statistics of the connection time to RSBs. Secondly, [27] targets to maximize the number of content files that can be distributed to vehicles along their trajectories. VTube targets to maximize the global system objective (such as average download delay of users). Towards this goal, we develop the mathematical framework to evaluate the file download delay given the contact intervals of vehicles to RSBs. Lastly, [27] relies on the infrastructure for content distribution without the assistance of V2V communications. VTube allows vehicles to share the download files.

3.3 System Model and Problem Formulation

In this section, we present the system model, including the RSB modelling, mobility of vehicles and the utility function of vehicular users. Based on the system model, we formulate the system design as an optimization problem. The main notations used are summarized in Table 3.1.

3.3.1 System Model

Model of RSBs

We consider the city as a bounded region as an example shown in Fig. 3.2, where a set $\mathbf{R}$ of RSBs are randomly deployed. Note that with different building environments and diverse communication capabilities, RSBs at different locations would have different radio coverage. Within their communication coverage, we consider RSBs to have the same data transmission rate to vehicle nodes, denoted by C_{V2R}. In this work, we allow vehicles to communicate with each other to cooperatively disseminate the downloaded contents to each other. Let C_{V2V} denote the data transmission rate of V2V communications. Each vehicle is equipped with a single-radio transceiver and communicate to only one node at each time. We make $C_{\mathsf{V2R}} > C_{\mathsf{V2V}}$, and vehicles prefer to downloading from RSBs if RSB connections are available. This is a working assumption as reported in [37], the throughput of the V2V communication in a real-world measurement is less than one fifth of the throughput of the vehicle-to-infrastructure communication.

Table 3.1: Summary of Notations

Notations	Description
$\mathbf{R}$	Set of RSBs in the region of interest
$\mathbf{F}$	Set of files published for download in the region of interest
$\mathbf{P}$	Popularity profile of files, where each element p_i, $i \in \mathbf{F}$, represents the probability that a user subscribes to download file i upon each download request
$\mathbf{A}$	Availability profile of files, where each element a_i, $i \in \mathbf{F}$, represents the portion of users which have file i stored in the local buffer
$\mathbf{B}$	Caching profile of files, where each element x_i, $i \in \mathbf{F}$, represents the probability that file i is stored in a randomly selected RSB
C_{V2R}	Communication data rate between RSBs and vehicles
C_{V2V}	Communication data rate between vehicles
B_{R}	Buffer size of RSBs
$U(\cdot)$	Utility function of vehicular users
$\mathcal{U}$	Global system objective (overall performance of network to optimize)
r	Download throughput of vehicles when vehicles are outside the communication range of RSBs
R	Download throughput of vehicles when vehicles are inside the communication range of RSBs
n	Number of vehicles which are able to transmit to the tagged node, or contend the channel for transmission with the tagged node
κ_i	Number of blocks in file i
t_i	Mean download delay of file i
$1/\lambda$	Mean sojourn time of vehicles inside the communication range of RSBs
$1/\mu$	Mean sojourn time of vehicles outside the communication range of RSBs
$1/\delta$	Mean file block download time of vehicles outside the communication range of RSBs
$1/\gamma$	Mean file block download time of vehicles inside the communication range of RSBs
$\Gamma(m,k)$	Mean first passage time of Markov process from state (m,k) to state $(\cdot,\kappa_i)$
$\lvert\cdot\rvert$	Cardinality of set
$\langle\cdot\rangle$	Mean value of a random variable
$\mathrm{Var}(\cdot)$	Variance of a random variable

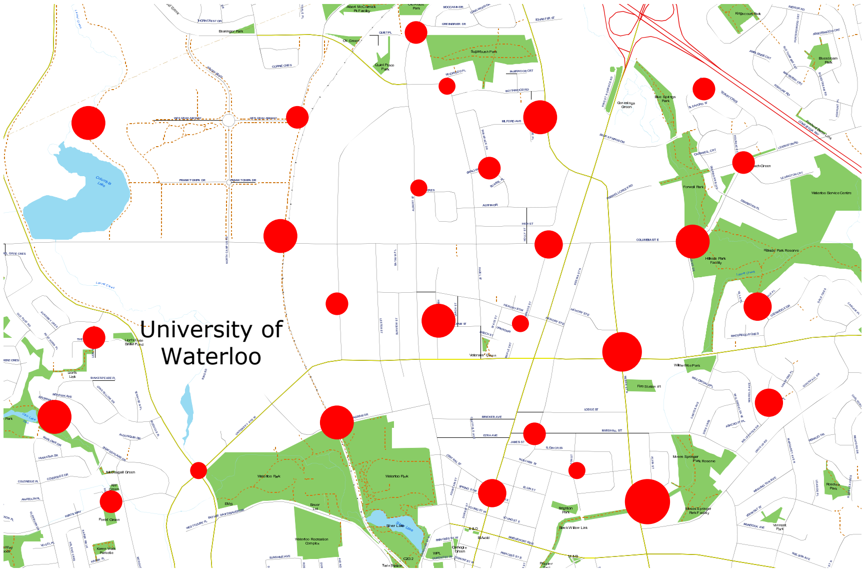


Figure 3.2: Example of the RSB deployments. RSBs are represented by red circles with diverse communication range due to the different communication environments.

Model of Node Mobility

The mobility of each vehicle node is represented by an *on-off* process based on its connectivity to RSBs: a vehicle node is in state 0 if it is outside the coverage of any RSB; otherwise, it is in state 1. Due to the random radio coverage and deployment locations of RSBs, we model the sojourn time of vehicles in state 1 and state 0 by the unpredictable, memoryless and continuous-time setting, following an exponential distribution with the mean value $1/\lambda$ and $1/\mu$, respectively.

Model of Files

Let $\mathbf{F}$ denote the integrated set of content files available for download in the region of interest. Throughout the work, each RSB is assumed to be manipulated by a distinct owner; the owner uploads contents to its RSB at periodic intervals following the exponential distribution with the mean Δ. RSBs have homogenous buffer size[5] which is denoted by L. When the buffer of RSBs overflows with excessive file uploading from vehicles, the oldest

[5] In practice, the RSBs would be produced by the same vendor with equal

files stored in the RSB will be evicted[6].

Throughout the work, we focus on the design of RSBs and assume that the buffer management at vehicle nodes are predefined and out of the control. In specific, the vehicles could have heterogenous and limited sized buffer storage, and randomly select files to evict if their buffer overflows. The pattern of V2V content swap is also predefined which could follow existing schemes, such as SPAWN [11].

With new files continually published at distributed RSBs and old files evicted from the network, $\mathbf{F}$ dynamically changes over time. In the system, each file is characterized by a three-tuple including file blocks, popularity and availability.

File Blocks Each content file in the system is divided into multiple non-overlapping file blocks for delivery. That is to say, in order to finish downloading a file, a vehicle node must collect all blocks of the requested file from either RSBs or other vehicles with the file stored. A vehicle node can only redistribute a file to the others after it has the entire file downloaded and recovered[7]. Let κ_i denote the number of blocks of file i, where $i \in \mathbf{F}$. For ease of analysis, we assume that all the blocks of files have equal size; with files having different numbers of blocks, they are heterogenous in size. For computation simplicity, L, C_{V2R} and C_{V2V} are normalized by the block size.

File Popularity and Availability Besides the number of file blocks, each content file in the system is characterized by another two parameters, namely popularity and availability.

Definition 1 *The popularity p_i of a file i, where $i \in \mathbf{F}$, represents the probability that a vehicle subscribes to download file i upon each download request which it issues. The* popularity profile *of* $\mathbf{F}$ *is a* $1 \times |\mathbf{F}|$ *probability vector* $\mathbf{P} = \{p_i; i \in \mathbf{F}\}$*, where* $|\cdot|$ *indicates the cardinality of set.*

Definition 2 *The availability a_i of a file i, where $i \in \mathbf{F}$, represents the probability that a randomly selected vehicle has file i cached in its buffer. The* availability profile *of* $\mathbf{F}$ *is a* $1 \times |\mathbf{F}|$ *probability vector denoted by* $\mathbf{A} = \{a_i; i \in \mathbf{F}\}$.

[6] It is interesting to investigate on the impacts of different buffer management schemes, e.g., least frequently used (LFU) and least recently used (LRU) on the system performance, which however is out the scope of this work.

[7] It can be extend by allowing vehicles to redistribute file blocks as long as certain blocks are downloaded in entirety. We study the simplest case and leave the extension for future works

Note that since $\mathbf{F}$ is dynamically changing over time, the popularity profile $\mathbf{P}$ and availability profile $\mathbf{A}$ are also varying over time. In this case, RSBs stochastically select files in $\mathbf{F}$ to cache in their buffer following the caching profile defined below.

Definition 3 *Let b_i denote the probability that a randomly selected RSB has file i stored in its buffer. The* caching profile *of $\mathbf{F}$ is a $1 \times |\mathbf{F}|$ probability vector denoted by $\mathbf{B} = \{b_i; i \in \mathbf{F}\}$.*

Mean Download Delay of Files

The performance of the system is characterized by the mean download delay of files. Let τ_i denote the mean download delay of file i which starts when a download request of file i is issued by a vehicle until the subscriber finishes downloading all the blocks of file i. Given the distribution of RSBs and density of vehicle nodes in the concerned region, the download delay τ_i is dependent on the availability of file i at RSBs, represented by b_i, and vehicles, represented by a_i.

3.3.2 System Utility Function

For each file i, we assume that there is an underlying utility function $U_i\left(\tau_i\right)$ that specifies the satisfaction of vehicular users on the download of file i provided the download delay τ_i. Moreover, it is nature to assume that $U\left(\tau_i\right)$ is a monotonically decreasing function of τ_i, i.e., reducing the download delay τ_i would monotonically increase the user's utility of file i.

The proposed system is designed to maximize a global system utility function $\mathcal{U}$. In essence, the system utility $\mathcal{U}$ represents the integrated utilities of vehicles. In general cases, it can be expressed as a weighted sum of individual user utilities over all files, mathematically,

$$\mathcal{U} = \sum_{i \in \mathbf{F}} w_i U\left(\tau_i\right), \tag{3.1}$$

where $w_i, i \in \mathbf{F}$, is a given positive weight. With different concerns, the system utility can be adapted to achieve different design goals, as following examples:

User-centric Content Distribution

In this scenario, by tuning the weighting factor of each file equal to the corresponding file popularity, the proposed system targets to optimize the user's download experience by maximizing the integrated user satisfactions on the file dissemination. Mathematically, the system utility is given as

$$\mathcal{U} = \sum_{i \in \mathbf{F}} p_i U\left(\tau_i\right). \tag{3.2}$$

Content-centric Content Distribution

The weighting factor w_i can be set to be a predefined value which reflects the importance of file i. For example, breaking news, important software patch, etc., can be assigned with the large weighting factors and accordingly attain high priorities to be stored in RSBs. This ensures those important files to be vastly stored and ubiquitously available.

Cost-centric Content Distribution

A practical concern of the proposed system is the physical cost of RSBs. With larger buffer storage of RSBs, more files can be cached in each RSB, rendering reduced download delay to users; nevertheless, it increases the cost of RSB hardware accordingly. Motivated by this concern, the system utility can be modified by introducing the cost function in (3.1) to strike a trade-off between the system performance and investment cost, as

$$\mathcal{U} = \sum_{i \in \mathbf{F}} w_i U\left(\tau_i\right) - \mathbf{C}\left(L\right), \tag{3.3}$$

where $\mathbf{C}\left(L\right)$ represents the physical cost of RSBs which is a non-decreasing function of the buffer size L. In practice, $\mathbf{C}\left(L\right)$ can be evaluated by $\mathbf{C}\left(\sum_{i \in \mathbf{F}} b_i \kappa_i\right)$ instead, where $\sum_{i \in \mathbf{F}} b_i \kappa_i$ represents the mean usage of RSB buffer storage. As such, the three designs of the system as aforementioned can be solved using the unified formulation as described below.

3.3.3 Problem Formulation

Given the network model introduced above, each RSB distributedly determine the optimal caching profile $\mathbf{B}$ of files to attain the maximal system utility $\mathcal{U}$, mathematically,

$$\begin{array}{rll} \textit{maximize} & \mathcal{U} & \\ \textit{subject to}: & \Pr(X \geq L) \leq \varepsilon, & \\ & b_i \in [0,1], & i \in \mathbf{F}, \end{array} \tag{3.4}$$

where X denotes the usage of the RSB buffer storage at any time, and $0 < \varepsilon << 1$ is a predefined constant. The constraint of (3.4) specifies that the overflow probability of each RSB should be no larger than ε.

Lemma 1 *Let X_i, $i \in \mathbf{F}$, be an independent binary random variable with the probability mass function*

$$\Pr\left(X_i = 1\right) = b_i, \quad \Pr\left(X_i = 0\right) = 1 - b_i,$$

which denotes whether file i is stored in a RSB or not. For $X = \sum_{i \in \mathbf{F}} X_i \kappa_i$ with $\kappa_i > 0$, which denotes the usage of RSB buffer storage, we have $E\left(X\right) = \sum_{i \in \mathbf{F}} b_i \kappa_i$. By denoting $v = \sum_{i \in \mathbf{F}} b_i \kappa_i^2$, we have

$$\Pr\left(X \geq E\left(X\right) + \psi\right) \leq \exp\left(-\frac{\psi^2}{2\left(v + \kappa\psi/3\right)}\right) \tag{3.5}$$

where $\kappa = \max\{\kappa_i ; i \in \mathbf{F}\}$.

Proof 1 *Refer to [38] (pp.25).*

Theorem 1 *Given the network modeling in previous subsections, the constraint of (3.4) is achieved when*

$$E\left(X\right) \leq L - \kappa\frac{2}{3}\log\epsilon - \sqrt{\kappa^2\frac{4}{9}\log^2\epsilon - 2\kappa L\log\epsilon}. \tag{3.6}$$

Proof 2 *As shown in Appendix 3.7.1.*

Denote by $\mathcal{L} = L - \kappa\frac{2}{3}\log\epsilon - \sqrt{\kappa^2\frac{4}{9}\log^2\epsilon - 2\kappa L\log\epsilon}$. With Theorem 1, (3.4) can be modified as

$$\begin{array}{lll} \mathsf{OPT} & \textit{maximize} & \mathcal{U} \\ & \textit{subject to}: & \sum_{i \in \mathbf{F}} b_i \kappa_i \leq \mathcal{L}, \\ & & b_i \in [0,1], \qquad i \in \mathbf{F}. \end{array} \tag{3.7}$$

In (3.7), with $\mathbf{A}$ and $\mathbf{P}$ provided, tuning the caching profile $\mathbf{B}$, i.e., content replications in RSBs, will adapt the download delay τ_i of each file i and accordingly lead to different system utility $\mathcal{U}$. In this work, our goal is to determine the solution of (3.7) in a distributed manner.

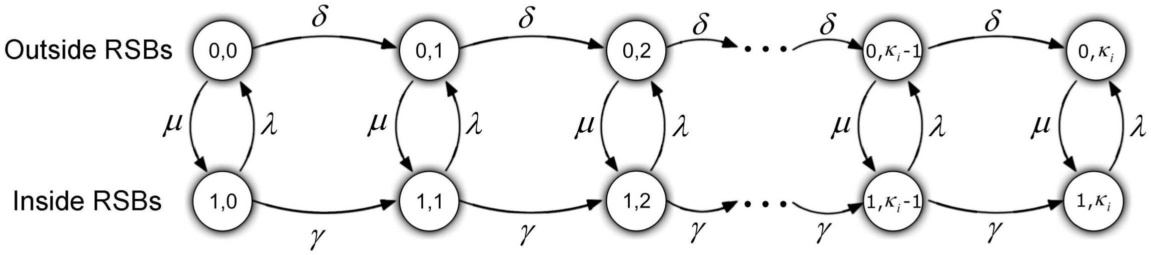


Figure 3.3: State space and transitions of the two-dimensional Markov process

3.3.4 Evaluation of File Download Delay

To solve (3.7), the foremost issue is to identify the relationship between the file download delay and the caching profile $\mathbf{B}$. To this end, we randomly select a vehicle node from the network (referred to as the tagged node) and evaluate its download delay of file i. Specifically, based on the system model described in Subsection 3.3.1, we represent the tagged vehicle node by a two-dimensional Markov process $(M_i(t), K_i(t))$. Here, $M_i(t) \in \{0,1\}$ represents the mobility of the vehicle node according to the on-off model described in Subsection 3.3.1, and $K_i(t) \in \{0,1,...,\kappa_i\}$ represents the number of file blocks that the tagged node has downloaded until time t. Fig. 3.3 shows the state space of the Markov process and all the non-null transitions. In what follows, we evaluate the transition rates of the Markov process according to the locations of the tagged node.

Tagged Vehicle Outside the Coverage of RSBs

When the tagged node is outside the coverage of any RSBs, it can only download from nearby vehicles using the V2V communications; at each time, we refer to the set of vehicles which are within the communication range of the tagged node as the neighbor nodes. Let n denote the number of the neighbor nodes of the tagged node; n as a random variable and let $\langle n\rangle$ and $\mathrm{Var}(n)$ denote its mean and variance. Assuming that the vehicular network has an ideal MAC where the channel airtime is fairly shared among the nearby vehicles, the throughput of the tagged node using the V2V communication is a function of n as,

$$r = \frac{C_{\mathsf{V2V}}}{n+1} Q_i(n), \tag{3.8}$$

where $Q_i(n) = 1-(1-a_i)^n$, representing the probability that at least one neighbor node of the tagged node has file i stored in its buffer, or equivalently, the probability that the tagged node can retrieve file i from its neighbor nodes. $(n+1)$ in (3.8) represents the number of vehicles fairly sharing the channel.

Let δ denote the transition rate from the state $(0, K_i(t))$ to the state $(0, K_i(t)+1)$, where $K_i(t) \in \{0, 1, ..., \kappa_i - 1\}$, as shown in Fig. 3.3. Assuming that the download time of one block using the V2V communication follows the exponential distribution, δ is equal to the mean V2V communication throughput $\langle r \rangle$ where r specified in (3.8). Taking the expectation on n in both sides of (3.8), we approximate $\langle r \rangle$ using the second order Taylor series approximation as shown in Lemma 2.

Lemma 2 *With the second order Taylor approximation, we have*

$$\langle r \rangle \approx r|_{\langle n \rangle} + \frac{1}{2}\mathrm{Var}(n) \left. \frac{d^2 r}{dn^2} \right|_{\langle n \rangle}. \tag{3.9}$$

Proof 3 *Refer to Appendix 3.7.2.*

Tagged Vehicle Inside the Coverage of RSBs

In this case, the tagged node can download the demanded blocks from either neighbor vehicles or the RSB. We assume that the tagged vehicle would select to download from RSBs with high priority, if the connected RSBs have the desired file stored; otherwise, it would download from neighbor vehicle nodes. This is because that RSBs have the greater communication capacity than vehicles [37]. In this scenario, given that file i is stored at the RSB with probability b_i, the download throughput of the tagged vehicle in this scenario is

$$R = b_i \frac{C_{\mathsf{V2R}}}{n+1} + (1 - b_i) r. \tag{3.10}$$

The first component on the right-hand-side of (3.10) represents the download rate from the RSB with the ideal MAC applied, and the second component on the right-hand-side of (3.10) represents the download rate using the V2V communications given that with probability $(1 - b_i)$ the RSB does not have the desired file i stored.

Let γ denote the transition rate from the state $(1, K_i(t))$ to the state $(1, K_i(t)+1)$, where $K_i(t) \in \{0, 1, ..., \kappa_i - 1\}$, as shown in Fig. 3.3. Similar to the previous case, we assume that the download time of one block inside the RSB follows the exponential distribution. Therefore, we have γ equal to $\langle R \rangle$ with R shown in (3.10). $\langle R \rangle$ can be approximated with the second order Taylor approximation as in Lemma 3.

Lemma 3 *With the second order Taylor approximation, we have*

$$\langle R\rangle \approx b_i\Phi + (1-b_2)\langle r\rangle, \tag{3.11}$$

where $\Phi = C_{\mathsf{V2R}}\left(\frac{1}{\langle n\rangle+1} + \frac{\mathrm{Var}(n)}{(\langle n\rangle+1)^3}\right)$

Proof 4 *Refer to Appendix 3.7.3.*

Mean First Passage Time

We evaluate the average file download delay by the mean first passage time stating from the state $K_i(0) = 0$, i.e., no blocks are downloaded, until the state $K_i(t) = \kappa_i$, i.e., the tagged node collects all the blocks. Let $\Gamma(m, k)$ denote the first passage time stating when the vehicle is in state (m, k) until all κ_i blocks are downloaded, mathematically,

$$\Gamma_i(m, k) = \min\{t > 0 |\, M_i(0) = m, K_i(0) = k \text{ and } K_i(t) = \kappa_i\}.$$

The mean download delay of file i is thus

$$\tau_i = \frac{1}{\lambda+\mu}\left(\lambda\Gamma_i(0,0) + \mu\Gamma_i(1,0)\right), \tag{3.12}$$

where $\frac{\lambda}{\lambda+\mu}$ and $\frac{\mu}{\lambda+\mu}$ are the limiting probabilities that the tagged node is outside and inside the coverage of RSBs, respectively, when the tagged node initiates the download subscription of file i. The expression of τ_i is shown in Theorem 2.

Theorem 2 *The average download delay of file i is*

$$\tau_i \approx \frac{b_i(\Phi-\delta)\lambda + \delta(\lambda+\mu) + \kappa_i(\lambda+\mu)^2}{b_i(\Phi-\delta)\mu(\lambda+\mu) + \delta(\lambda+\mu)^2}. \tag{3.13}$$

Proof 5 *Refer to Appendix 3.7.4.*

Corollary 1 *The download delay τ_i is a monotonic non-increasing, convex function of b_i if $\frac{\kappa_i}{\delta} \geq \frac{1}{\mu} - \frac{2}{\lambda+\mu}$.*

Proof 6 *Refer to Appendix 3.7.5.*

3.4 Protocol Description

By substituting (3.13) into (3.1), we can derive the solution of (3.7).

3.4.1 Sufficient Conditions for a Concave system utility Function

We make two assumptions as follows:

- ▷ $\frac{\kappa_i}{\delta} \geq \frac{1}{\mu}$, for all i. That is to say the average download time of a file when vehicles are outside the RSBs (evaluated as κ_i/δ) should be no smaller than the average sojourn time of vehicle nodes outside the RSBs (evaluated as $1/\mu$). Otherwise, the assistance of RSBs is negligible as the desired file can be downloaded easily through V2V communications only before vehicles entering into the coverage of any RSBs.
- ▷ $U(\tau_i)$ is a non-increasing, twice differentiable concave function of τ_i. As an example to explain the methodology, in this work we adopt

$$U(\tau_i) = -\tau_i \quad \text{and} \quad w_i = p_i,\ i \in \mathbf{F}, \tag{3.14}$$

for the simplicity.

According to Corollary (1) and Proposition (1), the system utility $\mathcal{U}$ is a concave function of b_i, and accordingly, the system utility maximization problem in (3.7) is a convex optimization problem.

Proposition 1 *Assuming that $U(\tau_i)$ is a non-increasing, twice differentiable concave function of τ_i, then $U(\tau_i)$ and system utility $\mathcal{U}$ are concave functions of b_i.*

Proof 7 *Refer to Appendix 3.7.6.*

3.4.2 Global Optimal Solution to (3.7)

Let $\mathbf{B}^*$ denote the optimal caching profile. By examining (3.7) with the Karush–Kuhn–Tucker (KKT) conditions as shown in Appendix 3.7.7, we have $\mathbf{B}^*$ as

$$b_i^* = \frac{\sqrt{p_i\left[\kappa_i\mu(\lambda+\mu)+\delta(\mu-\lambda)\right]}}{\sum_{j\in\mathbf{F}}\kappa_j\sqrt{p_j\left[\kappa_j\mu(\lambda+\mu)+\delta(\mu-\lambda)\right]}}\left(\mathcal{L}+\frac{\delta(\lambda+\mu)}{\mu(\Phi-\delta)}\sum_{j\in\mathbf{F}}\kappa_j\right)-\frac{\delta(\lambda+\mu)}{\mu(\Phi-\delta)},\quad b_i^*\in\mathbf{B}^*. \tag{3.15}$$

As such, given the availability profile $\mathbf{A}$ and popularity profile $\mathbf{P}$, each RSB should select a content file i to cache with a probability of b_i^*. We refer to this scheme as the global optimal replication in RSBs.

The global optimal replication provides the optimal solution to (3.7). However, note that both $\mathbf{A}$ and $\mathbf{P}$ are system parameters related to the file information across the whole network. They are not available to individual RSBs or vehicle nodes when the network size is large. Therefore, the global optimal replication scheme is not practical for large-scale real-world deployment. Nevertheless, the global optimal replication provides a benchmark for performance comparison with other replication schemes. In what follows, we propose a decentralized algorithm to determine the content replication at RSBs.

3.4.3 Distributed Content Replication

In this part, we design a distributed algorithm to enable RSBs to select the appropriate files to store according to (3.7) in a fully distributed manner. To achieve this goal, we approximate b_i^* by b_i^{d} as

$$b_i^{\mathsf{d}} = \frac{\mathcal{L}\sqrt{p_i\left[\kappa_i\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}}{\sum_{j\in\mathbf{F}}\kappa_j\sqrt{p_j\left[\kappa_j\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}}. \tag{3.16}$$

This can greatly simplify the algorithm design with modest performance degradation as verified by simulations.

To help RSBs distributedly select each file i with the probability b_i^{d} from the network, we adopt a random walk based algorithm over a file graph as follows:

File Graph

The file graph refers to as a graph connecting all the files stored in distributed vehicles. As an example shown in Fig. 3.4, each vertex in the graph represents a file stored in a vehicle node. Additionally, each vehicle has an anchor file, e.g., file j, which is selected from the locally stored files in vehicles and has the largest value of $\sqrt{p_j\Phi\left[\kappa_j\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}$ among the buffered files. Each vehicle node periodically broadcasts its anchor file information, including the availability a_j and download demand p_j, to the neighbor vehicles. How to measure the availability and download demand of files will be presented in Subsection 3.4.3. In the file graph, all files stored in the same vehicle node are fully connected, and the anchor files among neighboring vehicles are fully connected, as shown in Fig. 3.4.

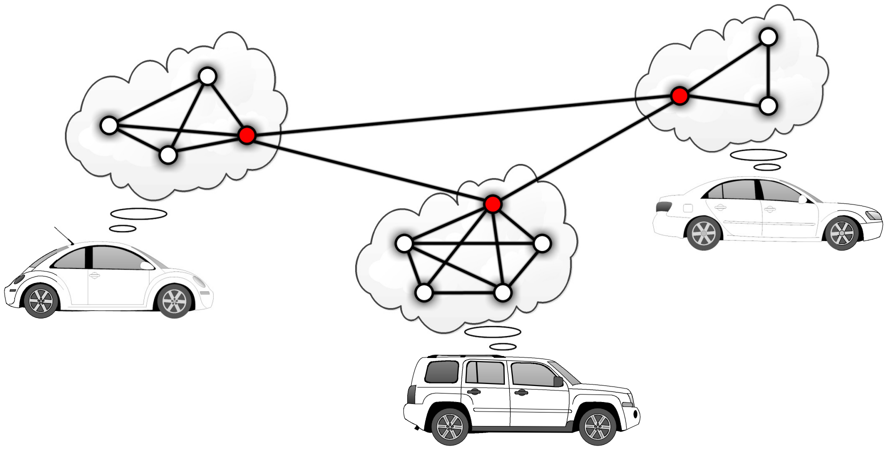


Figure 3.4: File graph in VTube

Therefore, the file graph has a two-tier architecture where the top tier connects the anchor files of vehicles and the underlying tier connects all the files inside a vehicle to its anchor file.

Random Walk Based File Selection

The file selection is realized by a random walk algorithm over the file graph as described in Algorithm 1. Specifically, to determine the files stored in RSBs, a RSB first issues a number η of random walkers to separate vehicles in the communication range. Each vehicle which receives a walker will then initiate the random walk process starting from its anchor file. The walker is forwarded stochastically on the file graph from one vertex (file) to another vertex (file) following the Metropolis-Hasting algorithm; the derivation of transition probabilities in the random walk is given in Appendix 3.7.8. Once the walker is forwarded to the anchor file, it is possible to be relayed to other anchor files stored in different vehicles. In this case, the walk is forwarded to other vehicles and proceeds the random walk algorithm. After being relayed for Time-To-Live (TTL) hops among files on the file graph including self-loops, the walker stops at a file which is then selected to be uploaded to RSBs.

In order to compute the transition probability of the walker, each vehicle needs to know the availability and download demand of the files stored in its buffer. In the proposed system, we enable vehicles to distributedly measure these parameters as follows:

Algorithm 1: Random walk algorithm starting from file x

```
/* $m$:  current file with walker                                   */
/* $h$:  hop account                                                */
/* $p$:  random number                                              */
/* $P_{mn}$:  transition probability from file $m$ to file $n$ shown in H-4)
   of Appendix 3.7.8                                                */
begin
    Initialization: $m \leftarrow x$; $h \leftarrow 0$; $p \leftarrow 0$;
    while $h < TTL$ do
        $p \leftarrow$ random number in $[0, 1]$ ;
        foreach file $n$ $(n \neq m)$ connected to file $m$ in the file graph do
            if $p \leq P_{mn}$ then
                $m \leftarrow n$;
                quit the foreach loop;
            else
                $p \leftarrow p - P_{mn}$;
        $h \leftarrow h + 1$;
Result: File $m$
```

Measurement of File Availability

The availability a_i of file i is only measured by the vehicles which needs to download file i. As each vehicle interested in file i continually issues the download requests to its neighboring vehicles, it can estimate the file availability a_i based on the replies with $\frac{\text{No. of vehicles having file } i \text{ stored}}{\text{Overall No. of vehicles contacted}}$. Whenever the vehicle, e.g., x, interested in file i meets another vehicle, e.g., y, which has file i stored, vehicle x would inform the measurement of a_i to vehicle y piggybacked with the download request. As such, vehicle y would receive multiple measurements of a_i. For each new measurement received, it would incorporate it with the previous measurement using the moving average. Once vehicle x finishes downloading file i, it can use its measurement on a_i to evaluate the availability of file i.

Measurement of Download Demand

The download demand d_i of file i is measured by the vehicles which have file i stored. As those vehicles keep receiving download requests from others and a portion of the requests are for file i, d_i can be estimated based on this information as $\frac{\text{No. of vehicles requesting to download file } i}{\text{Overall No. of download requests received}}$.

It is important to note that each vehicle only needs to know the available and download demand of the files it stores. As such, the distributed measurement will not impose much workload on the message exchange. To improve the measurements of availability and download demand of files, we can also make use of the RSBs. In this case, RSBs would collect the measurements from different vehicles driving through, and average the measurements towards a more accurate estimation, then announce them to vehicles. There would be other methods for more accurate measurements based on the vehicular sensor networks [14], which is out of the scope of this work.

3.4.4 Protocol Description

This part describes the detailed protocol design and implementation of the proposed system. In the system, each RSB, e.g., A, works in a fully distributed manner and conducts the following three operations:

File Publication

Whenever a new file is published at RSB A (uploaded by its owner), the RSB A issues η walkers to separate vehicles in its coverage. Each walker is relayed among files over the file graph embedded in the vehicular networks following Algorithm 1, and results in one file selected after the TTL hops. The vehicles with the selected files will then upload the files to the RSBs which they drive through. As such, RSBs are dynamically refreshed with new contents continuously uploaded; and this process is triggered by the publication of new files. The value of η will be discussed later. Note that in this phase, RSB A is only responsible to issue walkers to the vehicular network upon the publication of new files. The files selected by the walkers will be uploaded to RSBs in the communication range of the vehicles hosting the selected file, which may not be RSB A.

The value of η is set to make the overall number of files in the network stable. It is dependent on the rates at which new files are published to the network and the out-dated files are evicted from the network. Let $\mathcal{T}$ be the average life time of files in the network, where the life time represents the time duration that a file is stored in RSBs. Let θ be the average injection rate of new contents to the network at distributed RSBs. As each newly published file will initiate η walkers to the vehicular network and finally cause η files to be uploaded to RSBs from vehicles, the rate at which RSBs get new contents uploaded is in total $(1+\eta)\,\theta$. Let $\mathcal{N}$ denote the number of RSBs in the overall network. Mathematically, the rate at which the number of content files changes over time is

$$\frac{\partial\left|\mathbf{F}\right|}{\partial t}=(1+\eta)\,\theta\mathcal{N}-\frac{1}{\mathcal{T}}\mathcal{N}\mathcal{L}. \tag{3.17}$$

In the steady state with $\frac{\partial|\mathbf{F}|}{\partial t}=0$, we have

$$\eta=\frac{\mathcal{L}}{\mathcal{T}\theta}-1. \tag{3.18}$$

To compute η with (3.18), we assume that θ and $\mathcal{T}$ are known which can be measured at different RSBs distributively based on the history of file storage in RSBs. RSBs can also exchange the measurements among each other to improve the accuracy with the assistance of vehicles.

Retrieve Files from Vehicles

Whenever a vehicle with a selected file in the random walk algorithm comes into the coverage of RSB A, it will retrieve the file immediately from the vehicle. During this period, the channel of RSB A is used exclusively for the file retrieval. If there are multiple uploads simultaneously from different vehicles to RSB A, RSB A only processes one retrieval at one time until this retrieval completes. Once the selected file in a vehicle is uploaded, the vehicle will not upload this file to other RSBs unless this file is selected again in the random walk algorithm. In case that a vehicle moves out of the coverage of RSB A before it accomplishes the retrieval, RSB i would proceed the file retrieval again from other driving through vehicles which has the unfinished file stored. If its buffer is full, RSB A depletes the buffer by deleting the file which has been stored for the longest time[8].

Upload File to Vehicles

In the idle period of RSB i when it does not need to issue walkers to the network or retrieve files from vehicles, it uploads the cached file to the driving through vehicles upon their requests.

Each RSB in the system thus works in the three modes in a fully distributed manner. In what follows, we evaluate the performance of the proposed system compared to the centralized content replication.

[8] Other buffer management schemes such as LFU (least frequently used) and LRU (least recently used) can also be used which is out scope of the paper.

3.5 Simulations

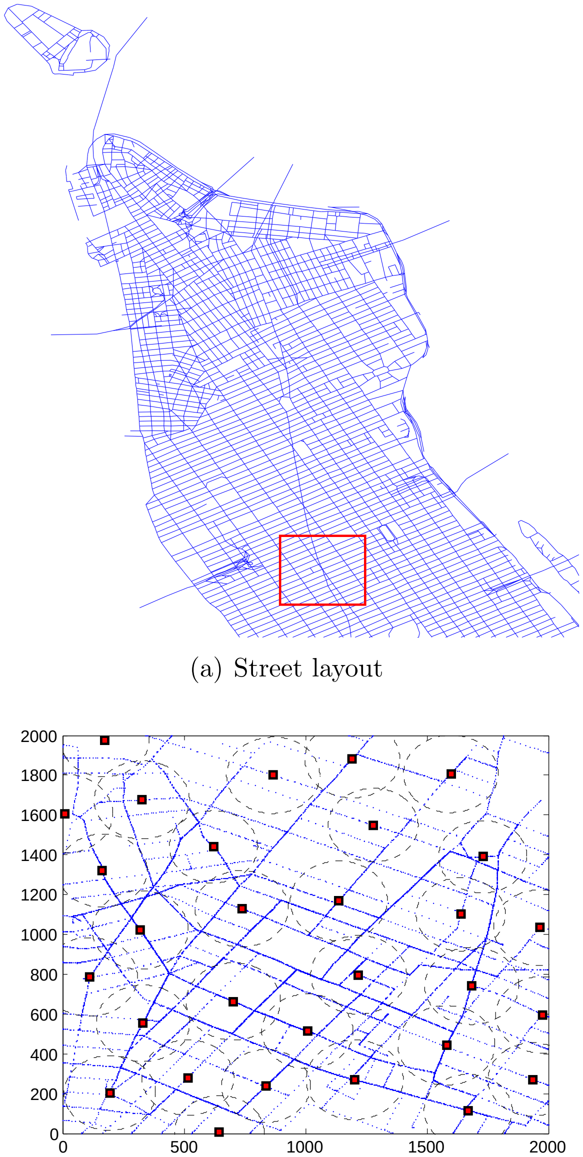
This section evaluates the performance of the proposed system using simulations based on a discrete event simulator coded in C++.

3.5.1 Simulation Setup

Our simulation is carried out over a 2 km×2 km regional road map on the Manhattan island with the contour of the street layout plotted in Fig. 3.5. Each road segment in Fig. 3.5(a) is of two lanes with the bidirectional vehicle traffic. Compromised to the complexity of simulations, we select a bounded region on the map for our simulations as shown in run Fig. 3.5(b). There are totally 29 RSBs deployed in the region with the communication range uniformly distributed within the range $[180, 200]$ meters. For each simulation run, 300 vehicles are involved in the content distribution. The mobilities of vehicles are generated by VANETMobisim [39], in which the destination of each trip is randomly selected, and the velocity of vehicles is controlled no larger than 60 km/s and adapted by the IDM-LC (Intelligent Driver Model with Lane Changes) mode. The coverage of V2V communications is set to be 150 meters. With this configuration, we have $\lambda = 29.46$s, $\mu = 12.19$s, $\langle n \rangle = 4.02$ and $\mathrm{Var}(n) = 6.18$. In each simulation run, 200 files are initially available for download in the network, which are randomly stored in RSBs and vehicles. Unless mentioned otherwise, all RSBs have the equal buffer storage to cache 3×10^3 file blocks, i.e., $3 \times 10^3/100 = 30$ files at most. Vehicles have equal buffer to cache 1×10^3 file blocks, i.e., 10 files at most. The download capacity of the vehicle to RSB communication, C_{V2R}, is 50 blocks/sec and that of the V2V communication, C_{V2V}, is 20 blocks/sec. Each vehicle can communicate with one other network component at most, and the parallel communication sessions are scheduled through the ideal MAC.

3.5.2 Verification of the Analysis

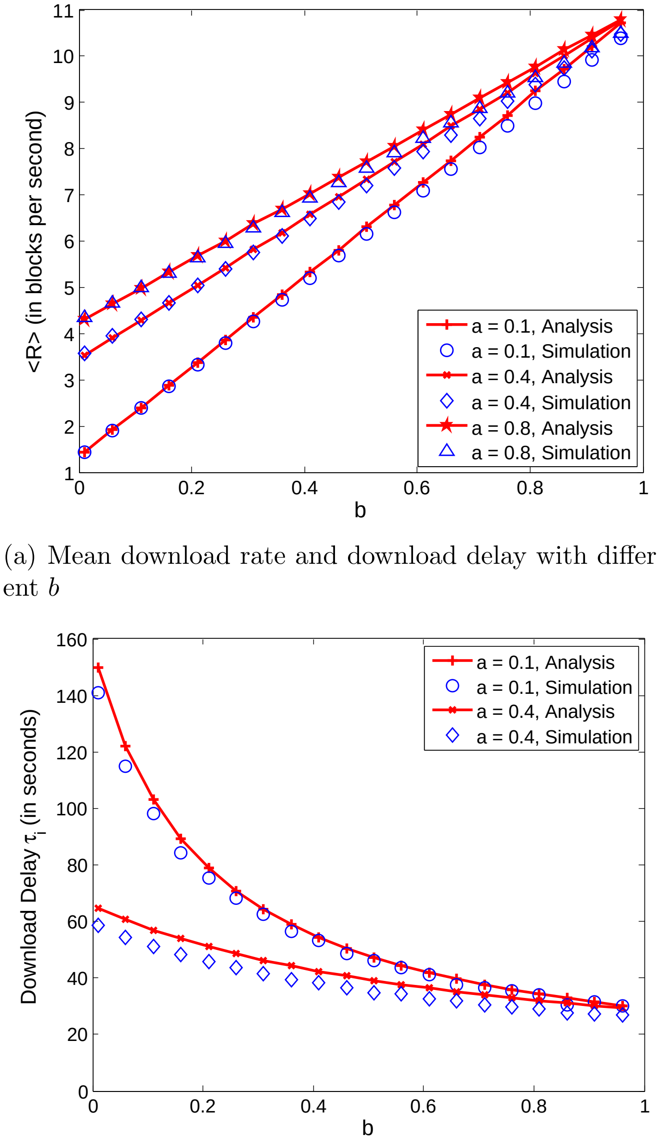
As the proposed protocol is based on the evaluation of file download delay, in the first experiment, we verify the accuracy of (3.9), (3.7.3) in evaluating the mean download rates $\langle r \rangle$ and $\langle R \rangle$, when vehicles are inside and outside RSBs, respectively, and the accuracy of (3.13) in evaluating the mean download delay of files. To this goal, we carry out the Monte Carlo simulations by investigating on the download performance of a tagged vehicle. We make the tagged vehicle subscribe to download a file, referred to as file i in this subsection, of file size to be 200 blocks and report the averaged results over 5000 simulation runs.



(a) Street layout

(b) RSB distribution on the road

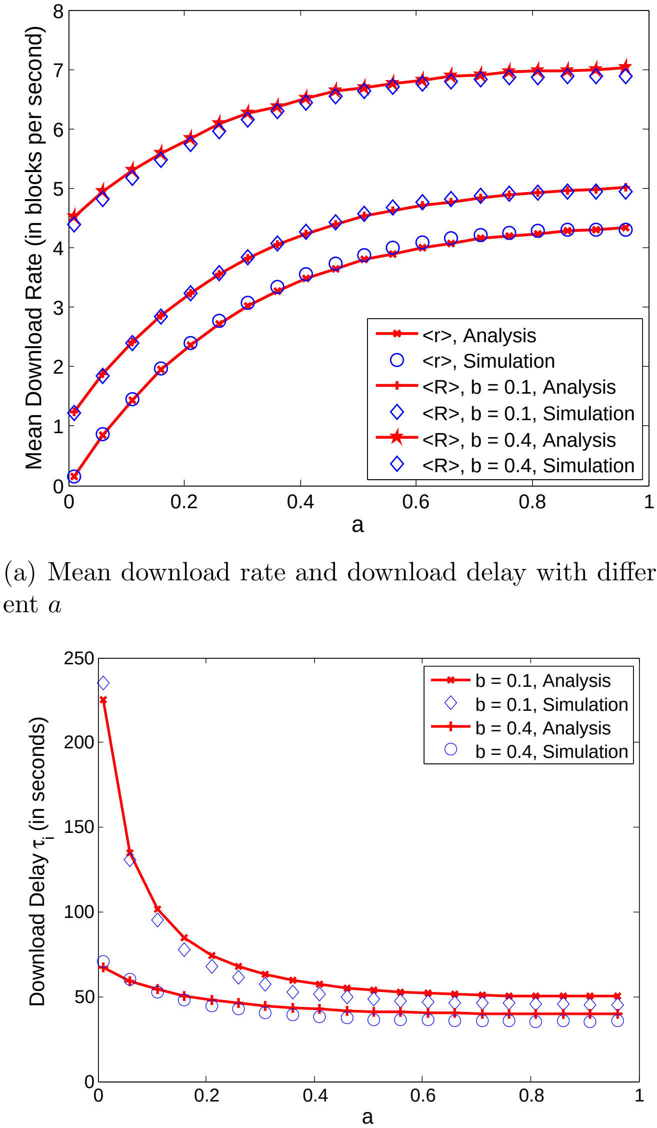
Figure 3.5: Street layout and RSB distribution



(a) Mean download rate and download delay with different b

(b) Global system utility with different C_{V2R}

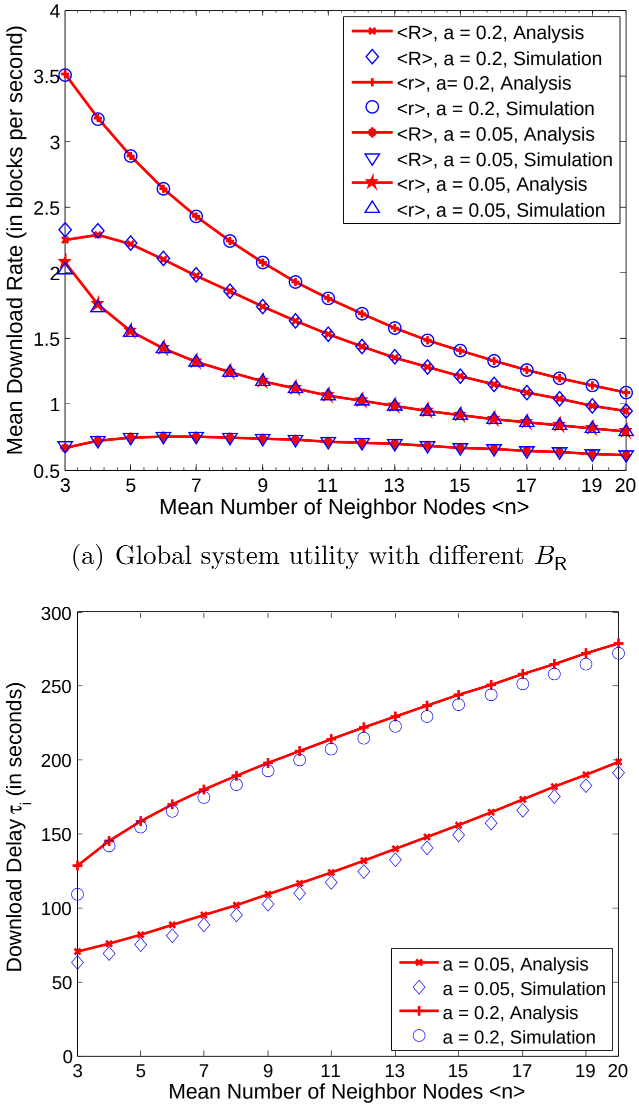
Figure 3.6: Performance with different parameters of RSBs



(a) Mean download rate and download delay with different a

(b) Global system utility with different C_{V2R}

Figure 3.7: Performance with different parameters of RSBs



(a) Global system utility with different B_{R}

(b) Global system utility with different C_{V2R}

Figure 3.8: Performance with different parameters of RSBs

Fig. 3.6(a) shows the values of $\langle r\rangle$ and $\langle R\rangle$ as a function of b_i when a_i is 0.1. As we can see from the figure, $\langle r\rangle$ remains the same with different b_i, and $\langle R\rangle$ increases linearly with b_i. The analysis in (3.9), (3.7.3) matches the simulations well. Fig. 3.6(b) shows the mean download delay the file with different b_i. As we can see, when b_i increases, the download delay τ_i reduces dramatically which can be characterized by (3.13). Moreover, τ_i is a convex function of b which validate Corollary 1. In addition, when a_i changes from 0.1 to 0.4, the download delay τ_i reduces significantly, as in this case more vehicles on the road have file i stored and therefore the tagged node can finish downloading faster.

Fig. 3.7(a) shows the values of $\langle r\rangle$ and $\langle R\rangle$ as a function of a_i when b_i is 0.1. As we can see, by increasing a_i, both $\langle r\rangle$ and $\langle R\rangle$ increases with a constant gap between the two curves which can be characterized by (3.7.3). As shown in Fig. 3.7(b) and indicated by (3.13), we can see that the mean download delay of file i reduces when a_i increases.

Fig. 3.8(a) shows the values of $\langle r\rangle$ and $\langle R\rangle$ as a function of the mean number of neighbor vehicles, i.e., $\langle n\rangle$, when b_i is 0.1. As we can see, by increasing $\langle n\rangle$, $\langle R\rangle$ reduces monotonically. This is because that with $\langle n\rangle$ increasing more vehicles share the capacity of RSBs and contend the channel with the tagged node. As in Fig. 3.8(a), in both cases when $a_i = 0.2$ and 0.05, $\langle r\rangle$ increases first when $\langle n\rangle$ increases and then reduces. This is because that when $\langle n\rangle$ increases, more neighbor vehicles may have the desired file i stored and upload the file to the tagged node. However, when $\langle n\rangle$ is large, indicating that more neighbor nodes are contending the channel with the tagged node, $\langle r\rangle$ reduces with $\langle n\rangle$ increasing. Fig. 3.8(b) shows the download delay of file i, τ_i, as a function of $\langle n\rangle$. As we can see, by increasing $\langle n\rangle$, τ_i increases. This is because that the download rate of the tagged node reduces as shown in Fig. 3.8(a).

3.5.3 Performance of Protocol

In this experiment, we validate the performance of the proposed protocol. To this end, we simulate a dynamic system in which each RSB periodically publishes a new file to the network at the intervals following the exponential distribution with the mean of 60 seconds. The index of files increases linearly according the publication time of the file in the network. The size of each file is accounted in the unit of blocks and is uniformly distributed between $[40, 100]$ blocks. Files have different popularity which follows the Zipf distribution; the popularity of the ith file in the network is as

$$p_i = \frac{1}{(\hat{\imath})^\alpha} / \sum\nolimits_{j=1}^{|\mathbf{F}|} \frac{1}{j^\alpha}, \tag{3.19}$$

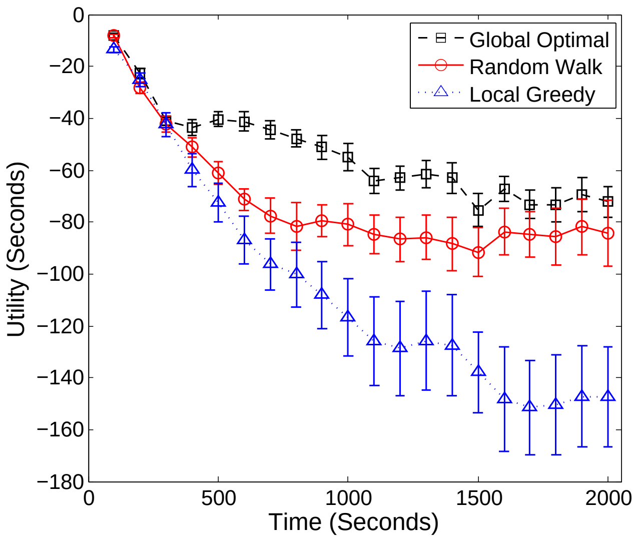
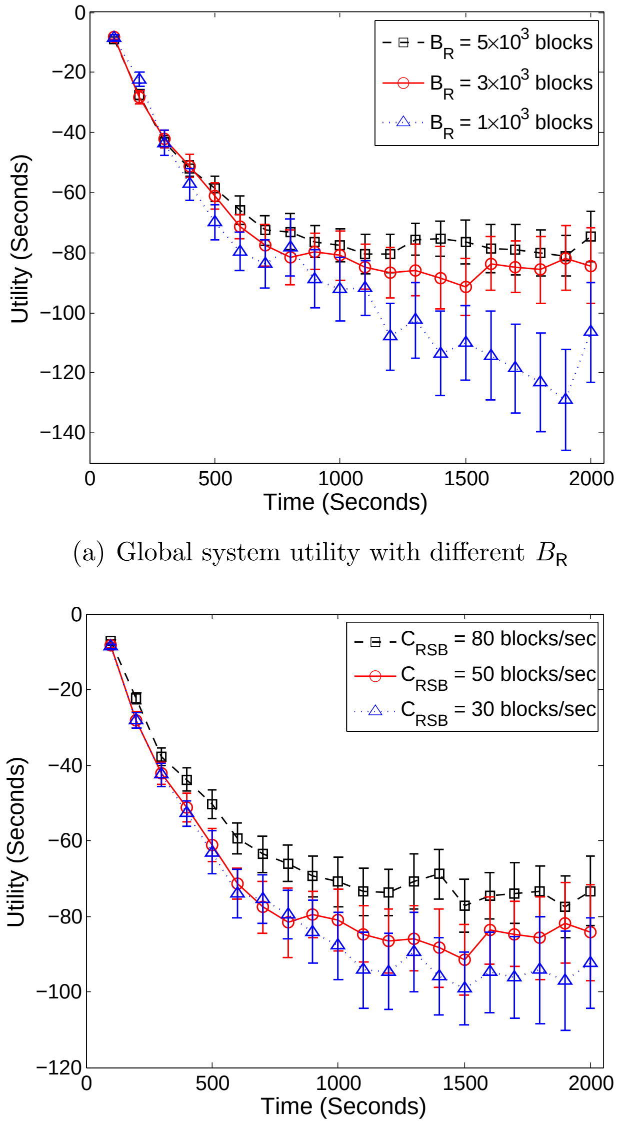


Figure 3.9: Comparison between global optimal, random walk based and local greedy content replication schemes

where α is a configurable parameter of the Zipf function. $\hat{i} = (i \,\mathsf{mod}\, 500)$ where mod denotes the modulo operation. In this case, the popularity of files renews whenever 500 new files are published. The lifetime of each file is set 200 seconds which results in 5 walkers generated per RSB according to (3.18) when a new file is published. RSBs selectively retrieve files from vehicles based on the content replication scheme presented in Section 3.4. When the buffers of RSBs overflow, the file with the longest lifetime stored in the buffer is evicted. Vehicles select files to download based on the Zipf distribution as aforementioned. Once the buffer of vehicles is full, a randomly selected file is evict to release the cache for new downloads.

In this experiment, we evaluate the utility function $\mathcal{U} = -\sum_{i\in\mathbf{F}} p_i\tau_i$ every 100 sections, which accounts for the summed download delay of files weighted by the file popularity within this period. We conduct 50 simulation runs upon each experiment and plot the mean results with the 95% confidence intervals.

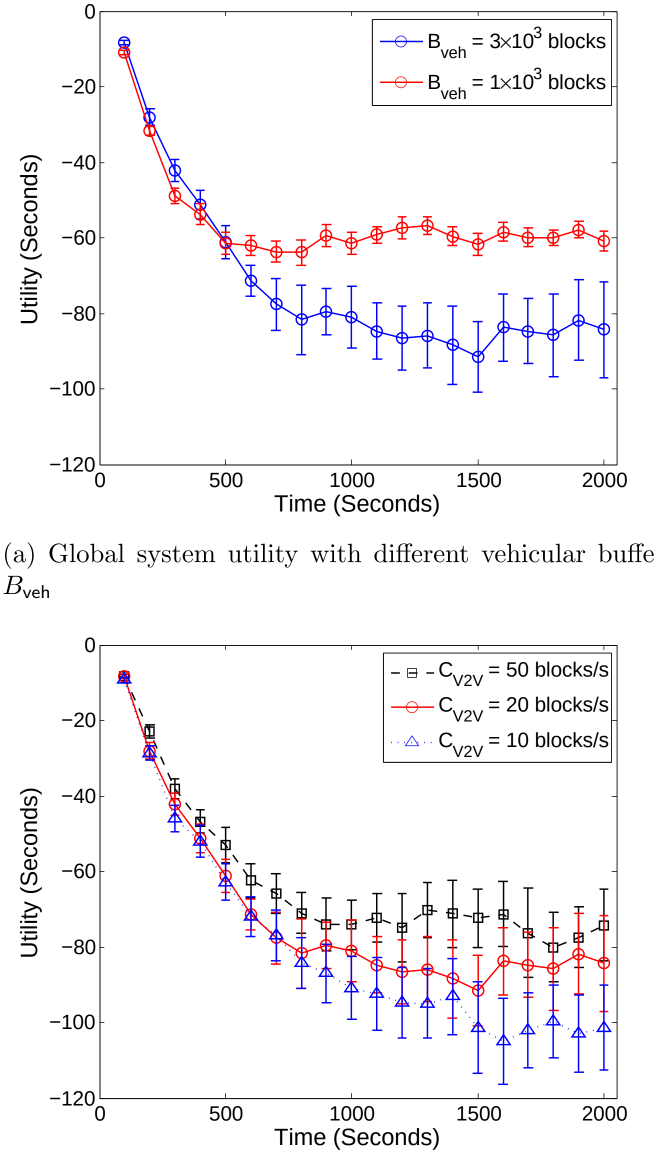
Fig. 3.9 shows the comparisons between the proposed random walk based content replication scheme and the global optimal and local greedy content replication schemes. The three schemes adopt the same content upload and download operations between vehicles and RSBs, except for the file selection strategies when individual RSBs retrieve files to store. Using the global optimal scheme, each RSB selects a file, e.g., i, from the drive through vehicles to store with the probability b_i^* as shown in (3.15). Using the local greedy



(a) Global system utility with different B_{R}

(b) Global system utility with different C_{V2R}

Figure 3.10: Performance with different parameters of RSBs



(a) Global system utility with different vehicular buffer B_{veh}

(b) Global system utility with different C_{V2V}

Figure 3.11: Performance with different parameters of RSBs

algorithm, each RSB selects a file with the largest value of file popularity p_i to store. The strategy of file selection using the random walk based algorithm is discussed in Section 3.4. As we can see from Fig. 3.9, the global optimal scheme has the best performance, followed by the random walk algorithm. The local greedy algorithm has the worst performance for two reasons. Firstly, by selecting files with the local maximal file popularity to store, the local greedy scheme is myopic and cannot optimize the overall performance of the network. Secondly, without considering the storage of vehicles, using the local greedy scheme, RSBs may store files which have already been vastly stored in vehicles, and therefore, cannot be efficiently utilized to cooperate with the vehicular storage towards maximal social welfare. Notably, in all the schemes, the global system utility $\mathcal{U}$ reduces over time and finally approaches to a stabilized value. This is because that at each interval, $\mathcal{U}$ is evaluated by summing up the weighted download delays of files downloaded. Therefore, in the early periods of simulations, only files with small delays are finished and accounted, leading to the small value of $\mathcal{U}$. As time goes, files with long download delays are accounted when they are finished, and $\mathcal{U}$ become stabilized in this case.

Figs. 3.11(a) and 3.11(b) show the global system utility with different buffer sizes and communication capacity of RSBs, respectively, with other parameters unchanged. As we can see, when the buffer size or communication capacity increase, the global system utility increases, indicating smaller download delay of files. However, by enhancing the buffer storage and communication capacity will lead to increased physical cost of RSBs which discourages the large-scale deployment of RSBs.

Figs. 3.11(a) and 3.11(b) show the global system utility when increasing the buffer size and V2V capacity of vehicles, respectively, with other parameters unchanged. As we can see, increasing the vehicles' buffer size or the capacity of V2V communications will significantly enhance the network performance, as more bandwidth and storage resource are available in the network. However, in practise, the capacity and buffer storage contributed by vehicles are out of control. This therefore calls for an effective incentive mechanisms to encourage contributions.

Fig. 3.12 shows the global system utility with different values of α in the Zipf function (3.19). $\alpha = 0$ indicates that all the files have the equal popularity; the larger α is, the faster that the popularity of files decreases when $\hat{\imath}$ increases. As we can see, when α increases, the global system utility reduces significantly. This is because there exists certain very popular files which are highly requested. As such, the RSBs becomes the upload bottleneck of the files which enlarge the download delay of vehicles. To combat this effect therefore requires to enlarge the upload capacity of RSBs.

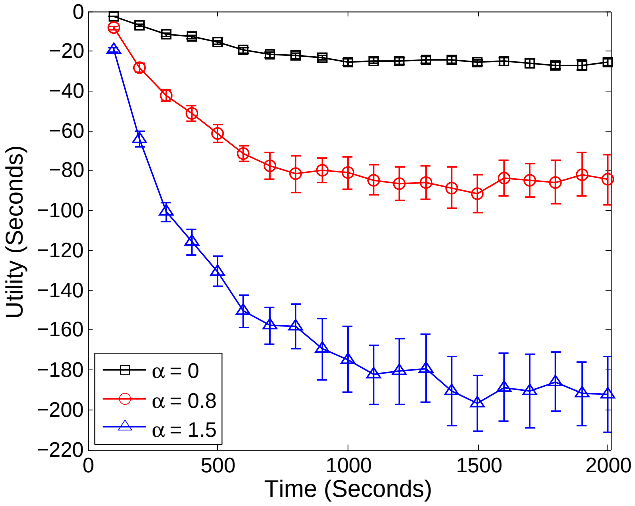


Figure 3.12: Performance with different values of α

3.6 Discussions and Conclusion

In this paper, we have presented the design of a distributed large-scale infrastructure for vehicular content distribution in urban areas. The proposed infrastructure is formed by RSBs deployed across the city which are managed by individual entities at different locations and form an integrated system towards the global system utility. To enable RSBs to work in a fully distributed manner and accordingly make the proposed system scalable to any network size, we have proposed a random-walk based content replication scheme at RSBs. Using extensive simulations, we have validated the superior performance of the proposed scheme.

The proposed infrastructure represents a new and practical solution on building the large-scale content distribution network for mobile users. Within this framework, there exists multiple interesting and open issues:

Connection to Internet: in this paper, we assume that RSBs are not connected to Internet. This can avoid the expensive bandwidth cost to RSB owners and encourage them to deploy RSBs. In practise, certain RSBs can be connected to the Internet and accordingly provide the Internet contents to entire infrastructure. In this case, the location of these Internet-connected RSBs is crucial to reduce the download delay of Internet contents to vehicles and therefore requires appropriate deployment.

Heterogenous Users: the proposed system can be extend to provide content distribution to different mobile users in different environments. For example, the RSBs can be deployed in a shopping mall to distribute store flyers to users with tablets, PDAs and laptops. In this case, different users may have different characteristics of mobility and requirements on the service quality. This dictates the system to take the distinct features of heterogenous user's QoS requirements into considerations.

Security Threat: without central control, the proposed system faces multiple security threats. For example, the buffer storage of RSBs can be abused to store and distribute harmful contents and even virus to vehicles. Moreover, the contents stored in RSBs can also be polluted with the misleading and mismatched content titles. The content population is severe in peer-to-peer networks but has not been addressed in vehicular content distribution networks. To combat the security issues, it is necessary for RSBs to quickly identify and filter the harmful and spam contents.

3.7 Appendix

3.7.1 Proof of Theorem 1

According to Lemma 1, given the caching profile $\mathbf{B}$, we have that the usage of RSB buffer storage X satisfies (3.5). As $v = \sum_{i \in \mathbf{F}} x_i \kappa_i^2 \leq \kappa E(X)$, where v and κ are as defined in Lemma 1. By substituting $v \leq \kappa E(X)$ into (3.5), we have

$$P\left(X \geq E(X) + \psi\right) \leq \exp\left(-\frac{\psi^2}{2\left(v + \kappa\psi/3\right)}\right) \leq \exp\left(-\frac{\psi^2}{2\left(\kappa E(X) + \kappa\psi/3\right)}\right).$$

By assuming $L = E(X) + \psi$ and substituting it into (3.5), we have

$$\Pr\left(X \geq L\right) \leq \exp\left(-\frac{\left(L - E(X)\right)^2}{2\left(\kappa E(X) + \kappa\left(L - E(X)\right)/3\right)}\right).$$

As such, the constraint of (3.4) can be achieved if

$$\exp\left(-\frac{\left(L - E(X)\right)^2}{2\left(\kappa E(X) + \kappa\left(L - E(X)\right)/3\right)}\right) \leq \varepsilon. \tag{A-1}$$

By solving (A-1), we have that the constraint of (3.4) is satisfied if

$$E(X) = \sum_{i \in \mathbf{F}} x_i \kappa_i \leq L - \kappa\frac{2}{3}\log\epsilon - \sqrt{\kappa^2\frac{4}{9}\log^2\epsilon - 2\kappa L\log\epsilon}. \tag{A-2}$$

3.7.2 Proof of Lemma 2

By applying the Taylor series expansion, the second order approximation of r as a function n can be represented as

$$r \approx r|_{\langle n\rangle} + (n - \langle n\rangle) \left.\frac{dr}{dn}\right|_{\langle n\rangle} + \frac{1}{2} (n - \langle n\rangle)^2 \left.\frac{dr^2}{dn^2}\right|_{\langle n\rangle}. \tag{B-1}$$

By taking the expectation on both sides of (B-1) with respect to n, we have

$$\langle r\rangle \approx r|_{\langle n\rangle} + \frac{1}{2}\mathrm{Var}(n) \left.\frac{dr^2}{dn^2}\right|_{\langle n\rangle},$$

where $\mathrm{Var}(n)$ denotes the variance of n.

3.7.3 Proof of Lemma 3

Similar to the proof of Lemma 2, by applying the Taylor series expansion, the second order approximation of R as a function of n represented as

$$R \approx b_i \left(G(\langle n\rangle) + (n - \langle n\rangle) \left.\frac{dG(n)}{dn}\right|_{\langle n\rangle} + \frac{1}{2} (n - \langle n\rangle)^2 \left.\frac{dG^2(n)}{dn^2}\right|_{\langle n\rangle} \right) + (1 - b_i)\, r, \tag{C-1}$$

where $G(n) = \frac{C_{\mathsf{V2R}}}{n+1}$.

By taking the expectation of (C-1) on both sides with respect to n, we have

$$\begin{aligned}\langle R\rangle &\approx b_i \left(G(\langle n\rangle) + \frac{1}{2}\mathrm{Var}(n) \left.\frac{dG^2(n)}{dn^2}\right|_{\langle n\rangle} \right) + (1 - b_i)\, \widetilde{\langle r\rangle} \\ &\approx b_i C_{\mathsf{V2R}} \left(\frac{1}{\langle n\rangle + 1} + \frac{\mathrm{Var}(n)}{(\langle n\rangle + 1)^3} \right) + (1 - b_2)\, \widetilde{\langle r\rangle}.\end{aligned}$$

3.7.4 Proof of Theorem 2

The mean first passage time from state $(0, k)$ can be represented in a recursive manner as

$$\Gamma(0,k) = \frac{1}{\mu + \delta} + \frac{\delta}{\mu + \delta}\Gamma(0, k+1) + \frac{\mu}{\mu + \delta}\Gamma(1, k), \tag{D-1a}$$

$$\Gamma(1,k) = \frac{1}{\lambda + \gamma} + \frac{\gamma}{\lambda + \gamma}\Gamma(1, k+1) + \frac{\lambda}{\lambda + \gamma}\Gamma(0, k), \tag{D-1b}$$

for $0 \le k \le \kappa_i - 1$, and

$$\Gamma\left(0, \kappa_i\right) = \Gamma\left(1, \kappa_i\right) = 0. \tag{D-2}$$

In (D-1a), the first term on the right-hand-side represents the mean time that the tagged node spends in state $(0, k)$. With probability $\frac{\delta}{\mu+\delta}$, the tagged node transits to state $(0, k+1)$ which has the mean first passage time $\Gamma\left(0, k+1\right)$; with the rest probability, it transits to state $(1, k)$ which has the mean first passage time $\Gamma\left(1, k\right)$. (D-1b) is derived in the same manner.

As such, we have

$$\begin{aligned} & \lambda\Gamma\left(0, k\right) + \mu\Gamma\left(1, k\right) \qquad\qquad \text{(D-3)} \\ = \;& \frac{A}{B} + \frac{\lambda+\mu}{B}\left(\lambda\delta\Gamma(0, k+1) + \mu\gamma\Gamma(1, k+1)\right) + \frac{\delta\gamma}{B}\left(\lambda\Gamma(0, k+1) + \mu\Gamma(1, k+1)\right), \end{aligned}$$

where $A = \left(\lambda+\mu\right)^2 + \lambda\gamma + \mu\delta, B = \lambda\delta + \gamma\mu + \delta\gamma$.

In particular, via (D-1a) we have

$$\lambda\delta\Gamma(0, k) + \mu\gamma\Gamma(1, k) = \left(\kappa_i - k\right)\left(\lambda+\mu\right). \tag{D-4}$$

By substituting (D-4) to (D-3), we have

$$\begin{aligned} & \lambda\Gamma\left(0, k\right) + \mu\Gamma\left(1, k\right) \\ = & \frac{A}{B} + \frac{\left(\kappa_i - k - 1\right)\left(\lambda+\mu\right)^2}{\Pi} + \frac{\delta\gamma}{B}\left(\lambda\Delta(0, k+1) + \mu\Delta(1, k+1)\right) \\ = & \cdots \\ = & \sum_{i=0}^{\kappa_i - k - 1}\left(\frac{\delta\gamma}{B}\right)^i \left(\frac{A}{B} + \frac{\left(\kappa_i - k - 1\right)\left(\lambda+\mu\right)^2}{B}\right) \\ = & \frac{1 - \left(\frac{\delta\gamma}{B}\right)^{\kappa_i - k}}{1 - \frac{\delta\gamma}{B}}\left(\frac{A}{B} + \frac{\left(\kappa_i - k - 1\right)\left(\lambda+\mu\right)^2}{B}\right). \end{aligned}$$

By plugging (D-5) into (3.12), we have

$$\tau_i = \frac{1}{\lambda+\mu}\left(\lambda\Gamma\left(0, 0\right) + \mu\Gamma\left(1, 0\right)\right) = \frac{1}{\lambda+\mu}\frac{1 - \left(\frac{\delta\gamma}{B}\right)^{\kappa_i}}{1 - \frac{\delta\gamma}{B}}\left(\frac{A}{B} + \frac{\left(\kappa_i - 1\right)\left(\lambda+\mu\right)^2}{B}\right). \tag{D-5}$$

As $\frac{\delta\gamma}{B} < 1$, when κ_i is large, we have $\left(\frac{\delta\gamma}{B}\right)^{\kappa_i} \approx 0$, and accordingly,

$$\tau_i = \frac{1}{\lambda+\mu}\frac{1}{1-\frac{\delta\gamma}{B}}\left(\frac{A}{B} + \frac{(\kappa_i - 1)(\lambda+\mu)^2}{B}\right) = \frac{\gamma\lambda + \delta\mu + \kappa_i(\lambda+\mu)^2}{\gamma\mu(\lambda+\mu) + \lambda\delta(\lambda+\mu)} \tag{D-6}$$

By substituting (3.11), $\widetilde{\langle R\rangle} = \gamma$ and $\widetilde{\langle r\rangle} = \delta$ into (D-6), we have

$$\tau_i = \frac{b_i(\Phi-\delta)\lambda + \delta(\lambda+\mu) + \kappa_i(\lambda+\mu)^2}{b_i(\Phi-\delta)\mu(\lambda+\mu) + \delta(\lambda+\mu)^2}.$$

3.7.5 Proof of Corollary 1

According to Theorem 2, we obtain the first and second order derivative of τ_i as

$$\begin{aligned}
\frac{d\tau_i}{db_i} &= -\frac{(\Phi-\delta)\left[\kappa_i\mu(\lambda+\mu) + \delta(\mu-\lambda)\right]}{\left[b_i\mu(\Phi-\delta) + (\lambda+\mu)\right]^2} \\
\frac{d^2\tau_i}{db_i^2} &= \frac{2\mu(\Phi-\delta)^2\left[\kappa_i\mu(\lambda+\mu) + \delta(\mu-\lambda)\right]}{\left[b_i\mu(\Phi-\delta) + \delta(\lambda+\mu)\right]^3}
\end{aligned}$$

The download delay τ_i is a convex function if $\frac{d^2\tau_i}{db_i^2} \geq 0$, i.e., $\frac{\kappa_i}{\delta} \geq \frac{1}{\mu} - \frac{2}{\lambda+\mu}$.

3.7.6 Proof of Proposition 1

Evaluating the first and second derivatives of $U(\tau_i)$ on b_i, we have

$$rCl\frac{dU(\tau_i)}{dx_i} = \frac{dU(\tau_i)}{d\tau_i}\frac{d\tau_i}{dx_i} \tag{F-1}$$

$$\frac{d^2U(\tau_i)}{dx_i^2} = \frac{d^2U(\tau_i)}{d\tau_i^2}\left(\frac{d\tau_i}{dx_i}\right)^2 + \frac{dU(\tau_i)}{d\tau_i}\frac{d^2\tau_i}{dx_i^2} \tag{F-2}$$

Since $\frac{dU(\tau_i)}{d\tau_i} \leq 0$ and $\frac{d\tau_i}{db_i} \leq 0$, with (F-1a) we have $\frac{dU(\tau_i)}{db_i} \geq 0$. Since $\frac{d^2U(\tau_i)}{d\tau_i^2} \leq 0$, $\frac{dU(\tau_i)}{d\tau_i} \leq 0$ and $\frac{d^2\tau_i}{db_i^2} \geq 0$, with (F-1b), we have $\frac{d^2U(\tau_i)}{db_i^2} \leq 0$, and therefore, $U(\tau_i)$ is a concave function of b_i. As the system utility W, in general, is the weighted sum of the $U(\tau_i)$ over $i \in \mathbf{F}$, we have $\frac{d^2W}{db_i^2} = w_i\frac{d^2U(\tau_i)}{db_i^2} \leq 0$. Therefore, W is also a concave function of b_i.

3.7.7 Optimal Solution of (3.7) According to KKT Conditions

As (3.7) is a convex optimization problem, the KKT conditions are both necessary and sufficient for the optimal solution. Let b_i^* denote the optimal solution of (3.7). Introducing Lagrangian multiplier ϖ for the constraint in (3.7), we list the KKT conditions of (3.7) as follows:

$$\left.\frac{\partial \mathcal{U}}{\partial b_i}\right|_{b_i^*} - \varpi\kappa_i = 0 \tag{G-1}$$

$$\varpi\left(\mathcal{L} - \kappa_i b_i^*\right) = 0 \tag{G-2}$$

$$\sum_{i\in\mathbf{F}} \kappa_i b_i^* \leq \mathcal{L} \tag{G-3}$$

$$b_i^*, \gamma \geq 0,\ i \in \mathbf{F} \tag{G-4}$$

where $\mathcal{U} = -\sum_{i\in\mathbf{F}} p_i\tau_i$ as specified in (3.1) and (3.14)

Assume that $\varpi = 0$. Substituting it into (G-1), we have $p_i \left.\frac{\partial\tau_i}{\partial b_i}\right|_{b_i^*} = 0$. This is not feasible as $\frac{\partial\tau_i}{\partial b_i} < 0$ and $p_i > 0$ for all $i \in \mathbf{F}$. Therefore, from (G-2) we have $\gamma > 0$ and $\mathcal{L} - \sum\limits_{i\in\mathbf{F}} b_i^* = 0$.

Substituting (3.13) into (G-1), we have

$$b_i^* = \frac{1}{\mu\left(\Phi - \delta\right)}\sqrt{\frac{p_i\left(\Phi-\delta\right)\left[\kappa_i\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}{\varpi\kappa_i}} - \frac{\delta\left(\lambda+\mu\right)}{\mu\left(\Phi-\delta\right)}. \tag{G-5}$$

Together with $\mathcal{L} - \sum\limits_{i\in\mathbf{F}} b_i^* = 0$, we have

$$\begin{aligned} b_i^* \quad = \quad & \mathcal{L}\frac{\sqrt{p_i\left[\kappa_i\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}}{\sum_{j\in\mathbf{F}}\kappa_j\sqrt{p_j\left[\kappa_j\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}} \\ & + \left(\frac{\sqrt{p_i\left[\kappa_i\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}\sum_{i\in\mathbf{F}}\kappa_i}{\sum_{j\in\mathbf{F}}\kappa_j\sqrt{p_j\left[\kappa_j\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}} - 1\right)\cdot\frac{\delta\left(\lambda+\mu\right)}{\mu\left(\Phi-\delta\right)}, \end{aligned} \tag{G-6}$$

and

$$\varpi = \left(\frac{\sum_{i\in\mathbf{F}}\sqrt{\kappa_i p_i\left[\kappa_i\mu\left(\lambda+\mu\right)+\delta\left(\mu-\lambda\right)\right]}}{\mathcal{L}\mu+\delta\left(\lambda+\mu\right)\sum_{i\in\mathbf{F}}\kappa_i}\right)^2$$

3.7.8 Transition Probability in the Random Walk Algorithm

The target of the random walk algorithm is to select a file i from the file graph with the probability b_i^{d} shown in (3.16).

Assume that there are totally V nodes presenting in the network and a_i of them having file i stored. Therefore, there are totally $a_i\,|V|$ copies of file i in the file graph. To select file i with probability b_i^{d}, one should sample each copy of file i in the graph with probability

$$\pi_i = \frac{b_i^{\mathsf{d}}}{a_i V}. \tag{H-1}$$

Using the Metropolis-Hasting algorithm, the transition of random walk constitutes two steps. In the first step, a candidate file, e.g., m, is selected from the neighboring files of the current file, e.g., n, which holds the walker based on the proposal probability

$$\alpha_{mn} = \frac{1}{s_m + 1}, \tag{H-2}$$

where s_m denotes the fanout of file m in the file graph. A neighboring file of file n is the file which is connected to file n in the file graph.

In the second step, file m is accepted as the next hop of the walker with the acceptance probability as

$$q_{mn} = \min\left\{\frac{\pi_n \alpha_{nm}}{\pi_m \alpha_{mn}}, 1\right\} = \min\left\{\frac{a_m b_n^{\mathsf{d}}\,(s_m+1)}{a_n b_m^{\mathsf{d}}\,(s_n+1)}, 1\right\}, \tag{H-3}$$

with the rest probability the walk will sojourn in file n for one hop.

To summarize, the transition probability from file m to file n is

$$P_{mn} = \alpha_{mn} q_{mn} = \begin{cases} \frac{1}{s_m+1}\min\left\{\frac{a_m b_n^{\mathsf{d}}(s_m+1)}{a_n b_m^{\mathsf{d}}(s_n+1)}, 1\right\}, & m \neq n, \\ 1 - \sum\limits_{m\neq n} P_{mn}, & m = n. \end{cases} \tag{H-4}$$

Chapter 4

MAC in Motion: Analysis of MAC Performance in the Highly Mobile Drive-Thru Internet

4.1 Introduction

Catering to the ever-increasing demand of ubiquitous Internet access and motivated by the widespread adoption of Wireless LAN, the drive-thru Internet has recently been proposed in [10] which adopts the "grass root" IEEE 802.11 access points (APs) deployed along the road in place of RSUs as shown in Figure 1.1 to provide Internet access to vehicles on the move. Through the real-world experiment (see Chapter 2 for details), it is shown in [10] that using the off-the-shelf IEEE 802.11b hardware, a vehicle could maintain a connection to a roadside AP for around 500 m and transfer 9 MB of data at 80 km/h using either TCP or UDP. CarTel in MIT [7] further extends the drive-thru Internet with city-wide trials in Boston. It is shown that the plethora IEEE 802.11b APs deployed in cities could provide vehicle nodes with the *intermittent* and *short-lived* connectivity, yet high throughput when the connectivity is available. Similar properties of the drive-thru Internet are also reported separately in [40, 41]. Meanwhile, prominent automobile corporations have also lunched important projects using the similar architecture for promoting vehicular Internet communications. For instance, Mercedes-Benz proposes to deploy the "InfoFuel" stations along the roads to fuel on-road vehicles with the high throughput Internet access using the IEEE 802.11a radio [4].

While being seriously pursued, the performance of IEEE 802.11 in the *high-speed large-*

scale drive-thru Internet scenario is still unclear due to the following reasons. *First*, compared with the small-scale indoor scenarios, the drive-thru Internet is typically a much larger network composed of tens or hundreds of users. Previous works in [10]-[41] largely adopt the experimental approach; limited by the hardware, their results are attained in small-scale networks only and can hardly provide insights into the large-scale case when a great number of vehicles compete for communications simultaneously. Therefore, we argue that a thorough theoretical framework which is accurate and scalable to different network scales is necessary to guide the real-world deployments. *Second*, originally designed for low mobility scenarios, the IEEE 802.11 adopts the contention-based distributed coordination function (DCF) as its MAC in which the transmission opportunity of stations are rendered in an opportunistic manner (refer to Section 4.3 for details). In the case of drive-thru Internet, as vehicles have volatile connectivity due to the fast mobility, whether DCF can fully utilize the cherished access time of users and provide them the guaranteed throughput is questionable. As the previous theoretical studies on DCF [42, 43] mainly focus on the static WLAN scenarios *without taking the node mobility into consideration*, they are not applicable to the drive-thru Internet scenario.

In this chapter, we focus on the investigation of the DCF performance in the highly mobile driven-thru Internet scenario by considering high node mobilities [44]. In particular, we aim at addressing the following questions: *what is the performance of DCF in the high-speed large-scale drive-thru Internet; in what fashion does the mobility affect the MAC throughput and, more importantly, how to remedy that*? Note that the newly emerged IEEE 802.11p WAVE standard adopts the DCF-based IEEE 802.11e EDCA scheme as the MAC of vehicular communications, to understand the performance of the fundamental DCF MAC is crucial to the real-world deployment of VANETs. On addressing these issues, we provide a systematic and theoretical treatment based on a Markov chain model which incorporates the mobility of vehicles in the analysis of DCF. Based on the Markov model, we unveil the impacts of mobility (characterized by the node velocity and moving direction) on the resultant system throughput and describe the optimal configuration of DCF to mitigate the negative effects of mobility towards best system performance. Our main contributions are two-fold:

- ▷ *Performance Evaluation*: We propose an accurate and scalable model to analytically evaluate the impacts of node mobility on the achievable system throughput in drive-thru Internet scenarios. The accuracy of the analytical model is demonstrated by extensive simulations. Moreover, we show that the throughput performance is solely dependent on the node velocity. Since velocity can be easily measured, vehicles are able to conveniently assess their throughput with local information only and then optimize the MAC in a fully distributed manner.

▷ *Protocol Enhancement*: Based on the developed model, we propose to further enhance the MAC throughput by adaptively adjusting the MAC in tune with the node mobility. In particular, we propose three guidelines of the DCF design in the highly mobile vehicular environment, and describe the optimal schemes to determine the channel access opportunity to fully utilize the transient connectivity of vehicles.

The remainder of this section is organized as follows: Section 5.2 provides an survey of related works and compares our work with that from the existing literature. Section 4.3 provides an overview of DCF in and discuss the problems when directly implementing it in vehicular communications. Section 5.3 describes the proposed analytical model in detail and Section 4.5 validates the accuracy of the analytical model using simulations. Section 4.6 proposes several enhancement schemes to boost the performance of DCF by accommodating the high mobility of nodes, and Section 5.7 closes the section with summaries.

4.2 Related Works

In this section, we highlight our contributions in the light of previous works.

Inspired by the pioneer work in [10], the drive-thru Internet has been further investigated in numerous measurement studies from different aspects [7, 40, 45]. While the measurement studies shed insightful lights for the real-world deployments, their focus is mainly on the link quality and transport performance between a vehicle and series of APs passed through. As a result, they do not consider the MAC layer contention when multiple drive-thru vehicles concurrently transmit and compete for the transmission resource. Even with promising link performance as shown in [7, 40], a coarse MAC would result in severe collisions and chaos of transmissions to the connection-limited vehicles; therefore, the elaborate analysis of MAC deserves.

In parallel to the measurement studies, a collection of works are devoted to improve the performance of drive-thru Internet from MAC [46, 47], routing [48], transport [49] and application layer [27]. Zhang et al. [46] proposes a cooperative MAC, namely VC-MAC, for vehicle communications which incorporates the cooperative relays among vehicles with the vehicle to roadside infrastructure communication. By harvesting the spatial and path diversity, VC-MAC significantly improves the throughput and service coverage to volatile fast-moving vehicles. Sikdar [47] devises a reservation-based MAC with the emphasis on the handoff among APs. Upon the arrival to a new AP, a node first waits for the beacon message from the AP which notifies the available transmission slots to vehicles. After receiving the beacon message, the node then requires to associate with the AP and reserves a time slot for transmission. In contrast to [46, 47], rather than proposing new MAC schemes with distinguished features, we target to an in-depth understanding of the legacy IEEE 802.11 DCF in the newly emerged vehicular environment. The reason is two-fold. First, DCF is the most practical and adopted MAC currently with the broad compatibility to various portable devices in different networks, e.g., hotspot networks in trains and buses [50]. Second, it is widely used in various projects like Fleetnet [51] and DieselNet [48] with proven performance.

On the other hand, some research works focus on investigating the impacts of node mobility on the throughput performance of drive-thru Internet. In [52], Tan et al. develop an analytical model to evaluate the download volume of vehicles per each drive-thru. Assuming the optimal MAC and fair share of airtime, the throughput of each vehicle is computed by averaging the service rate of AP on the population of vehicles. Since the population of vehicles on the road varies over time, the throughput of each node is stochastic and its density function is derived based on a Markov model. [52] considers the network as a flow of nodes. In comparison, our work investigates the throughput from a microscopic view by standing at the viewpoint of individual vehicles. Moreover, unlike [52] which assumes perfect MAC, we model the specific DCF in details and show the quantified impacts of mobility on the MAC throughput.

Furthermore, an extensive body of research has been devoted to the performance evaluation of IEEE 802.11 DCF for WLAN communications [42, 43]. However, as those works mainly focus on the indoor environment with small-scale and static stations, the examine of DCF in the high speed large-scale vehicular environment deserves a fresh treatment. To support the vehicular communications, the IEEE working group has recently proposed IEEE 802.11p [53] as a draft amendment to the IEEE 802.11 standard, namely WAVE (Wireless Access for Vehicular Environments). The new standard adopts IEEE 802.11e EDCA as the MAC. As DCF is the basis of EDCA, our analytical model can be easily extended to study 802.11p. We have considered a simple case in this work to better explain the theory.

4.3 DCF in the Drive-Thru Internet

Using DCF, each node with packets to transmit monitors the availability of the channel. If the channel is sensed idle for a period of distributed interframe space (DIFS), the transmission may proceed; otherwise, the node will wait until the end of the in-progress

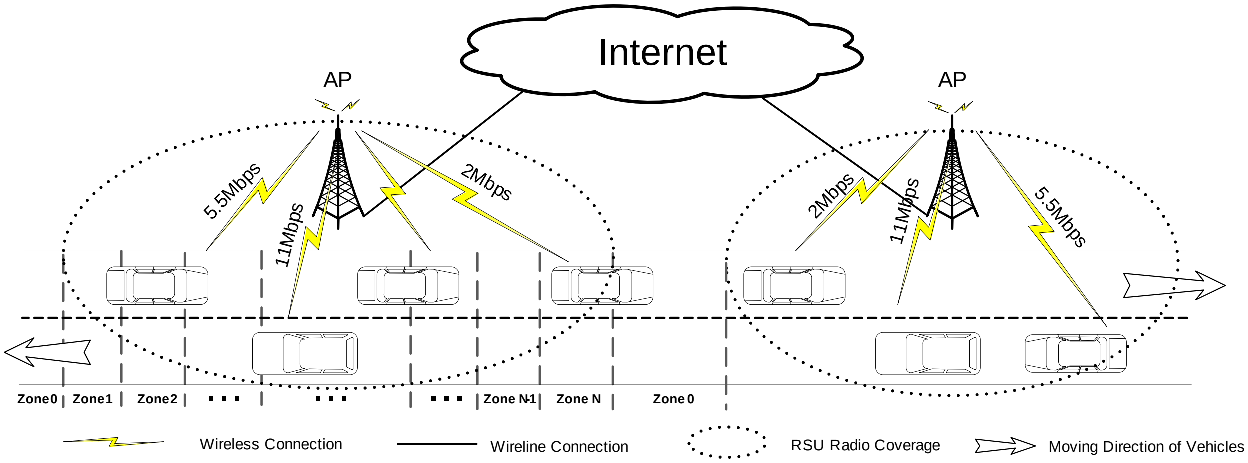


Figure 4.1: Drive-thru Internet in which the radio coverage of AP is divided into multiple zones according to the data modulation rates

transmission. To avoid the case that multiple nodes transmit simultaneously when the channel is released idle, DCF adopts the collision avoidance (CA) mechanism. Specifically, before transmission, each node uniformly selects a random discrete backoff time from the range $[0, W-1]$, where W is called the Contention Window (CW). To transmit packets after DIFS, a node first reduces the backoff time with constant step δ, and transmits only if the backoff time is 0. The countdown of backoff time is frozen once the channel becomes busy due to other node transmission, and resumes until the channel is idle for another DIFS. The size of CW, W, depends on the history of transmissions. At the first transmission attempt, W is set to a predefined value CW_{min}, the *minimum contention window*. Upon each unsuccessful transmission s, W is updated as $W = 2^s CW_{min}$ until W reaches a maximum value CW_{max}. s here is called backoff stage. More details of DCF can be found in [54].

The advantages of DCF are salient: *First*, it is fully distributed, which is particularly desirable in vehicular communications. As frequent handoffs and topology changes are made due to the high node mobility, the distributed behavior of DCF makes the system quite robust. *Second*, thanks to the binary exponential backoff, DCF is scalable and could be implemented for different traffic and road environments, e.g., urban and rural regions.

However, originally designed for stationary indoor networks, when used for the in-motion vehicular communications, the performance of DCF highly depends on the mobility of nodes, as we show in the following sections.

Moreover, with nodes at different locations to an AP, their channel conditions diverse,

Table 4.1: Summary of Notations

Symbols Associated with Zones and Vehicle Traffic	
$\mathbb{Z}$	Set of spatial zones in the coverage of an AP. z is an element in $\mathbb{Z}$.
r_z	Payload transmission rate (in Mbps) of vehicle node to AP at zone z.
d_z	Length of the spatial zone z (in meters).
λ	Mean arrival rate of vehicles to the road segment.
v	Mean velocity of vehicles (in km/h) in the road segment.
n_L	Number of lanes in the road segment.
k	Density of vehicles (in veh/km/lane) along the road segment.
k_{jam}	Traffic jam density (in veh/km/lane) at which the traffic flow comes to a halt.
v_f	Free-flow speed (in km/h).

Symbols Associated with Transmissions and Backoffs	
W_z	The minimum contention window size CW_{min} associated in zone z.
m	The maximum backoff stage.
τ_z	Conditional transmission probability of nodes in zone z.
$\pi_{z,s,b}$	Steady state probability of a vehicle in zone z with the backoff time and backoff stage equal to b and s, respectively.
T_{dec}	Time that the backoff time of the tagged node deducts by one.
p_{col}	Collision probability of the tagged node.
p_{suc}	Conditional probability that the in-progress transmission is successful given that the channel is busy.
$p_{\text{suc},z}$	Conditional probability that the in-progress transmission is by a node in zone z, given that the transmission is successful.
$p_{\text{col},z}$	Conditional probability that the in-progress transmission is collided and the longest transmission in the collision is from zone z.
$p_{\text{hcol},z}$	The homogenous collision probability.
$p_{\text{dcol},z}$	The diverse collision probability.
$E\left[Tx_{\text{suc},z}\right]$	The mean time of the successful transmission of the tagged node in zone z.
$E\left[Tx_{\text{col},z}\right]$	The mean time of the collided transmission of the tagged node in zone z.
$E\left[T_{\text{suc}}\right]$	The mean time of the successful in-progress transmission during the backoff of the tagged node.
$E\left[T_{\text{col}}\right]$	The mean time of the collided in-progress transmission during the backoff of the tagged node.
$T_{\text{suc},z}$	Successful transmission time of a packet from zone z.
T_{col}	Collision time of the in-progress transmission in the backoff of the tagged node.
s_z	Normalized nodal throughput of vehicles in zone z.
S	Overall throughput of nodes in the system.

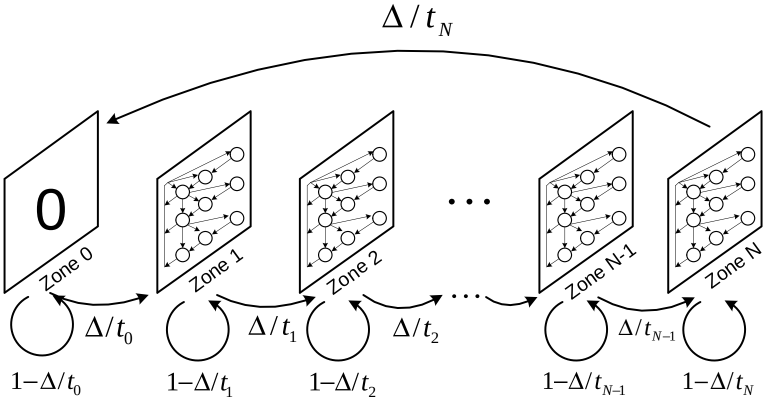


Figure 4.2: Three-dimensional Markov model for vehicle nodes

resulting in different data rates for reliable transmissions, as shown in Figure 4.1. In this case, DCF suffers from the *performance anomaly*, i.e., the system throughput is throttled to the minimum transmission rate among nodes [55]. To boost the throughput performance, existing literatures [56, 57, 58] largely adapt CWs according to node transmission rates. By assigning high-rate nodes the relatively small CWs and high packet transmission probability, the system throughput could be enhanced. Hadaller et al. [59] first consider the performance anomaly in the drive-thru Internet and propose a greedy algorithm where only nodes with the best SNR are allowed to transmit. Unlike [59], in this work, we provide a thorough theoretical study.

4.4 System Model and Throughput Evaluation

This section details our analytical model for the evaluation of DCF in the highly mobile drive-thru Internet scenario. The many symbols used in this chapter have been summarized in Table 4.1.

4.4.1 System Model

We consider the drive-thru Internet scenario, as shown in Figure 4.1, with nodes connecting to intermittent and serial APs along the road. We focus on the MAC layer under the assumption of perfect channel conditions (i.e., no transmission errors and hidden terminals) with line-of-sight communications. This assumption is typical in literature [42, 60, 56] to

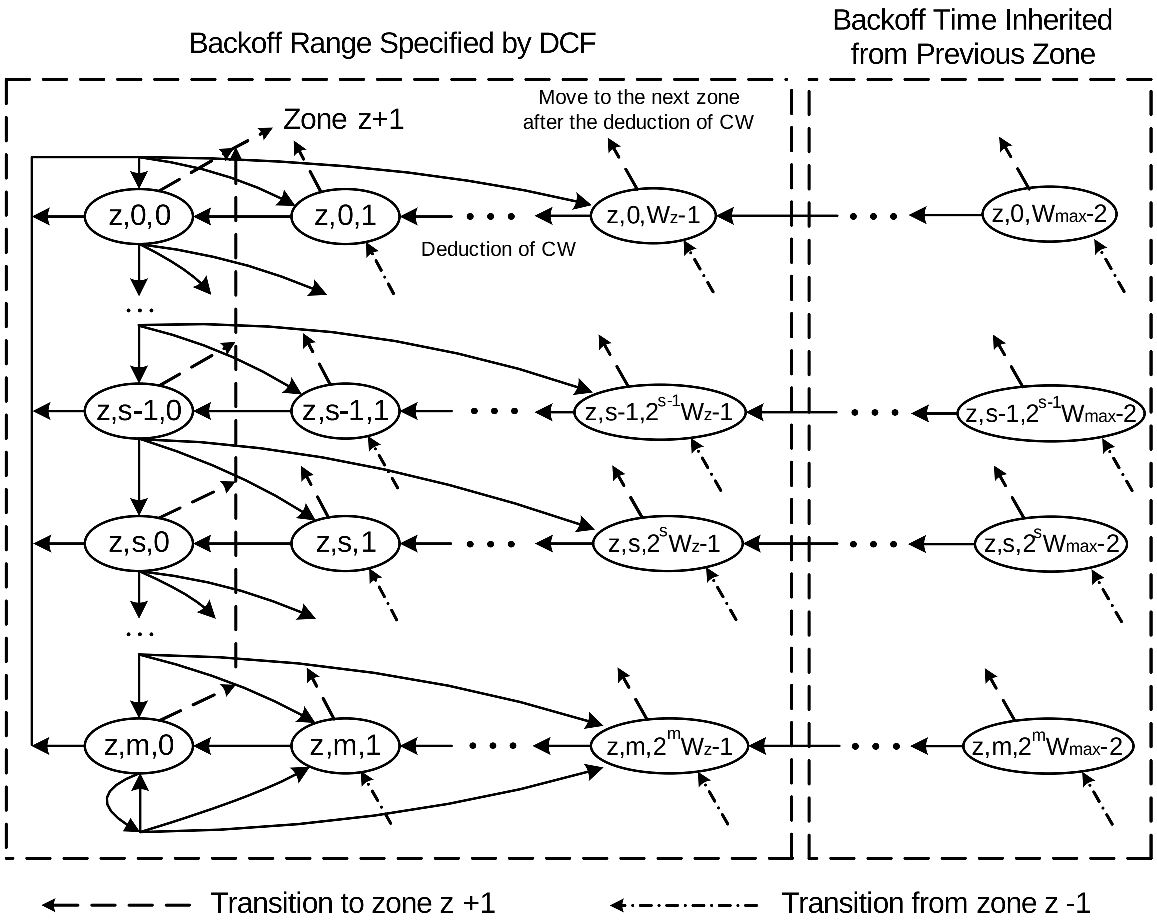


Figure 4.3: State space of CW in spatial zone z

evaluate the MAC performance. In this case, the SNR and modulation rates of vehicles are mainly determined by their distance to the AP. Field tests have validated the assumption by showing the strong correlation between distance and transmission rate in vehicular environment [10, 40, 61].

Without loss of generality, we divide the road into multiple spatial zones as shown in Figure 4.1. The session outside the coverage of APs is denoted by zone 0. Within the radio coverage of an AP, the road is divided into multiple zones denoted as $\mathbb{Z} = \{1, 2, .., N\}$ such that within each zone z, $z \in \mathbb{Z}$, vehicles have distinct payload transmission rates, denoted by r_z, according to their distance to AP. Let d_z denote the length of each spatial zone $z \in \mathbb{Z}$. With nodes traversing consecutive APs along the road, they are regarded to transit iteratively among the zones in $\mathbb{Z}$. The mobility of vehicles is then represented by the zone transitions using a Markov chain model (inspired by [62]) as shown in Figure 4.2 in which each state corresponds to one spatial zone. The time that nodes stay in each zone $z \in \mathbb{Z}$ is assumed to be geometrically distributed with mean duration of t_z, which is determined by the length of the partition zone and the average velocity, v, of vehicle nodes as $t_z = d_z/v$. As such, within a small duration, e.g., Δ, vehicles either move to the next zone with probability Δ/t_z, or remain in the current zone with the rest probability $1 - \Delta/t_z$. The limiting probability that a node is in zone z at any time is then $d_z/\sum_{n\in\mathbb{Z}} d_n$. With this model, the road could be of multiple bidirectional lanes[1], and nodes are allowed to have varying speeds but constant mean value.

Within the communication range of APs, packet transmissions are coordinated by the DCF scheme as described in Section 4.3. We consider the *saturated case* in that each node always has a packet to transmit. The packet length L is assumed to be fixed and same for all the nodes. To address the performance anomaly, we set CW_{min} dependent on the zones such that nodes in different zones transmit with differentiated probabilities. Let W_z denote the CW_{min} of nodes in zone z. Let m denote the maximum number of backoff stage in DCF, which is set to 7 by default in standard [54]. Throughout the work, we assume nodes are homogeneous and abide to the same W_z in zone z. We resist considering the general formulation with service differential to nodes as it could be obtained easily by extending the developed model and, more importantly, it risks making the model difficult to understand.

4.4.2 Markov Model of Moving Vehicles

To evaluate the DCF performance of individual vehicles, we examine a randomly tagged vehicle and represent its status by a three-dimensional Markov chain $\{Z(t), S(t), B(t)\}$ at time slot t. $Z(t)$ denotes the spatial zone that the node is currently in. $S(t)$ denotes the current backoff stage of the tagged node using DCF. $B(t)$ denotes the backoff time of the tagged node at the current time slot. A discrete and integer scale time is applied, where slot times t and $t+1$ correspond to the beginning of two consecutive backoffs of the tagged node. In other words, the Markov chain is embedded in the countdown of the backoff time. The principle of the three-dimensional Markov chain is sketched in Figure 4.2. Similar to [42], it is important to note that this discrete time does not directly map to the real system time; the duration between any two time slots is a random variable as the backoff time of the tagged node could be frozen for a random period.

[1] Due to the symmetric locations and payload transmission rates of zones along the AP, vehicles along different directions can be modeled using the same Markov chain.

Figure 4.3 plots the state transitions when the tagged node is in zone z. Here, $W_{\max}$ is the maximal W_z among all zones, i.e., $W_{\max} = max\{W_z | z \in \mathbb{Z}\}$. As shown in Figure 4.3, upon each transition, the tagged node would have its backoff time deducted by one. Meanwhile, the tagged node would move to the next zone probabilistically based on the mobility model described in the previous subsection. When the backoff time deducts to zero, the tagged node would initiate one transmission attempt. If the transmission is collided, the tagged node would backoff and selected a new backoff time based on the DCF mechanism as specified in Section 4.3; otherwise, the backoff stage is cleared to zero. After the transmission attempt, either successful or failed for transmission, the tagged node is possible to move to the next zone and select the backoff time based on the contention window size in the newly arrived zone.

As such, our model is distinct from Bianchi's [42] by considering the node mobility in three aspects. Firstly, after the deduction of the backoff time $B\left(t\right)$, the tagged node either stays in the current zone or moves to the next zone with renewed CW_{min} and transmission probabilities. Secondly, when the tagged node moves to a new zone, its backoff time $B(t)$ reduces smoothly as $B(t) = B(t-1) - 1$, if $B(t-1) \neq 0$, unrelated to zones. Therefore, if $B(t)$ is large, even though the tagged node arrives at a new zone with a very small CW_{min} in the next time slot, it can not benefit immediately. Lastly, the backoff stage is inherited when switching to the next zone with $S(t) = S(t+1)$, if the tagged node does not transmit during the zone transition. In other words, if the tagged node encounters severe collisions, the transmission history will be inherited in the new zone.

Given t_z, $z \in \mathbb{Z}$, the one-step non-null transition probabilities of the Markov chain from time slot t to $t+1$ are as follows:

(i) Arriving at AP (from zone 0 to zone 1):

$$P\left(1, 0, b | \mathbb{0}\right) = \frac{E\left[T_{\text{dec}}\right]}{t_0 W_1}, \quad b \in \left[0, W_1 - 1\right], \tag{H-1}$$

where $\mathbb{0}$ represents zone 0, and $E\left[T_{\text{dec}}\right]$ is the mean duration of one time slot given that the tagged node is not transmitting. $P\left(1, 0, b | \mathbb{0}\right)$ in (H-1) accounts for the transition probability that the tagged node moves from zone 0 to zone 1 and selects the backoff time b from the range $[0, W_1 - 1]$. This is because that within one time slot, with probability $E\left[T_{\text{dec}}\right] / t_0$, the tagged node move from zone 0 to zone 1 according to the geometrically distributed sojourn time in each zone. After reaching zone 1, the tagged node selects the initial $B(t)$ uniformly from $[0, W_1 - 1]$. As the zone transition and backoff time selection are independent, the overall transmission probability is hence $\frac{E[T_{\text{dec}}]}{t_0 W_1}$. In this work, we turn down the DCF in zone 0 – the backoff time set to infinity and the backoff stage cleared to

0 – as in this case nodes are out of the transmission range. As such, nodes in zone 0 have only one state whereas those in other zones have multiple states with different values of backoff time and stage.

(*ii*) Within the AP coverage (in zones 1 to N): Eq. (4.2) shows the transition probabilities when the tagged node is in the coverage of AP, where p_{col} is the collision probability when the tagged node transmits. $E\left[Tx_{\text{suc},z}\right]$ and $E\left[Tx_{\text{col},z}\right]$ are the mean time of one successful and collided transmission of the tagged node in zone z, respectively.

$$P\left(z,s,b|z,s,b+1\right)=1-\frac{E\left[T_{\text{dec}}\right]}{t_z}, \qquad z\in\left[1,N\right],s\in\left[0,m\right],b\in[0,2^sW_{\max}-1), \tag{H-2a}$$

$$P\left(z,s,b|z-1,s,b+1\right)=\frac{E\left[T_{\text{dec}}\right]}{t_{z-1}}, \qquad z\in\left[2,N\right],s\in\left[0,m\right],b\in[0,2^sW_{\max}-1), \tag{H-2b}$$

$$P\left(z,0,b|z,s,0\right)=\frac{1-p_{\text{col}}}{W_z}\left(1-\frac{E\left[Tx_{\text{suc},z}\right]}{t_z}\right), \qquad z\in\left[1,N\right],s\in\left[0,m\right],b\in\left[0,W_z-1\right], \tag{H-2c}$$

$$P\left(z,0,b|z-1,s,0\right)=\frac{1-p_{\text{col}}}{W_z}\frac{E\left[Tx_{\text{suc},z-1}\right]}{t_{z-1}}, \qquad z\in\left[2,N\right],s\in\left[0,m\right],b\in\left[0,W_z-1\right], \tag{H-2d}$$

$$P\left(z,s,b|z,s-1,0\right)=\frac{p_{\text{col}}}{2^sW_z}\left(1-\frac{E\left[Tx_{\text{col},z}\right]}{t_z}\right), \quad z\in\left[1,N\right],s\in[0,m),b\in\left[0,2^sW_z-1\right], \tag{H-2e}$$

$$P\left(z,s,b|z-1,s-1,0\right)=\frac{p_{\text{col}}}{2^sW_z}\frac{E\left[Tx_{\text{col},z-1}\right]}{t_{z-1}} \qquad z\in\left[2,N\right],s\in[0,m),b\in\left[0,2^sW_z-1\right], \tag{H-2f}$$

$$P\left(z,m,b|z,m,0\right)=\frac{p_{\text{col}}}{2^mW_z}\left(1-\frac{E\left[Tx_{\text{col},z}\right]}{t_z}\right), \qquad z\in\left[1,N\right],b\in\left[0,2^mW_z-1\right], \tag{H-2g}$$

$$P\left(z,m,b|z-1,m,0\right)=\frac{p_{\text{col}}}{2^mW_z}\frac{E\left[Tx_{\text{col},z-1}\right]}{t_{z-1}}, \qquad z\in\left[2,N\right],b\in\left[0,2^mW_z-1\right], \tag{H-2h}$$

$P\left(z,s,b|z,s,b+1\right)$ in (H-2a) accounts for the probability that the tagged node remains in the original zone z after its backoff time deducts by one. $P\left(z,0,b|z,s,0\right)$ in (H-2c) accounts for the probability that the tagged node transmits successfully and starts a new round of backoff. $1-\frac{E\left[Tx_{\text{col},z}\right]}{t_z}$ is the probability that the tagged node remains in the same zone during the collision time. $P\left(z,s,b|z,s-1,0\right)$ in (H-2e) accounts for the probability that the tagged node encounters the collision and backoffs by one stage, all in the original zone. In this scenario, with probability p_{col} that the transmission is collided and with probability $1/2^sW$ that a random backoff interval b is selected within the range $[0,2^sW-1]$. With probability $1-\frac{E[Tx_{\text{suc},z}]}{t_z}$, the tagged node does not switch zones in this slot time. (H-2g) shows the transition probabilities when the backoff stage reaches its upper bound m. (H-2b), (H-2d), (H-2f) and (H-2h) are the transition probabilities that the tagged node moves to the next zone in the new slot time.

(iii) Departing the AP (from zone N to zone 0),

$$P\left(\mathbb{0}|N,s,b\right)=\frac{E\left[T_{\text{dec}}\right]}{t_N},\quad s\in[0,m]\,,b\in\left[1,2^sW_{\max}-1\right], \tag{H-3a}$$

$$P\left(\mathbb{0}|N,s,0\right)=\frac{\left(1-p_{\text{col}}\right)E\left[Tx_{\text{suc},N}\right]+p_{\text{col}}E\left[Tx_{\text{col},N}\right]}{t_N},\quad s\in[0,m]\,. \tag{H-3b}$$

Eq. (4.3) indicates the transition probabilities that the tagged node departs from the zone N and enters zone 0 (out of AP coverage). In these transitions, (H-3a) is obtained in the same manner of (H-2a). (H-3b) accounts for the probability that the tagged node moves out of zone N after it transmits where $\left(1-p_{\text{col}}\right)E\left[Tx_{\text{suc},N}\right]+p_{\text{col}}E\left[Tx_{\text{col},N}\right]$ is the mean duration of its transmission time.

Let $\pi_{z,s,b}=\lim_{t\to\infty}\Pr\{Z(t)=z,S(t)=s,B(t)=b\}$ be the steady state probability of the Markov chain and $\pi=\{\pi_{z,s,b}\}$ denote the corresponding vector. Given the state transition probability matrix $\mathbf{P}$ with each non-null element shown in (4.1), (4.2) and (4.3), $\pi_{z,s,b}$ could be derived with the following balance equations

$$\begin{cases}\pi\mathbf{P}=\pi,\\ \pi_{\mathbb{0}}+\displaystyle\sum_{z=1}^{N}\sum_{s=0}^{m}\sum_{b=0}^{2^sW_{\max}-1}\pi_{z,s,b}=1.\end{cases} \tag{H-4}$$

4.4.3 Packet Transmission Time in the Contention

To solve (H-4), we first consider the expressions of $E\left[T_{\text{dec}}\right]$ and $E\left[Tx_{\text{col},z}\right]$ and $E\left[Tx_{\text{suc},z}\right]$ in (4.1), (4.2) and (4.3).

Let X denote the mean node population in the road segment[2], excluding the tagged node. Let λ denote the mean arrival rate of nodes to the road segment. According to Little's law,

$$X = \lambda \frac{\sum_{z\in\mathbb{Z}} d_z}{v} - 1, \tag{H-5}$$

where $\sum_{z\in\mathbb{Z}} d_z/v$ is the mean sojourn time of nodes in the coverage of AP. Let X_z denote the number of nodes in zone z, excluding the tagged node, then

$$X_z = \frac{X d_z}{\sum_{n\in\mathbb{Z}} d_n}, \tag{H-6}$$

where $d_z/\sum_{n\in\mathbb{Z}} d_n$ is the limiting probability that a node is in zone z.

Denote by τ_z the conditional transmission probability given that nodes are in zone z. Mathematically we have

$$\tau_z = \frac{\sum_{s\in[0,m]} \pi_{z,s,0}}{d_z/\sum_{n\in\mathbb{Z}} d_n}, \ \ z \in \mathbb{Z}. \tag{H-7}$$

Here, $\sum_{s\in[0,m]} \pi_{z,s,0}$ is the joint probability that a node is in zone z and transmits.

The conditional collision probability p_{col} of the tagged node in (4.2), given the tagged node is transmitting, is

$$p_{\text{col}} = 1 - \prod_{z=1}^{N} \left(1-\tau_z\right)^{X_z}. \tag{H-8}$$

Mean Duration of One Time Slot $E\left[T_{\text{dec}}\right]$

The mean duration of one time slot $E\left[T_{\text{dec}}\right]$, given that the tagged node is not transmitting, is comprised of the unit backoff time δ and mean frozen duration of the backoff time, as

$$E\left[T_{\text{dec}}\right] = \delta + p_{\text{suc}} E\left[T_{\text{suc}}\right] + (1-p_{\text{suc}}) E\left[T_{\text{col}}\right], \tag{H-9}$$

where p_{suc} is the probability that in-progress transmission is successful given that the channel is busy. $E\left[T_{\text{suc}}\right]$ and $E\left[T_{\text{col}}\right]$ are the mean time of the in-progress transmission with the transmission to be successful and collided, respectively.

$E\left[T_{\text{suc}}\right]$ in (H-9) can be represented as

$$E\left[T_{\text{suc}}\right] = \sum_{z\in\mathbb{Z}} p_{\text{suc},z} T_{\text{suc},z}. \tag{H-10}$$

[2] Each road segment includes the radio coverage of one AP and one zone 0 ahead of it.

where $p_{\text{suc},z}$ is the conditional probability that the in-progress transmission is by a node in zone z, given that the transmission is successful. Mathematically,

$$p_{\text{suc},z} = \frac{1}{p_{\text{suc}}} X_z \tau_z (1-\tau_z)^{X_z - 1} \prod_{n \in \mathbb{Z}, n \neq z} (1-\tau_n)^{X_n}. \tag{H-11}$$

$T_{\text{suc},z}$ in (H-10) is the successful transmission time when the in-progress transmitting node is in zone z. Mathematically,

$$T_{\text{suc},z} = L/r_z + SIFS + ACK/r_z + DIFS + \delta, \tag{H-12}$$

The collision time T_{col} of the in-progress transmission in (H-9) equals to the longest transmission time in the collision. Let $p_{\text{col},z}$ denote the probability that the longest transmission time is from nodes in zone z or its mirror zone $z_{\text{mir}} = N + 1 - z$ along the AP. Here, we jointly consider two zones z and z_{mir} as they have the same distance to AP and payload transmission rate.[3] Similar to [56], $p_{\text{col},z}$ could be computed as

$$p_{\text{col},z} = \begin{cases} \frac{1}{1-p_{\text{suc}}} \left(p_{\text{hcol},z} + p_{\text{dcol},z}\right), & \text{if } z \leq \lfloor (N-1)/2 \rfloor, \\ \frac{1}{1-p_{\text{suc}}} p_{\text{hcol},z}, & \text{if } z = \lceil N/2 \rceil. \end{cases} \tag{H-13}$$

$p_{\text{hcol},z}$ in (H-13) is called the homogeneous collision probability representing the probability that only nodes in zones z or z_{mir} transmit, where $z \leq \left\lceil \frac{N}{2} \right\rceil$. It is shown below as

$$\begin{aligned} p_{\text{hcol},z} = & \Big[\left(1 - (1-\tau_z)^{X_z} - X_z \tau_z (1-\tau_z)^{X_z - 1}\right) (1-\tau_{z_{\text{mir}}})^{X_{z_{\text{mir}}}} \\ & + \left(1 - (1-\tau_{z_{\text{mir}}})^{X_{z_{\text{mir}}}} - X_{z_{\text{mir}}} \tau_{z_{\text{mir}}} (1-\tau_{z_{\text{mir}}})^{X_{z_{\text{mir}}} - 1}\right) (1-\tau_z)^{X_z} \\ & + \left(1 - (1-\tau_z)^{X_z}\right) \left(1 - (1-\tau_{z_{\text{mir}}})^{X_{z_{\text{mir}}}}\right) \Big] \times \prod_{m=1, m \neq z, m \neq z_{\text{mir}}}^{N} (1-\tau_m)^{X_m}, \end{aligned} \tag{H-14}$$

which is comprised of three components: 1) the collided nodes are all from zone z; 2) the collided nodes are all from zone z_{mir}; and 3) the collision is from a mixture of nodes from both zones z and z_{mir}.

$p_{\text{dcol},z}$ in (H-13) is called diverse collision probability representing the probability that the collision is from at least one node in zones z or z_{mir}, where $z \leq \left\lfloor \frac{N}{2} \right\rfloor$, and one or more

[3] In case N is odd and $N + 1 - z = z$, z_{mir} is null with both its population $X_{z_{\text{mir}}}$ and transmission opportunity $\tau_{z_{\text{mir}}}$ to be 0.

nodes in other zones with larger transmission rate. The expression of $p_{\mathrm{dcol},z}$ is as

$$\begin{aligned} p_{\mathrm{dcol},z} &= \left[1-(1-\tau_z)^{X_z}(1-\tau_{z_{\mathrm{mir}}})^{X_{z_{\mathrm{mir}}}}\right]\left(1-\prod_{m=z+1}^{z_{\mathrm{mir}}-1}(1-\tau_m)^{X_m}\right) \\ &\quad \prod_{m=1}^{z-1}(1-\tau_m)^{X_m}\prod_{m=z_{\mathrm{mir}}+1}^{N}(1-\tau_m)^{X_m} \qquad \text{(H-15)} \end{aligned}$$

The mean collision time $E\left[T_{\mathrm{col}}\right]$ is then

$$E\left[T_{\mathrm{col}}\right]=\sum\nolimits_{z=1}^{\left\lceil\frac{N}{2}\right\rceil}T_{\mathrm{col},z}p_{\mathrm{col},z}, \qquad \text{(H-16)}$$

where $p_{\mathrm{col},z}$ is obtained in (H-13). $T_{\mathrm{col},z}$ is the packet collision time in zone z, mathematically,

$$T_{\mathrm{col},z}=L/r_z+DIFS+\delta, \qquad \text{(H-17)}$$

By substituting (H-10) and (H-16) in (H-9), we can obtain $E\left[T_{\mathrm{dec}}\right]$.

Mean Transmission Time $E\left[Tx_{\mathrm{suc},z}\right]$ and $E\left[Tx_{\mathrm{col},z}\right]$ of the Tagged Node

The successful transmission time $Tx_{\mathrm{suc},z}$ of the tagged node in zone z is deterministic as

$$E\left[Tx_{\mathrm{suc},z}\right]=T_{\mathrm{suc},z} \qquad \text{(H-18)}$$

where $T_{\mathrm{suc},z}$ is specified in (H-12).

The collision time $Tx_{\mathrm{col},z}$ of the tagged node is a random variable equal to the longest transmission time involved in the collision. Given that one collided node is the tagged node in zone z, the probability that the longest transmission is of nodes from zone z is hence

$$p_{\mathrm{ctag},z}=\frac{1}{p_{\mathrm{col}}}\prod_{n=1}^{z_{\mathrm{low}}-1}(1-\tau_n)^{X_n}\prod_{n=z_{\mathrm{up}}+1}^{N}(1-\tau_n)^{X_n}\left(1-\prod_{n=z_{\mathrm{low}}}^{z_{\mathrm{up}}}(1-\tau_n)^{X_n}\right),$$

when the collisions nodes are from zones closer to the AP than zones z and z_{mir}, where $z_{\mathrm{mir}}=N-z+1$, $z_{\mathrm{low}}=\min\{z,z_{\mathrm{mir}}\}$ and $z_{\mathrm{up}}=\max\{z,z_{\mathrm{mir}}\}$. Similar to (H-16), we jointly consider a zone z and its mirror zone z_{mir} along the AP.

The probability that the longest transmission time is from zone m or its mirror zone $m_{\text{mir}} = N + 1 - m$, where $m < z_{\text{low}}$ and $m_{\text{map}} > z_{\text{up}}$, is

$$p_{\text{ctag},m} = \frac{1}{p_{\text{col}}} \prod_{n=1}^{m-1} (1-\tau_n)^{X_n} \prod_{n=m_{\text{mir}}+1}^{N} (1-\tau_n)^{X_n} \left(1 - (1-\tau_m)^{X_m} (1-\tau_{m_{\text{mir}}})^{X_{m_{\text{mir}}}}\right),$$

i.e., nodes in zones farther than zones m and m_{mir} to the AP are not transmitting and at least one node in zones m or m_{mir} transmits.

The mean collision time $E\left[Tx_{\text{col},z}\right]$ of the tagged node in zone z is hence

$$E\left[Tx_{\text{col},z}\right] = \sum_{n=1}^{z_{\text{low}}-1} T_{\text{col},n} p_{\text{ctag},n} + T_{\text{col},z} p_{\text{ctag},z}, \tag{H-19}$$

with $T_{\text{col},z}$ given in (H-17).

4.4.4 Numerical Solution

By substituting (H-10), (H-16), (H-18) and (H-19) into (H-4), the steady state probability of the intermediate states $\pi_{z,s,b}$ could be represented by that of the boundary states $\pi_{z,s,0}$. Therefore, (H-4) could be managed as a self-contained non-linear system with unknowns $\pi_{z,s,0}$, where $z \in \mathbb{Z}, s \in [0, m]$, and solved numerically based on the fixed point equation

$$\pi_{z,s,0} = f\left(\pi_{z,s,0}\right). \tag{H-20}$$

As $\pi_{z,s,0} \in [0,1], \forall z \in \mathbb{Z}, s \in [0, m]$, the feasible region of the system $f(\cdot)$ in (H-20) is a compact convex set. According to the Brouwer's fix point theorem [63], (H-20) has at least one solution[4].

4.4.5 Derivation of the System Throughput

We evaluate the throughput performance in terms of nodal throughput s_z, representing the throughput achieved by individual node in a given zone z, and the system throughput S, representing the integrated throughput of all the nodes.

[4] In our simulations, the computation time of (H-20) is around 10 minutes with the scenario considered in the simulations.

The nodal throughput s_z is evaluated as the amount of packet payloads sent by individual node in each transmission in zone z, mathematically,

$$s_z = \frac{\tau_z(1 - p_{\text{col}})L}{(1 - \tau_z)\, E\left[T_{\text{dec}}\right] + \tau_z \times \text{mean trans. time in } z}, \tag{H-21}$$

where the mean trans. time in z is evaluated as $(1 - p_{\text{col}})\, E\left[Tx_{\text{suc},z}\right] + p_{\text{col}} E\left[Tx_{\text{col},z}\right]$.

This is because that within one time slot, the tagged node either backoffs or transmits. The former happens with probability $1 - \tau_z$. In this case, the channel could be either idle or used by others' transmission with the average duration $E\left[T_{\text{dec}}\right]$ specified in (H-9). The latter happens with probability τ_z. In this case, the transmission of the tagged node could be either successful or failed with mean duration of $(1 - p_{\text{col}})\, E\left[Tx_{\text{suc},z}\right] + p_{\text{col}} E\left[Tx_{\text{col},z}\right]$. Overall, the denominator in (H-21) computes the average length of one time slot. Within this duration, the tagged node transmits with probability τ_z and with probability $1 - p_{\text{col}}$ the transmission is successful. Upon each successful transmission, an average payload L is delivered.

With X_z nodes transmitting in zone z, the integrated system throughput S of the whole network is

$$S = \sum_{z \in \mathbb{Z}/\{0\}} X_z s_z. \tag{H-22}$$

4.4.6 Derivation of Network Size

The throughput characterized by (H-21) and (H-22) are dependent on the population of vehicles in each zone. In what follows, we show that the network size could be attained based on the node velocity only.

The mean arrival rate λ to AP and velocity v are in general linearly related as

$$\lambda = n_L k v. \tag{H-23}$$

where n_L here is the number of lanes in the road segment. k denotes the traffic density corresponding to the number of vehicles per unit distance in each lane along the road segment.

Moreover, based on Greenshield's model [64], the node density k linearly changes with the mean velocity v as

$$k = k_{jam}\left(1 - \frac{v}{v_f}\right), \tag{H-24}$$

Table 4.2: Parameter of Zones

Zone z	0	1	2	3	4	5	6	7
d_z (m)	20	25	30	40	60	40	30	25
r_z (Mbps)	0	1	2	5.5	11	5.5	2	1
$CW_{\min,z}$	Inf	128	64	32	16	32	64	128

Table 4.3: Default Setting of DCF and Road Traffic Parameters

Time slot δ	$50\mu s$
SIFS	$50\mu s$
DIFS	$128\mu s$
ACK	38 Bytes
Traffic jam density k_{jam}	120 veh/km/lane
Free-way speed v_f	160 km/h

where k_{jam} is the vehicle jam density at which traffic flow comes to a halt. v_f is the free-flow speed corresponding to the speed when the vehicle is driving alone on the road (usually taken as the road's speed limit).

Substituting (H-23) and (H-24) into (H-5), the mean node population in one road segment becomes

$$X = n_L k_{jam} \left(1 - \frac{v}{v_f}\right) \sum_{z \in \mathbb{Z}} d_z - 1, \tag{H-25}$$

with the tagged node excluded. Accordingly, the mean population in each zone X_z can be computed by substituting (H-25) into (H-6).

Given knowledge of n_L, k_{jam}, v_f and d_z, (H-25) indicates that the average network size is solely dependent on the velocity v. As a result, vehicles can estimate the achieved throughput via (H-21) and (H-22) by measuring its own velocity, and consequently they can conveniently adapt the DCF towards optimized performance which will be discussed in Section 4.6.

4.5 Model Validation

4.5.1 Simulation Setup

We validate our analytical models using simulations based on a discrete event simulator coded in C++. For evaluation purpose, we simulate a drive-thru Internet scenario as shown in Figure 4.1, in which an AP is deployed along the road and the vehicles passing through compete for communications using IEEE 802.11b. The whole road segment is divided into 8 zones as specified in Tables 4.2 and 4.3, with 7 zones in the radio coverage of AP and 1 zone representing the region outside the coverage of AP. The length and data rates of each zone are based on the extensive measurements reported in [65], also used in [52]. Unless otherwise mentioned, we simulate a road segment composed of 8 lanes. Along each lane vehicle nodes are uniformly deployed and moving at the constant velocity $v = 80$ km/h towards the same direction. By default, we set the traffic jam density k_{jam} and the free-way speed v_f as in Table 4.3 such that there are $X = 130$ vehicles on the road according to (H-25). Once reaching the end of the road segment, vehicles reenters the road as a new arrival starting from zone 0. Upon each renewal arrival, we clear the transmission history of vehicles with backoff stage set to 0. The vehicles are in the saturated mode with the packet size L of 1000 Bytes. Parameters of DCF are given in Table 4.3, which are used for both the simulations and the analysis. In each experiment, we carry out 30 simulation runs and plot the results with the 95% confidence interval.

We validate the developed analytical models in two scenarios: 1) equal CW with nodes transmit using the same CW_{min} in all zones, as in legacy IEEE 802.11b [54]; and 2) differentiated CWs where nodes transmit using different CW_{min} values in different zones. In each step, we change the node velocity v and network size X (by tuning k_{jam}) to show the impact of mobility and network size on the throughput performance.

4.5.2 Equal Contention Window (Legacy IEEE 802.11 DCF)

In this experiment, we set $CW_{min} = 32$ to all zones. In this case, nodes suffer from performance anomaly as described in Section 4.3 such that their throughput is throttled to the minimum value. This phenomenon is shown in Figure 4.4 which plots the nodal throughput s_n in different zones. As we can see, s_n is unrelated with the data rates in different zones. In the meantime, s_n reduces when the zone index increases. This is because that the mean backoff times of nodes increases with the increasing zone index as indicated in Figure 4.5(a). As a result, the transmission opportunity of nodes reduces with the

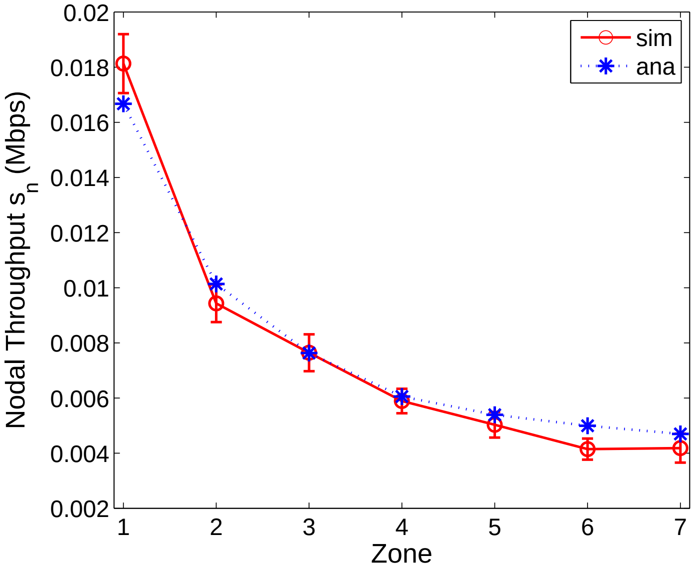
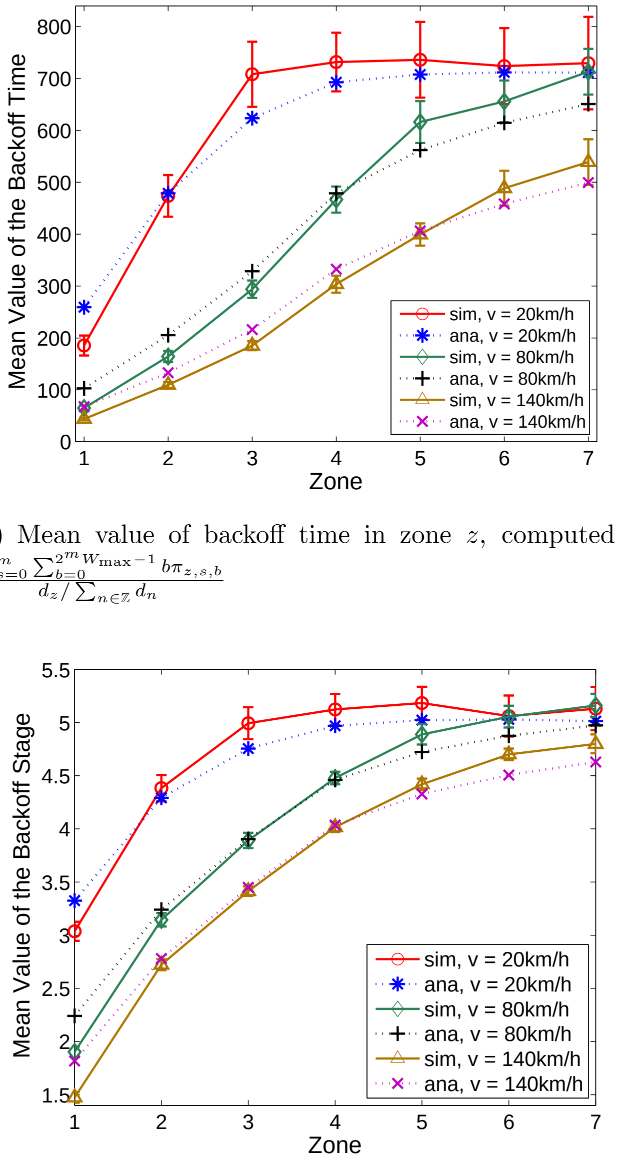


Figure 4.4: Nodal throughput s_n with equal contention window ($CW_{min} = 32$) in all zones and other parameters in Tables 4.2 and 4.3

increasing zone index. The reason for this phenomenon is that in our analytical models and simulations, the backoff stage is reset to 0 when nodes depart from the AP. As a result, the average backoff stage in zone 1 is smaller than that in other zones as in Figure 4.5(b), and the following zones are also affected due to the mobility.

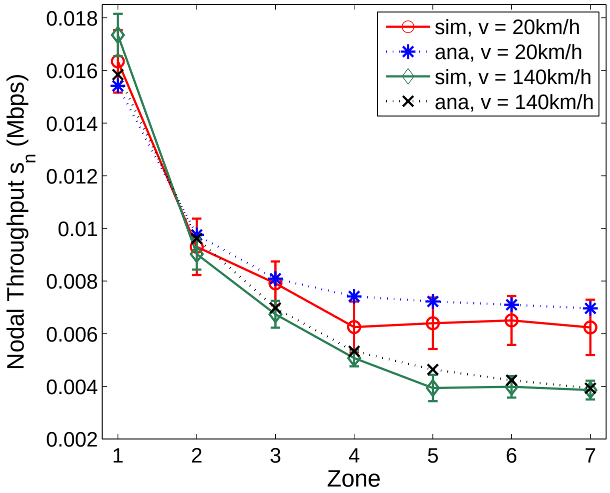
Figure 4.6 shows the throughput performance when node velocity increases from 20 km/h to 140 km/h and $X = 130$ nodes. As can be seen in Figure 4.6, both nodal throughput and system throughput reduce dramatically when the velocity increases. This is because that with increasing velocity both the mean backoff time and mean backoff stage in each zone reduce as indicated in Figure 4.5, resulting in increased collisions shown in Figure 4.7. The unintended backoff time is due to the high mobility of nodes. With enhanced mobility, nodes switch zones more often and therefore adapt their contention windows more frequently. As such, the small backoff stage in zone 1 affects the ensuing zones more easily. Despite having the largest transmission rate, nodes in zone 4 encounter the most frequent collisions which results in the large mean backoff time and stages and correspondingly throttles their throughput. The large backoff stage in zone 4 also propagates to the following zones when velocity increases, making the backoff times in $\{5, 6, 7\}$ larger than that in zones $\{1, 2, 3\}$ as shown in Figure 4.5.



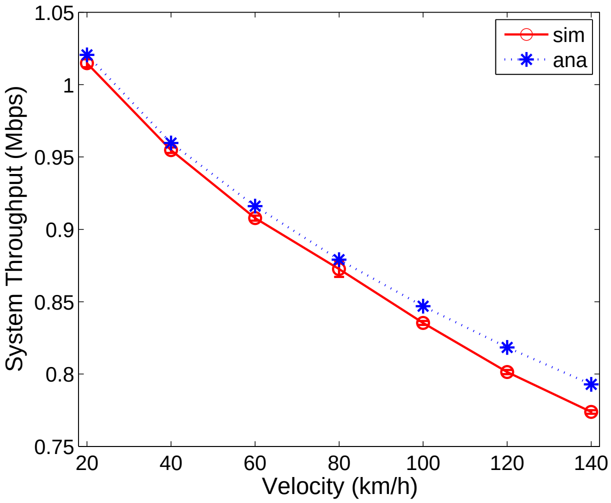
(a) Mean value of backoff time in zone z, computed as $\frac{\sum_{s=0}^{m}\sum_{b=0}^{2^m W_{\max}-1} b\pi_{z,s,b}}{d_z/\sum_{n\in\mathbb{Z}} d_n}$

(b) Mean value of backoff stage in zone z, computed as $\frac{\sum_{s=0}^{m}\sum_{b=0}^{2^m W_{\max}-1} s\pi_{z,s,b}}{d_z/\sum_{n\in\mathbb{Z}} d_n}$

Figure 4.5: Statistics of backoff time and stage with increasing node velocity, equal contention window ($CW_{min} = 32$) in all zones and constant network size $X = 130$ vehicles



(a) Nodal throughput s_n with increasing velocity



(b) System throughput S with increasing velocity

Figure 4.6: Throughput performance with increasing vehicle velocity, equal contention window ($CW_{min} = 32$) in all zones and constant network size $X = 130$ vehicles

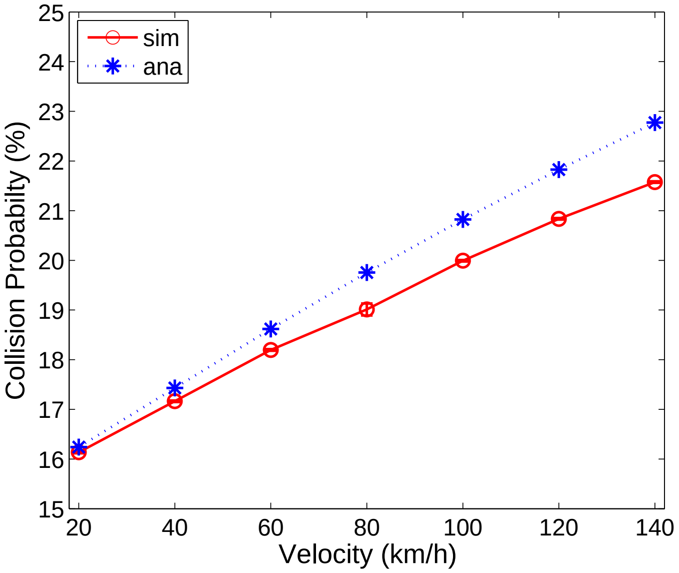


Figure 4.7: Packet loss probability with increasing node velocity, equal contention window ($CW_{min} = 32$) in all zones and constant network size ($X = 130$)

Figure 4.8 shows the impacts of network size on the throughput performance with constant node velocity $v = 80$ km/h. In this experiment, we increase k_{jam} from 40 veh/km/lane to 200 veh/km/lane, resulting in the increased network size from 43 to 216 vehicles. As a result, we can see that the system throughput reduces with increased network size. This is because that more intense collisions are encountered with the increasing number of competing nodes.

In summary, deploying equal CW_{min} in different zone would suffer from the performance anomaly. Moreover, the throughput performance is keenly dependent on the network size and node velocity. Increasing the velocity will result in the unintended backoff time distribution and enhanced packet collisions. Therefore, adapting the DCF according to the node velocity is necessary for guaranteed throughput.

4.5.3 Differentiated Contention Window Sizes among Zones

To address the performance anomaly and boost the system throughput, in this experiment, we let nodes in different zones have different $CW_{\min}$ values and investigate the impacts of network size and velocity on the throughput performance. The $CW_{\min}$ used is shown in

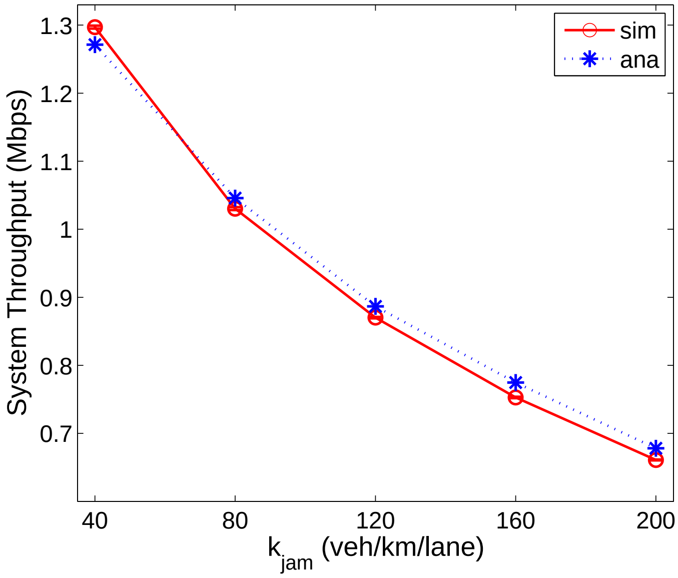


Figure 4.8: System throughput when increasing network size (by tuning k_{jam}), equal contention window ($CW_{min} = 32$) in all zones and constant node velocity $v = 80$ km/h

Table 4.2 which is devised based on [54]. The optimal selection of $CW_{\min}$ in different network sizes and node velocities will be discussed in the next section.

Figure 4.9 plots the nodal throughput with differentiated contention windows in zones. In this case, as nodes close to AP have relatively smaller CW_{min} and accordingly higher transmission probability, the nodal throughput is a bell-shape curve. Meanwhile, with nodes in front zones having relatively small backoff time as shown in Figure 4.10(a), the curve tilts to the right.

Figure 4.11 shows the throughput performance when node velocity increases. Similar to the equal contention window case, we can see that increasing the velocity also results in the monotonic decreasing of throughput. Moreover, the curve of nodal throughput in Figure 4.11(a) tilts even more severely with the reduced throughput in the back zones. To take a close examine, Figure 4.10(a) shows that mean backoff time changes dramatically when increasing the velocity, while the mean backoff stage changes slightly as shown in Figure 4.10(b). As we can see, with velocity increasing, both the mean value of backoff time and mean value of backoff stage reduce and they increase as zone index increases. This, on one hand, is because that zone 1 has the smallest mean backoff stage due to the renewal arrival to the AP. With increased velocity, the following zones are also affected from that

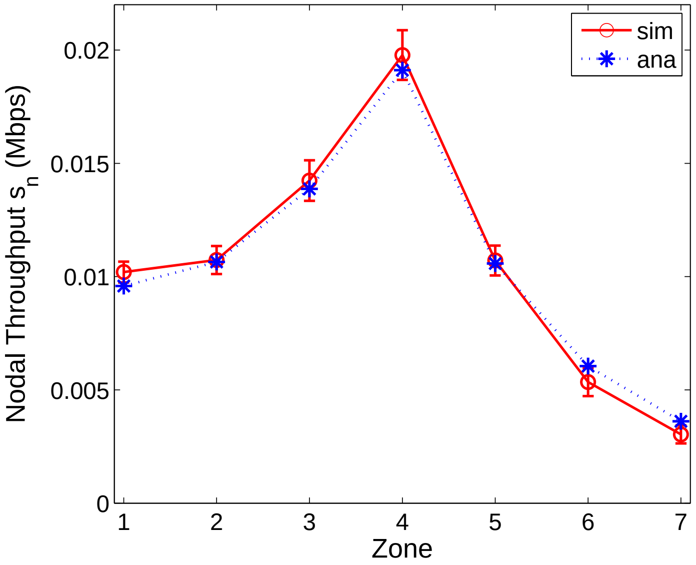
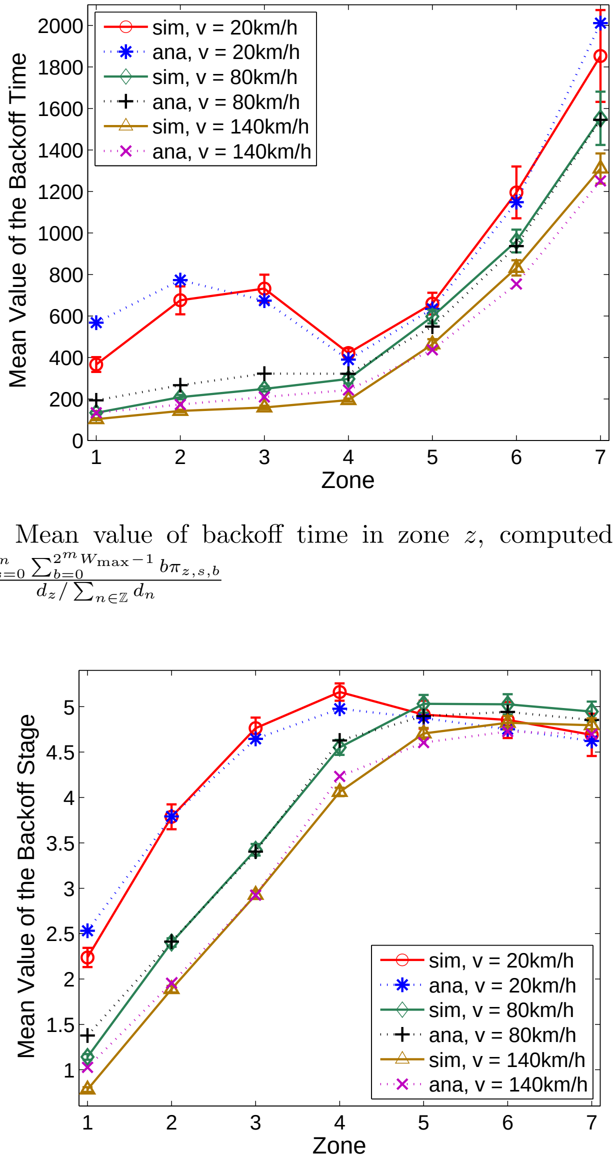


Figure 4.9: Nodal throughput s_n with differentiated CW_{min} in zones and other parameters in Tables 4.2 and 4.3

with the smaller backoff times than expectation. On the other hand, in zone 4, nodes have large backoff stages due to intensive transmissions and collisions. This affects the following zones as shown in Figure 4.10(b), resulting in large backoff times in those zones. As a direct result of the reduced mean backoff time, the collision probability increases, as shown in Figure 4.12, which finally leads to the reduced system throughput as indicated in Figure 4.11(b). In a nutshell, the high mobility result in fast transitions between zones which intensively affects the resulting backoff time and the throughput.

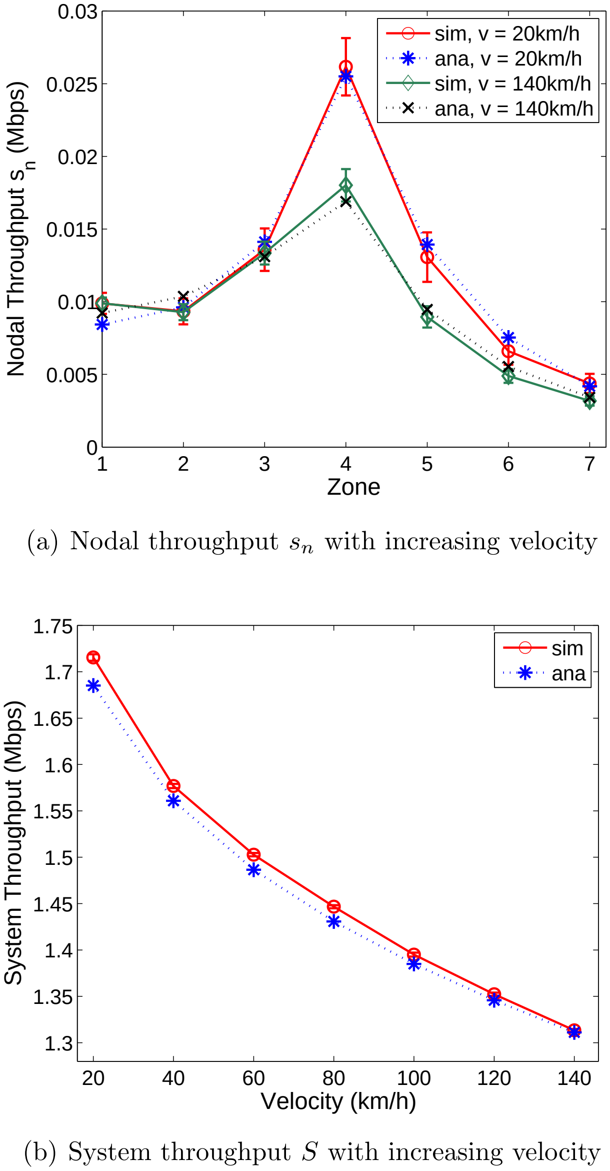
In the next experiment, we modify the zone length as specified in Table 4.2 and examine whether the above conclusions are still valid when the length of each zone is changed. Figure 4.15 plots the throughput performance when the length of each zone is enlarged to four times of the default value in Table 4.2 while other parameters remain unchanged (as specified in Table 4.3). As shown in Figure 4.15(a), when the zone length is enlarged, the nodal throughput is also a bell-shaped curve tilted to right which is similar to that in Figure 4.9. With increasing node velocity and fixed node density in Table 4.3, as indicated in Figure 4.15(b), the system throughput would also reduce which is similar to that in Figure 4.11(b). Moreover, as exhibited in Figure 4.15, both the nodal throughput and system throughput reduce when the zone length is enlarged. This is because that increasing



(a) Mean value of backoff time in zone z, computed as $\frac{\sum_{s=0}^{m}\sum_{b=0}^{2^m W_{\max}-1} b\pi_{z,s,b}}{d_z/\sum_{n\in\mathbb{Z}} d_n}$

(b) Mean value of backoff stage in zone z, computed as $\frac{\sum_{s=0}^{m}\sum_{b=0}^{2^m W_{\max}-1} s\pi_{z,s,b}}{d_z/\sum_{n\in\mathbb{Z}} d_n}$

Figure 4.10: Statistics of backoff time and stage with increasing node velocity, differentiated CW_{min} in zones and constant network size $X = 130$ vehicles



(a) Nodal throughput s_n with increasing velocity

(b) System throughput S with increasing velocity

Figure 4.11: Throughput performance with increasing node velocity, differentiated CW_{min} in zones and constant network size $X = 130$ vehicles

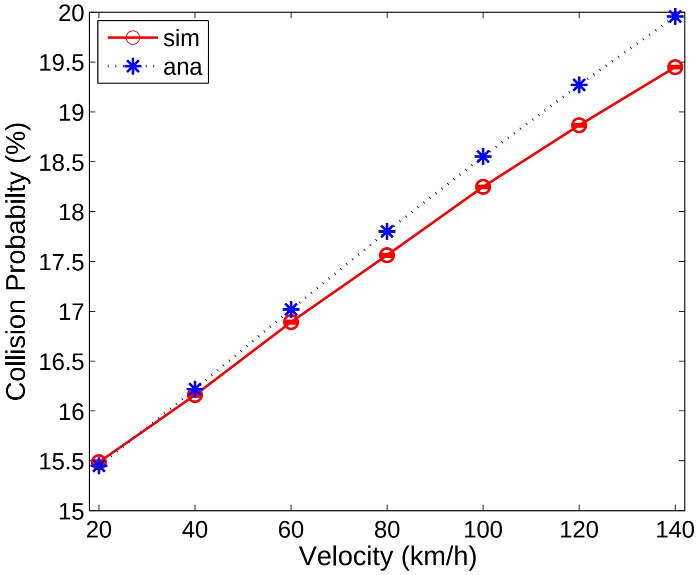


Figure 4.12: Collision probability with increasing velocity and adapted network size according to (H-25) with k_{jam}, v_f in Table 4.3

the zone length implies enlarging the coverage of AP. As the vehicle density remains the same, more vehicles are therefore contending for transmissions, which lead to much severer collisions of transmissions and the degraded throughput performance. Therefore, how to optimally adjust the CW with different road traffic parameters is crucial. Based on our model, we strive to address this issue in the next section of the chapter.

Recall that the network size could be estimated based on the node velocity via (H-25). In the last experiment of this section, we increase the node velocity with fixed k_{jam} and v_f as in Table 4.3. In this case, the network size adapts with the velocity, which simulates a road segment in different time periods. For example, with low velocity, more nodes are accumulated on the road according to (H-25), which simulates the busy hour traffic. With high velocity, vehicle traffic on the road is smooth with low density, similar to the late night scenario. As shown in Figure 4.13, since both velocity and network size affect the throughput, the resulting throughput is not monotonic when velocity increases. The network achieves the lowest throughput when node velocity is around 80 km/h which happens to be the prevalent speed in the urban freeway.

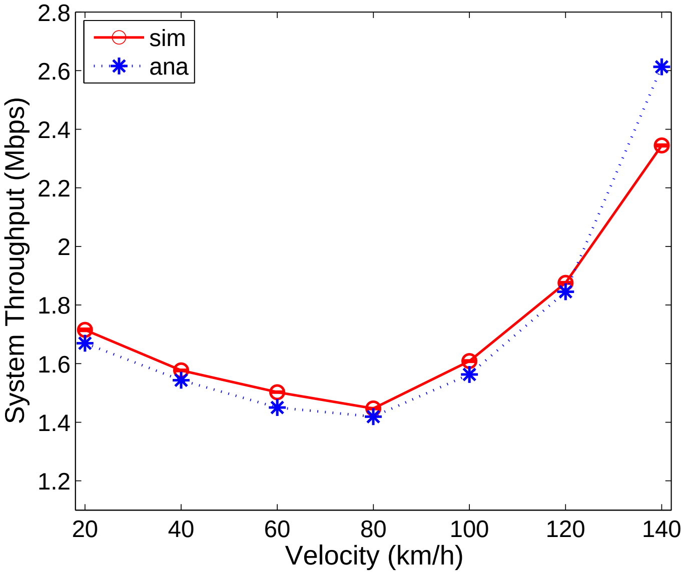


Figure 4.13: System throughput with increasing velocity and adapted network size with k_{jam}, v_f in Table 4.3

4.6 Protocol Enhancement

Based on the observations in the previous section, we propose the following assertions as the guideline of the selection of CW_{min} in different zones:

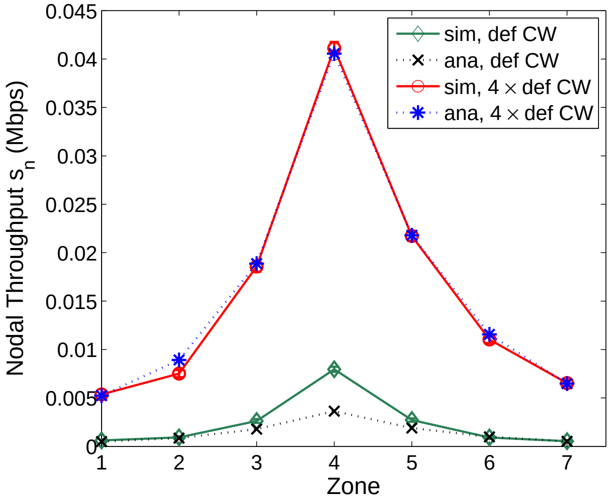
▷ CW_{min} should adapt to the payload transmission rates of vehicles according to their distance to AP.

▷ The maximum backoff stage m should be kept small to mitigate the impacts of fast zone transitions on the throughput.

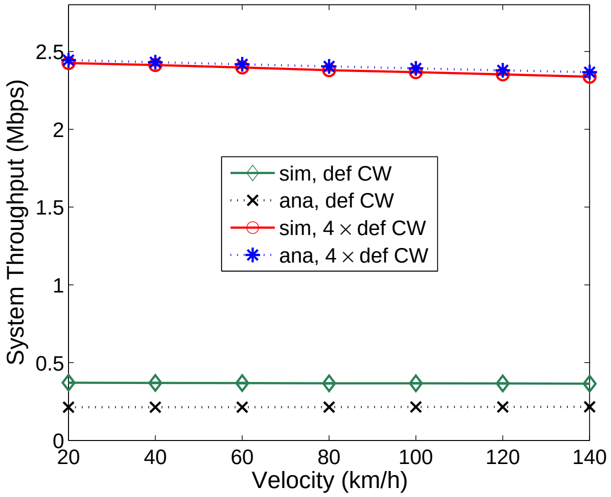
▷ CW_{min} should adapt to node velocity (and network size).

The reasoning behind the first assertion is obvious: to eliminate the performance anomaly, nodes with different transmission rates should be rendered with different channel access probabilities to fully utilize the transient high-rate connectivity.

The second assertion is rooted in the high mobility of nodes. As indicated in Figure 4.11, increasing velocity will reduce the throughput. This is because that in DCF, the value of

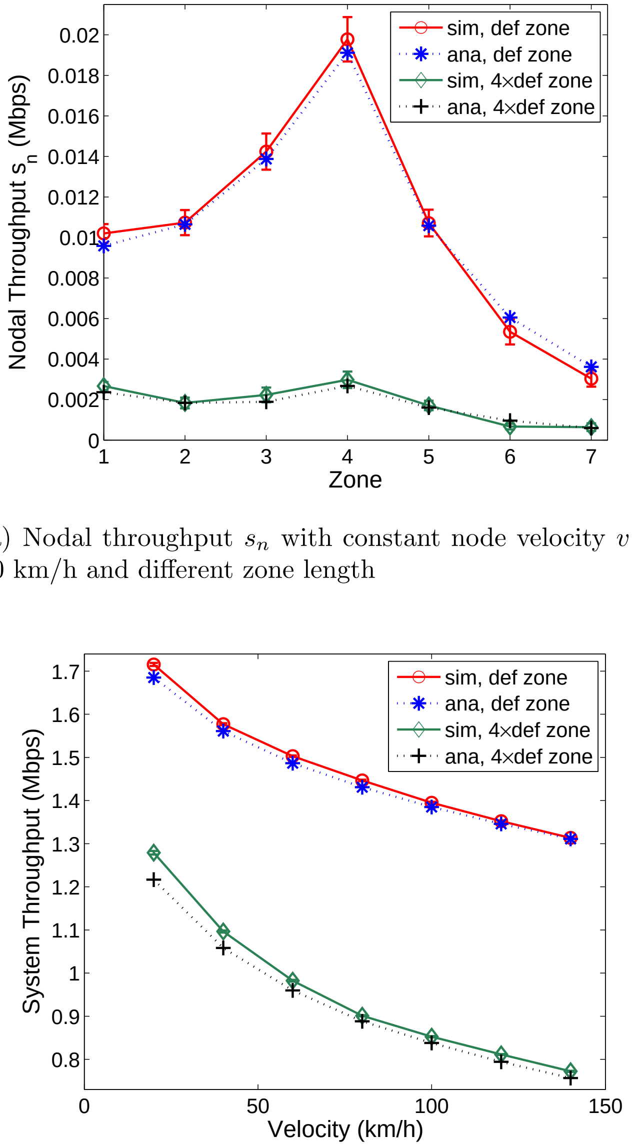


(a) Nodal throughput s_n with $m = 1$ and increasing velocity



(b) System throughput S with $m = 1$ and increasing velocity

Figure 4.14: Throughput performance when $m = 1$ with the constant network size $X = 130$ and increasing velocity



(a) Nodal throughput s_n with constant node velocity $v =$ 80 km/h and different zone length

(b) System throughput S with increasing velocities and d-ifferent zone length

Figure 4.15: Throughput performance with different zone length and other parameters in Tables 4.2 and 4.3

backoff stage records the transmission history of nodes. With the fast mobility and frequent zone transitions of nodes, the backoff stages in different zones influence each other, resulting in the unintended distribution of backoff times as in Figure 4.10. To minimize the mutual interference of backoff times among zones, we should keep m small.

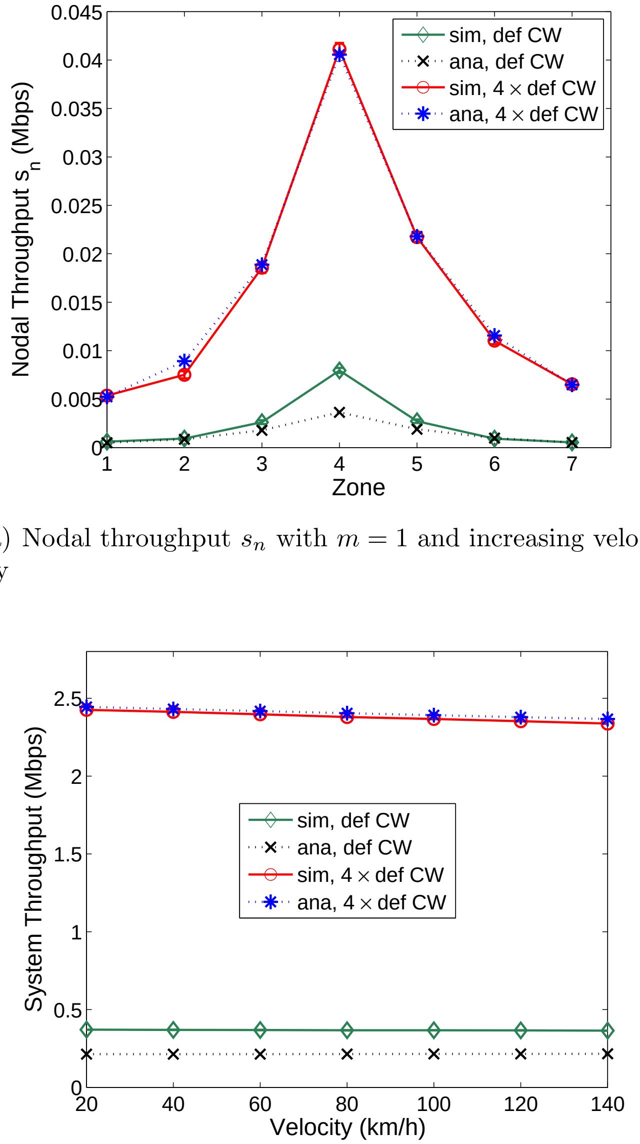
Figure 4.16 plots the throughput with $m = 1$ and increasing velocity. The network size is kept constant with $X = 130$ vehicles. "def CW" in Figure 4.16 refer to using the default CW_{min} values shown in Table 4.2. As we can see, increasing the velocity in this case does not affect the throughput much. Instead, the system throughput reduces significantly compared with the value in Figure 4.11. This is because that with a smaller m nodes have a smaller backoff time and transmit more frequently with more collisions. Our model is not very accurate when the contention window is small. This is because that in our simulation and the standard, it is possible that a node continually selects the backoff time to be 0 after each transmission and then transmits consecutively. In our analysis, however, we do not take this case into account when computing the slot time $E\left[T_{\text{dec}}\right]$ in (H-9). In real-world, the backoff time needs to be large enough to avoid collisions and consecutive transmissions are rare. Our model is accurate in this case, e.g., when CW is 4 times of the default CW as shown in Figure 4.16. In summary, as indicated in Figure 4.16, reducing m would make DCF unscalable and decrease the throughput. To compensate, we could estimate the network size based on the node velocity according to (H-25), and then adapt CW_{min} accordingly, which explains the third assertion.

Based on the above assertions and provided the node velocity, the optimal CW_{min} could be obtained by solving the optimization problem as

$$\begin{aligned} \underset{W_z}{maxmize} \quad & S \\ s.t., \quad & s_z \geq \eta_z, \quad z \in \mathbb{Z}/\{0\}. \end{aligned} \tag{H-26}$$

In (H-26), the objective is to maximize the system throughput. The constraint dictates that the nodal throughput in different zones must be above certain level. This, on one hand, is to guarantee the throughput fairness of nodes with different distance to AP. On the other hand, the upper layer applications and protocols may also need guaranteed throughput when nodes are in zones far away from the AP. For example, multimedia applications, e.g., VoIP and live streaming, typically pose a bound on the minimal transmission rate to maintain effective connections [66]. Upper-layer protocols, e.g., TCP, may also require a minimum rate of connection to ensure their functionalities, e.g., congestion control [67].

Eq. (H-26) is an integer programming problem. To reduce the computation complexity, we seek a sub-optimal solution as follows. Assuming that the backoff time in different zones



(a) Nodal throughput s_n with $m = 1$ and increasing velocity

(b) System throughput S with $m = 1$ and increasing velocity

Figure 4.16: Throughput performance when $m = 1$ with the constant network size $X = 130$ and increasing velocity

are independent with small m (the second assertion). According to (H-21), we have

$$\frac{s_x}{s_y} \approx \frac{\tau_x}{\tau_y} \approx \frac{W_y}{W_x}. \tag{H-27}$$

Incorporating (H-27), the constraint of (H-29) is satisfied when

$$s_x = \frac{\eta_x}{\eta_y} \times s_y \quad \text{and} \quad s_y \geq \eta_y, \quad x, y \in \mathbb{Z}/\{0\}, \tag{H-28}$$

As a result, instead of computing W_z in all zones as in (H-26), we could assume fixed ratios between nodal throughput according to (H-28), and tune CW_{min} in one zone, e.g., W_1, as

$$\begin{array}{cc} \underset{W_1>0}{maxmize} & S \\ s.t., & W_x = \frac{\eta_x}{\eta_1} \times W_1, \quad x \in \mathbb{Z}/\{0\}, \\ & s_1 \geq \eta_1, \end{array} \tag{H-29}$$

and adjust W_z in other zone z at the basis of W_1.

Let the ratios of CW_{min} in different zones be same as those in Table 4.2. Let the setting of simulators be same as those in Table 4.3, and let m be 1. Figure 4.17 plots the optimal value of CW_{min} based on (H-29) with the increasing velocities (decreasing network size according to (H-25)). Here, the resultant CW_{min} is represented by $W_z^*/W_z^0, \in \mathbb{Z}/\{0\}$, in which W_z^* denotes the optimal CW_{min} in zone z and W_z^0 is the CW_{min} as specified in Table 4.2. The system throughput with the optimal CW_{min} is plotted in Figure 4.18. Compared with Figure 4.13, using the optimal CWs the system throughput can improve for around 15% $\sim$ 45% in different velocities.

To implement (H-29), the optimal CWs could be computed off-line at different node velocities and then loaded into APs as a table. Based on the estimation of node velocity and network size in (H-25), APs could search the CW table and apply the optimal CWs correspondingly without much computations. As the network size is random in practice and varying over time [8], the CWs applied also need to be adapted timely. Note that the CW should be adapted slowly, e.g., at the intervals of hours, to capture the changing traffic density in the long term. Moreover, the CW table could be coarse, e.g., mapping the node velocity to appropriate CW at the step of 5 km/h. This avoids the frequent and unnecessary CW adaption and thus make the network unstable.

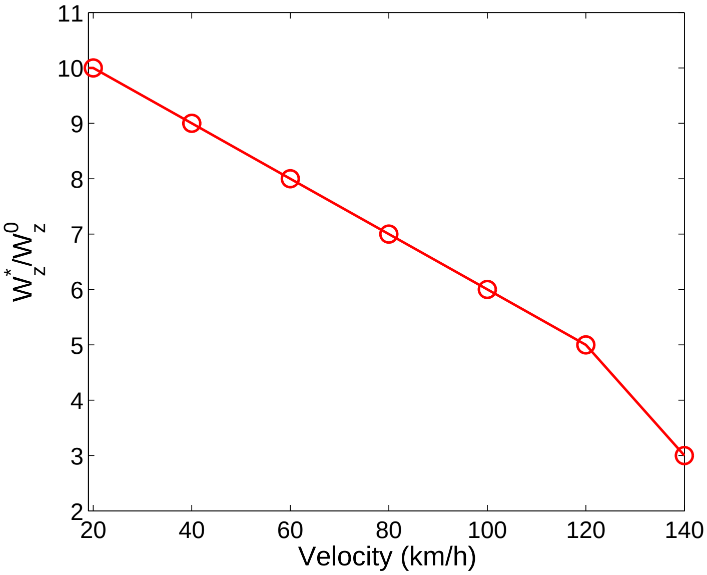


Figure 4.17: Optimal CW_{min} with increasing velocity

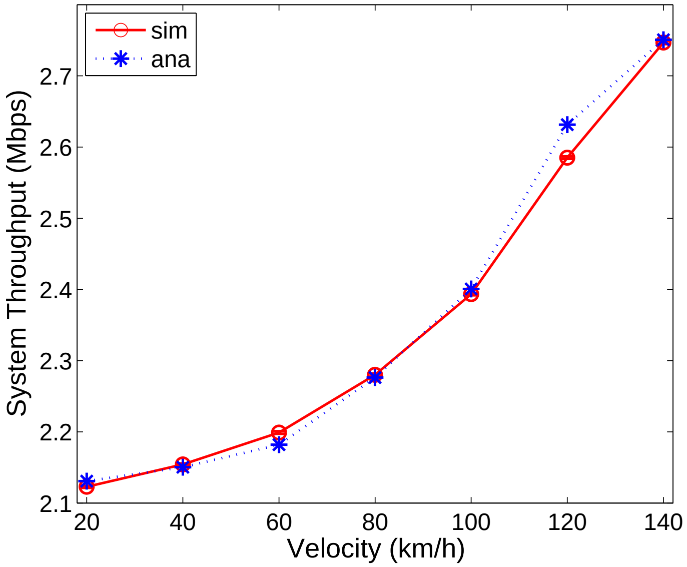


Figure 4.18: System throughput with optimal CW_{min} and increasing velocity

4.7 Summary

We conclude this chapter by reinforcing our observation that the high mobility of nodes significantly influences the performance of DCF, which results in unintended transmission probabilities rendered to nodes and finally degraded throughput performance. In this work, we have developed an accurate and scalable model to investigate the throughput performance under different velocities and network scales. We have shown that due to the mobility, the network size of the drive-thru Internet is solely dependent on the node velocity, which enables us to optimally configure the DCF by knowing the node velocity only. To enhance the MAC throughput in the drive-thru Internet scenario, we have proposed three assertions as the guideline of the DCF design, which are effective to mitigate the impacts of mobility. As an immediate next step, we plan to further extend our model to evaluate the QoS performance for multimedia applications and the QoS provision schemes in the high-speed drive-thru Internet scenario.

Chapter 5

Video in Motion: Impact of Dynamics on User Perceived Video Quality and Adaptive Receiver Adaption for Smooth Video Playback

5.1 Introduction

To provide the real-time multimedia services such as live and on demand video streaming, video/audio conferencing and video monitoring of road traffic etc., are, if not the most, exciting and value-added services to vehicles. This, however, represents several fundamental engineering challenges due to the stringent quality of service (QoS) requirements of multimedia applications [68] and the highly variable channels of vehicular communications [61]. In specific, the high node mobility results in the dynamically changing wireless channels with severe shadowing and multi-path fading, leading to the dramatically changing end-to-end throughput and delays. In the multimedia applications, video packets have strict deadlines of presentation. The varying transmission delays and intermittent connections of VANETs may result in the severe delays of packet download and missing of playback deadline, which consequently lead to the jerkiness or even frozen of video playback. To enable efficient multimedia streaming services to fast-motion vehicles, it is therefore crucial to understand how the network dynamics affect the user's perceived video quality and effectively accommodate the dynamic network delays in the video playback.

There are a variety of techniques appeared in the literature to address the problem

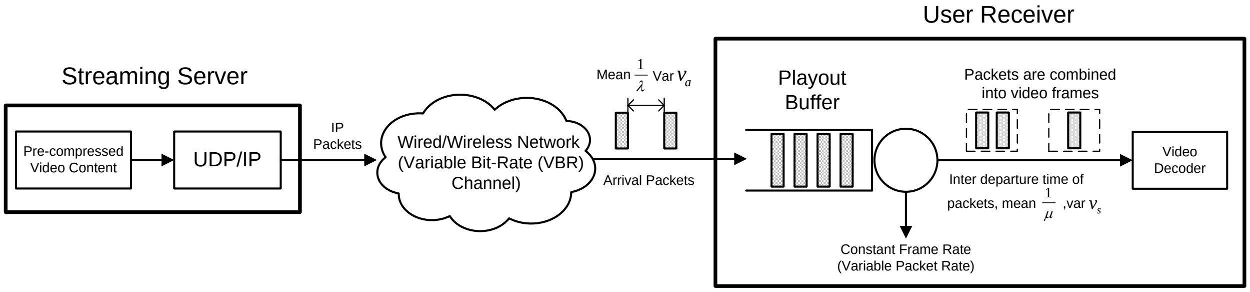

User Receiver
Streaming Server
Pre-compressed Video Content
UDP/IP
IP Packets
Wired/Wireless Network (Variable Bit-Rate (VBR) Channel)
Mean $\frac{1}{\lambda}$ Var v_a
Arrival Packets
Playout Buffer
Packets are combined into video frames
Inter departure time of packets, mean $\frac{1}{\mu}$,var v_s
Video Decoder
Constant Frame Rate (Variable Packet Rate)

Figure 5.1: Process of video streaming

of packet video delivery over time-varying dynamic channels [68, 69], including the rate-distortion optimized packet scheduling [70] and routing [71, 72, 73], power control and adaptive coding at the transmitter [74, 75, 76], playback rate and strategy adaption at the receiver [77, 78], etc. However, most of these approaches focus on the network-side issues which consider the network optimization in terms of throughput, delay and delay jitter; nevertheless they fail to fully investigate on the resultant video quality from the end user's perspective, which motivates our work.

In this study, we analyze and optimize the user perceived video quality in terms of the start-up delay, fluency of video playback and packet loss rate [79]. By intelligently using the user-side resources which are dependent on network dynamics, we design adaptive quality-driven video streaming schemes towards the maximal user utility. To this end, we first develop an analytical framework to reveal the impacts of network dynamics on the video quality perceived by the end user. In specific, to overcome the network dynamics, the playout buffer is usually deployed as shown in Figure 5.1, which stores the received video packets and delays the initial media playout for a short period until a certain playback threshold is reached. This short period constitutes the start-up delay. During the media playout, the packets are discharged from the playout buffer and injected to the media player for playback. As long as the playout buffer is non-empty, the continuous media playout is always guaranteed. Since the occupancy of the playout buffer is closely related to the user's viewing experience in terms of start-up delay and smoothness of playback, we study the video quality by analyzing the evolution of the queue length under the network dynamics. Secondly, we propose adaptive schemes to optimally determine the playback threshold driven by the users' required video quality. We strike a trade-off between the start-up delay, smoothness of playback and video packet loss, by adjusting the playback threshold. Specifically, to ensure smooth media playout, enough packets should be cached initially in the playout buffer to absorb the variations of packet arrivals. This, however, may incur an intolerable long waiting time to end users. Meanwhile, to prevent packet loss due to overflow of the finite buffer, the queue length should be maintained at a relatively low level with less packets buffered initially. However, low queue occupancy may cause frequent playback interruptions. Thus, the playback threshold should be adaptively and optimally determined under the specific requirements of video quality metrics. Based on the analytical framework, we formulate the playback threshold selection as a stochastic optimization problem driven by the specific video quality requirements, and provide the optimal solutions to guarantee the stochastic video performance to end users.

Our main contributions are in three-fold.

- ▷ *General Modeling*: We consider a general network setting characterized by the first two moments of statistics, i.e., the mean and variance of traffic arrival rate and the video playback rate. We study both infinite buffer and finite buffer cases by modeling the playout buffer as $G/G/1/\infty$ and $G/G/1/N$ queues, respectively. In this way, the proposed analytical model is general and suitable for a diverse range of video codecs, video streaming applications, and network scenarios.
- ▷ *Compact Solution*: We apply the diffusion approximation to derive the closed-form expressions of the video quality metrics in terms of the start-up delay, the number of playback frozens, and the packet loss rate, and represent the video quality metrics by the network statistics, i.e., the average throughput and delay jitters. With the obtained results, we can evaluate the impacts of network statistics on the user's video quality. In a reverse manner, given the user's specific requirements on video quality, we can also conveniently determine the demanded network throughput to support the required video quality. In this way, the achieved compact solutions pave the way for quality-driven network resource allocation.
- ▷ *Distributed Optimal Control*: With the network statistics as an input to our analytical model, we design adaptive playout buffer management schemes to optimally select the playback threshold to cater to users with different quality requirements. The proposed schemes are employed in a distributed manner via local estimations of users only without any assistance from the networks, which is hence particularly suitable for large scale network deployments.

The remainder of this chapter is organized as follows: Section 5.2 reviews the related works. The analytical framework is presented in details in Section 5.3 by considering both the infinite buffer and finite buffer cases. Section 5.4 describes adaptive playout buffer management schemes, and Section 5.5 validates the achieved analytical results and optimal control schemes by extensive simulations. Section 5.7 closes the chapter with the concluding remarks.

5.2 Related Work

Streaming media over unreliable and time-varying dynamic channels has attracted an extensive research attention in the last decade. Various network-adaptive schemes have been proposed [68]. Our work belongs to the scope of end-system centric solutions [69] which adapt the video from the receiver's perspective.

The end-system centric solutions refer to the adaptive video streaming mechanisms which adaptively modify the visual quality via the playback rate control or playout buffer management at the end-systems based on the occupancy of playout buffer or user's available bandwidth [77]. Liu et al. [78] propose an end-to-end playback rate adaptation scheme based on the layer coding technique. Each receiver actively measures its local available bandwidth and pass that to the server. Based on the echo information, the server then determines the appropriate number of layered streams conveyed to users and hence adapts the video compression rate according to the available bandwidth. By doing so, the visual quality degrades with enhanced compression ratio when the end-to-end bandwidth is insufficient; nevertheless, users can enjoy smooth playback. Galluccio et al. [80] describe an adaptive MPEG video streaming framework in which the wireless channel is modeled as a Rayleigh fading channel represented by a FSMC (finite state Markov chain). By analyzing the channel status via the Markovian model, the available bandwidth can be computed and the appropriate video playback rate is determined accordingly. Similar approach is adopted in [74] where channel coding is adapted in different channel conditions which are evaluated using FSMC. However, the source adaption schemes suffer from the scalability issue, as the server needs to respond to each individual user to resolve different quality requirements. When the network scales to a large size, the server can be easily overloaded. Moreover, most of the previous works on wireless multimedia transmission only consider a single-hop wireless channel which can be well modeled by FSMC. However, if multi-hop wireless transmissions and heterogeneous networks are considered, the analysis become invalid as accurate channel model in this case is generally not available.

To distribute the computation burden to the end users and hence enhance the network scalability, some approaches have been proposed to adapt the video playback at the end users. Kalman et al. [81] introduce the adaptive media playback (AMP) scheme at the end user, which can adaptively tune the video playout rate according to the playout buffer occupancy to ensure the smooth video playback. In specific, when the occupancy of playout buffer is above some threshold, the video playback rate will be increased to avoid the overflow of playout buffer. This leads to the effects of fast forward to the users. Laoutaris et al. [82] adopt the same mechanism but use the Markov decision process (MDP) to optimally determine the video playback rate at different channel conditions.

Another prevailing way of adaptive video streaming is by playout buffer management. In this scenario, the key issue is how to optimally determine the playback threshold to maximize the duration of continuous playback while minimizing the start-up delay. Liang et al. [83] establish a Markovian model to study the tradeoff between playback continuity and start-up delay. The wireless channel is modeled as a FSMC and the interplay between the channel statistics and playout buffer is provided under different buffer strategies. However, only the single-hop scenario is considered which can not be applied to multi-hop transmissions. Dua et al. [84] propose to adapt the playback threshold through a MDP. The channel is also modeled as a FSMC, where each successful transmission incurs certain profit. The playout buffer is managed to determine an optimal playback threshold to maximize the overall profit. In [85], the upper bound and lower bound of the jitter-free probability have been derived. However, this work only provides the probability of smooth playback without interruptions and does not capture the interruption frequency, which is of more interest. Instead of providing performance bounds, we provide the closed form expressions of the video quality.

Different from the previous efforts, we consider a general network scenario by modeling the playout buffer at the user end as a $G/G/1/\infty$ and $G/G/1/N$ queue, with both finite and infinite buffer cases. The analytical model could be applied to not only the single-hop wireless networks but also the multi-hop wired/wireless networks. Furthermore, since the network can be highly dynamic with intensive variance, we explicitly take delay jitter into consideration and show its impacts on the perceived video quality.

5.3 Analytical Framework

In this section, we first present the architecture of video streaming system. After that, we describe a general model for the playout buffer and develop an analytical framework to study the video quality perceived by the user, considering both infinite and finite buffer cases.

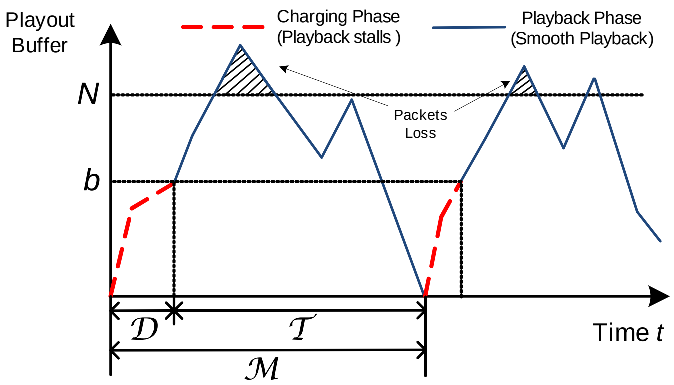


Figure 5.2: Evolution of the playout buffer during media playout with buffer size N

5.3.1 Model of Playout Buffer

Figure 5.1 shows a typical architecture of media streaming [86]. In Figure 5.1, the raw video contents are pre-compressed and saved in the storage devices. Upon the user's request, the media server retrieves the pre-stored content and then segmented it into packets. The video packets are transmitted over a lossy variable bit rate (VBR) heterogeneous network using the User Datagram Protocol (UDP)/IP protocol suite. With network dynamics, packets arrive at the user with variable delays. Without loss of generality, we assume that the inter-arrival time of video packets follows a given but arbitrary distribution with mean $\frac{1}{\lambda}$ and variance v_a. At the user end, the downloaded packets are first stored at the playout buffer, then combined into video frames and injected into the video player at the same cadence of frame rate that the video encoder generates. As the video frames are played at the constant rate, the service rate in terms of video packets is hence variable. We consider that the inter-departure time of video packets, determined by the instantaneous video playback rate, also follows a general distribution with the constant mean $\frac{1}{\mu}$ and variance v_s.

The playout buffer thus can be modeled as a $G/G/1/\infty$ queue when the buffer size is infinite or a $G/G/1/N$ queue when the buffer is finite. For the remaining part of this section, we analyze the evolution of the playout buffer in the infinite buffer and finite buffer cases, respectively, given network and playback statistics, i.e., λ, μ, v_a and v_s.

5.3.2 Infinite Buffer Case

We first consider the infinite playout buffer case, i.e., the buffer is infinitely large or large enough to accommodate the whole video file. This is typically true when the end host systems are personal computers with a large hard disk.

In general, the video playback process can be divided into two iterative phases, namely the charging phase and playback phase, as shown in Figure 5.2. The charging phase starts once the playout buffer becomes empty. In this case, the buffer is charged with continuously downloaded packets and the media playback is kept frozen until b packets are filled. Henceforth, we refer to b as the threshold of playback. Let R.V. (random variable) $\mathcal{D}$ denote the duration of the charging phase. The playback phase starts once the playback threshold b is reached and packets are discharged from the buffer for playback. Due to dynamic packet arrivals and departures of the buffer, the playback phase may stall when the playout buffer becomes empty again. Let R.V. $\mathcal{T}$ denote the duration of the video playback phase. The charging and playback phases iterate until the whole video is downloaded.

In this work, we evaluate the user's video quality in following two aspects: the *start-up delay* and *smoothness of video playback*. The former refers to the time period that users have to wait before video playback starts, which is the duration of the charging phase $\mathcal{D}$ in Figure 5.2. The latter is evaluated by the likelihood or frequency of playback frozens during the media playout. The trade-off between the two aspects of video quality is adapted by the playback threshold b. A larger threshold b results in a longer start-up delay, but makes the playback less likely freeze during the media playback. In what follows, we develop a mathematical framework to investigate this trade-off and evaluate the impacts of the threshold b and network statistics on the two quality metrics.

Diffusion Approximation

To evaluate the length of start-up delay $\mathcal{D}$ and the frequency of playback frozens, we model the playout buffer as a $G/G/1/\infty$ queue and resort to the diffusion approximation [87, 88] for compact solutions [89].

Denote the buffer size at time instant t by $B(t)$. The diffusion approximation method consists in replacing the discrete buffer size $B(t)$ by a continuous process $X(t)$ and model it as the Brownian motion,

$$dX(t) = X(t+dt) - X(t) = \beta dt + G\sqrt{\alpha dt}, \tag{H-1}$$

where $G \sim N(0,1)$ is a normally distributed random variable with zero mean and unit variance. β and α are called drift and diffusion coefficients, respectively, defined by

$$\begin{cases} \beta = E(\lim\limits_{\Delta t \to 0} \frac{X(t+\Delta t)-X(t)}{\Delta t}) = \lambda - \mu \\ \alpha = Var(\lim\limits_{\Delta t \to 0} \frac{X(t+\Delta t)-X(t)}{\Delta t}) = \lambda^3 v_a + \mu^3 v_s. \end{cases} \tag{H-2}$$

Let $p(x,t|x_0)$ denote the conditional probability density function (p.d.f.) of the buffer size $X(t)$ at time t,

$$p(x,t|x_0) = \Pr\left(x \leq X(t) < x + dx | X(0) = x_0\right), \tag{H-3}$$

where x_0 is the initial queue length. With the diffusion approximation, $p(x,t|x_0)$ can be characterized by the (forward) diffusion equation

$$\frac{\partial p(x,t|x_0)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 p(x,t|x_0)}{\partial x^2} - \beta \frac{\partial p(x,t|x_0)}{\partial x}, \tag{H-4}$$

with the initial condition

$$p\left(x,0|x_0\right) = \delta\left(x - x_0\right). \tag{H-5}$$

By applying the diffusion approximation, we can exploit the transient solution of the queue length by obtaining its p.d.f. at any time instant t.

Start-up Delay $\mathcal{D}$

We first evaluate the start-up delay by analyzing the charging phase. In the charging phase shown in Figure 5.2, the buffer is initially empty, i.e., $x_0 = 0$, and the playback is frozen, i.e., $\mu = v_s = 0$. This phase terminates when b packets are stored. The duration of charging phase or the start-up delay is thus given by

$$\mathcal{D} = \min\{t | X(0) = 0, X(t) = b, t > 0\}. \tag{H-6}$$

Note that $\mathcal{D}$ is a random variable. In what follows, we evaluate it by showing its density function and statistics.

To this end, we model the charging phase as a diffusion process with drift $\beta_D = \lambda$ and diffusion coefficient $\alpha_D = \lambda^3 v_a$ based on (H-2). Define $P_D(x,t|0)$ as the conditional CDF of the buffer size $X(t)$ in the charging phase. During this phase, the initial buffer is empty and the queue length $X(t)$ is less than b. Thus, we have

$$P_D(x,t|0) = \Pr\{X(t) \leq x | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\}, \tag{H-7}$$

The CDF of the start-up delay is given by

$$G_D(t) = \Pr\{\mathcal{D} \le t\} = 1 - P_D(b, t|0) = 1 - \int_0^b p_D(y, t|0)dy, \tag{H-8}$$

as $P_D(b, t|0)$ represents the probability that $X(t)$ is still below b at time t, i.e., $\mathcal{D}$ is greater than t. $p_D(x, t|0) = \Pr\{x \le X(t) < x + dx | X(0) = 0, X(\tau) < b \text{ for } 0 < \tau < t\}$ is the p.d.f. of $X(t)$ in the charging phase.

The p.d.f. of $\mathcal{D}$ is hence obtained as

$$g_D(t) = \frac{dG_D(t)}{dt} = -\frac{d}{dt} P_D(b, t|0) = -\frac{d}{dt} \int_0^b p_D(y, t|0)dy. \tag{H-9}$$

As $p_D(x, t|0)$ can be described by the diffusion approximation with the queue length never exceeding b [88], it follows the diffusion equation (H-4), as

$$\frac{\partial p_D(x, t|0)}{\partial t} = \frac{\alpha_D}{2} \frac{\partial^2 p_D(x, t|0)}{\partial x^2} - \beta_D \frac{\partial p_D(x, t|0)}{\partial x}, \quad x < b, \tag{H-10}$$

coupled with the initial condition

$$p_D(x, 0|0) = \delta(x), \tag{H-11}$$

and the boundary condition

$$p_D(b, t|0) = 0. \tag{H-12}$$

(H-12) is obtained by the event that the diffusion process terminates when $X(t) = b$. This is imposed by the absorbing barrier in the diffusion process [90].

Solving (H-10) with (H-11) and (H-12) yields[1]

$$p_D(x, t|0) = \frac{1}{\sqrt{2\pi\alpha_D t}} \left[\exp\left\{ -\frac{(x - \beta_D t)^2}{2\alpha_D t} \right\} - \exp\left\{ \frac{2\beta_D b}{\alpha_D} - \frac{(x - 2b - \beta_D t)^2}{2\alpha_D t} \right\} \right] \tag{H-13}$$

and

$$P_D(x, t|0) = \Phi\left(\frac{x - \beta_D t}{\sqrt{\alpha_D t}} \right) - \exp\left\{ \frac{2\beta_D b}{\alpha_D} \right\} \Phi\left(\frac{x - 2b - \beta_D t}{\sqrt{\alpha_D t}} \right), \tag{H-14}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$.

[1] The solution is obtained by the method of images as shown in [90, 91].

Substituting (H-14) into (H-8) and (H-9), we can obtain the CDF of $\mathcal{D}$,

$$G_D(t) = 1 - \Phi\left(\frac{b - \beta_D t}{\sqrt{\alpha_D t}}\right) + \exp\left\{\frac{2\beta_D b}{\alpha_D}\right\}\Phi\left(-\frac{b + \beta_D t}{\sqrt{\alpha_D t}}\right), \tag{H-15}$$

and its p.d.f.

$$g_D(t) = \frac{b}{\sqrt{2\pi\alpha_D t^3}}\exp\left\{-\frac{(b - \beta_D t)^2}{2\alpha_D t}\right\}. \tag{H-16}$$

The moment generating function (m.g.f.), represented by the Laplace transform, of $g_D(t)$ is, [92]

$$g_D^*(s) = E(e^{-st}) = \exp\left[\frac{b}{\alpha_D}\left\{\beta_D - \sqrt{\beta_D^2 + 2s\alpha_D}\right\}\right]. \tag{H-17}$$

Based on the m.g.f. $g_D^*(s)$, the mean and variance of the start-up delay with the playback threshold b can be derived accordingly,

$$E(\mathcal{D}) = -\frac{d}{ds}g_D^*(s)\bigg|_{s=0} = \frac{b}{\lambda}, \tag{H-18}$$

$$Var(\mathcal{D}) = \frac{d^2}{ds^2}g_D^*(s)\bigg|_{s=0} - E^2(\mathcal{D}) = bv_a. \tag{H-19}$$

(H-18) and (H-19) indicate that the expected value and variance of start-up delay increase linearly with the playback threshold b.

Playback Duration $\mathcal{T}$

The video playback starts immediately after the charging phase. With a longer playback duration $\mathcal{T}$, less playback frozens will be encountered, and hence the length of $\mathcal{T}$ is critical to the smoothness of media playback. Without loss of generality, we focus on one playback phase and model it as a diffusion process starting at time $t = 0$. As the playback phase terminates when the buffer becomes empty again, the playback duration is thus given by

$$\mathcal{T} = \min\{t | X(0) = b, X(t) = 0, t > 0\}. \tag{H-20}$$

Denote $g_T(t)$ and $G_T(t)$ as the p.d.f. and CDF of $\mathcal{T}$, respectively. Same as the start-up delay $\mathcal{D}$, we evaluate $\mathcal{T}$ by showing its density function.

Given that the buffer size is b at the beginning of the charging phase, the probability that the buffer size $X(t)$ is larger than x at time t is given by,

$$\begin{aligned} P_T(x,t|b) &= \Pr\{X(t) > x|X(0) = b, X(\tau) > 0 \text{ for } 0 < \tau < t\} \\ &= \int_x^\infty p_T(y|t,b)dy, \end{aligned} \tag{H-21}$$

where $p_T(y,t|b) = \Pr\{y \leq X(t) < y + dy|X(0) = b, X(\tau) > 0 \text{ for } 0 < \tau < t\}$ is the p.d.f. of $X(t)$ at time t in the playback phase, given the initial buffer size b.

Similar to the computation of start-up delay, we have

$$g_T(t) = -\frac{d}{dt}\int_0^\infty p_T\left(x,t|b\right)dx. \tag{H-22}$$

where $p_T\left(x,t\right)$ follows the diffusion equation,

$$\frac{1}{2}\alpha_T\frac{\partial^2 p_T(x,t|b)}{\partial x^2} - \beta_T\frac{\partial p_T(x,t|b)}{\partial x} = \frac{\partial p_T(x,t|b)}{\partial t}, \tag{H-23}$$

subject to the initial and boundary conditions

$$p_T(x,0|b) = \delta\left(x-b\right), \quad t = 0, \tag{H-24}$$

$$p_T(0,t|b) = 0, \qquad t > 0. \tag{H-25}$$

(H-25) is dictated by the events that the playback phase terminates when the buffer becomes empty. β_T and α_T can be derived from (H-2).

Solving the diffusion equations (H-23), (H-24) and (H-25), we have

$$p_T\left(x|t,b\right) = \frac{\exp\left\{\frac{\beta_T}{\alpha_T}\left(x-b\right) - \frac{\beta_T^2}{2\alpha_T}t\right\}}{\sqrt{2\pi\alpha_T t}}\left[\exp\left\{-\frac{\left(x-b\right)^2}{2\alpha_T t}\right\} - \exp\left\{-\frac{\left(x+b\right)^2}{2\alpha_T t}\right\}\right]. \tag{H-26}$$

Substitute (H-26) into (H-22), we have

$$g_T(t) = \frac{b}{\sqrt{2\pi\alpha_T t^3}}\exp\left\{-\frac{(\beta_T t + b)^2}{2\alpha_T t}\right\}, \tag{H-27}$$

and its m.g.f.

$$g_T^*(s) = \exp\left\{-\frac{b}{\alpha_T}(\beta_T + \sqrt{\beta_T^2 + 2\alpha_T s})\right\}. \tag{H-28}$$

Smoothness of Playback

With the p.d.f. of $\mathcal{T}$ in hand, we are now ready to evaluate the smoothness of playback in terms of two metrics, namely the stopping probability $\mathcal{P}$ and frequency of playback frozens $\mathcal{F}$.

Stopping Probability $\mathcal{P}$ The stopping probability $\mathcal{P}$ represents the probability that the playback freezes in the middle of media playout, mathematically,

$$\mathcal{P} = \Pr(t < S | B(0) = b, B(t) = 0), \tag{H-29}$$

where S denotes the length of the video file.

Substituting (H-27) into (H-29), we have

$$\mathcal{P} = \int_0^S g_T(t)dt. \tag{H-30}$$

To obtain the closed-form expression on $\mathcal{P}$, we approximate S to be infinity as

$$\mathcal{P} \approx \lim_{S \to \infty} \int_0^S g_T(t)dt = \lim_{s \to 0} g_T^*(s) = \begin{cases} 1, & \text{if } \beta_T \leq 0, \\ \exp\left\{ -\frac{2b}{\alpha_T} \beta_T \right\}, & \text{if } \beta_T > 0. \end{cases} \tag{H-31}$$

Note that the obtained stopping probability is conservative as in reality S is limited. However, this approximation does not generate much difference as S is considerably large compared to the video frame intervals.

Substitute (H-2) into (H-31), we have

$$\mathcal{P} \approx \begin{cases} 1, & \text{if } \lambda \leq \mu, \\ \exp\left\{ -\frac{2b}{\lambda^3 v_a + \mu^3 v_s} (\lambda - \mu) \right\}, & \text{if } \lambda > \mu. \end{cases} \tag{H-32}$$

(H-32) indicates that the video playback stops with probability 1 when the mean download rate λ is less than or equal to the video playback rate μ. In the mean time, even if the mean traffic arrival rate or video download rate λ exceeds the average video playback rate μ, it is still possible that video playback stops due to the variance of packet arrivals and playback. In the real-world deployments, $\lambda - \mu$ is normally controlled small to admit more users in the system. In this case, the stopping probability $\mathcal{P}$ is heavily dependent on the threshold b and the statistics of the network.

Number of Playback Frozens $\mathcal{F}$ In (H-32), we have shown that when the mean traffic arrival rate λ is smaller than the average video playback rate μ, the video playout will stop with probability one. To shed light on how serious the interruptions of playback are in this case, we derive the overall number of playback frozens, denoted by $\mathcal{F}$, encountered during the media playback. Let R.V. $\mathcal{M}$ denote the duration between two consecutive playback frozen events, and we have $\mathcal{M} = \mathcal{D} + \mathcal{T}$, as shown in Figure 5.2. It is obvious that $\mathcal{F}$ is negatively proportional to $\mathcal{M}$. In what follows, we show its density function and statistics.

Denote the p.d.f. of $\mathcal{M}$ as $g_M(t)$. Hence,

$$g_M(t) = g_D(t) \otimes g_T(t), \tag{H-33}$$

where $\otimes$ denotes convolution. The m.g.f. of $g_M(t)$ is thus given by

$$g_M^*(s) = g_D^*(s) \cdot g_T^*(s). \tag{H-34}$$

Substitute (H-17) and (H-28) into (H-34), we can obtain the mean and variance of $\mathcal{M}$ as

$$E\left(\mathcal{M}\right) = -\frac{d}{ds} g_M^*(s)\bigg|_{s=0} = \frac{-b\mu}{\lambda\left(\lambda-\mu\right)}, \quad \lambda < \mu, \tag{H-35}$$

and

$$Var\left(\mathcal{M}\right) = \frac{d^2}{ds^2} g_M^*(s)\bigg|_{s=0} - E^2(\mathcal{M}) = -b\frac{\mu^3(v_s+v_a)+3v_a\lambda\mu\left(\lambda-\mu\right)}{\left(\lambda-\mu\right)^3}, \quad \lambda < \mu. \tag{H-36}$$

Next, we use the diffusion approximation to obtain the p.d.f. of $\mathcal{F}$. Specifically, we assume that there is a virtual event buffer B_F which counts the events of playback frozen. Whenever an event of playback frozen happens, we increase the queue length of B_F by one. Thus, the buffer size of B_F at time t, denoted by $X_F(t)$, represents the number of playback frozens up to time t. The interarrival time between two consecutive increments of $X_F(t)$ is $\mathcal{M}$, where $X_F(t)$ is a non-decreasing function of time t. Denote by $P_F(x,t|0)$ the conditional CDF of $X_F(t)$ at time t, given the initial buffer size 0,

$$P_F(x,t|0) = \Pr\{X_F(t) \leq x | X_F(0)= 0\}. \tag{H-37}$$

Similarly, $X_F(t)$ can be approximated as a continuous function by applying diffusion equation, and its CDF is governed by

$$\frac{\partial P_F(x,t|0)}{\partial t} = \frac{\alpha_F}{2}\frac{\partial^2 P_F(x,t|0)}{\partial x^2} - \beta_F\frac{\partial P_F(x,t|0)}{\partial x}, \tag{H-38}$$

coupled with the boundary condition

$$\begin{cases} \lim_{x\to\infty} P_F(x,t|0) = 1, & t \geq 0, \\ \lim_{x\to 0} P_F(x,t|0) = 0, & t \geq 0. \end{cases} \tag{H-39}$$

where $\beta_F = \frac{1}{E(M)}$ and $\alpha_F = \frac{Var(M)}{E^3(M)}$ can be derived from (H-2), (H-35) and (H-36).

Solving (H-38) and (H-39), we have

$$P_F(x,t|0) = \Phi\left(\frac{x - \beta_F t}{\sqrt{\alpha_F t}}\right) - \exp\left\{\frac{2\beta_F x}{\alpha_F}\right\}\Phi\left(-\frac{x+\beta_F t}{\sqrt{\alpha_F t}}\right). \tag{H-40}$$

The mean and variance of the number of playback frozens at time t, when $\lambda < \mu$, can be approximated as

$$E(\mathcal{F}) \approx \beta_F t = -\frac{\lambda\,(\lambda-\mu)}{\mu b}t, \tag{H-41}$$

$$Var(\mathcal{F}) \approx \alpha_F t = \frac{\mu^2\lambda^3(v_s+v_a) + 3v_a\lambda^4\,(\lambda-\mu)}{b^2\mu^2}t, \tag{H-42}$$

as $\exp\{\frac{2\beta_F x}{\alpha_F}\}\Phi\left(-\frac{x+\beta_F t}{\sqrt{\alpha_F t}}\right)$ decreases dramatically when t is large.

5.3.3 Finite Buffer Case

In this subsection, we consider the case that the playout buffer is limited compared with the volume of video file. This is typical when the end users use personal devices with limited buffer size and hard disk such as handsets.

The start-up delay $\mathcal{D}$ obtained in the previous subsection is also valid in the finite buffer case as the start and termination conditions of the charging phase in both cases are the same. As shown in Figure 5.2, in the playback phase, the queue length of the playout buffer is upper bounded by the buffer size, denoted by N ($N > b$). When the playout buffer is full, the arrival video packets will be dropped, which not only degrades the user's video quality but also results in the bandwidth waste. Therefore, a key performance metric in this case is the packet loss probability due to buffer overflow. In this chapter, the packet loss probability and the buffer overflow probability are interchangeably used.

Let $\mathcal{L}$ denote the packet loss probability of the playout buffer,

$$\mathcal{L} = \lim_{t\to\infty} \Pr\left(B(t) = N\right). \tag{H-43}$$

The smoothness of playback is evaluated by the charging probability, denoted by $\mathcal{C}$, which is defined as the probability that the playback is frozen and the playout buffer is in the charging phase at any time instant.

We invoke the diffusion approximation to analyze playback phase in the finite buffer case and evaluate $\mathcal{L}$ and $\mathcal{C}$ in terms of network statistics and threshold of playback, as

$$\frac{\partial p\left(x,t|b\right)}{\partial t}=\frac{1}{2}\alpha_T\frac{\partial^2 p\left(x,t|b\right)}{\partial x^2}-\beta_T\frac{\partial p\left(x,t|b\right)}{\partial x}+\frac{\lambda}{b}\mathcal{C}\delta\left(x-b\right)+\mu\mathcal{L}\delta\left(x-N+1\right), \quad \text{(H-44)}$$

$$\lim_{x\to 0}\left[\frac{\alpha_T}{2}\frac{\partial p\left(x,t|b\right)}{\partial x}-\beta_T p\left(x,t|b\right)\right]=\frac{\lambda}{b}\mathcal{C}, \quad \text{(H-45)}$$

$$\lim_{x\to N}\left[\frac{\alpha_T}{2}\frac{\partial p\left(x,t|b\right)}{\partial x}-\beta_T p\left(x,t|b\right)\right]=-\mu\mathcal{L}, \quad \text{(H-46)}$$

subject to the initial and boundary conditions

$$\lim_{x\to 0^+} p\left(x,t|b\right)=0 \quad t>0,$$
$$\lim_{x\to N^-} p\left(x,t|b\right)=0 \quad t>0,$$

where $\delta(x)$ is the Dirac delta function; $p(x,t|b)=\Pr\left\{x\le X(t)<x+dx|X(0)=b\right\}$ is the conditional p.d.f. of the queue length $X(t)$.

The probability density in (H-44) is composed of two parts, the p.d.f. of the queue length $p\left(x,t|b\right)$ when $x\in(0,N)$ and the p.m.f. $\mathcal{L}$ and $\mathcal{C}$ on the two boundaries when buffer is full and in the charging phase, respectively. $\frac{\lambda}{b}\mathcal{C}\delta\left(x-b\right)$ in (H-44) evaluates the probability that the queue changes from the charging phase to the playout phase, where $\frac{\lambda}{b}$ is the mean rate of the change computed as $\frac{1}{E(\mathcal{D})}$. $\mu\mathcal{L}\delta\left(x-N+1\right)$ evaluates the probability that the queue length jumps from N to $N-1$ with packets being served at the mean rate μ. More details about the rational behind the equations can be found in [93].

It is important to note that the diffusion process of playback phase in (H-44) is different from that in (H-23). (H-23) describes a single playback phase that starts when b packets are stored in the charging phase and terminates once the playout buffer becomes empty. However, (H-44) represents the whole media playback process from the first playback to the instant when video is wholly downloaded. The reason we consider the whole session of video playback as one diffusion process in this case is that it facilitates to evaluate the long-term buffer overflow and underflow probability in the steady state. While in the infinite buffer case, we are more interested in the transient behavior of the queue to evaluate the duration of each start-up delay and playback phase.

Solving (H-44) at the steady state when $\lim\limits_{t\to\infty} \frac{\partial p(x,t|b)}{\partial t} = 0$, we have

$$p\left(x,\infty|b\right) = \begin{cases} \frac{\lambda\mathcal{L}}{b\beta_T}\left(e^{rx}-1\right), & 0 < x \le b, \\ \frac{\lambda\mathcal{L}}{b\beta_T}\left(1-e^{-rb}\right)e^{rx}, & b < x \le N-1, \\ \frac{\mu\mathcal{C}}{\beta_T}\left(1-e^{r(x-N)}\right), & N-1 < x < N, \end{cases} \tag{H-47}$$

where $r = \frac{2\beta_T}{\alpha_T}$, and the packet loss probability $\mathcal{L}$ and the charging probability $\mathcal{C}$ are given by

$$\mathcal{L} = \left(\frac{-\left(1-e^{-r}\right)\mu^2 b}{\lambda\beta_T\left(1-e^{-rb}\right)e^{r(N-1)}} + \frac{\lambda}{\beta_T}\right)^{-1}, \tag{H-48}$$

$$\mathcal{C} = \left(-\frac{\mu}{\beta_T} + \frac{\lambda^2}{\beta_T b\mu}\frac{e^{r(N-1)}\left(1-e^{-rb}\right)}{1-e^{-r}}\right)^{-1}. \tag{H-49}$$

The infinite playout buffer could be regarded as a special case of the finite playout buffer when $N \to \infty$. In specific, when $\beta_T < 0$, i.e., $\lambda < \mu$, with (H-48) and (H-49), we have

$$\lim_{N\to\infty} \mathcal{L} = 0, \quad \lambda < \mu, \tag{H-50}$$

as no packets will be lost when $N \to \infty$, and

$$\lim_{N\to\infty} \mathcal{C} = -\frac{\beta_T}{\mu}, \quad \lambda < \mu. \tag{H-51}$$

(H-51) matches the results of infinite buffer case as

$$\lim_{N\to\infty} \mathcal{C} = \frac{E(\mathcal{F}) \times E\left(\mathcal{D}\right)}{t}, \quad \lambda < \mu, \tag{H-52}$$

where $E\left(\mathcal{D}\right)$ and $E(\mathcal{F})$ are given in (H-18) and (H-41), respectively.

Therefore, we show the video quality in terms of start-up delay and smoothness of video playback with given statistics of the download and playback rates of video packets. The mean video playback rate μ is usually fixed and given for non-scalable video. When scalable video coding, e.g., layer-encoded video streaming [94], is used, the playback rate can be adjustable when different video layers are downloaded. The statistics of playback rate in this case are no longer predefined and therefore need to be measured at the receiver.

5.4 Quality Driven Playout Buffer Management

In this section, by exploiting the obtained video quality metrics from the analytical framework, we determine the optimal playback threshold to achieve the maximal user utility, based on different video requirements of end users. Towards this goal, we formulate the playback threshold selection as a stochastic optimization problem.

5.4.1 Infinite Buffer Case

Let $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{F}}$ denote the maximum tolerable start-up delay and number of playback frozens input by the users, respectively. Our objective is to manage the threshold of playback b to maximize the user perceived video quality within the tolerable range specified by $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{F}}$, mathematically,

$P1:$ if $\lambda > \mu$,

$$\begin{aligned} \min_{b} \quad & \mathcal{P} + \varpi_1 \left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right) \\ s.t. \quad & \Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} \leq \zeta, \\ & b > 0. \end{aligned} \tag{H-53}$$

$P2:$ if $\lambda \leq \mu$,

$$\begin{aligned} \min_{b} \quad & E(\mathcal{F}) + \vartheta_F Var(\mathcal{F}) + \varpi_2 \left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right) \\ s.t. \quad & \Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} \leq \zeta, \\ & \Pr\left\{\mathcal{F} > \widehat{\mathcal{F}}\right\} \leq \eta, \\ & b > 0. \end{aligned} \tag{H-54}$$

where $\omega_1, \omega_2 > 0$ are the weighting factors and ϑ_D, $\vartheta_F \geq 0$ are called risk aversion factors which are adjustable with respect to different user requirements. ζ, η are predefined scalers such that $0 < \zeta, \eta << 1$.

Scheme $P1$ is implemented when the mean packet arrival rate λ is larger than the mean video playback rate μ. In this case, with probability $1 - \mathcal{P}$ the video playback can be finished without any interruptions. The objective is hence to avoid playback frozens while minimizing the start-up delay. ϖ_1 in the utility function is a knob to balance the requirements between smooth playback and start-up delay. A larger ϖ_1 represents that users are more sensitive to the start-up delay, e.g., when watching a live soccer match. ϑ_D is called risk aversion factor which models the user's attitude to the variance of start-up delay[2]. When ϑ_D is large, the users are conservative and require more strict start-up delay with low variance. The constraint is represented by a stochastic bound that the resulting start-up delay must be within the tolerable region $\widehat{\mathcal{D}}$ imposed by the user with a high probability. The stochastic QoS is considered because providing absolute QoS guarantee may not be feasible and is typically difficult and costly for implementation in the time-varying environment [95].

The scheme $P2$ is employed when the mean packet arrival rate is insufficient to meet the playback. In this case, interruptions of playback are inevitable as shown by (H-32). The objective is to minimize the number of playback frozens and the incurred start-up delay. The utility functions and constraints are defined in the same fashion of $P1$.

Both $P1$ and $P2$ are probability-constrained stochastic optimization (also referred to as chance constrained programming) [96]. By substituting (H-15), (H-18), (H-19) and (H-32) into $P1$; (H-15), (H-18), (H-19), (H-40), (H-41) and (H-42) into $P2$, we have

$P1'$: if $\lambda > \mu$,

$$\begin{array}{cc} \min\limits_{b} & \exp\left\{-\frac{2b}{\lambda^3 v_a+\mu^3 v_s}\left(\lambda-\mu\right)\right\}+\varpi_1 b\left(\frac{1}{\lambda}+\vartheta_D v_a\right) \\ s.t. & \Phi\left(\frac{b-\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right)-\exp\left\{\frac{2b}{\lambda^2 v_a}\right\}\Phi\left(-\frac{b+\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right)\leq\zeta, \\ & b\geq 0, \end{array} \tag{H-55}$$

$P2'$: if $\lambda \leq \mu$,

$$\begin{array}{cc} \min\limits_{b} & \frac{A}{b}+\frac{\vartheta_F}{b^2}B+\varpi_2 b\left(\frac{1}{\lambda}+\vartheta_D v_a\right) \\ s.t. & \Phi\left(\frac{b-\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right)-\exp\left\{\frac{2b}{\lambda^2 v_a}\right\}\Phi\left(-\frac{b+\lambda\widehat{\mathcal{D}}}{\sqrt{\lambda^3 v_a\widehat{\mathcal{D}}}}\right)\leq\zeta, \\ & 1-\Phi\left(\frac{\widehat{\mathcal{F}}-\beta_F S}{\sqrt{\alpha_F S}}\right)+\exp\{\frac{2\beta_F\widehat{\mathcal{F}}}{\alpha_F}\}\Phi\left(-\frac{\widehat{\mathcal{F}}+\beta_F S}{\sqrt{\alpha_F S}}\right)\leq\eta, \\ & b\geq 0, \end{array} \tag{H-56}$$

where $A=-\frac{\lambda(\lambda-\mu)}{\mu}S, B=\frac{\mu^2\lambda^3(v_s+v_a)+3v_a\lambda^4(\lambda-\mu)}{\mu^2}S$ are positive scalers. Here, the statistics of network and video playback rate, i.e., λ, v_a, μ and v_s, and the video length S are known and used as inputs to the control scheme. This is reasonable as those network statistics can be measured in real time at the user end.

Both $P1'$ and $P2'$ are nonlinear programming problems which may be prohibitively expensive for practical real-time streaming systems. To reduce the computation complexity,

[2] This utility function is defined in the fashion of Markowitz mean-variance model which is widely used in portfolio optimization.

we apply the one-sided Chebyshev inequality, which states that for any R.V. χ and any positive real number x,

$$\Pr\{\chi - E(\chi) \geq x\} \leq \frac{Var(\chi)}{Var(\chi) + x^2}, \text{ for } \chi > E(\chi). \tag{H-57}$$

Using the Chebyshev inequality, together with (H-18), (H-19), (H-41) and (H-42), the constraints of $P2$ become

$$b \leq \widehat{\mathcal{D}}\lambda + \frac{v_a\left(1-\zeta\right) - \sqrt{\frac{4\widehat{\mathcal{D}}\zeta}{\lambda}v_a\left(1-\zeta\right) + v_a^2\left(1-\zeta\right)^2}}{2\zeta/\lambda^2}, \tag{H-58}$$

$$b \geq \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta\left(1-\eta\right)}}{\eta\widehat{\mathcal{F}}}, \tag{H-59}$$

where A and B are the same as those in (H-56). The details are shown in Appendix 5.8.4.

By replacing the constraints of $P1$ and $P2$ with (H-58) and (H-59), both $P1$ and $P2$ become convex optimization problems which could be solved efficiently. Note that computation complexity is reduced at the expanse of user's utility, because comparing with $P1'$ and $P2'$, the new constraints obtained with the Chebyshev inequality is more conservative, resulting in a smaller feasible region. However, a conservative but fast algorithm is desirable for practical use.

To ensure that the resultant video performance is within the tolerable region, the threshold of playback b must be within the range specified by (H-58) and (H-59). To make this condition satisfied, we could apply call admission control at the user end. In this way, the request of playback is reject directly at agent of the end host without sending it to the media server if there is no positive b to meet both (H-58) and (H-59). Thus, the network resources can be efficiently utilized to provision video quality for all admitted videos.

5.4.2 Finite Buffer Case

We optimize the video playback in the finite buffer case. Our objective is to control the threshold of playback b to minimize the interruptions of video playback due to buffer empty and the packet loss caused by the buffer overflow. The minimization problem, in this case, can be represented as

$$\begin{array}{ll} \min\limits_{b} & \rho_1\mathcal{L} + \rho_2\frac{\mathcal{C}\times S}{E(\mathcal{D})} + \varpi_1\left(E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})\right) \\ s.t. & \Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} \leq \zeta, \\ & b > 0, \end{array} \tag{H-60}$$

where ρ_1 and ρ_2 are the weighting factors of packet loss and charging probabilities, respectively. ϖ_1, ϑ_D and ζ are defined in the same manner as those in the infinite buffer case.

In (H-60), $\frac{\mathcal{C}\times S}{E(\mathcal{D})}$ represents the mean number of playback frozens where $\mathcal{C}\times S$ computes the overall time spent in the charging phase. The objective is to balance the trade-off between the video quality metrics, i.e., the smoothness of playback, packet loss and the encountered start-up delay.

In a summary, this section provides examples to apply the achieved analytical results to the optimal receiver buffer design. This leverages the property that the analytical results bridge the network throughput with the user perceived video quality as,

$$(\mathcal{D}, \mathcal{F}) = f\left(\lambda, v_a, \mu, v_s, b\right), \tag{H-61}$$

where the mapping function $f(\cdot)$ could be represented by the constraints of $P1$ and $P2$ (or (H-58) and (H-59)). In a reverse manner, we can also obtain the desired network resource with given user requirements as

$$(\lambda, v_a) = f^{-1}\left(\mathcal{D}, \mathcal{F}, \mu, v_s, b\right). \tag{H-62}$$

This is useful as the guideline of the network resource allocation to achieve specific video quality requirements. For non-scalable video coding, the video playback rate μ is usually fixed and only the playback threshold b is adjusted to adapt to the required video quality. In this case, the presented optimization framework can be applied directly. When layered-encoded video coding is used, the video playback rate could also be adjustable [94] and the problem can be extended to a joint optimization framework of playback threshold and playback rate (or video layers) selections. We will pursue the joint optimization problem in our future work.

5.5 Simulation Results

In this section, we verify the achieved analytical results using extensive simulations, based on a trace-driven discrete event simulator coded in C++.

5.5.1 Simulation Setup

We use two real VBR video clips, “Aladdin” and “Susi & Strolch”, from [97] encoded by MPEG-4 with diverse frame statistics. Each video clip lasts $S = 1$ hour and the sequences are encoded at a constant frame rate of 25 frames per second in the Quarter Common Intermediate Format (QCIF) resolution (176 × 144). The statistics of video frames are summarized in Table 5.1.

The simulated network is shown in Figure 5.1. In each simulation run, the simulator loads video frames from the video trace file and segment the variable size video frames into IP packets with the maximum size of $1,400$ Bytes. The available bandwidth of the network varies over time during which the overall variable bit rates of the channels are 10 Kbps, 500 Kbps, 2 Mbps, and 4 Mbps with probability 0.02, 0.48, 0.30 and 0.20, respectively. The average throughput is thus 1.64 Mbps; the mean and standard deviation of network delay are $\frac{1}{\lambda} = 35.4$ ms and $\sqrt{v_a} = 155.2$ ms, respectively. The video file is played at a constant rate 25 frames/sec by default and variable packet rates as shown in Table 5.1. For each scenario, we conduct 30 simulation runs and plot the mean results with the 95% confidence intervals.

5.5.2 Infinite Buffer Case

We first examine the case of infinite playout buffer.

Start-up Delay $\mathcal{D}$ and Playback Duration $\mathcal{T}$

In the first simulation, we verify the analysis of the start-up delay $\mathcal{D}$ and playback duration $\mathcal{T}$. We use the trace "Aladdin" in which $\lambda < \mu$ according to Table 5.1. In this case, the playback frozens are inevitable as indicated by (H-32).

The CDF of the start-up delays and playback durations with different buffer thresholds b are shown in Figure 5.3 and Figure 5.4, respectively. In Figure 5.3, the mean start-up delay increases with b and the corresponding CDF moves to the left. In addition, the variance of start-up delay increases accordingly as the CDF curve expands in width. Similarly, it can be seen in Figure 5.4 that both mean and variance of the playback duration increase with the threshold b. The simulation results well validate our analysis.

Stopping Probability $\mathcal{P}$

We verify the stopping probability $\mathcal{P}$ in the second simulation using the clip "Susi & Strolch" where $\lambda > \mu$. In this case, the video packets are downloaded at a faster rate than that of the playback, and the stopping probability $\mathcal{P}$ is less than one as shown in (H-32). We

Table 5.1: Statistics of Video Frames

Video Clip	Frame	Frame Size		Bit Rate		Inter-departure of pkts	
Name	Number	Mean (bytes)	Variance	Mean (bit/sec)	Peak	$\frac{1}{\mu}$(msec/pkt)	Variance
Aladdin	89998	7.7e+02	5.8e+05	1.5e+05	1.3e+06	33.6	102
Susi & Strolch	89998	5.8e+02	3.9e+05	1.2e+05	1.3e+06	36.2	70.4

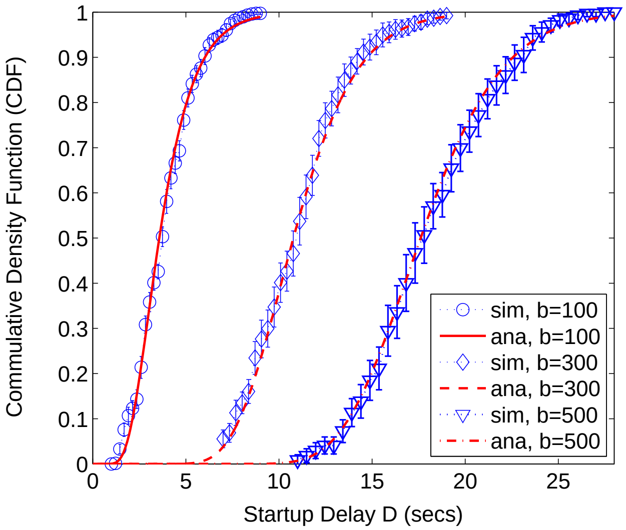


Figure 5.3: CDF of the start-up delay D for $b = 100, 300$ and 500 packets, respectively

conduct 30 experiments, and for each experiment we increase the buffer threshold b by 70 packets starting from 1 packet. Within each experiment, we conduct 500 simulation runs with each run terminated either when the playback frozen occurs or after the whole video is played without any interruptions. The simulation stopped by playback frozens are called frozen events. The probability of stopping is then computed as the total number of frozen events divided by 500. It is observed in Figure 5.5 that the probability of stopping decreases exponentially with the increase of buffer threshold b. The analytical results are slightly larger than the simulation results because the video length S is assumed infinity for analysis while S is 1 hour in the simulations.

Number of Playback Frozens $\mathcal{F}$

We study the number of playback frozens using the clip "Aladdin" with $\lambda < \mu$. In this simulation, we conduct 500 runs and measure the number of playback frozens. Figure 5.6 plots the CDF of the number of playback frozens when b is 100, 300 and 500 packets, respectively, at time $t = S$. The analysis obtained from (H-40) well match the simulation result. Meanwhile, we can see that when b increases, the CDF curve shifts to the left which implies that on average fewer events of playback frozens are encountered. However, the step size of each shift is different; the mean number of playback frozens decreases dramatically when b is initially small. The width of the CDF curves also becomes smaller with a larger

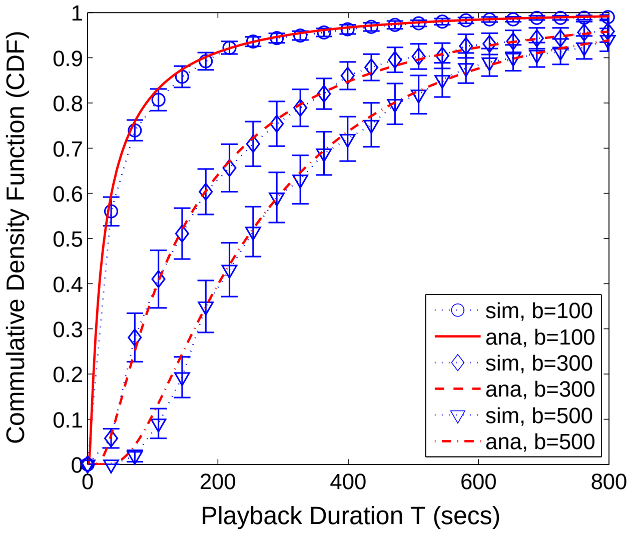


Figure 5.4: CDF of the playback duration T for $b = 100, 300$ and 500 packets, respectively, at time $t = S$

b, which implies that the variance of the number of playback frozens decreases when b increases.

Optimal Selection of Playback Threshold b

Based on the above analytical results, we apply Matlab `fmincon` to solve (H-53) and (H-54) subject to the constraints (H-58) and (H-59) for optimal playback threshold management. The default setting of parameters are in Table 5.2.

Table 5.2: Parameters in Optimal Playout Buffer Management (Infinite Buffer)

Scheme	Video Trace	$\widehat{\mathcal{D}}$	$\widehat{\mathcal{F}}$	ζ	η	ϖ_1	ϖ_2	ϑ_D	ϑ_F
$P1(\lambda > \mu)$	Susi & Strolch	120 sec	N/A	5%	N/A	0.01	N/A	1	N/A
$P2(\lambda \leq \mu)$	Aladdin	120 sec	20	5%	5%	N/A	0.1	1	1

We first show the impacts of weighting factors on the optimal selection of playback threshold. Figure 5.7 plots the optimal threshold b^* of $P1$ using trace “Aladdin”. In this scenario, we can see that the optimal threshold b^* decreases monotonically with the increasing weighting factor ϖ_1. This is because when ϖ_1 in (H-53) increases, the utility of

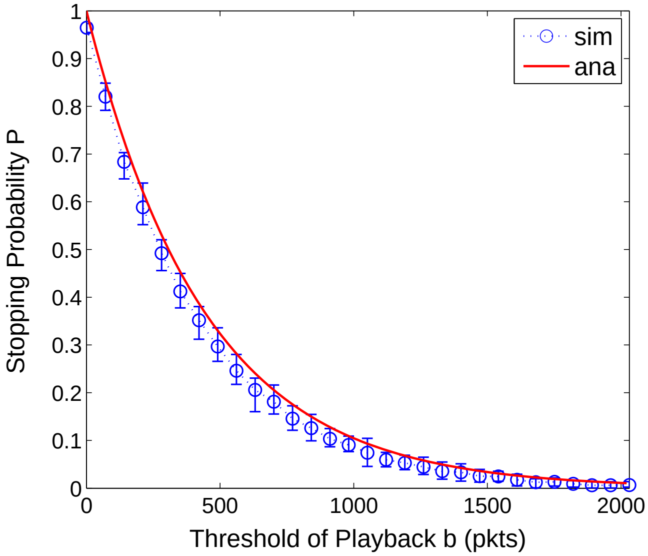


Figure 5.5: The simulated stopping probability $\mathcal{P}$

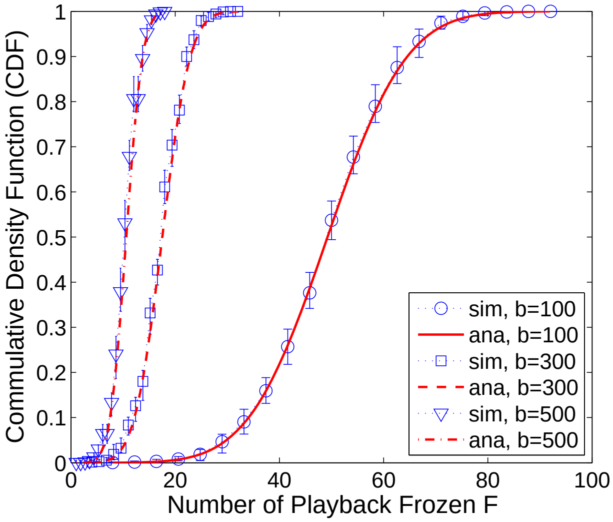


Figure 5.6: Number of playback frozens $\mathcal{F}$ for $S = 1$ hour, and $b = 100, 300$ and 500 packets, respectively

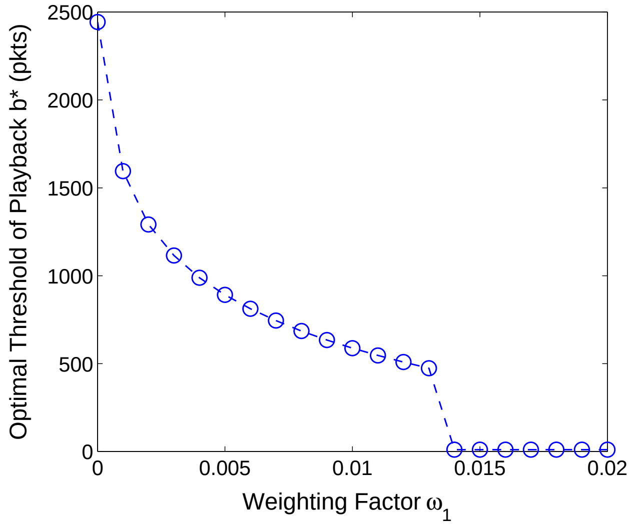


Figure 5.7: The optimal playback threshold b^* with the increasing weighting factor ϖ_1

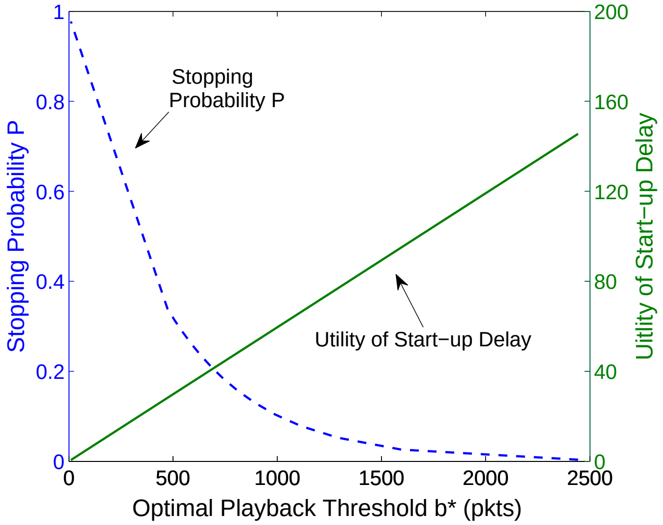


Figure 5.8: Trade-off between the stopping probability and start-up delay at different optimal playback thresholds b^*

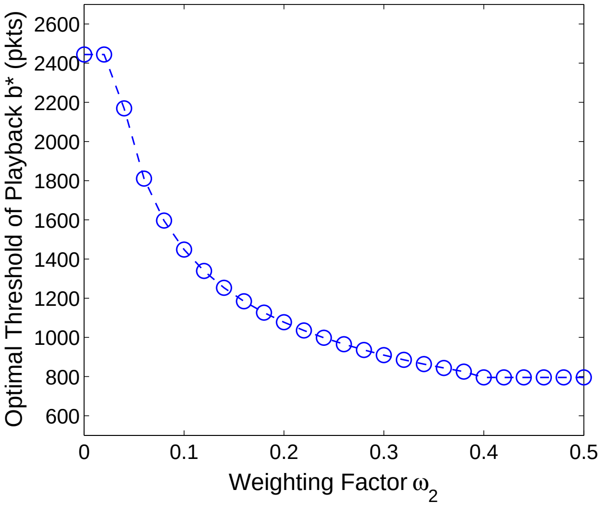


Figure 5.9: The optimal playback threshold b^* with the increasing weighting factor ϖ_2

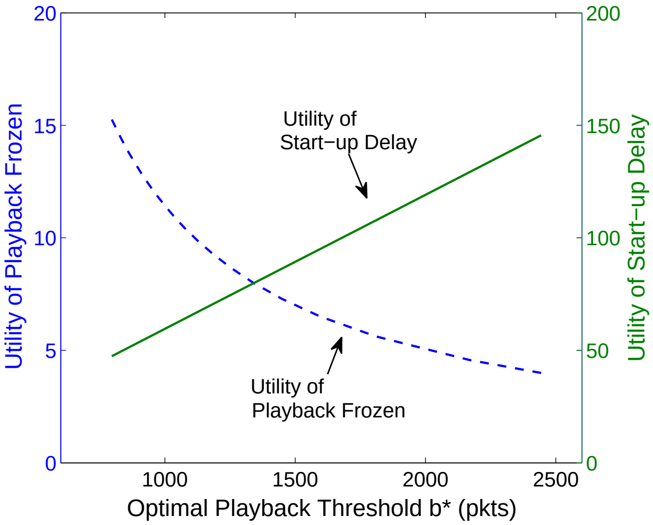


Figure 5.10: Trade-off between the number of playback frozens and start-up delay at different optimal playback thresholds b^*

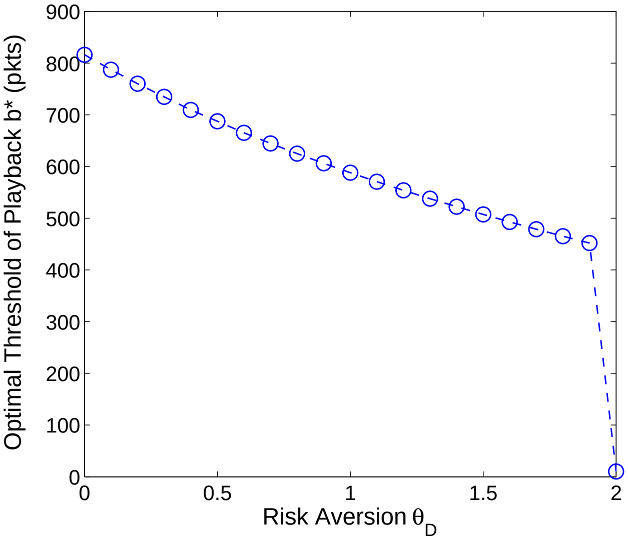


Figure 5.11: Impact of θ_D on the optimal selection of playback threshold b in $P1$

start-up delay, evaluated as $E(\mathcal{D}) + \vartheta_D Var(\mathcal{D})$, becomes more important in the objective and overwhelms the stopping probability. Therefore, the optimal threshold b^* is reduced accordingly to shrink the start-up delay at the cost of a higher playback frozen probability. The resultant stopping probability and utility of start-up delay at different optimal thresholds b^* are shown in Figure 5.8, where the utility of start-up delay is computed as $b^* \left(\frac{1}{\lambda} + \vartheta_D v_a\right)$ which is the portion of utility characterized by the start-up delay in (H-55). Figure 5.9 plots the optimal thresholds b^* of $P2$ with the increasing weighting factor ϖ_2. We can see that the corresponding b^* decreases, as the utility of start-up delay becomes more critical when ϖ_2 increases. When ϖ_2 is very small, b^* is upper bounded by 2500 packets to guarantee that the resulting start-up delay is within the tolerable value $\widehat{\mathcal{D}}$. When ϖ_2 is very large, b^* is lower bounded by 800 packets to assure that the tolerable value $\widehat{\mathcal{F}}$ is not violated. The resultant utilities of playback frozens and start-up delay at the different optimal thresholds of playback b^* are shown in Figure 5.10 where the utility of playback frozens is characterized by playback frozen in (H-56) computed as $\frac{A}{b^*} + \vartheta_F \frac{B}{(b^*)^2}$. The utility of start-up delay is computed in the same manner as that in Figure 5.8.

The impacts of risk aversion factors are shown in Figs. 5.11-5.13. Figs. 5.11 and 5.12 show the impacts of ϑ_D in schemes $P1$ and $P2$, respectively, with all the other parameters the same as in Table 5.2. With an increasing ϑ_D, the optimal playback threshold b decreases in both schemes, $P1$ and $P2$. This is because that users require a strictly small variance of start-up delay. Figure 5.13 shows the impact of ϑ_F in $P2$. With ϑ_F increasing, the

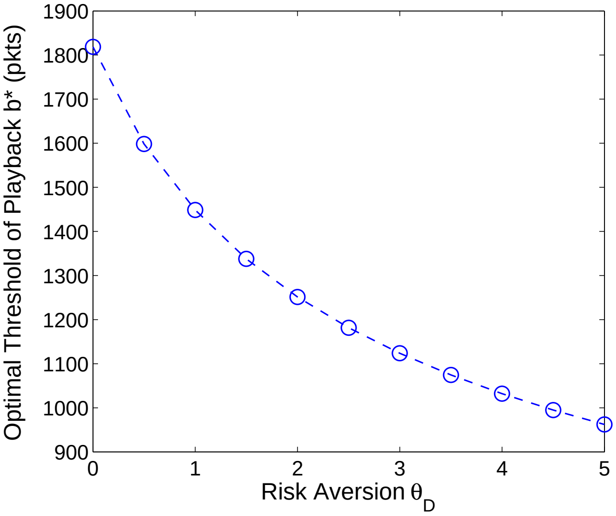


Figure 5.12: Impact of θ_D on the optimal selection of playback threshold b in $P2$

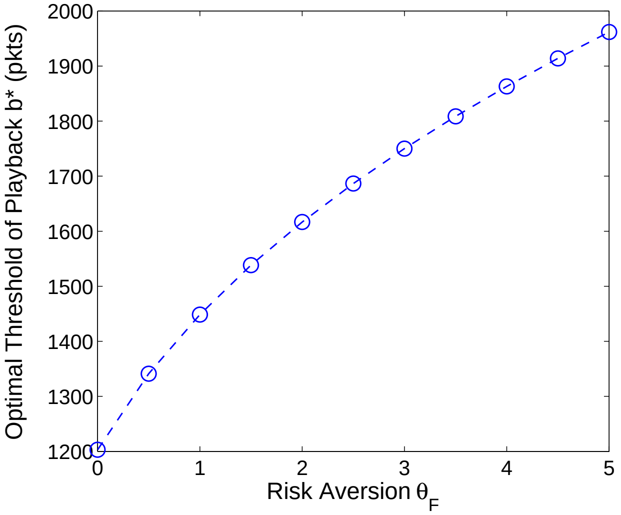


Figure 5.13: Impact of θ_F on the optimal selection of playback threshold b in $P2$

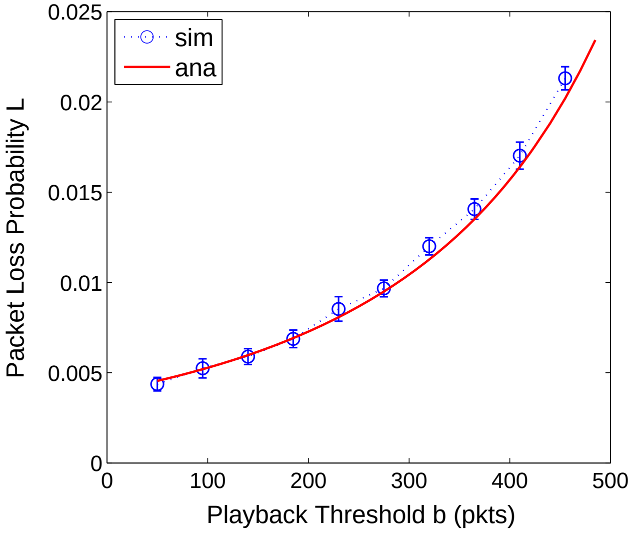


Figure 5.14: Packet loss probability with the increasing playback threshold b

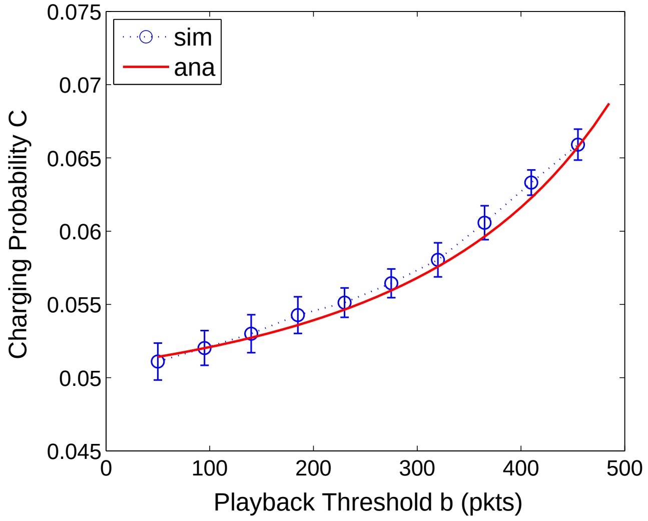


Figure 5.15: Charging probability with the increasing playback threshold b

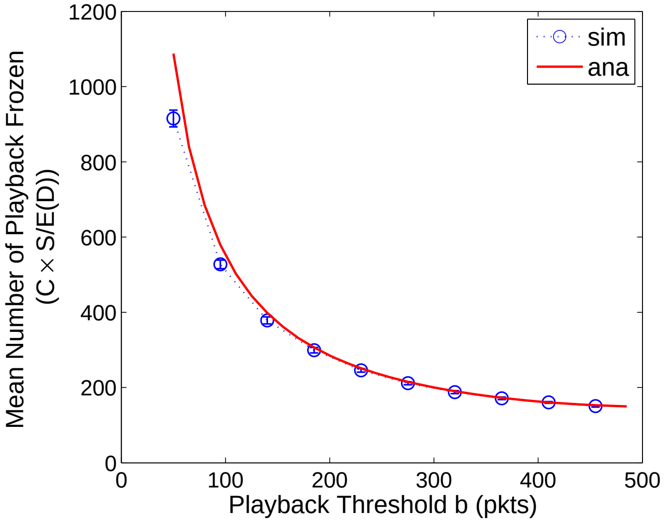


Figure 5.16: Mean number of playback frozens with the increasing playback threshold b

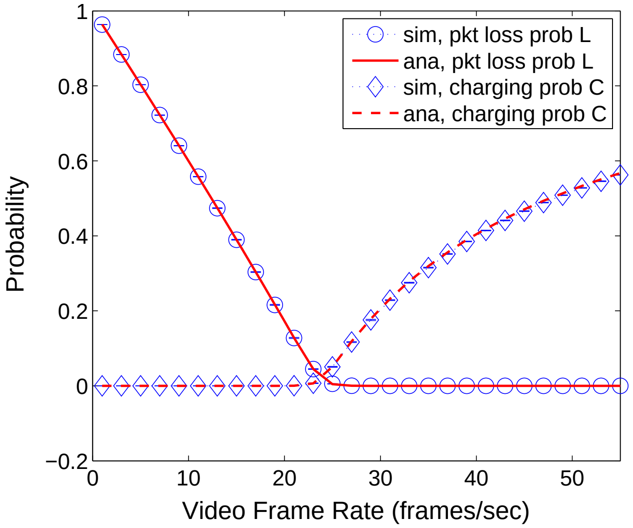


Figure 5.17: Packet loss and charging probabilities with the increasing frame rate

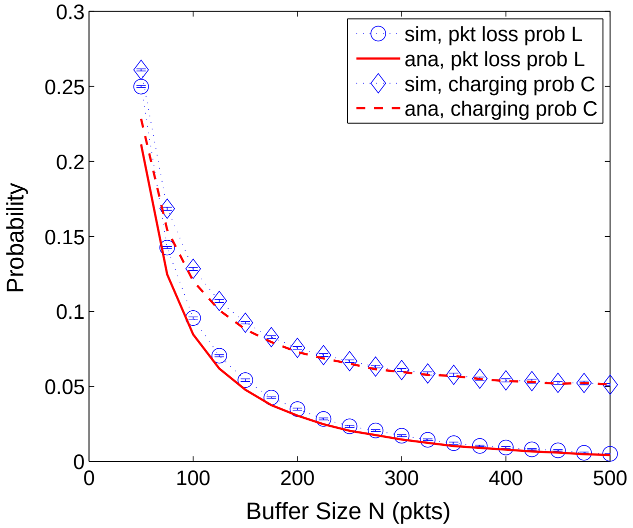


Figure 5.18: Packet loss and charging probabilities with the increasing buffer size

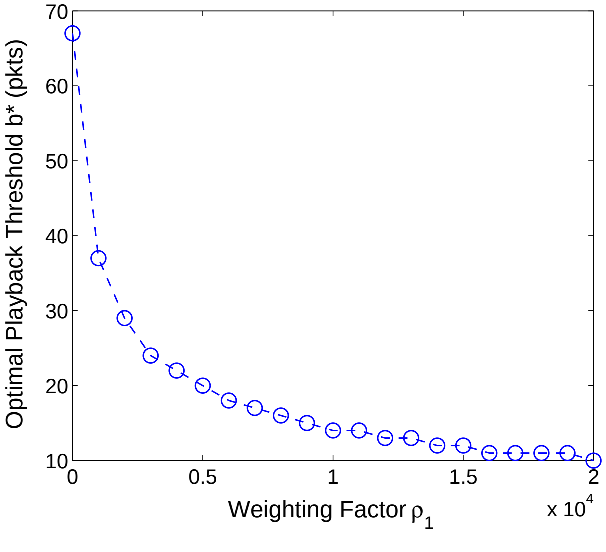


Figure 5.19: The optimal playback threshold b^* with the increasing weighting factor ρ_1

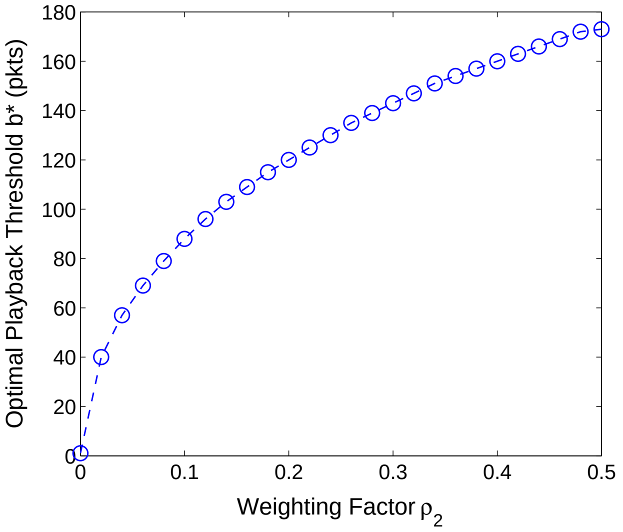


Figure 5.20: The optimal playback threshold b^* with the increasing weighting factor ρ_2

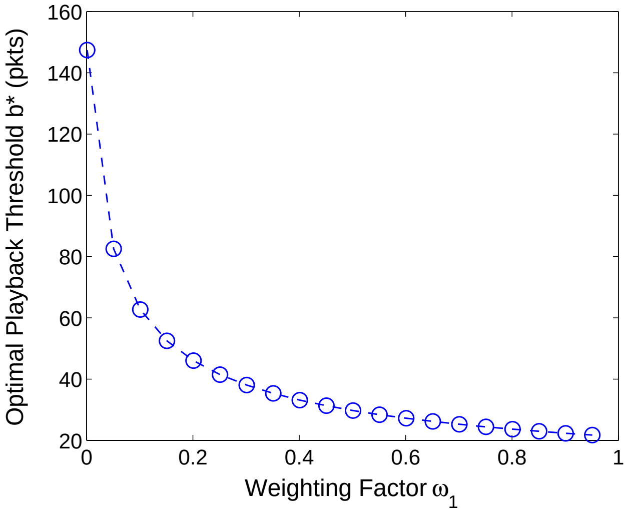


Figure 5.21: The optimal playback threshold b^* with the increasing weighting factor ϖ_1

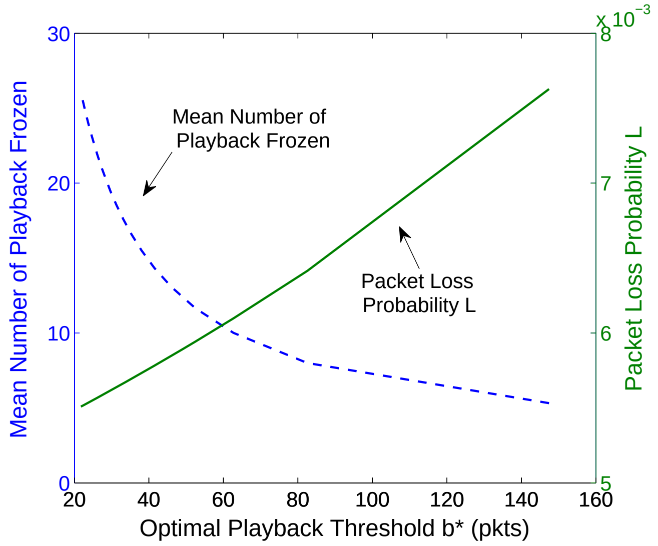


Figure 5.22: Trade-off between the mean number of playback frozens and the packet loss probability at different optimal playback thresholds b^*

optimal playback threshold increases monotonically which hence reduce the variance of interruption frequency of the playback.

5.5.3 Finite Buffer Case

Packet Loss Probability $\mathcal{L}$ and Charging Probability $\mathcal{C}$

We verify the analytical results of the packet loss probability $\mathcal{L}$ and charging probability $\mathcal{C}$ when the buffer size is finite. In each simulation, $\mathcal{L}$ is computed as the dropped packets due to buffer overflow divided by the total number of transmitted packets. $\mathcal{C}$ is evaluated as the overall time spent in the charging phase divided by the whole video session length, i.e., one hour. We set the buffer size N to be 500 packets by default and use the trace "Aladdin" in the experiment.

Figure 5.14 plots the packet loss probability $\mathcal{L}$ with increasing threshold b. With a larger b, more packets are buffered in the charging phase and therefore the buffer becomes more easily to get filled. As a result, $\mathcal{L}$ increases monotonically with b. Figure 5.15 plots the charging probability $\mathcal{C}$ under various thresholds b. It can be seen that $\mathcal{C}$ also increases monotonically with b. This is because that the charging probability $\mathcal{C}$ represents the probability that at any point the buffer is in the charging phase; with a larger b, each charging phase is elongated, making $\mathcal{C}$ increase accordingly. However, the mean number of playback frozens, computed as $\frac{\mathcal{C}\times S}{E(\mathcal{D})}$, reduces when b increases, as shown in Figure 5.16.

Figure 5.17 plots the packet loss probability $\mathcal{L}$ and charging probability $\mathcal{C}$ under different video frame rates (and playback rate). It can be seen that, with an increasing playback rate μ which implies more video frames are played in a unit time, $\mathcal{L}$ decreases and $\mathcal{C}$ increases. This is because with a faster playback, the buffer is more likely to become empty and less likely to overflow. The simulation verifies our analysis with various values of μ. Figure 5.18 shows the impacts of the playout buffer size N on the overflow and charging probabilities when playback threshold b is 50 packets. It can be seen that as the buffer size increases, both $\mathcal{L}$ and $\mathcal{C}$ decrease monotonically. This is because that with enhanced buffer capacity, less packets will be dropped due to buffer overflow and more packets are served for playback. This reduces the frequency that buffer becomes empty. However, unlike $\mathcal{L}$ which becomes 0 when N is large, $\mathcal{C}$ approaches to a non-zero value as $\lim_{N\to\infty} \mathcal{C} = -\frac{\lambda-\mu}{\mu} = 0.0536$, as derived in (H-50) and (H-51).

Optimal Selection of Playback Threshold b

After verifying the correctness of our analysis, we show how to invoke the analytical results to optimally determine the optimal playback threshold b^* using numerical examples. We solve (H-60) using the `fmincon` function of Matlab. The video trace used is “Susi & Strolch” and the default setting of the parameters are: $\widehat{\mathcal{D}} = 15\,\text{sec}, \rho_1 = 50, \rho_2 = 0.05, \varpi_1 = 0.1, \vartheta_D = 1$ and $\zeta = 0.05$.

Figure 5.19 shows the optimal playback threshold b^* with different values of ρ_1 and default setting of other parameters. It can be seen that the increasing ρ_1 leads to the decrease of b^* because a smaller playback threshold is preferred in order to avoid buffer overflow. Figure 5.20 shows the impact of ρ_2. When ρ_2 increases, the objective function is sensitive to the playback frozen and hence a larger b is desirable. Figure 5.21 plots the optimal selection of playback threshold b with the increasing ϖ_1. Similar to the case of infinite buffer, with ϖ_1 increasing, the optimal playback threshold b^* decreases monotonically to keep reducing the start-up delay. Figure 5.22 shows the tradeoff between the mean number of playback frozens computed by $\frac{\mathcal{C}\times S}{E(\mathcal{D})}$ and the packet loss probability $\mathcal{L}$.

5.6 Network Resource Allocation

In Section 5.4, we show how to adaptive control the receiver playout buffer towards the user desired video quality, with the given network throughput statistics. In this section, we discuss on the network resource in a reverse manner, i.e., given the user requirement on the video quality (startup delay and smoothness of playback) and value of b, how to allocate the network resource to provision the user desired service quality. In what follows, we first describe the mathematical framework, and then showcase the implementation of the framework in video streaming design in cognitive radio networks and vehicular networks.

5.6.1 Quality-Driven Network Resource Allocation for Heterogenous Media Flows

In the previous sections, we show how to Given the required user requirements on the media quality, we can explore the QoS mapping specified in (H-16) and (H-32) for efficient network resource allocation in wireless networks.

The network resource is generally shared by multiple users carrying multimedia traffic with diverse QoS requirements on the network resource. In general, the quality-driven network resource allocation problem can be formulated as the network utility maximization (NUM) problem,

$$\begin{array}{rl} \underset{\lambda_i, v_i}{Maximize} & \sum_i U_i\left(\mathcal{D}_i, \mathcal{P}_i\right) \\ Subject\ to: & \\ \text{Quality requirements:} & \left\{\begin{array}{ll} \Pr\left\{\mathcal{D}_i \geq \widehat{D}_i\right\} \leq \zeta_i, & \forall i \\ \mathcal{P}_i \leq \widehat{P}_i, & \forall i \end{array}\right. \\ \multicolumn{2}{c}{\text{Link and flow constraints of } \left(\lambda_i, v_i\right)} \end{array} \tag{H-63}$$

where the decision variables λ_i and v_i denote the end-to-end flow rate and network jitter of user i, respectively. The objective is to optimize the global network welfare where $U_i\left(\mathcal{D}_i, \mathcal{P}_i\right)$ denotes the utility of each user i. $U_i\left(\mathcal{D}_i, \mathcal{P}_i\right)$ could be the same as U in the buffer management problem in the previous subsection except that different users may have various weighting and risk aversion factors. As $\left(\mathcal{D}_i, \mathcal{P}_i\right)$ are dependent of $\left(\lambda_i, v_i\right)$ shown in (H-16) and (H-32), the NUM problem is to determine the optimal $\left(\lambda_i, v_i\right)$ to achieve the maximal overall utility. The first two constraints specify the required media quality of individual users. The third constraint specifies the feasible solutions of $\left(\lambda_i, v_i\right)$ constrained by the available link capacity and flow conservation (i.e., the connectivity and continuity of media flows). To solve the NUM problem, the required media performance of each user i in terms of $(\mathcal{D}_i, \mathcal{P}_i)$ is first translated into the demanded network resources in terms of (λ_i, v_i) using the QoS mapping in (H-16) and (H-32), and then the network resource is managed to meet the demanded (λ_i, v_i) so as to maximize the overall network utility.

5.6.2 Case Study

Smooth Video Delivery in Cognitive Radio Networks

The cognitive radio (CR) networks allow a group of CR users to dynamically access the idle spectrums when spectrums are not used by the li-censed users; the CR users are dictated to vacate the channels instantaneously once the licensed users are online. By doing so, the wasted spectrum can be recycled to improve the spectrum utilization. However, as CR users need to keep switching channels to avoid the possible interference to the licensed users, paired with mutual contention among CR users, the down-load of CR users tend to be turbulent and unstable, which poses significant challenges to the high-quality video streaming in CR networks.

To provide smooth video delivery to CR users in the dynamic system, we propose an adaptive channel spectrum allocation scheme in [98] based on the cross-layer framework (H-63). In specific, we consider two groups of users coexisting in the system, video users and best effort users. The former downloads the inelastic video traffic from the network, and the latter downloads elastic data traffic. Therefore, the two groups of users have distinct QoS requirements. The video users demand relatively static download rate to support the smooth video playback characterized by QoE; whereas the best effort users require lower bounded download rate to enable on-top applications. Based on the instantaneous channel status and different QoS requirements of users, we adaptively allocate the channel spectrums to users, which affects their download rates. Therefore, video users are rendered with different QoE which can be evaluated by (H-16) and (H-32). By feeding this to (H-63), the spectrum allocation is configured to maximize the integrated utility of all CR users. Figure 5.23 plots the resultant video frozen probability of video users when the proposed algorithm is applied, compared with the random and greedy channel allocations. As we can see, the proposed scheme can achieve much lower video frozen probability compared to traditional heuristics.

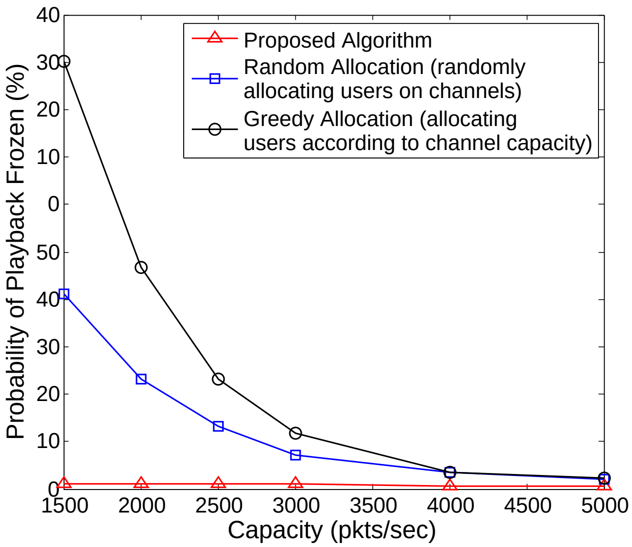


Figure 5.23: Video frozen probability with different channel allocation schemes in CR network

QoE-oriented Video Streaming in Vehicular Networks

In VANET, due to the limited coverage of infrastructure, Internet video streaming to the highly mobile vehicles typically involves both the V2I communication and the multi-hop V2V relays from the gateway to the destination vehicles. The intermittent connectivity of the video streaming path paired by the severe interference among vehicles make the smooth video streaming a very challenging task.

[99] has developed a QoE-oriented video stream routing protocol steaming from the cross-layer design framework in (H-63). Given the video playback rate and playback threshold b, the optimal packet retransmission scheme is proposed to attain the best QoE. In specific, due to the volatile wireless channel, coupled with the interference and contentions among vehicles in proximity, packet delivery to vehicles may suffer from severe packet losses. The retransmissions are used to correct the errors. This, however, prolongs the packet delivery and may lead to the underflow of playout buffer, which results in the frozen of video playback. On addressing this issue, [99] first evaluates the impact of packet retransmissions on the video download rate to each user and the resultant QoE (startup delay and probability of playback frozen) of users according to (H-16) and (H-32). The optimal retransmissions are designed based on (H-63) with the constraints subject to the tolerable QoE of users, i.e., upper bounded start-up delay and probability of frozen.

5.7 Summary

In this chapter, we have developed a mathematical framework to study the impacts of network dynamics on the perceived video quality of end users. We have evaluated the user's perceptual quality, in terms of the start-up delay, playback smoothness, and the packet loss probability, and represented them by the network statistics and the threshold of playback in closed-form expressions. We have shown how to invoke the analytical framework to guide the playout buffer design. The analytical results have been verified by extensive simulations.

We envision the future research in two dimensions. First, we will study efficient measurement schemes of the statistics of download and playback rates at the receiver and extend the analysis by taking the scalable video coding and time-dependent packet arrivals into account. Second, we will implement the proposed buffer management schemes in real-world applications and test them under different network scenarios, such as the buffer management for peer-to-peer on demand video streaming and handset buffer management for 3G cellular video streaming.

5.8 Appendix

5.8.1 Diffusion Approximation of $p_D(x,t)$

Let $0 < \tau_1 < \tau_2 < \cdots$ be the intervals between the arrival packets in the renew phase, which are i.i.d. random variables with mean and variance $\frac{1}{\lambda}$ and v_a, respectively. Let $S(n)$ represent the cumulative number of n arrivals, i.e.,

$$S(n) = \sum_{i=1}^{n} \tau_i, \tag{H-64}$$

Let $N(t)$ be the number of arrivals until time t, or equivalently, the queue length at time t. Let us consider the short period $[0, \Delta]$ with start of video downloading synchronized to time 0. Apparently,

$$\Pr\{S(n) < \Delta\} = \Pr\{N(\Delta) \geq n\}. \tag{H-65}$$

As $S(n)$ is the sum of i.i.d. random variables $\{\tau_i; i = 1, ..., n\}$, when Δ is sufficiently large, with central limit theory,

$$\Pr\{S(n) < \Delta\} \approx \Phi\left(\frac{\Delta - n\frac{1}{\lambda}}{\sqrt{nv_a}}\right), \tag{H-66}$$

where $\Phi(x)$ is the unit normal distribution with $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\frac{y^2}{2}\}dy$.

With (H-65), we have

$$\begin{aligned} \Pr\{N(\Delta) < n\} &= 1 - \Pr\{S(n) < \Delta\} \\ &\approx 1 - \Phi\left(\frac{\Delta - n\frac{1}{\lambda}}{\sqrt{nv_a}}\right) = \Phi\left(\frac{n-\lambda\Delta}{\sqrt{n\lambda^2 v_a}}\right). \end{aligned} \tag{H-67}$$

For sufficiently long period Δ, the distribution of $N(\Delta)$ is concentrated around $n \approx \lambda\Delta$. Substitute it into (H-67), yielding

$$\Pr\{N(t) < n\} \approx \Phi\left(\frac{n - \lambda t}{\sqrt{\lambda^3 t v_a}}\right). \tag{H-68}$$

Without the boundary conditions $N(t) > 0$ and $N(t) < b$ imposed, the p.d.f. of $N(t)$ is

$$p(x,t) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left\{-\frac{(x-\lambda t)^2}{2\alpha t}\right\}, \tag{H-69}$$

where $\alpha = \lambda^3 v_a$.

Apparently, $p(x,t)$ follows the diffusion equation as in (H-4). With the boundary condition $N(t) > 0$ and the absorbing state b imposed, the p.d.f. of queue length in the renew phase can hence be characterized by the diffusion equations shown in (H-10), (H-11) and (H-12).

5.8.2 Derivation of $g_D^*(s)$

The derivation of $g_D^*(s)$ is based on a standard result from [92] that

$$\mathscr{L}(\frac{k}{2\sqrt{\pi t^3}} \exp\{-\frac{k^2}{4t}\}) = \exp\{-k\sqrt{s}\}, \tag{H-70}$$

with $\mathscr{L}(f(t))$ denoting the Laplace transform of function $f(t)$. Denote by

$$f(k,t) = \frac{k}{2\sqrt{\pi t^3}} \exp\{-\frac{k^2}{4t}\}. \tag{H-71}$$

and its Laplace transform by

$$f^*(k,s) = \exp\{-k\sqrt{s}\}. \tag{H-72}$$

As

$$g_D(t) = \frac{b}{\sqrt{2\pi\alpha_D t^3}}\exp\{-\frac{(b-\beta_D t)^2}{2\alpha_D t}\} = f(\frac{b}{\sqrt{\alpha_D/2}}, t)\cdot e^{-\frac{\beta_D^2}{2\alpha_D}t}\cdot e^{\frac{\beta_D}{\alpha_D}b}, \tag{H-73}$$

by the property of Laplace transform, we hence have

$$g_D^*(s) = f^*(\frac{b}{\sqrt{\alpha_D/2}}, s+\frac{\beta_D^2}{2\alpha_D})\cdot e^{\frac{\beta_D}{\alpha_D}b} = \exp\left[\frac{b}{\alpha_D}\{\beta_D - \sqrt{\beta_D^2+2s\alpha_D}\}\right]. \tag{H-74}$$

5.8.3 Derivation of $\mathcal{C}$ and $\mathcal{L}$

With the diffusion approximation formulation as

$$\frac{1}{2}\alpha_T\frac{\partial^2 p\left(x,\infty|b\right)}{\partial x^2} - \beta_T\frac{\partial p\left(x,\infty|b\right)}{\partial x} + \frac{\lambda}{b}\mathcal{C}\delta\left(x-b\right) + \mu\mathcal{L}\delta\left(x-N+1\right) = 0, \tag{H-75}$$

$$\lim_{x\to 0}\left[\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x} - \beta_T p\left(x,\infty|b\right)\right] = \frac{\lambda}{b}\mathcal{C}, \tag{H-76}$$

$$\lim_{x\to N}\left[\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x} - \beta_T p\left(x,\infty|b\right)\right] = -\mu\mathcal{L}, \tag{H-77}$$

subject to the initial and boundary conditions

$$\begin{gathered} p\left(0,\infty|b\right) = 0, \\ p\left(N,\infty|b\right) = 0. \end{gathered}$$

By integrating (H-75), we have

$$\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x} - \beta_T p\left(x,\infty|b\right) = -\frac{\lambda}{b}\mathcal{C}\Phi\left(x-b\right) + \mu\mathcal{L}\Phi\left(x-N+1\right) + \frac{\lambda}{b}\mathcal{C}, \tag{H-78}$$

where $\Phi\left(x\right) = \begin{cases} 0, & x<0, \\ 1, & x\geq 0. \end{cases}$

$\triangleright$ When $0 < x < b$, according to (H-78), we have

$$\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x}-\beta_T p\left(x,\infty|b\right)=\frac{\lambda}{b}\mathcal{C}. \tag{H-79}$$

Solving (H-79), we have

$$p\left(x,\infty|b\right)=\frac{\lambda\mathcal{C}}{b\beta_T}\left(e^{rx}-1\right),$$

where $r=\frac{2\beta_T}{\alpha_T}$.

▷ When $b\le x\le N-1$, according to (H-78), we have

$$\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x}-\beta_T p\left(x,\infty|b\right)=0, \tag{H-80}$$

which yields

$$p\left(b,\infty|b\right)=\frac{\lambda\mathcal{C}}{b\beta_T}\left(1-e^{-rb}\right)e^{rx}. \tag{H-81}$$

▷ When $N-1\le x<N$, according to (H-78), we have

$$\frac{\alpha_T}{2}\frac{\partial p\left(x,\infty|b\right)}{\partial x}-\beta_T p\left(x,\infty|b\right)=-\mu\mathcal{L}, \tag{H-82}$$

which yields

$$p\left(x,\infty|b\right)=\frac{\mu\mathcal{L}}{\beta_T}\left(1-e^{r(x-N)}\right). \tag{H-83}$$

In summary,

$$p\left(x,\infty|b\right)=\begin{cases}\frac{\lambda\mathcal{C}}{b\beta_T}\left(e^{rx}-1\right), & 0<x\le b,\\ \frac{\lambda\mathcal{C}}{b\beta_T}\left(1-e^{-rb}\right)e^{rx}, & b<x\le N-1,\\ \frac{\mu\mathcal{L}}{\beta_T}\left(1-e^{r(x-N)}\right), & N-1\le x<N.\end{cases} \tag{H-84}$$

When $x=N-1$, with (H-81) and (H-83), we have

$$p\left(N-1,\infty|b\right)=\frac{\lambda\mathcal{C}}{b\beta_T}\left(1-e^{-rb}\right)e^{r(N-1)}=\frac{\mu\mathcal{L}}{\beta_T}\left(1-e^{-r}\right), \tag{H-85}$$

which yields

$$\mathcal{L} = \frac{\lambda \mathcal{C}}{b\mu}\left(1-e^{-rb}\right)e^{r(N-1)}\left(1-e^{-r}\right)^{-1}. \tag{H-86}$$

Since

$$\int_0^N p\left(x,\infty|b\right)dx + \mathcal{C} + \mathcal{L} = 1, \tag{H-87}$$

substituting (H-84) and (H-86) into (H-87), we thus have

$$\mathcal{C} = \left(-\frac{\mu}{\beta} + \frac{\lambda^2}{\beta b\mu}\frac{e^{r(N-1)}\left(1-e^{-rb}\right)}{1-e^{-r}}\right)^{-1} \tag{H-88}$$

$$\mathcal{L} = \left(\frac{-\left(1-e^{-r}\right)\mu^2 b}{\lambda\beta\left(1-e^{-rb}\right)e^{r(N-1)}} + \frac{\lambda}{\beta}\right)^{-1}. \tag{H-89}$$

5.8.4 Derivation of (H-58) and (H-59)

We show how to apply the Chebyshev inequality (H-57) to derive (H-58) and (H-59), respectively.

Based on (H-57), we have

$$\begin{aligned}\Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} &\leq \frac{Var(\mathcal{D})}{Var(\mathcal{D}) + \left(\widehat{\mathcal{D}} - E\left(\mathcal{D}\right)\right)^2} \\ &= \frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{\mathsf{b}}{\lambda}\right)^2}.\end{aligned} \tag{H-90}$$

To satisfy the constraint $\Pr\left\{\mathcal{D} > \widehat{\mathcal{D}}\right\} \leq \zeta$, we make

$$\frac{bv_a}{bv_a + \left(\widehat{\mathcal{D}} - \frac{\mathsf{b}}{\lambda}\right)^2} \leq \zeta, \tag{H-91}$$

which implies

$$b \le \widehat{\mathcal{D}}\lambda + \frac{v_a\left(1-\zeta\right) - \sqrt{\frac{2D\zeta}{\lambda}v_a\left(1-\zeta\right) + v_a^2\left(1-\zeta\right)^2}}{2\zeta/\lambda^2} \quad \text{or}$$
$$b \ge \widehat{\mathcal{D}}\lambda + \frac{v_a\left(1-\zeta\right) + \sqrt{\frac{2D\zeta}{\lambda}v_a\left(1-\zeta\right) + v_a^2\left(1-\zeta\right)^2}}{2\zeta/\lambda^2}. \tag{H-92}$$

As $\widehat{\mathcal{D}} \ge E\left(\mathcal{D}\right) = \frac{b}{\lambda}$, we have $b \le \widehat{\mathcal{D}}\lambda$. Together with (H-92), we have

$$b \le \widehat{\mathcal{D}}\lambda + \frac{v_a\left(1-\zeta\right) - \sqrt{\frac{2\widehat{\mathcal{D}}\zeta}{\lambda}v_a\left(1-\zeta\right) + v_a^2\left(1-\zeta\right)^2}}{2\zeta/\lambda^2}. \tag{H-93}$$

Apply the Chebyshev inequality to bound $\Pr\left\{\mathcal{F} > \widehat{\mathcal{F}}\right\}$, we have

$$\Pr\left\{\mathcal{F} > \widehat{\mathcal{F}}\right\} \le \frac{Var(\mathsf{F})}{Var(\mathcal{F}) + \left(\widehat{\mathcal{F}} - E\left(\mathcal{F}\right)\right)^2} = \frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{\mathsf{A}}{\lambda}\right)^2}. \tag{H-94}$$

To satisfy the constraint $\Pr\left\{\mathcal{F} > \widehat{\mathcal{F}}\right\} \le \eta$, we make

$$\frac{\frac{B}{b^2}}{\frac{B}{b^2} + \left(\widehat{\mathcal{F}} - \frac{\mathsf{A}}{\lambda}\right)^2} \le \eta, \tag{H-95}$$

which implies

$$b \le \frac{A}{\widehat{\mathcal{F}}} - \frac{\sqrt{B\eta\left(1-\eta\right)}}{\eta\widehat{\mathcal{F}}} \quad \text{or} \quad b \ge \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta\left(1-\eta\right)}}{\eta\widehat{\mathcal{F}}}. \tag{H-96}$$

In addition, b be set such that $\widehat{\mathcal{F}} \ge E\left(\mathcal{F}\right) = \frac{A}{b}$, i.e., $b \ge \frac{A}{\widehat{\mathcal{F}}}$. Substitute it into (H-96), we have

$$b \ge \frac{A}{\widehat{\mathcal{F}}} + \frac{\sqrt{B\eta\left(1-\eta\right)}}{\eta\widehat{\mathcal{F}}} \tag{H-97}$$

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This dissertation reports our research work on enabling the efficient and QoS (and QoE) guaranteed vehicular content distribution networks. With the distinguished large-scale and dynamic features of vehicular networks, our target is to explore the efficient, scalable and, more importantly, practical solutions. To this end, a comprehensive and systematic study which optimizes and even redesigns the communication layers and entities to address those new features and challenges is necessary. In this chapter, we summarize our research outcome from the perspectives of network scheduling, end user adaption and system architecture, respectively.

6.1.1 System Architecture Programming

In vehicular communications, RSUs behave as the data depots or injection points to provide the rich Internet contents and applications to vehicles on the road. In order to provide the ubiquitous coverage and guaranteed service to vehicles, the large-scale deployment of RSUs is necessary. This, however, is a daunting task due to the intense installation and maintenance cost. In Chater 3, our goal is to develop a practical and cost-effective solution on building the large-scale infrastructure dedicated for the vehicular communications. To this end, we design RSB which is a buffer device with the wireless transceiver to communicate with vehicles. The RSBs are installed and managed by distributed individuals in the city, such as grocery stores, movie theatres, schools, etc., on the purpose of distributing their own contents to users on the road, such as flyers, movie tailors, etc. Within this framework, we propose distributed content replication algorithm to manage distributed RSBs across the city to optimally utilize their buffer storage and self-organize as a whole to the enable QoS guaranteed content distribution to users. Using extensive simulations, we show the effectiveness of the proposed framework and algorithm.

6.1.2 Network Resource Allocation

In the presence of intense network dynamics, how to fully utilize the varying network resource towards the global welfare of the system is a fundamental issue of the vehicular networking. In order to conduct effective network control and resource allocation, the accurate evaluation of system performance which fully considers the unique features of the vehicular network is foremost. This motivates our research in Chapter 4 on the evaluation of the fundamental DCF MAC in the drive-thru Internet scenario (or V2R communication). In Chapter 4, we first establish a three-dimension Markov chain model which captures the vehicle mobility in the modeling of DCF operations. Using the developed model, the resultant nodal and system throughput can be derived with different vehicular velocities and configurations of the DCF MAC. In this manner, the developed model can provide insightful guidance for the real-world deployment of the drive-thru Internet systems. Based on the developed model, we then propose multiple enhancement mechanisms of DCF to further improve the MAC performance of the highly dynamic vehicular networks. Lastly, using extensive simulations, we verify the accuracy of the proposed analytical model and validate the performance of the proposed enhancement mechanisms.

6.1.3 End User Adaption

The basic goal of any communication systems is to provide the users with their *desired* service quality of applications. Therefore, to investigate on the specific user requirements and provide the corresponding service guarantee is the fundamental issue in the communication system design. This, however, is a very challenging issue due to the gap between QoE and QoS. The QoE refers to the general satisfaction of users on the service which is often subjective and obscure; it may change with different users and is typically affected by the a variety of personal and environmental factors such as modes, knowledge of users, darkness of the theater, hardware devices, etc. The QoS is defined from the network perspective in terms of data throughput, transmission delay and packet loss probability, etc. In order to adopt effective network resource allocation to achieve the guaranteed QoE at the end user, the foremost issue is to bridge QoS and QoE with an effective mathematical translation. In Chapter 5 of the dissertation, we focus on the real-time media streaming applications and establish an analytical model to represent the QoE metrics (in startup delay and smoothness of playback) by the QoS metrics (in mean and variance of download throughput). Our model applies the queueing theory to analyze the receiver playout buffer from the user's perspective. Using the statistics of the network download rate as the input to the model, we are able to derive the distribution of the startup delay and the likelihood of smooth playback throughout the video playout session. Moreover, based on the model, we are able optimally control the threshold of media playback and accordingly adapt the video playback strategy at the receiver based on the user's preferences given the available network bandwidth.

6.2 Future Work

Towards the efficient, low-cost and ubiquitous content distribution to vehicles, there are still many open issues remaining to be solved. Next we outline several interesting directions for future work.

6.2.1 Cooperative Communication of Heterogenous Network

It has already been a common practice that many wireless access technologies, such as WiFi and 3G/4G cellular networks, now coexist in the city. As the vehicular networking is still in its infancy stage, to utilize and cooperatively work with the existing mature access techniques is important not only to speed up the deployment of VANETs and but also to improve the service quality that can be provided to VANET users. However, different access technologies typically have diverse features in terms of the communication cost, availability and coverage. For example, the cellular networks can provide ubiquitous coverage but are at some monetary cost on bandwidth. The WiFi hotspots can provide the high-rate connections to the public but are managed and operated by private owners; they are only open to others at the prerequisite of not harming to the service quality of their own users. The RSUs are dedicated and exclusively used for vehicular communications, but may be sparsely deployed at the early stage of VANETs. As vehicles have a very large mobility region spanning several tens or hundreds of miles, to explore the random access opportunities of different access networks along the moving trajectory is a practical and important solution to enhance of performance of VANETs. For example, according to the deadline of transmissions, the data can be transmitted through different access networks.

In this case, the content with an urgent transmission deadline can be delivered immediately through the "always-on" cellular connections, whereas the contents without strict deadline can be delivered through the opportunistic connections to WiFi networks or RSUs. However, as the different access networks impose different cost on the communications and are various in availabilities and data rates, an efficient design at vehicles is necessary to optimally determine the communication strategy based on the specific QoS of users and mobility features of vehicles to maximize the service quality and meanwhile minimize the communication cost.

6.2.2 Vehicular Social Network Design

With the limited access to the infrastructure and Internet accordingly, the vehicular networks face the shortage of contents. For example, on the highway vehicular networks, the RSUs may be sparsely deployed whereas other networks, such as WiFi and cellular networks are not sufficient as a complementary. Although the V2V communications can be used to enhance the communication bandwidth for content delivery, without sufficient RSUs as the data depots and the high-rate Internet connections, Internet contents cannot be retrieved, and VANETs can hardly be attractive to users as there simply are no desirable contents to drive the download. In this case, we can rely on the social networks to encourage users to spontaneously create contents and distribute to the crowd in proximity. Specifically, the social networks, such as Facebook and Twitter, make them successful by enabling individuals to contribute the knowledge and content information to the community. When being transplanted to the vehicular networking, it can be used to explore the localized content information, such as traffic and weather conditions, local activities and news, among vehicular users on the road and distribute to each other using V2V connections only without the involvement of Internet and RSUs. However, the infrastructure-less vehicular social network may face the challenges imposed by the fast motion and transient connectivity of vehicles. Meanwhile, how to enable users to quickly identify the interested information and social friends in the network is also challenging.

6.2.3 Incentive Mechanism Design and Content Pollution Issue

Without sufficient infrastructure support, especially in the early stage, VANETs typically rely on distributed vehicles to contribute their buffer, power and bandwidth resources to cache and relay contents for each other. However, users tend to be selfish and are reluctant to help if doing so will not incur matching rewards to them. This may finally result in the tragedy of the commons, and therefore the incentive mechanism is necessary to encourage contributions and punish the free-riding and selfish behaviors. However, as VANETs are typically large-scale and distributed in nature with the high and wide-range node mobility, any centralized control is not practical which calls for a fully distributed and efficient design.

In addition to the free-riding issue, the content pollution issues can also be severe in vehicular networks. Specifically, the content pollution refers to the fake duplicates of contents which have the similar size and meta data, such as title, authors, as the original copy of contents, but are useless to users or even harmful with virus. The content pollution issue is severe in the peer-to-peer networks [100] which is the counterpart of the mobile ad hoc network in general and VANETs in particular in the wired network. As in vehicular networks, due to the anonymity of users and lack of the central controller, the content pollution may cause severe cost to users with the poor service quality or even hardware damage. Therefore, the design of incentive mechanisms should not only forbid the selfish behavior of users, but also need to timely identify and punish the malicious nodes which spread the polluted contents or conduct other security attacks.

References

[1] J. P. Hubaux, S. Capkun, and J. Luo, “The Security and Privacy of Smart Vehicles,” *IEEE Security & Privacy*, vol. 2, no. 3, pp. 49–55, May 2004.

[2] T. . Globe and M. Update, “Canadians’ Internet Use Exceeds TV Time,” Mar. 2010.

[3] P. Bouvard, L. Rosin, J. Snyde, and J. Noel, “The Arbition National In-Car Study,” *Arbitron/Edison Media Research*, Dec. 2003.

[4] J. Angel, “Mercedes-Benz Demos Wireless Network,” 2001. [Online]. Available: http://www.allbusiness.com/marketing-advertising/4446481-1.html

[5] G. Motors, “On Star.” [Online]. Available: http://www.onstar.com

[6] G. Wong, “Toyota to Launch Toyota Friend: a Car-Owner Social Network,” May 2011. [Online]. Available: http://www.ubergizmo.com/2011/05/toyota-friend-car-social-network

[7] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan, and S. Madden, “A Measurement Study of Vehicular Internet Access Using In Situ Wi-Fi Networks,” in *Proc. of ACM MobiCom*, 2006.

[8] F. Bai and B. Krishnamachari, “Spatio-Temporal Variations of Vehicle Traffic in VANETs: Facts and Implications,” in *Proc. of ACM VANET*, 2009.

[9] Y. L. Morgan, “Notes on DSRC & WAVE Standards Suite: Its Architecture, Design, and Characteristics,” *IEEE Communications Surveys & Tutorials*, vol. 12, no. 4, pp. 504–518, Fourth Quarter 2010.

[10] J. Ott and D. Kutscher, “Drive-Thru Internet: IEEE 802.11b for ”Automobile” Users,” in *Proc. of IEEE Infocom*, 2004.

[11] A. Nandan, S. Das, G. Pau, M. Gerla, and M. Y. Sanadidi, "Co-Operative Downloading in Vehicular Ad-Hoc Wireless Networks," in *Proc. of IEEE/IFIP WONS*, 2005.

[12] O. Trullols-Cruces, M. Fiore, and J. Barcelo-Ordinas, "Cooperative Download in Vehicular Environments," *IEEE Transactions on Mobile Computing*, vol. 11, no. 4, pp. 663 – 678, Apr. 2011.

[13] D. Niyato and P. Wang, "Optimization of the Mobile Router and Traffic Sources in Vehicular Delay-Tolerant Network," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 5095–5104, Nov. 2009.

[14] U. Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi, "Dissemination and Harvesting of Urban Data Using Vehicular Sensing Platforms," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 882–901, Feb. 2009.

[15] Z. Yang, M. Li, and W. Lou, "CodePlay: Live Multimedia Streaming in VANETs using Symbol-Level Network Coding," in *Proc. of IEEE ICNP*, 2010.

[16] M. Li, Z. Yang, and W. Lou, "Codeon: Cooperative Popular Content Distribution for Vehicular Networks Using Symbol Level Network Coding," in *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, Jan. 2010, pp. 223 –235.

[17] F. Ye, S. Roy, and H. Wang, "Efficient Data Dissemination in Vehicular Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 769 –779, May 2012.

[18] Q. Yan, M. Li, Z. Yang, W. Lou, and H. Zhai, "Throughput Analysis of Cooperative Mobile Content Distribution in Vehicular Network Using Symbol Level Network Coding," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 484–492, 2012.

[19] Y. Zhang, J. Zhao, and G. Cao, "Roadcast: a Popularity Aware Content Sharing Scheme in VANETs," in *Proc. of IEEE ICDCS*, 2009.

[20] E. Cohen and S. Shenker, "Replication Strategies in Unstructured Peer-to-Peer Networks," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 177–190, Oct. 2002.

[21] Y. Zhang and G. Cao, "V-PADA: Vehicle-Platoon Aware Data Access in VANETs," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2326 – 2339, Jun. 2011.

[22] N. Wisitpongphan and F. Bai and P. Mudalige and V. Sadekar and O. Tonguz, "Routing in Sparse Vehicular Ad Hoc Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 8, pp. 1538–1556, Oct. 2007.

[23] W. H. Yuen, R. D. Yates, and S. C. Mau, "Exploiting Data Diversity and Multiuser Diversity in Noncooperative Mobile Infostation Networks," in *Proc. of IEEE Infocom*, 2003.

[24] R. H. Frenkiel and T. Imielinski, "Infostations: The Joy of Many-Time Many-Where Communications," *Journal on Mobile Computing*, 1996.

[25] Y. Zhang, J. Zhao, and G. Cao, "On Scheduling Vehicle-Roadside Data Access," in *Proc. of ACM VANET*, 2007.

[26] A. Nandan, S. Das, B. Zhou, G. Pau, and M. Gerla, "AdTorrent: Digital Billboards for Vehicular Networks," in *Proc. of IEEE/ACM V2VCOM*, 2005.

[27] Y. Huang, Y. Gao, K. Nahrstedt, and W. He, "Optimizing File Retrieval in Delay-Tolerant Content Distribution Community," in *Proc. of IEEE ICDCS*, 2009.

[28] J. Wortham, "Customers Angered as iPhones Overload AT&T," The New York Times, Sept. 2009.

[29] "Devicescape Wi-Fi Report: Original Research on Wi-Fi Usage and Trends," 2009. [Online]. Available: http://www.devicescape.com/pdf/reports/ds-wfr-1Q-09$_$final.pdf

[30] M. Reardon, "Cities Deploying Wi-Fi Face Challenges," CNET News, May 2006.

[31] N. Post, "Toronto Hydro assailed for city-wide WiFi plan," Canada.com, March 2006.

[32] H. T. Cheng, H. Shan, and W. Zhuang, "Infotainment and Road Safety Service Support in Vehicular Networking: From a Communication Perspective," *Mechanical Systems and Signal Processing, Special Issue on Integrated Vehicle Dynamics*, vol. 25, no. 6, pp. 2020C–2038, Aug. 2011.

[33] U. Lee, J. S. Park, J. Yeh, G. Pau, and M. Gerla, "Code Torrent: Content Distribution Using Network Coding in VANET," in *Proc. of ACM Mobishare*, 2006.

[34] S. H. Lee, U. Lee, K. W. Lee, and M. Gerla, "Content Distribution in VANETs using Network Coding: the Effect of Disk I/O and Processing O/H," in *Proc. of IEEE SECON*, 2008.

[35] U. G. Acer, P. Giaccone, D. Hay, G. Neglia, and S. Tarapiah, "Timely Data Delivery in a Realistic Bus Network," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 3, pp. 1251 –1265, Mar. 2012.

[36] A. Nandan, S. Tewari, S. Das, M. Gerla, and L. Kleinrock, "Adtorrent: Delivering Location Cognizant Advertisements to Car Networks," in *Proc. of IEEE/IFIP WONS*, 2006.

[37] B. Yu and F. Bai, "ETP: Encounter Transfer Protocol for Opportunistic Vehicle Communication," in *Proc. of IEEE Infocom*, 2011.

[38] F. Chung and L. Lu, *Complex Graphs and Networks.* Amer Mathematical Society, 2006.

[39] J. Haerri, F. Filali, C. Bonnet, and M. Fiore, "VanetMobiSim: Generating Realistic Mobility Patterns for VANETs," in *Proc. of ACM VANET*, 2006.

[40] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, "Vehicular Opportunistic Communication Under the Microscope," in *Proc. of ACM MobiSys*, 2007.

[41] P. Bucciol, E. Masala, N. Kawaguchi, K. Takeda, and J. C. D. Martin, "Performance Evaluation of H. 264 Video Streaming Over Inter-Vehicular 802.11 Ad Hoc Networks," in *Proc. of IEEE PIMRC*, 2005.

[42] G. Bianchi, "Performance Analysis of the IEEE 802. 11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[43] L. X. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, "Voice Capacity Analysis of WLAN with Unbalanced Traffic," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 752–761, May 2006.

[44] T. H. Luan, X. Ling, and X. Shen, "MAC in Motion: Impact of Mobility on the MAC of Drive-Thru Internet," *IEEE Transactions on Mobile Computing*, vol. 11, no. 2, pp. 305 – 319, 2011.

[45] J. Zhao, T. Arnold, Y. Zhang, and G. Cao, "Extending Drive-Thru Data Access by Vehicle-to-Vehicle Relay," in *Proc. of IEEE VANET*, 2008.

[46] J. Zhang, Q. Zhang, and W. Jia, "VC-MAC: A Cooperative MAC Protocol in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1561–1571, Mar. 2009.

[47] B. Sikdar, "Characterization and Abatement of the Reassociation Overhead in Vehicle to Roadside Networks," *IEEE Transactions on Communications*, vol. 58, no. 11, pp. 3296–3304, Nov. 2010.

[48] X. Zhang, J. Kurose, B. N. Levine, D. Towsley, and H. Zhang, "Study of a Bus-Based Disruption-Tolerant Network: Mobility Modeling and Impact on Routing," in *Proc. of ACM MobiCom*, 2007.

[49] J. Ott and D. Kutscher, "A Disconnection-Tolerant Transport for Drive-Thru Internet Environments," in *Proc. of IEEE Infocom*, 2005.

[50] S. Pack, H. Rutagemwa, X. Shen, J. W. Mark, and K. Park, "Proxy-Based Wireless Data Access Algorithms in Mobile Hotspots," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 5, pp. 3165–3177, Sept. 2008.

[51] A. Festag, H. Fußler, H. Hartenstein, A. Sarma, and R. Schmitz, "FLEETNET: Bringing Car-to-Car Communication into the Real World," *Computer*, vol. 4, no. L15, p. 16, 2004.

[52] W. L. Tan, W. C. Lau, O. Yue, and T. H. Hui, "Analytical Models and Performance Evaluation of Drive-thru Internet Systems," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 207–222, Jan. 2010.

[53] "IEEE P802.11p/D5.0, Draft Amendment to Standard for Information Technology Telecommunications and Information Exchange Between Systems LAN/MAN Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Wireless Access in Vehicular Environments (WAVE)," 2008.

[54] "Institute of Electrical and Electronics Engineers, Standard 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," 2007. [Online]. Available: http://standards.ieee.org/getieee802/download/802.11-2007.pdf

[55] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance Anomaly of 802.11b," in *Proc. of IEEE Infocom*, 2003.

[56] D.-Y. Yang, T.-J. Lee, K. Jang, J.-B. Chang, and S. Choi, "Performance Enhancement of Multirate IEEE 802.11 WLANs with Geographically Scattered Stations," *IEEE Transactions on Mobile Computing*, vol. 5, no. 7, pp. 906–919, Jul. 2006.

[57] A. V. Babu and L. Jacob, "Fairness Analysis of IEEE 802.11 Multirate Wireless LANs," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 3073–3088, Sept. 2007.

[58] T. Joshi, A. Mukherjee, Y. Yoo, and D. P. Agrawal, "Airtime Fairness for IEEE 802.11 Multirate Networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 4, pp. 513–527, Apr. 2008.

[59] D. Hadaller, S. Keshav, and T. Brecht, "MV-MAX: Improving Wireless Infrastructure Access for Multi-Vehicular Communication," in *Proc. of ACM CHANTS*, 2006.

[60] F. Calì, M. Conti, and E. Gregori, "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," *IEEE/ACM Transactions on Networking*, vol. 8, no. 6, pp. 785–799, Dec. 2000.

[61] L. Cheng, B. E. Henty, D. D. Stancil, F. Bai, and P. Mudalige, "Mobile Vehicle-to-Vehicle Narrow-Band Channel Measurement and Characterization of the 5.9 GHz Dedicated Short Range Communication (DSRC) Frequency Band," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 8, pp. 1501–1516, Oct. 2007.

[62] K.-H. Liu, X. Shen, R. Zhang, and L. Cai, "Performance Analysis of Distributed Reservation Protocol for UWB-Based WPAN," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 902–913, Feb. 2009.

[63] S. Karamardian and C. B. García, *Fixed Points: Algorithms and Applications*. Academic press, 1977.

[64] M. A. Chowdhury and A. W. Sadek, *Fundamentals of Intelligent Transportation Systems Planning*. Artech House Publishers, 2003.

[65] G. Anastasi, E. Borgia, M. Conti, and E. Gregori, "Wi-Fi in Ad Hoc Mode: a Measurement Study," in *Proc. of IEEE PerCom*, 2004.

[66] B. Yu and C.-Z. Xu, "Admission Control for Roadside Unit Access in Intelligent Transportation Systems," in *Proc. of IEEE IWQoS*, 2009.

[67] P. Shankar, T. Nadeem, J. Rosca, and L. Iftode, "CARS: Context-Aware Rate Selection for Vehicular Networks," in *Proc. of IEEE ICNP*, 2008.

[68] B. Girod, J. Chakareski, M. Kalman, Y. J. Liang, E. Setton, and R. Zhang, "Advances in Network-adaptive Video Streaming," *Wireless Communications and Mobile Computing*, vol. 2, no. 6, pp. 549–552, 2002.

[69] Q. Zhang, W. Zhu, and Y.-Q. Zhang, “End-to-End QoS for Video Delivery Over Wireless Internet,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 123–134, Jan. 2005.

[70] P. A. Chou and Z. Miao, “Rate-Distortion Optimized Streaming of Packetized Media,” *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 390–404, Apr. 2006.

[71] J. Chakareski and P. Frossard, “Rate-Distortion Optimized Distributed Packet Scheduling of Multiple Video Streams Over Shared Communication Resources,” *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 207–218, Apr. 2006.

[72] S. Mao, Y. T. Hou, X. Cheng, H. D. Sherali, S. F. Midkiff, and Y.-Q. Zhang, “On Routing for Multiple Description Video Over Wireless Ad Hoc Networks,” *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 1063–1074, Oct. 2006.

[73] X. Tong, Y. Andreopoulos, and M. van der Schaar, “Distortion-Driven Video Streaming Over Multihop Wireless Networks With Path Diversity,” *IEEE Transactions on Mobile Computing*, vol. 6, no. 12, pp. 1343–1356, Dec. 2007.

[74] J. Xu, X. Shen, J. W. Mark, and J. Cai, “Adaptive Transmission of Multi-Layered Video Over Wireless Fading Channels,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 6, p. 2305, Jun. 2007.

[75] R. A. Berry and R. G. Gallager, “Communication Over Fading Channels with Delay Constraints,” *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[76] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, “Joint Source Coding and Transmission Power Management for Energyefficient Wireless Video Communications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 411–424, Jun. 2002.

[77] N. Laoutaris and I. Stavrakakis, “Intrastream Synchronization for Continuous Media Streams: A Survey of Playout Schedulers,” *IEEE Network*, vol. 16, no. 3, pp. 30–40, May 2002.

[78] J. Liu, B. Li, and Y.-Q. Zhang, “An End-to-End Adaptation Protocol for Layered Video Multicast Using Optimal Rate Allocation,” *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 87–102, Feb. 2004.

[79] T. H. Luan, L. X. Cai, and X. Shen, “Impact of Network Dynamics on User’s Video Quality: Analytical Framework and QoS Provision,” *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 64–78, Jan. 2010.

[80] L. Galluccio, G. Morabito, and G. Schembra, "Transmission of Adaptive MPEG Video Over Time-Varying Wireless Channels: Modeling and Performance Evaluation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2777–2788, Nov. 2005.

[81] M. Kalman, E. Steinbach, and B. Girod, "Adaptive Media Playout for Low-Delay Video Streaming Over Error-prone Channels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 841–851, Jun. 2004.

[82] N. Laoutaris, B. V. Houdt, and I. Stavrakakis, "Optimization of a Packet Video Receiver Under Different Levels of Delay Jitter: An Analytical Approach," *Performance Evaluation*, vol. 55, no. 3-4, pp. 251–275, Feb. 2004.

[83] G. Liang and B. Liang, "Balancing Interruption Frequency and Buffering Penalties in VBR Video Streaming," in *Proc. of IEEE Infocom*, 2007.

[84] A. Dua and N. Bambos, "Buffer Management for Wireless Media Streaming," in *Proc. of IEEE Globecom*, 2007.

[85] G. Liang and B. Liang, "Effect of Delay and Buffering on Jitter-Free Streaming Over Random VBR Channels," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1128 – 1141, Oct. 2008.

[86] D. Wu, Y. T. Hou, W. Zhu, Y. Q. Zhang, and J. M. Peha, "Streaming Video Over the Internet: Approaches and Directions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 282–300, Mar. 2001.

[87] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications.* John Wiley & Sons, 1976.

[88] G. Louchard and G. Latouche, *Probability Theory and Computer Science.* Academic Press Professional, Inc. San Diego, CA, USA, 1983.

[89] A. Duda, "Transient Diffusion Approximation for Some Queueing Systems," in *Proc. of ACM Sigmetrics*, 1983.

[90] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes.* Chapman & Hall/CRC, 1977.

[91] F. P. T. Czachorski, "Diffusion Approximation as a Modelling Tool in Congestion Control and Performance Evaluation," in *Proc. of HET-NET*, 2004.

[92] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions.* New York: Dover, 1965.

[93] E. Gelenbe, "On Approximate Computer System Models," *Journal of the ACM (JACM)*, vol. 22, no. 2, pp. 261–269, Apr. 1975.

[94] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Scalable Video Coding and Transport Over Broadband Wireless Networks," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 6–20, Jan. 2001.

[95] W. Kumwilaisak, Y. T. Hou, Q. Zhang, W. Zhu, C.-C. J. Kuo, and Y.-Q. Zhang, "A Cross-layer Quality-of-Service Mapping Architecture for Video Delivery in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1685–1698, Dec. 2003.

[96] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming.* Springer, 1997.

[97] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H. 263 Video Traces for Network Performance Evaluation," *IEEE Network*, vol. 15, no. 6, pp. 40–54, Nov. 2001.

[98] S. Li, T. H. Luan, and X. Shen, "Channel Allocation for Smooth Video Delivery over Cognitive Radio Networks," in *Proc. of IEEE Globecom*, 2010.

[99] M. Asefi, J. W. Mark, and X. Shen, "A Mobility-Aware and Quality-Driven Retransmission Limit Adaptation Scheme for Video Streaming over VANETs," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1817 –1827, May 2012.

[100] N. Christin, A. S. Weigend, and J. Chuang, "Content Availability, Pollution and Poisoning in File Sharing Peer-to-Peer Networks," in *Proc. of ACM EC*, 2005.

[101] N. Laoutaris and I. Stavrakakis, "An Analytical Design of Optimal Playout Schedulers for Packet Video Receivers," *Computer Communications*, vol. 26, no. 4, pp. 294–303, 2003.

[102] P. Rodriguez, I. Pratt, J. Chesterfield, R. Chakravorty, and S. Banjeree, "Mar: A Commuter Router Infrastructure for the Mobile Internet," in *Proc. of ACM MobiSys*, 2004.

[103] J. Chakareski, J. G. Apostolopoulos, S. Wee, W. Tan, , and B. Girod, "Rate-Distortion Hint Tracks for Adaptive Video Streaming," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1257–1269, Oct. 2005.

[104] Y. Sun, I. Sheriff, E. M. Belding-Royer, and K. C. Almeroth, "An Experimental Study of Multimedia Traffic Performance in Mesh Networks," in *Proc. of ACM WiTMeMo*, 2005.

[105] T. View, "Cross-Layer Wireless Multimedia Transmission: Challenges, Principles, and New Paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, Aug. 2005.

[106] R. Gass, J. Scott, and C. Diot, "Measurements of In-Motion 802.11 Networking," in *Proc. of IEEE WMCSA*, 2006.

[107] R. Mahajan, J. Zahorjan, and B. Zill, "Understanding WiFi-Based Connectivity From Moving Vehicles," in *Proc. of ACM IMC*, 2007.

[108] S. Pack, X. S. Shen, J. W. Mark, and J. Pan, "Mobility Management in Mobile Hotspots with Heterogeneous Multihop Wireless Links," *IEEE Communications Magazine*, vol. 45, no. 9, pp. 106–112, Sept. 2007.

[109] V. Lenders, G. Karlsson, and M. May, "Wireless Ad Hoc Podcasting," in *Proc. of IEEE SECON*, 2007.

[110] J. Zhao, Y. Zhang, and G. Cao, "Data Pouring and Buffering on the Road: A New Data Dissemination Paradigm for Vehicular Ad Hoc Networks," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 6, pp. 3266 – 3277, Nov. 2007.

[111] S. Phillips, "Financial Times: The Future Dashboard," 2001. [Online]. Available: http://specials.ft.com/ftit/june2001/FT3A72I0JNC.html

[112] J. Huang, Z. Li, M. Chiang, and A. K. Katsaggelos, "Joint Source Adaptation and Resource Allocation for Multi-User Wireless Video Streaming," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 582–595, May 2008.

[113] E. Karamad and F. Ashtiani, "A Modified 802.11-Based MAC Scheme to Assure Fair Access for Vehicle-to-Roadside Communications," *Computer Communications*, vol. 31, no. 12, pp. 2898–2906, Jul. 2008.

[114] X. Ling, Y. Cheng, X. Shen, and J. W. Mark, "Voice Capacity Analysis of WLANS with Channel Access Prioritizing Mechanisms," *IEEE Communications Magazine*, vol. 46, no. 1, p. 82, Jan. 2008.

[115] B. Sikdar, "A MAC Protocol for Vehicle to Roadside Networks," in *Proc. of IEEE ICC*, 2008.

[116] C. C. Vassilakis, N. Laoutaris, and I. Stavrakakis, "The Impact of Playout Policy on the Performance of P2P Live Streaming," in *Proc. of SPIE/ACM MMCN*, 2008.

[117] B. B. Chen and M. C. Chan, "MobTorrent: A framework for Mobile Internet Access From Vehicles," in *Proc. of IEEE Infocom*, 2009.

[118] Y. Zhang, J. Zhao, and G. Cao, "Service Scheduling of Vehicle-Roadside Data Access," *Mobile Networks and Applications*, vol. 15, no. 1, pp. 83 – 96, Feb. 2010.

[119] I. Leontiadis, P. Costa, and C. Mascolo, "Persistent content-based information dissemination in hybrid vehicular networks," in *Proc. of IEEE PerCom*. IEEE, 2009.

[120] P. Deshpande, A. Kashyap, C. Sung, and S. R. Das, "Predictive methods for improved vehicular WiFi access," in *Proc. of ACM Mobisys*, 2009.

[121] L. Hu, J. Y. L. Boudec, and M. Vojnoviae, "Optimal Channel Choice for Collaborative Ad-Hoc Dissemination," in *Proc. of IEEE Infocom*, 2010.

[122] J. Ahn, B. Krishnamachari, F. Bai, and L. Zhang, "Optimizing Content Dissemination in Heterogeneous Vehicular Networks."

[123] S. Ioannidis, L. Massoulie, and A. Chaintreau, "Distributed Caching Over Heterogeneous Mobile Networks," in *Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, 2010, pp. 311–322.

[124] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. M. Ni, "Impact of Traffic Influxes: Revealing Exponential Inter-Contact Time in Urban VANETs," *IEEE Transactions on Distributed and Parallel Systems*, vol. 22, no. 8, pp. 1258 – 1266, Aug. 2011.

[125] T. H. Luan, L. X. Cai, J. Chen, X. Shen, and F. Bai, "VTube: Towards the Media Rich City Life with Autonomous Vehicular Content Distribution," in *Proc. of IEEE SECON*, 2010.

[126] N. Lu, T. H. Luan, W. Wang, X. Shen, and F. Bai, "Capacity and Delay Analysis for Social-Proximity Urban Vehicular Networks," in *Proc. of IEEE Infocom*, 2012.