Channel Allocation for Smooth Video Delivery over Cognitive Radio Networks

by

Sanying Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2010

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

Video applications normally demand stringent quality-of-service (QoS) for the high quality and smooth video playback at the receiver. Since the network is usually shared by multiple applications with diverse QoS requirements, QoS provisioning is an important and difficult task for the efficient and smooth video delivery. In the context of cognitive radio (CR) networks, as the secondary or unlicensed users share a pool of bandwidth that is temporarily being unused by the primary or licensed users, there is an inevitable interference between the licensed primary users and the unlicensed CR devices. As a result, efficient and smooth video delivery becomes even more challenging as the channel spectrum is not only a precious resource, but also much more dynamic and intermittently available to secondary users.

In this thesis, we focus on the provision of guaranteed QoS to video streaming subscribers in CR network. In video streaming applications, a playout buffer is typically deployed at the receiver to deal with the impact of the network dynamics. With different buffer storage, users can have different tolerance to the network dynamics. We exploit this feature for channel allocation in CR network. To this end, we model the channel availability as an on-off process which is stochastically known. Based on the bandwidth capacity and the specific buffer storage of users, we intelligently allocate the channels to maximize the overall network throughput while providing users with the smooth video playback, which is formulated as an optimization framework. Given the channel conditions and the video packet storage in the playout buffer, we propose a centralized scheme for provisioning the superior video service to users. Simulation results demonstrate that by exploiting the playout buffer of users, the proposed channel allocation scheme is robust against intense network dynamics and provides users with the elongated smooth video playback.

Acknowledgments

First and foremost, I would like to express my deepest and sincerest gratitude and appreciation to my supervisor Professor Xuemin (Sherman) Shen for his continuous guidance, encouragement, support and for giving me the opportunity to do this work during my study at the University of Waterloo. Without his inspiration, deep insight, and invaluable advice, this dissertation would not have been possible. More importantly, he is not only an advisor, but also a role model who guided me to be a good person, which are of great benefit to me forever.

I would also like to thank Professor Liang-Liang Xie and Professor En-Hui Yang as my thesis readers for contributing their valuable time and effort in perfecting my work.

Special thanks to Mr. Tom H. Luan for his enlightening and patient guidance during this research. He provided me not only the knowledge but also the way to do good research.

I am also grateful to Dr. Lin X. Cai, Dr. Hangguan Shan, Dr. Ho Ting Cheng, Prof. Jiming Chen, Mr. Hao Liang, Mr. Xiaohui Liang, Mr. Rongxing Lu, Mr. Yongkang Liu, Mr. Wei Cai, Mr. Jian Qiao, Ms. Xiaoxia Zhang, and other friends and colleagues in the Broadband and Communication Research (BBCR) group for their warm friendships and invaluable comments and advice. It is my great honor to be a member of BBCR group at the University of Waterloo.

Many thanks to the administrative staff, Ms. Wendy Boles, Ms. Lisa Hendel, and Ms. Karen Schooley for their kindness, patience, and help.

Finally, my deepest gratitude and love belong to my parents for their support, encouragement, and endless love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Cognitive Radio

The current wireless networks follow a static spectrum assignment strategy where radio spectrum are allocated to license holders for exclusive usage privileges on a long-term basis for large geographical regions. With the increase in spectrum demand, this strategy has been challenged by the problem of spectrum scarcity, which makes the radio spectrum one of the most heavily regulated and expensive natural resources around the world. In Europe, the 3G spectrum auction yielded 35 billion dollars in England and 46 billion in Germany. The question is, however, whether spectrum is really so scarce. Although almost all spectrum suitable for wireless communications has been allocated, extensive Federal Communications Commission (FCC) measurements indicate that much of the radio spectrum is not in use for a significant amount of time, and at a large number of locations [1]. For instance, experiments conducted by shared spectrum company indicate 62% percent "white space" (unused space) below the 3GHz band, even in the most crowed area near downtown Washington, D. C., where both governmental and commercial spectrum usage

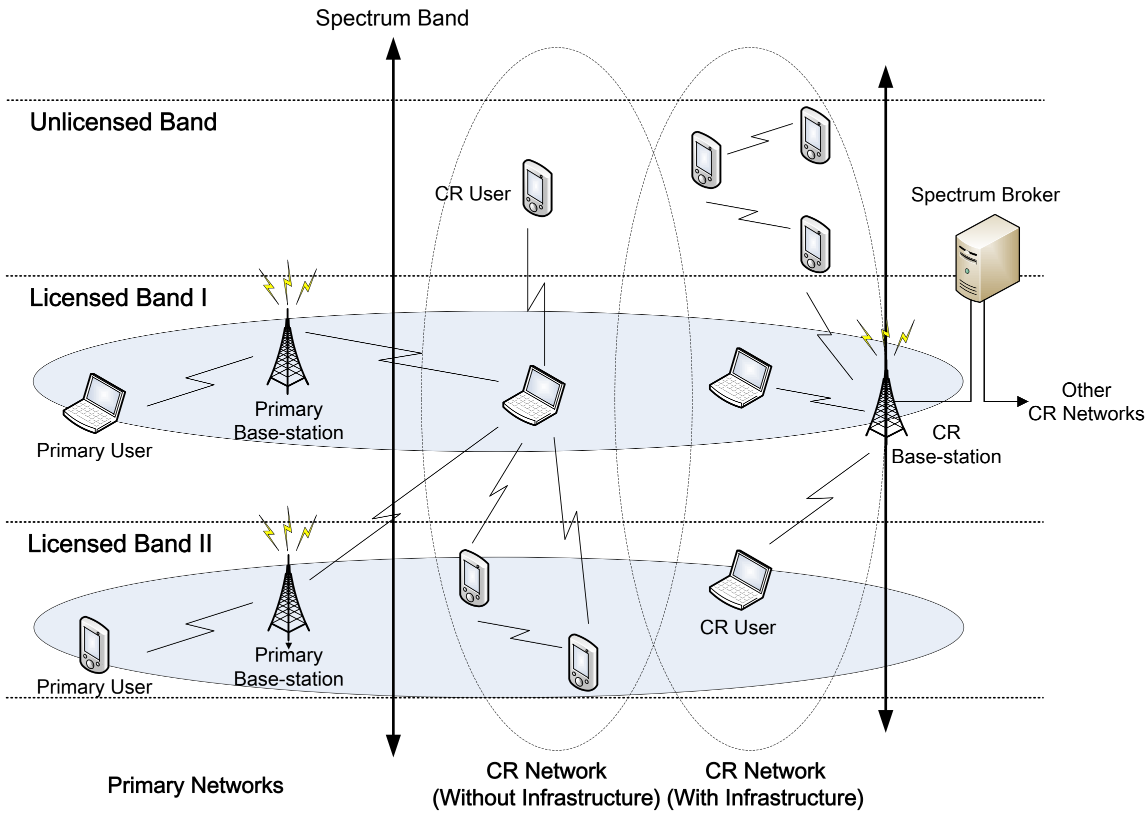

Figure 1.1.: Cognitive radio network architecture

are intensive [2]. In this experiment, a band is counted as white space if it is wider than 1MHz and remains unoccupied for 10 minutes or longer. Furthermore, spectrum usage levels vary dramatically in time, geographic locations, and frequency. A lot of the precious spectrum (below 5GHz), this is worth billion of dollars, and is perfect for wireless communications sits there silently. The limited available spectrum and inefficient spectrum utilization necessitate a new mechanism to exploit the existing spectrum in a opportunistic manner. Consequently, cognitive radio (CR) is proposed to solve this inefficiency problem in spectrum usage [3, 4, 5, 6, 7].

Cognitive radio is a radio that can change its transmission parameters based on what it learns from the environment in which it operates [8]. It is a technology that allows unlicensed users to operate in underutilized licensed frequency bands in an intelligent way without intruding the privileges of licensed users. As shown in Fig. 1.1, a typical infrastructure-based (or centralized) CR network architecture is comprised of two groups of device as the primary network and the CR network. The primary network is referred to as the legacy network that has an exclusive right to certain spectrum band. Examples include the common cellular and TV broadcast networks. The primary network is composed of primary users (PU), also called licensed users, which have a license to operate in the spectrum band, and the primary base stations, *e.g.*, the base station in a cellular system. Typically, the primary users and base stations do not have any CR capability for sharing spectrum with CR users. In contrast, the CR network does not have a license to operate in the desired band. Hence, the spectrum access is allowed only in an opportunistic manner. A CR network is composed of three basic elements as

- *CR user*: A CR user, also called secondary user (SU), has no spectrum license. Hence, additional functionalities are required to share the licensed spectrum band. In infrastructure-based networks, the CR users may be able to only sense a certain portion of the spectrum band through local observations. They do not make a decision on spectrum availability and just report their sensing results to the CR base station.

- *CR base station*: A CR base station is a fixed infrastructure component with CR capabilities. It provide single-hop connection without spectrum access licenses to CR users within its transmission range and exerts control over them. Through this connection, a CR user can access other networks. It also helps in synchronizing the sensing operations performed by the different CR users. The observations and analysis performed by the latter are fed to the central CR base station so that the decision on the spectrum availability can be made.

- *Spectrum broker*: A spectrum broker (or scheduling server) is a central network entity that plays a role in sharing the spectrum resources among different CR networks. It is not directly engaged in spectrum sensing. It just manages the spectrum allocation among different networks according to the sensing information collected by each network.

To successfully recycle the underutilized licensed frequency bands, the key of CR networks is to ensure the CR-enabled secondary users to operate in an intelligent way without any interfering to the privileges of licensed primary users. Note that the activity of primary users could be totally random and independent to the CR networks, it is a must for CR networks to punctually detect the spectrum availability and efficiently manage the use of spectrum. To achieve this goal, the functionality of CR networks consists of four major steps for effective spectrum management [9]:

- *Spectrum sensing*: To maximize the usage of available spectrum bands and to avoid any possible intrusion to primary users, CR users should timely monitor the available spectrum bands, exploit the unused spectrum, and detect the presence of primary users when they operate in a licensed band.

- *Spectrum decision*: Based on the collected information of spectrum availability from individual CR users, CR base stations can allocate a channel. This allocation not only depends on spectrum availability, but is also determined based on internal policies among CR users.

- *Spectrum sharing*: Because there may be multiple CR users trying to access the spectrum, CR network access should be coordinated to prevent multiple users[1] colliding in overlapping portions of the spectrum. In other words, a MAC is required to resolve the possible contention among CR users.

- *Spectrum mobility*: CR users are regarded as visitors to the spectrum. Hence, if the specific portion of the spectrum in use is required by a primary user, the communication must be continued in another vacant portion of the spectrum. In the centralized CR network, the mobility is coordinated by the CR base station.

Through the spectrum management, the CR networks allow users to actively monitor and exploit the temporally unused spectrum, which is referred to as a *spectrum hole* or *white space* [10]. As shown in Fig. 1.2, whenever a primary user occupies this band, the secondary user moves to another spectrum hole to avoid the interference to the primary user. However, due to the instability of channel availability and status, the quality-of-service (QoS) of CR users can hardly be guaranteed, which becomes a significant challenge in CR networks.

1.1.2 On-demand Media Streaming

The past few years have seen a proliferation use of various multimedia applications over wireline packet networks, such as interactive voice communications (voice over Internet Protocol (VoIP)), multimedia messaging, video conferencing, live TV broadcasting and

[1] In this work, we focus on the performance of SUs and abuse the term user to refer to the SU only.

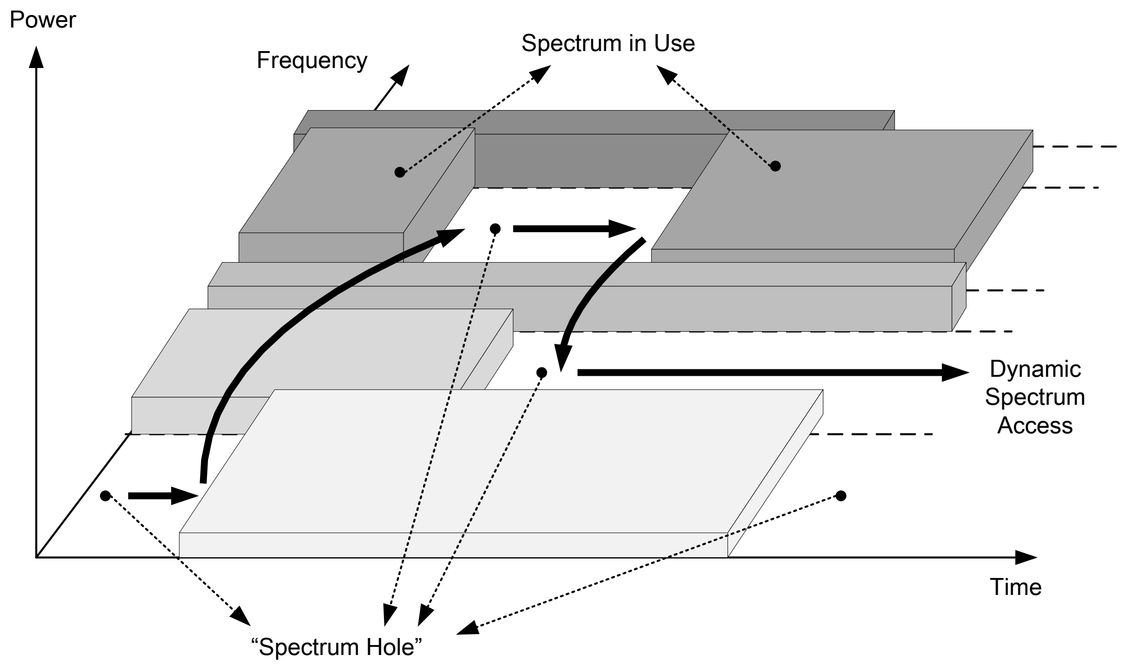

Figure 1.2.: Spectrum hole concept

on-demand multimedia streaming, *e.g.*, PPlive, PPStream, Youtube. In the meantime, the rapid diffusion of wireless devices in the past years, such as PDA, Smartphone, IPod, *etc.*, is increasingly propelling the shift of users toward wireless technologies.

Although wireless networks provide a low-cost and flexible infrastructure to multimedia applications, this infrastructure is unreliable and provides dynamically varying resources with only limited QoS support. Moreover, media streaming applications, such as TV broadcasting, has especial stringent QoS requirements compared with generic data communications [11]. In particular, media traffic is characterized by strong time sensitivity and inelastic bandwidth requirements. Media packets, in fact, must be available at the decoder before their playback time (deadline) to allow an undistorted media reconstruction. Packets that are not received before their deadline become useless. Excessive end-to-end delays might negatively affect user experience as well, for instance, impairing the ability to effectively interact with other users. Regarding bandwidth requirements, media data are

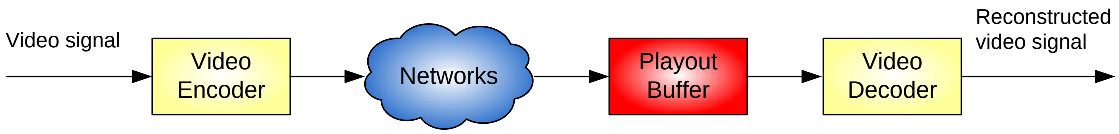

Figure 1.3.: Architecture of media streaming system

generally encoded at a fixed data rate, as in the voice and audio cases. If the bandwidth required by the compressed data exceeds the channel capacity, packet losses will occur causing distortions in the decoded data.

Fig. 1.3 shows the main components of a generic media streaming system. In addition to the encoder/decoder elements, a playout buffer called dejitter buffer is generally introduced before the media decoder to compensate the unequal packet delays caused by the network. The playout buffer allows to trade off a reduction of excessively delayed packets for an increase of the overall end-to-end delay.

From a communication point of view, media streaming can be further divided into two classes:

1. Live media streaming, such as TV/audio broadcasting,

2. On-demand streaming, also called video-on-demand (VoD), such as remote education, Youtube.

The major difference between the two groups is that the latter requires some form of interactions between the source and end hosts [12]. In specific, in the first group, the video packets are delivered one way from the source to destinations. All the users subscribed to the service receive the same media content and proceed the playback at the same rhythm. The latter group, however, allows users to select different media contents from a rich media database and interactively control the rhythm of playback: play/resume, stop/pause, stop/pause/abort, fast forward/rewind, fast search/reverse search, slow motion as identified in [13]. With more flexibility rendered to users, of the many applications of distributed media streaming systems, VoD has much appeal.

With different characteristics of media service, the QoS provision schemes for the two different media applications are also diverse. The live streaming service requires minimum start-up delay to ensure the instant information delivery from the source. In this case, the media playout buffer is kept small to reduce the latency of start-up; therefore, the media traffic requires static throughput support with persistent guaranteed data rate. The on-demand media streaming, however, can tolerate a fairly long start-up delay compared with the live streaming service, as long as the media can be smoothly played without interruptions. In other words, the playout buffer plays an important role now and how to make use of the playback buffer in network resource allocation becomes critical. Fig. 1.4 shows an example of the user playout buffer. As shown in Fig. 1.4, the smoothness of media playback is determined by the media playback rate, download throughput and the buffer storage of user. With abundant buffer storage, the user are resilient to the network dynamics which could benefit the resource allocation. Moreover, it is important to note that the video-on-demand allows users to arbitrarily select the position of media playback. Once the user switches to a new playback position, its previous buffer contents become useless and new content must be pulled immediately via the network. This may require intensive network resource in a short period.

1.2 Research Motivation and Objectives

Motivated by the ever-growing demand of wireless multimedia services and the increasing importance of CR networks in wireless communication, in this thesis, we investigate how to provide the large-scale video-on-demand service over CR networks. As VoD users moving from traditional wireless communication networks expect the same level of performance in CR networks, significant challenges must be addressed to provide successful video stream-

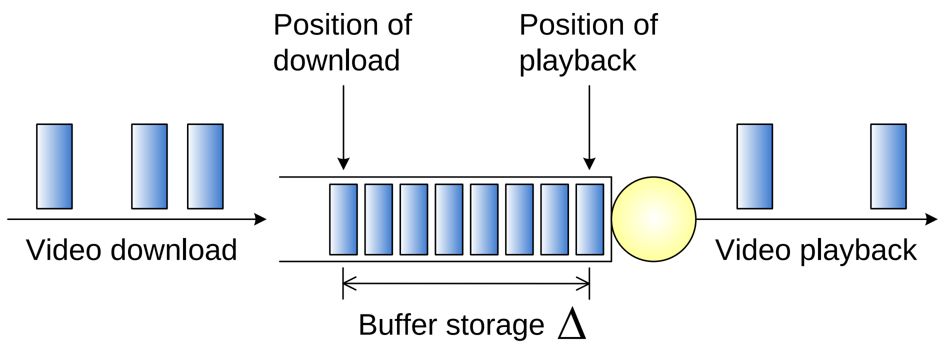

Figure 1.4.: Example of user's playout buffer

ing delivery. Moreover, because of the unique features of CR networks and specific QoS requirements of VoD services, existing schemes designed for VoD over traditional wireless networks or data and live steaming service over CR networks could not be easily extended, which we indicate as follows:

- To provide media streaming services on traditional wireless networks, researchers focus on how to exploit the subscribed spectrum effectively and efficiently, in order to maximize the system throughput and reduce video packet delay. However, in CR networks, the main variable which degrades the system performance is the dynamic nature of spectrum itself. Provision requested QoS in CR networks is difficult, not only because of the changing end-to-end available bandwidth of a certain channel caused by wireless fading or transmission error, but also because of the difficulty to detect and make use of such channel. In specific, a CR user could exploit a licensed channel only when the primary users are not using it, not to mention that multiple CR users may have detected the same available licensed channel and intend to occupy this channel at the same time. In addition, even when a CR user is transmitting, it has to monitor the channel status frequently and release the channel whenever detecting the primary user's presence, to avoid any interference to the primary user. The above mentioned situation makes the channel availability highly dynamic. Therefore, to provide guaranteed QoS to CR users, we first need to sense the channel status correctly, then we coordinate the different CR users operating in the licensed frequency band using QoS-oriented resource allocation scheme and intelligent medium access control (MAC) protocol. It worth noting that compared with other resource allocation used in traditional wireless networks, to maximize the channel usage and avoid interference to PUs, the provision of VoD in CR networks generally requires more frequent re-allocation since the transmission opportunities appear and subside occasionally. On the other hand, re-allocation introduces large overhead, which makes the system throughput even worse. Therefore, the resource allocation scheme need to be efficient, and there exists a tradeoff between instant resource allocation and smaller system overhead.

- Compared with other services over CR networks, such as data communication and video streaming delivery, VoD over CR networks has different QoS requirements. Therefore, the QoS-oriented resource allocation schemes for these applications are completely different. As mentioned before, data communication is delay-tolerant and best-effort (BE) communication, so the throughput requirements are not so stringent and the resource allocation scheme is relatively less complex. In most CR resource allocation works, CR users with data communication requirement usually monitor a fixed group of channels and decide to transmit to each other in a distributed way. Several channel sensing and decision mechanisms and MAC protocols are introduced to ensure CR users capture most available channels and avoid interference to PUs and each others. However, those schemes are not sufficient when deal with media streaming over CR networks. As media traffic is time-sensitive and requires more bandwidth, the resource allocation is much more complex. It is difficult for distributed algorithm to achieve the effective resources allocation to avoid any waste of bandwidth. In addition, media delivery relies on the video server, which determines the CR network is intuitively centralized. The CR base station serves as the central controller which decides the overall system resource allocation and cooperates the spectrum sharing among CR users. And among all the solutions aiming at media delivery over CR networks, most of them ignore the interaction between the video source and the end users, which are applicable to live streaming rather than VoD. Moreover, VoD could tolerate a relatively long start-up delay than live streaming, so the playout buffer plays an important role in tolerate the dynamics nature of CR networks. For instance, a VoD user with smaller playout buffer storage is more urgent for downloading video packets, so it tends to be allocated to a "good" channel; reversely, a VoD user with smaller playout buffer storage is less urgent for downloading video packets, so it tends to be allocated to a "bad" channel. As far as we know, the feature of playout buffer has not been well exploited in the resource allocation of CR networks.

As such, we are motivated to investigate how to provide the large-scale video-on-demand service over CR networks. To achieve this goal, we exploit the unique feature of VoD service, while considering the dynamic nature of CR networks. The objective of this research is to propose a centralized scheme for provisioning the superior video service to users, which is mainly a resource allocation problem formulated by an optimization framework. Given the stochastic characteristics of current spectrum resources combined with the video packet storage in users' playout buffer, we derived an algorithm to maximize the overall network throughput while providing users with the smooth video playback.

1.3 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 provides the detailed literature survey of existing works on video streaming over CR networks. We introduce the system model in Chapter 3, including the network architecture, activity model of primary users, MAC protocol of VoD users, wireless channel model, along with the QoS requirements of VoD users. Chapter 4 formulates the optimization problem for the large scale CR-based VoD service, and Chapter 5 presents the simulation results to evaluate the performance of the centralized algorithm. Finally, Chapter 6 closes this research with conclusions and future works.

Chapter 2

Background and Literature Review

In this chapter, we will highlight the issues related to the main component of a CR VoD system and discuss pertinent issues. We will evaluate previous and current research and show how this dissertation will address some of the limitations of previous work.

2.1 Video on Demand Systems and Techniques

Driven by the increasing expectations of customers for fully customizable and more convenient services, yet at low prices, VoD is expected to replace both broadcast-based TV programming and DVD movie rentals at stores. The market has already prepared for such a change, and some landmark steps have been taken. For example, many digital TV service providers offer VoD, and some TV channels, such as Rogers Cable TV, have launched VoD programming. Moreover, motivated by the availability of high user download bandwidth, many news networks (such as CNN and BBC) provide short, medium-quality, on-demand news clips on their web-sites. The popularity of video web-sites, such as YouTube, provides another clear indication of this direction in video delivery.

Unfortunately, the number of video streams that can be supported concurrently is highly constrained by the required real-time and high-rate transfers, which quickly consume server and network bandwidth resources.

Recognizing the constrained resource of users, different VoD systems provide user-specific service tailored to the available bandwidth and network traffic of viewers. In general, the VoD services can be categorized into four groups, based on the scheduling policies of data delivery and on the degree of interactivity [14]:

- *No Video-on-Demand (No-VoD)*: The No-VoD is a service where the user is passive and has no any control on what he/she watches. In this case, users have no video requests.

- *Pay-Per-View (PPV)*: In PPV the viewer signs up and pays to watch a specific program which is scheduled at predetermined times.

- *Near Video-on-Demand (NVoD)*: NVoD is a technique that groups users who request for the same video and serves them using one video stream to minimize the server bandwidth. The server is in control of when to serve the video. Video Cassette Recording (VCR) capabilities are provided, by using many channels delivering the different requested portions of the same video requested by the different users. In NVoD, the users have to wait to watch the video that he/she wishes, and do not have the full control over it.

- *Quasi Video-on-Demand (QVoD)*: QVoD [15] is a threshold-based NVoD. The server delivers a video when the number of user requests lager than a predefined threshold.

- *True Video-On-Demand (TVoD)*: In TVoD users have the 100% of control over the video session with full VCR capabilities, such as Youtube. The simplest way to achieve TVoD is to dedicate each channel to each user in the system.

Of the four categories, NVoD and TVoD are of the most interests, and meanwhile are of most difficulties. Research community face this conflict of full interaction and constrained bandwidth by utilizing the multicast facility, i.e., to stream the same video file to a large group of users of the same interest. The main classes of these techniques are Batching, stream merging, periodic broadcasting, and composite techniques. In the following, we will describe those techniques in details.

2.1.1 Batching

Batching [16, 17] simply services all waiting requests for a video using one full-length multicast stream, called regular stream. Thus, it requires only one download channel. Because streams are delivered at the video playback rate, the required user download bandwidth equals to the video playback rate.

2.1.2 Stream Merging Techniques

These techniques reduce the delivery costs by combining streams. They include, in increasing order of complexity and performance, Stream Tapping/Patching [18], Transition Patching [19], and Earliest Reachable Merge Target (ERMT) [20]. Each of these techniques requires two user download channels at the video playback rate. Patching expands the multicast tree dynamically to include new requests. A new request joins the latest regular stream for the video and receives the missing portion as a patch stream at the video playback rate. Hence, it requires additional buffer space and two download channels, leading to a total required download bandwidth of $2r$. When the playback of the patch is completed, the user continues the playback of the remaining portion using the data received from the multicast stream and already buffered locally. To avoid the continuously increasing patch lengths, regular streams are retransmitted when the required patch length for a new request exceeds a pre-specified value, called regular window (W_{reg}).

Transition Patching allows some patches to be sharable by extending their lengths. It introduces another multicast stream, called transition patch. As a motivating example

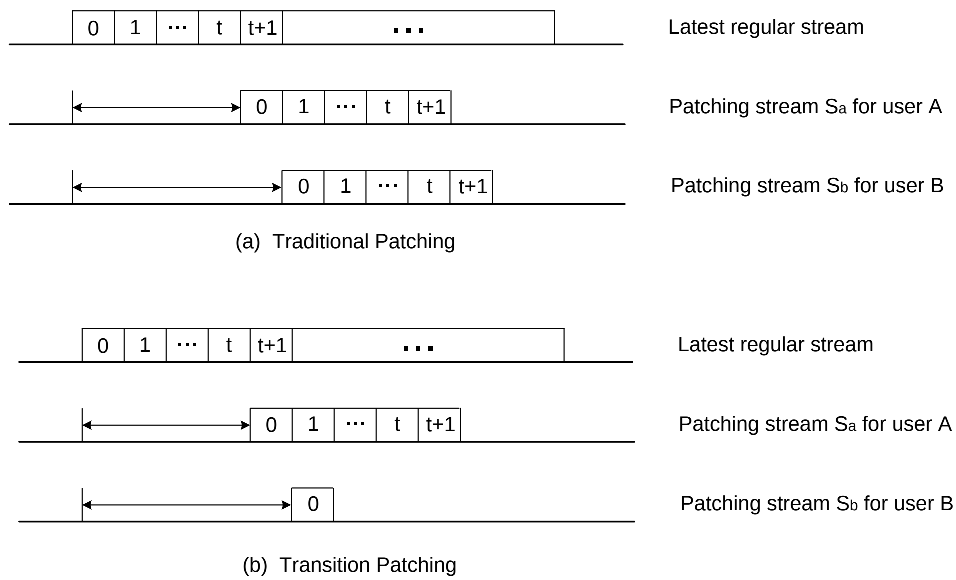

Figure 2.1.: A comparison between the traditional patching and transition patching

shown in Fig. 2.1, let's say the latest regular stream is started at time 0. At time t, a patching stream, S_a is initiated for user A. One time unit later, another patching stream, S_b, is scheduled for user B. If we use $v[t_1, t_2]$ to denote the segment of the video from the t_l time unit to the t_2 time unit, then the patch required by user A is $v[O, t]$, under the existing approach. Similarly, the patch delivered by S_b is $v[O, t+1)$. Totally, $2t+1$ time units of video data are delivered for the two users. However, if we make S_a, deliver $v[O, t+2]$, a patch with two extra time units, then user B needs only the first time unit of the video, i.e., $v[0, 0]$. User B can receive and play back the video as follows. One of its data loaders, say L_1, is used to download data from S_b, from which it receives video clip $v[O, O]$. At the same time, another loader, L_2, starts downloading data from S_b where it can receive video clip $v[1, t+2]$. Once all data from S_b is received, L_1 switches to download the remaining data from the regular stream.

ERMT also requires $2r$ of user download bandwidth and additional user buffer space. However, each stream is sharable by later-joined users, while traditional patching only shares regular streams with later users. In ERMT, a new client or newly merged group of clients listen to the closest stream that it can merge, which is called *target*, and receive the missing portion by a new stream called *merger*. After merging, the newly merged group of clients listen to the closest stream, and continues merging.

2.1.3 Periodic Broadcasting Techniques

Whereas stream merging suits a wider spectrum of video workloads, periodic broadcasting techniques can be more efficient in serving highly popular videos. These techniques divide each video into multiple segments and broadcast them periodically on dedicated channels. They include Skyscraper Broadcasting (SB) [21], Greedy Disk-Conserving Broadcasting (GDB) [22], Fibonacci Broadcasting (FB) [23], and Harmonic Broadcasting (HB [24]. In SB, GDB, and FB, the segments are of variable length and the channels have equal bandwidth. The user waits until the beginning of the next broadcast of the first segment and then receives data concurrently from two broadcast channels. The relative length of the nth segment compared to the first segment is determined using a technique-specific partitioning function, as shown in Table 2.1.3. To illustrate the main concept, if a 30-minute video is divided into three segments by FB, the length of the first segment and thus the maximum waiting time are $30/(1 + 2 + 3) = 5$ minutes. Note that the maximum waiting times decreases with the number of segments at the expense of increasing server bandwidth. In contrast, HB-based protocols have uniform-length segments and the channel bandwidth decreases with the segment number. Despite their high effectiveness, they require partition each video into a relatively large number of segments (to reduce the maximum delay).

2.1.4 Combining Stream Merging and Periodic Broadcasting

The modified version of GDB (called here MGDB) is used to deliver hot videos, but instead of making the user wait until the beginning of the next broadcast time, the user receives

Table 2.1: Segment Partitioning in Periodic Broadcasting

Technique	**Partitioning Function**
Skyscraper Broadcasting (SB)	$f(n) = \begin{cases} 1 & n = 1 \\ 2 & n = 2, 3 \\ 2f(n-1)+1 & n \bmod 4 = 0 \\ f(n-1) & n \bmod 4 = 1, 3 \\ 2f(n-1)+2 & n \bmod 4 = 2 \end{cases}$
Fibonacci Broadcasting (FB)	$f(n) = \begin{cases} 1 & n = 1, 2 \\ f(n-2)+f(n-1) & n > 2 \end{cases}$
Greedy Disk-Conserving Broadcasting (GDB)	$f(n) = \begin{cases} 1 & n = 1 \\ 2 & n = 2, 3 \\ 5 & n = 4, 5 \\ 12 & n = 6, 7 \\ 5f(n-4) & n > 7 \end{cases}$
Modified Greedy Disk-Conserving Broadcasting (MGDB)	$f(n) = \begin{cases} 1 & n \leq 3 \\ 2 & n = 4, 5 \\ 5 & n = 6, 7 \\ 12 & n = 8, 9 \\ 5f(n-4) & n > 9 \end{cases}$

the small missed portion by a patch. A user initially listens to one broadcast channel and its own patch, and then listens to two broadcast channels. The partitioning function of MGDB is shown in Table 2.1.3. If we define a function $h(m) = \sum_{i=1}^{m} f(i)$, then the length of the first segment is equal to $D/h(m)$, where D is the length of the entire video, and $h(m)$ is given by:

$$h(m) = \begin{cases} m, & m = 1, 2, 3, \\ 5, & m = 4, \\ 7, & m = 5, \\ 5^{(m-6)/4} \times 27/2 - 1.5, & m > 5, m \bmod 4 = 2, \\ 5^{(m-7)/4} \times 37/2 - 1.5, & m > 5, m \bmod 4 = 3, \\ 5^{(m-8)/4} \times 61/2 - 1.5, & m > 5, m \bmod 4 = 0, \\ 5^{(m-9)/4} \times 85/2 - 1.5, & m > 5, m \bmod 4 = 1. \end{cases}$$

Selective Catching extends Catching by delivering cold videos using Controlled Multicast [25], which works essentially as Patching. Another hybrid solution that combines Batching with SB was proposed in [26].

2.2 Cross Layer Design for User-specific QoS Provision

In addition to various VoD systems and techniques, provisioning user desired video quality has recently attracted more and more attention from the research community. To this end, an extensive body of cross layer designs have been proposed which tune the parameters residing in different network layers and entities to economically and efficiently utilize the constrained network resource for persistently enhanced video quality from the user's perspectively. In the context of CR networks, video streaming suffers from the constrained and dynamic varying network resource. As such, a flexible and efficient system which could fast adapt to the varying network status is particularly desirable. In what follows, we first elaborate on the fundamental concept behind the term “QoS” in the video streaming system. After that, we review several recent research works which invoke the cross layer design framework for QoS provision in the cognitive radio networks.

2.2.1 QoS from the Network Perspective

Although the word QoS from the network perspective is a vague and all-encompassing term, in most cases, it means the need for maintaining the following metrics at a certain level:

- Data rate (or bit rate, throughput)
- End-to-end delay (including the processing time, queueing delay and transmission delay at the end-system and network)
- Network jitter (or delay variation)
- Reliability (usually represented by the bit error rate (BER), packet loss rate, *etc.*)

These metrics measure the required network resource of the media service and are commonly used for efficient network resource allocation, such as QoS routing and scheduling; nevertheless, they are not direct to characterize the delivered video quality perceived by the users, and thus disregard the quality of media presentation from the user’s perspective. As the user’s satisfaction is the ultimate metric for evaluating a multimedia service, it is crucial to specify the QoS requirements from the user’s perspective and reflect them in the network resource allocation and QoS provision.

2.2.2 QoS from the Viewer Perspective

The perception of service quality from end-user point of view, however, is a wider and more subjective issue. In general, it could be specified from the following two aspects.

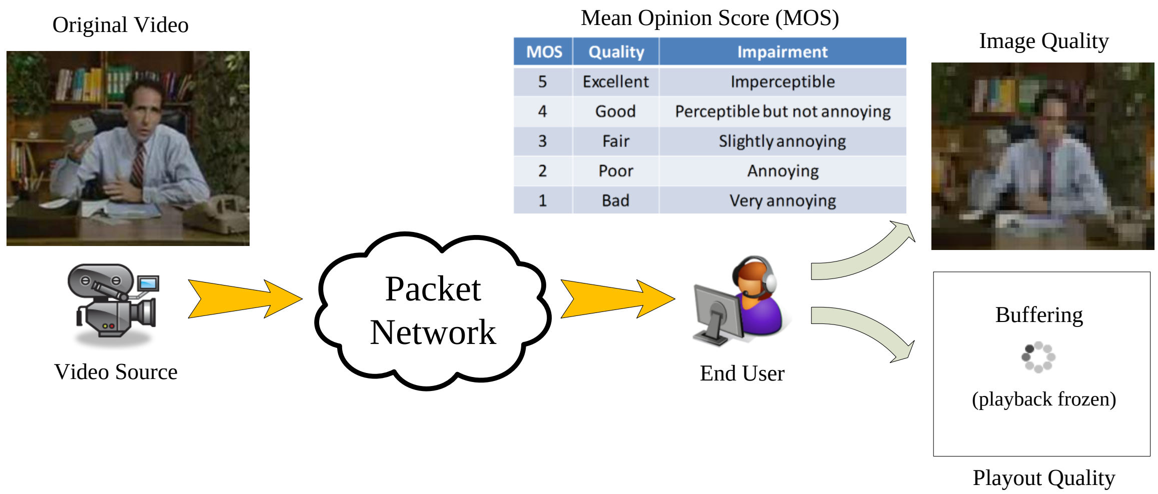

Figure 2.2.: Perceived video performance at the receiver

Media Image Quality

The media image quality is typically evaluated by the resolution distortion of the received video images at the end user, as shown in Fig. 2.2. The image distortion is mainly caused by the packet loss during the network transmission or receiver reception. There are two basic approaches to measure the resolution distortion, namely, subjective quality measures and objective quality measures.

Subjective Quality Measures In this approach, the source video and the transmitted video are presented in pairs to a large group of viewers who grade the media quality on a scale from 5 (the best) to 1 (the worst). The scale is called mean opinion score (MOS) as shown in Fig. 2.2. The subjective approach is the most reliable way of assessing the quality of the video pictures by capturing the impression of users watching the video. However, as the subjective quality measure involves a large group of viewers and special viewing equipments, it is not a convenient and cost-effective approach and can not provide in-service quality monitoring for real-time media system design and optimization.

Objective Quality Measures The approach typically uses the mean squared error (MSE) and peak signal-to-noise-ratio (PSNR) to measure and predict the perceived media image quality automatically without involving viewers. The MSE and PSNR have clear physical meanings, and are mathematically easy to apply to rate-distortion optimized communication systems. But they have been criticized for not correlating well with MOS values especially for low-quality video signals.

Media Playback Quality

The media playback quality reflects the user's overall viewing experience within the whole session of media presentation. It is represented by the start-up delay and the smoothness of media playback, which are closely related to the playback buffer used by the receiver to overcome the network dynamics. The start-up delay accounts for the time delay when users initiate a request until the video frames in the playout buffer are played out, and the smoothness of media playback is measured by the frequency of playback frozens when video stalls during the playback due to empty playout buffer. Apparently, a long start-up delay is annoying to users, especially when they are watching live programs. Frequent playback frozens should also be reduced as it ruins the continuity of the video.

Notice that for different users and applications, the requirements on the two playback quality metrics are diverse. For example, in live video broadcasting service, (*e.g.*, football game), users usually want to minimize the start-up delay; whereas for Video-on-Demand service, most users can tolerate a certain delay but prefer a smooth playback. Even for the same applications, users may have different preference on the playback quality, *e.g.*, some users may appreciate a shorter start-up delay even at the cost of frequent interruptions while some others prefer smooth playback at the sacrifice of the long waiting time. Therefore, to provide satisfactory user perceived quality, multimedia service provisioning should not only be application-aware, but also quality-driven which caters to individual user's preference on the perceived media quality.

2.3 Cross-Layer Design for QoS Provision in Cognitive Radio Networks

The mainstream CR research has been focused on spectrum sensing and dynamic spectrum access, (*i.e.,* the MAC and PHY) issues, as observed in [5, 27]. For example, the approach of iteratively sensing a selected subset of available channels has been adopted in the design of CR MAC protocols [28]. The important trade-off between the two types of sensing errors is addressed in depth in [29]. The equally important QoS issue has been considered only in a few papers [30, 31, 32], where the focus is on the so called "network-centric " metrics such as maximum throughput. In [33], the authors proposed an admission control scheme and channel allocation policy for ad hoc CR networks, based on a user dynamic ID numbering approach, to meet the QoS requirements of the CR network. Media streaming over CR networks, as one of the most important Internet services, has attracted more and more efforts from the research community.

[34] describes a live video streaming system over infrastructure-based cognitive radio networks. In [34], the CR users are composed of G multicast groups with each having N_g users, $g = 1, 2, ..., G$. The multicast users of each group may reside in different and fixed spectrum bands without the capability of spectrum mobility. The video files are encoded using the layered video coding and broadcasted to multicast groups from the CR base station. Upon each available spectrum band, the CR base station could only broadcast one video file to the most appropriate multicast group. The research is to determine which file should be delivered over certain spectrum band based on the different channel status and achieved video quality of users. Towards this goal, [34] propose an optimizer of the network which exploits three dimensions of the system: coding rate selection, spectrum selection and spectrum sharing. The first issue is to determine the optimal coding rates for broadcasting. In specific, using layered coding, the video is encoded to different layers, including base layer and enhanced layers, at different rates. To reconstruct the video, one

must download the base layer, while the enhanced layers are additive to the base layer for improved video quality. With more layers downloaded, users could enjoy enhanced video performance, but at the cost of higher bandwidth consumption. Therefore, as users have different channel status, the video should be coded at appropriate rates to ensure the playback. The goal of layer selection is to determine the best video coding rate to maximize the integrated video quality of multicast users. Secondly, as each channel may include users associated with different multicast groups, the channel selection determine which video file should be broadcasted over each channel towards the maximal network utility. This may depend on the population of the multicast group users in each channel and their integrated received video quality. Third, at different time slot, the channel could be used to broadcast different video files. Based on a TDMA MAC, the optimizer determines the access time of multicast groups on each channel spectrum. The proposed optimizer is an integer programming problem which is solved by an approximated algorithm. Unlike our work, [34] fails to consider the buffer storage of users which is particularly useful to fight with the channel dynamics.

In [35], a game-theoretic framework is described for resource allocation for multimedia transmissions in spectrum agile wireless networks. Similar to [34], the work is conducted in the infrastructure-based CR network governed by the central controller. However, [35] considers the selfishness of users where each user may provide the wrong information to the CR base station in order to attain more resource. The network is then modeled as a resource management game. Using game theoretical analysis, an incentive scheme is proposed to ensure the honesty of users. Based on this, a centralized mechanism-based scheme is described to determine the amount of transmission time to be allocated to various users on different frequency bands such that certain global system metrics are optimized.

[36] also studies the resource allocation for real-time streaming in CR networks. However, [36] focuses on the uplink where users competes for the channel access for video upload. Based on the buffer storage of SUs and their channel status, the base station

performs the channel scheduling and power allocation within independent OFDMA subchannels to minimize the packet loss of SUs caused by the buffer overflow. In specific, users with heavily loaded queues have the priority to upload, and users with nearly empty queues will be penalized, which is implemented by the proportional fair (PF) scheduling algorithm. The PF algorithm chooses the user to transmit on a certain subchannel that has the largest ratio of the transmission data rates and the average throughput. Moreover, a video encoder rate control is introduced to limit the video frame loss.

[37] proposed a dynamic channel selection solution to allow users to adapt their channel selection and maximize video qualities. Specifically, the secondary users have their own video to be uploaded, and each video is separated into several quality layers as in [34] and each layer could be transmitted on different frequency band, while the highest priority class is reserved for primary users. All video layers transmitted on each frequency band form a virtual queue, which reflects the impact on each secondary user from the action of other secondary users. Based on this virtual queue analysis, the secondary users select frequency channels considering the various QoS requirements such as rate requirements and delay deadlines, to maximize the transmitted video qualities.

[38] studies the QoS provision in an OFDMA network, where two groups of users, the best-effort (BE) and the real-time (RT) users, compete for the channel access. As a result, an optimization framework is proposed to achieve the maximal throughput of the network while satisfying the specific QoS requirements of different users.

However, all the above literatures are on live streaming which only exploit the adaptive coding rate but fail to consider the unique features of video-on-demand service. In this thesis, we provide a centralized algorithm for on-demand streaming over CR networks.

Chapter 3

System Model

3.1 Network Architecture

We consider an infrastructure-based single-hop CR network as shown in Fig. 3.1.

The overall spectrum band of the network is composed of N orthogonal channels indexed by n ($n = 1, 2, \ldots, N$) and each channel is allocated to one primary user (PU) also indexed by n. The N channels evolve independently and could be respectively occupied by N PUs who have exclusive access. Meanwhile, M secondary users (SU) indexed by m ($m = 1, 2, \ldots, M$) coexist with PUs in this network, and only access channels when PUs do not use that channel. To realize such opportunities of idle channel usage, SUs need to possess the capability to utilize the channels in an intelligent manner, i.e., to form a cognitive radio network.

The cognitive radio network composed of one central Base Station (BS) and M SUs. Without loss of generality, we assume that there are two groups of SUs: VoD users with video delivery and best effort (BE) users with data transfer. The VoD users download the subscribed video clips from remote video servers through the BS using the CR interface. The BE users also communicate with the same central BS using the same CR interface as VoD users, but with both uplink and downlink transmissions. Here, the BS is an over-

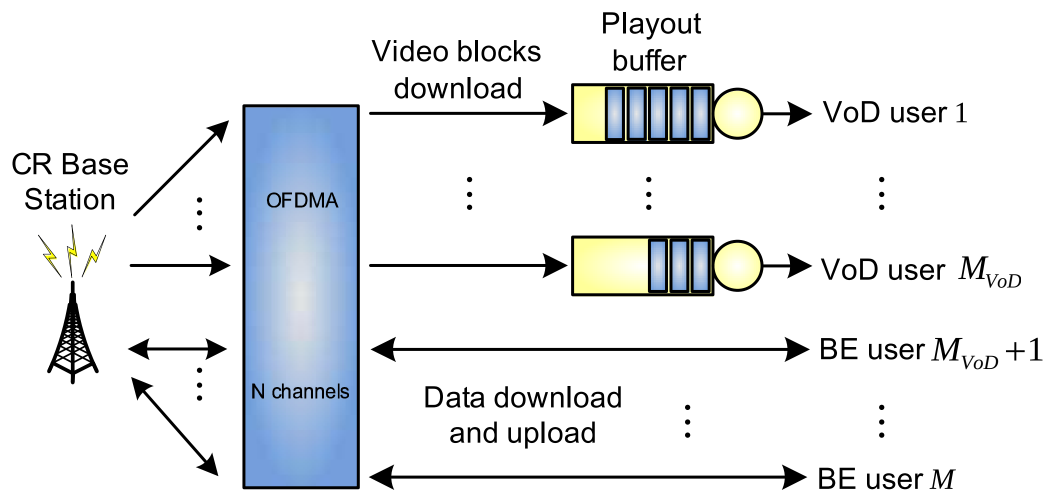

Figure 3.1.: Illustration of the infrastructure-based CR network

all entity which in fact consists of two components: CR base station and video content provider. The CR base station is responsible for all CR-related functions, such as collecting the information of spectrum resource statistics and observations of VoD users' buffer storage, allocating channels to CR users, coordinating the transmission between VoD users and video content provider, and maintaining single-hop connections to all CR users. It also synchronizes the sensing operation of different CR users. (In this thesis, the expressions "user", "CR user", and "secondary user" are interchangeably used.) Video content provider consists of massive storage and media controllers to store a large number of videos and serve simultaneous video requests to different VoD users, using the spectrum allocated by CR base station.

As shown in Fig. 3.1, let M_{VoD} denote the number of VoD users indexed from 1 to M_{VoD}. Each VoD user deploys the playout buffer before the decoder at the receiver end. Let M_{BE} denote the number of BE users indexed from $M_{\text{VoD}}+1$ to M. Hence, we have $M_{\text{VoD}} + M_{\text{BE}} = M$.

Let τ denote the maximal time duration which the PU could tolerate in presence of the interference from SUs. To avoid the interference from SUs to PUs, we slot the system time

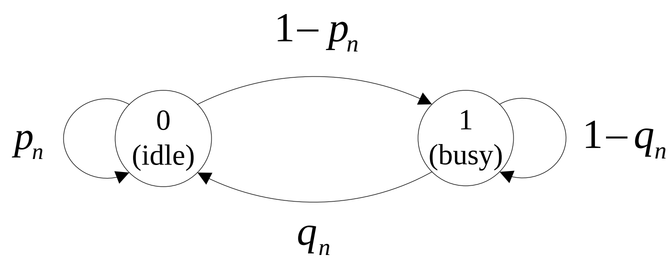

Figure 3.2.: On-OFF model for channel n ($n = 1, 2, \ldots, N$)

into discrete intervals of τ. In each time slot, SUs monitor channels actively and are allowed to transmit only when the channel is sensed idle. In other words, using τ as the spectrum sensing and transmission unit is to guarantee the privileges of licensed users. However, there is no need to allocate channels every τ seconds. To prevent unnecessary overhead and computational complexity, the BS allocates channel spectrums to SUs according to their specific QoS requirements and channel status every $L = T/\tau$ time slots, namely a *channel allocation epoch*. In the ensuing L time slots, SUs follow the allocation and monitor and transmit on their allocated channels in a distributed manner. The design of the channel allocation will be detailed later.

3.2 Channel Availability

According to the activity of PU, the availability of each channel n is abstracted by an ON-OFF model in which state "0" represents that the channel is available with PU silent and state "1" represents that the channel is busy with PU transmitting, as shown in Fig. 3.2. The transition takes place every slot with probability

$$P_n = \begin{bmatrix} p_n & 1-p_n \\ q_n & 1-q_n \end{bmatrix} \tag{3.1}$$

Let $\pi_{0,n}$ denote the limiting probability that channel n is in state "0". Based on accumulated observations, we could derive $\pi_{0,n}$ as

$$\pi_{0,n} = \frac{q_n}{1 + q_n - p_n} \tag{3.2}$$

3.3 MAC for Secondary Users

To avoid frequent communication with the central BS, we leave the channel sensing and the initiative for transmission to each secondary user itself. During every channel allocation epoch, SUs allocated to the same channel distributedly monitors channel status and compete the transmission opportunity when the channel is idle. Here we deploy the p-persistent MAC to coordinate this competition. In specific, in each slot, when a channel n is sensed idle, each SU allocated on this channel will issue a request for video or data transmission with probability p_n, which is related to the number of SUs on channel n. If no collision happens, the SU sending the request would be rendered for transmission in the ensuing slot time τ. Otherwise, this opportunity would be wasted. In this way, the channel spectrum is fairly shared by all of the SUs on channel n. Although we use p-persistent MAC as an example for its simplicity, it is convenient to modify our framework if other MAC protocols are applied [39].

In this thesis, we assume that SUs could sense the channel state instantaneously and correctly.

3.4 Channel Model

We assume that channel allocation epoch is shorter than the channel coherence time, so the channel gain remains constant during T seconds. And we assume that each SU m is able to measure its SNR upon channel n, denoted by $\rho_{m,n}$. Since channel gain is constant, $\rho_{m,n}$ is assumed unchanged, within the channel allocation epoch time T. By reason that SUs are located in different places and channels are fading independently, $\rho_{m,n}$ of SUs are distinct from one to another. Let $c_{m,n}$ denote the achievable transmission rate at which the SU m could transmit or receive over channel n. According to the Shannon capacity, we have

$$c_{m,n} = \frac{1}{T} \log_2 \left(1 + \rho_{m,n}\right). \tag{3.3}$$

3.5 QoS Requirements

The proposed channel allocation scheme aims at maximizing the overall network throughput while providing users with guaranteed QoS. Since we consider two groups of users, i.e., VoD and BE users, which have distinct QoS requirements from each other, we have to first formulate the QoS requirements for both of these two kinds of users.

3.5.1 QoS Requirements of BE users

The QoS requirements of BE users are represented by the mean throughput of the uplink and downlink transmissions, to prevent the large delay of their packets. For any BE user m, let D_m denote the demanded average transmission rate. Let d_m denote the mean upload and download rate of the BE user m. Mathematically, the QoS requirement of BE users is specified as

$$d_m \geq D_m, m = M_{\mathrm{VoD}} + 1, \ldots, M. \tag{3.4}$$

3.5.2 QoS Requirements of VoD users

The QoS requirement of VoD users is represented by the smoothness of the video playback which depends on the current storage of the playout buffer. As shown in Fig. 3.3, the playout buffer is deployed at the user end to store the downloaded video packets and sustain the smooth playback in the dynamic network. With packets downloaded and played out at variable rates, the playback will be frozen once the playout buffer becomes empty. Let

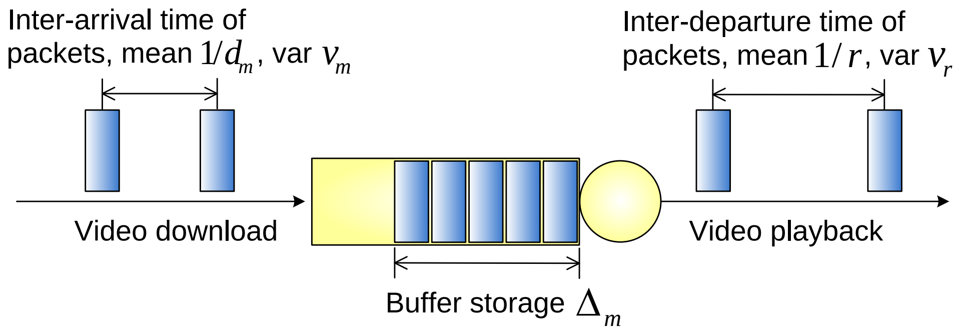

Figure 3.3.: Example of VoD user m's playout buffer

$\mathcal{P}_m$ denote the probability that the playback of VoD user m is frozen in T seconds, Δ_m the storage in the playback buffer of user m, $1/d_m$ and v_m the mean and variance of the inter-arrival time of the packets to user m's playout buffer, and $1/r$ and v_{r} the mean and variance of the inter-departure time of packets of user m's playout buffer, respectively. From [40], the probability of playback frozen of user m is heavily dependent on its buffer storage, download and playback rates as

$$\mathcal{P}_m = \int_0^T g(t)\, dt, \tag{3.5}$$

where $g(t)$ is the density function of the time duration in which video can be played smoothly, as

$$g(t) = \frac{\Delta_m}{\sqrt{2\pi\left(d_m^3 v_m + r^3 v_{\mathrm{r}}\right) t^3}} \exp\left\{-\frac{\left[\left(d_m - r\right)t + \Delta_m\right]^2}{2\left(d_m^3 v_m + r^3 v_{\mathrm{r}}\right)t}\right\}. \tag{3.6}$$

The QoS of each VoD user m is guaranteed by upper bounding $\mathcal{P}_m$ as

$$\mathcal{P}_m \leq \epsilon,\ m = 1, ..., M_{\mathrm{VoD}}, \tag{3.7}$$

where $0 < \epsilon \ll 1$ represents the level of playback smoothness. For ease of exposition, we assume ϵ to be constant and the same for all the users. It is, however, easy to extend by differentiating ϵ and QoS requirements for different VoD users.

Chapter 4

Optimal VoD Streaming over CR Networks

In this section, we first describe the system architecture, and then present the optimal channel allocation in details.

4.1 Description of System Structure

Fig. 4.1 depicts the basic structure of the VoD system which operates iteratively at the interval of channel allocation epochs (L slots). Each epoch is composed of two phases: beacon period and transmission period.

4.1.1 Beacon Period

The beacon period is at the beginning of each epoch, as shown in Fig. 4.1. Within this period, the central BS first collects the three-tuple profiles of each SU m, including the current buffer stage Δ_m, video playback rate (mean and variance of inter-departure time r and v_{r}) and the measured SNRs $\rho_{m,n}$ upon each channel n. After that, the BS performs the

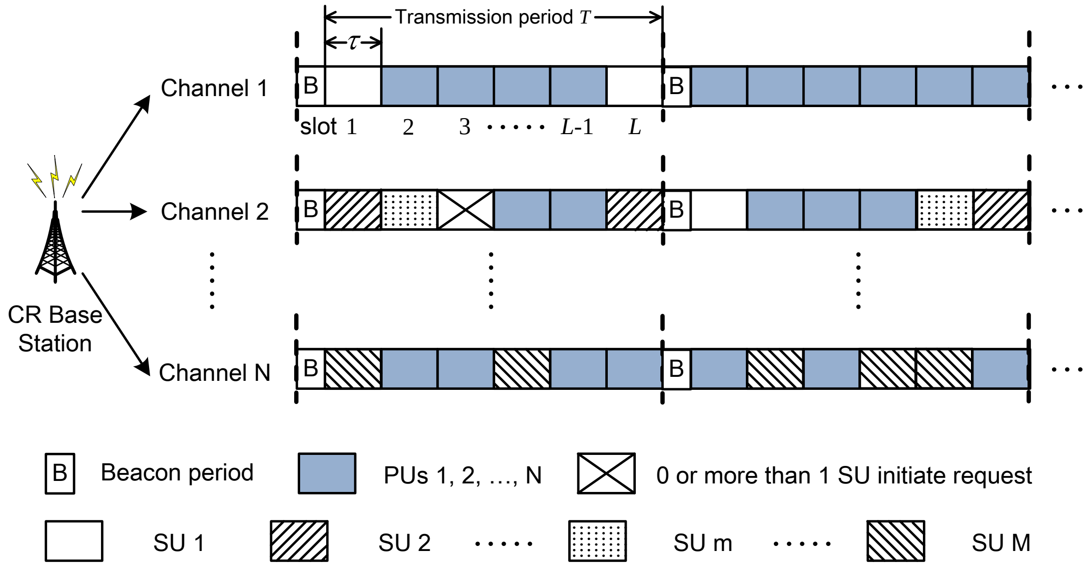

Figure 4.1.: Iterative channel allocation in the proposed framework

optimal channel allocation as follows: based on the user profile, the BS first calculates the Shannon capacity of users upon each channel in (3.3), and then figures out the mean and variance of the inter-arrival time of the transmitted packets, based on the afore calculated Shannon capacity, statistics of channel availability, and the competition among SUs. After that, the BS estimated the video quality of VoD users in (3.5) and acquired throughput of BE users. Based on the QoS requirements of SUs combined with the overall system throughput, the BS optimally allocates the channel to SUs which is described in Section 4.2, and then broadcast the allocation to SUs at the end of the beacon period.

4.1.2 Transmission Period

After the beacon period, each SU monitors the availability of the channel which it is assigned to at every slot τ and competes for the transmission opportunities using the p-persistent MAC once the channel is sensed idle. As shown in Fig. 4.1, each time slot can be either occupied by PU and SUs or unused if no SU decides to transmit or collision happens among multiple SUs.

4.2 Optimal Channel Allocation

We now formulate the optimal channel allocation of BS in the beacon period. Our goal is to maximize the overall system throughput, while satisfying the QoS requirements of both VoD and BE users, i.e.,

$$\begin{aligned}
\underset{a_{m,n}}{\text{maximize}} \quad & \sum_{m=1}^{M} d_m \\
s.t. \quad & \mathcal{P}_m \leq \epsilon, \qquad m = 1, ..., M_{\text{VoD}}, \\
& d_m \geq D_m, \qquad m = M_{\text{VoD}} + 1, \ldots, M, \\
& \sum_{n=1}^{N} a_{m,n} = 1, \quad m = 1, 2, ..., M,
\end{aligned} \tag{4.1}$$

where $a_{m,n}$ is binary as

$$a_{m,n} = \begin{cases} 1, & \text{if channel } n \text{ is allocated to user } m, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

The last constraint of (4.1) dictates that each SU can only be assigned to one channel in each channel allocation epoch.

To solve (4.1), in what follows, we represent $\mathcal{P}_m$ and d_m by the channel allocation $a_{m,n}$.

Let M_n denote the number of SUs allocated to channel n, mathematically,

$$M_n = \sum_{m=1}^{M} a_{m,n}. \tag{4.3}$$

With the knowledge of $a_{m,n}$, based on the p-persistent MAC protocol, the probability that a SU m transmits successfully over channel n in each slot τ is

$$P_{\text{suc},n} = p_n \left(1 - p_n\right)^{M_n - 1} \pi_{0,n}. \tag{4.4}$$

By setting $\frac{dP_{\text{suc},n}}{dp_n} = 0$ and getting $p_n = 1/M_n$, we obtain the maximum $P_{\text{suc},n}$ and the throughput accordingly.

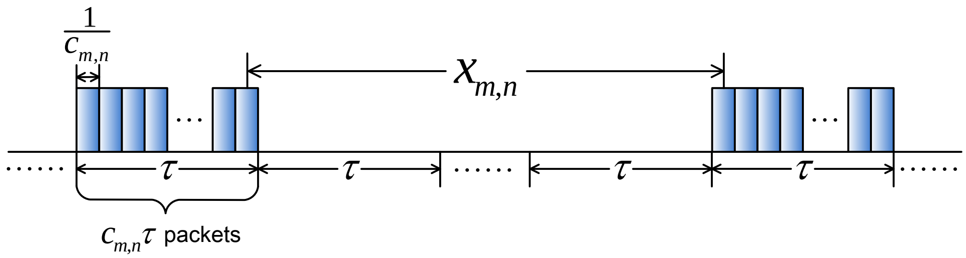

Figure 4.2.: Inter-arrival time of video packets

Let $x_{m,n}$ denote the time between two consecutive transmission slots of SU m on channel n, as shown in Fig. 4.2. As the SU contends for the transmission opportunities using the p-persistent MAC scheme, we have the probability mass function of $x_{m,n}$ as

$$P\left\{x_{m,n} = i\tau + \frac{1}{c_{m,n}}\right\} = P_{\text{suc},n}\left(1 - P_{\text{suc},n}\right)^{i}, \tag{4.5}$$

where $c_{m,n}$ is the capacity of user m on channel n, and $\frac{1}{c_{m,n}}$ is the interval between any two consecutive packets within one slot.

Next we try to obtain the mean and variance of the interval time between the consecutive downloaded packets, given the knowledge of $x_{m,n}$. As shown in Fig. 4.2, when user m successfully compete one slot τ, there could be a spurt of packets uploaded or downloaded by user m through channel n. Given the capacity of user m on channel n to be $c_{m,n}$, within one slot τ, the number of packets being transmitted would be $c_{m,n}\tau$. In this spurt of packets, the interval between any two consecutive packets is $1/c_{m,n}$. However, for the last packet in a spurt, the interval between itself and the next consecutive packet should be $x_{m,n}$. Therefore, based on the definition of mean, the mean interval time between the consecutive downloaded packets $\frac{1}{d_{m,n}}$ contributed by two parts: $(c_{m,n}\tau - 1)$ intervals with the value of $1/c_{m,n}$ and 1 interval with the value of $E\left[x_{m,n}\right]$, which is the expected interval between the current spurt and the next spurt. Then we have

$$\frac{1}{d_{m,n}} = \frac{1}{c_{m,n}} \cdot \frac{c_{m,n}\tau - 1}{c_{m,n}\tau} + E\left[x_{m,n}\right] \cdot \frac{1}{c_{m,n}\tau}. \tag{4.6}$$

From (4.5), the mean of $x_{m,n}$ is:

$$E\left[x_{m,n}\right] = \frac{\tau\left(1 - P_{\mathrm{suc},n}\right)}{P_{\mathrm{suc},n}} + \frac{1}{c_{m,n}}. \tag{4.7}$$

Substituting (4.7) into (4.6), we have

$$\frac{1}{d_{m,n}} = \frac{1}{P_{\mathrm{suc},n} c_{m,n}}. \tag{4.8}$$

The variance of the inter-arrival time of video packets over channel n to VoD user m is

$$v_{m,n} = \left(\frac{1}{c_{m,n}}\right)^2 \cdot \frac{c_{m,n}\tau - 1}{c_{m,n}\tau} + E\left[x_{m,n}^2\right] \cdot \frac{1}{c_{m,n}\tau} - \left(\frac{1}{d_{m,n}}\right)^2. \tag{4.9}$$

Then we have

$$v_{m,n} = \frac{c_{m,n}\tau\left(1 - P_{\mathrm{suc},n}\right)\left(2 - P_{\mathrm{suc},n}\right) - \left(1 - P_{\mathrm{suc},n}\right)^2}{P_{\mathrm{suc},n}^2 c_{m,n}^2}. \tag{4.10}$$

Given the channel allocation represented by $a_{m,n}$, the integrated mean of transmission rate and the variance of the inter-arrival time of the video packets of user m are, respectively,

$$\begin{aligned} d_m &= \sum_{n=1}^{N} a_{m,n} \cdot d_{m,n}, \\ v_m &= \sum_{n=1}^{N} a_{m,n} \cdot v_{m,n}. \end{aligned} \tag{4.11}$$

Substituting (4.11) into (3.5), we are able to evaluate the video performance of VoD users, and finally obtain the optimal channel allocation by solving (4.1).

4.3 Heuristic Algorithm

Since (4.1) is a nonlinear integer programming problem, it may be impossible to be solved for the real-time channel allocation, especially when the number of SUs scales to a large number. In this section, we propose a utility-based heuristic algorithm to determine the channel allocation by rendering SUs differentiated service according to their specific QoS requirements and contribution to the overall utility.

Our goal of the channel allocation is to maximize the overall system throughput while guarantee user specific QoS requirements at the same time. However, for VoD users, the QoS requirements, (*i.e.,* the smoothness of video playback, should be given priority compared with the contribution to the overall system throughput in the heuristic algorithm. This is because the video quality has a great influence on VoD users' satisfaction of the entire system; on the contrary, the downloading rate of the video packets are less important as long as the video is played smoothly. On the other hand, for BE users, the QoS requirements are less stringent, so we should pay more attention to improve the overall transmission throughput. From the above mentioned thoughts, our heuristic algorithm follows these rules:

1. VoD users have priority to be allocated over BE users.

2. VoD users with less buffer storage have priority to be allocated over the ones with more buffer storage.

3. Each VoD user chooses the channel on which the user has smaller playback frozen probability.

4. Each BE user chooses the channel on which the user has the larger throughput.

5. After a BE user's allocation, all the left BE users update their throughput on each channel.

6. Whenever a user is going to be allocated to a channel (candidate user), check if the first VoD user allocated to this channel, (*i.e.,* the VoD user with the smallest buffer storage in this channel, is harmed in terms of its QoS requirement. If so, we should lay this candidate user-channel allocation pair aside as the last choice, and see if the

channel on which the candidate user has larger playback frozen probability could be chosen.

7. If there are some users could not be allocated since all left choices would do harm to the afore-allocated VoD users, we increase the upper bound of the playback frozen probability.

Based on the above rule, we propose the heuristic algorithm as follows:

Fig. 4.3 and Fig. 4.4 show the pseudo code of the heuristic algorithm which is composed of two parts. As shown in Fig. 4.3, the first part is to allocate channels to VoD users subject to their QoS requirements. To this end, we first evaluate the urgency of download for VoD users according to their buffer storage, as in line 2, and then ranking channels in terms of the frozen probability of VoD users over each channel, as in lines 3-5. After that, we allocate channels to users according to the stringency of QoS requirements since line 6. From line 9 to line 19, each VoD user is allocated to the channel on which the user has the relatively small frozen probability; more importantly, the channel should be still available to be allocated to more VoD users without violating the QoS of users previously allocated to the channel. If no channels are available while some users still have not been allocated, we relax the QoS requirement of users by enlarging ϵ (line 21) and then continue the allocation.

After all the VoD users are accommodated, we allocate the channels to BE users in the second part of the algorithm, as shown in Fig. 4.4. As BE users have relatively loose QoS requirements, the emphasis is to enhance the system throughput. To this end, we first evaluate the download rates of BE users over each channel in lines 2-4 of part II, and then sort the rates in the descending order in line 5. The channel is allocated to a BE user from line 9 to line 19 if the user has comparatively large throughput over this channel and more importantly by inserting the BE user to this channel the QoS of VoD users specified in (3.7) are not violated. If there is no such allocation, we relax the QoS requirement of

Input: Profiles of SUs
Output: Channel allocation $a_{m,n}$ of SU m on channel n

I. VoD Users:

Define: k_n returns the first user ID in channel n;
Define: F_n returns the availability of channel; F_n is $TRUE$ if channel n is forbidden to be allocated to any more users, and otherwise, F_n is $FALSE$;
Define: L is a queue which stores a list of VoD users;
Define: P is a queue which stores the frozen probability of VoD users on different channels;

Set $F_n \leftarrow FALSE$ for all channel n; Set $k_n \leftarrow 0$ for all channel n; Insert all VoD users into L;
Sort L in ascending order according to VoD users' current storage of packets in their playout buffer;
for m **is** *the first* **to** *the last user in* L **do**
    Evaluate the frozen probability $p_{m,n}$ on channel n for all $n \in \{1, ..., N\}$; Sort $p_{m,n}$ in ascending order and insert sorted $p_{m,n}$ into a list P;
end
while P *is non-empty* **do**
    Set i to be the VoD user ID and j to be the channel ID which the first frozen probability $p_{i,j}$ in P associates with;
    if F_j *is FALSE* **then**
        if k_j *is* 0 **then**
            Set $k_j \leftarrow i$; Set $a_{i,j} \leftarrow 1$; Erase $p_{i,n}$ of user i for all channel $n \in \{1, ..., N\}$ in P;
        else
            Recalculate $p_{k_j,j}$ with $a_{i,j} = 1$;
            if $p_{k_j,j} \leq \epsilon$ **then**
                Set $a_{i,j} \leftarrow 1$; Erase $p_{i,n}$ of user i for all channel $n \in \{1, ..., N\}$ in P;
            else
                Set $F_j \leftarrow TRUE$;
                Move all $p_{m,j}$, $m \in \{1, ..., M\}$, to the end of queue P;
            end
        end
    else
        Set $\epsilon \leftarrow \epsilon + 0.01$;
        Reset $F_n \leftarrow FALSE$ for all channel $n \in \{1, ..., N\}$;
    end
end

Figure 4.3.: Proposed heuristic algorithm for VoD users allocation

II. BE Users:

Define: D is a queue which stores the download rate of BE users on different channels;

Set $F_n \leftarrow FALSE$ for all channel n;
for m **is** $M_{\text{VoD}} + 1$ **to** M **do**
  Calculate the download rate $d_{m,n}$ on channel n for all $n \in \{1, ..., N\}$, given the VoD users' allocation; Insert $d_{m,n}$ into a list D;
end
Sort D in the descending order;
while D *is non-empty* **do**
  Set i to be the BE user ID and j to be the channel ID which the first download rate $d_{i,j}$ in D associates with;
  if F_j *is FALSE* **then**
    if k_j *is* 0 **then**
      Set $a_{i,j} \leftarrow 1$; Erase $d_{i,n}$ of user i for all channel $n \in \{1, ..., N\}$ in D; Recalculate and sort the remaining $d_{m,n}$ in D with $F_n = FALSE$ in the descending order;
    else
      Recalculate $p_{k_j,j}$ with $a_{i,j} = 1$;
      if $p_{k_j,j} \leq \epsilon$ **then**
        Set $a_{i,j} \leftarrow 1$; Erase $d_{i,n}$ of user i for all channel $n \in \{1, ..., N\}$ in D; Recalculate and sort the remaining $d_{m,n}$ in D with $F_n = FALSE$ in the descending order;
      else
        Set $F_j \leftarrow TRUE$;
        Move all $d_{m,j}$, $m \in \{1, ..., M\}$, to the end of queue D;
      end
    end
  else
    Set $\epsilon \leftarrow \epsilon + 0.01$;
    Reset $F_n \leftarrow FALSE$ for all channel $n \in \{1, ..., N\}$;
    Recalculate remaining $d_{m,n}$;
    Sort D in the descending order;
  end
end

Figure 4.4.: Proposed heuristic algorithm for BE users allocation

VoD users by enlarging ϵ (line 21) and then continue the allocation following line 9 to line 19.

Chapter 5

Performance Evaluation

In this chapter, we evaluate the performance of the proposed CR VoD system using a discrete time, event-driven simulator coded in C++.

5.1 Simulation Setting

According to the discussion of the system model in Chapter 3, there exist two components in our simulation: channels and secondary users. For the former, we consider $N = 5$ orthogonal channels, which are described by two features: channel availabilities and channel capacity. For simplicity, we represent channel availabilities by the idle probabilities $\pi_{0,n}$, which are set as 0.3,0.4,0.5,0.6,0.7, respectively, with $n = 1, \ldots, 5$. And unless otherwise mentioned, the capacity of SUs on channels $c_{m,n}$ is uniformly selected in the range of $[1000, 2000]$ pkts/sec within each channel allocation epoch. For SUs, we consider $M = 50$ SUs contending for transmissions over all channels, and by default $M_{\text{VoD}} = 20$ SUs are VoD users. The download and upload of SUs is slotted at intervals of $\tau = 10$ ms and the channel allocation epoch L is 50. The tolerable frozen probability of VoD users ϵ is set to 0.05. The throughput performance of BE users is evaluated by the percentage of BE users which download at a rate larger than 30 pkts/sec.

For evaluation purpose, all the VoD users use the same VBR video trace "Aladdin" from [41] but starting from randomly selected sections of the video for playback. As such, the statistics of the video playback rate are the same for all VoD users with the mean rate $r = 30$ pkts/sec and variance $v_{\mathrm{r}} = 102$. The mean packet size is 630 Bytes for all users. In addition, each VoD user is associated with two system parameters: initial buffer storage and lifetime. The initial buffer storage represents the packet storages of VoD users at the start-up of their video playback when they join the system. It is fixed to 50 packets for all VoD users. The lifetime of each VoD user is uniformly distributed within the range of $[10, 30]$ seconds. Once its lifetime expires, a VoD user resets its playout buffer storage to the initial buffer storage, i.e., 50 packets, and selects a new lifetime, representing a new round of download. We simulate a dynamic network where VoD users dynamically join and depart from the network after watching the video clips. The lifetime of users is random as in real-world users may subscribe to videos of different length and they may quit in the middle of video playback.

5.2 Simulation Results

In this section, we present the extensive simulation results to demonstrate that the our proposed scheme satisfies the QoS requirements of both VoD users and BE users while achieving good overall throughput. In Section 5.2.1 and Section 5.2.2, we tune the channel capacity of users and the portion of VoD users, respectively, to evaluate the performance of the proposed heuristic algorithm when the network resource is surplus or deficient. We compare the heuristic algorithm with another two heuristic schemes, namely the random allocation and the greedy allocation. For the random allocation, the SUs are randomly allocated to each channel every allocation epoch. Using the greedy allocation, each SU is allocated to the channel with the largest throughput, evaluated by the product of channel capacity $c_{m,n}$ and channel idle probability $\pi_{0,n}$. Further, in Section 5.2.3, the relationship

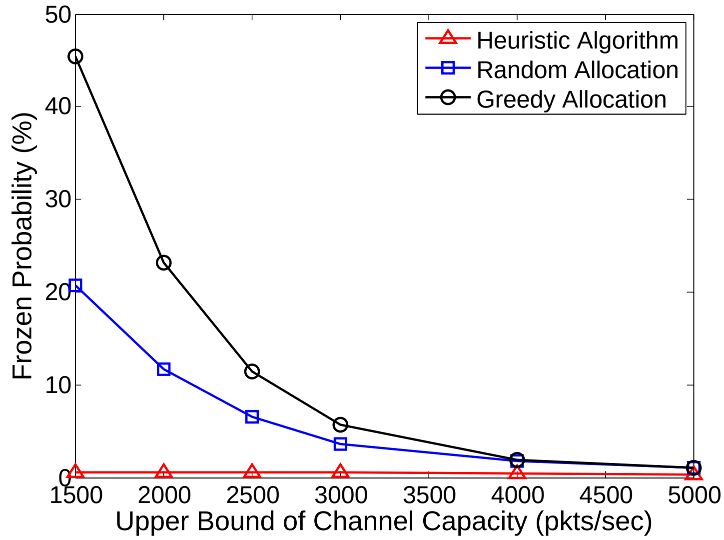

Figure 5.1.: Playback frozen probability of VoD users with the increased upper bound of channel capacity

between the length of channel allocation epoch T and the system performance is studied. For each scenario, we conduct 10 simulation runs with each run terminating at $t = 150$ seconds, and plot the mean results.

5.2.1 System Performance with Changing Channel Capacity

Fig. 5.1, Fig. 5.2 and Fig. 5.3 plot the performance of the users when enlarging the range of the channel capacity $c_{m,n}$ with increased upper bound and fixed lower bound.

As shown in Fig. 5.1, the VoD users using the heuristic algorithm have much lower playback frozen probability compared with the random and greedy allocations. With increased upper bounds of the channel capacity and hence enhanced mean capacity, the frozen probability of VoD users in both random and greedy allocations reduces dramatically; nevertheless, it is very stable for the proposed heuristic algorithm and is always lower than 1%. Fig. 5.2 plots the percentage of BE users in the network which have the throughput larger than 30 pkts/sec. It can be seen that the curves increase for all three allocation

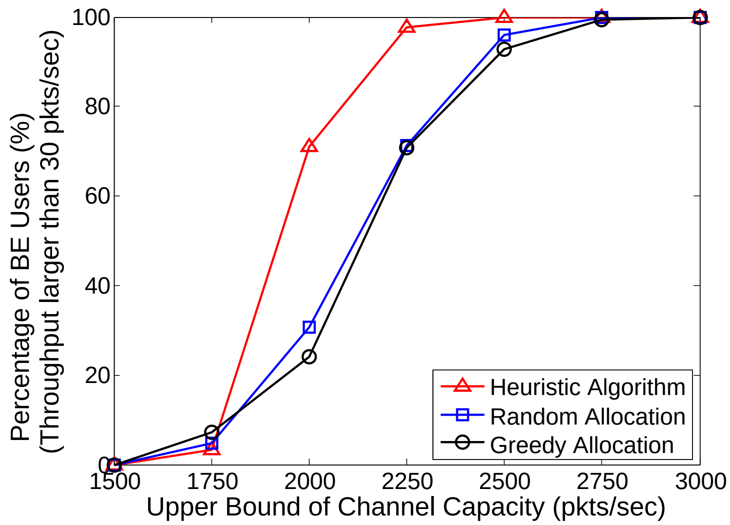

Figure 5.2.: Throughput of BE users with the increased upper bound of channel capacity

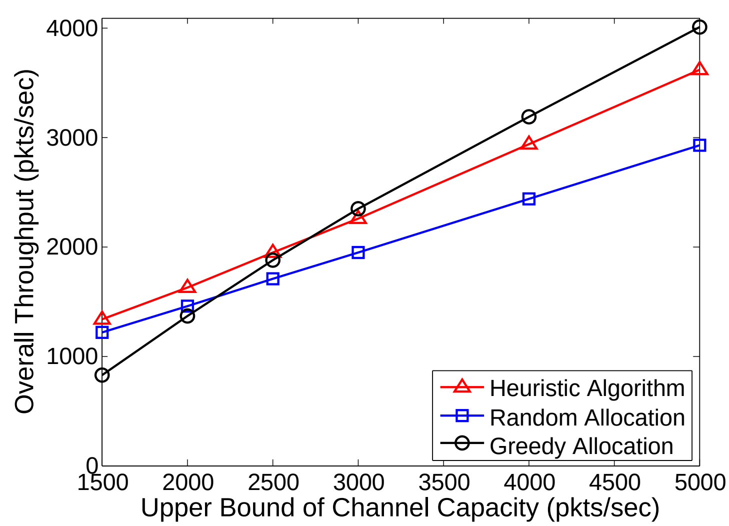

Figure 5.3.: Overall throughput with the increased upper bound of channel capacity

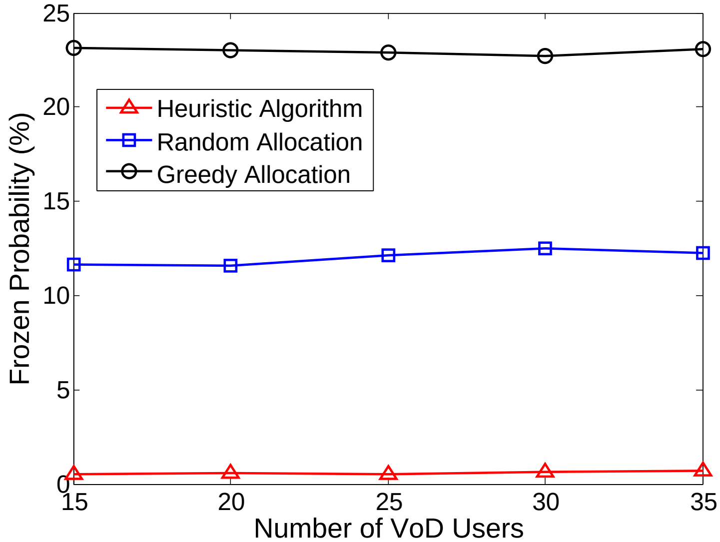

Figure 5.4.: Playback frozen probability of VoD users with the increased portion of VoD users and fixed overall population

schemes with enhanced capacity; however, the proposed heuristic algorithm performs the best, indicating that more BE users acquire the demanded throughput. Fig. 5.3 plots the resultant overall throughput. We can see that when the channel capacity increases, the heuristic algorithm outperforms the random allocation, but is worse than the greedy allocation. This is because the greedy allocation always assigns users to the channel with the best throughput performance. However, without catering to the specific QoS requirements of users, the greedy allocation tends to allocate a crowd of users to certain channels with the high availability, resulting in high collision probability to users. This leads to the poor performance in terms of the playback frozen probability of VoD users and the percentage of the satisfied BE users, as indicated in Fig. 5.1 and Fig. 5.2, respectively.

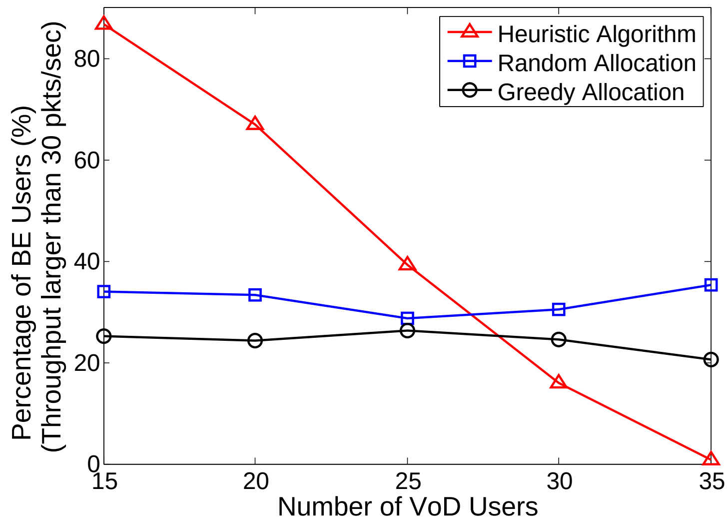

Figure 5.5.: Throughput of BE users with the increased portion of VoD users and fixed overall population

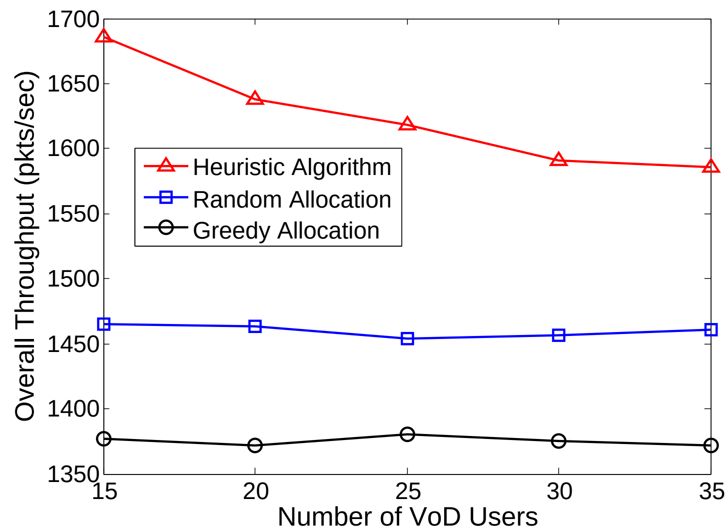

Figure 5.6.: Overall throughput with the increased portion of VoD users and fixed overall population

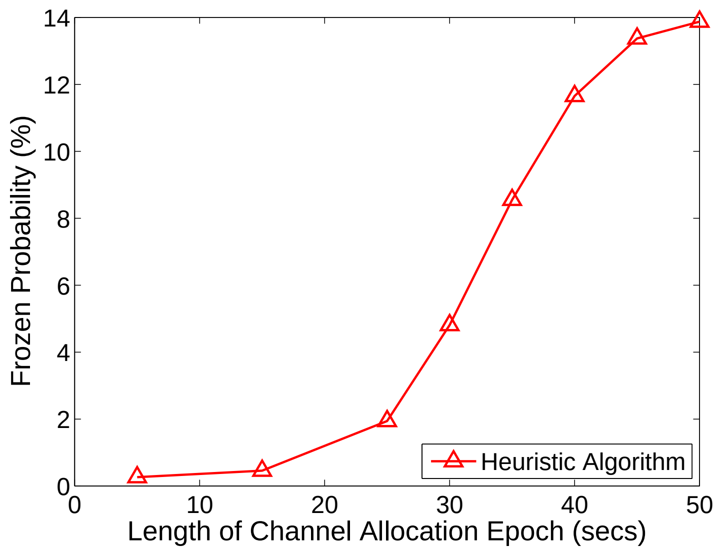

Figure 5.7.: Playback frozen probability of VoD users with the increased length of channel allocation epoch

5.2.2 System Performance with Changing the Portion of VoD Users

Fig. 5.4, Fig. 5.5 and Fig. 5.6 show the impacts of VoD user's population on the performance of the heuristic algorithm with the total number of users fixed to be 50. As shown in Fig. 5.4, the playback frozen probability of the heuristic algorithm remains stable when VoD user's population increases, while in Fig. 5.5, the throughput performance of BE users degrades when the network resource is deficient. The reason is that, the heuristic algorithm gives high priority to VoD users. The total throughput degrades slightly when VoD users' population increases, as shown in Fig. 5.6. In addition, the performance of random and greedy allocations remains the same because both random and greedy allocations do not differentiate VoD and BE users. Therefore, they are not sensitive to the change of VoD users' population.

5.2.3 System Performance with Changing Re-allocation Frequency

Fig. 5.7 plots the impact of channel allocation epoch T on the performance of the heuristic algorithm. We can see that the playback frozen probability increases as the channel allocation epoch T increases. This is because that frequent re-allocation would ensure the fairness in terms of VoD users' buffer storage, and consequently ensure only a small amount of buffers will be empty occasionally. In specific, the system could not provide all VoD users with sufficient bandwidth and superior download throughput simultaneously, which means, some of the VoD users have an increasing buffer storage and at the same time the others have a decreasing buffer storage. The allocation algorithm tends to allocate the VoD users with less buffer storage to a "good" channel, so that their buffer storage would increase and not be empty; and the allocation algorithm tends to allocate the VoD users with more buffer storage to a "poor" channel, since they have more tolerance to the network dynamics. If channel allocation epoch T is small enough, none of the buffers will stay in a "poor" channel for a long time to exhaust its buffer storage. However, larger channel allocation epoch requires less computational overhead, and makes the allocation process more efficient.

Chapter 6

Conclusions and Future Work

Due to the opportunistic use of channels, CR networks are extraordinarily dynamic. This directly threats to the smooth video playback of the users. In this dissertation, we propose to efficiently provision the network resource among SUs according to their specific video quality requirement. To this end, we develop an optimal channel allocation framework based on the cross-layer design. The proposed framework exploits the user diversity in terms of the tolerance to the network dynamics, and allocates the channels based on the required user specific video quality. As the optimal channel allocation is an integer programming problem which is impossible to be solved for the real-time communication, we proposed a heuristic algorithm to attain the sub-optimal solution which has a relatively low time complexity. Using extensive simulation evaluations, we have shown that proposed heuristic algorithm can outperform the conventional schemes with much reduced frozen probability and satisfactory throughput.

For the future work, we intend to further investigate the CR VoD system in two aspects:

- Suboptimal solution with guaranteed performance: while the proposed algorithm in the dissertation is practical and efficient based on the simulation evaluation, it can not quantitively guarantee the video performance of users. In the future, we intend to invoke the optimization theory in hunting for the schemes to provide the quantitively guaranteed video performance to users.

- Distributed channel allocation: the current framework is based on the centralized CR architecture where a central CR base station is available to carry out the channel allocation. However, such a system is not scalable when the network size increases to a large value. In the future, we intend to device the distributed channel allocation scheme by allowing SUs to distributedly select channels to transmit according to their QoS requirements. Without any central coordinations, the resulting channel allocation scheme would thus be more flexible and scalable.

Bibliography

[1] FCC, "ET docket no 02-135," Spectrum Poicy Take Force (SPTF) Report, November 2002.

[2] M. McHenry, "Spectrum white space measurements," Presented to New America Foundation Broad-Band Forum, June 2003.

[3] J. M. III and G. M. Jr, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[4] G. Staple and K. Werbach, "The end of spectrum scarcity," *IEEE Spectrum*, vol. 41, no. 3, pp. 48–52, 2004.

[5] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[6] P. J. Kolodzy, "Dynamic spectrum policies: promises and challenges," *CommLaw Conspectus*, vol. 12, p. 147, 2004.

[7] M. McHenry, "Frequency agile spectrum access technologies," in *FCC Workshop on Cognitive Radios*, vol. 19, no. 5, 2003.

[8] FCC, "ET docket no 03-222," Notice of Proposed Rule Making and Order, December 2003.

[9] I. F. Akyildiz, W. Y. Lee, M. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, 2008.

[10] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[11] A. Vogel, B. Kerherve, G. V. Bochmann, and J. Gecsei, "Distributed multimedia and QoS: A survey," *IEEE Multimedia*, vol. 2, no. 2, pp. 10–19, 1995.

[12] T. Little and D. Venkatesh, "Prospects for interactive video-on-demand," *IEEE Multimedia*, vol. 1, no. 3, 1994.

[13] V. O. K. Li, W. Liao, X. Qiu, and E. W. M. Wong, "Performance model of interactive video-on-demand systems," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 6, pp. 1099–1109, 1996.

[14] K. A. Hua, M. A. Tantaoui, and W. Tavanapong, "Video delivery technologies for large-scale deployment of multimedia applications," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1439–1451, 2004.

[15] E. L. Abram-Profeta and K. G. Shin, "Scheduling video programs in near video-on-demand systems," in *Proceedings of the Fifth ACM International Conference on Multimedia*, 1997, p. 369.

[16] C. C. Aggarwal, J. L. Wolf, and P. S. Yu, "The maximum factor queue length batching scheme for video-on-demand systems," *IEEE Transactions on Computers*, vol. 50, no. 2, p. 97, 2001.

[17] L. Gao, Z. L. Zhang, and D. Towsley, "Catching and selective catching: Efficient latency reduction techniques for delivering continuous multimedia streams," in *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, 1999, p. 206.

[18] Y. Cai and K. A. Hua, "Sharing multicast videos using patching streams," *Multimedia Tools and Applications*, vol. 21, no. 2, pp. 125–146, 2003.

[19] Y. Cai, W. Tavanapong, and K. A. Hua, "Enhancing patching performance through double patching," in *Proc. Int. Conf. Distributed Multimedia Systems.* Citeseer, 2003, pp. 72–77.

[20] D. Eager, M. Vernon, and J. Zahorjan, "Optimal and efficient merging schedules for video-on-demand servers," in *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, 1999, p. 202.

[21] K. A. Hua and S. Sheu, "Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems," *ACM SIGCOMM Computer Communication Review*, vol. 27, no. 4, pp. 89–100, 1997.

[22] L. Gao, J. Kurose, and D. Towsley, "Efficient schemes for broadcasting popular videos," *Multimedia Systems*, vol. 8, no. 4, pp. 284–294, 2002.

[23] A. Hu, "Video-on-demand broadcasting protocols: A comprehensive study," in *IEEE INFOCOM*, vol. 1. Citeseer, 2001, pp. 508–517.

[24] L. S. Juhn and L. M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Transactions on Broadcasting*, vol. 43, no. 3, pp. 268–271, 1997.

[25] L. Gao and D. Towsley, "Supplying instantaneous video-on-demand services using controlled multicast," in *Proc. IEEE International Conference on Multimedia Computing and Systems*, vol. 130. Citeseer, 1999.

[26] K. A. Hua, J. H. Oh, and K. Vu, "An adaptive video multicast scheme for varying workloads," *Multimedia Systems*, vol. 8, no. 4, pp. 258–269, 2002.

[27] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79–89, 2007.

[28] Q. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 785–796, 2008.

[29] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Transactions on Information Theory*, vol. 54, no. 5, p. 2053, 2008.

[30] T. Weingart, D. C. Sicker, and D. Grunwald, "A statistical method for reconfiguration of cognitive radios," *IEEE Wireless Communications*, vol. 14, no. 4, pp. 34–40, 2007.

[31] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 118–129, 2008.

[32] S. A. Jafar and S. Srinivasa, "Capacity limits of cognitive radio with distributed and dynamic spectral activity," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 529–537, 2007.

[33] A. Alshamrani, L. Xie, and X. S. Shen, "Adaptive admission control and channel allocation policy in cooperative ad hoc opportunistic spectrum networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1618–1629, 2010.

[34] D. Hu, S. Mao, and J. H. Reed, "On video multicast in cognitive radio networks," in *IEEE INFOCOM*, 2009.

[35] A. R. Fattahi, F. Fu, M. van der Schaar, and F. Paganini, "Mechanism-based resource allocation for multimedia transmission over spectrum agile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 601–612, 2007.

[36] D. Piazza, P. Cosman, L. B. Milstein, and G. Tartara, "A resource allocation algorithm for real-time streaming in cognitive networks," in *Proc. of IEEE WCNC*, 2009.

[37] H. P. Shiang and M. van der Schaar, "Dynamic channel selection for multi-user video streaming over cognitive radio networks," in *15th IEEE International Conference on Image Processing*, 2008, pp. 2316–2319.

[38] K. W. Choi, W. S. Jeon, and D. G. Jeong, "Resource allocation in OFDMA wireless communications systems supporting multimedia services," *IEEE/ACM Transactions on Networking*, vol. 17, no. 3, pp. 926–935, 2009.

[39] C. Cormio and K. R. Chowdhury, "A survey on MAC protocols for cognitive radio networks," *Ad Hoc Networks*, vol. 7, no. 7, pp. 1315–1329, 2009.

[40] T. H. Luan, L. X. Cai, and X. S. 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, 2010.

[41] 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, 2001.