COGNITIVE RADIO NETWORKS

Ian F. Akyildiz
Georgia Institute of Technology

Won-Yeol Lee
Georgia Institute of Technology

Kaushik R. Chowdhury
Northeastern University

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Cognitive Radio Networks - Theory and Applications
Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PART II SPECTRUM MANAGEMENT FOR COGNITIVE RADIO NETWORKS

INTRODUCTION

Today's wireless networks are regulated by a fixed spectrum assignment policy, i.e. the spectrum is regulated by governmental agencies and is assigned to license holders or services on a long term basis for large geographical regions. In addition, a large portion of the assigned spectrum is used sporadically. According to Federal Communications Commission (FCC) [72], temporal and geographical variations in the utilization of the assigned spectrum range from 15% to 85%. Although the fixed spectrum assignment policy generally served well in the past, there is a dramatic increase in the access to the limited spectrum for mobile services in the recent years. This increase is straining the effectiveness of the traditional spectrum policies.

The limited available spectrum and the inefficiency in the spectrum usage necessitate a new communication paradigm to exploit the existing wireless spectrum opportunistically [121]. Dynamic spectrum access is proposed to solve these current spectrum inefficiency problems. DARPA's approach on dynamic spectrum access network, the so-called NeXt Generation (xG) program aims to implement the policy based intelligent radios known as cognitive radios [65], [66].

Cognitive radio (CR), also known as Dynamic Spectrum Access Networks (DSANs), will provide high bandwidth to mobile users via heterogeneous wireless architectures and dynamic spectrum access techniques. The inefficient usage of the existing spectrum can be improved through opportunistic access to the licensed bands without interfering with the existing users. CR networks, however, impose several research challenges due to the broad range of available spectrum as well as diverse quality-of-service (QoS) requirements of applications. These heterogeneities must be captured and handled dynamically as mobile terminals roam between wireless architectures and along the available spectrum pool.

The key enabling technology of CR networks is the cognitive radio. Cognitive radio techniques provide the capability to use or share the spectrum in an opportunistic manner. Dynamic spectrum access techniques allow the cognitive radio to operate in the *best available channel*.

Once a cognitive radio supports the capability to select the best available channel, the next challenge is to make the network protocols adaptive to the available spectrum. Hence, new functionalities are required in a CR network to support this adaptivity. In summary, the main functions for cognitive radios in CR networks can be summarized as follows [6]:

- *Spectrum sensing:* Detecting unused spectrum and sharing the spectrum without harmful interference with other users.
- *Spectrum decision:* Capturing the best available spectrum to meet user communication requirements.
- *Spectrum sharing:* Providing the fair spectrum scheduling method among coexisting CR users.
- *Spectrum mobility:* Maintaining seamless communication requirements during the transition to better spectrum.

These functionalities of CR networks enable spectrum-aware communication protocols. However, the dynamic use of the spectrum causes adverse effects on the performance of conventional communication protocols, which were developed considering a fixed frequency band for communication. So far, networking in CR networks is an unexplored topic. In this book, we also capture the intrinsic challenges for networking in CR networks and lay out guidelines for further research in this area. More specifically, we overview the recent proposals for spectrum sharing and routing in CR networks as well as the challenges for transport protocols. Moreover, the effect of cross-layer design is addressed for communication in CR networks.

This book presents a definition, functions and current research challenges of the CR networks. In Chapter 1, we provide a brief overview of the cognitive radio technology. The CR network architectures on licensed band and on unlicensed band are presented in Chapter 2. Chapter 3 presents a novel spectrum management framework along with its research challenges, which is necessary to realize efficient and reliable communications in CR networks. In Chapters 4, 5, 6, and 7, we explain the existing work and challenges in spectrum sensing, spectrum decision, spectrum sharing, spectrum mobility, respectively. In Chapters 8, 9, 10, and 11, we investigate how CR features influence the performance of the upper layer protocols, i.e., medium access control (MAC), routing, transport, and security respectively. Finally, we explain the current effort on CR standardization in Chapter 12.

PART I

COGNITIVE RADIO NETWORKS : OVERVIEW

CHAPTER 1

INTRODUCTION TO COGNITIVE RADIO

Today's wireless networks are regulated by a fixed spectrum assignment policy, i.e. the spectrum is regulated by governmental agencies and is assigned to license holders or services on a long term basis for large geographical regions. In addition, a large portion of the assigned spectrum is used sporadically as illustrated in Figure 1.1, where the signal strength distribution over a large portion of the wireless spectrum is shown. The spectrum usage is concentrated on certain portions of the spectrum while a significant amount of the spectrum remains unutilized. According to Federal Communications Commission (FCC) [72], temporal and geographical variations in the utilization of the assigned spectrum range from 15% to 85%. Although the fixed spectrum assignment policy generally served well in the past, there is a dramatic increase in the access to the limited spectrum for mobile services in the recent years. This increase is straining the effectiveness of the traditional spectrum policies. The limited available spectrum and the inefficiency in the spectrum usage necessitate a new communication paradigm to exploit the existing wireless spectrum opportunistically [121]. In order to address the critical problem of spectrum scarcity, the FCC has recently approved the use of unlicensed devices in licensed bands. Consequently, dynamic spectrum access (DSA) techniques are proposed to solve these current spectrum inefficiency problems. This new area of research foresees the development of cognitive radio (CR) networks to further improve spectrum efficiency. The basic idea of CR networks is that the unlicensed devices (also called cognitive radio users or secondary users) need to vacate the band once the licensed device (also known as a primary user) is detected. CR networks, however, impose unique challenges due to the high fluctuation in the available

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

spectrum as well as diverse qualityof- service (QoS) requirements.

To provide a better understanding of CR networks, this chapter provides the overview of cognitive radio, mainly focusing on spectrum usage models, basic functionalities, physical layer requirements, and possible applications.

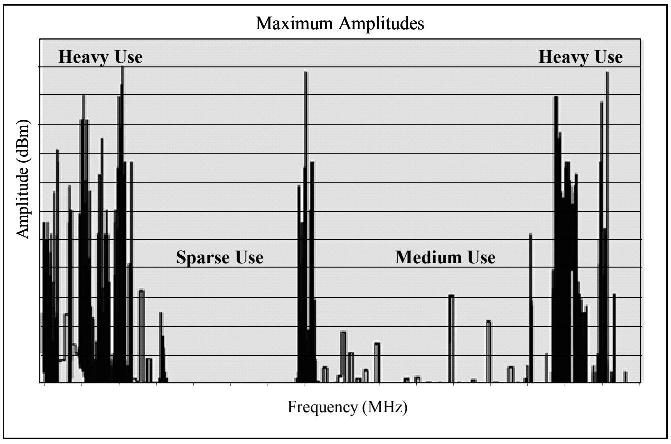


Figure 1.1 Spectrum utilization.

1.1 SPECTRUM USAGE MODELS

Since spectrum is a finite but reusable resources, it can be distributed to multiple users through the spectrum usage model. There are three general models for assigning spectrum usage rights as follows:

- *Commands and control model:* The traditional process of spectrum management in the United States, currently used for most spectrum within the Commission's jurisdiction, allocates and assigns frequencies to limited categories of spectrum users for specific government-defined uses. Service rules for the band specify eligibility and service restrictions, power limits, build-out requirements, and other rules.
- *Exclusive use model:* A licensee has exclusive and transferable rights to the use of specified spectrum within a defined geographic area, with flexible use rights that are governed primarily by technical rules to protect spectrum users against interference. Under this model, exclusive rights resemble property rights in spectrum, but this model does not imply or require creation of "full" private property rights in spectrum.
- *Commons model:* This model allows unlimited numbers of unlicensed users to share frequencies, with usage rights that are governed by technical standards or etiquettes but with no right to protection from interference. Spectrum is available to all users that comply with established technical "etiquettes" or standards that set power limits and other criteria for operation of unlicensed devices to mitigate potential interference.

As explained above, the main drawbacks of commands and control model, which is a legacy model for current wireless networks are that it is very slow to adapt, it is unfriendly to commercial interests, and it results in inefficient use of spectrum.

Thus, there has been growing interest in making more efficient use of spectrum by shifting from the conventional "command-and-control" spectrum usage model to more flexible "Exclusive Use" and "Commons" models.

1.2 COGNITIVE RADIO

In both exclusive use and the commons models, a basic question is how to share the available spectrum efficiently and fairly. Cognitive radio is the key technology that enables a CR network to share spectrum in a dynamic manner. The term, cognitive radio, can formally be defined as follows [72]:

> *A "Cognitive Radio" is a radio that can change its transmitter parameters based on interaction with the environment in which it operates.*

From this definition, two main characteristics of the cognitive radio can be defined [111], [210]:

- **Cognitive capability:** Cognitive capability refers to the ability of the radio technology to capture or sense the information from its radio environment. This capability cannot simply be realized by monitoring the power in some frequency band of interest but more sophisticated techniques are required in order to capture the temporal and spatial variations in the radio environment and avoid interference to other users. Through this capability, the portions of the spectrum that are unused at a specific time or location can be identified. Consequently, the best spectrum and appropriate operating parameters can be selected.

- **Reconfigurability:** The cognitive capability provides spectrum awareness whereas reconfigurability enables the radio to be dynamically programmed according to the radio environment. More specifically, the cognitive radio can be programmed to transmit and receive on a variety of frequencies and to use different transmission access technologies supported by its hardware design [136].

Figure 1.2 depicts how the cognitive radio concept can be realized through cognitive capability and reconfigurability. First, the cognitive radio identifies radio information through observation and learning processes and makes proper decisions accordingly. Based on these decisions, the cognitive radio reconfigures its software (e.g., communication protocols) and hardware (e.g., an RF front-end and an antenna).

The cognitive radio concept was first introduced in [124], [125], where the main focus was on the radio knowledge representation language (RKRL) and how the cognitive radio can enhance the flexibility of personal wireless services. The cognitive radio is regarded as a small part of the physical world to use and provide information from environment.

The ultimate objective of the cognitive radio is to obtain the best available spectrum through cognitive capability and reconfigurability as described before. Since most of the spectrum is already assigned, the most important challenge is to share the licensed spectrum without interfering with the transmission of other licensed users as illustrated in Figure 1.3. The cognitive radio enables the usage of temporally unused spectrum, which is referred to as *spectrum hole* or *white space* [111]. If this band is further used by a licensed user, the cognitive radio moves to another spectrum hole or stays in the same band, altering its transmission power level or modulation scheme to avoid interference as shown in Figure 1.3.

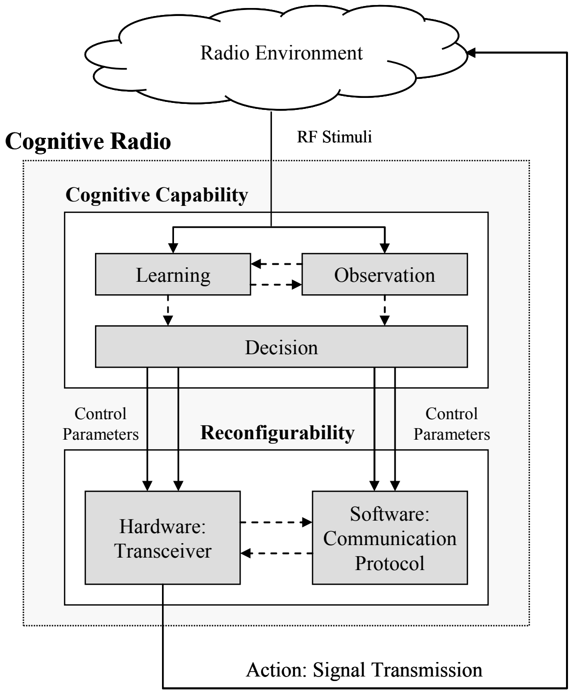


Figure 1.2 Cognitive radio concept.

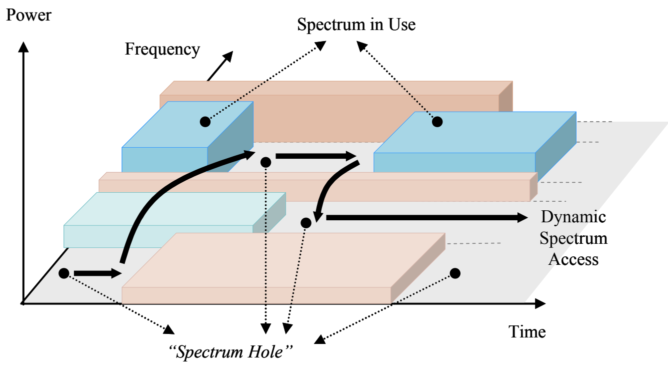


Figure 1.3 Spectrum hole concept.

1.2.1 Cognitive Capability

The cognitive capability of a cognitive radio enables real time interaction with its environment to determine appropriate communication parameters and adapt to the dynamic radio environment. The tasks required for adaptive operation in open spectrum are shown in Figure 3.1 [111], [125], [210], which is referred to as the *cognitive cycle*. In a cognition cycle, a radio receives stimuli (i.e., information about its operating environment) from the external environment through direct observation or signaling information exchanges and then reacts to these stimuli in four steps: orientation, planning, decision and acting. The radiofls action will influence the outside world and will in turn be received as stimuli by the other radios.

The outside world provides stimuli. Cognitive radio parses these stimuli to recognize the context of its communications tasks. Incoming and outgoing multimedia content is parsed for the contextual cues necessary to infer the communications context (e.g., urgency). Thus, for example, the radio may infer that it is going for a taxi ride (with some probability) if the user ordered a taxi by voice and is located in a foreign country. The Orient-stage decides on the urgency of the communications in part from these cues in order to reduce the burden on the user. Normally, the Plan-stage generates and evaluates alternatives, including expressing plans to peers and/or the network to obtain advice. The Decide stage allocates computational and radio resources to subordinate (conventional radio) software. The Act-stage initiates tasks with specified resources for specified amounts of time. Cognitive radio also includes some forms of supervised and unsupervised machine learning.

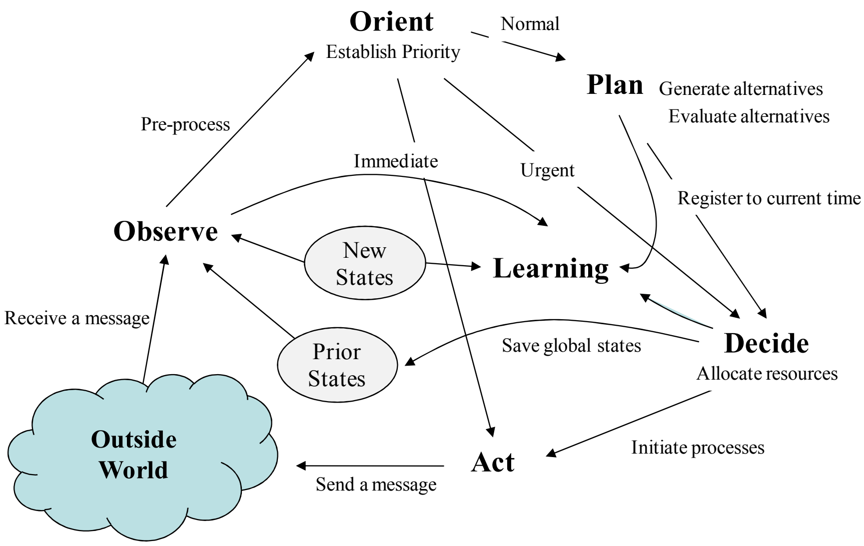


Figure 1.4 Cognitive Cycle [124].

.

1.2.2 Reconfigurability

Reconfigurability is the capability of adjusting operating parameters for the transmission on the fly without any modifications on the hardware components. This capability enables the cognitive radio to adapt easily to the dynamic radio environment. There are several reconfigurable parameters that can be incorporated into the cognitive radio [72] as explained below:

- *Operating Frequency:* A cognitive radio is capable of changing the operating frequency. Based on the information about the radio environment, the most suitable operating frequency can be determined and the communication can be dynamically performed on this appropriate operating frequency.

- *Modulation:* A cognitive radio should reconfigure the modulation scheme adaptive to the user requirements and channel conditions. For example, in the case of delay sensitive applications, the data rate is more important than the error rate. Thus, the modulation scheme that enables the higher spectral efficiency should be selected. Conversely, the loss-sensitive applications focus on the error rate, which necessitate modulation schemes with low bit error rate.

- *Transmission power:* Transmission power can be reconfigured within the power constraints. Power control enables dynamic transmission power configuration within the permissible power limit. If higher power operation is not necessary, the cognitive radio reduces the transmitter power to a lower level to allow more users to share the spectrum and to decrease the interference.

- *Communication technology:* A cognitive radio can also be used to provide interoperability among different communication systems.

The transmission parameters of a cognitive radio can be reconfigured not only at the beginning of a transmission but also during the transmission. According to the spectrum characteristics, these parameters can be reconfigured such that the cognitive radio is switched to a different spectrum band, the transmitter and receiver parameters are reconfigured and the appropriate communication protocol parameters and modulation schemes are used.

1.3 PHYSICAL ARCHITECTURE OF COGNITIVE RADIO

A generic architecture of a cognitive radio transceiver is shown in Figure 1.5 [136]. The main components of a cognitive radio transceiver are the radio front-end and the baseband processing unit. Each component can be reconfigured via a control bus to adapt to the time-varying RF environment. In the RF front-end, the received signal is amplified, mixed and A/D converted. In the baseband processing unit, the signal is modulated/demodulated and encoded/decoded. The baseband processing unit of a cognitive radio is essentially similar to existing transceivers. However, the novelty of the cognitive radio is the RF front-end. Hence, next, we focus on the RF front-end of the cognitive radios.

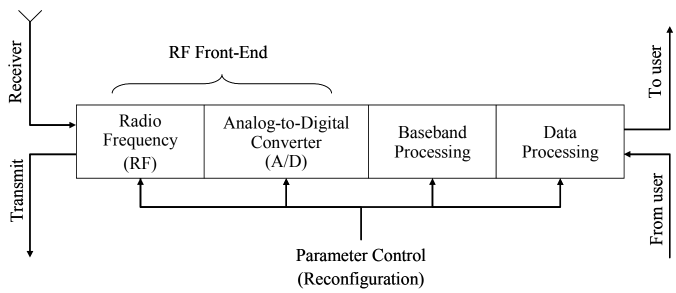


Figure 1.5 Cognitive radio transceiver [136].

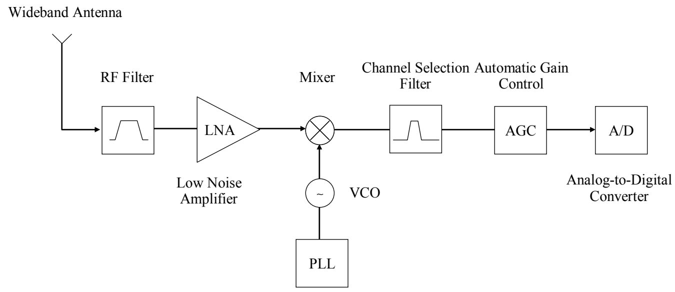


Figure 1.6 RF fornt-end for cognitive radio [31].

1.3.1 Physical Layer Technologies for Cognitive Radio

The transmission scheme should allow assignments of any frequency band to any cognitive user, and should be scalable with the number of users and bands. In order to keep the cognitive receiver demodulator fairly simple, it is desirable to restrict a single user transmission in a single frequency band. This constraint could be further justified by reduced transmission power of a single user rather than additive transmission power of many users, which would potentially cause interference to the active primary user in the vicinity. The modulation scheme based on orthogonal frequency division multiplexing (OFDM) is a natural approach that might satisfy desired properties [29]. OFDM has become the modulation of choice in many broadband systems due to its inherent multiple access mechanism and simplicity in channel equalization, plus benefits of frequency diversity and coding.

The transmitted OFDM waveform is generated by applying an inverse fast Fourier transform (IFFT) on a vector of data, where number of points N determines the number of sub-carriers for independent channel use, and minimum resolution channel bandwidth is determined by W/N, where W is the entire frequency band accessible by any cognitive user. The frequency domain characteristics of the transmitted signal are determined by the assignment of non-zero data to IFFT inputs corresponding to sub-carriers to be used by a particular cognitive user. Similarly, the assignment of zeros corresponds to channels not permitted to use due to primary user presence or channels used by other cognitive users. The output of the IFFT processor contains N samples that are passed through a digital-to-analog converter producing the wideband waveform of bandwidth W. A great advantage of this approach is that the entire wideband signal generation is performed in the digital domain, instead of multiple filters and synthesizers required for the signal processing in analog domain [29].

From the cognitive network perspective, OFDM spectrum access is scalable while keeping users orthogonal and non-interfering provided the synchronized channel access. However, this conventional OFDM scheme does not provide truly band-limited signals due to spectral leakage caused by sinc-pulse shaped transmission resulted from the IFFT operation. The slow decay of the sinc-pulse waveform, with first sidelobe attenuated by only 13.6dB, produces interference to the adjacent band primary users which is proportional to the power allocated to the cognitive user on the corresponding adjacent sub-carrier. Therefore, a conventional OFDM access scheme is not an acceptable candidate for wideband cognitive radio transmission. However, there are techniques that the resolve this problem by use of techniques like guard bands, power control, canceling carriers and window-based systems. In [29] a more detailed treatment of these techniques is provided.

1.3.2 Radio Frequency (RF) Front-End for Cognitive Radio

The novel characteristic of cognitive radio transceiver is a wideband sensing capability of the RF front-end. This function is mainly related to RF hardware technologies such as wideband antenna, power amplifier, and adaptive filter. RF hardware for the cognitive radio should be capable of tuning to any part of a large range of frequency spectrum. Also such spectrum sensing enables real-time measurements of spectrum information from radio environment. Generally, a wideband front-end architecture for the cognitive radio has the following structure as shown in Figure 1.6 [31]. The components of a cognitive radio RF front-end are as follows:

- *RF filter:* The RF filter selects the desired band by bandpass filtering the received RF signal.
- *Low noise amplifier (LNA):* The LNA amplifies the desired signal while simultaneously minimizing noise component.
- *Mixer:* In the mixer, the received signal is mixed with locally generated RF frequency and converted to the baseband or the intermediate frequency (IF).
- *Voltage-controlled oscillator (VCO):* The VCO generates a signal at a specific frequency for a given voltage to mix with the incoming signal. This procedure converts the incoming signal to baseband or an intermediate frequency.
- *Phase locked loop (PLL):* The PLL ensures that a signal is locked on a specific frequency and can also be used to generate precise frequencies with fine resolution.
- *Channel selection filter:* The channel selection filter is used to select the desired channel and to reject the adjacent channels. There are two types of channel selection filters [215]. The *direct conversion receiver* uses a low-pass filter for the channel selection. On the other hand, the *superheterodyne receiver* adopts a bandpass filter.
- *Automatic gain control (AGC):* The AGC maintains the gain or output power level of an amplifier constant over a wide range of input signal levels.

In this architecture, a wideband signal is received through the RF front-end, sampled by the high speed analog-to-digital (A/D) converter, and measurements are performed for the detection of the licensed user signal. However, there exist some limitations on developing the cognitive radio front-end. The wideband RF antenna receives signals from various transmitters operating at different power levels, bandwidths, and locations. As a result, the RF front-end should have the capability to detect a weak signal in a large dynamic range. However, this capability requires a multi-GHz speed A/D converter with high resolution, which might be infeasible [31], [29].

The requirement of a multi-GHz speed A/D converter necessitates the dynamic range of the signal to be reduced before A/D conversion. This reduction can be achieved by filtering strong signals. Since strong signals can be located anywhere in the wide spectrum range, tunable notch filters are required for the reduction [31]. Another approach is to use multiple antennas such that signal filtering is performed in the spatial domain rather than in the frequency domain. Multiple antennas can receive signals selectively using beamforming techniques [29].

As explained previously, the key challenge of the physical architecture of the cognitive radio is an accurate detection of weak signals of licensed users over a wide spectrum range. Hence, the implementation of RF wideband front-end and A/D converter are critical issues in CR networks.

1.4 COGNITIVE RADIO APPLICATIONS

Before entering into a deeper technical discussion on cognitive radios, we investigate the possible scenarios to deploy CR networks in the following:

LEASED NETWORK

The primary network can provide a leased network by allowing opportunistic access to its licensed spectrum with the agreement with a third party without sacrificing the service quality of the primary user [241]. For example, the primary network can lease its spectrum access right to a mobile virtual network operator (MVNO). Also the primary network can provide its spectrum access rights to a regional community for the purpose of broadband access.

COGNITIVE MESH NETWORK

Wireless mesh networks are emerging as a cost-effective technology for providing broadband connectivity [7]. However, as the network density increases and the applications require higher throughput, mesh networks require higher capacity to meet the requirements of the applications. Since the cognitive radio technology enables the access to larger amount of spectrum, CR networks can be used for mesh networks that will be deployed in dense urban areas with the possibility of significant contention [148]. For example, the coverage area of CR networks can be increased when a meshed wireless backbone network of infrastructure links is established based on cognitive access points (CAPs) and fixed cognitive relay nodes (CRNs) [11]. The capacity of a CAP, connected via a wired broadband access to the Internet, is distributed into a large area with the help of a fixed CRN. CR networks have the ability to add temporary or permanent spectrum to the infrastructure links used for relaying in case of high traffic load.

EMERGENCY NETWORK

Public safety and emergency networks are another area in which CR networks can be implemented [174]. In the case of natural disasters, which may temporarily disable or destroy existing communication infrastructure, emergency personnel working in the disaster areas need to establish *emergency networks*. Since emergency networks deal with the critical information, reliable communication should be guaranteed with minimum latency. In addition, emergency communication requires a significant amount of radio spectrum for handling huge volume of traffic including voice, video and data. CR networks can enable the usage of the existing spectrum without the need for an infrastructure and by maintaining communication priority and response time.

MILITARY NETWORK

One of the most interesting potential applications of an CR network is in a military radio environment [186]. CR networks can enable the military radios choose arbitrary, intermediate frequency (IF) bandwidth, modulation schemes, and coding schemes, adapting to the variable radio environment of battlefield. Also military networks have a strong need for security and protection of the communication in hostile environment. CR networks could allow military personnel to perform spectrum handoff to find secure spectrum band for themselves and their allies.

COGNITIVE RADIO FEMTOCELLS

Femtocells, also known as home base stations, are consumer-installed, low-power, short-range access points used for increased indoor cellular coverage to provide high data rates for cellular users. Femtocells are usually connected with the macrocell (BS) through a broadband wired connection such as cable or DSL line, or through a wireless back haul link. The most critical issue in femtocell is to determine a spectrum sharing strategy to share the spectrum allocated to its corresponding macrocell between an orthogonal basis where the femtocell and the macrocell share different sections of the allocated spectrum to the macrocell and a nonorthogonal basis where the femtocell reuses the spectrum allocated to its macrocell. The obvious tradeoff between these two strategies is increased cell capacity versus increased interference between the macrocell and femtocells, and among the different femtocells.

In this regard, femtocells necessitate the CR technology to use the spectrum to the macrocells in a spectrally efficient way of utilizing the macrocell allocated spectrum [205]. Cooperation between the different femtocells within a certain macrocell is vital for interference avoidance and efficient resource allocation. Since the transmission range of CR femtocells is limited, cooperative relays for self-coexistence of multiple CR femtocells is required for effective cooperative spectrum sharing. For example, two nearby femtocells might observe different channel availability in a macrocell allocated spectrum. In order for these two femtocells to optimally utilize these vacant channels, they need to use an interference-aware resource allocation scheme in order to avoid interference to each other and to the macrocell. They can also rely on cooperative relays to further improve spectrum utilization with other femtocells outside their transmission range.

VEHICULAR COGNITIVE RADIO NETWORKS

Vehicular communication networks have received wide attention in the past decade as a way to support interesting applications such as driving safety, accident avoidance, and in-car infotainment, among others. Here we consider vehicular CR networks, where vehicle-to-vehicle (V2V) and vehicle-to-roadside (V2R) infrastructure communications are carried out opportunistically over some vacant spectrum. Due to high mobility of vehicles, vehicular CR networks face more challenges compared to fixed CR networks. Vehicles and roadside infrastructure are used as cooperative relays for sharing heterogeneous spectrum. A resource allocation scheme is needed for vehicles to optimally share rapid switching channels as the result of high mobility.

CHAPTER 2

COGNITIVE RADIO NETWORK ARCHITECTURE

Existing wireless network architectures employ heterogeneity in terms of both spectrum policies and communication technologies [121]. Moreover, some portion of the wireless spectrum is already licensed to different purposes while some bands remain unlicensed. For the development of communication protocols, a clear description of the CR network architecture is essential. In this chapter, the CR network architecture is presented such that all possible scenarios are considered.

2.1 BASIC ARCHITECTURE

Figure 2.1 shows the reference architecture of CR networks, where CR networks are operated under the mixed spectrum environment that consists of both licensed and unlicensed bands. Also, CR users can either communicate with each other in a multi-hop manner or access the base-station. The components of the CR network architecture, as shown in Figure 2.1, can be classified in two groups as *primary networks* and *CR networks*.

The *primary network* (or *licensed network*) is referred to as the legacy network that has an exclusive right to a certain spectrum band. Examples include the common cellular and TV broadcast networks. On the contrary, the *CR network* (or *secondary network, unlicensed network, dynamic spectrum access network*) does not have a license to operate in the desired band. Hence, the spectrum access is allowed only in an opportunistic manner. The followings are the basic components of primary networks [6]:

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

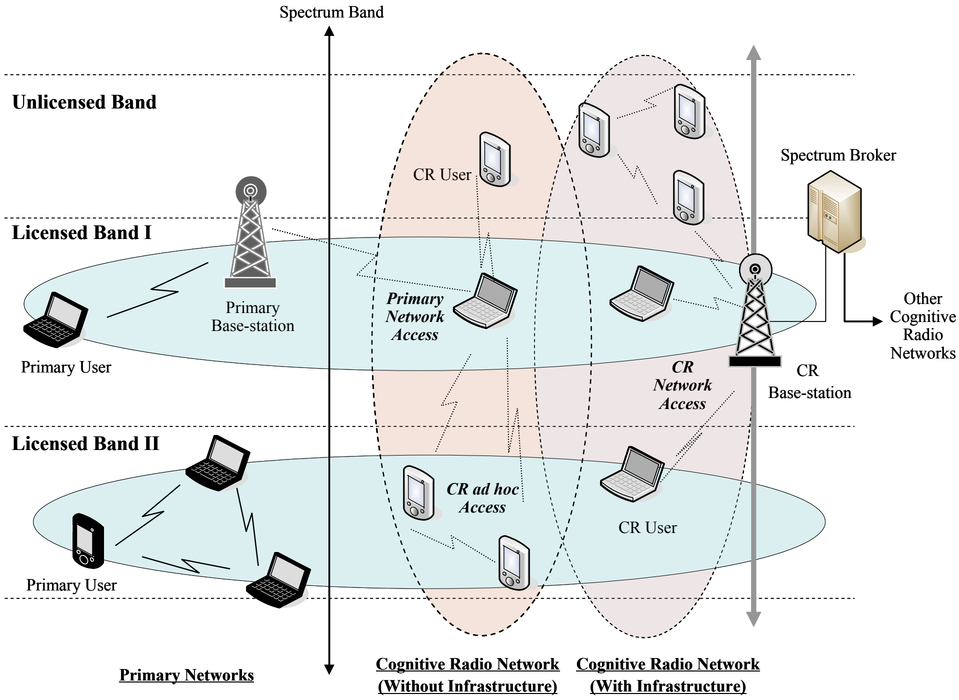


Figure 2.1 CR network architecture.

- *Primary user:* A primary user (or licensed user) has a license to operate in a certain spectrum band. This access can only be controlled by the primary base-station and should not be affected by the operations of any other unlicensed users. Primary users do not need any modification or additional functions for coexistence with CR base-stations and CR users.
- *Primary base-station:* A primary base-station is the fixed infrastructure network component which has a spectrum license such as base station transceiver system (BTS) in a cellular system. In principle, the primary base-station does not have any CR capability for sharing spectrum with CR users. However, the primary base-station may be requested to have both legacy and CR protocols for the *primary network access* of CR users, which is explained below.

The basic elements of the CR network are defined as follows [6]:

- *CR user:* A CR user (or unlicensed user, secondary user, dynamic spectrum access user) has no spectrum license. Hence, additional functionalities are required to share the licensed spectrum band.
- *CR base-station:* A CR base-station is the fixed infrastructure component with CR capabilities. The CR base-station provides single hop connection to CR users without spectrum access license. Through this connection, a CR user can access other networks.
- *Spectrum Broker:* A spectrum broker (or spectrum manager) is the central network entity that plays a role in sharing the spectrum resources among different CR net-

works. The spectrum broker can be connected to each network and can serve as a spectrum information manager to enable coexistence of multiple CR networks [21], [122], [284].

As shown in Figure 2.1, the CR users have the opportunity to perform three different access types as explained next:

- *CR Network Access:* CR users can access their own CR base-station both on licensed and unlicensed spectrum bands. Since all interactions occur inside the CR network, their spectrum sharing policy can be independent of that of the primary network.
- *CR Ad-hoc Access:* CR users can communicate with other CR users throug8h ad-hoc connection on both licensed and unlicensed spectrum bands.
- *Primary Network Access:* CR users can also access the primary base-station through the licensed band. Unlike other access types, CR users need an adaptive MAC protocol, which enables roaming over the multiple primary networks with different access technologies.

According to the reference architecture shown in Figure 2.1, various functionalities are required to support the heterogeneity in CR networks. In Sections 2.2 and 2.3, we describe the CR network functions to support the heterogeneities of the network environment, and the spectrum environment, respectively.

2.2 NETWORK HETEROGENEITY

2.2.1 Infrastructure-based (Centralized) Network

In this architecture, some powerful entity such as base-station exerts ownership and control over the nodes within its range. The observations and analysis performed by each CR user feeds to the central CR base-station so that decisions can be made by the base-station on how to avoid interfering with primary networks.

2.2.2 Ad-hoc (Distributed) Network

CR AH HOC NETWORKS (CRAHNS)

CR Ad-hoc networks (CRAHNs) do not have a central network entity such as a base-station or an access point. Thus, each CR user should have all functionalities for dynamic spectrum access. In this architecture, these functionalities are executed either in a non-cooperative or in a cooperative manner. In case of the non-cooperation, CR users perform functions independently through their local observations. In the cooperative method, information from multiple CR users are incorporated for dynamic spectrum access through the exchange of observations among CR users.

The changing spectrum environment and the importance of protecting the transmission of the licensed users of the spectrum mainly differentiate classical ad hoc networks from CRAHNs. The following are these unique features of CRAHNs compared to classical ad hoc networks:

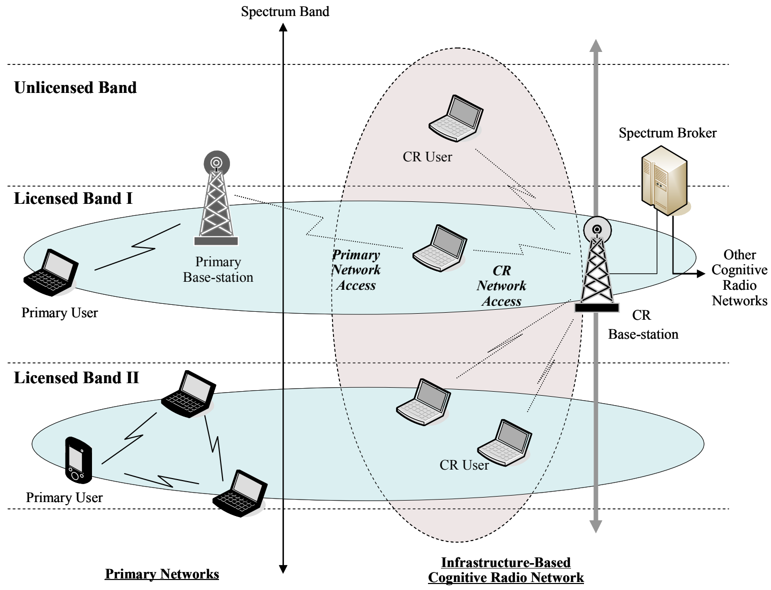


Figure 2.2 Infrastructure-based CR network architecture.

- *Choice of Transmission Spectrum:* In CRAHNs, the available spectrum bands are distributed over a wide frequency range, which vary over time and space. Thus, each user shows different spectrum availability according to the primary user (PU) activity. As opposed to this, classical ad hoc networks generally operate on a pre-decided channel that remains unchanged with time. For the ad hoc networks with multi-channel support, *all* the channels are continuously available for transmission, though nodes may select few of the latter from this set based on self-interference constraints. A key distinguishing factor is the primary consideration of protecting the PU transmission, which is entirely missing in classical ad hoc networks.

- *Topology Control:* Ad hoc networks lack centralized support, and hence must rely on local coordination to gather topology information. In classical ad hoc networks, this is easily accomplished by periodic beacon messages on the channel. However, in CRAHNs, as the licensed spectrum opportunity exists over large range of frequencies, sending beacons over all the possible channels is not feasible. Thus, CRAHNs are highly probable to have incomplete topology information, which leads in an increase in collisions among CR users as well as interference to the PUs.

- *Multi-hop/Multi-Spectrum Transmission:* The end-to-end route in the CRAHN consists of multiple hops having different channels according to the spectrum availability. Thus, CRAHNs require collaboration between routing and spectrum allocation in establishing these routes. Moreover, the spectrum switches on the links are frequent based on PU arrivals. As opposed to classical ad hoc networks, maintaining end-to-end QoS involves not only the traffic load, but also how many different channels and possibly spectrum bands are used in the path, the number of PU induced spectrum change events, consideration of periodic spectrum sensing functions, among others.

- *Distinguishing Mobility from PU Activity:* In classical ad hoc networks, routes formed over multiple hops may periodically experience disconnections caused by node mobility. These cases may be detected when the next hop node in the path does not reply to messages and the retry limit is exceeded at the link layer. However, in CRAHNs, a node may not be able to transmit immediately if it detects the presence of a PU on the spectrum, even in the absence of mobility. Thus, correctly inferring mobility conditions and initiating the appropriate recovery mechanism in CRAHNs necessitate a different approach from from the classical ad hoc networks.

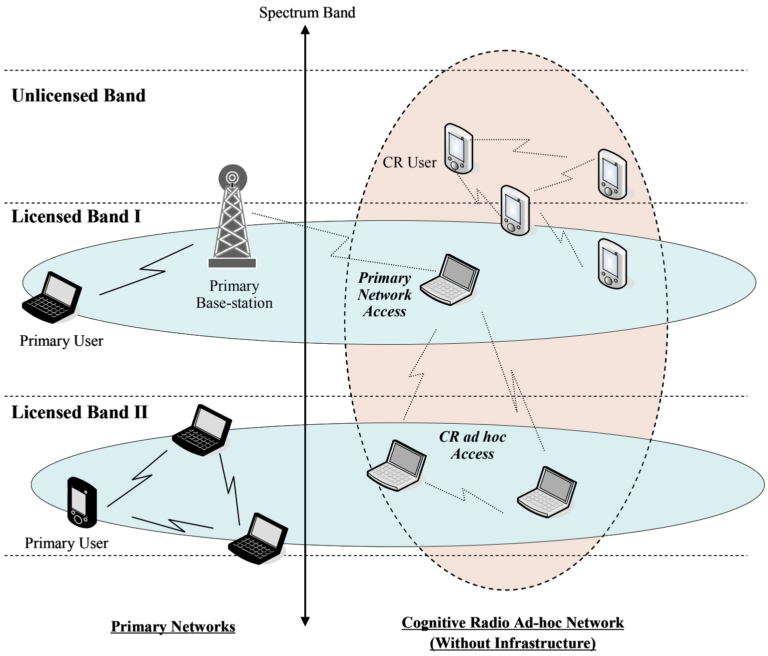


Figure 2.3 CR ad-hoc network architecture.

CR MESH NETWORKS

The wireless mesh network is a special case of ad hoc networks, and is emerging as a cost-effective technology for providing broadband connectivity [7]. However, as the network density increases and the applications require higher throughput, mesh networks require higher capacity to meet the requirements of the applications. Since the cognitive radio technology enables the access to larger amount of spectrum, CR networks can be used for mesh networks that will be deployed in dense urban areas with the possibility of significant contention.

The components of cognitive mesh networks are as follows:

- *Cognitive Mesh Router:* It serves as the Access Point supporting several users in a residential setting or along the road. The MRs serve as the routing backbone, forwarding packets till the Gateway that provides the final link to the Internet is reached.

- *Cognitive Mesh Client:* MCs are free to either associate themselves with a MR in a cluster, or form their own ad-hoc network. The ad-hoc network is connected to the backbone via any closest peer and this poses some interesting research issues:

- *Gateway:* This mesh routers can be connected to the Internet or other wireless/wired networks such as cellular and WiFi networks.

For example, the coverage area of CR networks can be increased when a meshed wireless backbone network of infrastructure links is established based on cognitive access points (CAPs) and fixed cognitive relay nodes (CRNs). The capacity of a CAP, connected via a wired broadband access to the Internet, is distributed into a large area with the help of a fixed CRN. CR networks have the ability to add temporary or permanent spectrum to the infrastructure links used for relaying in case of high traffic load.

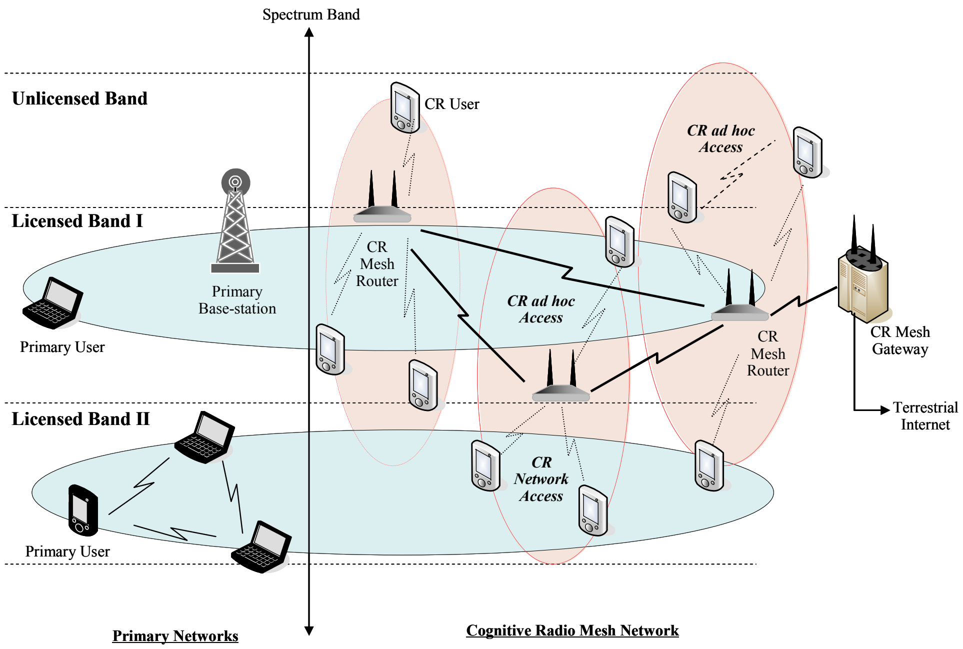


Figure 2.4 CR mesh network architectures.

2.3 SPECTRUM HETEROGENEITY

As explained before, CR network can operate in both licensed and unlicensed bands. Hence, the functionalities required for CR networks vary according to whether the spectrum is licensed or unlicensed. Accordingly, in this section, we classify the CR network operations as *CR network on licensed band* and *CR network on unlicensed band*. The CR network functions are explained in the following sections according to this classification.

2.3.1 CR Network on Licensed Band

As shown in Figure 1.1, there exist temporally unused spectrum holes in the licensed spectrum band. Hence, CR networks can be deployed to exploit these spectrum holes through cognitive communication techniques. This architecture is depicted in Figure 2.5, where the CR network coexists with the primary network at the same location and on the same

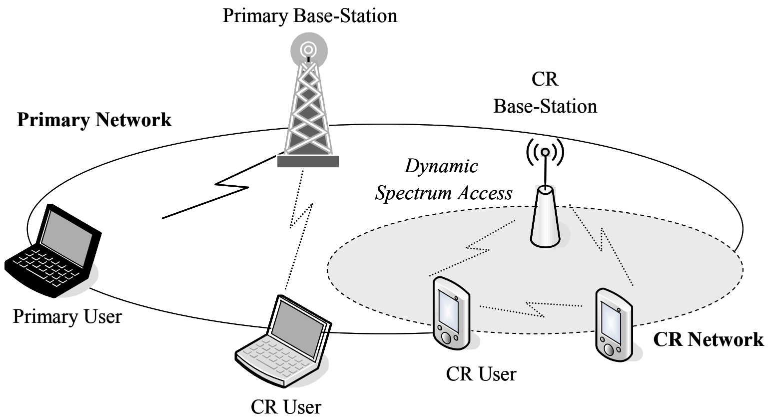


Figure 2.5 CR network on licensed band.

spectrum band.

There are various challenges for CR networks on licensed band due to the existence of the primary users. Although the main purpose of the CR network is to determine the best available spectrum, CR functions in the licensed band are mainly aimed at the detection of the presence of primary users. The channel capacity of the spectrum holes depends on the interference at the nearby primary users. Thus, the interference avoidance with primary users is the most important issue in this architecture. Furthermore, if primary users appear in the spectrum band occupied by CR users, CR users should vacate the current spectrum band and move to the new available spectrum immediately, called *spectrum handoff*.

2.3.2 CR Network on Unlicensed Band

Open spectrum policy that began in the industrial scientific and medical (ISM) band has caused an impressive variety of important technologies and innovative uses. However, due to the interference among multiple heterogeneous networks, the spectrum efficiency of ISM band is decreasing. Ultimately, the capacity of open spectrum access, and the quality of service they can offer, depend on the degree to which a radio can be designed to allocate spectrum efficiently.

CR networks can be designed for operation on unlicensed bands such that the efficiency is improved in this portion of the spectrum. The *CR network on unlicensed band* architecture is illustrated in Figure 2.6. Since there are no license holders, all network entities have the same right to access the spectrum bands. Multiple CR networks coexist in the same area and communicate using the same portion of the spectrum. Intelligent spectrum sharing algorithms can improve the efficiency of spectrum usage and support high QoS.

In this architecture, CR users focus on detecting the transmissions of other CR users. Unlike the licensed band operations, the spectrum handoff is not triggered by the appearance of other primary users. However, since all CR users have the same right to access the spectrum, CR users should compete with each other for the same unlicensed band. Thus, sophisticated spectrum sharing methods among CR users are required in this architecture.

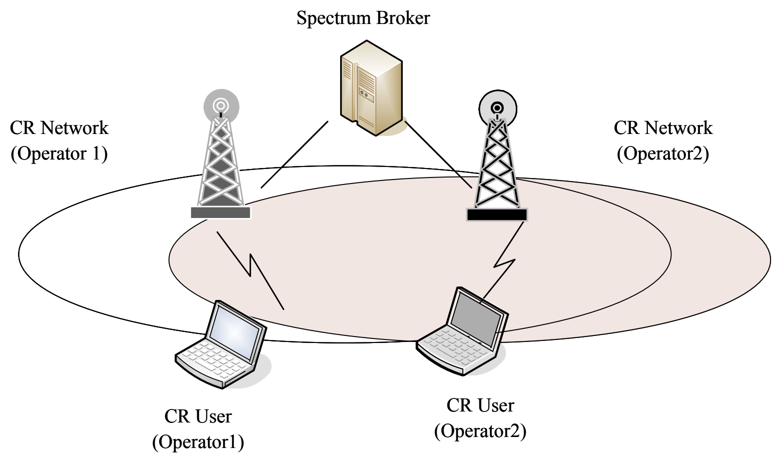


Figure 2.6 CR network on unlicensed band.

If multiple CR network operators reside in the same unlicensed band, fair spectrum sharing among these networks is also required.

PART II

SPECTRUM MANAGEMENT FOR COGNITIVE RADIO NETWORKS

CHAPTER 3

SPECTRUM MANAGEMENT FRAMEWORK

3.1 COGNITIVE CYCLE

CR networks impose unique challenges due to the coexistence with primary networks as well as diverse QoS requirements. Thus, new spectrum management functions are required for CR networks with the following critical design challenges:

- *Interference Avoidance:* CR network should avoid interference with primary networks
- *QoS Awareness:* In order to decide an appropriate spectrum band, CR networks should support QoS-aware communication, considering dynamic and heterogeneous spectrum environment.
- *Seamless Communication:* CR networks should provide seamless communication regardless of the appearance of the primary users.

In order to adapt to dynamic spectrum environment, and hence address the above challenges, the CR network necessitates the spectrum-aware operations, which form a cognitive cycle [5]. As shown in Figure 3.1, the steps of the cognitive cycle consist of four spectrum management functions: *spectrum sensing*, *spectrum decision*, *spectrum sharing*, and *spectrum mobility*. To implement CR networks, each function needs to be incorporated

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

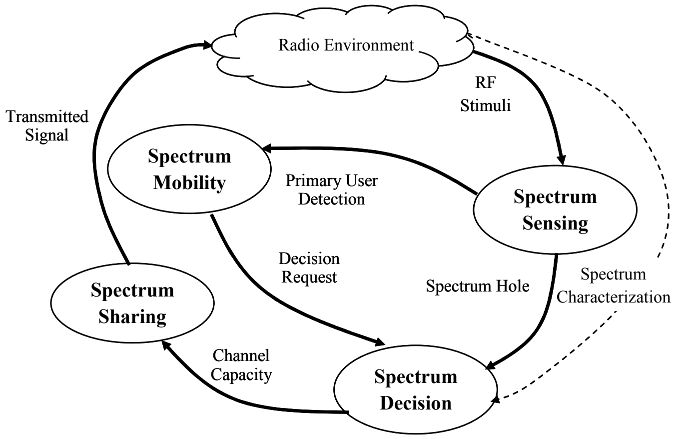


Figure 3.1 Cognitive Cycle.

into the classical layering protocols. The following are the main features of spectrum management functions:

1. *Spectrum Sensing:* A CR user can be allocated to only an unused portion of the spectrum. Therefore, a CR user should monitor the available spectrum bands, and then detect spectrum holes. Spectrum sensing is a basic functionality in CR networks, and hence it is closely related to other spectrum management functions as well as layering protocols to provide information on spectrum availability.

2. *Spectrum Decision:* Based on the spectrum availability, CR users can allocate a channel. This allocation not only depends on spectrum availability, but it is also determined based on internal (and possibly external) policies.Once the available spectrums are identified, it is essential that the CR users select the most appropriate band according to their QoS requirements. It is important to characterize the spectrum band in terms of both radio environment and the statistical behaviors of the PUs. In order to design a decision algorithm that incorporates dynamic spectrum characteristics, we need to obtain *a priori* information regarding the PU activity. Furthermore, in CRAHNs, spectrum decision involves jointly undertaking spectrum selection and route formation.

3. *Spectrum Sharing:* Since there may be multiple CR users trying to access the spectrum, their transmissions should be coordinated to prevent collisions in overlapping portions of the spectrum. Spectrum sharing provides the capability to share the spectrum resource opportunistically with multiple CR users which includes *resource allocation* to avoid interference caused to the primary network. Furthermore, this function necessitates a *CR Medium Access Control (MAC) protocol*, which facilitates the *sensing control* to distribute the sensing task among the coordinating nodes as well as *spectrum access* to determine the timing for transmission.

4. *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 needs to be continued in another vacant portion of the spectrum.If a PU is detected in the specific portion of the spectrum in use, CR users should vacate

the spectrum immediately and continue their communications in another vacant portion of the spectrum. For this, either a new spectrum must be chosen or the affected links may be circumvented entirely. Thus, spectrum mobility necessitates a *spectrum handoff* scheme to detect the link failure and to switch the current transmission to a new route or a new spectrum band with minimum quality degradation. This requires collaborating with spectrum sensing, neighbor discovery in a link layer, and routing protocols. Furthermore, this functionality needs a *connection management* scheme to sustain the performance of upper layer protocols by mitigating the influence of spectrum switching.

3.2 SPECTRUM MANAGEMENT FRAMEWORK

According to the network architecture, CR networks require different spectrum management functionalities. The spectrum management frameworks for the infrastructure-based CR network and CR ad hoc networks are illustrated in Figures 3.2 and 3.3, respectively, It is evident from the significant number of interactions that the spectrum management functions necessitate a cross-layer design approach.

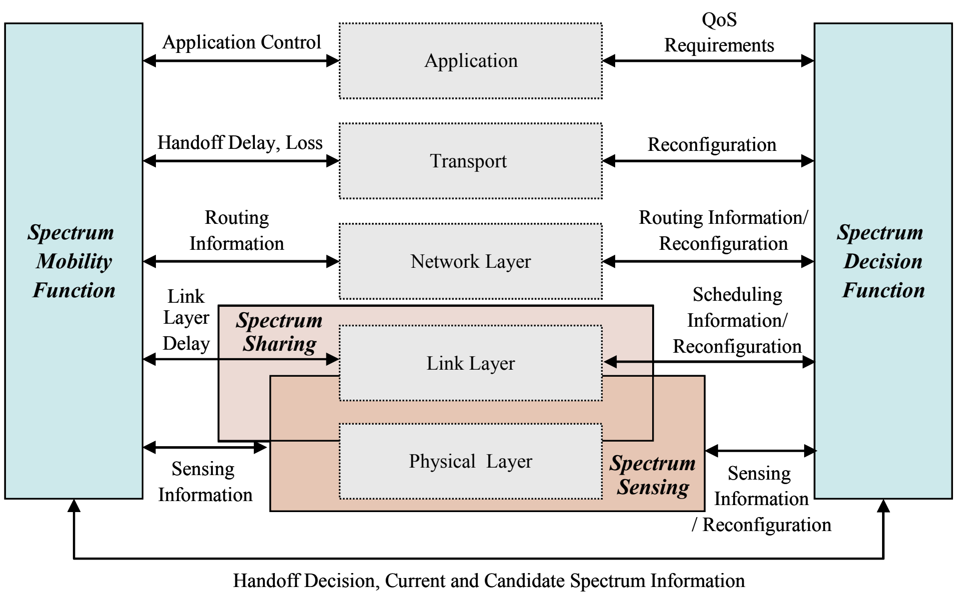


Figure 3.2 Spectrum management framework for infrastructure-based cognitive radio networks.

This spectrum management framework needs to be implemented differently according to the network architecture. In the infrastructure-based CR networks, the observations and analysis performed by each CR user feed the central CR base-station, so that it can make decisions on how to avoid interfering with primary networks. According to this decision, each CR user reconfigures its communication parameters, as shown in Figure 3.4 (a). On the contrary, in CR ad hoc networks, each user needs to have all CR capabilities and is responsible for determining its actions based on the local observation, as shown in Figure 3.4 (b). Since the CR user cannot predict the influence of its actions on the entire network with

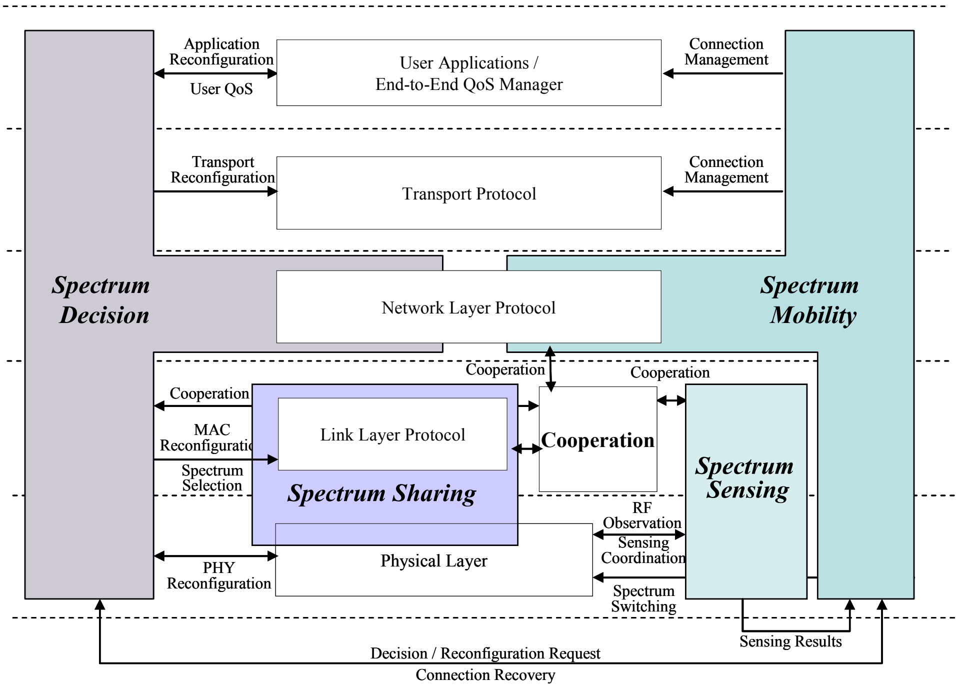


Figure 3.3 Spectrum management framework for cognitive radio ad hoc networks.

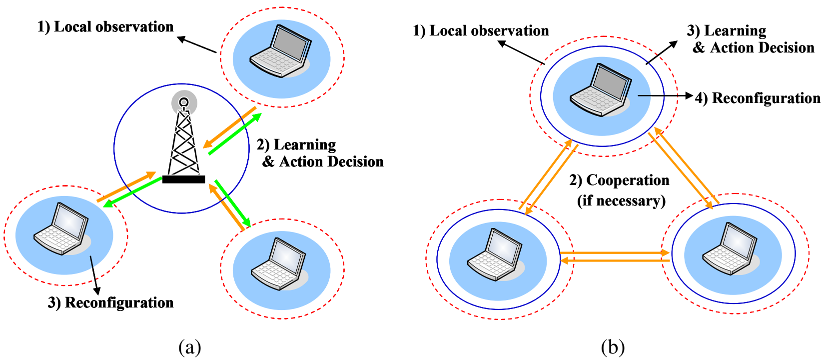


Figure 3.4 Comparison between CR capabilities for (a) infrastructure-based CR networks, and (b) CR ad hoc networks.

its local observation, all of spectrum management functions in CRAHNs are based on cooperative operation to broaden the knowledge on the network. In this scheme, all decisions are made based on the observed information that is gathered from their neighbors [3] [4].

In the following chapters, we investigate how these spectrum management functions are integrated into the existing layering functionalities in CR networks and address the challenges of them. In this thesis, all proposed solutions are focused on the development of CR networks that require no modification of primary networks.

CHAPTER 4

SPECTRUM SENSING

A cognitive radio is designed to be aware of and sensitive to the changes in its surrounding, which makes spectrum sensing an important requirement for the realization of CR networks. Spectrum sensing enables CR users to exploit the unused spectrum portion adaptively to the radio environment. The main objective of spectrum sensing is to provide more spectrum access opportunities to CR users without harmful interference to the primary networks. To this end, CR hardware should be able to opportunistically identify the unused portions of the spectrum and use them for communication. However, these licensed bands, also defined as primary bands should be immediately vacated if the legitimate or primary users are detected.

If CR networks can obtain the information regrading the spectrum usage by directly interacting with the primary network, they do not require the sensing capability. However, this scheme requires tremendous modifications to the legacy primary network, which is impractical because of high deployment cost. For this reason, CR networks are generally assumed to be deployed independently from primary networks without any communication channel between them. In this scenario, to identify spectrum availability, CR networks should rely only on their local observations obtained by spectrum sensing. Consequently, the accurate sensing of the wireless spectrum is a key challenge in realizing CR technology.

The spectrum sensing capability is generally required in the following cases:

- When CR users begin to transmit or vacate the spectrum because of the appearance of the primary user, they should find available spectrum holes over a wide frequency range for their transmission, called *out-of-band sensing*.

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury

978-0470-68852-6

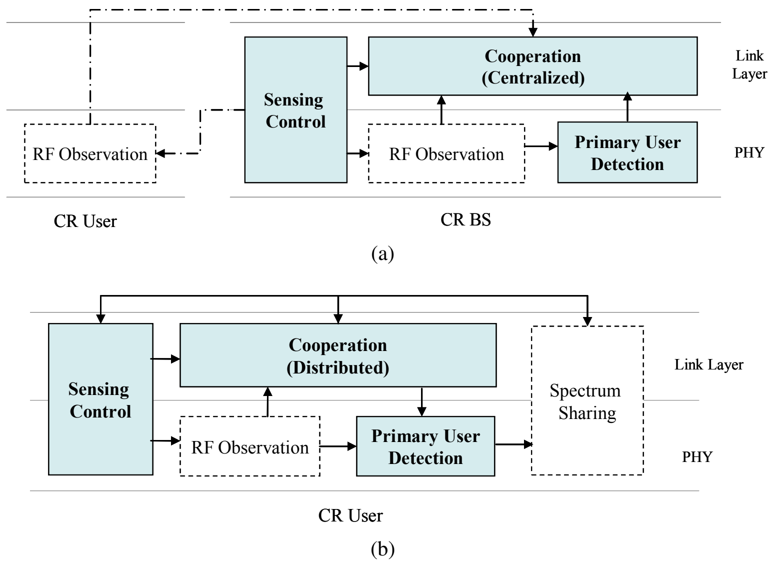


Figure 4.1 Functional block diagram for spectrum sensing: (a) infrastructure-based CR networks, and (b) CR ad hoc networks.

- During the transmission, CR users keep monitoring the spectrum band that they are in use and detect the presence of primary users to avoid harmful interference to primary users, referred to as *in-band sensing*.

Although both cases are commonly based on the capability to reliably obtain the spectrum availability, they require different design principles. Out-of-band sensing mainly focuses on a fast discovery of the unused spectrum portions among multiple licensed bands that spread over a wide frequency range. On the other hand, in-band sensing emphasizes the protection of primary networks from harmful interference caused by the transmissions of CR networks. The main difference in implementation between both schemes lies in how to control the detection capability. More details are explained in Section 4.4.

BASIC FUNCTIONALITIES

As shown in Figure 4.1, the CR network necessitates the following functionalities for spectrum sensing:

- *PU Detection:* The CR user observes and analyzes its local radio environment. Based on these location observations of itself and its neighbors, CR users determine the presence of PU transmissions, and accordingly identify the current spectrum availability.

- *Cooperation:* The observed information in each CR user is sent to base-station or exchanged with its neighbors, and spectrum availability is determined accordingly. Through this cooperation, sensing accuracy is significantly improved.

- *Sensing Control:* The PU detection functionality is controlled and coordinated by a sensing controller, which considers two main issues on i) how quickly a CR user can find the available spectrum band over a wide frequency range for their transmissions, and ii) how long and how frequently a CR user should sense the spectrum to achieve sufficient sensing accuracy during the transmission and detect the presence of transmission in primary networks to avoid interference.

ARCHITECTURAL FEATURES
Spectrum sensing functionalities are implemented differently according to the network architecture. In infrastructure-based networks, the base-station plays a role in coordinating the operations of sensing operation through the synchronized sensing schedule. Sensing parameters determined through sensing control are applied to the sensing operations of all CR users. By considering all sensing information gathering from CR users, the base-station determines spectrum availability in its coverage, as shown in Figure 4.1 (a). On the other hand, due to the lack of strict coordination, CR ad hoc users perform sensing operations independently of each other, leading to an adverse influence on sensing performance. In the worst case, the sensing operations of one CR user may be interfered by the transmission of other neighboring CR users since CR users may not be able to distinguish the signals from primary and CR users. Thus, spectrum sensing is closely coupled with spectrum sharing, especially medium access control (MAC) protocols, as depicted in Figure 4.1 (b).

In the following sections, we first explain three main functions that comprises spectrum sensing: PU detection, cooperation, and sensing control, and then present the state-of-the-art solutions in implementing spectrum sensing. In addition, we introduce alternatives to spectrum sensing currently suggested in cognitive radio, and finally also provides an in-depth discussion of the open research topics in spectrum sensing.

4.1 SPECTRUM SENSING: OVERVIEW

4.1.1 Spectrum Opportunity

The main objective of spectrum sensing is to find more spectrum opportunities while avoiding interference to primary networks. Here, the *spectrum opportunity* can be modeled in terms of *time*, *frequency*, *space*, and *power* as follow:

- *Time:* In general, a spectrum band is not continuously used by primary users. Thus, in the time dimension, spectrum opportunity is defined as "a band of frequencies that are not being used by the primary user of that band at a particular time" [282]. This dimension depends on in-band sensing capability.

- *Frequency:* Unused spectrum portions are generally spread over a wide frequency range. Thus, spectrum opportunities in the frequency domain are identified by the operations of PU detection over multiple bands, which is mainly related to out-of-band sensing.

- *Space:* This dimension is related to the propagation loss in space. According to the geographical location of the primary transmitter and the receiver, spectrum availability is different from one place to another even in the same spectrum band. This spectrum opportunities can be utilized by adjusting the transmission power of CR users. Beamforming schemes also allow CR users to exploit spectrum opportunities in space by transmitting in a specific direction not to interfere with the transmission of primary users.

- *Power:* Time, frequency space dimensions in spectrum opportunity involve the determination of the presence of the primary signals in the spectrum, called *overlay approach*. However, in the power dimension, even though primary users occupy a spectrum band, CR users can also exploit it through spread spectrum or frequency hopping capabilities, which is referred to as *underlay approach*. In this case, the addition of interference from CR transmissions does not affect the transmission of primary users. This is a basic idea of a interference temperature model, which is explained in Section 4.1.3.

In the literature, most of spectrum sensing schemes have investigated spectrum opportunities in terms of *time*, *frequency*, and *space*. The underlay approach based on the power dimension includes resource allocation issues, which is one of the main functionalities in spectrum sharing, and hence is explained in Chapter 6 in detail. This chapter mainly focuses on spectrum sensing based on the overlay approach.

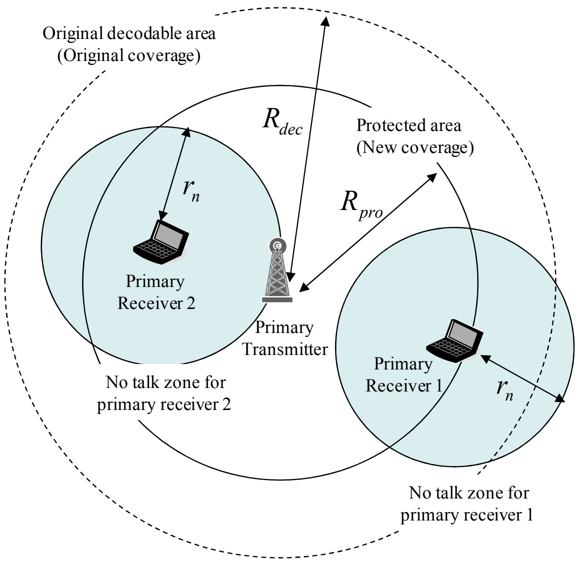


Figure 4.2 Optimal no talk zone.

4.1.2 Spectrum Sensing Model

OPTIMAL NO-TALK ZONE

Spectrum opportunities in either time or frequency guarantee no primary user activity in a certain area, but may cause interference to the primary receivers within the transmission range of the CR user. To avoid this hidden node problem, CR users need to exploit the space dimension of the spectrum opportunity by introducing a new concept, called *no talk zone*. The *no-talk zone* is defined as the area where CR users should not transmit to protect the reception of the primary receiver [226].

As Figure 4.2 shows the spectrum opportunities based no-talk zone. Here R_{dec} is the radius within which the receiver is able to decode the signal from the primary transmitter, and

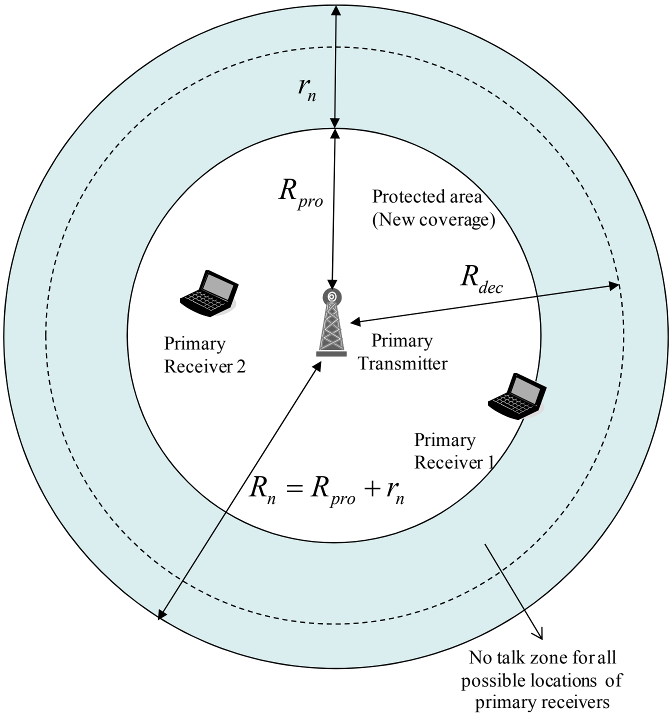


Figure 4.3 Global no talk zone considering all possible locations of primary receivers.

R_{pro} is the minimum radius where the decodability of the primary receiver should be guaranteed. Note that the use of spectrum opportunity in the space dimension inevitably leads to decrease in the decodable region from R_{dec} to R_{pro} due to the interference from CR users.

The no-talk zone is closely related to the interference range of a CR transmitter, r_n, which is defined as the maximum distance from a primary receiver at which the incurred interference is still considered harmful. This interference range is obtained by the transmission power of the CR user, and SINR threshold Γ. Γ, is the predetermined threshold below which primary receivers are not capable of decoding the signal from the primary transmitter. This threshold is generally imposed by the regulatory bodies, and depends on the primary receiver's robustness toward interference. For example, the thresholds for analog and digital TVs are 34dB and 23dB, respectively.

The primary receiver can tolerate the interference caused by a CR user as long as their spatial separation is greater than r_n. Thus, the optimal no-talk zone should be determined based on the location of each primary receiver with radius r_{n} centered at the primary receiver as shown in Figure 4.2.

GLOBAL NO-TALK ZONE

In reality, it is impossible to be aware of the exact location of the primary receiver, and hence the optimal no-talk zone may not be feasible. Instead, all possible locations for primary receivers are considered in determining a global no-talk zone, as shown in Figure 4.3. Thus, the radius of the global no-talk zone centered at the primary transmitter, R_{n}, is determined

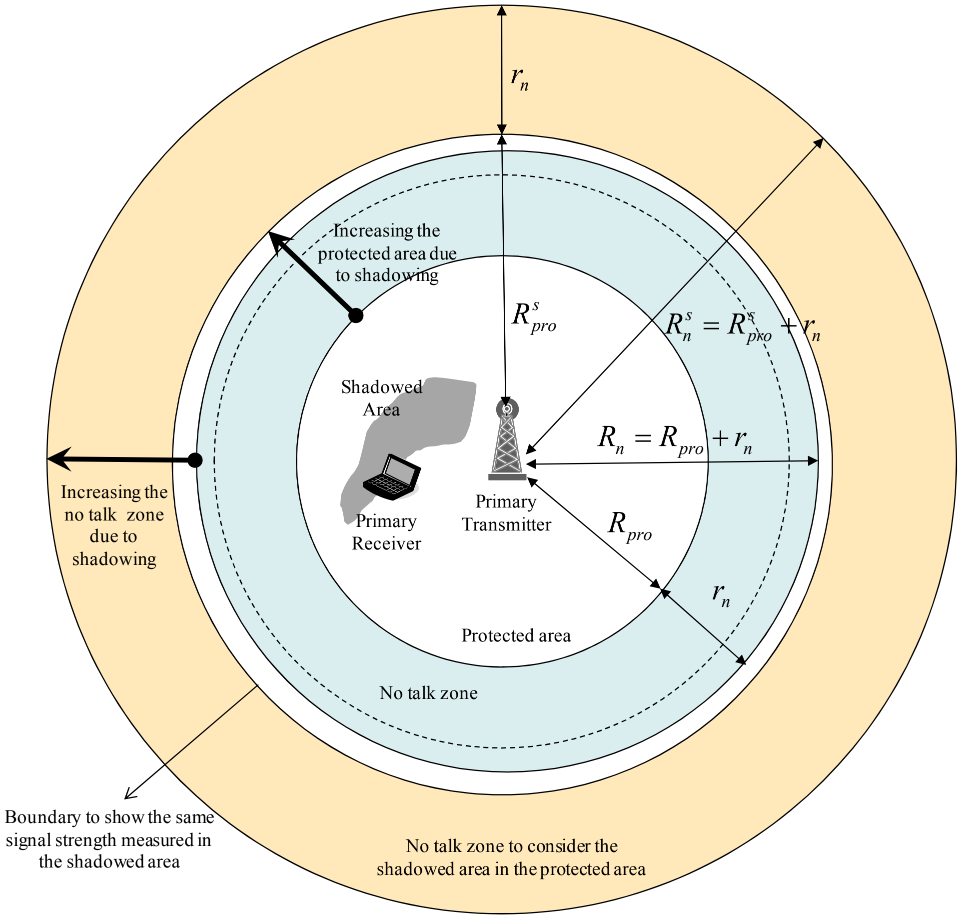


Figure 4.4 No talk zone that considers shadowing in the protected area

as $R_{\rm pro} + r_{\rm n}$.

As a result, the CR user should be able to detect active primary transmitters within a radius of $R_{\rm pro} + r_{\rm n}$. Thus, the minimum SNR at which the CR user should be still capable of detecting the primary signal, $\gamma_{\rm min}$, also called the *detection sensitivity*, can be expressed as,

$$\gamma_{\rm min} = \frac{P_{\rm p}(R_{\rm pro} + r_{\rm n})^{-\alpha}}{N_0 W} \tag{4.1}$$

where $P_{\rm p}$ is the transmission power of the primary user, and α is the attenuation coefficient. N_0 is the one-sided noise power spectral density, and W the bandwidth of spectrum. Compared to the optimal no-talk zone, the global no-talk zone causes significant loss in spectrum opportunities due to the unnecessary no-talk zone where the transmission of CR users does not interfere with any primary receiver.

Furthermore, this no-talk zone concept can be applied to the shadowed and faded radio environment. Due to the low signal strength observed in the shadowed region, CR user in

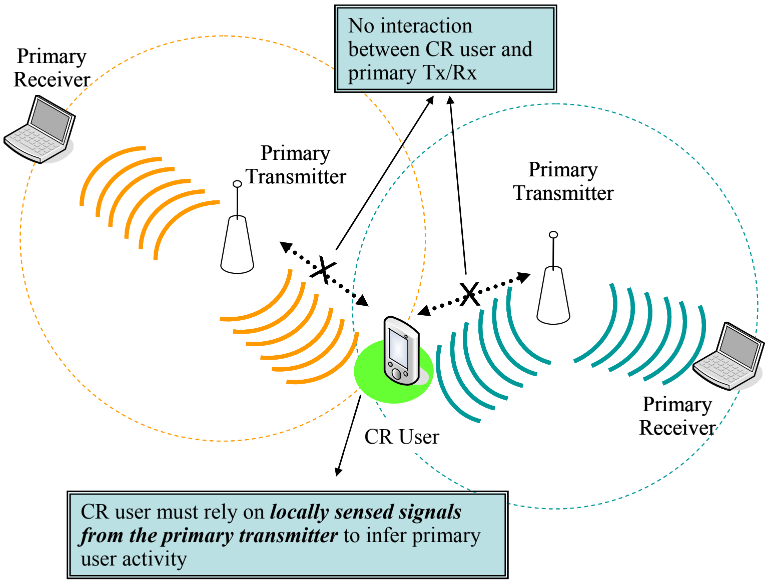


Figure 4.5 Primary transmitter detection.

that area may not detect the presence of the primary signal correctly. To avoid this problem, the protected region can be extended to the radius $R^{\mathrm{s}}_{\mathrm{pro}}$ that shows the same signal strength observed in the shadowed area. Accordingly the no-talk zone R_{n} increases by $R^{\mathrm{s}}_{\mathrm{pro}} - R_{\mathrm{pro}}$, as shown in 4.4.

4.1.3 Spectrum Sensing Classification

As explained above, spectrum opportunity can be modeled as four dimensions: time, frequency, space, and power, all of which are tightly coupled with each other. These spectrum opportunities can be identified by PU detection techniques, which are generally classified into three groups: (1) *primary transmitter detection*, (2) *primary receiver detection*, and (3) *interference temperature management*. In the following sections, we describe these spectrum sensing methods for CR networks and discuss the open research topics in this area as described in the following.

TRANSMITTER DETECTION

Since CR users are usually assumed not to have any real-time interaction with the primary transmitters and receivers, they cannot know the exact information on current transmissions within the primary networks Thus, in transmitter detection, in order to distinguish between used and unused spectrum bands, CR users detect the signal from a primary transmitter through the only local observations of CR users, as shown in Figure 4.5. Thus, CR users should have the capability to determine if a signal from the primary transmitter is locally present in a certain spectrum.

Due to the lack of interactions between the primary users and the CR users, the transmitter detection techniques rely on the weak signals from only the primary transmitters. This introduces critical hidden node issues in spectrum sensing as follow:

- *Receiver Uncertainty Problem:* First, transmitter detection techniques alone cannot avoid the interference to primary receivers because of the lack of primary receiver's information as depicted in Figure 4.6. In transmitter detection, when both CR transmitter and receiver do not detect the presence of primary transmissions, the CR transmitter begins to transmit, but may cause interference at nearby primary receiver located within the transmission range.

- *Shadowing Problem:* Moreover, the multi-path fading and shadowing effects also leads to increase in interference to primary networks. If CR users cannot have a good line-of-sight to a primary transmitter, and located in the shadowed area, it may not be able to detect the primary transmitter, as shown in Figure 4.7.

Both problems can be partially resolved by exploiting sensing information from other users. More details are explained in Section 4.3.

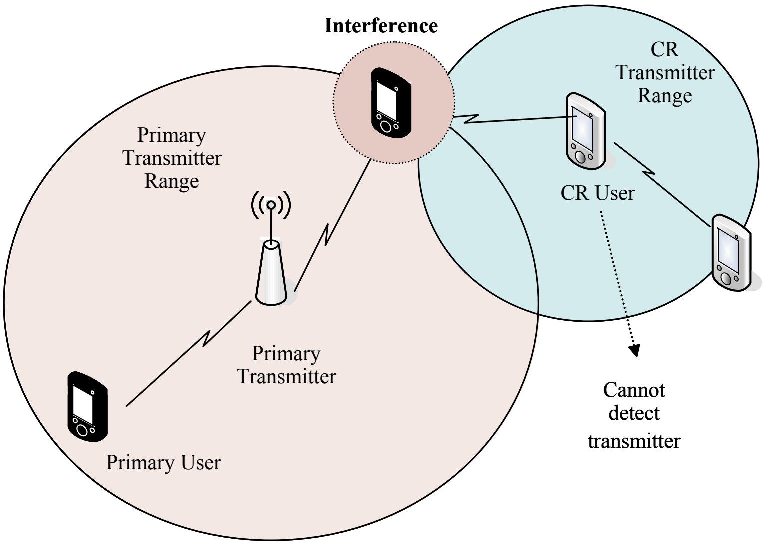


Figure 4.6 Transmitter detection problem: receiver uncertainty.

RECEIVER DETECTION

Although cooperative detection reduces the probability of interference, the most efficient way to detect spectrum holes is to detect the primary users that are receiving data within the communication range of a CR user. As depicted in Figure 4.8, the primary receiver usually emits the local oscillator (LO) leakage power from its RF front-end when it receives the signals from the primary transmitter. In order to determine the spectrum availability, a primary receiver detection method exploits this local oscillator (LO) leakage power instead of the signal from the primary transmitter and detects the presence of the primary receiver directly [271].

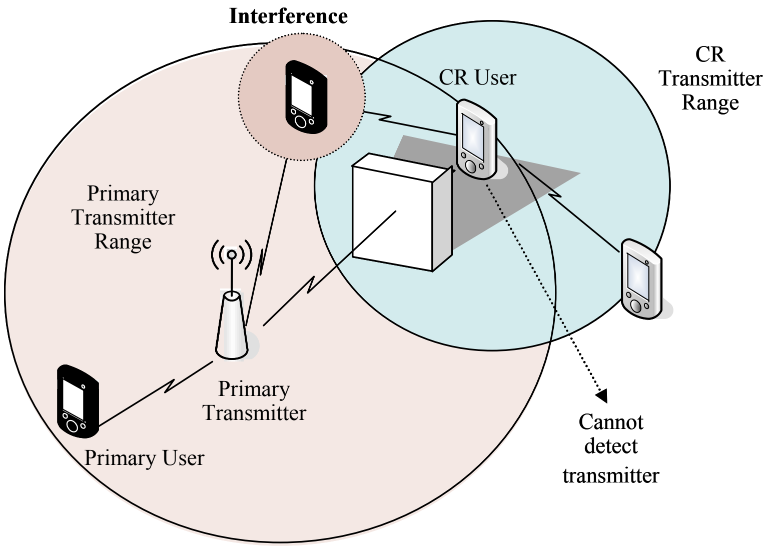


Figure 4.7 Transmitter detection problem: shadowing uncertainty.

Detecting this leakage power directly with a CR would be impractical for two reasons. Firstly, it would be difficult for the receive circuitry of the CR to detect the LO leakage over larger distances. From the calculations shown on the next page, it can be shown that a distance of 20m, it would take on the order of seconds to detect the LO leakage with a high probability. For a practical system, the detection would need to be made on the order of milliseconds at worst. The second reason that it would be impractical to detect the LO leakage directly is that the LO leakage power is very variable, depending on the receiver model and year. If the CR used this variable power level to estimate proximity to the primary receiver, there would be too much error introduced by this variability. Such an approach may be feasible for TV receivers only or need further hardware such as a supporting sensor network in the area with the primary receivers.

INTERFERENCE TEMPERATURE MANAGEMENT

Both transmitter and receiver detection methods explained above, are based on the *overlay access approach* where the CR user accesses the network using a portion of the spectrum that has not been used by licensed users. However, with this approach that mainly focuses on determining the presence of primary user's transmission, it cannot exploit all spectrum opportunities that the primary network provides.

To solve this problem, CR networks can adopt a underlay access scheme that exploits the spread spectrum techniques developed for cellular networks. Once a spectrum allocation map has been acquired, an CR user begins transmission such that its transmit power at a certain portion of the spectrum is regarded as noise by the licensed users. This technique requires sophisticated power allocation techniques based on spreading spectrum and can utilize increased bandwidth compared to overlay techniques.

In the *underlay access approach*, the most important issue is how to determine the transmission power margin for CR users that does not influence on the transmission of primary

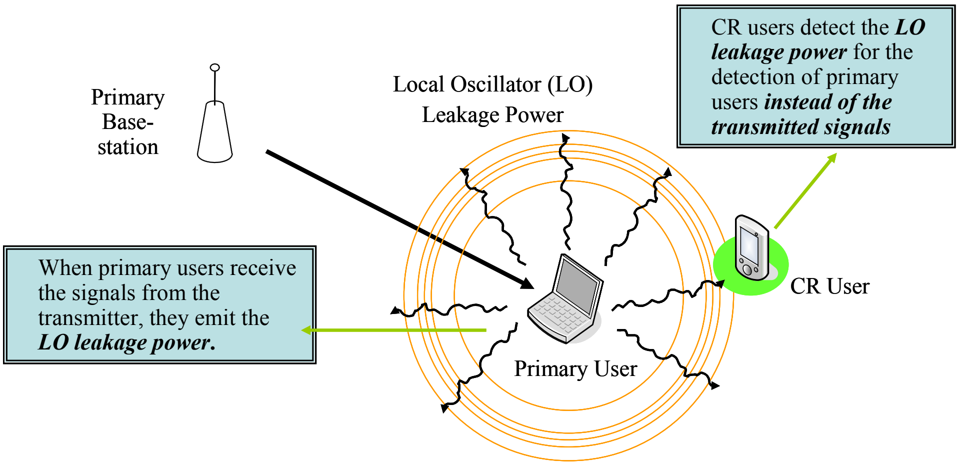


Figure 4.8 Primary receiver detection.

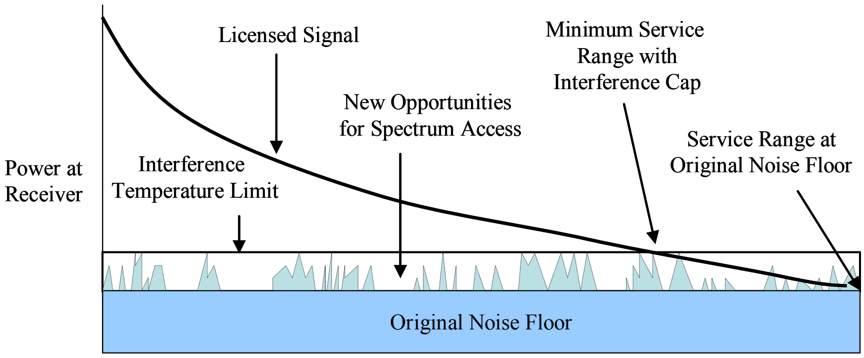


Figure 4.9 Interference temperature management [73].

users. To this end. the interference temperature model is proposed by FCC. Traditionally, interference can be controlled at the transmitter through the radiated power and location of individual transmitters through the radiated power, the out-of-band emissions and location of individual transmitters. However, interference actually takes place at the receivers, as shown in Figure 4.6. Recently, a new model for measuring interference, referred to as *interference temperature* has been introduced by the FCC [73].

Figure 4.9 shows the signal of the primary transmitter designed to operate out to the distance at which the received power approaches the level of the noise floor. The noise floor is location specific based on the additional interfering signals at that point. As shown in Figure 4.9, this model suggests an interference temperature limit, which is the amount of new interference that the primary receiver could tolerate. As long as CR users do not exceed this limit, they can use the spectrum band.

Although this model is best fit for the objective of spectrum sensing, the difficulty lies in accurately determining the interference temperature limit for each location-specific case. A CR user is naturally aware of its transmit power level and can determine its precise location with the help of a positioning system. With this ability, however, its transmission could cause significant interference at a neighboring receiver on the same frequency. Currently, there exists no practical way for a cognitive radio to measure or estimate the interference temperature at nearby primary receivers. Since primary receivers are usually passive devices, a CR user cannot be aware of the precise locations of primary receivers. Furthermore, if CR users cannot measure the effect of their transmission on all possible receivers, a useful interference temperature measurement may not be feasible. Also, with the increase in the interference temperature limit, the SNR at the primary receiver decreases, resulting in decreasing primary network's capacity and coverage as depicted in Figure 4.9. Another important challenge is the determination of the interference temperature limit. Since CR users try to control their transmissions according to this limit, an accurate model for the determination of the interference temperature limit is necessary. However, the optimal value of this limit may depend on the density and the traffic characteristics of the primary users. Furthermore, the physical layer characteristics such as the modulation and transmit power as well as the operating frequency of the primary users also affect the limit. Consequently, adaptive techniques for the determination of the interference temperature limit are necessary.

Consequently, FCC determined that this concept is not workable, and hence abandoned it in 2007 [71]. However, the basic concept of underlay and interference temperature is still advantageous to improve the spectrum efficiency in CR networks, and hence has widely explored combining with the power allocation in spectrum sharing. More details are presented in Chapter 6.

4.2 PU DETECTION

Since CR users are usually assumed not to have any real-time interaction with the primary transmitters and receivers, they cannot know the exact information on current transmissions within the primary networks Thus, in transmitter detection, in order to distinguish between used and unused spectrum bands, CR users detect the signal from a primary transmitter through the only local observations of CR users, as shown in Figure 4.5.

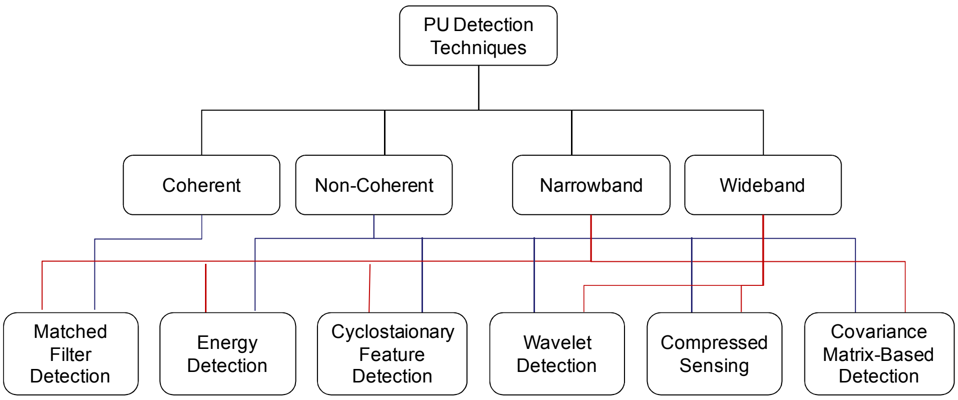


Figure 4.10 Classification of primary user detection techniques [162].

From the perspective of signal detection, PU detection techniques can be classified into two broad categories: *coherent* and *noncoherent detection*. In coherent detection, the primary signal can be coherently detected by comparing the received signal with a priori knowledge of primary signals. In non-coherent detection, no a priori knowledge is required for detection. Another way to classify PU detection techniques is based on the bandwidth of the spectrum of interest for sensing: narrowband and wideband. The classification of sensing techniques is shown in Figure 4.10. All PU detection schemes except the matched filter detection adopt the non-coherent detection. Wavelet and compressed detection schemes are mainly for wideband sensing while the rest of them are primarily used for narrowband sensing. Each PU detection scheme is investigated in this section, followed by the basic detection theory.

4.2.1 Basic Detection Theory

In CR network each user should have the capability to determine if a signal from the primary transmitter is locally present in a certain spectrum. For this PU detection, the following hypothesis model is generally used:

$$x(t) = \begin{cases} n(t) & H_0 \\ hs(t) + n(t) & H_1 \end{cases} \tag{4.2}$$

where $x(t)$ is the signal received by the CR user, $s(t)$ is the transmitted signal of the primary user, $n(t)$ is a zero-mean additive white Gaussian noise (AWGN) and h is the amplitude gain of the channel. H_0 is a null hypothesis, which states that there is no licensed user signal in a certain spectrum band. On the other hand, H_1 is an alternative hypothesis, which indicates that there exist some primary user signal.

HYPOTHESIS TESTING

Hypothesis Testing is a statistical test to determine the presence or absence of PU. Figure 4.11 shows a main idea of the hypothesis testing in detection theory. Let $f(y|H_0)$ and $f(y|H_1)$ be the distributions of observation Y under hypotheses, H_0 and H_1, respectively. If the observation (test statistic) Y is greater than threshold λ, the detector determines that the primary signal is present. Otherwise, it considers that the spectrum is idle. However, the detector does not always make a correct decision because of the stochastic characteristics of the test statistic Y. If the signal is absent, but the detector chooses H_1 by mistake, a false alarm occurs, leading to missing spectrum opportunity. On the other hand, even though the signal is present, the detector may not determine it correctly, and make miss-detection, resulting in a collision with primary users. These detection performance can be characterized by the probabilities of detection and false alarm, P_{d}, and P_{f}, which depends on the decision criterion, i.e., choice of the threshold λ.

The following are two different testing schemes for the binary decision on the presence of signal $s(t)$

- *Maximum Likelihood (ML) Detection (or Neyman-Pearson Detection):* The detection probability P_{d} and false alarm probability P_{f} can be expressed as follows:

$$P_{\mathrm{d}}(\lambda) = Pr[Y > \lambda | H_1] \tag{4.3}$$

$$P_{\mathrm{f}}(\lambda) = Pr[Y > \lambda | H_0] \tag{4.4}$$

This detection criterion is called *a maximum likelihood (ML) detection* or *Neyman-Pearson detection*. This is equivalent to the following likelihood ratio test (LRT) given by

$$\Lambda(y) = \frac{f(y|H_1)}{f(y|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_{\text{LRT}} \tag{4.5}$$

where λ_{LRT} is the detection threshold for LRT. If λ in Eqs. (4.3) and (4.4) is obtained by $f(\lambda|H_0) = f(\lambda|H_1)$, λ_{LRT} is determined to be unity. The detector declares H_1 if $\Lambda(y) > \lambda_{\text{LRT}}$ and declares H_0 otherwise.

- *Maximum A Posteriori (MAP) Detection (or Bayesian Detection):* If the detector has information on the occurrence of the primary signals, i.e., the probability of the period used by primary users, P_{on} and the probability of the idle period, P_{off}, we can implement a more accurate detection scheme, called *a maximum a posteriori (MAP) detector* or *Bayesian detector*. From the definition of MAP detection, the detection probability P_{d} and false alarm probability P_{f} can be expressed as follows:

$$P_{\text{d}}(\lambda) = Pr[Y > \lambda|H_1]P_{\text{on}} \tag{4.6}$$

$$P_{\text{f}}(\lambda) = Pr[Y > \lambda|H_0]P_{\text{off}} \tag{4.7}$$

The LRT of the MAP detector is represented as

$$\Lambda(y) = \frac{f(y|H_1)}{f(y|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_{\text{off}}}{P_{\text{on}}}\lambda_{\text{LRT}} \tag{4.8}$$

The detector declares H_1 if $\Lambda(y) > \frac{P_{\text{off}}}{P_{\text{on}}}\lambda_{\text{LRT}}$ and declares H_0 otherwise.

Although a maximum a posteriori (MAP) detector known to be optimal [204], a maximum likelihood (ML) detection has been widely used for PU detection without considering the probabilities of ON and OFF states [46] [98] [183] [247].

RECEIVER OPERATING CHARACTERISTIC (ROC)

The relationship between both probabilities is shown in Figure 4.13, called receiver operating characteristic (ROC) curve. The operating point on the curve is determined by changing the threshold. As shown in Figure 4.13, the detector has different ROC curves according to the signal-to-noise ratio (SNR), i.e., received signal strength $hs(t)$. Generally, the detection scheme has the lowest operating SNR to satisfy both target probabilities of detection and false alarm. For example, IEEE 802.22 requires a detector working at -20 dB SNR, where the lowest received primary DTV signal detected is -116dBm.

Beside the SNR, a sensing time also influence the ROC curve. The operating SNR and sensing time are closely coupled in determining the ROC curve. If the detector want to operate at the lower SNR while satisfying the probabilities of false alarm and detection, it needs to have a longer sensing time.

SEQUENTIAL TESTING

In the previously discussed hypothesis testing methods such as the NP-based LRT, the number of required samples for testing is fixed, which corresponds to the fixed sensing time.

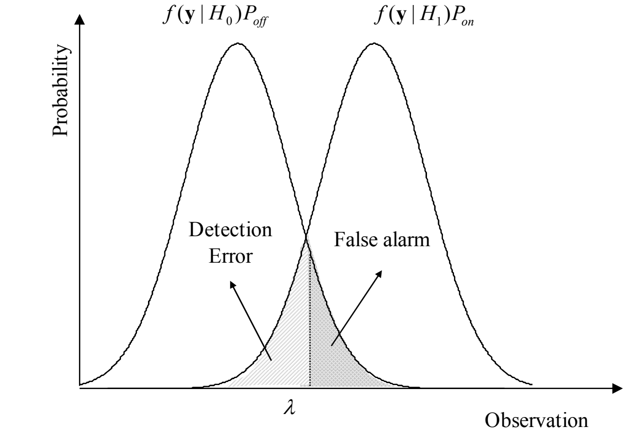


Figure 4.11 Probabilities of detection and false alarm in ML detection.

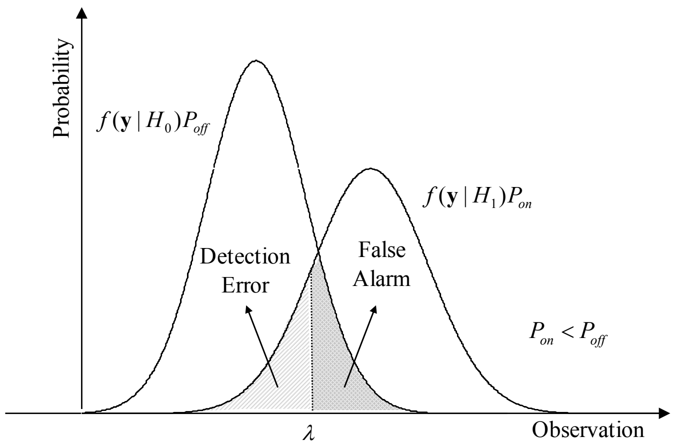


Figure 4.12 Probabilities of detection and false alarm in MAP detection.

On the contrary, the sequential hypothesis test is an approach of statistical inference, whose characteristic feature is that the number of samples required for testing is not determined in advance of the experiment, but is variable according to the detection performance.. The sequential probability ratio test (SPRT) is a special class of the sequential testing scheme that can minimize the sensing time subject to the detection performance constraints [261].

Suppose x_i is the i^{th} observation of the received signal. The measurements from each sensing period is observed sequentially until a change in channel is observed. If the decision is made,the process is terminated; otherwise, another observation is taken. If observation $\mathbf{x}$ is an independent and identically distributed (i.i.d.) random variable, log likelihood ratio (LLR) for sequential test is obtained as follow:

$$Y_i = \log \frac{Pr\{x_1|H_1\}Pr\{x_2|H_1\}\ldots Pr\{x_n|H_1\}}{Pr\{x_1|H_0\}Pr\{x_1|H_0\}\ldots Pr\{x_i|H_0\}} = \sum_{k=1}^{i} \log \frac{Pr\{x_k|H_1\}}{Pr\{x_k|H_0\}} \tag{4.9}$$

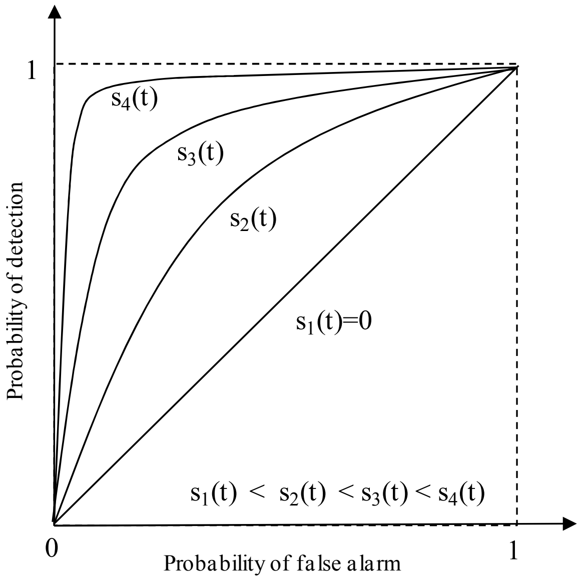


Figure 4.13 Classification of spectrum sensing techniques.

The sequential detection is based on the following criterion:

$$\begin{aligned} &Y_i \geq \lambda_1 \Rightarrow \text{accept } H_1 \\ &Y_i \leq \lambda_0 \Rightarrow \text{accept } H_0 \\ &\lambda_0 \leq Y_i \leq \lambda_1 \Rightarrow \text{take another observation } H_1 \end{aligned} \tag{4.10}$$

If the log likelihood ratio is greater than λ_1, the detector decides on H_1 while if it is smaller than λ_0, it decides on H_0. When the ratio falls between the two thresholds, it waits for the next observation, as the currently available information is not sufficient to achieve the final decision that satisfies the target constraints. In this case, the process is repeated until the decision can be determined.

Here λ_0 and λ_1 is approximated based on requirements on false alarm and detection probabilities [261] as follow:

$$\lambda_0 \approx \log \frac{1 - \tilde{P}_{\mathrm{d}}}{1 - \tilde{P}_{\mathrm{f}}} \tag{4.11}$$

$$\lambda_1 \approx \log \frac{\tilde{P}_{\mathrm{d}}}{\tilde{P}_{\mathrm{f}}} \tag{4.12}$$

where $\tilde{P}_{\mathrm{d}}$ and $\tilde{P}_{\mathrm{f}}$ are the tolerated probabilities of detection and false alarm, respectively.

The main advantage of the SPRT is that it requires less samples on the average than those fixed-sample testing methods to achieve the same detection performance. It is proven that the SPRT is optimal in minimizing the average number of independent samples and the corresponding average sensing time. The disadvantage of the SPRT includes the cost for obtaining samples and the possibly large number samples needed to reach the decision resulting in long sensing time [14]. In [305] [306], a sequential detection scheme with the SPRT is proposed for cooperative sensing. In this method, the fusion center sequentially accumulates the log-likelihood statistics from cooperating CR users and determines when to stop taking more sequential observations and make a cooperative decision.

4.2.2 Matched Filter Detection

Three schemes are primarily used for PU detection: *matched filter detection*, *energy detection* and *cyclostationary feature detection* [31]. Among them, the optimal detector in stationary Gaussian noise is the matched filter when the information of the primary user signal is known to the CR user.

The *matched filter* is the linear optimal filter used for coherent signal detection to maximize the signal-to-noise ratio (SNR) in the presence of additive stochastic noise. As shown in Figure 4.14, it is obtained by correlating a known original PU signal $s(t)$ with a received signal $r(t)$ where T_{m} is the symbol duration of PU signals. Then the output of the matched filter is sampled at the synchronized timing, i.e., every T_{M}. If the sampled value $\mathbf{Y}$ is greater than the threshold λ, the spectrum is determined to be occupied by the PU transmission. This detection method is known as an optimal detector in stationary Gaussian noise.

For more practical implementation, a pilot signal of PU systems is used for the matched filter detection [32]. In this method, PU transmitters send the pilot signal simultaneously with data, as illustrated in Figure 4.15. CR users have the perfect knowledge of the pilot signal, and are able to perform its coherent processing. The pilot signal is generally orthogonal to the data signal, and hence can be processed independently.

BASIC MODEL

We consider the matched filter to detect the signal in additive white Gaussian noise (AWGN). Then, test statistic $\mathbf{Y}$ follows the Gaussian distribution as follows [32]:

$$Y \sim \begin{cases} \mathcal{N}(0, N\sigma_{\mathrm{s}}^2\sigma_{\mathrm{n}}^2) & H_0 \\ \mathcal{N}(N\sigma_{\mathrm{s}}^2, N\sigma_{\mathrm{s}}^2\sigma_{\mathrm{n}}^2) & H_1 \end{cases} \tag{4.13}$$

where σ_{s} is the average power of the signal, σ_{n} is the noise power, and N is the number of samples.

Then the probabilities of detection and false alarm are obtained as follows:

$$P_{\mathrm{d}}(\lambda) = Pr[Y > \lambda | H_1] = Q(\frac{\lambda}{\sqrt{N\sigma_{\mathrm{s}}^2\sigma_{\mathrm{n}}^2}}) \tag{4.14}$$

$$P_{\mathrm{f}}(\lambda) = Pr[Y > \lambda | H_0] = Q(\frac{\lambda - N\sigma_{\mathrm{s}}^2}{\sqrt{N\sigma_{\mathrm{s}}^2\sigma_{\mathrm{n}}^2}}) \tag{4.15}$$

where λ is the decision threshold.

COMPLEXITY

By eliminating threshold λ in Eqs. (4.14) and (4.15), we can obtain the number of samples that the matched filter requires to achieve given target detection and false alarm probabilities as follow:

$$N = [Q^{-1}(P_{\mathrm{d}}) - Q^{-1}(P_{\mathrm{f}})]^2(\gamma)^{-1} = O(\gamma)^{-1} \tag{4.16}$$

where γ is the signal-to-noise ratio (SNR), and is obtained by $\sigma_{\mathrm{s}}^2/\sigma_{\mathrm{n}}^2$. Thus, the matched filter detection shows a fast sensing time, which requires $O(1/\gamma)$ samples to achieve a given target detection probability [31] [226]. However, the matched filter necessitates not only a priori knowledge of the characteristics of the PU signal but also the synchronization

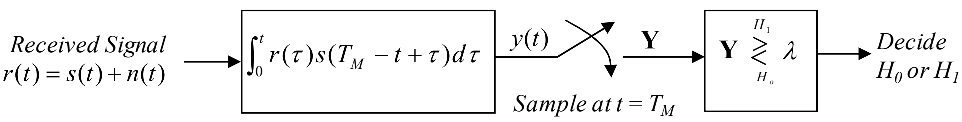


Figure 4.14 Block diagram of the matched filter detection.

between the PU transmitter and the CR user. If this information is not accurate, then the matched filter performs poorly. Furthermore, CR users need to have different multiple matched filters dedicated to each type of the PU signal, which increases the implementation cost and complexity.

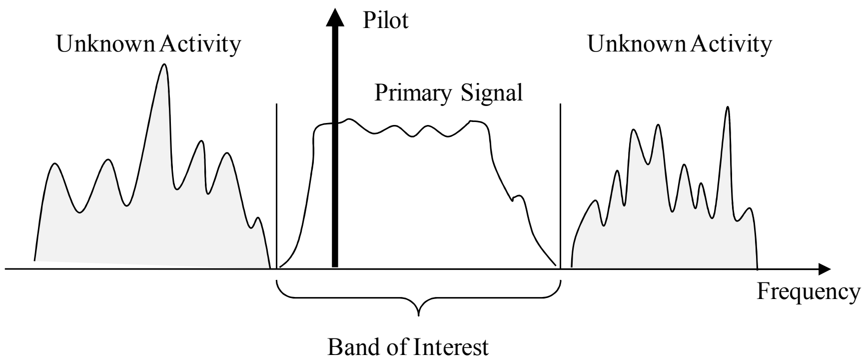


Figure 4.15 Matched filter detection based on the pilot signal [32].

4.2.3 Energy Detection

If the receiver cannot gather sufficient information about the primary user signal, for example, if the power of the random Gaussian noise is only known to the receiver, the optimal detector is an energy detector. In the energy detection, CR users sense the presence/absence of the primary users based on the energy of the received primary signal. In order to measure the energy of the received primary signal, the received signal is squared and integrated over the observation interval t_s. Finally, the output of the integrator is compared with a threshold to decide if a primary user is present (Figure 4.16).

BASIC MODEL

Furthermore, in reality, energy detection is implemented with the sampled discrete signals as shown in Figure 4.17. Accordingly, the hypothesis model presented in Eq. (4.2) can be modified as follow:

$$Y = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} n[i]^2 & H_0 \\ \frac{1}{N}\sum_{i=1}^{N} (x[i] + n[i])^2 & H_1 \end{cases} \tag{4.17}$$

where $x[i]$ and $n[i]$ are the i^{th} samples of the received signal $x(t)$, and Gaussian noise $n(t)$ in Eq. (4.2), respectively. N is the number of samples. If the energy detection can be applied in a non-fading environment, the probability of detection $P_{\rm d}$ and false alarm $P_{\rm f}$ are given as follows [67]:

$$P_{\rm d} = Pr\{Y > \lambda | H_1\} = Q_m\left(\sqrt{2\gamma}, \sqrt{\lambda}\right) \tag{4.18}$$

$$P_{\rm f} = Pr\{Y > \lambda | H_0\} = \frac{\Gamma(m, \lambda/2)}{\Gamma(m)} \tag{4.19}$$

where γ is the SNR, $\Gamma(.)$ and $\Gamma(.,.)$ are complete and incomplete gamma functions, and $Q_{\rm m}()$ is the generalized Marcum Q-function. $m = t_{\rm s}W$ is the time bandwidth product where W is the bandwidth of the spectrum band. From the above functions, while a low $P_{\rm d}$ would result in missing the presence of the primary user with high probability which in turn increases the interference to the primary user, a high $P_{\rm f}$ would result in low spectrum utilization since false alarms increase the number of missed opportunities. Since it is easy to implement, the recent work on detection of the primary user has generally adopted the energy detector [226].

ENERGY DETECTION IN FADING ENVIRONMENTS

In [98], the shadowing and the multi-path fading factors are considered for the energy detector. In this case, when the amplitude gain of the channel, h, varies due to the shadowing/fading, $P_{\rm d}$ gives the probability of the detection conditioned on instantaneous SNR as follows:

$$P_{\rm d} = \int_x Q_{\rm m}(\sqrt{2\gamma}, \sqrt{\lambda}) f_\gamma(x) dx \tag{4.20}$$

where $f_\gamma(x)$ is the probability distribution function of SNR under fading. Since $P_{\rm f}$ is independent of γ, it is the same as Eq. 4.19.

As mentioned above, the output of the integrator in the energy detector is known as the Chi-square distribution [67]. However, if the number of samples is large, we can use the central limit theorem to approximate the Chi-square distribution as Gaussian distribution [247].

$$Y \sim \begin{cases} \mathcal{N}(N{\sigma_n}^2, 2N{\sigma_n}^4), & H_0 \\ \mathcal{N}(N({\sigma_n}^2 + {\sigma_s}^2), 2N({\sigma_n}^2 + {\sigma_s}^2)^2), & H_1 \end{cases} \tag{4.21}$$

where ${\sigma_n}^2$ is the variance of the noise, and ${\sigma_s}^2$ is the variance of the received signal $s(t)$. According to the Nyquist sampling theorem, the minimum sampling rate, should be $2W$. Hence the number of observations N over $t_{\rm s}$ can be represented as $t_s/\Delta t = 2t_sW$ where Δt is the sampling time.

From Eq. (4.6), (4.7), and (4.21), $P_{\rm f}$ and $P_{\rm d}$ in MAP-based energy detection can be derived in terms of the Q function as follows:

$$P_{\rm f}(N) = Q\left(\frac{\lambda - N{\sigma_n}^2}{\sqrt{2N{\sigma_n}^4}}\right) \tag{4.22}$$

$$P_{\rm d}(N) = Q\left(\frac{\lambda - N({\sigma_s}^2 + {\sigma_n}^2)}{\sqrt{2N({\sigma_s}^2 + {\sigma_n}^2)^2}}\right) \tag{4.23}$$

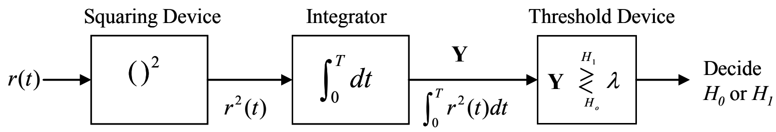


Figure 4.16 Block diagram of continuous-time energy detection.

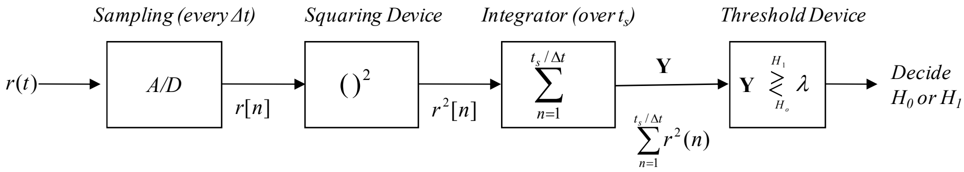


Figure 4.17 Block diagram of discrete-time energy detection.

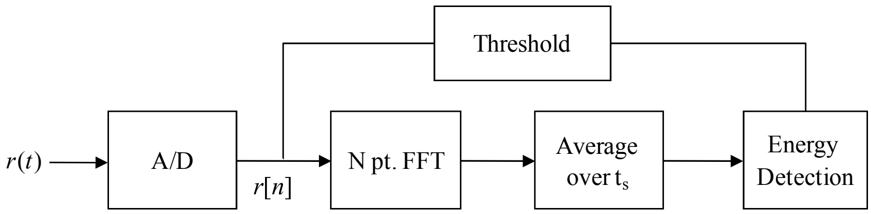


Figure 4.18 Implementation of an energy detector using Welch periodogram averaging [31].

Similarly, the detection probability under fading can be derived as follows:

$$P_{\rm d} = \int_x Q(\frac{\lambda - N({\sigma_s}^2 + {\sigma_n}^2)}{\sqrt{2N({\sigma_s}^2 + {\sigma_n}^2)^2}}) \cdot f_\gamma(x)dx \tag{4.24}$$

COMPLEXITY

From Eqs. (4.22) and (4.23), we can see that each spectrum band has different detection and false alarm probabilities according to the spectrum information, α, β, and W, as well as the observation time t_s.

By eliminating λ in Eqs. (4.22) and (4.23), we can obtain the number of samples N as follows:

$$N \simeq 2[Q^{-1}(P_{\rm f}) - Q^{-1}(P_{\rm d})(1+\gamma)]^2\gamma^{-2} \tag{4.25}$$

If γ is low, then $(\gamma + 1)$ can be approximated to 1, and hence the number of samples can be re-written as

$$N \simeq 2[Q^{-1}(P_{\rm f}) - Q^{-1}(P_{\rm d})]^2\gamma^{-2} = O(1/\gamma^2) \tag{4.26}$$

Thus, the energy detector requires $O(1/\gamma^2)$ samples for a given detection error probability. If CR users need to detect weak signals (SNR: -10dB to -40 dB), the energy detection suffers from longer detection time compared to the matched filter detection [31]. This shows that if the detector is aware of noise power, it can detect the primary signal at lower signal by increasing the number of samples by satisfying the target probabilities of detection and false alarms.

An energy detector can also be implemented similar to a spectrum analyzer by averaging frequency bins of a Fast Fourier Transform (FFT), as outlined in Figure 4.18 [31]. The processing gain is proportional to FFT size N and observation/averaging time $t_{\rm s}$. Increasing

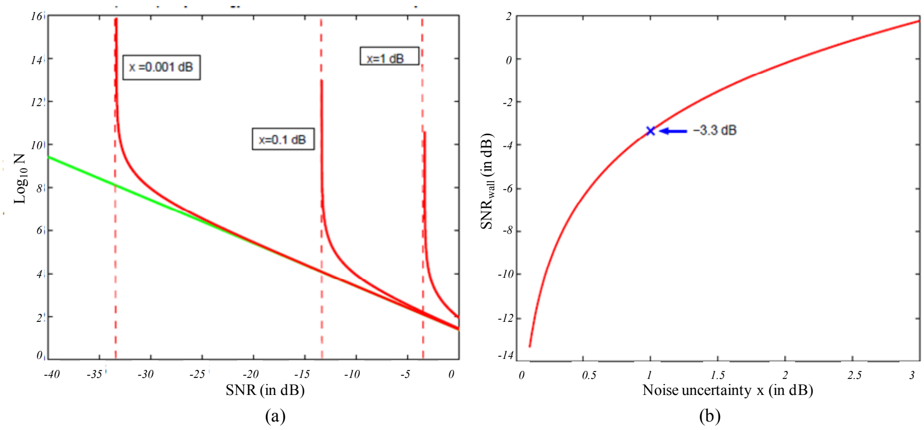


Figure 4.19 SNR wall for energy detection under noise uncertainty $x = 10\log_{10}\rho$ dB: (a)the number of samples under noise uncertainty, and (b) position of SNR wall [246].

N improves frequency resolution which helps narrowband signal detection. Also, longer averaging time reduces the noise power thus improves SNR.

SNR WALL

As mentioned above, the energy detection depends only on the SNR of the received signal, and hence its performance is susceptible to uncertainty in noise power. Noise is usually assumed to be a stationary white Gaussian, but in reality, background noise is an aggregation of interference sources such as thermal noise, leakage power from adjacent bands, interference from other users in the vicinity, etc. Thus, the noise power is uncertain and can be modeled as having any distribution in the range of $[(1/\rho)\sigma_{\rm n}^2, \rho\sigma_{\rm n}^2]$ where ρ represents the size of noise uncertainty, and is greater than unity [246]. Then, given the detection threshold λ, the minimum noise power, $1/\rho\sigma_{\rm n}^2$ yields the maximum false alarm probability, and the maximum noise power, $\rho\sigma_{\rm n}^2$ does the minimum detection probability. Both probabilities are obtained as follow:

$$P_{\rm f} = Q\left(\frac{\gamma - \rho\sigma_{\rm n}^2}{\sqrt{\frac{2}{N}}\rho\sigma_{\rm n}^2}\right) \tag{4.27}$$

$$P_{\rm d} = Q\left(\frac{\gamma - \rho(S + \frac{1}{\rho}\sigma_{\rm n}^2)}{\sqrt{\frac{2}{N}}(S + \frac{1}{\rho}\sigma_{\rm n}^2)}\right) \tag{4.28}$$

where S is the power of the received signal $h \cdot s(t)$.

Combining Eq. (4.26) with Eqs. (4.27) and (4.28), the number of samples to satisfy both detection and false alarm probabilities is obtained as follow:

$$N \approx \frac{2[Q^{-1}(P_{\rm f}) - Q^{-1}(P_{\rm d})]}{\gamma - (\rho - \frac{1}{\rho})^2} \tag{4.29}$$

Figure 4.19(a) shows how the number of samples in Eq. varies according to the SNR where N goes to infinity as the SNR approaches $10log_{10}(\rho - \frac{1}{\rho})$. This implies that no matter

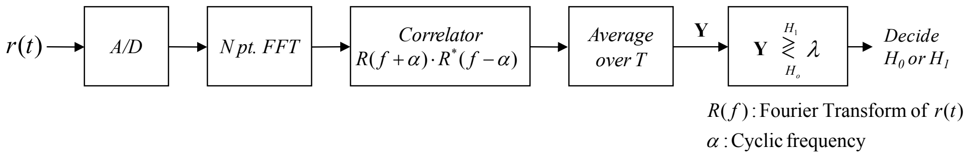


Figure 4.20 Block diagram of feature detection [31].

how many samples the energy detector has, it cannot detect the primary signal reliably if and only if the SNR is less than the certain SNR level, called the *SNR wall*. The SNR wall is expressed as follow:

$$SNR_{\text{wall}} = 10log_{10}(\rho - \frac{1}{\rho})(\text{in } dB) \tag{4.30}$$

The SNR wall is proportional to the noise uncertainty, which is shown in in Figure 4.19(b). Thus, increase in noise uncertainty significantly degrades the performance in energy detection.

While the energy detector is easy to be implemented, it can only determine the presence of the signal but cannot differentiate signal types. Thus, the energy detector often generates the false detection triggered by the unintended signals. Especially in CR ad hoc network, energy detection necessitates the synchronization over the sensing operations of all neighbors, i.e., each CR user should be synchronized with the same sensing and transmission schedules. Otherwise, CR users cannot distinguish the received signals from primary and CR users, and hence the sensing operations of the CR user will be interfered by the transmissions of its neighbors.

4.2.4 Cyclostationary Feature Detection

Feature detection determines the presence of PU signals by extracting their specific features such as pilot signals, cyclic prefixes, symbol rate, spreading codes, or modulation types from its local observation. These features introduce built-in periodicity in the modulated signals, which can be detected by analyzing a spectral correlation function. The feature detection leveraging this periodicity is also called *cyclostationary detection* [91].

BASIC PROPERTIES

The basic properties of cyclostationary random processes are presented as follows [245]. Cyclostationary processes are random processes whose statistical properties, i.e., mean and autocorrelation, vary periodically with time. Assume that $x(t)$ is a complex cyclostationary process with zero mean, its autocorrelation are periodic in time t with the same period T_0, and can be represented as a Fourier series as follows:

$$\begin{aligned} R_{\text{x}}(t - \frac{\tau}{2}, t + \frac{\tau}{2}) &= E[x(t + \frac{\tau}{2})x^*(t - \frac{\tau}{2})] \\ &= \sum_{\alpha} R_{\text{x}}^{\alpha}(\tau)e^{j2\pi\alpha t} \end{aligned} \tag{4.31}$$

$$R_{\text{x}}^{\alpha}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_{\text{x}}(t - \frac{\tau}{2}, t + \frac{\tau}{2})e^{-i2\pi\alpha t}dt \tag{4.32}$$

Table 4.1. Cyclic frequencies of various signals

Type of Signal	Cyclic frequencies
Analog television	Cyclic frequencies at multiples of the TV-signal horizontal line-scan rate (15.75 kHz in USA, 15.625 kHz in Europe)
AM signal $s(t) = a(t)\cos(2\pi f_0 t + \phi_0)$	$\pm f_0$
PM and FM signals $s(t) = \cos(2\pi f_0 t + \phi_0)$	$\pm f_0$
Amplitude shift keying $s(t) = \sum_{n=-\infty}^{\infty} (a_n p(t - nT_s - t_s)) \cos(2\pi f_0 t + \phi_0)$	k/T_0 $(k \neq 0)$, $\pm 2f_0 + k/T_0$ $(k = 0, \pm 1, \pm 2, \ldots)$
Phase Shift Keying $s(t) = \cos(2\pi f_0 t + \sum_{n=-\infty}^{\infty} (a_n p(t - nT_s - t_s)))$	$k/T_0(k \neq 0)$, $\pm 2f_0 + k/T_0(k = 0, \pm 1, \pm 2, \ldots$

where $\alpha = \frac{k}{T_0}$, $k = 0, 1, 2, ...$, and T is the sensing time. The Fourier coefficient R_{x}^{α}, which depends on the lag parameter τ, is called the *cyclic autocorrelation function*, and α is called the *cycle frequency*. For $\alpha = 0$, it reduces to the conventional autocorrelation function $R_{\mathrm{x}}^{0}(\tau)$. $R_{\mathrm{x}}^{0}(\tau)$ represents the DC component of the lag-product waveform $x(t-\frac{\tau}{2})x^*(t+\frac{\tau}{2})$ for each value of τ , and $R_{\mathrm{x}}^{\alpha}(\tau)$ can be thought of as the AC component corresponding to the frequency α. The Fourier transform of $R_{\mathrm{x}}^{\alpha}(\tau)$ is called the *spectral correlation function(SCF)*, which is given as follow:

$$S_{\mathrm{x}}^{\alpha}(f) = \int_{-\infty}^{\infty} R_{\mathrm{x}}^{\alpha}(\tau) e^{i2\pi f \tau} d\tau \tag{4.33}$$

Power spectral density is a special case of a spectral correlation function for $\alpha = 0$. For $\alpha \neq 0$, $S_{\mathrm{x}}^{\alpha}(f)$ can be considered as the density of the correlation between spectral components at frequencies $f + \frac{\alpha}{2}$ and $f - \frac{\alpha}{2}$. The cyclic frequencies for some modulated analog and digital signals are described in Table 4.1. [91]. Here noise is modeled as wide-sense stationary. It is easy to show that if $n(t)$ is wide-sense stationary, then its spectral correlation function $S_{\mathrm{n}}^{\alpha}(f) = 0$ for $\alpha \neq 0$ [90].

The following is the example to show how to obtain the spectral correlation function [91]:

- Example: Consider the noise-free AM signal $x(t) = s(t)$,

$$s(t) = a(t)\cos(2\pi f_c t + \phi_0) \tag{4.34}$$

where $a(t)$ is a stationary process. It is easily shown using Eq. (4.32) that

$$R_s^\alpha(\tau) = \begin{cases} \frac{1}{2} R_a^0(\tau)\cos(2\pi f_c \tau) & \alpha = 0 \\ \frac{1}{4} R_a^0(\tau) e^{\pm j2\phi_0} & \alpha = \pm 2f_c \end{cases} \tag{4.35}$$

and $R_s^\alpha(\tau) = 0$ for all other values of α. Or using Eq. (4.31), we obtain

$$R_s(t + \tau/2, f - \tau/2) = \frac{1}{2} R_a^0(\tau)[\cos(2\pi f_c t) + \cos(4\pi f_c t + 2\phi_0)] \tag{4.36}$$

from which Eq. (4.35) follows using Eq. (4.32). Fourier transformation of Eq. (4.35) yields

$$S_s^\alpha(f) = \begin{cases} \frac{1}{4}[S_a^0(f - f_c) + S_a^0(f + f_c)] & \alpha = 0 \\ \frac{1}{4} S_a^0(f) e^{\pm j2\phi_0} & \alpha = \pm 2f_c \end{cases} \tag{4.37}$$

Thus, the only spectral components that are correlated are those whose frequencies are separated by $|\alpha| = 2f_c$. This is easy to see intuitively since multiplication of $a(t)$ by $\cos(2\pi f_c t + \phi_0,)$ as in Eq. (4.34) shifts each spectral component in $a(t)$ up and down by the amount f. Thus, the spectral components at $f + f_c$ and $f - f_c$ in $s(t)$ are one and the same as the spectral component at f in $a(t)$. So they are obviously correlated.

In cyclostationary feature detection, the spectrum correlation of the received signal $r(t)$ is averaged over the interval T, as shown in Figure 4.20. Then, the test statistic is compared with the threshold to determine the presence of PU signals, similar to energy detection [31]. According to the the cyclostationary feature detection can be classified into *multi-cycle detection* and *single cycle detection* as follow:

MULTI-CYCLE BASED DETECTION

For weak Gaussian signals in white Guassian noise, the maximum-likelihood (ML) signal detection criterion leads to the following approximate sufficient statistic:

$$z_{\mathrm{ML}}(t) = \int_{t-T/2}^{t+T/2} \int_{t-T/2}^{t+T/2} R_x(u, v) y(u) y^*(v) du dv \tag{4.38}$$

where $y(t)$ is received signal and $R_x(u, v) = E[x(u)x^*(v)]$ is the autocorrelation function of the transmitted signal, $x(t)$. If the signal is cyclostationary, the ML detector sufficient statistic can be expressed as Eq. (4.38) It can be shown that this quadratic form is asymptotically (input SNR $\longrightarrow 0$) optimal even when the weak signal of interest is not Gaussian [89]. Thus, for weak-signal detection, the optimum detector implements a quadratic transformation of the received data and compares the resultant statistic to a threshold.

Assuming that the signal is cyclostationary, and sampled at a high enough rate, the ML detection problem is converted into a discrete time binary hypothesis testing problem with the following sufficient statistic:

$$z_{\mathrm{MC}}(N) = \sum_{\alpha} \int_{-f_s/2}^{f_s/2} S_x^{\alpha_k}(f)^* S_y^{\alpha_k}(f) df \tag{4.39}$$

where $S_x^{\alpha_k}(f)$ and $S_y^{\alpha_k}(f)$ are the Spectral Correlation Functions of $x(t)$ and $y(t)$, respectively, at the k^{th} center frequency (CF) $\alpha_k = k/T$. N is the number of samples and f_s is the sampling rate.

$S_y\alpha_k(f)$ can be defined as follow:

$$S_y(f) \triangleq \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N} Y_{\mathrm{T}}(n, f+\alpha/2)Y_{\mathrm{T}}^*(n, f-\alpha/2) \tag{4.40}$$

where $Y_{\mathrm{T}}(n, f)$ is the discrete-time Fourier transform of the signal $y[n]$ at sample n, which is obtained by

$$Y_{\mathrm{T}}(n, f) = \sum_{u=n-T/2}^{n+T/2} y[n]e^{-j2\pi \frac{f}{f_s}u} du \tag{4.41}$$

SINGLE-CYCLE BASED DETECTION

In order to reduce complexity and delay, the *single-cycle detector* measures the signal power only for one specific CF, i.e., the sufficient coefficient of the single-cycle detector is given by [91]

$$z_{\mathrm{SC}}(N) = \int_{-f_s/2}^{f_s/2} S_x^{\alpha}(f)^* S_{y_{\mathrm{T}}}(f) df \tag{4.42}$$

The statistic of the single cycle detector, $z_{\mathrm{sc}}(N)$ is a measure of the amount of spectral correlation present in the received waveform, whereas that of energy detector in Eq. (4.17), is a measure of the amount of energy present in the received waveform.

If $y[n] = x[n] + n[n]$, $S_{\mathrm{y}}^{\alpha}(f)$ is expressed as $S_{\mathrm{x}}^{\alpha}(f) + S_{\mathrm{n}}^{\alpha}(f)$ whenever $x[n]$ and $n[n]$ are independent random processes. Furthermore, as the sample size $N \to \infty$, it is known that $S_{\mathrm{x}}^{\alpha}(f) \to S_{\mathrm{x}}^{\alpha}(f)$. Therefore, as $N \to \infty$ the test statistic $Z_{\mathrm{sc}}(N|H_1) \to \int_{-\frac{f_{\mathrm{s}}}{2}}^{\frac{f_{\mathrm{s}}}{2}} |S_{\mathrm{x}}^{\alpha}(f)|^2 df$ and $Z_{\mathrm{sc}}(N|H_0) \to \int_{-\frac{f_{\mathrm{s}}}{2}}^{\frac{f_{\mathrm{s}}}{2}} S_{\mathrm{n}}^{\alpha}(f) S_{\mathrm{x}}^{\alpha}(f)^* df = 0$ as $S_{\mathrm{w}}^{\alpha}(f) = 0$ [245].

This shows that by increasing the number of samples N, detection can be made possible irrespective of the signal to noise ratio. In other words we have shown that the single cycle feature detector is robust to noise power level uncertainties.

For a signal which does not exhibit cyclostationarity, the cyclic autocorrelation function is below the threshold level for all $\alpha \neq 0$. Anyway, if $\alpha = 0$ the cyclic autocorrelation function reduces to the conventional autocorrelation function and power spectral density function, respectively. The cyclic frequencies α are typically related to the symbol rate and the carrier frequency of the signal [93].

EXAMPLES

The following are the example cases to exploit cyclostationary feature detection:

- *Digital TV Signal:* Since the DTV signal is digitally modulated, it shows the cyclostationary feature. Recently, Goh et al. [101], Han et al. [110], and Chen et al. [45] studied cyclostationary detection of ATSC and DVB-T DTV signals and investigated its performance via simulation, since the derivation of the miss-detection and false alarm probabilities of cyclostationary detectors for complex modulation schemes (e.g., 8-VSB) are known to be mathematically intractable [225].

- *Neural Network based Pattern Recognition:* In [78], the enhanced feature detection scheme combining cyclic spectral analysis with pattern recognition based on neural networks is proposed. The distinct features of the received signal are extracted using cyclic spectral analysis and represented by both spectral coherent function and spectral correlation density function. The neural network, then, classifies signals into different modulation types.

- *OFDM Signals:* The cyclostationary detector is applied to OFDM signals and particularly IEEE.802.11g compliant signals. The algorithm [127] jointly exploits the correlation induced by the cycle prefix and the fact that this correlation is time periodic, i.e. the fact that the OFDM signal is a cyclostationary signal. For each OFDM symbol, a part of its end is copied at its beginning, which is the so.called cyclic prefix. This induces a correlation between the OFDM signal and its time.shifted version. The cyclostationary detector calculates the autocorrelation function $R_y(u, m)$ of the received OFDM signal $y(u)$. $R_y(u, m)$ is a periodic function of u with period $\alpha_0^{-1} = N + D$, where N is the number of OFDM carriers and D is the length of cyclic prefix. As this function depends on u in a periodic way, the signal $y(u)$ is not a stationary but a cyclostationary signal. Its autocorrelation function can be written as a Fourier series. The Fourier coefficient $R_y^{(k\alpha_0)}(N)$ is called the cycle correlation coefficient at cyclic frequency $k\alpha_0$ and at time lag N for nonzero integer k. Therefore, the distinct features of OFDM signals with different number of carriers, symbol periods, and cyclic prefix lengths can be detected and identified in the cyclic frequency domain. In [195], it is shown that the feature detection enables the detection of the presence of the Gaussian minimum shift keying (GMSK) modulated GSM signal (PU signal) in the channel under severe interference from the orthogonal frequency division multiplexing (OFDM) based wireless LAN signal (CR signal) by exploiting different cyclic signatures of both signals.

- *UMTS FDD signals:* DS-CDMA signals can be detected by exploiting the baseband cyclostationary properties come from the redundancy between frequency components separated by multiples of the symbol rate, i.e. the cyclic feature appears at $\alpha = 1/(SF \cdot T_c)$, where SF is the spreading factor and T_c is the time chip duration. However, UMTS FDD standard employs, in addition to user specific spreading, so called scrambling sequences, in order to improve the correlation characteristics of the signals and provide base station identification [1]. Scrambling take place over multiple symbols, with period equal to 10 ms, removing the cyclostationarity with the symbol rate. Nevertheless in UMTS standard, user signals have always the same chip rate, even if the individual SF and symbol rates differ. Thus $\alpha_c = 1/T_c(3.84Mchip/s)$ is a common cyclic frequency to all downlink signals and the most appropriate to detect the received signal. An analytical formulation of the cyclic autocorrelation function for a UMTS FDD signal at $\alpha_c = 1/T_c$ can be found in [194].

  The cyclostationary feature detector exploits the cyclic frequency common to all downlink signals in a UMTS cellular scenario, which comes from the UMTS chip rate, assuming the CR user knows the UMTS carrier frequencies and bandwidths. For that, the proposed detector, using a periodogram approach, relies on second order statistics, based on spectrum cyclic density function. The output of the detector, after all signal processing, is a detection statistic, d, in dB, which represents the ratio between the power of the cyclostationary feature measured at cyclic frequency, α_c, and the estimated noise floor measured at α_n. Simulation results, considering an AWGN channel, show that for an SNR of -10 dB and an observation time of at least 30 ms it is possible to assure a 99.9probability of detection while having a negligible probability of false alarm, which is also possible for 10 ms of observation time if the SNR is at least -5 dB. An extensive analysis of the sensitivity of the algorithm to realistic impairments (synchronization, frequency offset, multipath) is extensively discussed in [196].

The main advantage of the feature detection is its robustness to the uncertainty in noise power. Furthermore, it can distinguish the signals from different networks. This method allows the CR user to perform sensing operations independently of those of its neighbors without synchronization. Although feature detection is most effective for the nature of CR Networks, it is computationally complex and requires significantly long sensing time.

4.2.5 Other Detection Schemes

As explained in the previous subsections, three detection schemes - matched filter detection, energy detection, and cyclostationary feature detection have been widely investigated for spectrum sensing in CR networks. Besides, following schemes are newly introduced for the transmitter detection:

COMPRESSED SENSING

Energy or cyclostationary detection is based on a set of observations sampled by ADC at Nyquist rate in the band of interest. Due to hardware limitation on the sampling speed, these sensing techniques are primarily used in detecting narrowband signals. To sense multiple spectrum bands spread over a wide frequency range, CR users may need to be synchronized and cooperate to sense one band at a time. Thus, there are important technical challenges in wideband spectrum sensing, including both hardware and algorithmic problems. One way to perform wideband spectrum sensing is to employ a bank of tunable narrow-band bandpass filters at the radio front-end to sense one narrow frequency band at a time. Simple algorithms such as energy or feature detection allow the detection of active users in one narrowband. As wireless communication systems of today operate on portions of spectrum having a lot of narrow frequency bands, this solution requires an excessively large number of radio frequency components.

Recent advances in compressed sensing [69] [34] [249] enables the sampling of the wideband signals at sub-Nyquist rate to relax the ADC requirements. Based on the assumption that the spectrum is under-utilized and the PU signals are sparse in wideband spectrum, the detection of primary signals in wideband spectrum is similar to the reconstruction of sparse signal in compressed sensing. Thus, the techniques of compressed sensing provide promising solutions to reliably recover wideband signals and facilitate wideband sensing at the reasonable computational complexity.

In compressed sensing, a sparse signal can be recovered by random sampling at sub-Nyquist rate as long as the sampling matrix satisfies the restricted isometry property [34] [35] [285]. For multi-step compressed sensing [249], the first step is to generate measurements x_t of size $K \times 1$ by sub-Nyquist-rate random sampling. If r_t of size $M \times 1$ is the discrete-time vector of the received wideband signal $r(t)$, the compressed sensing process can be represented by $x_t = S^T r_t$, where S^T is the $M \times K$ projection matrix, $K < M$. The second step is to reconstruct wideband spectrum $r_f = F_M r_t$ from x_t, where F_M is M-point discrete Fourier transform. To achieve this, efficient reconstruction methods such as basis pursuit can be used to solve the following convex optimization problem with the sparseness constraint in r_f [248]:

$$\hat{r}_f = \arg\min_{r_f} \|r_f\|_1 \quad \text{subject to : } x_t = (S^T F_M^{-1}) r_f \tag{4.43}$$

Once the spectrum is reconstructed, the wavelet-based spectrum detection [283] can be used to estimate the band locations in wideband spectrum and frequency response amplitude such that the bands occupied by PUs can be identified. In wideband cooperative sensing based on compressed sensing [248] [286] [285], CR users individually performs compressed sensing, cooperatively estimate the wideband spectrum by exchanging spectrum estimates, and iteratively reach a cooperative decision by exchanging local decisions. The wideband cooperative sensing schemes are discussed in Section 4.3.6.

The advantage of the compressed sensing is two-fold. First, it enables to reduce the number of samples to be stored and processed. Secondly, the reconstruction property holds for any random sampling pattern. The random sampling pattern can model multiple unsynchronised sensing devices sampling the radio signal. Then, these devices can share their computational capabilities to reconstruct the spectrum and to detect vacant bands. Unlike collaborative detection, this method does not require synchronization between collaborative sensors.

WAVELET DETECTION

Wavelet transform is a multi-resolution analysis mechanism where an input signal is decomposed into different frequency components, and then each component is studied with resolutions matched to its scales. Unlike the Fourier transform, using sines and cosines as basic functions, the wavelet transforms use irregularly shaped wavelets as basic functions and thus offer better tools to represent sharp changes and local features. For signal detection over wide-band channels, the wavelet approach offers advantages in terms of both implementation cost and flexibility in adapting to the dynamic spectrum, as opposed to the conventional use of multiple narrow-band bandpass filters.

In order to identify the locations of vacant frequency bands, the entire wide-band is modeled as a train of consecutive frequency subbands where the power spectral characteristic is smooth within each subband but changes abruptly on the border of two neighboring subbands [283]. By employing a wavelet transform of the power spectral density (PSD) of the observed signal $y(t)$, the singularities of the PSD $S_y(f)$ can be located and thus the vacant frequency bands can be found. The wavelet detection approach is particularly useful when detecting non-contiguous bands in the wide spectrum.

Conventional spectrum estimation techniques which are based on Short Time Fourier Transform (STFT) suffer from familiar problems such as low frequency resolution, variance and high side lobes/leakages. In this approach, however, the signal spectrum over a wide frequency band is decomposed into elementary building blocks of non-overlapping subbands that are well characterized by local irregularities in frequency.

The main attraction for wavelets in this application is in their ability to analyze singularities and irregular structures and the tradeoffs they provide in terms of the time-frequency resolution trade-offs. One critical challenge of implementing the wavelet approach in practice is the high sampling rates for characterizing the large bandwidth.

COVARIANCE MATRIX-BASED DETECTION

This scheme is based on the covariance matrix of signals received at CR users [287] [288]. Let $\mathbf{x}[k]$, $\mathbf{y}[k]$, and $\mathbf{n}[k]$ be the vectors of signal components in Eq. (4.17). Each vector consists of the L latest outputs at time k, which are expressed as follow:

$$\begin{aligned}\mathbf{x}[k] &\stackrel{\text{def}}{=} [x[k], x[k-1], \ldots, x[k-L+1]]^T, \\ \mathbf{y}[k] &\stackrel{\text{def}}{=} [y[k], y[k-1], \ldots, y[k-L+1]^T \\ \mathbf{n}[k] &\stackrel{\text{def}}{=} [n[k], n[k-1], \ldots, k[k-L+1]]^T\end{aligned} \tag{4.44}$$

Then the statistical covariance matrices of $\mathbf{x}[k]$, $\mathbf{y}[k]$ show the following relation:

$$\mathbf{R}_y = \mathbf{R}_x + \sigma_n^2 \mathbf{I} \tag{4.45}$$

where σ_n^2 is the variance of the noise, and $\mathbf{I}$ is the identity matrix. $\mathbf{R}_x$ and $\mathbf{R}_y$ are the covariance matrices of $\mathbf{x}[k]$, $\mathbf{y}[k]$, respectively, which are obtained by

$$\begin{aligned}\mathbf{R}_x &= E[\mathbf{x}[k]\mathbf{x}^T[k]] \\ \mathbf{R}_y &= E[\mathbf{y}[k]\mathbf{y}^T[k]]\end{aligned} \tag{4.46}$$

From this, maximum and minimum eigenvalues of $\mathbf{R}_y$ can be derived as follow:

$$\begin{aligned}\lambda_{\max} &= \rho_{\max} + \sigma_n^2 \\ \lambda_{\min} &= \rho_{\min} + \sigma_n^2\end{aligned} \tag{4.47}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of $\mathbf{R}_y$ and $\rho_{\max}$ and $\rho_{\min}$ are those of $\mathbf{R}_x$.

These statistical relations lead to the following observations:

- $\rho_{\max} = \rho_{\min}$ if and only if $\mathbf{R}_x = \delta\mathbf{I}$. However, this case is not likely to happen if signal $x[n]$ is present.
- If there is no signal,i.e., $\mathbf{R}_x = 0$, $\lambda_{\max} = \lambda_{\min}$
- Otherwise, i.e., $\mathbf{R}_x \neq \delta\mathbf{I}$ and $\mathbf{R}_x \neq 0$, $\lambda_{\max}\lambda_{\min} > 1$

Thus, the detector can use the ratio $\lambda_{\max}\lambda_{\min}$, to determine the presence of the signal.

In practice, a finite number of samples is available, and hence the sample covariance matrix can be used for detection instead of the statistic covariance matrix. Based on the sample covariance matrix, two detection methods are proposed as follows:

Algorithm 1: Maximum-minimum eigenvalue (MME) detection

- *Step 1:* Compute the sample covariance matrix of the received signal

$$R_x(N_s) \stackrel{def}{=} \frac{1}{N_s} \sum_{n=L-1}^{L-2+N_s} \hat{x}(n)\hat{x}\dagger(n), \tag{4.48}$$

where N_s is the number of collected samples.

- *Step 2:* Obtain the maximum and minimum eigenvalue of the matrix $R_x(N_s)$, that is, $\lambda_{\max}$ and $\lambda_{\min}$.
- *Step 3:* Decision: if $\lambda_{\max}/\lambda_{\min} > \gamma_1$, signal exists ("yes" decision); otherwise, signal does not exist ("no" decision), where $\gamma_1 > 1$ is a threshold.

Algorithm 2: Energy with minimum eigenvalue (EME) detection

- *Step 1:* The same as that in Algorithm 1.
- *Step 2:* Compute the average power of the received signal $T(N_s)$ (defined in (7)), and the minimum eigenvalue $\lambda_{\min}$ of the matrix Rx(Ns).
- *Step 3:* Decision: if $T(N_s)/\lambda_{\min} > \gamma_2$, signal exists ("yes" decision); otherwise, signal does not exist ("no" decision), where $\gamma_2 > 1$ is a threshold.

Given the false alarm probability, thresholds for MME and EME, γ_1 and γ_2 are derived as follow:

$$\begin{aligned}\gamma_1 &= \frac{(\sqrt{N}+\sqrt{L})^2}{(\sqrt{N}-\sqrt{L})^2}\cdot(1+\frac{(\sqrt{N}+\sqrt{L})^{(-2/3)}}{(NL)^{1/6}}F_1^{-1}(1-P_{\mathrm{f}}))\\ \gamma_2 &= \sqrt{(\frac{2}{N}})(Q^{-1}(P_{\mathrm{f}})+1)\frac{N}{(\sqrt{N}-\sqrt{L})^2}\end{aligned} \tag{4.49}$$

where $F_1(\cdot)$ is the cumulative distribution function (CDF) of the Tracy-Widom distribution of order 1, and $Q(\cdot)$ is the Q-function, which indicates tail probability of the standard normal distribution. Detailed derivations are found in [288]. In both algorithms, the thresholds are not related to noise power, unlike energy detection. The threshold can be pre-computed based only on N, L and P_{f}, irrespective of signal and noise.

In addition to using the eigenvalues of the covariance matrix described above, CR users can also use other properties of the matrix for detection [287]. If the signal $x(t)$ is not present, $R_y = \sigma_w^2 I_l$. Hence the off-diagonal elements of R_y are al zeros. If there is signal and the signal samples are correlated, R_y is not a diagonal matrix. Hence, some of the off-diagonal elements of R_y should be non-zeros. Denote r_{nm} as the element of matrix R_y as the nth row and mth column, and let

$$T_1 = \frac{1}{L}\sum_{n=1}^{L}\sum_{m=1}^{L}|r_{nm}| \tag{4.50}$$

$$T_2 = \frac{1}{L}\sum_{n=1}^{L}|r_{nm}| \tag{4.51}$$

Then, if there is no signal, $T_1/T_2 = 1$. If the signal is $T_1/T_2 > 1$. Hence the ratio T_1/T_2 can be used to detect the presence of the signal.

In practice, the statistical covariance matrix can only be calculated using a limited number of signal samples. Based on the sample covariance matrix, two following signal detection methods are proposed.

Algorithm 3: Covariance absolute value (CAV) detection

- *Step 1:* Compute the autocorrelation of the received signal $\lambda(l)$, $l = 0, 1, \ldots, L-1$, and form the sample covariance matrix.

- *Step 2:* Compute

$$\hat{T}_1 = \frac{1}{L} \sum_{n=1}^{L} \sum_{m=1}^{L} |\hat{r}_{nm}| \tag{4.52}$$

$$\hat{T}_2 = \frac{1}{L} \sum_{n=1}^{L} |\hat{r}_{nm}| \tag{4.53}$$

  where $\hat{r}_{nm}(N)$ are the elements of the sample covariance matrix $\hat{R}_y(N)$.

- *Step 3:* Determines the presence of the signal based on $\hat{T}_1(N)$, $\hat{T}_2(N)$, and the threshold γ_3, i.e., if $\hat{T}_1(N)/\hat{T}_2(N) > \gamma_3$, the signal exists, otherwise, the signal does not exist.

It is proved in [287] that for a given P_{f}, the threshold should be chosen as

$$\gamma_3 = \frac{1 + (L-1)\sqrt{\frac{2}{N\pi}}}{1 - Q^{-1}(P_{\mathrm{f}})\sqrt{\frac{2}{N}}} \tag{4.54}$$

Simulations shows that the method has similar performance to the MME. An advantage of the CAV is the reduction of computational complexity, because CAV does not need to compute the eigenvalues of the covariance matrix.

The difference between conventional energy detection and EME is as follows: energy detection compares the signal energy to the noise power, which needs to be estimated in advance, while EME compares the signal energy to the minimum eigenvalue of the sample covariance matrix, which is computed from the received signal only.

Similar to energy detection, both MME and EME only use the received signal samples for detections, and no information on the transmitted signal and channel is needed. Such methods can be called blind detection methods. The major advantage of the proposed methods over energy detection is as follows: energy detection needs the noise power for decision while the proposed methods do not need.

4.3 COOPERATION

Spectrum sensing is a key function of cognitive radio to prevent the harmful interference to licensed users and identify the available spectrum for improving the spectrum utilization. However, if each CR user depends only on its local observations in PU detection, detection performance in practice is often compromised by the following issues:

- First, the observation range of each CR user is small and typically less than its transmission range. Thus, even though CR users find the unused spectrum portion, their

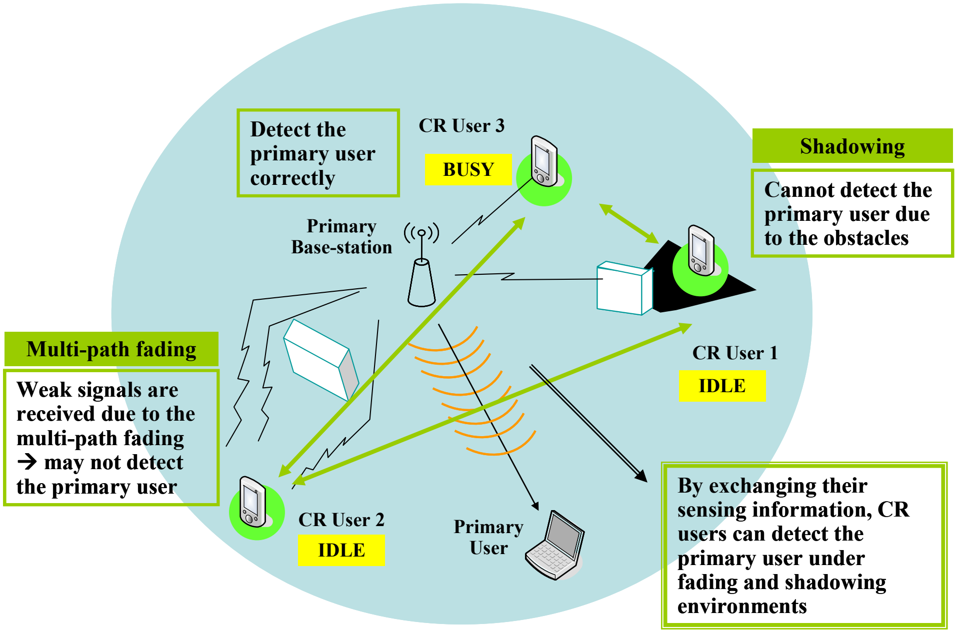


Figure 4.21 Cooperative transmitter detection under highly faded and shadowed environment.

transmission may cause interference at the primary receivers inside their transmission range, the so-called receiver uncertainty problem.

- Furthermore, multipath fading, and shadowing, can inevitably compromise the accuracy of PU detection in spectrum sensing. If the CR user receives a weak signal with a low signal-to-noise ratio (SNR) due to multi-path fading, or it is located in a shadowing area, it cannot detect the signal of the PUs.

Thus, spectrum sensing necessitates an efficient cooperation scheme to mitigate the multipath fading and shadowing effects as well as to prevent interference to PUs outside the observation range of each CR user. The idea of cooperative sensing is to exploit the spatial diversity of the different CR users in observing the spectrum. By cooperation among CR users, CR users share the sensed information to assist in making a combined decision that can be more accurate than the individual decisions.

Assume there are three CR users as illustrated in Figure 4.21. Since CR user 1 receives a weak signal (with a low SNR) due to the multi-path fading, it cannot detect the signal of the primary transmitter. CR user 2 is in the shadowing area so it cannot detect the primary user, either. Moreover, CR user 3 suffers from the receiver uncertainty problem because it is unaware of the PU transmission and the existence of primary receiver. Only CR user 4 detects the signal of the primary user correctly. As a result, CR users 1, 2, and 3 may interfere with the reception at primary receivers if they transmit based on their local observations. However, by exchanging sensing information with CR user 4, CR users 1, 2, and 3 can detect the existence of the primary user even though they are under fading and shadowing environments. This shows that cooperative sensing can effectively combat receiver uncertainty problem, multipath fading, and independent shadowing, leading to a large cooperative gain.

CLASSIFICATIONS

Cooperative spectrum sensing can be classified into three categories based on how cooperating CR users share the sensing data in the network: centralized [255] [98] [251], distributed [155], and relay-assisted [87] [88] [291].

- *Centralized Cooperative Sensing:* In centralized cooperative sensing, a central identity called fusion center (FC) controls the operations of cooperative sensing. First, the FC selects a channel or a frequency band of interest for sensing and instructs all cooperating CR users to individually perform local sensing and report the sensing results. Then the FC combines the received local sensing information, determines the presence of PUs, and diffuses the decision back to cooperating CR users. As shown in Figure 4.22(a), CR0 is the FC and CR1-CR5 are cooperating CR users performing local sensing and reporting the results back to CR0. For local sensing, all CR users are tuned to the selected licensed channel or frequency band where a physical point-to-point link, called a sensing channel, is established between the PU transmitter and each cooperating CR user for the observation of the primary signal. For data reporting, all CR users are tuned to a control channel where a physical point-to-point link, called a report channel, is established between each cooperating CR user and the FC for sending the sensing results. Note that the centralized cooperative sensing can occur in either centralized or distributed CR networks. In centralized CR networks, a CR base station (BS) is naturally the candidate of a FC. Alternatively, in CR ad hoc networks (CRAHNs) where a CR BS is not present, any CR user can act as a FC to coordinate cooperative sensing and combine the sensing information from the cooperating neighbors.i

- *Distributed Cooperative Sensing:* Unlike centralized cooperative sensing, distributed cooperative sensing does not reply on a FC for making the cooperative decision. In this case, CR users communicate among themselves and converge to a unified decision on the presence or absence of PUs by iterations. Figure 4.22(b) illustrates the cooperation in the distributed manner. After local sensing, CR1-CR5 share the local sensing results with other users within their transmission range. Based on a distributed algorithm, each CR user combines its own sensing results and the received sensing information from others, and decide whether or not the PU is present with local criteria. If the decision can not be made, CR users send their combined results to other and repeat this process until the algorithm is converged and a decision is reached. In this manner, this distributed scheme may take several iterations to reach the unanimous cooperative decision.

- *Relay-Assisted Cooperative Sensing:* In addition to centralized and distributed cooperative sensing, the third scheme is relay-assisted cooperative sensing. Since both sensing channel and report channel are not perfect, a CR user observing a weak sensing channel and a strong report channel and a CR user with a strong sensing channel and a weak report channel, for example, can complement and cooperate with each other to improve the performance of cooperative sensing. In Figure 4.22(c), CR1, CR4, and CR5, who observe strong PU signals, may suffer from a weak report channel. CR2 and CR3, who have a strong report channel, can serve as relays to assist in forwarding the sensing results from CR1, CR4, and CR5 to the FC. In this case, the report channels from CR2 and CR3 to the FC can also be called relay channels.

Note that although Figure 4.22(c) shows a centralized structure, the relay-assisted cooperative sensing can exist in distributed scheme. In fact, when the sensing results need to be forwarded by multiple hops to reach the intended receive node, all the intermediate hops are relays. Thus, if both centralized and distributed structures are one-hop cooperative sensing, the relay-assisted structure can be considered as multi-hop cooperative sensing. In addition, the relay for cooperative sensing here serves the different purpose from the relays in cooperative communications [154], where the CR relays are used for forwarding the primary user traffic.

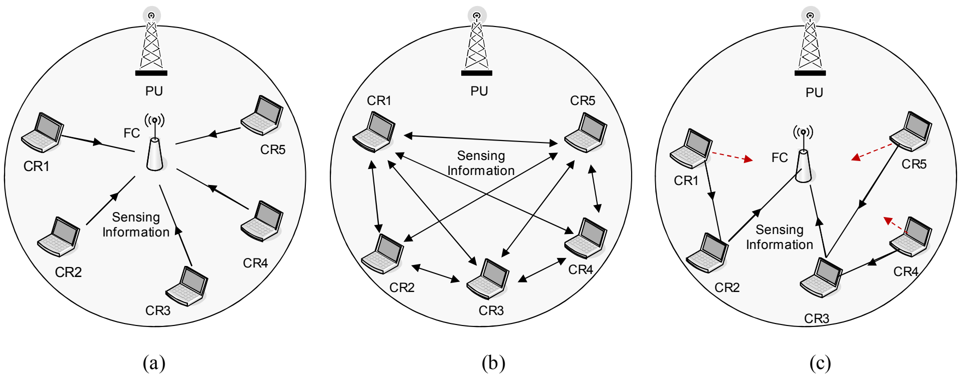


Figure 4.22 Classification of cooperative spectrum sensing: (a) centralized, (b) distributed, and (c) relay-assisted [162].

COOPERATION FRAMEWORK

Cooperation in spectrum sensing is generally considered as a four-step process: *user selection*, *local sensing*, *reporting*, and *fusion* as follow:

- *User selection :* This capability deals with how to optimally select the cooperating CR users and determine the proper cooperation footprint/range to maximize the cooperative gain and minimize the cooperation overhead.

- *Local sensing* This scheme is used to sense the RF environment, taking observation samples, and employing signal processing techniques for detecting the PU signal or the available spectrum. The choice of the sensing technique has the effect on how CR users cooperate with each other. This is related to PU detection techniques, which is explained in Section 4.2.

- *Reporting* concerns about how the sensing results obtained at cooperating CR users can be efficiently and reliably reported to the fusion center or shared with other CR users via the bandwidth-limited and fading-susceptible control channel, which is previously explained in Section 4.2.

- *Fusion:* Fusion is the process of combining the reported or shared sensing results for making the cooperative decision. Based on their data type, the sensing results can be combined by signal combining techniques or decision fusion rules.

- *Knowledge base:* The use of *knowledge base* helps to improve the detection performance. The information in the knowledge base is either a priori knowledge or the knowledge accumulated through the experience. The knowledge may include PU and CR user locations, PU activity models, and received signal strength (RSS) profiles

ADVANTAGES

The main advantages of cooperative spectrum sensing can be summarized in three folds as follow:

1. The improvement of detection performance in cooperative sensing is one type of cooperative gain. As shown in Figure 4.23, the performance degradation due to fading can be overcome by cooperative sensing such that the receiver sensitivity is close to the same level of nominal path loss without increasing the implementation cost of CR devices [183]. All these gains resulted from cooperative sensing are called cooperative gains.

2. From the perspective of sensing hardware, the requirement of the receiver sensitivity indicates the capability of detecting weak signals. Owing to multipath fading and shadowing the signal-to-noise ratio (SNR) of the received primary signal can be extremely small and the detection of which becomes a difficult task. However, to detect such a weak signal in a deep fade, the receiver will be imposed a strict sensitivity requirement that greatly increase the complexity of the implementation and the associated hardware cost. More importantly, the detection performance cannot be improved by increasing the sensitivity beyond a certain threshold, known as a SNR wall [246]. Fortunately, with cooperative sensing to combat fading, such a harsh requirement can be considerably reduced.

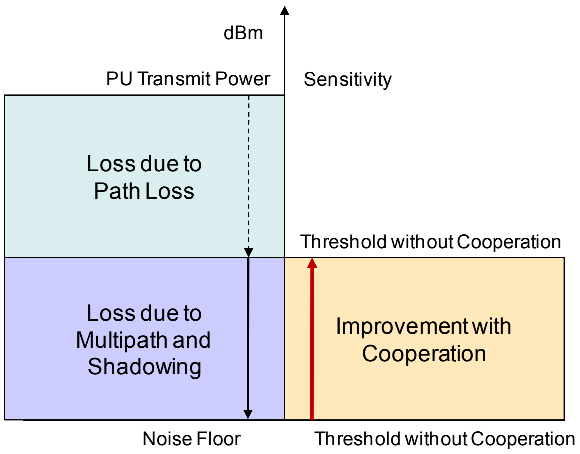


Figure 4.23 Improvement of sensitivity with cooperative sensing [183].

3. However, the cooperative gain obtained from cooperative sensing is not limited to improved detection performance and relaxed sensitivity requirement. For example, if sensing time can be reduced due to cooperative sensing, CR users will have more time for data transmissions so as to improve their throughput. In this case, the improved throughput is also one type of cooperative gain. Thus, a well-designed cooperation mechanism for cooperative sensing can significantly contribute to a variety of achievable cooperative gains.

In this section, we investigate the state of the art in cooperative sensing by first analyzing the cooperation models along with the fundamental components of cooperative sensing such as user selection, fusion, reporting, and knowledge base, and then presenting implementation issues in cooperative wideband sensing and impacting factors of incurred cooperation overhead. In addition, we identify open research challenges related to each issue in cooperative sensing along with the discussion.

4.3.1 Cooperation Model

The modeling in cooperative sensing is primarily concerned with how CR users cooperate to perform spectrum sensing and achieve the optimal detection performance. The most popular and dominating approach is originated from the parallel fusion (PF) model in distributed detection and fusion [253]. Nevertheless, recent studies [260], [262] model the behaviors of cooperating CR users in cooperative sensing by using game theory [187]. The PF models aim to achieve the detection performance by using the distributed signal processing techniques to determine how observations are combined and tested and how decisions are made. Unlike the PF models, game theoretical models focus on improving the sensing-parametric utility function by analyzing the interactions and the cooperative or non-cooperative behaviors of CR users. In this subsection, we discuss these two approaches to the modeling of CR user cooperation.

PARALLEL FUSION MODEL

In parallel fusion model, all CR users are assumed to be synchronized by the FC for sensing the channel or the frequency band of interest and reporting the sensing results. The FC combines the reported local sensing data and makes a cooperative decision, and then this decision is broadcast to all cooperating CR users. This cooperative model follows the same four-step process: user selection, local sensing, data reporting, and fusion. Most of cooperative spectrum sensing schemes [98] [251] [256] adopted the PF model or the variations of this model.

GAME THEORETIC MODEL

In game theoretical models, cooperative sensing is modeled as a game with a set of players, which are the cooperating CR users. Depending on the nature of the game, the behaviors of cooperating CR users are modeled differently. For example, in a coalitional game [260], CR users cooperate in the form of groups, called coalitions while in an evolutionary game [262], CR users are selfish users who may choose to cooperate or not cooperate depending on their own benefits.

In [260], cooperative sensing is modeled as a non-transferable (N, v) coalitional game, where N is the set of cooperating CR users and v is the utility function. The coalitional game is said to have non-transferable utility because each CR user has its own utility within the coalition. The utility of a coalition S is defined as [260]

$$v(S) = P_{d,S} - C(P_{f,S}) \tag{4.55}$$

where $P_{d,S}$ and $P_{f,S}$ are the detection and false alarm probabilities, respectively, of coalition S, and $C(P_{f,S})$ is the cost function of $P_{f,S}$ defined by a logarithmic barrier penalty function [16]. In this model, CR users can autonomously collaborate and self-organize into disjoint independent coalitions while taking into account the tradeoff between achieving maximum P_d and cost incurred in reducing P_f.

The cooperative sensing is performed in each coalition. To improve the detection performance and respond to PU activity and topology change, CR users merge or split the coalitions if the utility of the merged or split coalitions is larger than the original coalition partitions. The cooperative game model is then realized by a distributed algorithm containing three phases:

1. *Local Sensing:* Each individual CR user performs spectrum sensing locally and makes binary decisions.

2. *Adaptive Coalition Formation:* CR users interact in order to assess whether to share their sensing results with nearby coalitions. An iteration of sequential merge-and-split rules occur in the network whereby each coalition decides to merge or split if the merging or splitting results in the utility improvement.

3. *Coalition Sensing:* After the merge-and split process, the CR users in the same coalition report their local decisions to the coalition head, which can use a fusion rule to make a final cooperative decision.

In [262], distributed cooperative sensing is modeled as an evolutionary game to study the cooperative and non-cooperative behaviors of selfish CR users to maximize their own throughput. In this non-cooperative spectrum sensing game, a CR user can select an action from the action set $\{C, D\}$, where C represents that the CR user contributes to cooperative sensing and D represents that the CR user denies the participation in cooperation. On one hand, CR users can achieve stable throughput by contributing to cooperative sensing at the expense of reduced throughput due to less time for its own transmissions. On the other hand, CR users may choose not to participate in cooperative sensing to enhance its own throughput at the risk of obtaining zero throughput when no one contributes to cooperative sensing. Thus, by using replicator dynamics in evolutionary game theory, CR users interact with each other and learn the best strategy of whether or not to cooperate in cooperative sensing. When a group of heterogeneous CR users s_j with throughput C_{s_j} and the received primary signal signal-to-noise ratio (SNR) γ_j is considered in distributed cooperative sensing, the utility of each action C or D can be defined as the function of the sensing time, the number of cooperating CR users, the probabilities of detection and false alarm, and the chosen fusion rule, in addition to C_j and γ_j. The evolution dynamics of the probability of CR user s_j choosing strategy $h \in \{C, D\}$ at time t is denoted by $x_{h,s_j}(t)$ and given by [262]

$$\dot{x}_{h,s_j}(t) = [\hat{U}_{s_j}(h, x_{-s_j}) - \hat{U}_{s_j}(x)]x_{h.s_j} \tag{4.56}$$

where $\hat{U}_{s_j}(h, x_{-s_j})$ is the average utility of s_j choosing h, x_{-s_j} is the set of strategies chosen by other CR users (excluding s_j), and $\hat{U}_{s_j}(x)$ is the average utility of s_j choosing mixed strategy x_{s_j} . From Eq. 4.56, the growth rate $\dot{x}_{h,s_j}/x_{h,s_j}$ is proportional to the average utility difference of choosing pure strategy h over mixed strategy x_{h,s_j} . Thus, CR user s_j will choose h with higher probability if a higher utility can be achieved by selecting h. By the approximation of $s_j(h, x_{-s_j})$ and $s_j(x)$, a distributed learning algorithm is proposed in [262] to iteratively update the probability of choosing actions in distributed cooperative sensing and converge to the stable equilibrium. As a result, the general cooperation strategy for distributed cooperative sensing is obtained as follows. Without compromising its throughput, s_j may gradually increase (decrease) the probability of contributing to cooperative sensing x_{C,s_j} if initial x_{C,s_j} is low (high). In addition, s_j

can take advantage of other CR users with better detection performance by reducing x_{C,s_j} and cooperate with other CR users to improve detection performance by increasing x_{C,s_j}.

4.3.2 User Selection

The selection of CR users for cooperative sensing plays a key role in determining the performance of cooperative sensing because it can be utilized to improve cooperative gain and address the overhead issues. For example, when cooperating CR users experience correlated shadowing, it is shown in [183] that selecting independent CR users for cooperation can improve the robustness of sensing results. Moreover, removing malicious users from cooperation ensure the security and the reliability of the network. In Section 4.3.7, we discuss how user selection can be used to address overhead issues such as correlated shadowing, cooperation efficiency, security, energy, and mobility. In this subsection, we present the centralized and cluster-based user selection schemes in cooperative sensing.

In log-normal shadowing, the observations of two closely located CR users may be correlated due to their proximity. In this case, CR users experience similar shadowing effects called spatially correlated shadowing. In [98], it is shown that spatial correlation in shadowing can degrade the detection performance and compromise the achievable cooperative gain. In [99], it is further shown that having a small number of CR users over a large distance may be more effective than a large number of closely located users in correlated shadowing scenarios. Hence, it is important to select independent CR users for cooperation to minimize the effect of correlated shadowing. Due to its importance, spatially correlated shadowing needs to be considered in cooperative sensing. To evaluate the correlation between mobile users, a correlation model [98] [105] derived from empirical data with decaying exponential function is commonly used to determine the spatial correlation in urban and suburban environments. In general, the spatial correlation between two CR users is inversely proportional to the distance between these two.

CENTRALIZED SELECTION

The centralized selection relies on the user selection at the FC before the cooperative sensing is performed. Based on certain a priori knowledge such as the location estimates of CR users, the FC are able to select independent users for cooperation to counter the effect of correlated shadowing or other overhead.

In [233], three user selection algorithms are proposed for cooperative sensing to address the shadow correlation problem in a cellular system as follow:

1. The first algorithm aims to select a set of cooperating users with the minimum correlation measure among them by a greedy approach. Specifically, users with the largest summed correlation with respect to the remaining users are successively removed one at a time from the set until the desired number of CR users for cooperation is reached. Based on the knowledge of CR user locations, the correlation can be evaluated from the distance between two CR users.
2. Starting with the BS only in the set of cooperating users, the second algorithm selects users by successively adding uncorrelated users to the set if the selected users are located at a distance greater than the decorrelation distance d_0 from all existing members of the set.

3. The third algorithm finds K cooperating users within the radius r of the BS that satisfy the desired probability of uncorrelated K users with only the radius information from the BS to users. This method makes use of the probability of correlated shadowing between two users to compute the number of users that can be accommodated in circular cells of different sizes.

The complexity in partitioning the users into two groups: uncorrelated users and correlated users can be evaluated by two bounds: sphere packing upper bound and random selection lower bound. As in sphere packing on a hexagonal lattice, the upper bound indicates the maximum number of users that experience uncorrelated shadowing in a cell area. The lower bound is obtained by the expected number of randomly placed CR users in a cell. All aforementioned user selection algorithms perform better than the lower bound.

CLUSTER-BASED SELECTION

Centralized user selection may incur high overhead such as control channel bandwidth, energy efficiency, and reporting delay when a large number of CR users need to cooperate in sensing and report the results to the FC. To alleviate this problem, grouping the cooperating users into clusters [24] [2] [106] [137] or coalitions [260] for cooperative sensing is an effective approach to reduce the cooperation range and the incurred overhead.

In [2], four clustering methods are considered for user selection depending on the availability of location information. First, random clustering is adopted where the CR users are randomly divided into clusters of equal size when the positions of both CR users and PUs are not available. Second, reference-based clustering is based on CR user positions with respect to a given reference. In statistical clustering, clusters are formed by using the statistical information and the proximities of CR users when only the positions of CR users are known. Lastly, in distance-based clustering, only k out of K CR users closer to the PU in a cluster participate in cooperative sensing when the positions of both CR users and PUs are known.

In [24], clustering is utilized to exploit user selection diversity to improve the detection performance through reporting channels under Rayleigh fading. In each cluster, the CR user with the largest reporting channel gain is selected as the cluster head (CH) to reduce the reporting errors. The CH collects local sensing data from the members of the cluster and forward the results to the FC. The results show that this clustering method outperforms the conventional cooperative sensing scheme.

In [106], a cluster-based cooperative sensing scheme is proposed to address control channel bandwidth and sensing delay problems. The CHs are selected by the BS according to the distance from the BS and the received PU signal power. Since the overhead is reduced as the number of clusters is decreased, the method minimizes the number of clusters subject to required sensing performance. The results show that the proposed clustering method outperforms the K-mean clustering scheme.

In [137], a cluster-and-forward scheme is proposed to address the energy efficiency issue. To balance the energy consumption of users, CR users dynamically form clusters with the CH selected from the user with the largest channel gain at each time step. Moreover, to further improve the energy efficiency, the CHs take turn to act as the FC. The results show that, for each cluster size, there is an optimal number of clusters that can save the largest amount of energy.

4.3.3 Fusion Model

How to make decision based on local observations collected from multiple users is an important issue in cooperative sensing. The decision can be made through two following methods [156]:

- *Data Fusion (Soft Decision):* Collects all local observations from multiple users, and then makes the final decision.

- *Decision Fusion (Hard Decision):* Collect decisions made by each user, and makes the final decision by fusing the individual decision.

Obviously, using data fusion at the FC can achieve the best detection performance at the cost of control channel overhead. In this subsection, both fusion models are mainly discussed. combining techniques, and then focus on the fusion.

DATA FUSION

If the raw data from all receivers are sent to a central processor, the previously discussed methods for multi-antenna sensing can be directly applied. However, communication of raw data may be very expensive for practical applications. Hence, in many cases, users only send processed/compressed data to the central processor.

A simple cooperative sensing scheme based on the energy detection is the combined energy detection. For this scheme, each user computes its received source signal (including the noise) energy as and sends it to FC, which sums the collected energy values using a linear combination to obtain the following test statistic:

$$T(y) = \sum_{i=1}^{M} w_i T_i(y) \tag{4.57}$$

where M is the number of cooperating users, and w_i is the combining coefficient, with $w_i > 0$ and $\sum_{i=1}^{M} w_i = 1$. If there is no information on the source signal power received by each user, $w_i = 1/M$ for all i. If the source signal power received by each user is known, the optimal combining coefficients can be found [209] [156]. For the low-SNR case in energy detection, it can be shown that the optimal combining coefficients are given by [156]

$$w_i = \frac{|h_i|^2}{\sum_{i=1}^{M} h_i^4} \tag{4.58}$$

Existing receiver diversity techniques such as equal gain combining (EGC) and maximal ratio combining (MRC) can be utilized for data fusion of local observations or test statistics. In [169], an optimal soft combination scheme based on NP criterion is proposed to combine the weighted local observations. The proposed scheme reduces to EGC at high SNR and reduces to MRC at low SNR. Since such a soft combining scheme results in large overhead, a softened two-bit hard combining scheme is also proposed in [169] for energy detection. In this method, there are three decision thresholds dividing the whole range of test statistics into four regions. Each CR user reports the quantized two-bit information of its local test statistics. This method shows the comparable performance with the EGC scheme with less complexity and overhead.

Due to the computational complexity of the LRT-based fusion methods that involves quadratic forms, an efficient linear combination of local test statistics is proposed in [209]. In this method, the local test statistics are weighted by weighting coefficients, which are optimized based on the target P_f and P_d requirements of the CR network. Since the combining weights affect the PDF of the global statistic, a modified deflection coefficient (MDC) is introduced to measure the effect of the PDF on the detector performance. Simulation results show that maximizing the MDC can result in better detection probability. This heuristic algorithm can significantly reduce the computationally complexity of obtaining the global decision with a slight degradation in the detection performance. Overall, the optimal linear combination strategy is subject to performance degradation when the channel noise level increases.

DECISION FUSION

In decision fusion, each user sends its one-bit or multiple-bit decision to a central processor, which deploys a fusion rule to make the final decision. Specifically, if each user only sends one-bit decision ("1" for signal present and "0" for signal absent) and no other information is available at the central processor, some commonly adopted decision fusion rules are described as follows [42].

- *Optimal Decision Fusion Rule:* Let I_i be the binary decision from the ith time slot, where $I_i \in 0, 1$ for $i = 1, ..., M$. The optimal decision fusion rule is the Chair-Varshney fusion rule [92], which is a threshold test of the following statistic:

$$\Lambda_0 = \sum_{i=1}^{M} [I_i \log \frac{P_{\mathrm{d}}^{(i)}}{P_{\mathrm{f}}^{(i)}} + (1 - I_i) \log \frac{1 - P_{\mathrm{d}}^{(i)}}{1 - P_{\mathrm{f}}^{(i)}}] + \log \frac{P(H_1)}{P(H_0)} \tag{4.59}$$

  If $\Lambda_0 \geq 0$, then the primary user is present; otherwise, there is no primary user.

- *Logical OR Rule:* LO rule is a simple decision rule described as follows: if one of the decisions says that there is a primary user, then the final decision declares that there is a primary user. Mathematically, define $\Lambda = \sum_{i=1}^{M} I_i$, if $\Lambda_0 \geq 1$, then the primary user is present; otherwise, there is no primary user. Assuming that all decisions are independent, the probability of detection and probability of false alarm of the final decision are, respectively,

$$P_{\mathrm{d}} = 1 - \prod_{i=1}^{M} (1 - P_{\mathrm{d},i}) \tag{4.60}$$

$$P_{\mathrm{f}} = 1 - \prod_{i=1}^{M} (1 - P_{\mathrm{f},i}) \tag{4.61}$$

  As explained above, in this rule, the spectrum band is decided to be available only if no primary user activity is detected. Even if only one primary user activity is detected, CR users cannot use this spectrum band [183]. From this detection criterion, the cooperative detection probability $P_{\rm d}$ of N CR users is obtained by $1-(1-P_{\rm d})^M$ where $P_{\rm d}$ is the detection probability of the individual CR user.

While this decision strategy surely increases the detection probability, it increases the cooperative false alarm probability, $P_{\rm f} = 1-(1-P_{\rm f})^M$ where $P_{\rm f}$ is the false alarm probability of the individual CR user, which leads to lose more spectrum opportunities. Furthermore, cooperative approaches cause adverse effects on resource-constrained networks due to the overhead traffic.

- *Logical AND Rule:* LA rule works as follows: if all decisions says that there is a primary user, then the final decision declares that there is a primary user. Mathematically, define $\Lambda = \prod_{i=1}^{M} I_i$, if $\Lambda = 1$, then the primary user is present; otherwise, there is no primary user. Again, assuming that all decisions are independent, the probability of detection and probability of false alarm of the final decision are, respectively,

$$P_{\rm d} = \prod_{i=1}^{M} P_{{\rm d},i} \tag{4.62}$$

$$P_{\rm f} = \prod_{i=1}^{M} P_{{\rm f},i} \tag{4.63}$$

- *Majority Rule:* Another decision rule is based on majority of the individual decisions. If half of the decisions or more say that there is a primary user, then the final decision declares that there is a primary user. The number of detections follows the binomial distribution $\mathcal{B}(N, \bar{P}_d)$. Similarly, the number of false alarms also shows the binomial distribution $\mathcal{B}(N, \bar{P}_f)$. Thus, in order to determine the detection threshold M_{th} based on the majority rule, we, mathematically, define $\Lambda = \sum_{i=1}^{M} I_i$, if $\Lambda \geq \lceil \frac{M}{2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer not less then x, then the primary user is declared to be present; otherwise, there is no primary user. Assuming that all decisions are independent, and supposing that $P_{{\rm d},1} = P_{{\rm d},2} = ... = P_{{\rm d},M} = P_{{\rm d},0}$ and $P_{{\rm f},1} = P_{{\rm f},2} = ... = P_{{\rm f},M} = P_{{\rm f},0}$, the probability of detection and probability of false alarm of the final decision are given by,

$$P_{\rm d} = \sum_{i=0}^{M-M_{\rm th}} \binom{M}{\lceil \frac{M}{2} \rceil + i} (1-P_{{\rm d},0})^{M-\lceil \frac{M}{2} \rceil - i} \cdot P_{{\rm d},0}^{\lceil \frac{M}{2} \rceil + i} \tag{4.64}$$

$$P_{\rm f} = \sum_{i=0}^{M-M_{\rm th}} \binom{M}{\lceil \frac{M}{2} \rceil + i} (1-P_{{\rm f},0})^{M-\lceil \frac{M}{2} \rceil - i} \cdot P_{{\rm f},0}^{\lceil \frac{M}{2} \rceil + i} \tag{4.65}$$

respectively, where $M_{\rm th}$ is the decision threshold, $\lceil \frac{M}{2} \rceil$, and $\binom{M}{k} = \frac{k!}{k!(M-k)!}$

Furthermore, in order to determine the detection threshold M_{th} to balance between the detection error probability and the false alarm probability, we exploit the following strategy:

$$1 - P_{\rm bd}(M_{\rm th}) = P_{\rm bf}(M_{\rm th}) \tag{4.66}$$

where $P_{\rm bd}$ is the binomial cumulative distribution function (CDF) of the number of detections, and $P_{\rm bf}$ is the binomial CDF of the number of false alarms.

If the simple fusion rule is not used, advanced fusion techniques can be devised to utilize the statistical knowledge for decision fusion. In [251], a linear-quadratic (LQ) fusion method is proposed to consider the correlation between CR users in cooperative sensing. With the binary local decisions reported by cooperating CR users, this method provides a suboptimal solution to the decision fusion problem by using the partial statistical knowledge: the second-order statistics of the local decisions under H1 and the fourth-order statistics under H0. Based on the deflection criterion, the LQ detector compares a LQ function of the local decisions with a predetermined threshold and achieves better error probability with a higher value of deflection. The results show that the proposed scheme outperforms the Counting Rule in correlated shadowing.

Although cooperative sensing can achieve better performance, there are some issues associated with it. First, reliable information exchanges among the cooperating users must be guaranteed. In an ad hoc network, this is by no means a simple task. Second, most data fusion methods in literature are based on the simple energy detection and flat-fading channel model, while more advanced data fusion algorithms such as cyclostationary detection, space-time combining, and eigenvalue-based detection, over more practical propagation channels need to be further investigated. Third, existing decision fusions have mostly assumed that decisions of different users are independent, which may not be true because all users actually receive signals from some common sources. Next. the theoretical work on cooperative spectrum sensing reveals important tradeoffs for the design of CR networks. As expected, it has been shown that cooperative settings result in higher utilization of the spectrum as well as fairness. However, this advantage may not be so high considering the cost of cooperation due to frequent information exchange among users. At last, practical fusion algorithms should be robust to data errors due to channel impairment, interference, and noise.

4.3.4 Reporting

In cooperative sensing, a common control channel (CCC) is commonly used by CR users to report local sensing data to the FC or share the sensing results with neighboring nodes. As a result, control channel and reporting is one of the elements of cooperative sensing. The control channel can be implemented as a dedicated channel in licensed or unlicensed band, or an underlay ultra-wideband (UWB) channel [31]. A medium access control (MAC) scheme for multiple access is generally used by all cooperating CR users to access the control channel. From the perspective of physical layer, a physical point-to-point link from a cooperating CR user to the FC is called a reporting channel.

For control channel and reporting, three major requirements must be satisfied in cooperative sensing: bandwidth, reliability, and security. In this subsection, bandwidth and reliability requirements are investigated, and the security requirement for control channel jamming is addressed in Chapter 11.

BANDWIDTH REQUIREMENT

The bandwidth of the control channel is identified in [183] as the key impact factor of the performance in cooperative sensing. The amount of local sensing data that can be transmitted to the FC is limited by the control channel bandwidth. In [24], the problem of cooperative sensing under control channel bandwidth constraints is addressed by censoring and quantizing local sensing data. Each cooperating CR user performs the censoring by reporting the result only if the local decision is determined by the SPRT test. Thus, censoring reduces the unnecessary reporting and the usage of control channel bandwidth. In [300], a bandwidth-efficient combination scheme is proposed to enable the simultaneous reporting to the FC with the fixed required control channel bandwidth in cooperative sensing, regardless of the number of cooperating CR users. The test statistics for testing the superposition of all received local sensing data are devised for Gaussian and Rayleigh fading reporting channels.

RELIABILITY REQUIREMENT

In addition to bandwidth requirement, the reliability of the control channel has the great impact on cooperative sensing performance. Like data channels, the control channel is susceptible to multipath fading and shadowing. Hence, the channel impairments must be considered in the reliability issue of control channel. While early studies [31] [98] [255] assume a perfect error-free control channel in cooperative sensing, recent studies investigate the effect of Gaussian noise [209], multipath fading [291], and correlated shadowing [216] on the control channel and the sensing performance.

In [291], a transmit diversity-based cooperative sensing method is proposed to address the performance degradation caused by reporting channels under fading. Due to the reporting errors, the results show that the probability of false alarm Q_f is lower bounded and linearly increases with the probability of reporting errors. In addition, a censor-and-relay method is proposed for the FC to censor the received results from unreliable reporting channels. The CR users who do not have good reporting channels are instructed to forward their sensing results to those neighbors in good reporting channel conditions. These neighbors then report its own results and relay other's forwarded results through orthogonal control channels to avoid the mutual interference.

In [217] [216], the issue of correlated log-normal shadowing on the reporting channel is investigated. The results show that the performance degradation caused by the shadowing correlation on the reporting channel is similar to that on the sensing channel.

4.3.5 Knowledge Base

The performance of cooperative sensing schemes largely depends on the knowledge of PU characteristics such as traffic patterns, location, and transmit power. The PU information, if available in a database, can facilitate the PU detection. The database that stores all the knowledge of the RF environments is called a knowledge base. Knowledge base is an indispensable element of cooperative sensing because it can be utilized to assist, complement, or even replace cooperative sensing for detecting PU signals and identifying the available spectrum. Knowledge base serves as two roles in cooperative sensing: i) to enhance the detection performance by utilizing the accumulated knowledge and the learned experience such as statistical models in the database and ii) to alleviate the burden of cooperative sensing by retrieving the spectrum information such as a list of PU-occupied channels from the database. The knowledge base can provide PU information such as locations, tracking, transmit power, and activity in the forms of spatial-temporal-spectral maps for cooperative sensing. In this subsection, we discuss the following knowledge base approaches: radio environment map (REM) [296], received signal strength (RSS) profiles [182], channel gain map [144] [145], and power spectral density (PSD) map [10].

RADIO ENVIRONMENTAL MAPS (REMs)

Radio environment map (REM) [296] is a central database that can, among other things, be used as the infrastructure in CR networks to provide radio environment information for spectrum access, such as the locations of CR users, available spectrum, spectrum regulation and policies, shadowing areas, and PU signal types. In cooperative sensing, all the environment information, if available, can be accessed and utilized by each CR user to improve the detection performance in local sensing and in cooperative sensing. However, REMs may lead to large communication overhead due to a large amount of information transferred among CR users. More details are explained in Section 5.1.4.

SPATIAL RECEIVED SIGNAL STRENGTH PROFILES

In [65], a mechanism to establish spatial received signal strength (RSS) profiles is proposed for cooperative sensing. In this scheme, each cooperating CR user accumulates the RSS samples to establish the distribution of test statistics at each CR user location. When all these temporal profiles from different CR user locations are combined at the FC, the spatial RSS profile is constructed and can be used as the detection criterion at the FC. In cooperative sensing, the FC can determine the presence of PU signals if the observed RSS values are similar to those in the profile. When PU signals are not present, each CR user estimates the noise power distribution for RSS profiles. The training period of performing RSS profiling should be long enough to accurately estimate the RSS distributions. The frequency of updating the RSS profile can be determined based on the time variation of the RSS profiles.

POWER SPATIAL DENSITY MAP

In [10], a distributed cooperative sensing scheme based on power spectral density (PSD) maps is proposed for CRAHNs. In this scheme, CR users locally collect PSD samples and cooperatively estimate the basis expansion coeffcients of the PSD map by exchanging messages with one-hop neighbors. The consensus on the estimates is reached by using the distributed least-absolute shrinkage and selection operator (D-Lasso) algorithm. In addition, the exponentially weighted moving average (EWMA) method is utilized to track the slowly varying PSDs. Due to the narrow-band PSDs of PU signals in the wideband spectrum and the sparsely located PUs with active signals in a given area, the sparsity in both frequency and space are also exploited to formulate the non-negative Lasso criterion for ‘1-norm minimization of the unknowns. With the constructed PSD maps, this method is able to adapt to the environment change and track the locations and the power of PU transmitters.

CHANNEL GAIN MAPS

In [66, 67], a cooperative sensing scheme by using channel gain maps is proposed to track the PU locations and their transmit power. In this scheme, each CR user maintains a map of channel gain that consists of path loss, shadowing, and fading components. By extending the Kalman filter with the linear spatial interpolator, the Kriged Kalman filtering is used for tracking shadow fading at any point in an area. Similar to [68], cooperative sensing is formulated as a sparse regression problem with time-weighted non-negative Lasso to exploit the sparsity of PU locations. Based on the established channel gain maps, a centralized algorithm and a distributed algorithm using alternating direction method of multipliers (ADMoM) are used for tracking PU locations.

4.3.6 Wideband Cooperative Sensing

Conventional cooperative sensing exploits the spatial diversity of cooperating CR users and focuses on the sensing of one frequency band during each round of cooperation. To determine the spectrum availability in multiple channels or bands, CR users need to be synchronized to switch to another band and perform cooperative sensing separately in each band. This process can incur significant switching delay and synchronization overhead. Alternatively, CR users can cooperatively sense multiple channels or frequency bands to reduce the total sensing time for all users. In this subsection, we discuss multi-band cooperative sensing [208] [223] and wideband cooperative sensing [248] [285].

MULTI-BAND COOPERATIVE SENSING

In multi-band cooperative sensing, CR users cooperate to sense multiple narrow bands instead of focusing on one band at a time. In [208], a spatio-spectral joint detection (SSJD) scheme is proposed for combining the statistics of sensing K bands from M spatially distributed CR users. The FC calculates the test statistic and make a cooperative decision in each band. The weight coefficients and detection thresholds of all bands are obtained by jointly maximizing the aggregate CR throughput in each band subject to miss detection and false alarm probability constraints. To enable the multi-band sensing at each CR user, an energy detector is required for each band of interest. As a result, the method may incur higher hardware cost when the number of bands for cooperative sensing is large.

In [223], a parallel cooperative sensing scheme is proposed to enable the multi-channel sensing by optimally selected cooperating CR users. Unlike the multi-band sensing scheme in [208], each of cooperating CR users senses a different channel. By this method, multiple channels can be cooperatively sensed in each sensing period. The objective is to maximize the CR throughput while minimizing the sensing overhead such as the sensing time and the number of required CR users for cooperation.

WIDEBAND COOPERATIVE SENSING

As discussed in Section 3.2, compressed sensing techniques facilitate wideband sensing with the sampling at sub-Nyquist rate. Based on the assumption that the wideband spectrum is sparsely occupied by PUs, the spectrum of the wideband signal can be reconstructed for PU detection. Thus, we focus on the wideband cooperative sensing schemes utilizing compressed sensing [248] [286] [285].

In [248], a distributed cooperative sensing scheme is proposed for wideband sensing in CRAHNs. In this scheme, each CR user performs compressed sensing locally, determines the local spectral estimates, and sends the spectrum state vectors to one-hop neighbors. By using the distributed average consensus method, each CR user iteratively updates its spectrum state vectors with the weighted sum of the difference values between the CR user and its neighbors. As a result, the spectrum state vectors converge to the average statistic at each CR user for PU detection. Similarly, the spectral estimates can be obtained cooperatively by the consensus averaging.

In [286], the work in [248] is extended to consider the spectrum occupied by CR users, called spectral innovation, in addition to PUs in wideband sensing. The accuracy of spectrum estimation is improved by utilizing the spectral orthogonality between PUs and CR users. Based on the work in [248] [286], a distributed consensus optimization scheme is proposed in [285] for wideband sensing in CRAHNs. After compressed sensing, each CR user finds spectrum estimates by performing the consensus optimization for global optimality and broadcasts it to one-hop neighbors. This process is repeated until the convergence is reached. The average consensus technique incorporated in the constraints ensure the fast convergence. In addition, this method is also considered in the presence of spectral innovation.

4.3.7 Cooperation Overhead

The exploitation of spatial diversity in cooperative sensing results in significant improvement in detection performance. The performance improvement as the result of cooperation is termed diversity gain or cooperative gain [183]. Regardless of the improvement of detection performance, cooperation among CR users may also introduce a variety of overhead that limits or even compromises the achievable cooperative gain. The overhead associated with all elements of cooperative sensing are called cooperation overhead. In this section, we consider the issues of achievable cooperative gain and incurred cooperation overhead in cooperative sensing. These issues, called dominating factors of the cooperative gain and cooperation overhead, include 1) reporting delay, 2) synchronization, 3) control channel requirements, 4) mobility, 5) energy efficiency, and 6) wideband sensing, which are extensively discussed as follows.

REPORTING DELAY

In cooperative sensing, sharing local sensing data with the FC or other CR users incurs reporting delay. This is the overhead because it does not exist in spectrum sensing with no cooperation. In addition to transmission delay from the cooperating CR user to the FC, there are many reasons that can result in reporting delay. First, if cooperating CR users transmit on the control channel by a random access scheme, it is possible that the control messages sent from different CR users collide and the re-transmission is required. Moreover, delivering the sensing data by multiple hops such as the case in the relay-assisted cooperative sensing incurs extra reporting delay. Thus, reporting delay is the overhead that should be considered in cooperative sensing schemes.

In [99], the authors address the issue of cooperation processing tradeoff in cooperative sensing. The tradeoff is formulated as an optimization problem to minimize the total sensing time subject to constraints of false alarm and detection probabilities. The total sensing time to be minimized includes the integration time of the energy detector for local processing and the reporting time, proportional to the number of cooperating CR users, for cooperation. The results show that, for higher detection sensitivity, the longer integration time is generally required. However, with cooperation, the increasing number of cooperating CR users reduces the required sensing time to achieve the same level of detection sensitivity, even the reporting delay is longer in this case.

In [161], a reinforcement learning-based cooperative sensing scheme is proposed to minimize the cooperative sensing delay and improve the detection probability in spatially correlated shadowing. By considering the reporting delay and spatial correlation among CR users in calculating the reward functions, the learning algorithm effectively finds the optimal solution to obtain the optimal sensing /report sequence and minimize the total reporting delay from all cooperating CR users while the detection performance is improved in correlated shadowing.

SYNCHRONIZATION ISSUE

In addition to delays, many cooperative sensing schemes [255] require the synchronization of all the cooperating users and rely on simultaneous reporting of the CR users to perform the likelihood ratio testing. For example, due to the lack of the capability for distinguishing the PU signal from CR signals, spectrum sensing with energy detectors requires a scheduled quiet period for simultaneous local sensing operations. However, the synchronization may not be easily achieved for a large amount of CR users in CRAHNs. Thus, many asynchronous cooperative sensing methods [239] [301] [300] are proposed to deal with this issue. In [239], a sliding window algorithm is proposed to resolve the synchronization issue by detecting the change point sequentially in the sensing reports received within an observation window. Similar to SPRT, the window is advanced if more sensing reports are required to make a decision. Compared to WSPRT method [42], this method is able to achieve higher detection accuracy and reduce the detection time with and without misbehaving users. In [301] [302], a probability-based combination scheme is proposed to combine asynchronous reports at the FC. Based on the knowledge of PU ON/OFF period distribution and Bayesian decision rule, the conditional probability of the sensing reports received at different time and their combined likelihood ratio can be calculated to make the final decision.

MOBILITY

Most existing cooperative sensing techniques do not consider the movement of PUs and CR users during cooperative sensing. However, the mobility of PUs and CR users may have the impact on the detection performance in cooperative sensing as follow:

- *Primary User Mobility:* For large-scale PUs such as TV powers or cellular base station s, it is a reasonable assumption that the PUs are stationary. On the other hand, for small-scale PUs, such as wireless microphones in IEEE 802.22 or radios in emergency and military networks, the PUs can be mobile. The detection of small-scale PUs by an individual CR user is a challenge owing to their small transmit power and mobility. Thus, cooperative sensing with the assistance of PU tracking methods and the spectrum knowledge base could be the solution to this problem. To the best of our knowledge, no solution has yet been proposed to consider the impact of PU mobility on cooperative sensing.

- *CR User Mobility:* Intuitively, the movement of a CR user creates the spatial diversity in the observations taken on the move. As a result, a mobile CR user can improve the detection performance with its local samples and require less cooperation from others to reduce the cooperation overhead, depending on the speed and the direction of the movement and the location of the cooperating CR users. However, it is also likely that the mobility creates the correlation among CR users if the distance between CR users may be reduced by CR user movement. In addition, the network topology changes as CR users move. In this case, CR users may need to join or leave the group of cooperating CR users similar to the merge and split of coalitions in [260]. Thus, all the cooperation overhead due to mobility must be considered in cooperative sensing of mobile CR users.

In [182], the impact of mobility on spectrum sensing is investigated. For a single mobile CR user with energy detection, it is shown that the mobility increases the spatio-temporal diversity in the received PU signals. Without the cooperation from other users, the CR user mobility can improve the detection performance with the increasing moving speed. This is because the observations are less correlated as the speed is increased. Moreover, higher mobility speed can reduce the frequency of scheduled sensing for a given detection performance. This reduces the frequency of periodic sensing and the overall sensing time. It is also implied that it is more efficient to cooperate with other users than scheduling multiple times of sensing when CR users are slowly moving. Conversely, when CR users are moving at high speed, it is more efficient to sense individually multiple times than cooperate with other users. In addition, the number of cooperating CR users can be decreased if the number of times to perform sensing is increased. This results in the tradeoff between cooperation and scheduling. However, the mobility speed also reduces the average received signal strength. Thus, the degradation in sensitivity of energy detection must be compensated by the spatio-temporal diversity. These results imply that CR users can reduce the cooperation overhead with the speed of mobility if the independent observations with the spatio-temporal diversity can be obtained.

ENERGY EFFICIENCY

In cooperative sensing, CR users involves activity such as local sensing and data reporting that consumes additional energy. The energy consumption overhead can be significant if the number of cooperating CR users or the amount of sensing results for report is large. Thus, energy efficiency should be considered in cooperative sensing schemes. To address this issue, existing solutions reduce energy consumption by two main approaches: reducing the amount of reporting data by censoring [24] [164] and improving energy efficiency by optimization [203] [175].

- *Censoring:* Censoring was introduced in sensor networks as an energy-efficient technique for distributed detection [23] [8]. In cooperative sensing, it is used to limit the amount of reported sensing data according to certain criteria or constraints. Since the censoring criteria are chosen to refrain cooperating CR users from transmitting unnecessary or uninformative data, the energy efficiency can be improved in cooperative sensing. In addition, censoring can also lower the control channel bandwidth requirement (Section 4.3.4) due to the reduced number of control messages.

  In [24], a simple censoring method is proposed to decrease the average number of sensing bits reported to the FC. Similar to the SPRT, the energy detector output O_i of CR user i is compared to two thresholds λ_1 and λ_2, $\lambda_1 < \lambda_2$. If O_i is smaller than λ_1 or larger than λ_2, decision 0 or 1 is determined, respectively, and sent to the FC. Otherwise, no decision is made and this sensing output is censored from reporting. The results show that even though the Q_f may degrade due to the possibility that the sensing outputs of all CR users are censored, the amount of reported local decisions can be dramatically reduced. Thus, the energy efficiency can be traded off with Q_f . In [163] [164], a censoring scheme with communication rate constraints is proposed to reduce energy consumption in cyclostationarity-based cooperative sensing. In this scheme, CR users send the test statistic from the cyclostationary detector $\mathcal{T}^{(i)}$ to the FC if the following constraint is satisfied:

$$p(\mathcal{T}^{(i)} > t_i|H_0) \leq k_i, \quad i = 1, ..., L \tag{4.67}$$

where t_i is the threshold such that the probability of CR user i sends the test statistic under null hypothesis H_0 is i, i is the communication rate constraint of user i for reporting sensing data, and L is the number of cooperating CR users. As a result, energy efficiency is improved by independently selecting i for each user i based on the required detection performance. It is proven in [8] that, for $t_i = 0$, the probability of miss detection is minimized if the communication rate constraints i in Eq. 4.67 are chosen such that the probability of false alarm is less or equal to $1 - \prod_{i=1}^{L}(1 - k_i)$.

- *Energy Minimization:* Another approach to improve energy efficiency is to optimize the CR performance with energy constraints [47] or minimize energy consumption with detection performance constraints [203] [175]. In [203], the energy efficiency problem is addressed by energy minimization under detection performance constraints. Specifically, this method investigates the tradeoff between the two aspects of sensing time. On one hand, longer sensing time consumes more energy of each CR user. On the other hand, longer sensing time can improve detection performance at each CR user and reduce the number of cooperating users and the associated energy consumption overhead. Thus, this method finds the optimal sensing time and the optimal number of cooperating users to balance the energy consumption in local sensing and the energy overhead due to cooperation for required detection performance.

In [175], a sleeping and censoring combined scheme is proposed to jointly optimize the energy consumption cost under the detection constraints. Specifically, to find the optimal sleeping rate μ and the censoring thresholds, λ_1 and λ_2, the optimization problem is formulated as

$$\begin{aligned} &\min_{\mu,\lambda_1,\lambda_2} \sum_{i=1}^{N}(C_{s_i} + C_{t_i}(1-\rho)) \\ &\text{subject to}:\ Q_f \le \alpha, Q_d \ge \beta \end{aligned} \tag{4.68}$$

where C_{s_i} and C_{t_i} are the energy cost of CR user i in sensing and transmission, respectively, $\rho = Pr(\lambda_1 < E_i < \lambda_2)$ is probability of CR user i's energy detector output Ei being censored, and N is the number of cooperating CR users. The results show that this method significantly reduces energy consumption with or without a priori knowledge of PU activity. Moreover, for $\alpha = 0.1$ and $\beta = 0.9$, as the number of cooperating users increases, the optimal sleeping rate increases dramatically to minimize the overall energy consumption in cooperative sensing.

4.4 SENSING CONTROL

The main objective of spectrum sensing is to find more spectrum access opportunities without interfering with primary networks. To this end, the sensing operations of CR users are controlled and coordinated by a sensing controller, which considers two main issues on 1) how long and frequently CR users should sense the spectrum to achieve sufficient sensing accuracy in in-band sensing, and 2) how quickly CR user can find the available spectrum band in out-of-band sensing, which are summarized in Figure 4.24.

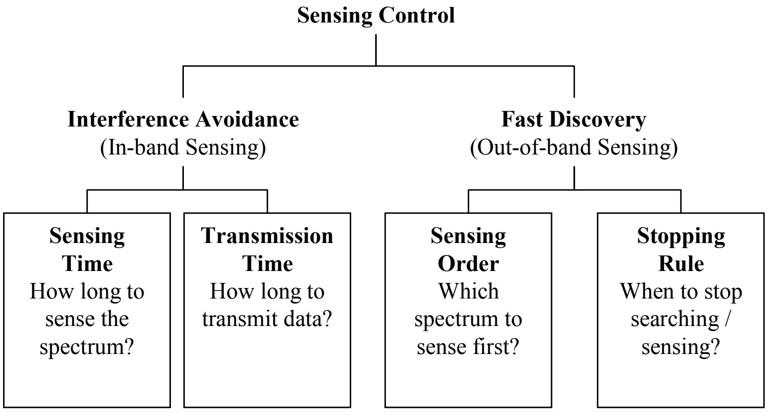


Figure 4.24 Configuration parameters coordinated by sensing control.

4.4.1 In-band Sensing Control

The first issue is related to *in-band sensing*, which aims at maximum spectrum opportunity as well as interference avoidance. The in-band sensing generally adopts the periodic sensing structure where CR users are allowed to access the spectrum only during the transmission period followed by sensing (observation) period. In the periodic sensing, longer sensing time leads to higher sensing accuracy, and hence to less interference. But as the sensing time becomes longer, the transmission time of CR users will be decreased. Conversely, while longer transmission time increases the access opportunities, it causes higher interference due to the lack of sensing information. Thus, how to select the proper sensing and transmission times is an important issue in spectrum sensing. In this section, first, we investigate a periodic sensing model, and then explain three different approaches for in-band sensing.

PERIODIC SENSING MODEL

With energy detection, mostly used in the spectrum sensing, CR users are not able to perform the transmission and sensing tasks at the same time. Thus, due to this hardware limitation, CR users necessitate a *periodic sensing* structure where sensing and transmission operations are performed in a periodic manner with separate observation period and transmission period. In this structure, CR users should stop their transmissions during the sensing time to prevent false alarms triggered by unintended CR signals This periodic sensing structure introduces the following design issues [150]:

- *Interference avoidance:* Interference in CR networks depends on the sensing accuracy, which is determined by the observation time. However, in periodic sensing, CR users cannot sense the spectrum bands during the transmission time, which leads to the increase in interference. Thus, for the interference avoidance, both the observation time and the transmission time need to be considered in the periodic spectrum sensing method.

- *Sensing efficiency:* The main objective of CR networks is efficient spectrum utilization. Thus, the spectrum sensing functionality should provide more transmission opportunities to CR users. However, during the observation period, the transmission

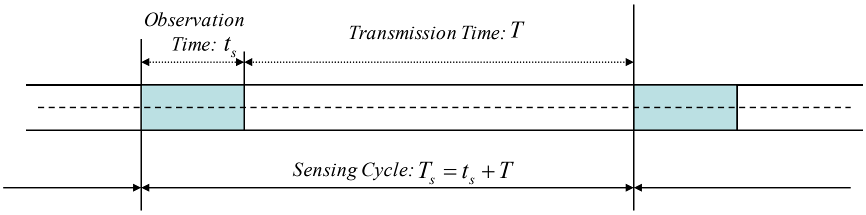


Figure 4.25 Periodic sensing structure.

of CR users is not allowed, which inevitably decreases the transmission opportunities of CR users, leading to the so-called *sensing efficiency* issue.

The in-band sensing generally adopts a periodic sensing structure where CR users are allowed to access the spectrum only during the transmission period followed by sensing (observation) period. As explained above, there is a tradeoff between interference and sensing efficiency. For interference avoidance, the observation time needs to be long enough to achieve sufficient detection accuracy, i.e., longer observation time leads to higher sensing accuracy, and hence to less interference. But as the observation time becomes longer, the transmission time of CR users will be decreased. Conversely, while a longer transmission time enhances the sensing efficiency, it causes higher interference due to the lack of sensing information. Hence, *observation time* and *transmission time* are the sensing parameters that mainly influence both the spectrum efficiency and interference avoidance. Thus, the proper selection of these sensing parameters is the most critical factor influencing the performance of CR networks.

Consider a typical sensing scenario in which a single CR user monitors a single spectrum band. The CR user alternately senses the spectrum and transmits data with observation time t_s and transmission time T, as shown in Figure 4.25. To determine these sensing parameters accurately, we need to consider the interference constraint and the sensing efficiency at the same time. To this end, we introduce the following definitions [150]:

Definition 1: The *interference ratio* T_I is the expected fraction of the ON state (i.e., the transmission time of primary networks) interrupted by the transmission of CR users..

Definition 2: The *maximum outage ratio* T_P is the maximum fraction of interference that primary networks can tolerate.

Definition 3: The *sensing efficiency* η is the ratio of the transmission time over the entire sensing cycle, defined as follows:

$$\eta = \frac{T}{T + t_s} \tag{4.69}$$

The objective of spectrum sensing is to achieve accurate detection probability as well as high sensing efficiency. Since both metrics are related to the sensing parameters T and t_s, the sensing parameter decision can be expressed as the optimization problem to maximize the spectrum efficiency satisfying interference constraint T_P as follows:

$$\begin{aligned} &\text{Find: } T, t_s \\ &\text{Maximize: } \eta = \frac{T}{T + t_s} \\ &\text{Subject to: } T_I \leq T_P \end{aligned} \tag{4.70}$$

SENSING TIME OPTIMIZATION

Sensing time optimization is investigated in [99] and [264]. In [264], the sensing time is determined to maximize 'the channel efficiency while maintaining the required detection probability, which does not consider the influence of a false alarm probability.

Assume the amount of sampling symbols in a detection cycle is N. Among N symbols, n symbols are used for channel detection. The performance of the detector above is determined by the detection probability P_{d}, and the false alarm probability P_{f}. Since the detection probability P_{d} is directly linked to the amount of interference that the PU encounters, it is generally set to the certain fixed level by the PU. Accordingly, the channel efficiency of the CR user, η, can be defined as the ratio of the amount of time that the idle channel can be detected and used by the CR user to the detection cycle, which is obtained by

$$\eta = \frac{N-n}{N}(1-P_{\mathrm{f}}) \tag{4.71}$$

In order to maximize the channel efficiency, P_{f} and n should be considered. Since P_{f} is known to a function of n, the optimization problem can be expressed as:

$$\hat{n} = \arg\max_{n=1,2,\ldots,N} \frac{N-n_s}{N}(1-f(n_s)) \tag{4.72}$$

In [99], the sensing time is optimized for a multiple spectrum environment so as to maximize the throughput of CR users. For an ongoing CR transmission, the probability that the user has to vacate the channel after sensing cycle T_s, is obtained by,

$$\begin{aligned} P_{\mathrm{e}} &= P_{\mathrm{f}} P_{00}(T_{\mathrm{s}}) + P_{\mathrm{d}} P_{01}(T_{\mathrm{s}}) \\ &\simeq P_{\mathrm{f}} P_{00}(T_{\mathrm{s}}) + P_{01}(T_{\mathrm{s}}) \end{aligned} \tag{4.73}$$

where $P_{00}(T_{\mathrm{s}})$ and $P_{01}(T_{\mathrm{s}})$ are the conditional probabilities that a channel will be idle and busy after T_s, respectively. given that it is currently idle.

Let r be the throughput of the CR user during the data transmission period T. Noting that, with probability P_e, each channel-monitoring period will be followed by a channel search period with average duration of $\bar{T}_{\mathrm{search}}$, the average throughput of the CR user, r, may be written as,

$$\begin{aligned} \bar{r} &= r\frac{T}{t_{\mathrm{s}} + P_{\mathrm{e}}\bar{T}_{\mathrm{search}} + T} \\ &= r\frac{T_{\mathrm{s}} - t_{\mathrm{s}}}{T_{\mathrm{s}} + P_{\mathrm{e}}\bar{T}_{\mathrm{search}}} \end{aligned} \tag{4.74}$$

Therefore, the optimum sensing time, $\hat{t}_{\mathrm{s}}$, maximizing the average throughput, may be found by solving the following optimization problem,

$$\hat{t}_{\mathrm{s}} = \arg\max_{t^{\mathrm{s}}>0} \frac{T_{\mathrm{s}} - t^{\mathrm{s}}}{T_{\mathrm{s}} + P_{\mathrm{e}}\bar{T}_{\mathrm{search}}} \tag{4.75}$$

TRANSMISSION TIME OPTIMIZATION

In [156], for a given sensing time, the transmission time is determined to maximize the throughput of the CR network while the packet collision probability for the primary network is under a certain threshold.

Suppose the sensing duration is τ and the frame duration is T. Denote C_0 as the throughput of the CR network when it operates in the absence of primary users, and C_1 as the throughput when it operates in the presence of primary users. For example, if there is only one point-to-point transmission in the CR network and the SNR for this CR link is $SNR_{\mathrm{s}} = P_{\mathrm{s}}/N_0$, where P_{s} is the received power of the CR user and N_0 is the noise power.

Let P_{p} be the interference power of primary user measured at the CR receiver, and assume that the primary users signal and CR users signal are Gaussian, white and independent of each other. Then $C_0 = \log_2(1+SNR_{\mathrm{s}})$ and $C_1 = \log_2(1+\frac{P_{\mathrm{s}}}{P_{\mathrm{p}}+N_0}) = \log_2(1+\frac{SNR_{\mathrm{s}}}{1+SNR_{\mathrm{p}}})$, where $SNR_{\mathrm{p}} = P_{\mathrm{p}}/N_0$. Obviously, we have $C_0 > C_1$. Note if the primary users signal is non-Gaussian, the above formula for C_1 can be treated as the lower bound of achievable rate for CR link when the primary user is active.

There are two scenarios for which the CR network can operate at the primary users frequency band.

- Scenario I: When the primary user is not present and no false alarm is generated by the CR user, the achievable throughput of the CR link is $\frac{T-\tau}{T}$.
- Scenario II: When the primary user is active but it is not detected by the CR user, the achievable throughput of the CR link is $\frac{T-\tau}{T}C_1$.

The probabilities for which Scenarios I and II happen are $(1 - P_{\mathrm{f}}(\epsilon.\tau))P_{\mathrm{off}}$ and $(1 - P_{\mathrm{d}}(\epsilon.\tau))P_{\mathrm{off}}$, respectively. If we define

$$R_0(\epsilon.\tau)) = \frac{T-\tau}{T}C_0(1 - P_{\mathrm{f}}(\epsilon.\tau))P_{\mathrm{off}} \tag{4.76}$$

and

$$R_1(\epsilon.\tau)) = \frac{T-\tau}{T}C_1(1 - P_{\mathrm{d}}(\epsilon.\tau))P_{\mathrm{on}} \tag{4.77}$$

then the average throughput for the CR network is given by

$$R(\tau) = R_0(\epsilon.\tau)) + R_1(\epsilon.\tau) \tag{4.78}$$

Obviously, for a given frame duration T, the longer the sensing time τ , the shorter the available data transmission time $(T - \tau)$. On the other hand, for a given target probability of detection, $\bar{P}_{\mathrm{d}}$, the longer the sensing time, the lower the probability of false alarm, which corresponds to the case that the CR network can use the channel with a higher chance.

The objective of sensing-throughput tradeoff is to identify the optimal sensing duration for each frame such that the achievable throughput of the CR network is maximized while the primary users are sufficiently protected. Mathematically, the optimization problem can be stated as

$$\begin{aligned} &\max_{\tau} R(\tau) = R_0(\epsilon.\tau)) + R_1(\epsilon.\tau) \\ &\text{subject to}: \ P_{\mathrm{d}}(\epsilon.\tau) \geq \bar{P}_{\mathrm{d}} \end{aligned} \tag{4.79}$$

where $\bar{P}_{\mathrm{d}}$ is the target probability of detection with which the primary users are defined as being sufficiently protected.

In practice, the target probability of detection $\bar{P}_\mathrm{d}$ is chosen to be close to but less than 1, especially for low SNR regime. For instance, in IEEE 802.22 WRAN, we choose $\bar{P}_\mathrm{d} = 0.9$ for the SNR of -20dB. It is pointed out that if the primary users require 100% protection in its frequency band, it will then be not allowed for the secondary usage in that frequency band. Also, we suppose the activity probability P_on of primary users is small, say less than 0.3, thus it is economically advisable to explore the secondary usage for that frequency band. Since $C_0 > C_1$, the first term in the right hand side of (18) dominates the achievable throughput. Therefore the optimization problem can be approximated by

$$\begin{aligned} &\max_{\tau} \bar{R}(\tau) = R_0(\epsilon.\tau)) \\ &\text{subject to}: \ P_\mathrm{d}(\epsilon.\tau) \geq \bar{P}_\mathrm{d} \end{aligned} \tag{4.80}$$

For a given sensing time, we may choose a detection threshold such that $P_\mathrm{d}(\epsilon.\tau) = \bar{P}_\mathrm{d}$. We may also choose a detection threshold $\epsilon_1 < \epsilon_0$ such that $P_\mathrm{d}(\epsilon.\tau) > \bar{P}_\mathrm{d}$. Obviously, $P_\mathrm{f}(\epsilon_1.\tau) > P_\mathrm{f}(\epsilon_0.\tau)$. Thus from Eqs. (4.76) and (4.77), we have $R_0(\epsilon_1.\tau) < R_0(\epsilon_0.\tau)$ and $R_1(\epsilon_1.\tau) < R_1(\epsilon_0.\tau)$. Therefore, the optimal solution to Eq. (4.80) is achieved with equality constraint in Eq. (4.80). Finally, $R_0(\epsilon_1.\tau) + R_1(\epsilon_1.\tau) < R_0(\epsilon_0.\tau) < R_1(\epsilon_0.\tau)$, thus the optimal solution to Eq. (4.79) is also achieved when the equality constraint in Eq. (4.79) is satisfied. However, this method does not consider a false alarm probability for estimating collision probability and throughput.

JOINT SENSING AND TRANSMISSION TIME OPTIMIZATION

All efforts stated above mainly focus on determining either optimal sensing time or optimal transmission time. On the other hand, in [150], a theoretical framework is developed to optimize both sensing and transmission times simultaneously in such a way as to maximize the transmission efficiency subject to interference avoidance constraints where both parameters are determined adaptively depending on the time-varying cooperative gain.

Analytical Interference Model: In order to optimize sensing parameters satisfying the interference constraint, we need to specify the relation between the interference ratio T_I and sensing parameters. In periodic sensing, interference can be expected to occur in the following cases [150]:

- *Interference on Busy State Sensing,* I_on*:* When the spectrum band is busy but CR users do not detect the primary user signal, CR users begin to transmit and interference can occur during the transmission period (Figure 4.26 (a)).
- *Interference on Idle State Sensing,* I_off*:* Even though the spectrum band is idle and CR users detect it correctly, there still exists the possibility that the primary user (PU) activity appears during the transmission period (Figure 4.26 (b)).

This scheme assumes the MAP based energy detection where the decision threshold λ is determined so as to equalize false alarm and miss-detection probability, i.e., $P_\mathrm{on} - P_\mathrm{d} = P_\mathrm{off}$. P_f and P_on and P_off represent busy and idle probabilities, respectively.

The interference I_on has two different interference patterns according to the transmission time T. If T is relatively short, interference is highly likely to persist over the entire transmission period with probability $P_\mathrm{on}(T)$ where $P_\mathrm{on}(T)$ represents the probability that the busy state will not change to the idle state during T. However, if T is long enough, busy and idle states occur alternately during T, and hence interference converges to $P_\mathrm{on} \cdot T$

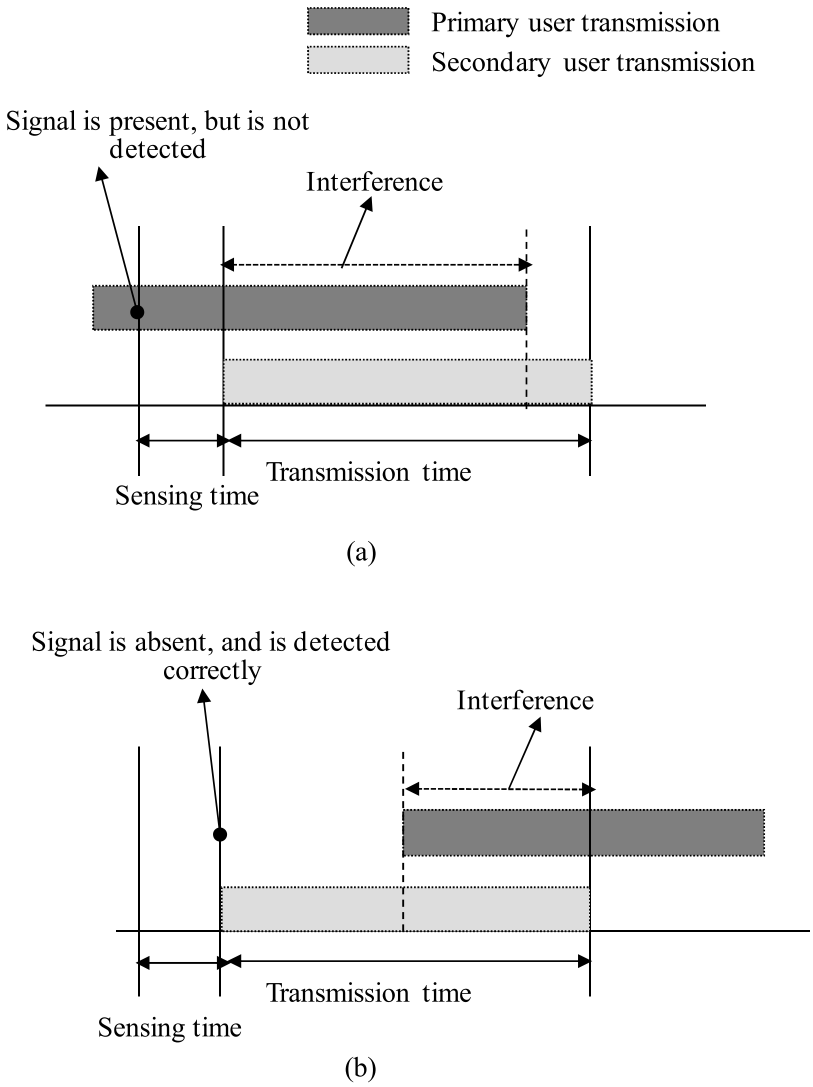


Figure 4.26 Interference model in periodic sensing: (a) interference in busy state, and (b) interference in idle state

with probability $1 - P_{\text{on}}(T)$. Thus, the expected interference during the transmission time T, $E[I_{\text{on}}]$, can be expressed as follows:

$$\begin{aligned} E[I_{on}] &= (P_{\text{on}} - P_{\text{d}})(P_{\text{on}}(T) \cdot T + (1 - P_{\text{on}}(T)) \cdot P_{\text{on}} \cdot T) \\ &= P_{\text{f}}(P_{\text{on}}(T) \cdot T + (1 - P_{\text{on}}(T)) \cdot P_{\text{on}} \cdot T) \end{aligned} \quad (4.81)$$

Similarly, I_{off} occurs only when one or more PU activities occur during the transmission time, which converges approximately to $P_{\text{on}} \cdot T$ with probability $1 - P_{\text{off}}(T)$ as follows:

$$E[I_{\text{off}}] = (P_{\text{off}} - P_f)(P_{\text{off}}(T) \cdot 0 + (1 - P_{\text{off}}(T)) \cdot P_{\text{on}} \cdot T) \quad (4.82)$$

where $P_{\text{off}}(T)$ represents the probability that the idle state will not change during T.

By combining $E[I_{\text{on}}]$ and $E[I_{\text{off}}]$, we can obtain the expected interference ratio T_I as follows:

$$T_{\text{I}} = \frac{E[I_{\text{on}}] + E[I_{\text{off}}]}{T \cdot P_{\text{on}}} \tag{4.83}$$

Sensing Parameter Optimization: Based on the MAP-based energy detection and the interference model, sensing parameters are optimized as follows:

- *Observation Time:* The observation time t_s can be represented as follows:

$$t_{\text{s}} = \frac{1}{W \cdot \gamma^2}[Q^{-1}(\frac{P_{\text{f}}}{P_{\text{off}}}) + (\gamma+1)Q^{-1}(\frac{P_{\text{f}}}{P_{\text{on}}})]^2 \tag{4.84}$$

  where $\gamma = {\sigma_{\text{s}}}^2/{\sigma_{\text{n}}}^2$ represents the signal-to-noise ratio (SNR).

- *Operating Region for Transmission Time:* From the Eq. (4.70) and (4.83), we obtain the inequality $T_{\text{I}}(T, P_{\text{f}}) < T_{\text{P}}$. Then boundary function of the operating region P_{f}^{op} can be expressed in terms of T, and T_{P} as follow:

$$P_{\text{f}} < P_{\text{f}}^{\text{op}}(T, T_{\text{P}}) \tag{4.85}$$

- *Optimization Procedure:* T and t_{s} have the same false alarm probability P_{f}. Thus, this optimization can be simplified as the problem to find an optimal false alarm probability P_f to maximize the sensing efficiency. According to the T, P_{f} is calculated using the boundary function $P_{\text{f}}^{\text{op}}(T, T_{\text{P}})$, and t_{s} is obtained accordingly. By searching all possible transmission time T, the optimal P_{f} can be obtained so as to have a maximum sensing efficiency.

4.4.2 Out-of-Band Sensing Control

When a CR user needs to find new available spectrum band (out-of-band sensing), a *spectrum discovery time* is another crucial factor to determine the performance of CR networks. Thus, this spectrum sensing should have a coordination scheme not only to discover as many spectrum opportunities as possible but also to minimize the delay in finding them. This is also an important issue in spectrum mobility to reduce the switching time, which will be explained in Chapter 7.

SENSING ORDER

First, the proper selection of spectrum sensing order can help to reduce the spectrum discovery time in out-of-band sensing.

Sensing Sequencing Optimization over Homogeneous Spectrum Bands: In [143], the optimal channel-sequencing algorithm is proposed to minimize delay in searching for an idle channel. Assume that SUs must sense $N-1$ foreign channels one by one until they can find an idle one. As a simple search-sequence, the channels may be arranged in an ascending order of channel utilizations u^i, which is not an optimal solution. Instead, spectrum sensing must consider P_{idle}^i the probability that channel i would be idle at a certain time t based on the previous samples. By setting t to the channel-switching triggering time, the optimal sensing sequence can be obtained as follows:

$$\begin{cases} P_{\text{idle}}^i(t) = Pr(Z^i(t) = 0 \mid \text{all previous samples}),\ \forall i \\ \text{Search channels in describing order of } P_{\text{idle}}^i(t) \end{cases} \tag{4.86}$$

Assume that CR users sense ON-OFF alternating channels. According to the renewal theory, CR users only need the most recent sample from each channel to derive $P^i_{\text{idle}}(t)$. Hence, $P^i_{\text{idle}}(t)$ becomes the transition probability between the most recent sample and its following sample at t. Then, $P^i_{\text{idle}}(t) = Pr(Z^i(t) = 0|Z^i(s^i) = d^i) = P^i_{d^i0}(t - s^i)$ is the most recent sensing time on channel i and $P^i_{d^i0}(t - s^i)$ is the transition probability between two samples d^i (at s^i) and 0 (at t). Since $d^i = 0$ or 1, P^i_{00} and P^i_{01} are considered.

The renewal theory suggests that $P^i_{11}(\Delta^i),\ \Delta^i = t - s^i$, is expressed as

$$P^i_{11}(\Delta^i) = \int_{|Delta^i}^{\infty} \frac{F_{T^i_{\text{ON}}}(u)}{E[T^i_{\text{ON}}]} du + \int_0^{|Delta^i} h^i_{10}(u) F_{T^i_{\text{ON}}}(\Delta^i - u) du \tag{4.87}$$

where $h^i_{10}(u)$ is the renewal density of the OFF state given that the renewal process started from the ON state. It is proven in [20] that $h^{i^*}_{10}(s)$ is expressed as

$$h^{i^*}_{10}(s) = \frac{f^*_{T^i_{\text{OFF}}}(s)[1 - f^*_{T^i_{\text{ON}}}(s)]}{E[T^i_{\text{ON}}] \cdot s[1 - f^*_{T^i_{\text{ON}}}(s) f^*_{T^i_{\text{OFF}}}(s)]} \tag{4.88}$$

For example, for a channel with Erlang-distributed ON/ OFF periods, as shown in (2), we have

$$\begin{aligned} P^i_{00}(\Delta^i) &= \frac{1}{2} + \frac{1}{2} e^{-\Delta^i} \cos(|Delta^i) \\ P^i_{10}(\Delta^i) &= \frac{1}{2} - \frac{1}{2} e^{-\Delta^i} \cos(|Delta^i) \end{aligned} \tag{4.89}$$

On the other hand, for a channel with exponentially distributed ON/OFF periods as shown in (3), we get

$$\begin{aligned} P^i_{00}(\Delta^i) &= (1 - u^i) + u^i \cdot e^{-(\lambda_{T^i_{\text{OFF}}} + \lambda_{T^i_{\text{ON}}})Delta^i)} \\ P^i_{10}(\Delta^i) &= (1 - u^i) - (1 - u^i) \cdot e^{-(\lambda_{T^i_{\text{OFF}}} + \lambda_{T^i_{\text{ON}}})Delta^i)} \end{aligned} \tag{4.90}$$

Then, the complete optimal channel-sequencing algorithm is given below.

1. $\forall i$, except that the channel to switch from

$$\text{calculate} P^i_{\text{idle}} = \begin{cases} P^i_{00}(\Delta^i), \ \text{if} d^i = 0 \\ P^i_{10}(\Delta^i), \ \text{if} d^i = 1 \end{cases} \tag{4.91}$$

$$\text{where} \begin{cases} d^i: \ \text{most recent sample of channel } i \\ \Delta^i: \ \text{elapsed amount of time since the most} \\ \qquad \text{recent sensing until channel switching} \end{cases} \tag{4.92}$$

2. Optimal sensing order. Sense $N - 1$ channels in descending order of P^i_{idle}

In case one round of channel search for all $N - 1$ channels cannot find any idle channel, an instant replay of the optimal channel searching is unlikely to find an idle channel because two consecutive samples collected within a short time window on one channel have non-negligible correlation, as will be shown in (4) in Section 6. Therefore, we recommend

$N-1$ channels to be searched again after T_{retry} seconds, which is a design parameter of the algorithm. In such a case, a new idle channel will be found by the research of the channels or by regular periodic sensing. In either case, once an idle channel is found, the channel switching procedure completes and CR users resume their communication on the new channel.

Sensing Sequence Optimization over Heterogeneous Spectrum Bands: The above method addresses a sensing-sequence that sorts channels in descending order of the idle probability. However, such a sequence only maximizes the chance of finding an idle channel, instead of minimizing the overall discovery-delay. Furthermore, the heterogeneity of licensed channels is not considered.

To address these problems, an optimal sensing-sequence is proposed to minimize the latency in finding the target amount of opportunities in backup channel list (BCL) [142]. This scheme considers heterogeneous channel characteristics, including signal detection time T_{I}^{i} , channel capacity C_i, and the probability P_{idle}^{i} of a channel to be idle. The optimal sequence is derived for channels with homogeneous capacities, i.e., $C_i = C_i, \ \forall i$. For a more general case (i.e., channels with heterogeneous capacities), a necessary condition for optimality is derived. It is also shown that finding the optimal sequence is NP-hard, and hence, a suboptimal sensing-sequence algorithm of polynomial time complexity is proposed.

Here, this method propose an efficient sensing-sequence of backup channels that incurs a small delay in discovering as much opportunities as a CR network needs. In building such a sequence, the heterogeneous characteristics of backup channels are considered by using a tuple of $\{T_i, C_i, P_{\text{idle}}^{i}\}$. T_{I}^{i} may differ between channels because it depends on the type of PU signals. C_i can be a physical bandwidth or Shannon capacity which varies with the time-varying channel condition (e.g., fading) and interference temperature [54]. $P_{\text{idle}}^{i}\}$ depends on the channel's ON/OFF usage pattern and hence varies with channels. In [143], P_{idle}^{i} is derived based on alternating renewal channels.

Suppose there are $N(< M)$ backup channels with their $\{T_i, C_i, P_{\text{idle}}^{i}\}$ known, and B_{req} is the amount of opportunities required for a CR network to support spectrum demands from its CR users. Then, upon triggering an opportunity discovery, the CR network needs to discover as much opportunities as $B = B_{\text{req}} - B_{\text{in-band}}$ where $B_{\text{in-band}}$ is the sum of in-band channels' capacities at the time of opportunity discovery. Note that $B_{\text{in-band}} = 0$ in the time-driven channel reuse model.

Let $S = \{s_1, s_2, ..., s_{\text{N}}\} \in \mathbf{S}$ be an ordered list of N channels, where s_j is the channel index of jth channel in the sequence (i.e., s_j : positive integer, $1 \le s_j \le N$) and S is the set of all possible channel sequences ($|\mathbf{S}| = N!$). Suppose T_{I}^{i}, C_i and $P_{\text{idle}}^{i} = \Pr(\Theta_i = 0)$ are known a priori, where $\Theta_i \in \{0, 1\}$ is the binary state of channel i, $i \in \{1, 2, ..., N\}$ ('0' means the channel is idle). The objective is to determine the optimal sensing-sequence S^* that minimizes the average delay in finding idle channels whose cumulative capacity exceeds B. This can be stated formally as

$$\begin{aligned} &\text{Find } S^* = \arg\min_{S\in\mathbf{S}} E_\tau[\sum_{i=s_1}^{s_\tau} T_{\mathrm{I}}^i] \\ &\text{Subject to} \sum_{i=s_1}^{s_\tau-1} C_i \cdot I_{\Theta_i} < B, \text{ and } \sum_{i=s_1}^{s_\tau} C_i \cdot I_{\Theta_i} \geq B \end{aligned} \tag{4.93}$$

where I_{Θ_i} is an indicator function such that

$$I_{\Theta_i} = \begin{cases} 1, & \text{if } \Theta_i = 0, \\ 0, & \text{otherwise} \end{cases} \tag{4.94}$$

Note that τ is a random variable, and hence, the expected delay (i.e., average sensing-time) is considered in the objective function.

Channel-Search Optimization: During the channel-search, the CR transmission is ceased as the primary channels have to be sensed one-by-one until an idle channel is found. Thus, in order to minimize the delay incurred by the CR user, the average duration of this period, $\bar{T}_{\text{search}}$, has to be minimized. Let t^s denote the time spent for sensing each channel (i.e. the integration time of the energy detector) during the channel-search period. While choosing a smaller t^s allows for faster sensing of each channel, it results in a higher probability of false-alarm, thereby reducing the chance of successful identification of a white space. As will be illustrated shortly, this tradeoff between the quality and the speed of sensing may be optimally balanced to minimize $\bar{T}_{\text{search}}$. Return of the primary user on the channel being used for the CR transmission triggers a sequential search of the remaining $N-1$ channels. The probability that a channel is declared idle and is acquired for the CR transmission, P_a, is given by,

$$P_a = (1 - P_f^s)(1-u) + (1-P_d)u \simeq (1-P_f^s)(1-u) \tag{4.95}$$

$u = P_{\text{off}}$ where P_f^s denotes the probability of false-alarm during the channel-search. $(1-P_f^s)(1-u)$ corresponds to the successful identification of a white space (i.e. no false-alarm) while $(1-P_d)u$ represents the case where the channel is falsely deemed idle due to the non-detection of the primary signal. A tight regulatory constraint on detection probability ensures that such false declarations are negligible and we may approximate P_a as in Eq. (4.95). Moreover, the number of primary channels, N, is assumed to be sufficiently large such that during the channel-search, with high probability, at least one of the primary channels is idle, that is, $u^{N-1} \ll 1$.

However, with high probability, a CR user may still be unable to acquire an idle channel if P_f^s is too high. Thus, the following constraint is set to ensure a successful channel-search with high probability,

$$u^{N-1} \leq (1-P_a)^(N-1) \leq \epsilon \ll 1 \tag{4.96}$$

where ϵ is chosen according to the QoS requirements of the CR user. The average time to find an idle channel is then given by,

$$\begin{aligned}\bar{T}_{\text{search}} &= P_a T^s \sum_{k=1}^{N-1} k(k - P_a)^{k-1} \\ &= T^s \left[\frac{1 - (1 - P_a)^N}{P_a} - N(1 - P_a)^{N-1} \right] \\ &\simeq \frac{T^s}{P_a} = \frac{T^s}{(1 - P_f^s)(1 - u)}\end{aligned} \tag{4.97}$$

where the approximation follows from Eq. (4.96). Combining Eqs. (4.96) and (4.97), we may formulate the optimum sensing time as follows,

$$\begin{aligned}\hat{T}^s &= \arg\min_{T^s > 0} \frac{T^s}{(1 - P_f^s)(1 - u)} \\ \text{subject to: } & P_f^s \leq 1 - \sqrt[n-1]{\epsilon} \frac{1-}{1 - u}\end{aligned} \tag{4.98}$$

STOPPING RULE

Moreover, if the CR user senses more spectrum bands, it is highly probable to detect a better spectrum band while resulting in longer spectrum searching time. To exploit this tradeoff efficiently, a well-defined stopping rule of spectrum searching is essential in out-of-band sensing. In [131], an optimal stopping time is determined to maximize the expected capacity of CR users subject to the maximum number of spectrum bands a CR user can use simultaneously.

There are multiple channels under consideration, and each channel is occupied by random primary traffic, which exposes itself as a spectrum opportunity with certain probability. According to the Shannon theory, for a single CR user, the theoretical throughput upper bound is proportional to the bandwidth used: $R = W \log(1 + SNR)$, where R is the data rate, W is the transmission bandwidth and SNR is the received signal strength and noise rate. Therefore if a CR user can exploit more channels and fully utilize them, significant throughput increase can be achieved. For the protection of primary users and for the exploitation of the spectrum opportunities, CR users must sense channels before they can actually use them. Further negotiation between a sender and its receiver is also needed for exchanging their channel availability information. These operations consume the effective transmission time of CR users. Therefore, there is a tradeoff between exploring more idle channels and encountering more sensing overhead, which is of great importance in the design of a multiple channel cognitive MAC protocol.

The spectrum sensing decision problem can be formulated as an optimal stopping problem. Here we briefly introduce the theory of stopping rule and optimal stopping [49]. Stopping rule is defined by two objects:

1. a sequence of random variables,X_1, X_2, ... , whose joint distribution is assumed to be known,
2. a sequence of real-valued reward functions, $y_0, y_1(x_1), y_2(x_1, x_2), \ldots, y(x_1, x_2, \ldots)$.

Given these two objects, the associated stopping rule problem is described as follows. The sequence of X_1,X_2, ... can be observed for as long as possible. For each $n = 1, 2, \ldots$, after

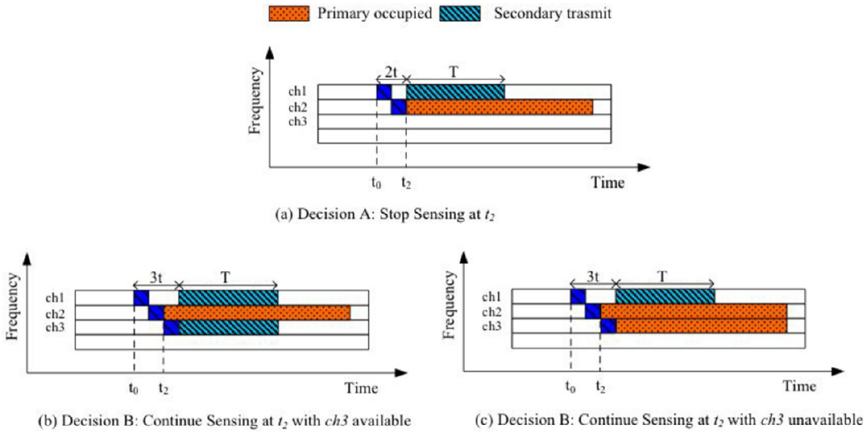


Figure 4.27 Stopping rule (need to be re-drawn.

observing $X_1 = x_1$, $X_2 = x2_,$ ..., $X_n = x_n$, the decision is either to stop and receive the known reward $y_n(x_1, ..., x_n)$, or to continue and observe X_{n+1} for further decision. If the decision is not to take any observations, the received reward is a constant amount, y_0. If never stopping, the received reward is $y(x_1, x_2, ...)$. The goal is to choose a time to stop such that the expected reward is maximized.

A stopping rule problem has a finite horizon if there is a known upper bound on the number of stages at which one may stop. If stopping is required after observing X_1,X_2, ..., X_{N}, the problem has a horizon N. A finite horizon problem is a special case of the general stopping rule problem with $y_{\text{N}+1} = ... = y_\infty = -\infty$. Finite horizon stopping rule problems can be solved by the method of backward induction [49]. Since spectrum sensing must stop at stage N, it first finds the optimal rule at stage $N-1$. Knowing the optimal rule at stage $N-1$ CR users find the optimal stopping rule at stage $N-2$ and so on, until back to the initial stage. In particular, Let

$$V_{\text{N}}^{(\text{N})} = y_{\text{N}}(x_1, x_2, ..., x_{\text{N}}) \tag{4.99}$$

and then inductively for

$$V_n^{(\text{N})} = \max[y_n(x_1, x_2, ..., x_n), E[V_{n+1}^{(\text{N})}(x_1, x_2, ..., x_n, X_{n+1}) \mid X_1 = x_1, ..., X_n = x_n]] \tag{4.100}$$

Let X_n denote the 0-1 (occupied-idle) state of the nth channel probed and the probability $\Pr(X_n = 1) = p$ is assumed to be equal for every channel. Let y_n denote the payoff of stopping probing and transmission after probing n channels. y_n is a function of the aggregated channel availability and depends on the radio technology. Here constraints for the cognitive radio are generalized: the maximum number of adjacent channels a single CR user can simultaneously use is W, the maximum number of spectrum fragments it can aggregate is F. For a band of spectrum with adjacent channels $\{i, i+1, ..., j\}$, we denote the number of fragments as $\text{Frag}(i, j)$. Let b_n be the maximum number of usable channels within n adjacent channels (starting from 1), subject to the above constraints (W, F), namely

$$b_n(x_1, ..., x_n) = \max_{1 \le i \le jn} \sum_{k=1}^{j} x_k \quad \text{Frag}(i, j) \le F \tag{4.101}$$

The function y_n can be written as

$$\begin{aligned} y_n(x_1, ..., x_n) &= \frac{T}{T + nT} b_n(x_1, ..., x_n) \\ &= \frac{c}{c + n} b_n(x_1, ..., x_n) \end{aligned} \tag{4.102}$$

where $c = T/t$. assuming that each available channel presents a unit of data rate, then y_n is actually the total effective data rate during the time interval $T + nt$ after making the stopping and transmission decision.

Assume the maximum number of channels a user can probe before make a stopping decision is at most K, which means this is a finite horizon problem, solvable by using the backward induction principle. Denote

$$\begin{aligned} V_{\rm K}^{(\rm K)}(x_1, ..., x_{\rm K}) &= \frac{c}{c + K} b_{\rm K}(x_1, ..., x_{\rm N}) \\ &= \frac{c}{c + n} b_n(x_1, ..., x_n) \end{aligned} \tag{4.103}$$

then

$$\begin{aligned} &E[V_{\rm K}^{(\rm K)}(x_1, x_2, ..., x_{\rm K-1}, X_{\rm K}) \mid X_1 = x_1, ..., X_{\rm K-1} = x_{\rm K-1}] \\ &= \frac{c}{c + K}[p \cdot b_{\rm K}(x_1, ..., x_{\rm K-1}, 1) + q \cdot b_{\rm K}(x_1, ..., x_{\rm K-1}, 0)] \end{aligned} \tag{4.104}$$

where p, q are the probabilities of $X_k = 1$ and $X_k = 0$ respectively; and inductively for $n = K - 1$ backward to $n = 2$,

$$\begin{aligned} &V_{\rm K}^{(\rm K)}(x_1, ..., x_n) \\ &= \max[y_n(x_1, ..., x_n), E[V_{n+1}^{(\rm K)}(x_1, ..., X_{n+1}, X_{n+1}) \mid X_1 = x_1, ..., X_n = x_n]] \\ &E[V_n^{(\rm K)}(x_1, ..., x_{n-1}, X_n) \mid X_1 = x_1, ..., X_{n-1} = x_{n-1}] \\ &= p \cdot V_n^{(\rm K)}(x_1, ..., x_{n-1}, 1) + q \cdot V_n^{(\rm K)}(x_1, ..., x_{n-1}, 0) \end{aligned} \tag{4.105}$$

Obviously, CR users should have a sensing at the beginning, with an observed result x_1, and then compare y_1 with $E[V_2]$, make the decision, and so on. At each stage, $\{E[V_n]\}$ defines the optimal stopping rule.

4.5 SPECTRUM SENSING IN IEEE 802.22

IEEE 802.22 standard is known as cognitive radio standard because of the cognitive features it contains. The standard is still in the development stage. One of the most distinctive features of the IEEE 802.22 standard is its spectrum sensing requirement [120]. IEEE 802.22

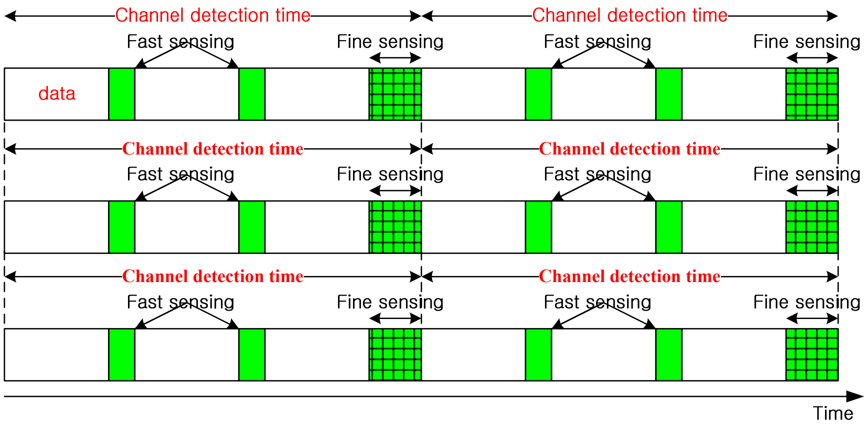


Figure 4.28 2-stage spectrum sensing in IEEE 802.11 [120].

based wireless regional area network (WRAN) devices sense TV channels and identify transmission opportunities. The functional requirements of the standard require at least 90% probability of detection and at most 10% probability of false alarm for TV signals with -116 dBm power level or above [240].

As shown in Figure 4.28, the spectrum sensing in IEEE 802.22 is envisioned to be based on two stages: fast and fine sensing [58]. In the fast sensing stage, a coarse sensing algorithm is employed, e.g. energy detector. The fine sensing stage is initiated based on the fast sensing results. Fine sensing involves a more detailed sensing where more powerful methods are used. Several techniques that have been proposed and included in the draft standard include energy detection, waveform-based sensing (PN511 or PN63 sequence detection and/or segment sync detection), cyclostationary feature detection, and matched filtering. A base station (BS) can distribute the sensing load among subscriber stations (SSs). The results are returned to the BS which uses these results for managing the transmissions. Hence, it is a practical example of centralized collaborative sensing.

Another approach for managing the spectrum in IEEE 802.22 devices is based on a centralized method for available spectrum discovery. The BSs would be equipped with a global positioning system (GPS) receiver which would allow its position to be reported. The location information would then be used to obtain the information about available TV channels through a central server. For low-power devices operating in the TV bands, e.g. wireless microphone and wireless camera, external sensing is proposed as an alternative technique. These devices periodically transmit beacons with a higher power level. These beacons are monitored by IEEE 802.22 devices to detect the presence of such low-power devices which are otherwise difficult to detect due to the low-power transmission.

4.6 ALTERNATIVE TECHNIQUES TO SPECTRUM SENSING

The FCC's recent proposal discussed three possible techniques unlicensed devices might use to determine whether white space spectrum is available for use at a given location:

- Passive sensing ("listen-before-talk") to detect the presence of a TV signal
- Geolocation using GPS or some other technology, followed by a check of a database to determine what frequencies are in use nearby
- Use of separate beacon transmitters that would indicate what spectrum is unavailable in a local area

So far, we've studied spectrum sensing techniques to find spectrum opportunities. However, this is considered as a rather time- and power-consuming operation, and hence a variety of controversial issues have emerged regrading its feasibility. In this section, we investigate other two techniques alternative to spectrum sensing.

The primary network can also provide the current status of its spectrum a central database as suggested in [107]. Whenever the CR network wants to access the spectrum, it looks up the database, and determines the spectrum availability. Furthermore, apart from mobilizing a rather time- and power-consuming operation such as spectrum sensing, the Cognitive Pilot Channel (CPC) concept has been proposed in European project E2R as a solution for providing the terminal with the necessary radio awareness at a given time and place, in a possible flexible spectrum management context [62]. In this concept, regional spectrum availability and relevant information are broadcasted to the CR network through the CPC.

4.6.1 Geo-Location Database Techniques

As recent FCC ruling removes the spectrum sensing requirement in TV white space, CR devices are enabled to access PU activity and spectrum information from a remote spectrum database [77]. This ruling raises the new challenges in using the on-demand service and web-based processing techniques such as cloud computing to provide CR users the fast, secure, scalable, and energy-efficient access to remote geo-location database.

Sensing-only devices do not generally utilize spectrum as efficiently as geo-location enabled devices, due to the large margins in incumbent detection thresholds that must be built into sensing-only devices. Geo-location enabled devices have knowledge of the specific interference protection requirements of each licensed incumbent, which allows varying levels of protection to be applied, maximizing utilization of the spectrum.

DATABASE INFORMATION

In addition to licensed transmitter information stored in incumbent databases, interference protection requirements such as protected service contour levels and required interference protection ratios may also be stored in the incumbent database, at whatever granularity (e.g., per-station type, per-region, per-individual station) is desired as follow:

- *Protected service contour levels:* Each type of licensed incumbent system described above has specific interference protection requirements. Each TV station has a commonly regulated protected service area. Generally, it is assumed that operation of TV white space devices (WSDs) is not allowed co-channel within these pre-defined service areas. To this end, the protected service contour levels need to be defined in terms of a minimum TV signal E-field strength. These levels along with the specific transmitter parameters effectively define a given station's protected service area.

- *Required interference protection ratio:* The required interference protection ratios need to be met for all white space device (WSD) operation scenarios, including for both sensing-only WSDs and geo-location enabled WSDs. These requirements help determine what WSD transmit power levels are permissible to avoid causing interference to incumbent systems. Generally, sensing-only WSDs must not only determine if an incumbent is present by detecting incumbent signals over a specified threshold such as the -116 dBm level for DTV signal detection, as proposed by IEEE 802.22, but they must also adjust their maximum transmit power levels downward based on sensed adjacent channel incumbent signal levels.

All WSDs must avoid co-channel operation within a stationfls protected service contour. However, WSDs may operate outside of protected service contours, as long as they can meet all of the required interference protection ratios at the nearest edge of each stationfls service contour. In this case, the WSD signal strength must be computed at the nearest contour edge of each station based on the protected service contour levels. Based on these information, each WSD determines the maximum allowable WSD transmit power.

GEO-LOCATION DATABASE IMPLEMENTATION

One method to significantly reduce the implementation complexity of stored geo-location databases is to utilize location uncertainty concepts. Starting with the computed high-resolution grid points $p_{x,y}$ of allowed power levels per location, the lower resolution °quantized– grid points $P_{X,Y}$ are assigned the minimum allowed power within the new resolvable area quanta (A) centered on the low-resolution grid points (X,Y). Basically, for a given bounded geographic region A (or desired spatial resolution), the worst case higher resolution WSD operating point is chosen per channel over that region. In this case, the lower bound of all higher spatial resolution computed maximum allowable transmit power levels px,y are taken into account for each lower resolution stored database point $(P_{X,Y})$ [107]:

$$P_{X,Y} = \min_{x,y \in A(X,Y)} (p_{x,y}) \tag{4.106}$$

This conservative method ensures that a WSD can transmit at a given (lower bound) power level without violating interference constraints as long as it operates in the bounded geographic region (A).

Using these techniques, a WSD could store database results at a resolution well below the required (i.e., FCC mandated) spatial resolution, thus reducing geo-location database implementation complexity while still meeting the required database operating requirements. The general trade-off is the larger the WSD location uncertainty region, the lower the allowed WSD transmit power level. The method can be used to significantly (e.g., >10x) reduce the size of stored geolocation databases. In addition, variable spatial resolutions (A) may be stored in the database, with lower resolutions being reserved for areas where the WSD is less likely to operate in, and higher resolutions utilized for common WSD operating regions [75]. Alternatively, highly mobile WSDs may utilize location uncertainty techniques to reduce real-time geolocation database queries (by only requiring database access at larger spatial intervals than specified when in motion). The described methods can be utilized to store very large geolocation databases, or reduce communications bandwidth to a geo-location database server (e.g., for mobile WSDs).

SPECTRAL UTILIZATION EFFICIENCY

All WSDs can be judged by how efficiently they utilize the white spaces. Geo-location database enabled WSDs will utilize the TVWS spectrum much more efficiently than their sensing-only counterpart WSDs. This is primarily due to the ability of geo-location enabled

WSDs to accurately determine protected service contours. Sensing-only WSDs must sense incumbent signals down to very low levels (e.g., -116 dBm for DTV transmissions) in order to combat hidden-node effects and other localized sensing phenomena (e.g., fading, shadowing, building penetration losses, etc.). In addition, portable sensing-only units will suffer from low antenna heights, and possibly low antenna gains (due to form factor restrictions and polarization mismatch), all of which make sensing incumbent signals more challenging, and require lower incumbent detection thresholds.

Assuming that the required UHF band DTV incumbent detection level is -116 dBm (or about 16dBu) due to the above issues, using F(50,90) DTV signal propagation modeling at 9m antenna heights, the average DTV transmitter protected service area over-estimation will be roughly 3.5x for an omnidirectional DTV transmitter antenna. Another way to look at this is that an area equal to 2.5 times the DTV stationfls service area is nominally unused by sensing-only WSDs due to the extremely low detection levels needed to combat the above signal reducing effects. Since this unused area is technically outside of the protected service area for the station, a vast majority of it could have been utilized. Considering that an average full-power DTV station (e.g., 400kW ERP, 400m HAAT) can have a coverage area of about 30,000km2, this means a significant under-use of white space spectrum will occur in sensing-only devices [107].

This analysis also assumes that the sensing-only WSD does intelligent detection of the DTV signal to distinguish it from other background noise and interference sources (which will almost certainly exceed -116 dBm levels, even in narrow observation bandwidths). Note again that the IEEE 802.22 standard has proposed the -116 dBm DTV detection level for outdoor, horizontally polarized antennas at 9 m nominal height.

Furthermore, certain incumbent signals, such as low-power and full-power TV transmissions have drastically different protection requirements, though the same modulation is utilized over the air. For example, a low power low-VHF band digital TV transmitter has a 15 dB lower protected service contour level than its full-power DTV counterpart. Sensing-only WSDs have no method to distinguish they both utilize identical ATSC modulation. Therefore, a WSD will be forced to protect low-power DTV stations as if they were their full-power counterparts. Using fourth law propagation, this 15 dB contour difference would result in an omni-directional protected service area reduction of 82%, meaning that a sensing-only WSD effectively let 4/5 of the equivalent full-power DTV stationfls coverage area lie fallow when the station is truly a low-power transmitter. In general, for a given indistinguishable difference in protected service contour levels ($|Deltac$) to a sensing-only WSD, the normalized percentage of wasted (unused) WSD operational area (A_w) is nominally

$$A_w = 100 \cdot (1 - (10^{-\Delta c/40})^2) \tag{4.107}$$

for an incumbent with a circular or omni-directional service area (assuming fourth-law propagation).

4.6.2 Cognitive Pilot Channel Technique

Within E^2R Project, reconfigurability features have been developed to cope with the context of heterogeneous cooperative radio access technologies, for the benefit of users, operators, service providers and regulators. This approach combining the diversity of complementary and collaborative systems led to the on-going definition of a Cognitive Pilot Channel

(CPC) [62]. An out-band CPC (could it be via a common world wide frequency band or other way) could help the connection of the user terminal to the most appropriate network, that would be available anywhere anytime.

TECHNICAL OVERVIEW

A worldwide based out-band Cognitive Pilot Channel (CPC) can be carried on an agreed frequency (frequencies). In the past administrative reasons within ITU led to the Physical Pilot Channel (i.e. identification of a harmonized frequency band world wide) in a standby status, when looking for a specific synchronization/roaming enabling channel for the IMT-2000 family. E^2R investigations framework and the broader set of different types of collaborative technologies considered in the project provides new optimistic challenging perspectives in the B3G context, even if it is clear, that any attempt to set up an international agreed CPC would necessarily require regulatory effort.

In the context of heterogeneous radio network environment, it is necessary for reconfigurable radio terminals to be able to initiate a new user session to set the conditions allowing a connection to the most suitable access point of the most appropriate Radio Access Technology (RAT). In particular after "power on" the mobile does not know which RAT may be the most appropriate or in which frequency bands potential RAT(s) are operating. This last point will be even more critical in the long term when new regulatory approaches to spectrum usage will allow the implementation of Dynamic Spectrum Allocation (DSA) and Flexible Spectrum Management (FSM) (which includes Spectrum Pooling). In this case, the mobile terminal will have to initiate a communication in a spectrum context which is completely unknown due to dynamic reallocation mechanisms. Without any information about the location of RATs within the considered frequency range reachable from the mobile terminal (frequency domains related to the collaborative radio systems), a scanning process of the whole frequency range will be necessary to discover the local and time dependant spectrum constellation.

In that context, an out-band Cognitive Pilot Channel (CPC) should provide relevant information (at least available RATs and their associated operating frequency bands) to a mobile terminal so as it can initiate a communication session in an optimal way, regarding time, situation and location. This would allow a number of meaningful advantages from several stakeholder viewpoints and in particular:

- The provided information would simplify the selection procedure, avoiding a large band scanning,
- The user terminal would benefit from lower battery consumption,
- It would be an appropriate solution in a DSA/FSM situation, hence the advantages for operators and spectrum management regulators, in a dynamically changing radio environment.

For that purpose, E^2R proposes to define an out band CPC carried on an out band harmonized frequency, common to locally existing PLMNs, whose role would be to help mobiles, at switch on, to choose the PLMN/RAT to camp on. It seems quite realistic to envisage that the information forwarded through this common downlink channel should remain of non strategic nature for operators, as those data are publicly broadcasted. The technical solutions proposed hereafter consider this need for confidentiality guaranty.

The minimal set of information that can be put in this CPC is, for each PLMN (operator), the frequency of the "preferred" RAT within his PLMN. Having the list of all existing operators and preferred RATs, the mobile selects the most suitable one to camp on (according to its profile and information stored in its SIM). This reduces the amount of information sent by the CPC, and consequently also, the risk that the bandwidth needed to convey the CPC information is not sufficient. Indeed, in addition to the difficulty identification of frequency allocations, the CPC bandwidth should be fixed and not subject to reallocations DSA mechanisms. Moreover, the PLMN/RAT choice procedure is simplified, and the information update can be dynamically processed in line with each operator strategy. In addition, operators do not need to broadcast confidential information.

In the following some preliminary studies on possible solutions for CPC are reported. Further investigations and/or other solutions are foreseen in E^2R II. The selection procedure using the CPC would consist of the following steps:

- At “switch on”, the mobile listens first to the out band CPC,
- Getting the list of all existing operators and preferred RATs, the mobile selects the most suitable one to camp on.

The information update can be dynamically processed in line with each operator strategy. The characteristics of the out-band Cognitive Pilot Channel must be such that the terminal can always receive it, even without any knowledge about the radio environment. The out-band Cognitive Pilot Channel should then be contained in a fixed frequency band, known in advance. The out-band Cognitive Pilot Channel can be broadcast on a wide zone including a great number of meshes. The out-band Cognitive Pilot Channel contains the data for all the meshes of this area. For each mesh, the out-band Cognitive Pilot Channel contains the operators available at this mesh, there preferred technologies and the corresponding frequency bands. The information corresponding to a given mesh can be written as follows:

- Mesh: [Operator O 1: technology T 1, frequency F I, 1, 1; technology T 2, frequency fi, 1, 2 ...

  [Operator O 2: technology T 1, frequency F I, 2, 1; technology T 2, frequency fi, 2, 2...] ,...

The RAT selection procedure using the information of the out-band Cognitive Pilot Channel is summarized in Figure 4.29.

- A terminal located in a certain mesh switches on.
- It determines itself its localization (e.g.: GPS/GALILEO)
- It is supposed that the mobile knows the technology in which the out-band Cognitive Pilot Channel is implemented, as well as the corresponding frequency band.
- Thanks to the knowledge of its position, the terminal is able to extract from the out-band Cognitive Pilot Channel the information on technologies available in its mesh, the operators deploying these RATs and the corresponding frequency bands.
- The terminal thus establishes its connection with the relevant operator and the network.

If the mobile has a subscription with a particular operator, it can seek for this operator in the list of the operators existing in its mesh. It can thus use information corresponding to its operator to select the network.

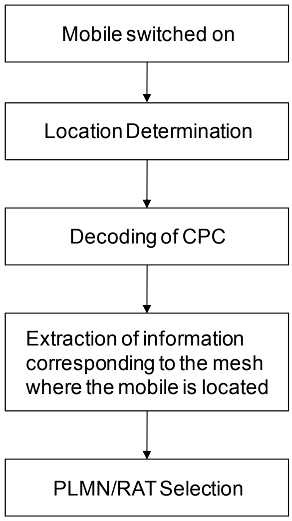


Figure 4.29 Block diagram for RAT selection using CPC [62].

4.7 SPECTRUM SENSING CHALLENGES

Spectrum sensing constitutes one of the most important components of the cognitive radio operation as highlighted in this chapter. The accuracy and the overhead of the spectrum sensing are two main issues in this area. The solutions discussed so far in this chapter provide valuable insight to the challenges and potential solutions in spectrum sensing. Nevertheless, there still exist several open research challenges that need to be investigated for the development of accurate and efficient the spectrum sensing solutions. We discuss these challenges in detail in this section.

4.7.1 Multi-user CR Networks

CR networks usually reside in a multi-user environment, which consists of multiple CR users and primary users. Furthermore, CR networks can also be co-located with other CR networks competing for the same spectrum band. However, current interference models [20], [73] do not consider the effect of multiple CR users. Multi-user environment makes it more difficult to sense the primary users and to estimate the actual interference. First, the effects of the transmission of other CR users are unknown to a specific CR user. Consequently, it is hard to estimate the total interference that would be caused at a primary receiver. Second, the transmissions of other CR users may prevent a specific CR user from detecting the activity of a primary transmitter and regard the primary user transmission as noise. This leads to degradation in sensing accuracy. Spectrum sensing functions should be developed considering the possibility of multi-user/network environment. In order to solve the multi-user problem, the cooperative detection schemes can be considered, which exploit the spatial diversity inherent in a multi-user network.

4.7.2 Physical Layer Constraints

Spectrum sensing techniques require efficient physical layer capabilities in terms of wide-band sensing and rapid spectrum switching. However, the constraints of the physical layer need to be known to design practical sensing algorithms. As discussed previously, the fact that the cognitive radio cannot sense and transmit simultaneously is one of the factors in the design of spectrum sensing algorithms. This fact has been considered in [150] to optimally schedule the transmission and sensing without degrading the sensing accuracy. As an alternative, the effect of using multiple radios has been investigated in [231], where a two transceiver operation is considered such that a transceiver always listens to the control channel for sensing. This operation improves the system performance, however, the complexity and device costs are high.

Another constraint is the limited spectrum sensing capabilities of cognitive radios. In other words, scanning the whole spectrum takes time. Since sensing consumes energy this process has to be carefully scheduled. One of the main requirements of CR networks is the detection of the primary users in a very short time [226]. Since sensing time is important, OFDM-based CR networks are known to be excellent fit for the physical architecture of CR networks [6] [247] [267]. Since multi-carrier sensing can be exploited in OFDM-based CR networks, the overall sensing time can be reduced. Once a primary user is detected in a single carrier, sensing in other carriers is not necessary. In [247], a power-based sensing algorithm in OFDM networks is proposed for detecting the presence of a primary user. It is shown that the overall detection time is reduced by collecting information from each carrier. However, this necessitates the use of a large number of carriers, which increases the design complexity. Hence, novel spectrum sensing algorithms need to be developed such that the number of samples needed to detect the primary user is minimized within a given detection error probability. In this sense, cooperative spectrum sensing mechanisms can be exploited to overcome the constraint of each cognitive radio. The superiority of cooperative techniques in terms of system performance has already been demonstrated in many studies [36] [115] [170] [293]. On the other hand, such a collaboration increases the communication overhead and may lead to overall system performance degradation when channel capacity or energy consumption is considered. Consequently, effective spectrum sharing techniques that enable efficient collaboration between different CR nodes in terms of spectrum sensing information sharing are required.

4.7.3 Adaptive Spectrum Sensing

As explained in the beginning of this chapter, the requirements of spectrum sensing solutions may depend on the network architecture. While centralized solutions focus on efficient information collection from multiple sensing devices and optimally allocating spectrum for users, distributed architectures lead to frequent information exchange between each CR user. Consequently, the nature of the spectrum sensing solution may differ depending on the architecture. However, considering that CR user devices will need to adapt to any network setting, whether being centralized or distributed, adaptive spectrum sensing solutions are crucial for rapid proliferation of the CR technology. As a result, a single CR device can be used in different network settings with a single, adaptive spectrum sensing solution.

Adaptive techniques are also necessary for different underlying physical layer functionalities. As explained above, physical layer constraints significantly affect the performance of spectrum sensing solutions. Moreover, it is clear that the realization of cognitive radio networks will lead to the implementation of different CR devices by different companies similar to the current case with WLANs. To provide a seamless spectrum sensing for higher networking layers, spectrum sensing solutions need to be adaptive to the physical layer capabilities.

4.7.4 Compressed Sensing

Compressed sensing is a promising wideband sensing technique in cooperative sensing. However, it also gives rise to many open research challenges:

- *Near Far Problem:* Due to the sub-Nyquist-rate sampling and insufficient number of samples, a weak PU signal with a nearby strong signal may not be properly reconstructed for detection in widenband spectrum. Thus, it is a challenge to achieve the detection sensitivity by compressed sensing in wideband spectrum.

- Implementation Issues: Compressed sensing is achieved by the random sampling of wideband signals. To realize random sampling, new ADC architecture with non-uniform timing and pseudo-random clock generator design such as the one in [13] [211] are needed. Since the complex clocking system will be the key factor of random sampling performance, how these implementation issues in compressed sensing affect cooperative sensing needs further investigation.

4.7.5 Security

From the primary user point of view, CR users can be regarded as *malicious* devices that *eavesdrop* on the channel that the primary user is transmitting. In a sense, spectrum sensing techniques resemble eavesdropping attacks. In order to preserve the privacy of the users spectrum sensing techniques need to carefully designed. This is particularly important considering the economics that lie behind the primary networks. Since each primary users own the particular spectrum, the traffic flowing through this spectrum needs to be protected. Spectrum sensing techniques, however, necessitate the knowledge of the existence of primary users for efficient operation. Consequently, spectrum sensing techniques should be designed in a way that they are aware of the *existence* of the ongoing traffic but cannot determine the *content* of the traffic. Moreover, these techniques need to be implemented so that any CR user that performs spectrum sensing will not be regarded as malicious by the already existing security protocols in primary networks.

4.7.6 Support of Asynchronous Sensing

If each user has independent and asynchronous sensing and transmission schedules, it can detect the transmissions of other CR users as well as primary users during its sensing period. However, with the energy detection, which is most commonly used for spectrum sensing, CR user cannot distinguish the transmission of CR and Primary users, and can detect only the presence of a transmission. As a result, the transmission of CR users detected during sensing operations causes false alarm in spectrum sensing, which leads to a decrease in spectrum opportunities. Thus, how to coordinate the sensing cooperation of each CR user to reduce these false alarms is an important issue in spectrum sensing.

4.7.7 Spectrum Sensing in Cellular Networks

Apart from TV bands, cellular frequencies represent one of the most viable ways of achieving DSA . both because they are widely used throughout the world and also because engineering devices and data applications for these frequencies are well understood. On the other hand, compared to TV bands, cellular spectrum usage is expected to be much more dynamic. Thus, data-driven studies are necessary to design DSA systems that optimize such usage. Additionally, looking from a completely different perspective, the low bit rates involved in wireless voice transmission make it an attractive application for secondary usage. Hence, understanding the nature of this application can also help drive secondary usage markets.

4.7.8 Cooperative Spectrum Sensing

The dominance of parallel fusion models in the literature result in the needs of proposing novel models in cooperative sensing for new applications. Thus, the open challenges regarding cooperation models include the following:

- *Modeling of Cooperation Overhead:* Most existing models for cooperative sensing are centered at the detection performance, that is cooperative gain. Only a few cooperation overhead issues have been discussed in proposed schemes. For example, in [262], only the number of cooperating CR users and the sensing time-throughput tradeoff are considered in forming utility functions. While cooperative gain is important in the model, proper modeling of cooperation overhead can reveal realistic achievable cooperative gain. Thus, the modeling of cooperation overhead is still an open challenge in the modeling for cooperative sensing.

- *Modeling of Primary User Cooperation:* Most existing models for cooperative sensing focus on the detection of single large-scale primary user such as TV base station and assume that the PUs do not cooperate with CR users. However, in certain applications such as military CR networks, these assumptions may not be true, since the PUs may be motivated to cooperate with CR users and the PUs may be connected in an ad hoc manner. As a result, new models that reflect the cooperation between PUs and CR users for cooperative sensing and cooperative communications such as the one in [289] are desired. In addition, the detection of small-scale mobile PUs is a known open challenging research problem, which will need a new model for cooperative sensing.

- *Reliability in Common Control Channel:* Apart from the unrealistic assumption of using perfect control channel in cooperative sensing, recent studies have focused on the cooperative sensing performance with the consideration of imperfect control channels. However, how to design a control channel resilient to channel impairments, robust to PU activity, and bandwidth-efficient for delivering sensing data is a nontrivial task.

.

4.7.9 User Selection

User selection is critical for cooperation performance. However, devising user selection scheme is nontrivial, especially when the geolocation information is unavailable.

Cooperation footprint [183] is the area where CR users cooperate with each other. Since cooperative gain is obtained from spatial diversity, cooperation footprint is an important parameter to evaluate the performance and the overhead in cooperative sensing. Thus, user selection schemes should consider the distribution of CR users and the the area covered by their cooperation, not just the distance between the CR users. However, deriving the exact footprint of cooperation from the user selection is a challenge.

4.7.10 Knowledge Base for Security

Most existing knowledge base methods are used to identify PU characteristics such as locations, power, and activity. To address security issues in cooperative sensing, the database should include other knowledge such as the behavior model of CR users and the model for jammer identification. Although it is a challenge to cooperatively establish accurate statistical models for security purpose, the knowledge derived from these models can significantly improve the security in cooperative sensing.

4.7.11 Sensing Delay

The challenges to improve cooperative sensing d sensing delay elay are as follows:

- *Multiple tradeoffs in Cooperative Sensing Delay:* The sensing-throughput tradeoff analysis in cooperative sensing should consider not only the sensing time and CR throughput, but also the report delay and the delay for synchronization or asynchronous reporting. Thus, the challenge is to balance the tradeoff between the CR throughput and cooperative sensing delay, which consists of multiple delay components depending on the cooperative sensing schemes.

- *Delay Analysis in Distributed Schemes:* Distributed cooperative sensing schemes usually require iterative process to reach the cooperative decision. The cooperative sensing delay is dominated by the report delays if the number of iterations for convergence is large. As a result, the delay analysis and the convergence of the distributed cooperative algorithm should be jointly considered.

4.7.12 Energy Conservation Techniques

As the advent of the green communications era, efficient energy conservation techniques in cooperative sensing are indispensable. The open research challenges are the following:

- *Energy Efficient User Selection:* Censoring techniques only reduces the energy consumption on reporting sensing data. However, the energy is still consumed by sensing even if the result is censored. Thus, it is a challenge to properly select the CR users for cooperation such that all the sensing results are informative and the energy spent on unnecessary sensing operations is saved.

- *Modeling of Energy Consumption:* Existing methods simply model the energy consumption in sensing, sleeping, and transmission/reporting as fixed values. However, many factors will affect the degree of energy efficiency in these operations. For example, different sensing techniques and sensing interval will consume different amount of energy. In addition, energy consumption in reporting may depend on the transmit power level adapted to channel conditions. Thus, more accurate energy model for cooperative sensing is needed.

4.7.13 Mobility

Despite some preliminary studies, there are still many unanswered questions regarding the impact of mobility on cooperative sensing. For example, what is the optimal way to perform cooperative sensing if CR users are moving? If CR users can be stationary or mobile, how to select the cooperating CR users? How to perform cooperative sensing in a stable and reliable manner at mobile or vehicular speed? We find that the research challenges on mobility issue in cooperative sensing are

- *PU Mobility and Tracking:* Due to the mobility of PUs, the tracking of PU movement becomes an important problem in cooperative sensing. The accurate tracking of PUs relies on an efficient localization method with location estimation. The development of an effective location estimation method based on the received signal strength values of PU signals remains a challenge.

- *Impact of Mobility Parameters:* It is a challenge to identify the mobility parameters that affect the detection performance, and their relations with cooperative gain and cooperation overhead. For example, mobility may increase or decrease the correlation among CR users and thus improving or degrading the detection performance in cooperative sensing. The possible parameters may include the mobility speed, the direction of movement, the doppler frequency, the density of CR users, or a profile that contains the moving trajectory and locations of CR users.

CHAPTER 5

SPECTRUM DECISION

After identifying spectrum opportunities through spectrum sensing, the CR network may have multiple available spectrum bands, which show different characteristics. Thus, CR networks require capabilities to decide on the best spectrum among the available bands, and corresponding communication configurations according to the QoS requirements of the applications. This notion is called *spectrum decision* and constitutes a rather important but yet unexplored topic.

Spectrum decision consists of the following main functionalities:

- *Spectrum Characterization:* Based on the RF observation, CR users identify not only the characteristics of each available spectrum but also its PU activity.

- *Decision Making:* The CR user finds the best spectrum band for each hop on the determined end-to-end route so as to satisfy end-to-end QoS requirements (*spectrum selection*), and accordingly reconfigure both hardware and software in the transceiver (*action decision*).

CR users require spectrum decision in the beginning of the transmission. Through RF observation, CR users characterize available spectrum bands by considering the received signal strength, interference, and the number of users currently residing in the spectrum, which are also used for resource allocation in classical wireless networks. However, in CR networks, each user observes heterogeneous spectrum availability that is varying over time and space according to the PU activities. This changing nature of the spectrum

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury

978-0470-68852-6

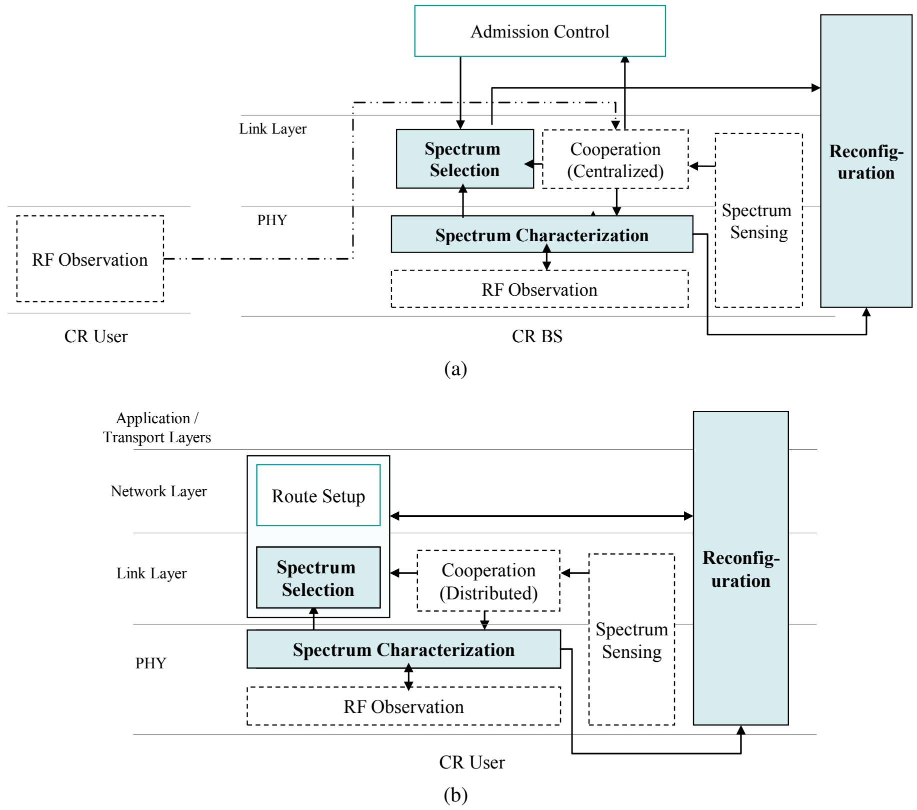


Figure 5.1 Functional block diagram for spectrum decision: (a)infrastructure-based CR networks, and (b) CR ad hoc networks.

usage needs to be considered in the spectrum characterization. Based on this characterization, CR users determine the best available spectrum band to satisfy its QoS requirements.

In infrastructure-based network, spectrum decision mainly focuses on selecting a proper spectrum band for a single hop to the base-station by considering current network utilization and the QoS requirements of a new incoming user. If the base-station cannot find the spectrum to satisfy the QoS requirements of the incoming user or adding the incoming user will expect significant quality degradation of current users, the base-station does not accept this incoming users through the *admission control*. Once the base-station admits the user, it allocates the best spectrum to the user as explained in Figure 5.1 (a).

Unlike infrastructure-based CR networks, CR ad hoc networks have unique characteristics in spectrum decision due to the nature of multi-hop communication. Spectrum decision needs to consider the end-to-end route consisting of multiple hops. Furthermore, available spectrum bands in CR networks differ from one hop to aother. As a result, the connectivity is spectrum-dependent, which makes it challenging to determine the best combination of the routing path and spectrum. Thus, spectrum decision in ad hoc networks should interact with routing protocols, which is shown in Figure 5.1 (b).

In the following sections we investigate functionalities for spectrum characterization, and decision making module in more detail:

5.1 SPECTRUM CHARACTERIZATION

In CR networks, the available spectrum holes show different characteristics, which vary over time. To determine a proper spectrum band corresponding to QoS requirements, it is important for CR users to identify the characteristics of each spectrum band. In order to describe the dynamic nature of CR networks, each spectrum hole should be characterized in terms of not only the time-varying RF environment and but also the primary user activity and the spectrum band information such as operating frequency and bandwidth.

In this section, we investigate RF channel parameters representing heterogenous radio environment and PU activities uniquely found in CR networks, and accordingly derive a novel CR channel model. Based on these information, we also introduce a new concept of database, radio environment map (REM), which stores all information regarding spectrum characteristics, and provides them to CR users.

5.1.1 Radio Environment

In CR networks, all available spectrum bands are spread over a wide frequency range, and hence exhibit different characteristics, which is varying over time. Thus, each spectrum band should be characterized by considering both the time-varying radio environment and the spectrum parameters such as operating frequency and bandwidth. Hence, it is essential to define parameters that can represent a particular spectrum band as follows:

- *Operating Frequency Range:* CR users are aware of the bandwidth and of the frequency range of the primary networks, which significantly influences the performance of CR networks. For example, while a higher frequency band has wider coherent bandwidth for high data, it shows a higher attenuation of radio signals, leading to a smaller coverage compared to a lower frequency band. Furthermore, a wider bandwidth requires a strict hardware requirement such as a high speed A/D converter with high resolution.

- *Interference:* Interference is a main parameter to determine the spectrum capacity. Unlike the conventional wireless network, CR networks should consider not only the interference at the CR user, but also the interference at the primary receiver to avoid the interruption of primary networks, i.e., the permissible power of a CR user can be obtained from the amount of the interference at the primary receiver such that the interference contraints .

- *Path Loss:* The path loss is closely related to the distance and operating frequency. As the operating frequency increases, the path loss increases, which results in a decrease in the transmission range. If transmission power is increased to compensate for the increased path loss, interference at other users may increase. Geolocation information and weather conditions also influence the propagation model.

- *Wireless Link Errors:* Depending on the modulation type, channel coding scheme, and the interference level of the spectrum band, the error rate of the channel changes.

- *Link Layer Delay:* To address different path loss, wireless link error, and interference, different types of link layer protocols are required at different spectrum bands. This results in different link layer delays.

- *Primary User Activity:* This is defined as the traffic statistics of the primary networks, which will be explained more in detail in Section 5.1.2.

Furthermore, the spectrum policy is also an important factor to characterize the spectrum band. Generally, each spectrum band has its own spectrum rule imposed by regulators, which specify, for example, how much limit on maximum transmit power that does not depend on activity in the band, how much interference CR devices are allowed to impose on the primary system, and what etiquette must devices follow in a commons based on coexistence. The following are some examples on the spectrum policy that should be considered spectrum characterization:

- *Minimum Signal-to-Noise Ratio (SNR):* To determine spectrum availability, CR users need statistical information on the received primary signals. The minimum SNR is the least signal level needed to decode the received signals, depending on the modulation type, channel coding and multiple access methods of primary user networks.

- *Interference Constraint:* Since CR users cannot monitor the spectrum continuously, CR networks do not guarantee perfect interference avoidance to the primary networks. Instead, CR networks exploit the interference constraint, which can be defined as either maximum interference level or maximum interference probability that primary networks can tolerate. Although the former is the most suitable for the objective of the opportunistic transmission, the latter is more practical since there is no practical way to measure the amount of the interference at the nearby primary receivers. The interference constraint is dependent on the radio access technology of the primary network.

- *Spectrum Etiquette:* Most of spectrum bands have their own spectrum etiquettes, which determine the coordination of spectrum access among multiple CR users, and control the selfish behavior of CR users. The spectrum etiquette can improve overall spectrum utilization especially in unlicensed bands, and hence should be considered in spectrum characterization.

It is desirable to identify the spectrum bands combining all the characterization parameters described above for accurate spectrum decision. However, a complete analysis and modeling of spectrum in CR networks is yet to be developed.

5.1.2 Primary User Activity

For an efficient spectrum utilization, the CR network needs to be aware of the traffic statistics of primary networks in each spectrum, called *PU activity*. In the following subsections, we introduce the several models regrading the PU activity.

TWO-STATE ON-OFF MODEL

The main interest of CR networks is to determine the presence of the primary users in the licensed band. Thus, the primary user activity in spectrum i can be modeled as a

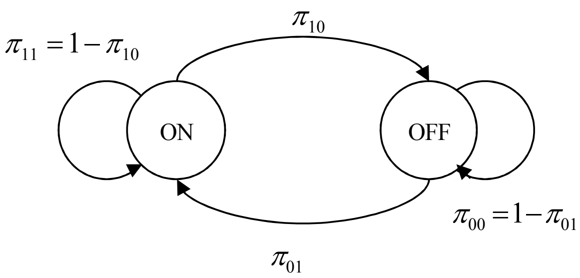


Figure 5.2 2 state On-OFF model for PU activity.

two-state birth-death Markov chain, which is shown in Figure 5.2.

In this model, An ON (busy) state represents the period used by primary users and an OFF (idle) state represents the unused period, and each transition is generally modeled as a Poisson process with death rate (from busy to idle) α_i and birth rate (from idle to busy) β_i. Thus, idle and busy duration can be modeled as an exponential distribution with average durations, $1/\beta_i$ and $1/\alpha_i$ respectively. Each transition probability can be obtained as follow:

$$\begin{aligned} \pi_{01} &= 1 - e^{-\beta_i t} \\ \pi_{10} &= 1 - e^{-\alpha_i t} \\ \pi_{00} &= 1 - \pi_{01} = e^{-\beta_i t} \\ \pi_{11} &= 1 - \pi_{10} = e^{-\alpha_i t} \end{aligned} \tag{5.1}$$

Then each state probability can be obtained as follows:

$$\begin{aligned} P_i^{\text{on}} &= \frac{T_i^{\text{on}}}{T_i^{\text{on}} + T_i^{\text{off}}} = \frac{\beta_i}{\alpha_i + \beta_i} \\ P_i^{\text{off}} &= \frac{T_i^{\text{off}}}{T_i^{\text{on}} + T_i^{\text{off}}} = \frac{\alpha_i}{\alpha_i + \beta_i} \end{aligned} \tag{5.2}$$

where T_i^{on} and T_i^{off} are average busy and idle durations.

The PU activity itself is modeled as a continuous Markov chain. However, since the PU activity can be observed by periodic sensing operations in CR users, it can be seen as a discrete-time Markov chain by CR users. The transition matrix for the discrete-time Markov chain can be express as follow:

$$\Pi = \begin{bmatrix} e^{-\beta_i \Delta t} & 1 - e^{-\beta_i \Delta t} \\ 1 - e^{-\alpha_i \Delta t} & e^{-\alpha_i \Delta t} \end{bmatrix}. \tag{5.3}$$

where Δt is the sensing period. Then n step transition matrix, which represents the transition probabilities after nth sensing operation, can be obtained as Π^n.

EMPIRICAL MODEL

There are some efforts to model the PU activity in specific spectrum bands based on field experiments. In [273], the characteristics of primary usage in cellular networks are presented based on the call records collected by network systems, instead of real measurement. This analysis shows that an exponential call arrival model is adequate to capture the PU activity while the duration of wireless voice calls does not follow an exponential distribution. Furthermore, it is shown that a simpler random walk can be used to describe the PU activity under high traffic load conditions. In [95], a statistical traffic model of wireless LANs based on a semi-Markov model is proposed to describe the temporal behavior of wireless LANs. Through empirical studies, it is shown that a hyper-Erlang distribution of the busy duration provides the best fitness to both stationary UDP traffic and non-stationary HTTP traffic in wireless LANs. However, the complexity of this distribution hinders its practical implementation in CR functions.

ONLINE PRIMARY USER ACTIVITY MODEL

The above approaches are fixed models based on offline measurements. Hence, they do not adequately capture the time varying nature of the PU activity. Furthermore, in recent studies, the PU activity is assumed to follow the Poisson model with exponentially distributed inter-arrivals. However, the Poisson model fails in capturing the bursty and spiky characteristics of the monitored data [199]. In Figure 5.3, it is shown that the actual PU activity fluctuates around the ON horizontal level, which is not exactly tracked by the Poisson approximation. Some of these fluctuations result in durations, where the PU is actually absent (shown by the dashed lines). These durations, mistakenly classified as an ON period by the Poisson model, serve as missed transmission opportunities for the CR users as the band is not utilized. Since the Poisson model does not consider correlations and similarities within data, it is incapable of identifying such fluctuations. This leads to fewer cases of correct spectrum hole detection, thus causing a degradation in CR network performance. Consequently, it is desired to detect these missed transmission opportunities while achieving less interference simultaneously.

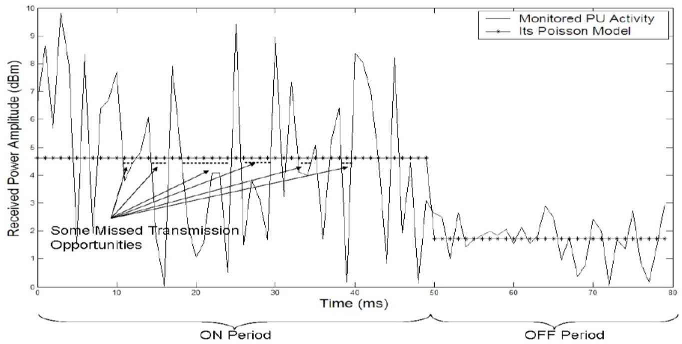


Figure 5.3 Missed transmission opportunities caused by the Poisson modeling [33].

In order to accurately track the changing PU activity a novel real-time based PU activity model for CR networks is developed in [33]. Here, the PU signal samples are first collected over a pre-determined duration. Then, the observed PU signals are clustered together, if they are greater than a threshold. Based on this clustering, the current PU arrival-departure rates can be estimated. The duration of collecting the signal samples, as well as the threshold for classifying the observed value as a legitimate PU signal are calculated in this work. However, this approach needs several PU signal samples collected at one centralized location. Thus, this needs to be extended for CRAHNs, so that each CR user may form individual clusters of the PU signals, based on their local observation, which can then be combined to give the complete PU activity model. Moreover, the additive white Gaussian noise (AWGN) channel model used in the proposed approach does not incorporate the effects of fading and shadowing, which can lower the accuracy of the PU activity prediction.

[Proposed Framework]

Our proposed model consists of two main modules: *PU activity monitoring module* and *clustering-modeling module* which are illustrated in Figure 5.4.

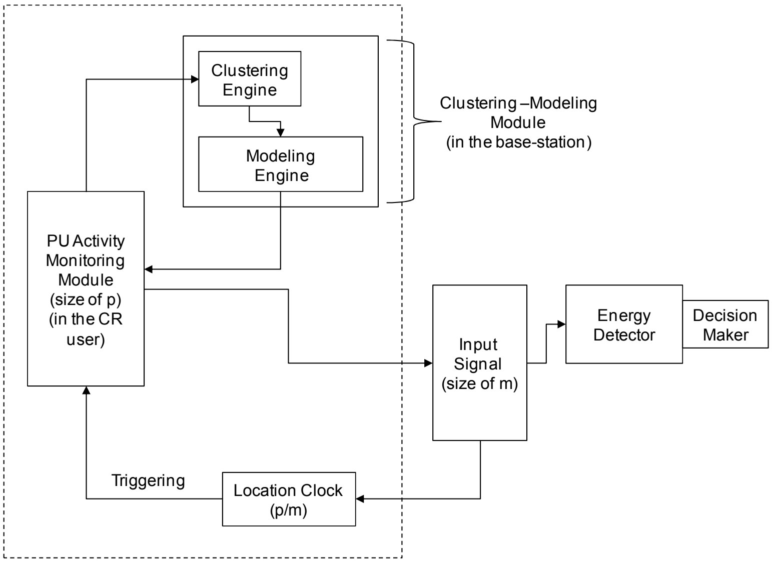


Figure 5.4 The block diagram of the online PU activity model [33].

The PU activity monitoring module, which is implemented in each CR user, monitors the spectrum band to take p consecutive samples of PU activity. Once the monitoring is finished, this module gives the monitored PU activity vector $\underline{q}$ for modeling and analysis to the *clustering-modeling module* which is implemented in the base station. The *clustering-modeling module* activates its *clustering engine* where the monitored PU activity samples are accumulated into *clusters* using a *first-difference filtering* procedure enhanced with temporal correlation calculations. As a result, a new clustered PU activity vector with clusters is generated and then input to the *modeling engine* as seen in Figure 5.4. In this engine, a correlation based modeling scheme produces the new modeled PU activity and parameterizes PU activity characteristics, i.e., P_{off}, the probability occupying the spectrum and P_{on}, the probability of PU presence. The newly generated PU activity characteristics and the modeled PU activity vector $\underline{r}$ are input back to the *PU activity monitoring module* in the CR user. Then, the modeled PU activity is input to the energy detector which takes the modeled PU activity vector size of $m << p$, and realizes its energy detection for spectrum sensing. As the energy detector operates in a loop with p/m iterations, the time is controlled by a local clock. Therefore, at the end of an iteration, the energy detector triggers the *PU activity monitoring module* using the local clock for a new analysis. More details on *clustering modeling module* are in the following subsections.

[Clustering-Modeling]

The monitored PU activity is input to the *clustering modeling module* in the base station. The module has two engines to process the monitored PU activity: the *clustering engine* and the *modeling engine* which are explained below.

- *Clustering Engine:* At the beginning of the clustering process, the *clustering engine* receives the monitored PU activity vector $\underline{q}$ from the *PU activity monitoring module*. Since the monitored PU activity $\underline{q}$ is input to the *clustering engine module*, Assume that the modeled PU activity vector $\underline{r}$ is identical to the monitored PU activity vector $\underline{q}$ at the beginning of the *clustering engine*, i.e. $\underline{q} = \underline{r}$. Then, all the consecutive samples (the current sample $r(m)$ and the last sample $r(m-1)$) are passed through the first-difference finite impulse response (FIR) filter. In the next step, the filter output $D(m)$ is checked with δ-test, which is a set of first-difference filtering procedures. If the δ-test is successful, the ρ-test is applied. Consequently, the modeled PU activity sample $r(m)$ is placed in the existing cluster $C(k)$ with its predecessor $(r(m-1))$ if both tests are successful, whereas any fail from these two tests leads the sample $r(m)$ to form a new cluster $C(k+1)$.

  As a result, only the modeled PU activity sample $r(m)$, which is $close$ to its predecessor $r(m-1)$ (successful in δ-test) and highly correlated with the last two samples $r(m-1), r(m-2)$(successful in ρ-test, which is a correlation calculation procedure), is placed in the same cluster with its predecessor $r(m-1)$. Furthermore, by using clusters, groups of first-difference filtered PU activity samples which have different correlation statistics are separated. In other words, spiky and bursty characteristics of the modeled PU activity are more accurately distinguished by cluster exploitation, leading the CR user to detect the PU activity fluctuations more precisely, hence causing less interference.

- *Modeling Engine:* This engine is used to model the clustered PU activity provided by the *clustering engine*. In the modeling engine, the pair $(C(k), C(k+1))$ is decided to be characterized by a decision region after passing by correlation slope and ρ-test. Consequently, the clusters of the pair, $C(k+1)$ and its predecessor $C(k)$, are decided individually to become either busy or idle using the specific decision of the region in which the pair is allocated. As the output of the modeling engine, the total number of idle clusters r_{off}, the total number of busy clusters r_{on}, the modeled PU activity vector $\underline{r}$, PU activity characteristics P_{on} and P_{off}, as well as the calculation of P_{f} and T are input to the PU Activity monitoring module as seen in Figure 5.4.

  Using the modeling engine, each cluster pair $(C(k), C(k+1))$ is analyzed independently, thus the fluctuations in PU activity are better classified. This leads to more accurate detection of the transmission opportunities and an increase in the CR network performance.

5.1.3 Cognitive Radio Channel Model

Channel capacity, which can be derived from the parameters explained above, is the most important factor for spectrum characterization. Recent work focuses on estimation of channel capacity in CR networks. Usually SNR at the receiver has been used for the capacity estimation. However, since SNR considers only local observations of CR users, it is not enough to avoid interference at the primary users. Besides channel capacity, other important statistics such as data loss rate and the outage probability need to be specified for characterizing each spectrum band, In this subsection, we introduce a CR capacity, and corresponding statistics by considering a unique feature of CR networks, PU activities.

CR CAPACITY

In the CR network, the available spectrum bands are not contiguous and may be spread over a wide frequency range with different bandwidth. Here we assume the CR network has multiple orthogonal non-interfering spectrum bands. For more flexible manipulation of heterogenous spectrum bands, we employ an orthogonal frequency division multiplexing (OFDM) as the physical layer technology.

Assume that each spectrum band i has a different bandwidth B_i Hz, consisting of multiple sub-carriers. Each sub-carrier can be assigned to different CR users. Moreover, each user can be allocated to the different number of sub-carriers in every time slot to control the data rate and error probability individually for each user. If a user k can be assigned to all sub-carriers in spectrum i with bandwidth B_i, the channel capacity in OFDM can be obtained as follows:

$$r_i(k) = \int_0^{B_i} \log_2(1 + \frac{|H_i^k(f)|^2}{N_i^k(f) + I_i^k(f)} P_i^k(f))df \tag{5.4}$$

where $H_i^k(f)$, $P_i^k(f)$, $N_i^k(f)$, and $I_i^k(f)$ denote the channel frequency response, the transmission power spectral density, the noise power spectral density, and interference corresponding to a user k at a spectrum band i, respectively.

Usually, each sub-carrier has a different channel gain and a noise level which are time-varying. However, in case of the long-term spectrum characteristics, both fast and frequency selective fading effects are mitigated, and hence we can say $H_k(f)/(N_k(f) + I_k(f))$ in the same spectrum band is identical over a long-term period. If $P_i^k(f)$ is also identical in frequency, a normalized channel capacity $c_i(k)$ (bits/sec/Hz) of spectrum band i can be expressed as $c_i(k) = r_i(k)/B_i$.

However, in CR networks, each spectrum i cannot provide its original capacity $c_i(k)$. First, CR users cannot have a reliable spectrum permanently and need to move from one spectrum to another according to the PU activity, which introduces the so-called *spectrum switching delay*. During the switching time, the transmission of the CR user is temporarily disconnected. Here, spectrum switching delay includes times for the spectrum decision process in the base-station, signaling for the new channel establishment, and RF front-end reconfiguration. In IEEE 802.22 Wireless Regional Area Network (WRAN), switching delay is required to be less than 2 sec [120]. Also conventional mobile broadcasting systems, for example, Qualcomm's MediaFLO, show an average physical layer channel switching delay up to 1.5 sec [40]. Depending on the development of the hardware technology, it

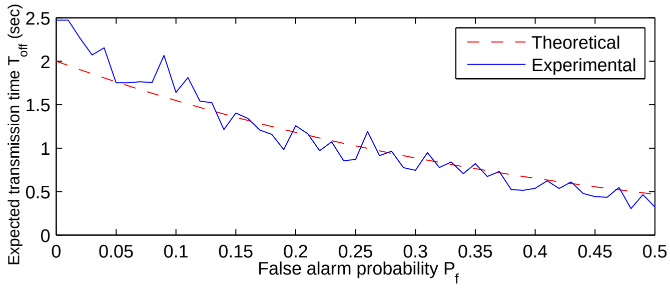


Figure 5.5 Expected transmission time in imperfect sensing.

will be much shorter but still be a significant factor to influence the network performance. Furthermore, CR users are not allowed to transmit during sensing operations, leading to the periodic transmissions with sensing efficiency η_i [150].

These unique features in CR networks shows a significant influence on the spectrum capacity $\mathbf{C_i(k)}$. To describe all these stochastic activities, we introduce a new capacity notion, the so-called *CR capacity* $C_i^{\mathrm{CR}}(k)$, which is defined as the expected normalized capacity of user k in spectrum i. Here we consider two scenarios for CR capacity - a single spectrum environment, which does not consider spectrum switching, multi spectrum environment, which allows to switch a new spectrum when detecting PU activity, as follows [152]:

- *Single-Spectrum:*

$$C_i^{\mathrm{CR}}(k) = E[\mathbf{C_i(k)}] = P_i^{\mathrm{off}} \cdot \eta_i \cdot c_i(k) \tag{5.5}$$

- *Multi-Spectrum:*

$$C_i^{\mathrm{CR}}(k) = E[\mathbf{C_i(k)}] = \frac{T_i^{\mathrm{off}}}{T_i^{\mathrm{off}} + \tau} \cdot \eta_i \cdot c_i(k) \tag{5.6}$$

  where τ represents the spectrum switching delay, and T_i^{off} is the expected transmission time without switching in spectrum i. Since CR users face to the spectrum switching after the idle period, the first term in Eq. (5.6) represents the transmission efficiency when CR users occupy spectrum i.

In the perfect spectrum sensing where both false alarm and detection error probabilities are zero, T_i^{off} is obtained as $1/\beta_i$, which is the average idle period based on the ON-OFF model. On the contrary, in case of imperfect sensing, CR capacity should account for the influence of sensing capability. Let Δt be a sensing period. Then, the average number of sensing slots in the idle period n_{s} is $\lceil 1/\beta_i/\Delta t \rceil$. From this, the expected transmission time can be obtained as follows:

$$\begin{aligned} T_i^{\text{off}} &= \Delta t \cdot \sum_{k=1}^{n_{\text{s}}-1} k \cdot (1-P_i^{\text{f}})^k \cdot P_i^{\text{f}} + \frac{1}{\beta_i} \cdot (1-P_i^{\text{f}})^{n_{\text{s}}} \\ &= \Delta t \cdot [\frac{(1-P_i^{\text{f}})(1-(1-P_i^{\text{f}})^{n_{\text{s}}-1})}{P_i^{\text{f}}} \\ &- (n_{\text{s}}-1) \cdot (1-P_i^{\text{f}})^{n_{\text{s}}}] + \frac{1}{\beta_i}(1-P_i^{\text{f}})^{n_{\text{s}}} \end{aligned} \tag{5.7}$$

where P_i^{f} represents a false alarm probability of spectrum i at each sensing slot. Here T_i^{off} can be expressed as the sum of the expected durations until when the false alarm is first detected in each slot. As P_i^{f} increases, T_i^{off} decreases, resulting in decrease in CR capacity, which is described in Figure 5.5. Here, we consider a cooperative sensing scheme based on 'OR' fusion, where its detection error probability converges to 0 as the number of users increases [156]. Thus, the detection error probability can be ignored in estimating CR capacity.

CR DATA LOSS RATE

Real-time applications are sensitive to delay and jitter. Moreover, they require a reliable channel to support a sustainable rate during the session time. Thus, real-time applications have strict constraints on the delay bound and the sustainable rate. within the delay bound. Even though the network can support sustainable rate R_s on average, packets can be delayed and finally discarded in the receiver due to the variation of channel capacity.

Unlike conventional wireless networks, the CR network has unique delay factors. When CR users either sense or switch the spectrum, they need to stop transmission temporarily, which prevents the real-time application from maintaining its sustainable rate, leading to delay and jitter. To observe the effect of the delay uniquely shown in CR networks, the buffering scheme is assumed to be optimized to absorb delay factors in conventional wireless networks, such as application layer, link layer, and transmission delays. Then, the additional delay factors uniquely introduced by CR networks can directly lead to data losses. For this reason, the data loss rate can be used to evaluate the service quality of real-time applications. Also real-time applications are assumed to have a set of discrete sustainable rates and to adjust their rates through the negotiation flexibly.

In the CR network, each spectrum band has two discrete capacity states, 0 and $c_i(k) \cdot w_i(k)$ according to its PU activity, as explained in Section 5.1. Here $c_i(k)$ and $w_i(k)$ are the normalized capacity and the bandwidth of spectrum i for user k, respectively. Thus, when N spectrum bands are assigned to a CR user k, the total capacity $\mathbf{R_T(k)}$ has 2^N states according to the PU activities of the selected spectrum bands. Thus, each state m has the following state probability:

$$P_m(k) = \prod_{i \in \mathcal{I}_m} \frac{T_i^{\text{off}}}{T_i^{\text{off}} + \tau} \cdot \eta_i \prod_{i \in \mathcal{B}_m} (1 - \frac{T_i^{\text{off}}}{T_i^{\text{off}} + \tau} \cdot \eta_i) \tag{5.8}$$

where $\mathcal{I}_m$ and $\mathcal{B}_m$ are the sets of idle spectrum bands and busy spectrum bands at state m, respectively.

Let the sustainable rate of user k be $R_{\rm s}(k)$ and the capacity of each state m be $\widehat{R}_m(k)$. From the assumption that the data loss occurs when channel capacity is below $R_{\rm s}(k)$, the data loss rate can be defined as the ratio of the expected capacity loss to the sustainable rate $R_{\rm s}(k)$ as follows:

$$\begin{aligned} P_{\rm loss}(k) &= \frac{R_{\rm s}(k) - \sum_{m=1}^{2^N} \min(R_{\rm s}(k), \widehat{R}_m(k)) P_m(k)}{R_{\rm s}(k)} \\ &= \frac{\sum_{m=1}^{2^N} |R_{\rm s}(k) - \widehat{R}_m(k)| P_m(k)}{2R_{\rm s}(k)} \end{aligned} \tag{5.9}$$

VARIANCE OF CR CAPACITY

From the capacity state probability, derived in Eq. (5.8), the variance of the total capacity $\mathbf{R_T(k)}$ can be derived as follows:

$$\mathrm{Var}[\mathbf{R_T(k)}] = \sum_{m=1}^{2^N} (\widehat{R}_m(k) - R_{\rm s}(k))^2 \cdot P_m(k) \tag{5.10}$$

By comparing Eq. (5.9) with Eq. (5.10), we can see that the variance of the total capacity $\mathrm{Var}[\mathbf{R_T(k)}]$ is proportional to the data loss rate $P_{\rm loss}(k)$. As a result, we can use the capacity variance for resource allocation, instead of the data loss rate. To apply the variance in Eq. (5.10) for the optimization, we need another form of the variance expressed in terms of the bandwidth $w_i(k)$ and the normalized capacity $c_i(k)$ of each spectrum. Since the spectrum is independent with each other, the variance of the total capacity in the selected spectrums can be expressed as follows:

$$\begin{aligned} \mathrm{Var}[\mathbf{R_T(k)}] &= \mathrm{Var}[\sum_{i \in \mathcal{S}} \mathbf{C_i(k)} \cdot w_i(k)] = \sum_{i \in \mathcal{S}} \mathrm{Var}[\mathbf{C_i(k)} \cdot w_i(k)] \\ &= \sum_{i \in \mathcal{S}} (E[(\mathbf{C_i(k)} \cdot w_i(k))^2] - E[\mathbf{C_i(k)} \cdot w_i(k)]^2) \\ &= \sum_{i \in \mathcal{S}} ((c_i(k)^2 \cdot w_i(k)^2 \cdot \frac{T_i^{\rm off}}{T_i^{\rm off} + \tau} \cdot \eta_i \\ &\quad - (c_i(k) \cdot w_i(k) \cdot \frac{T_i^{\rm off}}{T_i^{\rm off} + \tau} \cdot \eta_i)^2) \\ &= \sum_{i \in \mathcal{S}} \frac{T_i^{\rm off} \eta_i (T_i^{\rm off} + \tau - T_i^{\rm off} \eta_i)}{(T_i^{\rm off} + \tau)^2} c_i(k)^2 w_i(k)^2 \end{aligned} \tag{5.11}$$

where $\mathbf{C_i(k)}$ is the random variable to represent the capacity of spectrum i for user k. $\mathcal{S}$ is the set of the selected bands.

RESOURCE OUTAGE PROBABILITY

To model PU activities in the spectrum, a two-state Markov chain can be used with the transition probabilities from idle to idle $x_i^{00} = 1 - e^{-\beta_i \Delta t}$, from idle to busy $x_i^{01} = e^{-\beta_i \Delta t}$, from busy to idle $x_i^{10} = e^{-\alpha_i \Delta t}$, and from busy to busy $x_i^{11} = 1 - e^{-\alpha_i \Delta t}$, where Δt is a sensing period. Then, the idle probability of spectrum i after $r\Delta t$, $P_i^{\rm idle}(r)$, can be expressed as either one of the following probabilities [151]:

$$P_i^{\text{i2i}}(r) = \frac{x_i^{10}}{x_i^{01} + x_i^{10}} + (1 - x_i^{01} - x_i^{10})^r \cdot \frac{x^{01}}{x^{01} + x_i^{10}}$$
$$P_i^{\text{i2b}}(r) = \frac{x_i^{10}}{x_i^{01} + x_i^{10}} - (1 - x_i^{01} - x_i^{10})^r \cdot \frac{x^{10}}{x^{01} + x_i^{10}} \quad (5.12)$$

where $P_i^{\text{i2i}}(r)$ and $P_i^{\text{i2b}}(r)$ are the expected idle probabilities after $r\Delta t$ when current spectrum states are idle and busy, respectively. If a false alarm probability P_i^{f} is considered, the idle probability of spectrum i can be expressed as either $(1-P_i^{\text{f}})P_i^{\text{i2i}}(r)$ or $(1-P_i^{\text{f}})P_i^{\text{i2b}}(r)$.

Based on these probabilities, the expected resource outage probability is derived as follows: Since the network has M spectrum bands, it has 2^M states according to the status of each band. Let $\mathcal{L}$ be a set of states that experience resource outage, i.e., that $W^{\text{av}} < W_{\text{min}}$. $\mathcal{I}_n$ represents a set of idle spectrum bands at state n. Then, resource outage happens when all spectrum bands in $\mathcal{I}_n, n \in \mathcal{L}$ are idle and the rest of bands $i \notin \mathcal{I}_n, n \in \mathcal{L}$ are busy. From this, the resource outage probability after $r\Delta t$, $P_{\text{out}}(r)$ can be derived as follows:

$$P_{\text{out}}(r) = \sum_{n \in \mathcal{L}} \prod_{i \in \mathcal{I}_n} P_i^{\text{idle}}(r) \prod_{i \notin \mathcal{I}_n} (1 - P_i^{\text{idle}}(r)) \quad (5.13)$$

Based on this probability, the expected resource outage probability during $r\Delta t$, P_{out} can be obtained as $\sum_{r'=1}^{r} P_{\text{out}}(r^{'})/r$.

5.1.4 Radio Environment Maps

It is clear from the above discussion that CR networks are highly dependent on the state of the environment (radio emission state, topological state, etc.). Much of the work on cognitive radios has been based on making inferences about that environmental state solely at the time of decision making, especially when focussing on dynamic spectrum access (or anything else of highly dynamic nature). However, it has also become clear that any additional and temporally consistent knowledge of the environment can be used to significantly improve the accuracy and performance of the decision-making process. Such information includes typical behavior of other transmitters in the area as well as propagation conditions, just to name a few examples, usually assumed to be stored in a database-like system, locally or globally. The term, radio environment map (REM) is typically used to characterize such a database [296]. The key to the REM design is to decide what type of information must be stored and how this would be available to the various radios.

The REM, supported by distributed CR nodes and/or network infrastructure, is envisioned as the large.scale navigator for CRs. It provides cognitive services to the associated internal networks as well as a useful awareness of external networks such as legacy systems. REM covers multi-domain environmental information such as geographical features, available services, spectral regulations, location of various entities of interest (radios, reflectors, obstacles) plus radio.equipment capability profiles, relevant policies and past experiences. The REM information can be updated with observations from CR nodes and disseminated throughout CR networks [205].

There are numerous different types of information that the REMs could store. The simplest examples considered in the standards, such as IEEE 802.22, are transmitter locations and explicit protection zones around them. Also localization information of CR users themselves are important both for assessing their relations with the protection zones, and for more general policy issues. Furthermore, radio propagation information is the main factor to determine resource management and channel access decisions Thus, various models are commonly used to describe in a qualitative and quantitative manner the effects that the environment has on radio communications.

Some of the most common parameters that the radio propagation models use are:

- Street width and building height (used in, for example, the Walfisch-Ikegami propagation model)
- Weather conditions
- Location and height of obstacles (e.g. building floor number is used to calculate the building penetration loss)
- Vegetation and tree heights (used in, for example Weissberger's model is used to estimate the path loss due to the presence of trees)
- Antenna heights of the base stations (used by most propagation models such as the Okumura-Hata model)
- Area type such as urban, sub-urban, highway, open rural, indoor, office, airport, etc.

Another type of information the REMs could store is the presence or absence of wireless services; this is something that differs greatly from an area to another. For example, radio broadcasts are usually found in areas where the population density is high. Thus, such services are usually present in urban or suburban areas and not densely present in open rural areas. CRs can exploit the absence of such services in an opportunistic spectrum.reuse manner, as mentioned in the above Section on scenarios. Spectral regulations are also an example of heavily geographically.dependent information that changes slowly and thus could easily be stored in a REM.

Topology information of primary and secondary networks would also be useful in resource management and DSA.related decisions. In the REM case, "topology" is typically used to describe the location and the connectivity of the nodes. Instead of raw location information, statistical description of locations can be used as well. For example, interference or total received power is typically dominated by contributions from few of the closest transmitters. Thus, distributions of the distances to the nearest neighbors can already be used as a basis for a number of algorithms. Such distributions estimated from location data are examples of spatial statistics, methods to describe the structure of locations without enumerating all the individual coordinates.

In general, according to [218], three types of topological information can be stored in a REM:

- Raw location data, either known precisely or estimated using localization techniques.
- Statistics of measurements, such as the pair correlation function of the node locations.
- Models of the various phenomena arising in a network such as the formation of connections.

Statistics and models have the advantage that they describe the raw data with only few parameters (information-compression effect). In principle, this reduces the burden of storing large volumes and eases the calculation requirements for decision.making purposes. Finally, such statistics and models ease the burdening of the limited wireless resources as less information need be transferred between network nodes.

The activities of PUs can provide further insights into the radio environment. In [270] it is shown that the spectrum use is clustered in the frequency domain, which might prove valuable for making approximate predictions for opportunistic spectrum use. Although the individual user activities are obviously important in any decision.making process, there is no straightforward way of storing such information. Brute-force storage of raw measurement data is impractical as the volume increases very quickly. In [219], this problem is partially addressed by proving that spatial statistics and random fields can be used instead to model spectral maps, thus reducing the volume of needed and stored data.

5.2 DECISION MAKING ENGINE

5.2.1 Overview

Once the available spectrum bands are characterized, the most appropriate spectrum band should be selected considering the QoS requirements and the spectrum characteristics (*spectrum selection*). Accordingly, the transmission mode and the bandwidth for the transmission can be reconfigured (*action decision*).

Since there is no guarantee that a spectrum band will be available during the entire communication of a CR user, the spectrum selection functionality considers the primary user activity, i.e., how often the primary user appears on the spectrum band. Besides spectrum selection, spectrum decision involves reconfiguration in CR networks. The protocols for different layers of the network stack must adapt to the channel parameters of the operating frequency. Once the spectrum is decided, CR users need to select the proper communication modules such as physical layer technology and upper layer protocols adaptively dependent on application requirements as well as spectrum characteristics, and then reconfigure their communication system accordingly. In the following subsections, we investigate two main functionalities in the decision making engine - spectrum selection, and action decision in more detail.

5.2.2 Spectrum Selection

Once the available spectrum bands are characterized, the most appropriate spectrum band should be selected. Based on user QoS requirements and the spectrum characteristics, the data rate, acceptable error rate, delay bound, the transmission mode, and the bandwidth of the transmission can be determined.

Because of the operation of primary networks, generally, CR users cannot obtain a reliable communication channel for long durations. Moreover, CR users may not detect any single spectrum band to meet the user's requirements. Therefore, CR users can adopt the multi-radio transmissions where each transceiver (radio interface) tunes to different non-contiguous spectrum bands for different users and transmits data simultaneously. This method can create a signal that is not only capable of high data throughput, but is also immune to the interference and the PU activity. Even if a PU appears in one of the current spectrum bands, the rest of the connections continue their transmissions unaffected [30] [152]. In addition, transmission in multiple spectrum bands allows lower power to be used in each spectrum band. As a result, less interference with PUs is achieved, compared to the transmission on single spectrum band. As a result, less interference with PUs is achieved, compared to the transmission on single spectrum band. For these reasons, spectrum decision should support multiple spectrum selection capabilities. For example, how to determine the number of spectrum bands and how to select the set of appropriate bands are still open research issues in CR networks.

Furthermore, as stated previously, since the entire communication session consists of multiple hops with heterogeneous spectrum availability, the spectrum selection rule is closely coupled with routing protocols in CRAHNs. Since there exist numerous combinations of route and spectrum between the source and destination, it is infeasible to consider all possible links for spectrum decision. In order to determine the best route and spectrum more efficiently, spectrum decision necessitates the dynamic decision framework to adapt to the QoS requirements of the user and channel conditions. Furthermore, in recent research, the route selection is performed independent of the spectrum decision. Although this method is quite simple, it cannot provide an optimal route because spectrum availability on each hop is not considered during route establishment. Thus, joint spectrum and routing decision method is essential for CRAHNs, as described later in Chapter 9.

Uniqueness in Spectrum Selection

The objective of the spectrum selection in spectrum decision is similar to that of resource allocation spectrum sharing, which is presented in Chapter 6, in the sense that spectrum decision provides QoS guarantees as well as resource allocation. All of the previous research explained above has mainly addressed spectrum sharing issues where all operations are performed within the same spectrum band or across contiguous channels. Furthermore, to adapt the fast time-varying channels, they are generally designed as a short-term operation, such as a packet-based or a time-slot based scheduling.

However, CR networks necessitate an additional resource allocation capability when primary users are detected or CR users newly begin their sessions, which are relatively long-term events. Thus, this capability should consider longer-term channel characteristics, compared to spectrum sharing. In addition, since available spectrums are distributed over a wide frequency range, this function needs to be implemented as an inter-spectrum operation. However this operation inevitably introduces an additional switching delay leading to service quality degradation. Thus, it is not desirable to extend existing spectrum sharing solutions designed to adapt to the fast time-varying channel to the long-term inter-spectrum operation. This unique challenge in CR networks has not been addressed in previous research. Here the design objective of the spectrum decision framework is to decouple all inter-spectrum functionalities totally from spectrum sharing.

Spectrum Decision Framework

Here, spectrum decision is considered as an event-based functionality, i.e., the CR network decides on the proper spectrum bands in the following events:

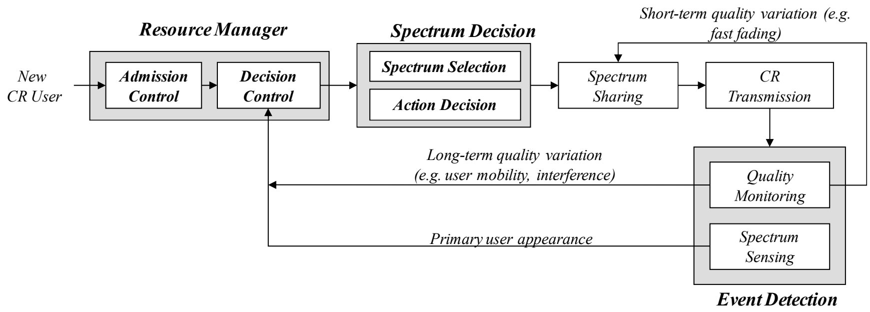


Figure 5.6 Spectrum decision framework.

- CR user appearance: When a new CR user appears in the CR network, it needs to be assigned to new spectrum bands for its transmission.
- Primary user appearance: When a primary user appears in the spectrum band, CR users should move to the new spectrum bands.
- Channel quality degradation: When channel condition becomes worse, CR users want to switch to a better spectrum band.

To consider all decision events effectively, the CR network necessitates a unified framework for spectrum decision. Figure 5.6 shows the proposed framework for spectrum decision.

A detailed description of this framework is as follows: By considering current spectrum conditions, a resource manager determines if the CR network accepts a new incoming CR user or not. If a new CR user is allowed to transmit, it is assigned to the proper spectrum bands through spectrum decision. Since there may be the multiple CR users competing the same spectrum, spectrum sharing coordinates those multiple accesses to prevent the collisions, and accordingly to achieve the maximum capacity. In the event detection, the current spectrum bands and users connections are monitored to detect decision events. The event detection consists of two main tasks: spectrum sensing and quality monitoring. When events are detected, the CR network reconfigures its resource allocation to maintain the service quality. In case of short-term channel variations such as fast fading, the CR network reallocates resources within the spectrum band through spectrum sharing. If a primary user is detected or the current spectrum band cannot provide the predetermined service quality any longer over a long-term period, the CR network switches the spectrum through the resource manager and the spectrum decision. In the proposed framework, CR users perform only event detection. Based on information gathered from CR users, the base-station decides on spectrum availability and performs spectrum decision as explained above.

Because of the PU activities, available spectrum bands show time-varying characteristics in the CR network. Thus, with the only proposed decision schemes, the CR network is not able to exploit spectrum resources efficiently, and hence results in the violation of the guaranteed service quality. As a result, the CR network necessitates an additional resource management scheme to coordinate the proposed spectrum decision methods adaptively with bandwidth fluctuations. The main objectives of the proposed resource management are as follows:

- The CR network is capable of determining the acceptance of a new incoming CR user without any effect on the service quality of currently transmitting users.
- During the transmission, the CR network needs to maintain the service quality of currently transmitting users by considering the fluctuation of the available bandwidth.
- Since real-time users usually have a higher priority in spectrum access, best effort users may not have enough resources. Thus, the CR network may be required to balance the bandwidth between both applications.

Consequently, the spectrum decision framework provides a hierarchical QoS guaranteeing scheme: spectrum sharing to allocate the channel and transmission power for short-term service qualities and spectrum decision to determine the best spectrum for maintaining service quality over a long term period. In this paper, we mainly focus on the decision functionalities: spectrum decision and resource management. Spectrum sharing will be explained in Chapter 6.

Spectrum Selection Principles: Examples

Applications can be divided into several classes based on the type of services supported. Each of these classes will have different QoS requirements, meaning different constraints and objective functions in the spectrum decision framework. The following are the spectrum selection principles for two example service classes- real-time and best effort applications:

- *Real-time applications:* Real-time applications need to have more reliable and time-invariant communication channels to satisfy strict service requirements, such as delay constraints and sustainable rates. However, how to maximize the total network capacity is still a crucial problem. To address these issues together, it is essential to guarantee the service quality of real-time applications with minimum spectrum resources. Thus, the spectrum decision problem can be formulated as an optimization to minimize bandwidth utilization subject to the constraint of the sustainable rate, data loss rate, and number of transceivers.
- *Best-effort application:* The objective of typical scheduling methods for best effort applications is to maximize the network capacity. The spectrum decision for best-effort applications has the same objective, but additionally needs to exploit the PU activity and long-term channel characteristics.

5.2.3 Action Decision

Action decision is responsible for find the optimal combination of communication protocols and parameters according to the local observation and the selected spectrum band. Here the action includes upper layer protocols as well as communication parameters in physical and MAC layers. This functionality is related to reconfigurability in cognitive radio. This also is an important issue but has not been widely investigated so far.

Action decision mainly focuses on optimization to suit the applications' and users' needs. Optimization occurs in many dimensions and is subject to a variety of constraints. We first introduce the optimization process for action decision, and then develop a basic approach to its implementation [220].

- *Optimization for QoS:* Optimization must take place in three major areas: the user/application domain, radio environment, and network interactions.

  1. *User/Application:* While most of current work in cognitive radio emphasizes spectrum management and agility, the user remains the fundamental component of the cognitive radio. The cognitive radio should adapt as best it is able to reflect the applications' and users' preferences for quality of service (QoS). Understanding this optimization domain and providing an effective solution remains one of the most difficult problems in the practical implementation of cognitive radio.

     The action decision module should understand and quantify the user's needs. Furthermore, the module has to translate these needs into specific radio actions. The radio could continuously poll the user for this information, or the user could define a desired level of QoS. Either of these two approaches involves user intervention and assumes that the user is expected to understand what he or she needs from the radio. The challenge is to develop passive methods to monitor the activities of the user and from these intelligently develop and model user behavior and QoS needs.

  2. *Radio Environment:* Another important area of optimization is related to the link quality of the point-to-point communications between nodes. The propagation channel has a direct impact on the QoS that a link delivers. Channel conditions, such as the fading type, fading level, Doppler spread, and path length will greatly impact bit errors which ultimately translate to QoS. How to choose the proper waveform like modulation, channel coding, interleaving, and spreading is an important issue.
  3. *Network Level Interaction:* When extending the optimization problem to the network level, it is important to both quantify the cognitive radio interactions and measure the impact of each radio on others. When networks are interacting and competing for spectrum resources, decisions on waveform adaptation must properly reflect and respect the needs and operations of other radios and networks in the same RF environment. Frequency agility is one way of working in this optimization field, but many other techniques also exist. Orthogonal spreading and modulations allow spectrum reuse in coding, antenna directivity allows spatial reuse, and perhaps cooperative timing schemes will allow spectrum sharing in the time domain.

  Each of the optimization domains interact strongly with each other; changing to a more robust channel coding method may improve robustness in a bad channel, but the cost in latency may negatively affect the user's QoS. Later sections of this paper will show how these interactions occur and present methods for understanding, representing, and optimizing overall performance with respect to them.

- *Constraints:* While the optimization on the PHY and MAC layers tries to build a waveform for the point-to-point link that maximizes the QoS, a cognitive radio must always respect any local regulatory limitations. Spectrum and power are two areas of major concern here. Devices operating in radio. different frequency bands will obviously be subject to different restrictions, for example, those of Part 15 of the FCC specifications [74] or those governing operation in a satellite band. Certain bands would have strong restrictions against certain power level transmission(GPS), locations (TV broadcast), or time (public safety). For example, any operation in TV white spaces must ensure non-interference with the licensed operators through either sensing techniques or a policy/regulatory database. Furthermore, the hardware architecture always provides limitations for every decision that the cognitive engine can make. For example, while OFDM might provide a high data rate and flexible spectrum occupancy, the complexity involved to transmit and receive such a waveform may not make it a practical consideration. A system with limited battery life might find a narrowband modulation technique more suitable for the power limitations while still satisfying the other QoS requirements.

The action decision module is realized through the combination of artificial intelligence and flexible (probably software defined) radio architectures. The intelligent radio must have sensors to read in the external information, actuators to effect changes (in the waveform), and an intelligent core to determine actions [221]. Sensors take in information about user, propagation, and network QoS requirements, and actuators implement the waveform to affect the required QoS. The intelligent core is an intelligent learning machine that develops the relationships between the environmental information and how to develop the waveform. In radio terminology, we refer to the actuators as knobs (turn the knob of a radio to adjust the carrier frequency), and sensors as meters (read the signal power).

- *Defining the parameters (Knobs):* Knobs are those parameters which determine the output waveform of the radio. Table 5.1. lists the PHY and, to a more limited extent, the MAC layer kobs.

- *Defining the objective functions (meters):* Meters are parameters that allow the radio to recognize its operational environment. From this awareness, it can extrapolate or extract information that relates to any problems with the link quality and calculate how far from the desired QoS the radio is. Some meters can be inherently present within the radio while others can be created by manipulating the information or creating the objective functions. The cognitive radio should be striving to meet the QoS requirements exactly, not going higher or lower. Lower QoS is obviously bad. Higher QoS is also bad because it wastes resources that could drain battery power or reduces the available resources for other users.

  The meters of interest on the PHY and MAC layers include: bit error rate (BER), signal to noise plus interference ratio (SNIR), data rate, occupied bandwidth, spectral efficiency, latency, computational complexity, and power consumption.

The meters listed above play a critical role as part of the optimization process of the cognitive radio. As discussed before, these objective functions require multi-dimensional analysis as they affect QoS on different levels. These objective functions are also not mutually exclusive; changing a single parameter affects all others in some way, to a greater or lesser extent. Figure 5.7 shows a single-layer graph of these interactions and how one objective affects others, possibly through an indirect path.

Based on these basic parameters and objective functions, we introduce several approaches to determine the proper combination of communication parameters in the following subsections.

Table 5.1. PHY and MAC knobs [220]

Symbol	Meaning
S	Signal power
B	Bandwidth
R_s	Symbol rate
Mod	Modulation type
M	Modulation order
PSF	Pulse shape filter type
α, β	Roll-off factor for root-raised cosine or Gaussian filters

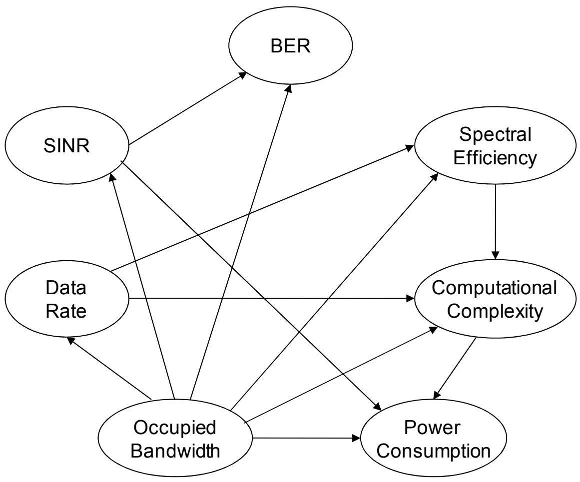


Figure 5.7 Objective function interaction and dependency [220].

Table 5.2. Objective functions [220]

Objective	Affecting Knobs
BER	$f_{\text{BER}}(S, N, I, B, R)$
SNIR	$f_{\text{SINR}}(S, N, I)$
Data Rate	$f_{\text{DR}}(R, M)$
Occupied Bandwidth	$f_{\text{OB}}(R_s, M, \alpha)$
Spectral Efficiency	$f_{\text{SE}}(R_s, M, \alpha)$
Computational Complexity	$f_{\text{CC}}(f_{\text{DR}} f_{\text{SE}}, f_{\text{OB}})$
Power	$f_{\text{P}}(f_{\text{CC}}, f_{\text{OB}}, f_{\text{SINR}})$

Each meter reflects a dependency on the knobs of Table 5.1., either directly or through another meter. Table 5.2. develops these meters as generic objective functions showing the dependency between knobs and meters. These objective functions then become the basis for evaluating the effect of the set of knobs in the optimization algorithm.

GENETIC ALGORITHM BASED APPROACH

Cognitive cycle is known to be embodied using a multiple-objective genetic algorithm (GA) for both efficient optimization of radio configuration. In this section, we investigate the genetic algorithm based approach for spectrum decision This can be used as the basis for machine learning, which will be explained in the next subsection.

Basics in Genetic Algorithm

In a genetic algorithm, a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

Once the genetic representation and the fitness function are defined, GA proceeds to initialize a population of solutions randomly, then improve it through repetitive application of mutation, crossover, inversion and selection operators.

Simple generational genetic algorithm pseudo-code is presented as follow:

1. Choose the initial population of individuals
2. Evaluate the fitness of each individual in that population
3. Repeat on this generation until termination: (time limit, sufficient fitness achieved, etc.)
   - Select the best-fit individuals for reproduction
   - Breed new individuals through crossover and mutation operations to give birth to offspring
   - Evaluate the individual fitness of new individuals
   - Replace least-fit population with new individuals

Crossover is a genetic operator that combines two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. On the contrary, *mutation* is a genetic operator that alters one ore more gene values in a chromosome from its initial state. This can result in entirely new gene values being added to the gene pool. With these new gene values, the genetic algorithm may be able to arrive at better solution than was previously possible.

Genetic Optimization for cognitive radio

The genetic formulation is not only convenient for talking about radio adaptation and behavior, but genetic algorithms offer many advantages to the optimization problem. Genetic algorithms (GAs) are powerful and flexible optimization algorithms. They are flexible in their analysis of problems as long as the chromosome and the objective functions are defined properly for a particular domain. The internal behavior of the algorithm is largely domain independent as crossover and mutation operations have little to do with the specifics of the optimization problem. They are also powerful especially in this application, because of their convergence behavior. It is widely accepted that GAs take a long time to find the optimal solutions, but, conversely, they take a short time to find very good solutions. Similarly, the radio domain does not require the absolute best solution at any given time, just a good enough solution to maintain a communication link. As radio environments change, there cognitive radio must track the changes, and GAs have been shown to excel at this task.

Finally, the multi-objective optimization problem defines one of the greatest challenges to cognitive radio realization. Multi-objective optimization problems pose problems that standard optimization techniques cannot often handle, and so they have become their own field of study. GAs have proven themselves well-suited to multi-objective optimization. Below, we explain in greater detail how to apply GAs to the cognitive radio multi-objective optimization problem.

- *Defining the Radio as a Chromosome:* The first two tasks in creating a genetic algorithm are to define the chromosome structure and develop the fitness, or objective, functions to measure the fitness of the chromosomes. The objective functions have been discussed already and summarized in Table 5.2.. The chromosome must then represent the radio in such a way that it fully defines the radio's behavioral traits and their interdependencies and is useful in the optimization process.

  Because the chromosomes of a GA are simply vectors of data structures, the genes are represented by data types define genes in a slightly different way, where the basic data structure of the chromosome is a bit, and arbitrary collections of bits are combined to define a gene.

  A chromosome defined in this way provides an extra level of flexibility to better represent a given radio platform. The idea is to use the minimum number of bits to represent all possible values that a radio knob setting may take without losing the level of precision necessary. A gene for frequency in a spectrally agile radio covering a few GHz may therefore take twenty or more bits to represent all possible frequency values, but the same radio may only have a half a dozen modulation values to choose from, which will require three or four bits.

- *Genetic Operators: Crossover and Mutation:* The genetic algorithm then performs crossover and mutation in standard ways. There is both a crossover and mutation rate, and the algorithm has a flexible number of crossover points to use. If crossover occurs on two parents to create an offspring, a set of random crossover points are selected in the chromosome. These crossover points can actually split parameter values, possibly cutting the frequency gene in half. However, this cut would preserve the higher frequency value in one offspring and the lower frequency value in the other offspring.

- *Multi-objective Optimization over Single-objective Optimization:* The choice of the radio parameters at all layers affects the radio's behavior in many dimensions. Bit error rate(BER), bandwidth, power consumption, and network latency are Just a few examples. Each of these dimensions has some relationship to the QoS, and these relationships change in their relative importance, depending on the application being used as shown in Figure 5.7. To define this multi- objective optimization problem properly, a multi- objective genetic algorithm (MOGA) is used as a powerful algorithmic approach to adapting a radio autonomously.

  The radio user has some desirable operation in mind that values certain goals more than others, such as the minimum latency requirement of a video conference. Each optimization dimension in the radio is associated with a weight to delineate the relative importance of the goals in the decision-making process.As the MOGA analyzes each dimension, optimization in the higher-weighted dimensions leads to a solution tailored to the user's preferences.

  In wireless communications, like many real-world problems, the interdependence of the operation aspects to each other and to various performance requirements makes it difficult to analyze the system in terms of any one single objective, creating a highly complex search space. Furthermore, the needs of the user and of the network cannot all be met simultaneously, and these needs can change dramatically over time or between applications.

  Because it could potentially depend on the user and the application, the search space is more complex. For certain users or applications, different objectives will mean different levels of quality. Given that the overall optimization goal is to provide the best quality of service to the user, there is no single search space that can account for all the variations in needs and wants from a given radio-user relationship.

  From this analysis, a few important points have been developed about ways to analyze the multiple objectives used in optimizing a radio:

  - There are many objectives, making a large N-dimensional search space
  - Different objectives may only be relevant to certain applications/needs;
  - The needs and subjective performances of users and applications vary
  - The external environmental conditions determines what objectives are valid and how they are analyzed

- *Multi-Objective Optimization Methods:* A multi-objective optimization algorithm is a mathematical method for choosing the set of parameters that best optimizes over the set of objective functions.

  A basic for formula for defining multi-objective optimization is shown below

$$
\begin{aligned}
\min / \max \{\bar{y}\} =& f(\bar{(x)}) = [f_1(\bar{(x)}), f_2(\bar{(x)}), \ldots, f_n(\bar{(x)})] \\
\text{subject to: } & \bar{x} = (x_1, x_2, \ldots, x_m) \in X \\
& \bar{y} = (y_1, y_2, \ldots, y_m) \in Y
\end{aligned}
\tag{5.14}
$$

Where there are n dimensions in the search space and $f_n(\bar{x})$ defines the mathematical function to evaluate dimension n. Both x, the set of input parameters, and y, the set of dimensions, may be constrained to some space, X and Y. The optimal solutions lie on the Pareto front, which is the set of input parameters, x, that is non-dominated in any dimension, which is often a trade-off of goals.

Multi-objective genetic algorithms define multiple fitness functions according to specific objectives and calculate their values on the population of competing chromosomes. The algorithm evolves in order to pick the one (or more) chromosome with the best performance to balance the combined fitness. Through the evolving process, which consists both of genetic operations and selection, the Pareto front moves so that the optimal solution provides the most efficient performance for the user's QoS requirements under radio-domain and regulatory constraints. Here, efficiency and optimization mean providing a QoS without over-maximizing, which may waste radio resources such as spectrum and power. For example, a user sending email does not need a 100Mbps link with a 30 dB carrier to noise ratio.

- *Genetic Selection Method:* Multi-objective genetic algorithms have been around for decades, but many methods exist to realize the process of selection and fitness evaluation. The most promising techniques involve an analysis of the solution space with respect to the Pareto front, the set of non-dominated solutions in the population.

  In the genetic algorithm for wireless communications, a Pareto ranking method of selection and evaluation has been chosen. Each member of the population is compared to all other members in all objective domains. Any member that is not dominated in all objectives by another member is a non-dominated solution and given a rank of 1. Other members are ranked by the number of solutions that dominate them. Fitness is therefore a discrete value between 1 and the total population size (where a solution might be dominated by all other solutions). The population can then be sorted in terms of this fitness.

  The biggest problem in multi-objective optimization is making the final decision. Evolving the Pareto front to the set of optimum solutions must then be translated in the final generation to a single decision that is a trade-off between the objectives. The chosen solution must therefore represent the user's and application's preferences. This is where the weights enter the algorithm. By weighting, the importance of each objective can help pinpoint a solution on the Pareto front that properly represents the objectives. In a situation where data rate is more critical to the user/application than minimizing occupied bandwidth, choosing a solution that has a higher symbol rate at the expense of a larger bandwidth may be the better one.

MACHINE LEARNING-BASED APPROACH

In wireless networks, cognitive radios offer the idea of intelligent radios that can adapt to their environment. Most of the first research was focused on policy-based radios that are hard-coded with a list of rules of how the radio should behave in certain scenarios. However, cognitive radios should have the context-awareness and intelligence abilities necessary to respond to the environment in an efficient manner and this is exactly what machine learning tries to achieve. For example, a cognitive radio can observe the transmission activity of the primary user on different channels, which enables the building of knowledge about primary userfls activity. When the channel has to be accessed by the secondary user, the knowledge is used in order to achieve the desired performance.

A question may arise concerning when machine learning techniques are more appropriate than game theory. If the problem framework can be viewed as a set of players (cognitive radios) competing in a noncooperative way for the same resources, then difficulties appear when trying to represent such a stochastic game using classical game theory due to the incomplete information at the players involved. Machine learning algorithms usually perform better in this context.

Machine learning algorithms can be mainly divided into three different categories: supervised learning, unsupervised learning and reinforcement learning. Their characteristics are summarized as follows:

- Supervised learning: the algorithms are developed to deduce a function from training data, which is composed of input vectors and their corresponding outputs. The objective is always to predict an output given any valid input. Two are the main type of problems where supervised learning is applied: regression and classification. When the output is a continuous value it is called a regression problem whereas if the task is to predict and assign a label of the input object it is called a classification problem.

- Unsupervised learning: in this class of problems there is also a set of training data but no correct output is specified. In contrast, the objective of the algorithms is just to identify the structure of the input data by detecting that certain patterns occur more often than others. Clustering is the most common example.

- Reinforcement learning: this approach is concerned with how an agent should take actions in an environment in order to maximize its long-term reward. A sequence of actions mapped to states, which is referred to as a policy, is obtained using the reinforcement learning algorithm. The agent focuses on the on-line performance while tries to find the best tradeoff between exploration (i.e. explore all feasible actions and consequences) and exploitation (i.e. make use of the accumulated knowledge).

Of the three families of algorithms, reinforcement learning is the most appropriate one for cognitive radio networks because it can be used without training data and it is aimed at maximizing the long-term reward. However, there also exist solutions in the field of supervised and unsupervised learning that are useful in the context of cognitive radio.

SPECTRUM-ADAPTIVE TRANSMISSION

In [206], the adaptive protocols are developed to determine the transmission power as well as the best combination of modulation and error correction code for a new spectrum band by considering changes in the propagation loss. In this work, two low-complexity protocols for cognitive radios are proposed for responding to slow variations in the channel, such as changes in shadow loss.

- At the start of a new session, which may be in a different frequency band than the previous session, the protocol must adjust the transmitter power to provide reliable communications with minimal energy consumption and minimal interference to other radios. The power-adjustment protocol enables successful communication in a new frequency band with an unknown propagation loss while avoiding the transmission of excessive power that would cause interference in unintended receivers and disrupt other sessions in the network.

- As the session progresses, the protocol must adjust the transmissions to compensate for changes in channel conditions. The adaptive transmission protocol chooses the combination of modulation and error-control code that is best suited to the channel state without having to estimate the state.

Initial Power Adjustment

When a new session begins, the source's transmitter power may be much higher than necessary or much lower than required for reliable communication, especially if the session's frequency band has not been used recently by the source and destination.

It is necessary to adjust the transmitter power as quickly as possible after the session is underway to obtain an acceptable packet error probability without causing interference to other sessions in the network. For the proposed protocol, each session begins with the code-modulation combination that has the highest information rate.

During the power-adjustment phase, a receiver statistic is included in each acknowledgment packet. A simple interval test is performed on the statistic to determine the power for the next packet. If the initial power is too low, it may be that the destination is not even aware that a packet was sent. Thus, for each unacknowledged packet that is sent during the power-adjustment period, the source automatically increases the power by a fixed amount (e.g., 5 dB). The termination of the power-adjustment phase is determined by a stopping condition that is applied to the demodulator statistics for the initial sequence of packets in the session. Once the power-adjustment phase is completed, the adaptive transmission protocol takes over and compensates for changes that might occur in the channel during the remainder of the session. In general, the response to a deterioration in channel conditions is to switch to a more robust, lower-rate, code-modulation combination. An improvement in channel conditions is exploited by switching to a code-modulation combination that provides a higher information rate.

Adaptive Transmission Protocol

After the initial power adjustment is complete, the adaptive transmission protocol governs the choice of modulation and error-control coding.

The receiver statistic that is employed for adaptive coding and modulation is either the error count or the iteration count; the two are approximately equally effective. For each packet that decodes correctly, the protocol performs an interval test to determine which code-modulation combination to use for the next packet. For each packet that does not decode correctly, the next packet is sent with the next lower-rate code-modulation combination than was used for the failed packet. If the failed packet used the lowest rate code-modulation combination, then the next packet uses the same combination again. After several consecutive failures with the combination of lowest rate, the protocol may be required to increase the transmitter power if that is permitted by the spectrum access protocol.

For each adaptation statistic, each code-modulation combination is associated with three decision intervals, $\mathcal{I}_{-1} = (\infty, \gamma_1)$, $\mathcal{I}_0 = [\gamma_1, \gamma_2]$, and $\mathcal{I}_1 = (\gamma_2, \infty)$. The decision interval

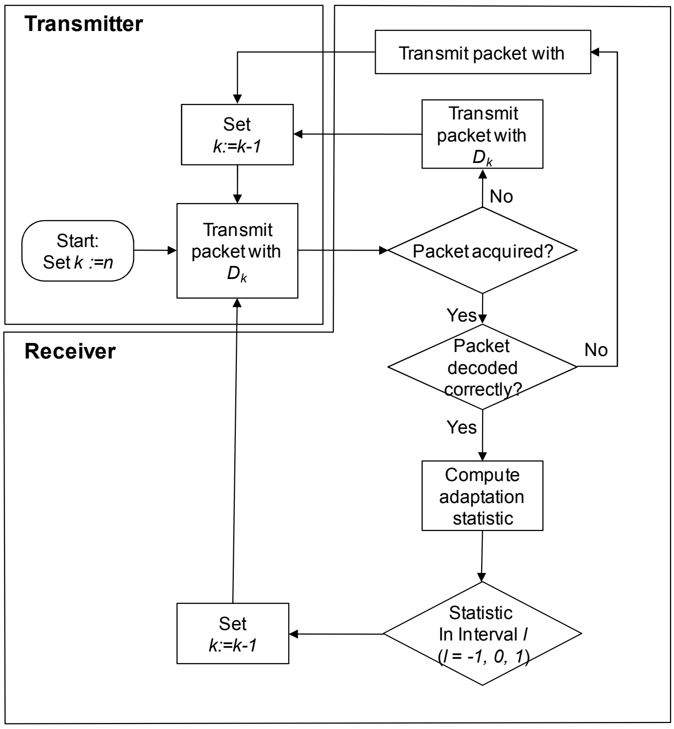


Figure 5.8 Flow chart for the adaptive transmission protocol.

endpoints, γ_1 and γ_2, define completely the decision intervals. The endpoints can be adjusted during the session by a learning process within the adaptive transmission protocol, so we distinguish between the design intervals, which are the nominal intervals selected by the system designer, and the adjusted intervals, which are derived from what the cognitive radio learns from its past decisions as the session progresses.

Because the adaptation statistics for the iteration count and the error count decrease as the channel quality increases, the description of the adaptation decision process is the same for the two. Let z_i denote the statistic obtained from the receiver as a result of the reception of the ith packet; that is, z_i is either the error count or the iteration count for the ith packet, depending on which receiver statistic is used for the session. Suppose that code-modulation combination D_k was used for packet i. The code-modulation combination for packet i+1 is D_k if $z_i \in \mathcal{I}_0$, D_{k+1} if $z_i \in \mathcal{I}_{-1}$, and D_{k-1} if $z_i \in \mathcal{I}_1$. The interval tests are equivalent to testing $z_i < \gamma_1$ and $z_i > \gamma_2$ (if neither, then $z_i \in \mathcal{I}_0$). The complete operation of the adaptive transmission protocol is illustrated in Figure 5.8.

5.3 SPECTRUM DECISION CHALLENGES

For the development of spectrum decision function, several challenges still remain unsolved.

- *Decision Model:* The estimation of spectrum capacity based on the signal to noise ratio (SNR) is not sufficient to characterize the spectrum band in CR networks. Also applications require different QoS requirements. Thus, design of application- and spectrum-adaptive spectrum decision models is still an open issue.

- *Spectrum Decision over Heterogeneous Spectrum Bands:* Currently, certain spectrum bands are already assigned to different purposes while some bands remain unlicensed. Thus, CR network should support spectrum decision operations on both the licensed and the unlicensed bands.

- *PU Activity Modeling:* Most of the current research on spectrum sensing are based on a simple ON-OFF model for PU activities, which cannot capture the diverse characteristics of all existing primary networks. This inaccurate model for primary networks leads to an adverse influence on spectrum sensing resulting in either lower spectrum access opportunities or higher interference to the primary networks. Some of the empirical models on PU activities [95] [273] are not computationally feasible in practical situations. Thus, we need to develop more practical PU activity models by considering the characteristics of access technologies as well as traffic types.

- *Joint Spectrum Decision and Reconfiguration Framework:* Once the available spectrum bands are characterized, the most appropriate spectrum band should be selected by considering the QoS requirements (sustainable rate, delay, jitter, average session time, acceptable loss rate, etc) and the spectrum characteristics. However, according to the reconfigurable transmission parameters such as modulation type, error control scheme, and communication protocol, these spectrum characteristics change significantly. Sometimes, with only reconfiguration, CR users can maintain the quality of the current session. For example, even if SNR is changed, bit rate and bit error rate (BER) can be maintained by exploiting an adaptive modulation, instead of changing spectrum and route. Hence, there is a need for a joint spectrum decision and reconfiguration framework to find the optimal combination of the spectrum band and parameter configuration according to applications with diverse QoS requirements.

CHAPTER 6

SPECTRUM SHARING

The shared nature of the wireless channel necessitates coordination of transmission attempts between CR users. In this respect, spectrum sharing provides the capability to maintain the QoS of CR users without causing interference to the primary users by coordinating the multiple access of CR users as well as allocating communication resources adaptively to the changes of radio environment. Thus, spectrum sharing includes many functionalities of a medium access control (MAC) protocol and resource allocation in classical wireless networks. However, the unique characteristics of cognitive radios such as the coexistence of CR users with PUs and the wide range of available spectrum incur substantially different challenges for spectrum sharing in CR networks.

BASIC FUNCTIONALITIES

Figure 6.1 depicts the functional blocks for spectrum sharing in CR networks. Unlike spectrum decision, spectrum sharing mainly focuses on resource management within the same spectrum with the following functionalities:

- *Resource Allocation:* Based on the QoS monitoring results, CR users select the proper channels (*channel allocation*) and accordingly adjust their transmission power (*power allocation*) so as to achieve QoS requirements as well as resource fairness. Furthermore, resource allocation in CR networks is constrained by interference to other CR and primary users.

- *Spectrum Access:* It enables multiple CR users to share the spectrum resource by determining who will access the channel or when a user may access the channel. This

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

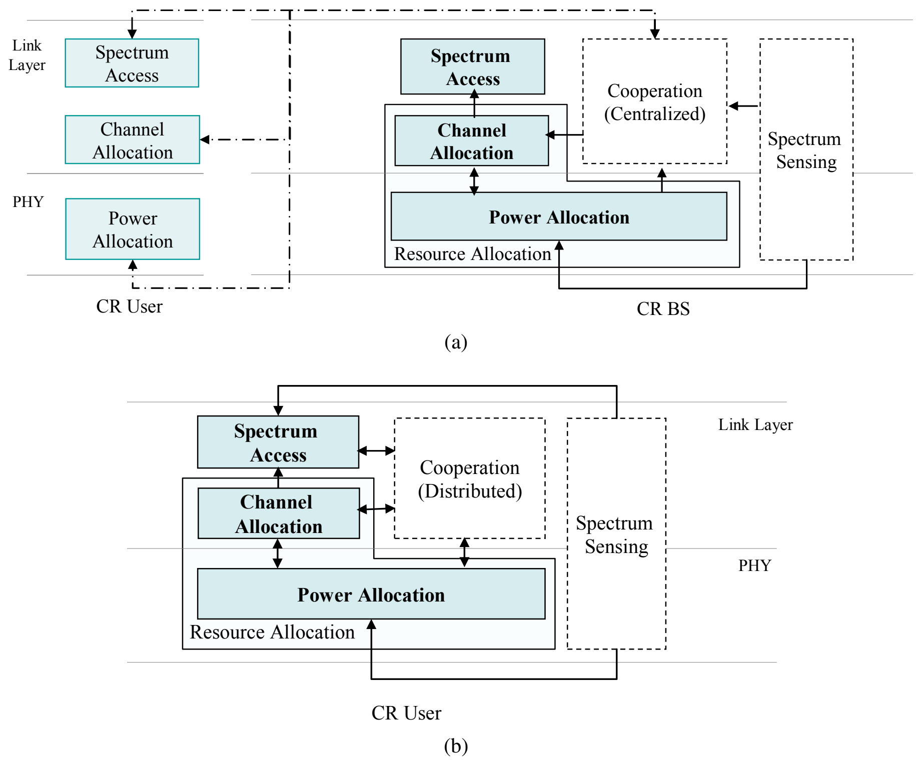


Figure 6.1 Functional block diagram for spectrum sharing: (a) infrastructure-based CR networks, and (b) CR ad hoc networks.

functionality includes a MAC capability. However, unlike classical MAC protocols, the spectrum access in CR networks is closely coupled with spectrum spectrum sensing, especially in sensing control described in Section 4.4.

Once a proper spectrum band is selected in spectrum decision, communication channels in that spectrum need to be assigned to a CR user while determining its transmission power to avoid the interference to the primary network (*resource allocation*). Then, the CR user decides when the spectrum should be accessed to avoid collisions with other CR users (*spectrum access*).

ARCHITECTURAL FEATURES

The infrastructure-based network can provide sophisticated spectrum sharing method with support of the base-station. Thus, it can exploit time slot-based scheduling and dynamic channel allocation to maximize the total network capacity as well as achieve fair resource allocation over CR users. Furthermore, through the synchronization in sensing operation, the transmission of CR users and primary users can be detected separately, which decouples sensing operation with spectrum sharing. Generally, CR networks use a periodic sensing scheme where CR users are allowed to transmit only during the transmission period followed by the sensing (observation) period. In this architecture, the transmission period is synchronized over all CR users. Thus, spectrum sharing needs to focus on channel allocation or time-slot-based scheduling within this transmission period. Also spectrum sharing just exploits the spectrum availability and is not directly related to spectrum sensing. Similar to spectrum sensing and decision, all sharing operations in CR users are coordinated by the base-station, as illustrated in Figure 6.1.

On the contrary, in CR ad hoc networks, the sensing schedules are not synchronized over all users due to lack of the central network entity, and hence are independent of each other. Thus, instead of the periodic sensing, CR ad hoc users may adopt the aperiodic or on-demand sensing triggered by only spectrum sharing operations, i.e., when CR users want to transmit or are requested their spectrum availability by neighbor users. Furthermore, sensing and transmission intervals, determined by the sensing control in spectrum sensing, influence the performance of spectrum access. For these reasons, spectrum sensing should be integrated into spectrum sharing, especially in spectrum access functionality, which is shown in Figure 6.1 (b).

Among the functionalities, spectrum access is closely related to MAC protocols, and hence is explained in Chapter 8 in more detail. In the following subsections, we describe unique features in spectrum sharing, especially focusing on resource allocation in CR networks. In Sections 6.1 and 6.2, we classify the spectrum sharing techniques and describe the fundamental approaches to these techniques, respectively. These work provide insight about how a spectrum sharing functionality can be designed. Accordingly, in Sections 6.3 and 6.4, we overview the solutions for spectrum sharing among multiple coexisting CR networks (*inter-network spectrum sharing*), and inside a CR network (*intra-network spectrum sharing*), respectively. Finally, in Section 6.5, the open research issues for spectrum sharing in CR networks are discussed.

6.1 OVERVIEW OF SPECTRUM SHARING TECHNIQUES

The existing solutions constitute a rich literature for spectrum sharing in CR networks, which can be mainly classified in five aspects: i.e., according to their *spectrum policy, network architecture, entities' behavior, sharing strategy*, and *sharing scope* as shown in Figure 6.2.

First, since spectrum is a finite but reusable resources, it can be distributed to multiple users through the spectrum usage model. There are two general models for assigning spectrum usage rights as follows:

- *Licensed Spectrum Sharing:* A licensee, i.e., a primary user has exclusive and transferable rights to the use of specified spectrum within a defined geographic area, with flexible use rights that are governed primarily by technical rules to protect primary users against interference. Under this model, exclusive rights resemble property rights in spectrum, but this model does not imply or require creation of "full" private property rights in spectrum.

- *Open Spectrum Sharing:* This model allows unlimited numbers of unlicensed users to share frequencies, with usage rights that are governed by technical standards or etiquettes but with no right to protection from interference. Spectrum is available to all users that comply with established technical "etiquettes" or standards that set power limits and other criteria for operation of unlicensed devices to mitigate potential interference.

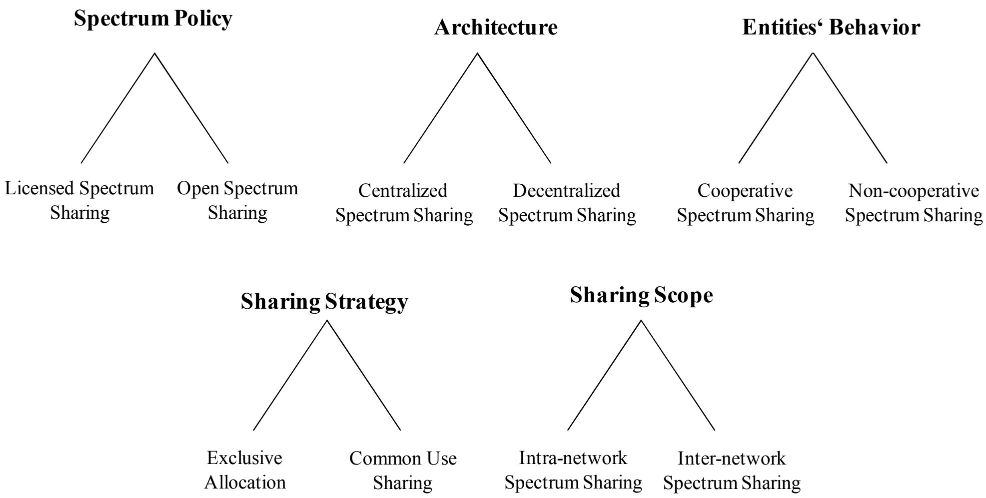


Figure 6.2 Classifications of spectrum sharing in CR networks.

There exist temporally unused spectrum holes in the licensed spectrum band. Hence, CR networks can be deployed to exploit these spectrum holes through cognitive communication techniques. In this case, the CR network coexists with the primary network at the same location and on the same spectrum band. The spectrum sharing scheme in this case should consider the interference avoidance issue to the primary network, along with efficient resource allocation among CR users. In addition, CR networks can be designed for operation on unlicensed bands such that the efficiency is improved in this portion of the spectrum. Since there are no license holders, all network entities have the same right to access the spectrum bands. Multiple CR networks coexist in the same area and communicate using the same portion of the spectrum.

The second classification for spectrum sharing techniques in CR networks is based on the architecture, which can be described as follows:

- *Centralized Spectrum Sharing:* In these solutions, a centralized entity controls the spectrum allocation and access procedures. Each entity in the CR network forward their measurements about resource allocation to the central entity and this entity constructs a resource allocation map.
- *Distributed Spectrum Sharing:* Distributed solutions are mainly proposed for cases where the construction of an infrastructure is not preferable. Accordingly, each node is responsible for the resource allocation and access is based on local (or possibly global) policies.

The third classification for spectrum sharing techniques is based on the behavior of the CR user. More specifically, the operations for spectrum sharing can be either *cooperative* or *non-cooperative* as explained below:

- *Cooperative Spectrum Sharing:* Cooperative (or collaborative) solutions consider the effect of the node's communication on other nodes. In other words, the interference measurements of each node are shared among other nodes. Furthermore, the spectrum sharing algorithms also consider this information. While all the centralized solutions can be regarded as cooperative, there also exist distributed cooperative solutions.

- *Non-cooperative Spectrum Sharing:* Contrary to the cooperative solutions, non-cooperative (or non-collaborative, selfish) solutions consider only the node at hand. While non-cooperative solutions may result in reduced spectrum utilization, the minimal communication requirements among other nodes introduce a tradeoff for practical solutions.

Next, the fourth classification for spectrum sharing in CR networks is based on the access strategy as explained below:

- *Exclusive Allocation:* Spectrum resource can be assigned to only one user to avoid interference to other neighbor users, which mainly focuses on the allocation of non-overlapping orthogonal channels.

- *Common use sharing:* This solution allows multiple users to access the same spectrum at the same time. Thus, in this approach, power allocation is the most important part to increase the capacity with less interference to other users. In the power allocation, the CR user needs to adjust its transmission power by considering co-channel (or inter-user) interference. In addition, power allocation should be based on the PU activities in its transmission not to violate the interference constraints. Cooperation among neighbors helps to enhance the performance of spectrum sharing, especially in power allocation which should be aware of the PU activities in the transmission range.

Finally, spectrum sharing techniques are generally focused on two types of solutions: spectrum sharing inside a CR network (intra-network spectrum sharing) and among multiple coexisting CR networks (internet-work spectrum sharing), as explained in the following:

- *Intra-network spectrum sharing:* These solutions focus on spectrum allocation between the entities of a CR network, as shown in Figure 6.3. Accordingly, the users of a CR network try to access the available spectrum without causing interference to the primary users. Intra-network spectrum sharing poses unique challenges that have not been considered previously in wireless communication systems.

- *Internet-work spectrum sharing:* The CR architecture enables multiple systems to be deployed in overlapping locations and spectrum, as shown in Figure 6.3. So far the internet-work spectrum sharing solutions provide a broader view of the spectrum sharing concept by including certain operator policies.

6.2 BASIC APPROACHES FOR SPECTRUM SHARING

In this section, we overview the basic theories used for spectrum sharing. Most of existing techniques in spectrum sharing are the combinations of these approaches discussed in this section.

6.2.1 Game-Theoretic Approach

In general, game theoretic approaches have been exploited to determine the communication resources of each user in CR networks [70] [128] [193]. Each CR user has a common interest to use the spectrum resources as much as possible. However, CR users have competing

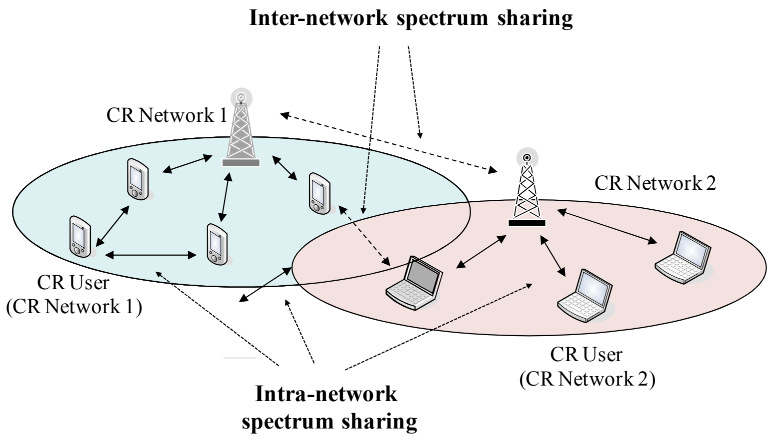


Figure 6.3 Inter-network and intra-network spectrum sharing in CR networks.

interests to maximize their own share of the spectrum resources. i.e., the activity of one CR user can impact the activities of the others. Furthermore, the rational decisions of a CR user must be undertaken while anticipating the responses of its rivals. Game theory provides an efficient distributed spectrum sharing scheme by describing the conflict and cooperation among CR users, and hence allowing each CR user to rationally decide on its best action.

BASIC COMPONENTS

In game theory, the output (*outcomes*) of the process (*game*) is the function of the inputs (*actions*) from several different decision makers (*players*) who may have potentially conflicting objectives (*preferences*) with regards to the outcome of the process. In CR networks, each game component: players, preferences, actions, and outcomes are interpreted as follows: 1) *players* will be either CR users or PUs, 2) the *preferences* can be considered as the communication metrics that must be optimized, such as throughput or delay, and are expressed in the form of a utility function, 3) *actions* represent the choice of the communication resources (channel, transmission power) made by a player, and 4) *outcomes* is the observed performance of the network (SNR, bandwidth allocation) as a result of the individual actions.

GAME MODELS

Let a finite set of players be $\mathbf{N}$, a set of actions (or strategies), $\mathbf{A}_i$, for each player i, and payoff/utility function, $u_i : \mathbf{A} \rightarrow \mathbf{R}$, which measures the outcome for player i determined. Here $\mathbf{R}$ is a set of real number and $\mathbf{A}$ is a set of all possible combinations of actions denoted by $\mathbf{A}_1 \times \mathbf{A}_2 \times \cdots \times \mathbf{A}_{\mathrm{N}}$, called *action profile*.

Depending on the relationship between these components, game theoretic approaches can exploit diverse game models. Among them, the following game models are mainly considered for spectrum sharing in CR networks [191].

- *Normal (or Strategic) Form Game:* This is a simple and basic model in game theory. In this model, all players make their decisions simultaneously and this process occurs only once for each player. Furthermore, they are assumed to be aware of not only their own utility functions but also the utility functions for all the other players in the game. This type of game can be specified by 3-tuple $\Gamma = \langle \mathbf{N}, \mathbf{A}, \{u_j\} \rangle$.

- *Repeated Game:* This model is defined as a sequence of stages, where each stage is the a normal form game. Based on the past actions, current observations, and future expectations, players determine their actions at each stage. The actions of each player are assumed to be synchronized. In this model, the action strategies can be updated in each stage adapting to the actions and outcomes observed previously. Based on the outcome of each stage of the game, the players can incorporate punishment and reward strategies, which are well-suited for wireless networks. If a player deviates from the previously negotiated strategy, the other players choose their actions so as to reduce the outcome of the offending player. This game is specified by 4-tuple $\Gamma = \langle \mathbf{N}, \mathbf{A}, \{u_j\}, \{d_j\} \rangle$, where d_j is the decision rule.

- *Asynchronous Myopic Repeated Game:* A *myopic repeated game* is a repeated game where the strategy update of a player is based on only its observation of the game at the most recent stage. Since players in a myopic repeated game are not able to consider future outcomes in determining the current actions, they employ simpler myopic strategies, instead of complex multi-stage strategies used in general repeated games. Here, all decisions at each stage are made simultaneously, similar to the classical repeated games. However, the myopic repeated games model may not be feasible for distributed wireless networks, such as CRAHNs. This is because CRAHNs may require random or asynchronous decisions due to the absence of a central network entity. In this case, an *asynchronous myopic repeated game* provides a better model for spectrum sharing, in which decisions do not have to be made synchronously. In this model, the actions of each player adapt to the most recent state of networks under a variety of different decision timings. This game is specified by 5-tuple $\Gamma = \langle \mathbf{N}, \mathbf{A}, \{u_j\}, \{d_j\}, T_j \rangle$, where T_j is the decision timing.

- *Mixed (or Probabilistic) Strategy Game:* Some of normal form games may not have a steady-state solution, called *Nash equilibrium* where no selfish CR user has incentive to unilaterally change its action. To overcome this limitation, game theoretic approaches introduce a mixed strategy game, where players employ their strategies based on the probabilities of each action. This approach achieves the Nash equilibrium even though it does not exist in pure strategies. This can be specified by 3-tuple $\Gamma = \langle \mathbf{N}, \Delta(\mathbf{A}), \{\mathbf{U}_j\} \rangle$, where $\{\mathbf{U}_j\}$ is the expected utility of user j, and $\Delta(\mathbf{A})$ is all possible mixed strategy tuples. Here, the mixed strategy for user i is defined as the probability of each action for user i.

NASH EQUILIBRIUM

In game theory, a Nash equilibrium is a key concept to analyze the outcome of the strategic interaction of multiple decision makers, which is defined as a set of action strategies if any player cannot do better by unilaterally changing its strategy. This concept provides a way of predicting what will happen if several selfish users are simultaneously making decisions, and if the decision of each user is dependent on the those of the others.

This concept is informally explained as follow: Each player determines if it can have any benefit by changing its strategy, assuming that it is aware of the current strategies of other players, which is unchangeable. If any player would benefit by the change, that set of strategies is not a Nash equilibrium. Consequently, in the Nash equilibrium, each strategy is a best response to all other strategies.

Mathematically, the Nash equilibrium of a strategic game $\langle N, (\mathbf{A}_i), (u_i) \rangle$ is expressed as a set of action strategy profile $\mathbf{a}^* \in \mathbf{A}$ of actions such that

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*), \forall i \in N, \forall a_i \in \mathbf{A}_i \tag{6.1}$$

where a_i denotes the strategy of player i and a_{-i} denotes the strategies of all players other than player i.

The definition implies that no player can improve its payoff by a unilateral deviation from the Nash equilibrium, given that the other players maintain their strategies. As a result, the Nash equilibrium can be defined as the best-response strategy of each player, as follows:

$$a_i^* \in B_i(a_{-i}^*) \quad \forall i \in N \tag{6.2}$$

where B_i is the best-response function of player i, i.e.,

$$B_i(a_{-i}) = \{a_i \in \mathbf{A}_i : u_i(a_{-i}, a_i) \geq u_i(a_{-i}, a_i')\} \quad \forall a_i' \in \mathbf{A}_i \tag{6.3}$$

EXAMPLE

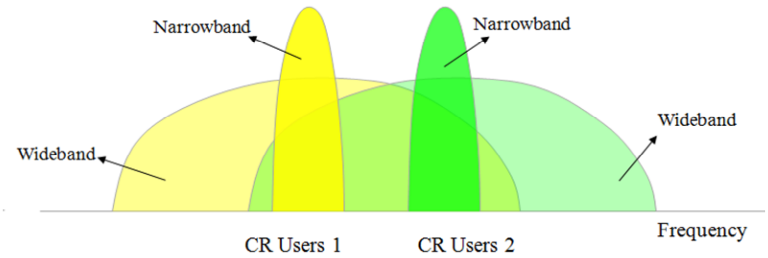
The following is the example to show how the game theory is applied to resource allocation in wireless networks [191]. Suppose two CR users are operating in the same environment and are attempting to maximize their throughput. Each user can implement two different waveforms- one a low-power narrowband waveform n, the other a higher power wideband waveform w. If both radios choose to implement their narrowband waveforms- action vector (n, n)- the signals will be separated in frequency and each radio will achieve a throughput of 9.6 kbps. If one of the radios implements its wideband waveform while the other implements its narrowband waveform - action vectors (n, w) or (w, n) - then interference occurs with the narrow band signal achieving a throughput of 3.2 kbps and the wideband signal a throughput of 21 kbps. If both radios implement wideband waveforms, then each radio experiences a throughput of 7 kbps. Here, the choice of the wideband waveform at each user (w, w) is the Nash equilibrium since other player cannot improve its throughput by unilaterally changing its strategy from this case.

These waveforms can be visualized in the frequency domain and represented in matrix form as shown in Figure 6.4. Without going into the analysis of this game, the insightful reader may already anticipate that this algorithm tends to lead to less than optimal performance.

Although the game theoretic approaches can achieve the Nash equilibrium, they cannot guarantee the Pareto optimum, leading to lower network capacity.

6.2.2 Auction-Based Approach

Generally, a non-cooperative game with incomplete information is known to be complex and difficult to solve since each selfish player does not know the perfect strategy profile of others. However, auction theory encourages each selfish user to reveal its private information as a bid, guaranteeing that it is not harmful to disclose the private information. Based



(a)

CR user 1		CR users 2 (kbps)	
		Narrowband	Wideband
	Narrowband	(9.6,9.6)	(3.2, 21)
	Wideband	(21,3.2)	(7,7)

(b)

Figure 6.4 Game theory example: (a) actions (narrowband and wideband waveforms) and (b) payoff matrix [191].

on bids, auction theory provides efficient allocation of scare resources including the sale of single-item indivisible goods (e.g. a painting), single-item in multi-unit bundles [230] and multi-item, multi-unit bundles (e.g. bonds) [258].

In auction theory, bids express user's preference for various outcomes. Auctioneers use auction-clearing algorithms, and accordingly determine the price and allocation maximizing their revenue on the basis of bids from the players (bidders). For example, in single-item single-unit auction, an item is assigned to a bidder with the highest bid. However, the clearing algorithm of multi-unit auctions is much more complex since multiple winners split the items [222]. Thus, how to design efficient bidding processes and fast clearing algorithms is the most important issue in auction theory. Diverse communication resources such as spectrum (or channel) or power can be considered as auction items, but for simplicity, we mainly focus on spectrum auction in this section.

BASIC SPECTRUM AUCTION FRAMEWORK

In spectrum auction, an *auctioneer* performs a periodic auction of channels to n *bidders* in a geographic region, as shown in Figure 6.5. Assume that the channels have uniform characteristics and values. Bidders submit the number of channels they demand and the per-channel prices they want to pay. Given the bids, the auctioneer determines winners and prices to maximize their profit. The following are the main components consisting the basic spectrum auction model:

- *Channel request (d_i):* It represents the number of channels requested by bidder i. *Strict requests* allow a bidder to accept to receive either d_i channels or 0 channel. On the contrary, in *range requests*, a bidder accepts any x channels such that $0 \leq x \leq d_i$.

- *Per-channel bid (b_i):* It denotes the per-channel bid submitted by bidder i. $B = \{b_1, b_2, \ldots, b_n\}$ is the set of bids submitted by all the bidders.

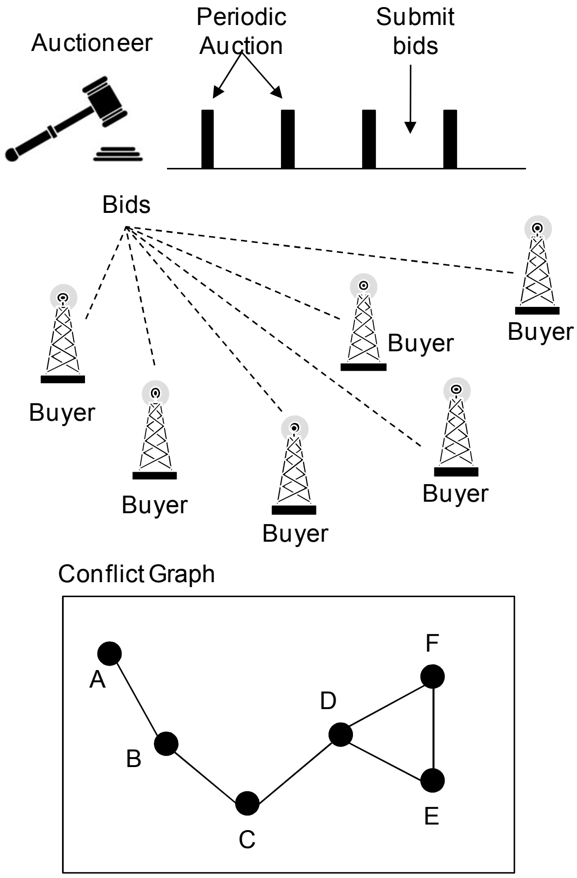


Figure 6.5 Dynamic auction scenario [222].

- *Per-channel true value (v_i):* It describes the true price i is willing to pay for each channel. Each bidder i has its own valuation, which is generally private information, and is known only to the bidder itself.

- *Clearing price (p_i):* Given bid set B, the auctioneer determines the winners of the auction, and charges price p_i for channels allocated to each winner i, called the *clearing price*.

- *Bidder utility (u_i):* This utility represents the residual worth of the channels allocated to bidder i. This is obtained as $v_i \cdot d_i^a - p_i$ if bidder i takes d_i^a channels from the auction. If the bidder obtains nothing, this utility is zero.

Although the spectrum auction has a similar framework to conventional multi-unit auctions, it shows a unique characteristic [299]. A spectrum band can be spatially reused concurrently, i.e., multiple conflicting bidders cannot use the same channels simultaneously, but well-separated bidders can. While a conventional auction with n bidders and k channels can only have at most k winners, the spectrum auction can have more than k winners.

Figure 6.5 shows a sample scenario for spectrum auction where wireless access points provide network access for their associated users [222]. For example, when A and B are located closely to each other, as shown in Figure 6.5, their associated users receive signals from both access points, causing interference to their communications. To avoid interference, A and B should not use the same spectrum band. Let F_{A} and F_{B} represent the spectrum allocation maps in A and B, respectively, such that $F_{\mathrm{A}} = \{s_1^{\mathrm{A}}, s_2^{\mathrm{A}}, \ldots, s_M^{\mathrm{A}}\}$ where $s_k^{\mathrm{A}} = 1$ if the k^{th} channel is assigned to A, and otherwise 0. M is the total number of spectrum bands. Then, the interference constraint of bidder i can be expressed as $F_{\mathrm{A}} \cap F_{\mathrm{B}} = \varnothing$ or $s_k^{\mathrm{A}} s_k^{\mathrm{B}} = 0$, $\forall k \in [1, M]$. Then, the interference constraint of bidder i can be expressed as $F_{\mathrm{A}} \cap F_{\mathrm{B}} = \varnothing$ or $s_k^{\mathrm{A}} s_k^{\mathrm{B}}$. Then, the auction clearing problem can be translated into aa special case of the non-linear integer programming problem problem to find optimal spectrum allocation to maximize auctioneer's revenue based on this interference constraint, which can be defined as follow [222]:

$$
\begin{aligned}
&\text{Maximize :} \quad \sum_{i \in \text{bidders}} f_i p_i(f_i, b_i) \\
&\text{subject to :} \\
&\qquad f_i \leq 1 \\
&\qquad |F_i| = d_i \text{ (strict requests) or } |F_i| \leq d_i \text{ (range requests)} \\
&\qquad F_i \cap F_j = \varnothing, \forall j \in C_i \quad \text{(Interference Connstraint)}
\end{aligned}
\tag{6.4}
$$

where $f_i = |F_i|/M$ represents the normalized spectrum assigned to bidder i. $p_i(f_i)$ is the per-unit price that the bidder i pays if he obtains f_i unit of spectrum. C_i is the set of access points that conflict bidder i.

The clearing price p_i is generally expressed as a function of b_i, and f_i, and can be determined in a diverse way according to auctioneer's strategy. The following is two basic principles to determine the clearing price [222]:

- *Uniform pricing:* The auctioneer determines a per-unit price p and applies it to all winning bidders. The auction clearing problem is to determine a market-clearing price that maximizes the auctioneer's revenue.
- *Discriminatory pricing:* The auctioneer charges different prices p_i to different bidders i.

In one time auctions, the uniform pricing is simple and shows better fairness to bidders while achieves lower revenue than the discriminatory pricing. However, it is suspect to collusion among the bidders, also leading to decease in the revenue of the auctioneer.

TRUTHFUL SPECTRUM AUCTION

In the auction, bidders have to strategize over others on how to bid to increase. Due to this market manipulation, the auction may not provide the best strategy in terms of either the revenue of the auctioneer or the utility of bidders. Any additional process to prevent market manipulation leads to significant overhead to both the auctioneer and bidders, which discourages bidders from participating the auction.

To solve this problem, the auction should be truthful by guaranteeing that if a bidder bids the true valuation of the resource, its utility will not be less than that when it lies. Hence, the dominating strategy for a bidder is to bid true valuation. To bidders, a truthful auction eliminates the expensive overhead of strategizing about other bidders and prevents market manipulation. Thus it can attract a wide range of network nodes/establishments to engage in the marketplace. To the auctioneer, by encouraging bidders to reveal their true valuations, a truthful auction can help the auctioneer increase its revenue by assigning spectrum to the bidders who value it the most. For this reason, many truthful auction schemes have been proposed, including the sealed-bid secondary-price [254], and Vickrey-Clarke-Groves (VCG) auctions [55] [103].

For single-item auctions, the classical truthful mechanism is the Vickrey or second-price auction [254], in which the lowest bidder is selected as a winner and is paid the second-lowest bid. This is later extended to cases where each bidder requests more than one items [153] [185]. In general, a spectrum auction based on Vickrey's mechanism defines rules for two consecutive steps: the rule to determine winners, and the rule to determine prices as follow:

[Allocation]

1. Sort the bids in descending order and set each bidder's available channel set.
2. Allocate channel m to the first bidder i in the sorted order using the lowest indexed channel in i's available channel set, remove i from the list, and remove channel m from bidder i's conflicting neighbors' available channel sets.
3. Repeat 2 until all bidders are considered.

[Pricing]

1. Charge winner i the highest bid of its unallocated conflicting neighbors. If there is no such neighbor, charge 0.

A VCG auction is a generalization of a Vickrey auction for multiple items, where the winners are decided in such a way that the social welfare is maximized, and the price charged to each winner is equal to his/her "social opportunity cost" to the whole system. Let $M = \{t_1, \ldots, t_m\}$ be the set of items, $N = \{b_1, \ldots, b_n\}$ be the set of bidders. The VCG auction can be described as follows:

1. Each bidder i is asked to submit its true valuation as bid b_i.
2. The auctioneer determines optimal resource allocation $F^* = \{f_1^*, \ldots, f_M^*\}$ that maximizes the total utility $V_{\max} = \sum_{j=1}^{M} b_j(F^*)$ achieved in a VCG auction with M items and N bidders, with a given set of bids submitted by the bidders.
3. The auctioneer computes the maximum total valuation if user i is excluded from the auction, i.e., $V_{\max/i} = \max_{F/f_i} \sum_{j \neq i} b_j(F^*)$ for each $i \in M$ where $F = \{f_1, \ldots, f_M\}$ is a set of resource allocation.
4. The auctioneer then charges user m the amount $V_{\max/i} - \sum_{j \neq i} V_j(F^*)$, which is the decrement in sum valuation over all other users from including user i in the auction.

6.2.3 Graph Theoretic Approach

For orthogonal spectrum or channel allocation, a graph theoretic approach can be used by mapping spectrum channels into colors, and assigning them to users, which are represented as vertices in a graph. In this approach, spectrum allocation is modeled as a bidirectional graph $G = (V, L, E)$, where V is a set of vertices denoting the users that share the spectrum, L is the available spectrum or the color list at each vertex, and E is a set of undirected edges between vertices representing interference between any two vertices. For any two vertices $u, v \in V$, a m-colored edge exists between u and v if $C_{u,v} = 1$. The edges depend on the interference constraint C, which is determined by the spectrum usage of nearby primary users and the transmit power of user u and v on channel m.

The spectrum allocation problem is equivalent to coloring each vertex using a number of colors from its color list to maximize system utility. The coloring scheme is constrained by that if a colored edge m exists between any two vertices, they cannot simultaneously use color m. This is a variant of the traditional graph coloring problem. In the traditional problem, graphs are colorless, colors have the same reward, and two connected vertices only have one colorless edge; in the spectrum allocation problem, vertices can be connected via multiple colored edges, refereed to as *color-sensitive graph coloring (CSGC)*.

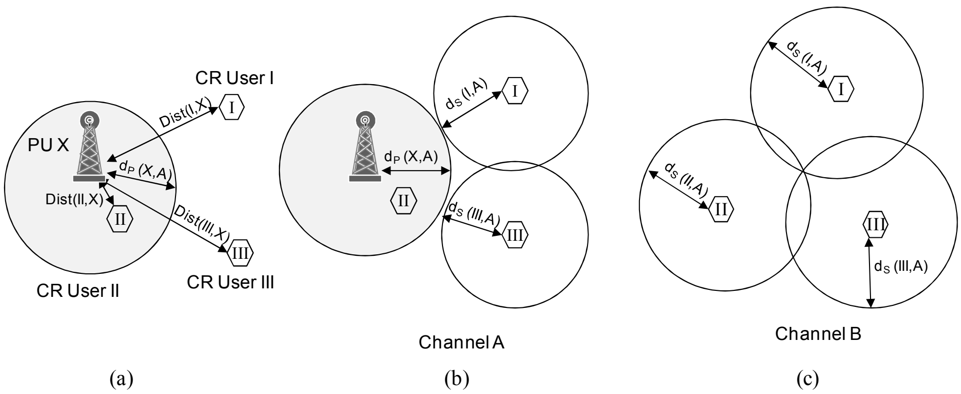


Figure 6.6 Spectrum availability changing with the presence of primary users. (a) Topology (b) availability of channel A; (c) availability of channel B [202].

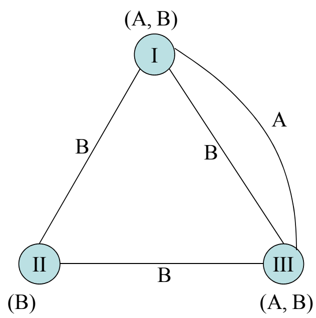


Figure 6.7 CSGC example: (a) example topology with its spectrum availability, and (b) graph model [202].

Figure 6.6 illustrates an example deployment where inactive broadcast (TV) spectrum is utilized to provide wireless connections to a residential community. The broadcast spectrum is divided into two channels (marked by A and B). In this example, broadcast stations (x) are primary users and wireless access points (I, II and III) are CR users. Each primary user x occupies one channel m which is associated with a protection area with radius $dP(x,m)$. Any radiation from CR users falling into it would interfere with the primary user. Each CR user n can adjust its interference range. $dS(n,m)$ by tuning its transmit power on channel m to avoid interfering with primary users.

Figure 6.7 illustrates the reduced CSGC graph that corresponds to the network from Figure 6.6. Channel A is available to CR user I and III, so that in the corresponding CSGC, vertex I and III have A on their color list. Since the transmission areas of I and III on channel A overlap, they can conflict on channel A, and there is a color A edge between I and III. Channel$B$ is available for three users and they all conflict with each other. Hence, B is on each vertex's color list and a color B edge exists between any two vertices.

Overall a conflict graph G can be used to model the network setup of each deployment of primary and CR users, reducing spectrum allocation to a graph coloring problem. This graph theoretic model is generally used jointly with other approaches such as game theoretic, auction-based, and optimization-based approaches Note that CSGC only optimizes color assignment for a fixed topology. If the topology changes (e.g. due to user movement), the graph coloring algorithm needs to be repeated.

6.2.4 Optimization-Based Approach

The optimization approach is generally perceived as a centralized spectrum sharing scheme where the central network entity determines resource allocation to maximize its utility functions. This optimization can be applied to two cases - power allocation and spectrum allocation as follows:

OPTIMIZATION FOR POWER ALLOCATION

The objective of this approach is to find power allocation to maximize the system utility such as total network throughput under power constraints [70]. Assume that M systems, each of which is formed by a single transmitter-receiver pair, coexist in the same area. Under this system, a Gaussian interference channel in discrete time is defined as follow:

$$y_i[n] = \sum_{j=1}^{M} h_{j,i} x_j[n] + z_i[n] \quad i = 1, ..., M \tag{6.5}$$

where $x_i, y_i, z_i \in C$ and the noise processes are i.i.d. over time with $z_i \sim \mathcal{CN}(0, N_0)$.

Here each system treats the received interference as noise. Then the maximum rate that system i can achieve for specific power allocations1 can be determined as follow:

$$R_i = \int_0^W \log\left(1 + \frac{c_{i,i} p_i(f)}{N_0 + \sum_{j \neq i} c_{j,i} p_j(f)}\right) df \tag{6.6}$$

where $p_i(f)$ is the power spectral density of the input signal of system i, and where for convenience we defined $c_{i,j} = |h_{i,j}|^2$. Note that due to the power constraints, $p_i(f)$ must satisfy:

$$\int_0^W p_i(f) df \le P_i \tag{6.7}$$

where P_i is the average power constraint of user i.

Assume that all the parameters are know to all the systems performing the optimization. Then, the achievable rate region, $\mathcal{R}$, is obtained by

$$\begin{aligned}\mathcal{R} = \{\mathbf{R} : R_i = \int_0^W \log\left(1 + \frac{c_{i,i}p_i(f)}{N_0 + \sum_{j\neq i} c_{j,i}p_j(f)}\right)df \text{ and} \\ \int_0^W p_i(f)df \leq P_i \text{ with } p_i(f) \geq 0 \text{ for } i = 1, \dots, M\}\end{aligned} \quad (6.8)$$

where $\mathbf{R} = (R_1, R_2, \dots, R_M)$ and let $\mathcal{R}^*$ be the set of Pareto optimal points of $\mathcal{R}$:

$$\begin{aligned}\mathcal{R}^* =& \{(R_1, R_2, \dots, R_M) \in \mathcal{R} : R_i \geq \tilde{R}_i, \\ &\forall (R_1, \dots, R_{i-1}, \tilde{R}_i, R_{i+1}, \dots, R_M) \in \mathcal{R},\ \text{for } i = 1, \dots, M\}\end{aligned} \quad (6.9)$$

Here, a rate allocation is Pareto optimal (or efficient) if it is not possible to increase the rate of any system without decreasing the rate of some other system.

The spectrum sharing problem is to determine a set of power allocations $p_i(f)$ for the M systems, that maximizes a given global utility function $U(\mathcal{R})$ while satisfying the power constraints. This maximization results in allocations that are fair and efficient in a cooperative scenario, i.e. free from the problem of incentives.

The choice of the utility function will strongly influence the fairness in the resulting allocations. For any utility function that is component-wise monotonically increasing in $(R_1, \dots, R_M)$, the optimal rate allocation must occur in a point of the boundary R.. So it is of interest to obtain a simple characterization for $\mathcal{R}$ and $\mathcal{R}^*$. At first glance, computing $\mathcal{R}$ requires to search over all possible power allocations $p_i(f)$ that satisfy the power constraint. Since $p_i(f)$ are functions with arbitrarily many degrees of freedom, the computation of $\mathcal{R}$ seems to be an infinite dimensional problem. However the following theorem shows that we can restrict attention to piecewise constant power allocations, and as a result, the problem of computing $\mathcal{R}$ has finite dimension.

OPTIMIZATION FOR CHANNEL ALLOCATION

This approach assumes that environmental conditions such as user location and available spectrum are static during the time it takes to perform spectrum assignment. This corresponds to a slow varying spectrum environment where users quickly adapt to environmental changes by re-performing network-wide spectrum allocation. Therefore, this scheme focus on a model for a fixed topology.

Assume that a network of N CR users indexed from 0 to $N-1$ competing for M spectrum channels indexed 0 to $M-1$. Each CR user can be a transmission link or a broadcast access point. The channel availability and rewards for each CR user can be calculated based on the location and channel usage of nearby primary users. The key components of this model are defined as follows [202]:

- *Channel availability:* $L = \{l_{n,m} | l_{n,m} \in \{0,1\}\}_{NM}$ is a N by M binary matrix representing the *channel availability:* $l_{n,m} = 1$ if and only if channel m is available at user n.
- *Channel reward:* $B = \{b_{n,m}\}_{NM}$, a N by M matrix representing the *channel reward:* $b_{n,m}$ represents the maximum bandwidth/throughput that can be acquired

(assuming no interference from neighbors) by user n using channel m. Then, the reward can be the coverage of a CR user using a channel:

$$b_{n,m} = d_S(n,m)^2, d_{\min} \le d_S(n,m) \le d_{\max} \tag{6.10}$$

where $d_S(n,m)$ is the transmission range of CR user n at channel m.

Or the reward can be the capacity of channel m:

$$b_{n,m} = \log(1 + f(d_S(n,m)), d_{\min} \le d_S(n,m) \le d_{\max} \tag{6.11}$$

Here the signal to noise ratio (SNR) is assumed to be a function of $d_S(n,m)$). Obviously, $b_{n,m} = 0$ if $l_{n,m} = 0$.

- *Interference constraint:* Let $\mathcal{C} = \{c_{n,k,m} | c_{n,k,m} \in \{0,1\}\}_{NNM}$, a N by N by M matrix, represents the interference constraints among CR users. If $c_{n,k,m} = 1$, users n and k would interfere with each other if they use channel m simultaneously. The constraint depends on channel availability, i.e., $c_{n,k,m} \le l_{n,m} \cdot l_{k,m}$ and $c_{n,n,m} = 1 - l_{n,m}$.

- *Conflict free channel assignment:* $A = \{a_{n,m} | a_{n,m} \in \{0,1\}, a_{n,m} \le l_{n,m}\}_{NM}$ is a N by M binary matrix that represents the assignment: $a_{n,m} = 1$ if channel m is assigned to user n. A conflict free assignment needs to satisfy all the

$$a_{n,m} + a_{k,m} \le 1, \text{ if } c_{n,k,m} = 1, \forall n, k < N, m < M \tag{6.12}$$

Let $\Lambda(L,C)_{N,M}$ denote the set of conflict free spectrum assignments for a given set of N users and M spectrum bands and constraints C.

- *Radio interface limit:* $C_{\max}$ represents the maximum number of channels that can be assigned to a CR user. The assignment at each user n needs to satisfy $\sum_{m=0}^{M-1} a_{n,m} \le C_{\max}$. represents the reward vector that each user gets for a given channel assignment.

- *User reward:* $F = \{\beta_n = \sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m}\}_{N \times 1}$ represents the reward vector that each user gets for a given channel assignment.

- *Network utilization:* The channel allocation is to maximize network utility function $U(\mathcal{F})$.

Given the model above, the spectrum assignment problem can be defined by the following optimization function:

$$A^* = \underset{A \in \Lambda(L,C)_{N,M}}{\arg} \max U(\mathcal{F}) \tag{6.13}$$

A variety of utility functions can be used in the channel allocation problem, some of which are explained as follow:

- *Max-Sum-Reward:* This maximizes the total spectrum utilization in the system regardless of fairness. The optimization problem is expressed as:

$$U_{\text{sum}} = \sum_{n=0}^{N-1} \beta_n = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m} \tag{6.14}$$

- *Max-Min-Reward:* This maximizes the spectrum utilization at the bottleneck user, or the user with the least allotted spectrum. The optimization problem is expressed as:

$$U_{\min} = \min_{0 \le n \le N} \beta_n = \min_{0 \le n \le N} \sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m} \tag{6.15}$$

  Roughly, Max-Min-Reward driven allocation gives the most poorly treated user (i.e. the user who receives the lowest reward) the largest possible share, while not wasting any network resources. This is the simplest notion of fairness.

- *Max-Proportional-Fair:* Consistent with prior work [118] [167] [189] [228], we consider and address fairness for single-hop flows. The corresponding fairness-driven utility optimization problem is expressed as:

$$u_{\text{fair}} = \sum_{n=0}^{N-1} \log \beta_n = \sum_{n=0}^{N-1} \log(\sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m}) \tag{6.16}$$

  The essence of proportional fair is that if for any other feasible assignment A' and the associated β'_n, the aggregate of proportional changes in user reward is zero or negative:

$$\sum_{n=0}^{N-1} \log \frac{\beta'_n - \beta_n}{\beta_n} \le 0 \tag{6.17}$$

  To make it comparable to $U_{\min}$ and U_{sum}, we modify the fairness utility to

$$u_{\text{fair}} = (\prod_{n=0}^{N-1} \log \beta_n)^{\frac{1}{N}} = (\prod_{n=0}^{N-1} \log \sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m})^{\frac{1}{N}} \tag{6.18}$$

Note that under the same assignment, $1/N \cdot U_{\text{sum}} \ge U_{\text{fair}} \ge U_{\min}$.

This approach is generally based on the graph theoretic model that is explained in Section 6.2.3. A GMS problem is to color each vertex using a number of colors from its color list, and find the color assignment that maximizes system utility. The coloring is constrained by that if an edge exists between any two distinct vertices, they can't be colored with the same color. Most importantly, the objective of coloring is to maximize system utility. This is different from traditional graph color solutions that assign one color per vertex. Notice that the solution to this graph coloring problem is to maximize system utility for a given graph, i.e. a given topology and channel availability. This characterizes the optimal solution for a static environment.

The optimal coloring problem is known to be NP-hard [94]. Efficient algorithms to optimize spectrum allocation for a given network topology exist. In [297], the authors presented a set of sequential heuristic based approaches that produce good coloring solutions. The algorithm starts from empty color assignment and iteratively assign colors to vertices to approximate the optimal assignment. In each stage, the algorithm labels all the vertices with a non-empty color list according to some policy-defined labeling. The algorithm picks the vertex with the highest valued label and assigns the color associated with the label to the vertex. The algorithm then deletes the color from the vertex's color list, and from the color lists of the constrained neighbors. The color list and the interference constraint of a vertex keep on changing as other vertices are processed, and the labels of the colored vertex and its neighbor vertices are modified according to the new graph. The algorithm can be implemented using a centralized controller who observes global topology and makes decisions, or through a distributed algorithm where each vertex performs a distributed voting process. Results in [202] [297] show that the heuristic based algorithms perform similarly to the global optimum (derived off-line for simple topologies), and the centralized and distributed algorithms perform similarly.

6.3 INTRA-NETWORK SPECTRUM SHARING

As mentioned in Section 6.1, spectrum sharing can be classified in many aspects. In the next two sections, we primarily focus on sharing scope, and divide existing spectrum sharing techniques into intra-network and inter-network spectrum sharing. A significant amount of work on spectrum sharing focuses on intra-network spectrum sharing, where the users of a CR network try to access the available spectrum without causing interference to the primary users. In this section, we overview the existing work and the proposed solutions in this area while providing a classification of existing solutions in terms of the classification provided in Section 6.1.

6.3.1 Centralized Spectrum Sharing

In the centralized spectrum sharing, resource allocation procedures are controlled by a central network entity, such as a base-station. Here, the central network entity is responsible for identifying spectrum opportunities and allocating them to its users while satisfying a target performance specified by its policy. Since the resource allocation is determined based on the local observation and resource requests collected from CR users, cooperative behaviors are natural in the centralized spectrum sensing. In this subsection, we show how radio resources are managed within the CR network in a centralized manner.

GRAPH-COLORING OPTIMIZATION

In [202], a graph coloring based collaborative spectrum access scheme is proposed, where a topology-optimized allocation algorithm is used. However, whenever the network topology changes according to the node mobility, the network needs to completely recompute spectrum assignments for all users, resulting in a high computational and communication overhead. This scheme is based on the exclusive allocation, and can be used for both licensed and open spectrum sharing. Furthermore, it supports both cooperative and noncooperative spectrum sharing as well.

This method is basically originated from a progressive minimum neighbor first (PMNF) scheme that is proposed as a sequential heuristic solution to graph coloring for generalized channel assignment [213]. This algorithm 1) assigns each vertex a unique label, 2) colors the vertex with the highest label with the lowest indexed color without violating the constraints, 3) removes the colored vertex and associated edges from the graph, and then 4) repeats until all the vertices are colored. The objective of PMNF is to minimize the total number of colors required to assigned to each vertex, and hence the basic idea of the algorithm is to color the “most difficult” vertices first. The worst case performance of PMNF significantly outperforms other heuristic approaches.

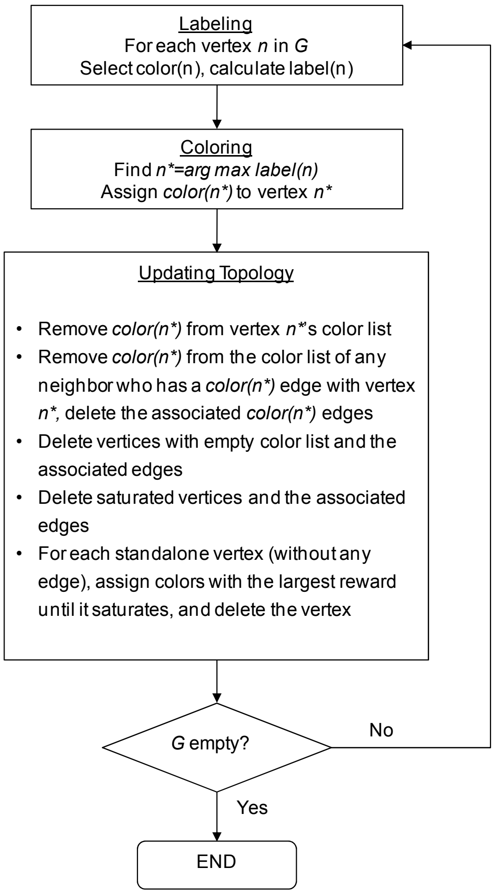


Figure 6.8 Flow chart of coloring [202].

This conventional scheme can be modified to support conflict constraints of channel allocation in a defined geographic region, as shown in Figure 6.8. The procedure for this algorithm is as follow:

1. A vertex is considered to be "saturated" if its channel capacity has reached $C\max$. In each stage, the algorithm labels all the non-saturated vertices with a nonempty color list according to the pre-defined labeling rule.

2. The algorithm chooses the vertex with the highest valued label and assigns the color associated with the label to the vertex.

3. The algorithm removes the color from the vertex's color list, and also from the color lists of its neighbors. It also deletes all the edges of the assigned color from the the vertex in the color graph.

4. The algorithm repeats above procedures until every vertex's color list becomes empty or every vertex saturates.

Note that our graph coloring problem aims at maximizing utility while the conventional graph coloring problem [213] focuses on minimizing the number of colors used. Unlike

PMNF, this scheme chooses to color the "most valuable" vertices first, i.e. the vertices that contribute to the system utility the most.

Based on channel allocation model explained in Sections 6.2.3 and 6.2.4, the labeling rule can be designed as follow: First, for each vertex n, its m color-specific degree, $D_{n,m}$, is the number of conflict edges that user n shares with its neighbors for channel m. This is the number of neighbors who cannot simultaneously use channel m with user n, i.e.:

$$D_{n,m} = \sum_{k=0,k\neq n}^{N-1} c(n,k,m) \cdot l_{n,m} \cdot l_{k,m} \tag{6.19}$$

$D_{n,m}$ is a good measure of the impact to neighbors when a color is assigned to a vertex. The following are the relevant labeling values corresponding to each of the utility functions described in Section 6.2.4.

- **Max sum reward**

1. *Collaborative-Max-Sum-Reward (CSUM):* This rule aims to maximize the sum reward defined in Eq. (6.15). When a vertex n is assigned with a color m, its contribution to the sum reward in a local neighborhood can be computed as $b_{n,m}/(D_{n,m}+1)$ since some of its neighbors cannot use this color. In this scheme, the vertex n is labeled according to

$$\begin{aligned} label_n &= \max_{m\in l_n} \frac{b_{n,m}}{D_{n,m}+1} \\ color_n &= \arg undersetm \in l_n \max \frac{b_{n,m}}{D_{n,m}+1} \end{aligned} \tag{6.20}$$

where D_n represents the color list available at vertex n at this assignment stage. This rule considers the tradeoff between spectrum utilization (in terms of selecting the color with the largest reward) and interference to neighbors (in terms of degree). This rule is collaborative, since it takes into account the impact to neighbors.

2. *Non-collaborative-Max-Sum-Reward (NSUM):* This rule aims to improve the sum of reward without considering the impact of interference to neighbors. The vertex with the maximum reward will be colored, i.e. a vertex n is labeled with

$$\begin{aligned} label_n &= \max_{m\in l_n} b_{n,m} \\ color_n &= \arg \max_{m\in l_n} b_{n,m} \end{aligned} \tag{6.21}$$

Compared to CSUM, this rule is relatively selfish. It is non-collaborative, since each vertex only considers its own reward and ignores impact on the overall system. Max min reward

3. *Collaborative-Max-Min-Reward (CMIN):* This rule tries to distribute colors uniformly among vertices to improve the minimum reward that a vertex can get, while considering interference to neighbors. This rule tries to solve Max-Min optimization as defined in Eq. (6.16). In each stage, a vertex n is labeled

according to

$$\begin{aligned} label_n &= -\sum_{m=0}^{N-1} a_{n,m} \cdot b_{n,m} \\ color_n &= \arg\max_{m \in l_n} \frac{b_{n,m}}{D_{n,m}+1} \end{aligned} \tag{6.22}$$

where $a_{n,m}$ represents the reward obtained at n before this assignment stage. Note that unlike CSUM and NSUM, the label depends on the reward obtained in previous stages. In each stage, the vertex with the minimum accumulated reward will be colored with the color that maximizes utilization while considering interference. If two vertices have the same label, then the vertex with larger $\max_{m \in l_n} b_{n,m}/(D_{n,m}+1)$ value gets a higher label.

4. *Non-collaborative-Max-Min-Reward (NMIN):* This rule is a non-collaborative version of CMIN where the impact of interference is not considered in the vertex labeling and coloring, i.e.,

$$\begin{aligned} label_n &= -\sum_{m=0}^{N-1} a_{n,m} - b_{n,m} \\ color_n &= \arg\max_{m \in l_n} b_{n,m} \end{aligned} \tag{6.23}$$

In each stage, the vertex with the minimum accumulated reward will be colored with the color that has the largest reward. If two vertices have the same label, then the vertex with larger $\max_{m \in l_n} b_{n,m}$ is assigned with a higher label.

- **Max proportional fair:**

1. *Collaborative-Max-Proportional-Fair (CFAIR):* This rule aims to achieve a specific fairness among vertices, corresponding to Eq. (6.18). It is well known that proportional fair scheduling assigns resource to the user with the highest $r_n/\hat{R}_n$, where rn represents the reward generated by using a time slot and $\hat{R}_n$ is the average reward that the user n has received in the past [15] [257]. The concept of proportional fair scheduling is applied to this problem by viewing colors as time slots. In each stage, each vertex n is labeled according to

$$\begin{aligned} label_n &= \frac{\underset{m \in l_n}{b_{n,m}/(D_{n,m}+1)}}{\sum_{m=0}^{M-1} a_{n,m} \cdot b_{n,m}} \\ color_n &= \arg\max_{m \in l_n} b_{n,m}/(D_{n,m}+1) \end{aligned} \tag{6.24}$$

where $label_n$ represents the ratio of the maximum interference-weighted reward from using a color and the accumulated reward in past stages. This rule is in general different from the traditional proportional fair rule as it captures the difference in the impact of interference generated by a color assignment.

2. *Non-collaborative-Max-Proportional-Fair (NFAIR):* This is a non-collaborative version of the CFAIR rule. Each vertex n is labeled according to

$$\begin{aligned} label_n &= \frac{\underset{m\in l_n}{b}{}_{n,m}}{\sum_{m=0}^{M-1} a_{n,m}\cdot b_{n,m}} \\ color_n &= \arg\max_{m\in l_n} b_{n,m} \end{aligned} \tag{6.25}$$

When all the channels have uniformed bandwidth, i.e. $b_{n,m} = 1$, this rule becomes NMIN rule.

The algorithm described above assumes a central allocation server with knowledge about all users in the system. In a centralized architecture, a central spectrum server makes decisions on channel assignment. The server collects location, power, spectrum and interference information from both primary and CR users, and runs the assignment algorithm to distribute channels among CR users. It then broadcasts the assignments on a predefined channel. CR users listen to the broadcast and communicate using their assigned channels.

While a central server can optimize across network-wide information, there are two serious limitations to this approach. First, this approach requires a communication path between the spectrum server and all CR users, i.e. all users need interference-free access to a pre-assigned dedicated control channel, possibly in a licensed band. In addition, as networks grow in density, a pre-defined control channel will limit the bandwidth available for control messages. Second, the server processing complexity will scale at least polynomially with the number of devices. Any central spectrum server will quickly become a computational bottleneck.

SPECTRUM SEVER-BASED SCHEDULING

In this scheme, a spectrum server allocates a schedule for a set of links operating in non-overlapping frequency range [212]. This method mainly focuses on channel allocation, and is designed for open spectrum and common use sharing techniques.

Consider a wireless network with N nodes forming L logical links. The network can be represented as a directed graph $G(V, E)$, where the nodes in the network are represented by the set of vertices V of the graph and the links are represented by a set of directed edges E. A directed edge from a node m to node n implies that n wishes to communicate data to node m. The spectrum server coordinates the activity of the set of L links to share the spectrum efficiently. Define the set of transmission modes $\mathcal{T} = \{1, 2, \ldots, M\}$, where M denotes the number of possible transmission modes. Then the mode activity vector t_i of mode i is a binary vector, indicating the on-off activity of the links. If $t_i = (t_{1i}, t_{2i}, \ldots, t_{Li})$ is a mode activity vector, then

$$t_{li} = \begin{cases} 1 & \text{link } l \text{ is active undertransmission mode } i \\ 0 & \text{otherwise} \end{cases} \tag{6.26}$$

Note that there are M possible transmission modes including the mode in which all links are off.

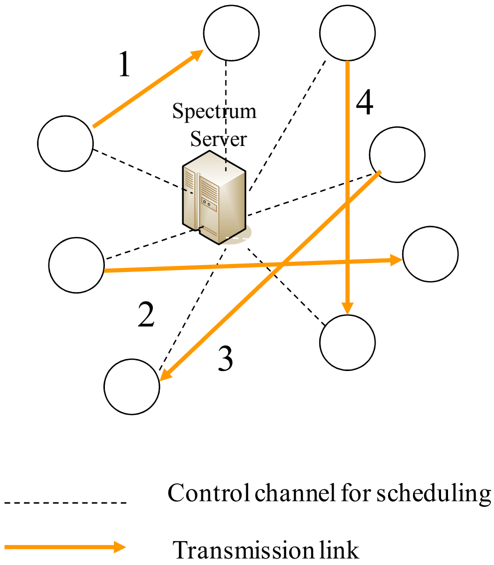


Figure 6.9 System architecture for spectrum sever-based scheduling.

Figure 6.9 shows a representative network with 4 links. Then the mode activity matrix $T = \{t_{li} | l = 1, \dots, L, i = 1, \dots, M\}$ is expressed as follow:

$$T = \begin{bmatrix} 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1 \\ 0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1 \\ 0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1 \end{bmatrix}. \tag{6.27}$$

If links 1 and 2 are active, and links 3 and 4 are off, then the transmission mode is $[1100]$.

Let the transmitter power on a link l be P_l. If G_{lk} is the link gain from the transmitter of link k to the receiver of link l and σ_l^2 is the noise power at the receiver of link l, the SIR γ_{li} at the receiver of link l in transmission mode i is given by

$$\gamma_{li} = \frac{t_{li} G_{ll} P_l}{\sum_{k \in E, k \neq l} t_{ki} G_{lk} P_k + \sigma_l^2} \tag{6.28}$$

The link gain between a transmitter and receiver takes into account the path loss and attenuation due to shadow fading. The link gains between each transmitter and receiver are known to the spectrum server. The data rate in each link l at mode i, c_{li}, depends on the signal-to-interference ratio (SIR) in that link, γ_{li}, which is obtained as $\log(1 + \gamma_{li})$ Furthermore, the transmitter is assumed to be able to vary its data rate, possibly through a combination of adaptive modulation and coding.

Let x_i be the fraction of time that transmission mode i is active and r_l be the average data rate of link l. Each link has a minimum average data rate requirement $r_l^{\min}$. The average data rate in link l is the time average of the data rates of all the transmission modes that include link l. Thus,

$$r_l = \sum_i c_{li} x_i, \tag{6.29}$$

or in vector form,

$$\mathbf{r} = \mathbf{C}\mathbf{x} \tag{6.30}$$

where $C = [c_{li}]$ is an $L \times M$ matrix with non-negative entries, such that column i indicates the rate obtained by each link in mode i.

The proposed scheme considers transmitters with a fixed power on-off modulation and devises schedules that maximize the system throughput. The optimization problem, subject to minimum rate constraints in the individual links, is posed as a linear integer program. If the link gains are known to the spectrum server, it can schedule the transmissions among the links to maximize the system throughput. It is shown that when there is no minimum rate constraint, a fixed set of links (called the dominant mode) which maximizes the sum rate is operated all the time. In order to offset the inherent unfairness in the above solution, this method introduces a minimum rate constraint and characterize the resulting loss in sum rate when compared to the case when there is no minimum rate constraint. The method also employs alternate fairness criteria by designing scheduling algorithms that achieve max-min fairness and proportional fairness. The max-min fair rate allocation can be obtained in one step by solving a linear program which maximizes the minimum common rate among the links. The proportional fair schedule is obtained by solving a non-linear convex optimization program. Each optimization problem is described as follow:

- *Maximum Sum Rate Scheduling:* This scheduling scheme focuses on maximizing the sum of the average data rates over all links $l = 1, 2, \ldots, L$, subject to constraints on the minimum rate for each link. The optimization problem can be posed as the linear program (LP):

$$\begin{aligned} \text{maximize}: \ & \mathbf{1}^T\mathbf{r} \\ \text{subject to}: \ & \mathbf{r} = \mathbf{C}\mathbf{x} \\ & \mathbf{r} > \mathbf{r}_{\min} \\ & \mathbf{1}^T\mathbf{x} = 1 \\ & \mathbf{x} \geq 0 \end{aligned} \tag{6.31}$$

  where $\mathbf{r}_{\min}$ is the set of the minimum average data rate requirement of each link.

- *Max-Min Fair Rate Scheduling:* The maximum sum rate scheduling is biased towards links that have the best quality (i.e., least interference) and is unfair to the other links that are not a part of the dominant transmission mode.

$$\begin{aligned} \text{maximize}: \ & \mathbf{r}_{\min} \\ \text{subject to}: \ & \mathbf{r} = \mathbf{C}\mathbf{x} \\ & \mathbf{r} \geq \mathbf{r}_{\min}\mathbf{1} \\ & \mathbf{1}^T\mathbf{x} = 1 \\ & \mathbf{x} \geq 0 \end{aligned} \tag{6.32}$$

- *Proportional Fair Scheduling:* While the max-min fair schedule leads to global fairness, the proportional fair criterion focuses on the fairness of individual links. The proportionally fair vector is the one that maximizes the sum of logarithms of the utility functions. Hence, the proportional fair rates is obtained by the following non-linear optimization problem with linear constraints:

$$\begin{aligned}\text{maximize}: \ & \sum_l \log r_l \\ \text{subject to}: \ & \mathbf{r} = \mathbf{C}\mathbf{x} \\ & \mathbf{1}^T\mathbf{x} = 1 \\ & \mathbf{x} \geq 0\end{aligned} \tag{6.33}$$

AUCTION-BASED INTRA-NETWORK SPECTRUM SHARING

This scheme proposes auction mechanisms for sharing spectrum among a group of users, subject to a constraint on the interference temperature at a measurement point [117]. The users access the channel using spread spectrum signaling and so interfere with each other. Each user receives a utility that is a function of the received signal-to-interference plus noise ratio (SINR). This scheme mainly focuses on power allocation based on open spectrum sharing.

In this method, two auction mechanisms are proposed for allocating the received power. The first is an auction where users are charged for received SINR, combined with logarithmic utilities. This auction leads to a weighted max-min fair SINR allocation. The second is an auction for power, which maximizes the total utility when the bandwidth is large enough and the receivers are co-located.

[System model]

Spectrum with bandwidth B is to be shared among M spread spectrum users, where a user refers to a transmitter and an intended receiver pair. User i's valuation of the spectrum is characterized by a utility $U_i(\gamma_i)$, where γ_i is the received SINR at user i's receiver. They primarily consider the case where each user's utility is given by $U_i(\gamma_i) = U_i(\theta_i, \gamma_i))$, where θ_i is a user-dependent parameter. As a particular example, the logarithmic utility $U_i(\gamma_i) = \theta_i \ln(\gamma_i)$ is considered. For each user i, $U_i(\gamma_i)$ is assumed to be increasing, strictly concave, and twice continuously differentiable in γ_i . Utilities that satisfy this assumption are commonly used to model elastic data applications [236]. For each i, the received SINR is given by

$$\gamma_i = \frac{p_i h_{ii}}{n_0 + \frac{1}{B}(\sum_{j\neq i} p_j h_{ji})} \tag{6.34}$$

where p_i is user i's transmission power, h_{ij} is the channel gain from user i's transmitter to user j's receiver, and n_0 is the background noise power that is assumed to be the same for all users. To satisfy an interference temperature constraint, the total received power at a specified measurement point must satisfy

$$\sum_{i=1}^{M} p_i h_{i0} \leq P \tag{6.35}$$

where h_{i0} is the channel gain from user i's transmitter to the measurement point. The system model is shown in Figure 6.10. A power allocation is Pareto optimal if no user's utility can be increased without decreasing another user's utility.

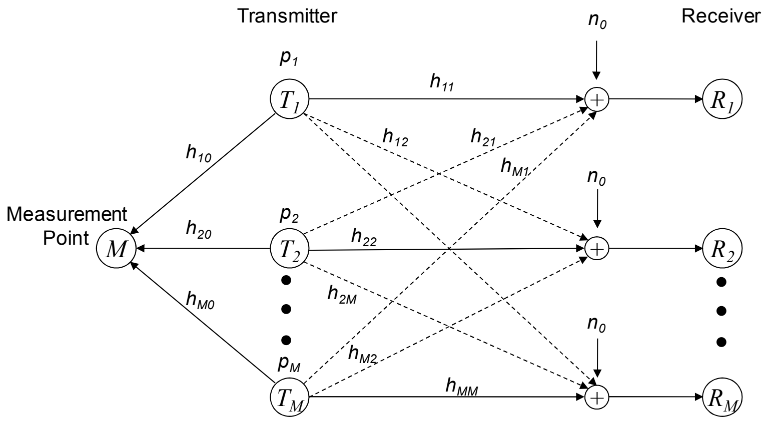


Figure 6.10 System architecture for M transmitter-receiver pairs [117].

A special case is when the receivers are co-located with the measurement point. This could model a situation where a service provider purchases the spectrum usage rights from the auctioneer and provides service from a single access point. In this case, $h_{ij} = h_{i0}$ for all $i, j \in \{1, \ldots, M\}$, and user i's received power is defined as $p_i^r = p_i h_{i0}$. In a Pareto optimal allocation for this co-located receiver case, SINR for each user i is expressed as follow:

$$\gamma_i = \gamma_i(p_i^r) = \frac{p_i^r}{n_0 + \frac{1}{B}(P - p_i^r)} \tag{6.36}$$

Here user i's utility $U_i(\gamma_i(p_i^r))$ under a Pareto optimal allocation does not depend on how the power is allocated among the interferers.

[One-dimensional auctions with pricing]

Each user's utility is assumed to be private information, i.e., only known to the user himself. The auctioneer must then design a mechanism for allocating power without having this knowledge a priori. Also the auctioneer may not have a priori knowledge of the channel gains, h_{ij}'s. One such mechanism is the VCG auction, which is explained in Section 6.2.2.

However, the VCG auction may not be suitable in this context for several reasons: (i) In order to completely specify the users' utilities, in particular, the SINR in (1), for each user i, the channel gains h_{ij} for all $i, j \in \{1, ..., M\}$ must be measured by the users and reported to the auctioneer. This might be a heavy burden for the users in a large network. (ii) The auctioneer must solve $M + 1$ optimization problems, which are typically non-convex due to the interference. This becomes computationally expensive for large M, and may not be suitable for online allocations. For these reasons, the auction mechanisms require less information exchange and less computation for the auctioneer.

In the proposed SINR-and power-based auctions, users submit one-dimensional bids representing their willingness to pay, and the auctioneer simply allocates the received power in proportion to the bids. The users then pay an amount proportional to their SINR (or power). The auctioneer announces a nonnegative reserve bid β, and uses a corresponding reserve power that interferes with the other users [117].

Regarding the information structure of the auction, this scheme assumes that it is a complete information game, i.e., all users' utilities and all channel gains are known to all users. The simultaneous auction algorithm is specified as follow:

1. The auctioneer announces a reserve bid $\beta \geq 0$, and a price $\pi^s > 0$ (in an SINR auction) or $\pi^p > 0$ (in a power auction).

2. After observing β, π^s (or π^p), user $i \in 1, \ldots, M$ submits a bid $bi \geq 0$.

3. The auctioneer keeps reserve power p_0, and allocates to each user i a transmission power p_i so that the received power at the measurement point is proportional to the bids, i.e.,

$$\begin{aligned} p_i h_{i0} &= \frac{b_i}{\sum_{j=1}^{M} b_i + \beta} P \\ p_0 &= \frac{\beta}{\sum_{j=1}^{M} b_i + \beta} P \end{aligned} \tag{6.37}$$

   The resulting SINR for user i is

$$\gamma_i = \frac{p_i h_{ii}}{n_0 + \frac{1}{B}\left(\sum_{j \neq i} p_j h_{ji} + p_0 h_{0i}\right)} \tag{6.38}$$

   where h_{0i} is the channel gain from the auctioneer (measurement point) to user i's receiver. If $\sum_{i=1}^{M} b_i + \beta = 0$, then $p_i = 0$.

4. In an SINR (power) auction, user i pays $C_i = \pi^s \gamma_i$ ($C_i = \pi^p p_i h_{i0}$).

A bidding profile is the vector containing the users' bids $b = (b_1, \ldots, b_M)$. The bidding profile of user i's opponents is defined as $b_{-i} = (b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_M)$, so that $b = (bi; b_{-i})$. In the preceding auctions, each user i submits a bid bi to maximize his surplus function

$$S_i(b_i; b_{-i}) = U_i(\gamma_i(b_i; b_{-i})) - C_i. \tag{6.39}$$

An NE of the auction is associated with a bidding profile b. such that $S_i(b_i^*; b_{-i}^*) \geq S_i(b_i'; b_{-i}^*)$ for any $b_i' \in [0, \infty)$ and any user i. Define user i's best response given b_{-i} as the set

$$\mathcal{B}_i(b_{-i}) = \{\hat{b}_i | \hat{b}_i = \arg_{b_i \in [0,\infty)} \max S_i(b_i; b_{-i})\} \tag{6.40}$$

i.e., the set of b_i's that maximize $S_i(b_i; b_{-i})$ given a fixed b_{-i}. The NE bidding profile b^* is a fixed point, i.e., no user has the incentive to deviate unilaterally. The existence and uniqueness of an NE are shown in the following to depend on β and π^s (or π^p).

These auction mechanisms differ from some previously proposed auction-based network resource allocation schemes (e.g., [135] [173]) in that the bids here are not the same as the payments. Instead, the bids are signals of willingness to pay. The auctioneer can therefore influence the NE by choosing] and π^s (or π^p). This alleviates the typical inefficiency of the NE, and allows us to reach Pareto optimal, and in some cases, socially optimal solutions.

BELIEF-ASSISTED PRICING GAME BASED ON DOUBLE AUCTION RULE

This scheme considers the spectrum allocation in wireless networks with multiple selfish legacy spectrum holders and unlicensed users as multi-stage dynamic games [129]. A belief-assisted dynamic pricing approach is proposed to optimize overall spectrum efficiency while keeping the participating incentives of the users based on double auction rules. Unlike the auction scheme mentioned above [117], this scheme is related to exclusive channel allocation over the licensed spectrum bands, and hence can be classified as exclusive allocation and the licensed spectrum sharing.

[System Model]
This method considers the wireless networks where multiple primary users and CR users operate simultaneously. Every primary user has the license of using a certain spectrum range, which can be divided into non-overlapping orthogonal channels.

In this system model, all users are assumed to be selfish and rational, that is, their objectives are to maximize their own payoffs, not to cause damage to other users. However, users are allowed to cheat whenever they believe cheating behaviors can help them to increase their payoffs. With regard to CR users, in order to have the rewards of achieving certain communication goals, they want to utilize more spectrum resources. The selfishness of both primary and CR users will prevent them from revealing their private information such as acquisition costs or reward payoffs, which makes traditional spectrum allocation approaches not applicable under this scenario.

Therefore, novel spectrum allocation approaches need to be developed, which not only optimize the spectrum efficiency but also extract the private information from the selfish parties through certain mechanisms to assist the optimization of spectrum allocation. Specifically, this method considers the collection of the available spectrums from all primary users as a spectrum pool, which totally consists of N non-overlapping channels. Assume there are J primary users and K CR users, indicated by the set $P = \{p_1, p_2, \ldots, p_J\}$ and $S = \{s_1, s_2, \ldots, s_K\}$, respectively. The channels authorized to primary user p_i is represented by $A_i = \{a_i^j\}, j \in \{1, 2, \ldots, n_i\}$, where a_i^j represents the channel index in the spectrum pool and n_i is the total number of channels that belong to user p_i. Define A as the set of all the channels in the spectrum pool. Moreover, denote the acquisition costs of user p_i's channels as the vector $C_i = \{c_i^{a_j}\}_{j\in\{1,2,\ldots,n_i\}}$, where the jth element represents the acquisition cost of the jth channel in A_i. A payoff vector of CR user s_i is defined as $V_i = \{v_i^j\}_{j\in\{1,2,\ldots,N\}}$, where the jth element is the reward payoff if this user successfully leases the jth channel in the spectrum pool.

[Pricing Game Model]
This method models the dynamic spectrum allocation problem as a pricing game to study the interactions among the players, i.e., the primary and CR users. If primary user p_i reaches agreements of leasing all or part of her/his channels to CR users, the payoff function of this primary user can be written as follows.

$$U_{p_i}(\phi_{A_i}, \alpha_i^{A_i}) = \sum_{j=1}^{n_i} (\phi_{a_i^j} - c_i^{a_j}) \alpha_i^{a_i^j} \tag{6.41}$$

where $\phi_{A_i} = \{\phi_{a_i^j}\}_{j\in\{1,2,\ldots,n_i\}}$ is the payment vector that user p_i obtains from the CR user by leasing the channel a_i^j in the spectrum pool. Note that $\alpha_i^{A_i} = \{\alpha_i^{a_i^j}\}_{j\in\{1,2,\ldots,n_i\}}$

and $\alpha_i^{a_i^j} \in \{0,1\}$, which indicates if the jth channel of user p_i has been allocated to a CR user or not. Similarly, the payoff function of CR user s_i can be modeled as follows.

$$U_{s_i}(\phi_A, \beta_i^A) = \sum_{j=1}^{N} (v_i^j - \phi_j)\beta_i^j \tag{6.42}$$

where $\phi_A = \{\phi_j\}_{j \in \{1,2,\ldots,N\}}$, $\beta_i^A = \{\beta_i^j\}_{j \in \{1,2,\ldots,N\}}$. Note that $\beta_i^j \in \{0,1\}$ illustrates if CR user s_i successfully leases the jth channel in the spectrum pool or not. Hence, the strategies of the primary users and CR users are actually defined by $\alpha_i^{A_i}$ and β_i^A, respectively.

Since the players may have conflict interests with each other, this dynamic spectrum sharing game can be modeled as a multi-stage non-cooperation game. To be specific, from the primary users' point of view, they want to earn the payments by leasing the unused channels, which not only cover their spectrum acquisition costs but also gain as much extra payments as possible; from the CR users' point of view, they aim to accomplish their communication goals by providing the least possible payments to lease the channels; while from the network designers' point of view, they attempt to maximize the network performance. Therefore, the spectrum users involved in the spectrum sharing process construct a non-cooperative pricing game [86] [197].

In the spectrum allocation pricing game, the primary users can be viewed as the principles, who attempts to sell the unused channels to the CR users. The CR users are the bidders who compete with each other to buy the permission of using primary users' channels, by which they may gain extra payoffs for future use. In the proposed pricing game, multiple sellers and buyers coexist, which indicates the double auction scenario. It means that not only the CR users but also the primary users need to compete with each other to make the beneficial transactions possible by eliciting their willingness of the payments in the forms of bids or asks. The most important property of double auction mechanism is its high efficiency. Moreover, it can respond quickly to changing conditions of auction participants. However, in order to achieve the full efficiency of the double auction mechanism, a lot of messages need to be exchanged among the auction participants, which can be easily implemented by powerful central authorities Therefore, the double auction should aim to develop an efficient pricing approach for spectrum allocation, which uses simple message exchanges to quickly and accurately coordinate the spectrum sharing.

[Belief-Assisted Dynamic Pricing For Efficient Spectrum Allocation]
Assume that the available channels from the primary users are leased for usage of certain time period T. Also, the cost of the primary users and reward payoffs of the CR users are assumed to remain unchanged over this period. Before this spectrum sharing period, the auction mechanism has a trading period τ , within which the users exchange their information of bids and asks to achieve agreements of spectrum usage. The time period $T + \tau$ is considered as one stage in our pricing game. Accordingly, the interactions of the players in static pricing games are described as follow: The users' goals are to maximize their own payoff functions. As for the primary users, the optimization problem can be

written as:

$$\begin{aligned} O(p_i) = & \max_{\phi_{A_i}, \alpha_i^{A_i}} U_{p_i}(\phi_{A_i}, \alpha_i^{A_i}), \ \forall i \in \{1, 2, \ldots, J\} \\ & \text{subject to :} \\ & U_{\hat{s}_{a_i^j}}(\{\phi_{-a_i^j}, \phi_{a^j+i}\}), \beta_i^A) \geq U_{\hat{s}_{a_i^j}}(\{\phi_{-a_i^j}, \tilde{\phi}_{a_i^j}\}), \beta_i^A), \\ & \hat{s}_{a_i^j} \neq 0, a_i^j \in A_i \end{aligned} \tag{6.43}$$

where $\tilde{\phi}_{a_i^j}$ is any feasible payment and $\phi_{-a_i^j}$ is the payment vector excluding the element of the payment for the channel a_i^j. Note that $\hat{s}_{a_i^j}$ is defined as follows.

$$\hat{s}_{a_i^j} = \begin{cases} s_k & \text{if} \beta_k^{a_i^j} = 1, \\ 0 & \text{if} \beta_k^{a_i^j} = 0, \forall k \in \{1, 2, \ldots, K\}. \end{cases} \tag{6.44}$$

Thus, Eq. (6.43) is the incentive compatible constraint [147]. It means that the CR users have incentives to provide the optimal payment because they cannot have extra gains by cheating on the primary users. Similarly, the optimization problem can be written for the CR users as follows.

$$\begin{aligned} O(s_i) = & \max_{\phi_A, \beta_i^A} U_{s_i}(\phi_A, \beta_i^A), \ \forall i \in \{1, 2, \ldots, K\} \\ & \text{subject to :} \\ & U_{\hat{p}_j}(\{\phi_{-j}, \phi_j\}), \beta_i^A) \geq U_{\hat{p}_j}(\{\phi_{-j}, \tilde{\phi}_j\}), \beta_i^A), \\ & \hat{p}_j \neq 0, \beta_i^j = 1 \end{aligned} \tag{6.45}$$

where $\hat{p}_j$ is defined as

$$\hat{p}_j = \begin{cases} p_k & \text{if} \beta_j^j = 1, j \in A_k, \alpha_k^j = 1 \\ 0 & \text{otherwise}, \forall k \in \{1, 2, \ldots, J\}. \end{cases} \tag{6.46}$$

Similarly, Eq.(6.45) is the incentive compatible constraint for the primary users, which guarantees that the primary user will give the usage permission of their channels to the CR users so that they can receive the optimal payments.

As explained in Eqs. (6.43)and (6.45), to obtain the optimal allocation and payments, a multi-objective optimization problem needs to be solved, which becomes extremely complicated due to the game setting in the proposed method that only involves incomplete information. Considering the double auction scenarios of the proposed pricing game, Competitive Equilibrium (CE) [100] is a well-known theoretical prediction of the outcomes. It is the price at which the number of buyers willing to buy is equal to the number of sellers willing to sell. Alternatively, CE can also be interpreted as where the supply and demand match [147]. The supply function can be defined as the relationship between the acquisition costs of primary users and the number of corresponding channels; the demand function can be defined as the relationship between the reward payoffs of CR users and the number of corresponding channels, as shown in Figure 6.11. Note that CE is also proved to be Pareto optimal in stationary double auction scenarios [119]. To achieve the CE the traditional continuous bid/ask interactions among players will involve a great amount of

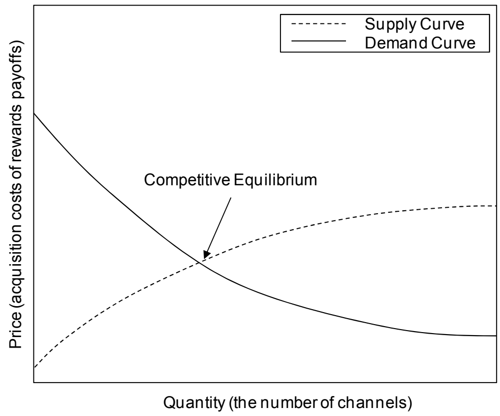


Figure 6.11 Illustration of supply and demand functions.

message exchanges and require powerful centralized control, which may not be applicable to wireless networking scenarios due to the limited bandwidth of control channels.

Considering network dynamics due to mobility, channel variations or wireless traffic variations, the CR users may have different reward payoffs of acquiring stages. Or, the CR users may face various channel fading conditions within different spectrum ranges or during different time periods, which changes their payoff values v_i^j at different time stages. Moreover, the costs of primary users will also change over time due to network dynamics. For instance, if the legacy users themselves have larger spectrum demands, some legacy channels may not be available for leasing anymore, which actually indicates an infinite leasing cost of those channels in this pricing model. In brief, $c_i^{a_j}$ and v_i^j need to be considered as random variables in dynamic scenarios, which satisfy the probability density functions (PDF) $f_c(c)$ and $f_v(v)$, respectively. Therefore, based on dynamic network conditions, the spectrum sharing is modeled as a multi-stage dynamic pricing game. Let γ be the discount factor of the multi-stage game. Based on Eqs. (6.43)and (6.45), the objective functions for the primary users and CR users can be rewritten as follows.

$$\tilde{O}(p_i) = \max_{\phi_{A_i}, t, \alpha_{i,t}^{A_i}} E_{c_i^{a_j}, v_i^j}[\sum_{t=1}^{\infty} \gamma^t \cdot U_{p_i,t}](\phi_{A_i}, t, \alpha_{i,t}^{A_i}) \tag{6.47}$$

$$\tilde{O}(s_i) = \max_{\phi_{A_i}, t, \beta_{i,t}^{A}} E_{c_i^{a_j}, v_i^j}[\sum_{t=1}^{\infty} \gamma^t \cdot U_{s_i,t}](\phi_{A_i}, t, \beta_{i,t}^{A}) \tag{6.48}$$

where the subscript t indicates the tth stage of the multi-stage game.

Furthermore, the above problem need to be further modeled as a dynamic programming process [12] [129] to obtain optimal sequential strategies by considering some state constraints such as the number of channels to be allocated at every stage or the residual monetary budget. However, the major difficulty of dynamic spectrum sharing lies in that how to efficiently and quickly update the spectrum sharing strategies adapt to the changing network conditions only based on local information. To address this issue, a belief-assisted dynamic pricing approach is developed, which can responds quickly to networking dynamics while only introducing limited overhead as well as approach CE outcomes.

Since the proposed pricing game belongs to the non-cooperation games with incomplete information [197], the players need to build up certain beliefs of other players' future possible strategies to assist their decision making. Considering that there are multiple players with private information in the pricing game and what directly affect the outcome of the game are the bid/ask prices, it is more efficient to define one common belief function based on the publicly observed bid/ask prices than generating specific belief of every other player's private information. Hence, the primary/CR users' beliefs are considered as the ratio their bid/ask being accepted at different price levels [100]. At each time during the dynamic spectrum sharing, the ratio of asks from primary users at x that have been accepted can be written as follows.

$$\tilde{r}_p(x) = \frac{\mu_A(x)}{\mu(x)} \tag{6.49}$$

where $\mu(x)$ and $_A(x)$ are the number of asks at x and the number of accepted asks at x, respectively. Similarly, at each time during the dynamic spectrum sharing, the ratio of bids from CR users at y that have been accepted is

$$\tilde{r}_s(x) = \frac{\eta_A(x)}{\eta(x)} \tag{6.50}$$

where $\eta(y)$ and $\eta_A(y)$ are the number of bids at y and the number of accepted bids at y, respectively. Usually, $\tilde{r}_p(x)$ and $\tilde{r}_s(y)$ can be accurately estimated if a great number of buyers and sellers are participating in the pricing at the same time.

However, in our pricing game, only a relatively small number of players are involved in the spectrum sharing at the specific time. The beliefs, namely, $\tilde{r}_p(x)$ and $\tilde{r}_s(y)$ cannot be practically obtained so that we need to further consider using the historical bid/ask information to build up empirical belief values. Based on the characteristics of double auction, the following observations are obtained:

- If an ask $\tilde{x} < x$ is rejected, the ask at x will also be rejected;
- If an ask $\tilde{x} > x$ is accepted, the ask at x will also be accepted;
- If a bid $\tilde{y} > x$ is made, the ask at x will also be accepted.

Based on the above observations, the players' beliefs can be further defined as follows using the past bid/ask information.

- *Definition 1:* Primary users' beliefs: for each potential ask at x, define

$$\hat{r}_p(x) = \begin{cases} 1 & x = 0 \\ \dfrac{\sum_{w \geq x} \mu_A(w) + \sum_{w \geq x} \eta(w)}{\sum_{w \geq x} \mu_A(w) + \sum_{w \geq x} \eta(w) + \sum_{w \leq x} \mu_R(w)} & x \in (0, M) \\ 0 & x \geq M \end{cases} \tag{6.51}$$

  where $\mu_R(w)$ is the number of asks at w that has been rejected, M is a large enough value so that the asks greater than M will not be accepted. Also, it is intuitive that the ask at 0 will be definitely accepted as no cost is introduced.

- *Definition 2:* CR users' beliefs: for each potential bid at y, define

$$\hat{r}_s(x) = \begin{cases} 1 & y = 0 \\ \dfrac{\sum_{w \leq y} \eta_A(w) + \sum_{w \leq y} \mu(w)}{\sum_{w \leq y} \eta_A(w) + \sum_{w \leq y} \mu(w) + \sum_{w \geq y} \eta_R(w)} & y \in (0, M) \\ 0 & y \geq M \end{cases} \quad (6.52)$$

  where $\eta_R(w)$ is the number of bids at w that has been rejected. And, it is intuitive that the bid at 0 will not be accepted by any primary users.

Generally, before the double auction pricing game converges to CE, there may exist a gap between the highest bid and lowest ask, which is called the spread of double auction. The spread reduction rule (SRR) states that any ask that is permissible must be lower than current lowest ask, i.e., outstanding ask [100], and then either each new ask results in an agreed transaction or it becomes the new outstanding ask. A similar argument can be applied to bids. By defining current outstanding ask and bid as ox and oy, respectively, we let $\bar{r}_p(x) = \hat{r}_p(x) \cdot I_{[0,ox)}(x)$ for each x and $\bar{r}_s(y) = \hat{r}_s(x) \cdot I_{(oy,M]}(y)$ for each y, which are modified belief function considering the SRR. Note that $I_{(a,b)}(x)$ is defined as

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{if } x \in (a, b) \\ 0 & \text{Otherwise} \end{cases} \quad (6.53)$$

By using the belief function $\bar{r}_p(x)$, the payoff maximization of selling the ith primary user's jth channel can be written as

$$\max_{x \in (oy, ox)} E[U_{p_i}(x, j)], \quad (6.54)$$

where $U_{p_i}(x, j)$ represents the payoff introduced by allocating the jth channel when the ask is x, and then $E[U_{p_i}(x, j)] = (x - c_i^{a_j}) \cdot \bar{r}_p(x)$. Similarly, as for the CR user s_i, the payoff maximization of leasing the jth channel in the spectrum pool can be written as

$$\max_{y \in (oy, ox)} E[U_{s_i}(y, j)], \quad (6.55)$$

where $U_{s_i}(y, j)$ represents the payoff introduced by leasing the jth channel in the spectrum pool when the bid is y, and then $E[U_{s_i}(y, j)] = (v_i^j - y) \cdot \bar{r}_s(y)$. Therefore, by solving the optimization problem for each primary and CR user using Eqs. (6.54) and (6.55), respectively, primary and CR users can make the optimal decision of spectrum allocation at every stage conditional on dynamic spectrum demand and supply.

Based on the above discussions, the proposed belief-assisted dynamic pricing algorithm for spectrum allocation can be expressed as follow:

1. Initialize the users' beliefs and bids/asks
   - The primary users initialize their asks as large values close to M and their beliefs as small positive values less than 1;
   - The CR users initialize their bids as small values close to 0 and their beliefs as small positive values less than 1.
2. Belief update based on local information:

- Update primary and CR users' beliefs using Eqs. (6.51) and (6.52), respectively

3. Optimal bid/ask update:

   - Obtain the optimal ask for each primary user by solving Eq. (6.54);
   - Obtain the optimal bid for each CR user by solving Eq. (6.55).

4. Update leasing agreement and spectrum pool:

   - If the outstanding bid is greater than or equal to the outstanding ask, the leasing agreement will be signed between the corresponding users;
   - Update the spectrum pool by removing the assigned channel.

5. Iteration:

   - If the spectrum pool is not empty, go back to Step 2.

RELAY-BASED SPECTRUM SHARING

In order to improve the spectrum utilization, cooperative relays have been recently introduced to spectrum sharing in CR networks [290], where a relay node with rich available spectrum bands acts as a bridge for communication between a source and a destination nodes. Figure 6.12 shows a typical cooperative relay system for cooperative spectrum sharing, where CR users co-exist with multiple PUs, PUs 1, 2, 3, and 4, and their corresponding licensed spectrum bands, CHs 1, 2, 3, and 4, respectively. Without loss of generality, we consider a three-terminal CR relays system, which consists of source, relay, and destination. As shown in the Figure 6.12, those available spectrum bands can support dual-hop transmission (CHs 1 and 2), relay transmission (CH 3), and direct transmission (CH 4). Therefore, spectrum heterogeneity observed at source, relay, and destination nodes brings new challenges to cooperative relays.

In [130], relays have been used for balancing traffic request and spectrum resource. In [102], the idea of using unused bands via relay nodes has been proposed to increase spectrum utilization. In [181], the method of using common bands via relay nodes to enhance the signal-to-noise ratio has been studied. However, existing works with a single relay consider separate end-to-end transmissions between the relay node and other nodes. In other words, how to perform cooperative relays with all available spectrum bands at these three terminals has not been addressed. Thus, the overall end-to-end performance can be further improved by advanced cooperative relay design.

6.3.2 Intra-Network Spectrum Sharing for Ad Hoc (Distributed) Networks

If the CR network does not have a centralized support, each CR user should determine their action by himself or herself depending on its local observations. To overcome the drawback caused by the limited knowledge of network topology and spectrum availability, all spectrum sharing functions are based on cooperative operations, where CR users determine their actions based on observed information exchanged with their neighbors. The cooperative spectrum sharing is theoretically more advantageous in the distributed network environment since the uncertainty in a single user's observation can be minimized through collaboration. In this subsection, spectrum sharing schemes based on distributed

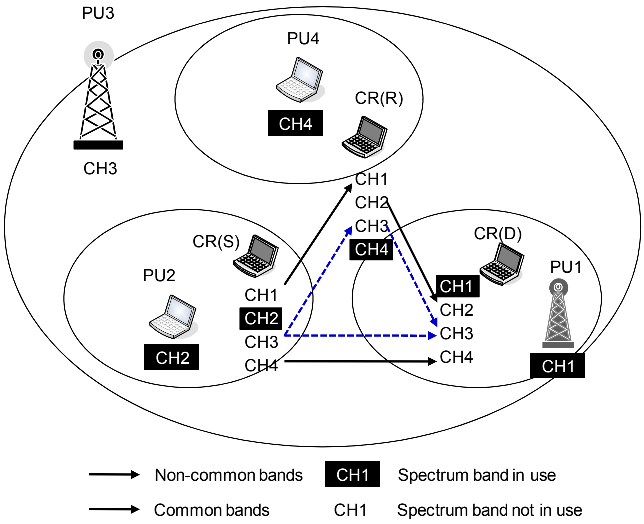


Figure 6.12 Cognitive Radio Relay

and cooperative operations are investigated.

INTERFERENCE COMPENSATION

In this work, both single channel and multi-channel asynchronous distributed pricing (SC/MC-ADP) schemes are proposed, where each node announces its *interference price* to other nodes [116]. Using this information from its neighbors, a node can first allocate a channel and in case there exist users in that channel, then, determine its transmit power. While there exist users using distinct channels, multiple users can share the same channel by adjusting their transmit power. Furthermore, the SC-ADP algorithm provides higher rates to users when compared to selfish algorithms where users select the best channel without any knowledge about their neighbors' interference levels. outperforms underlay techniques.

[System Model]

The method consider a set of $\mathcal{K} = \{1, \ldots, K\}$ spectrum agile users that seek to share a set of $\mathcal{M} = \{1, \ldots, M\}$ available channels (open spectrum bands). Each user corresponds to a distinct pair of nodes: one dedicated transmitter and one dedicated receiver, and is constrained to transmit over at most one spectrum band; this could be due to policy and/or technical limitations. For simplicity, every spectrum band is modeled as having the same bandwidth and the same background noise power of n_0. Over the time-period of interest, the channel gains are assumed to be fixed and that the users want to transmit continually. For channel m, the gain between user k's transmitter and user j's receiver is denoted by h_{kj}^m. An example of a network with four users (pairs of nodes) is shown in Figure 6.13. Here T_k and R_k denote the transmitter and receiver for user k, respectively.

Let $\varphi(k) \in M$ denote the spectrum band selected by user k. In addition to selecting a band, each user can determine its transmission power $p_k^{\varphi(k)}$ within the band. This transmission power must lie be in a feasible set $\mathcal{P}_k^{\varphi(k)} = [\check{P}_k^{\varphi(k)}, \hat{P}_k^{\varphi(k)}]$, with $0 \leq \check{P}_k^{\varphi(k)} \leq \hat{P}_k^{\varphi(k)}$. The power constraints may vary with the selected band, for example to model

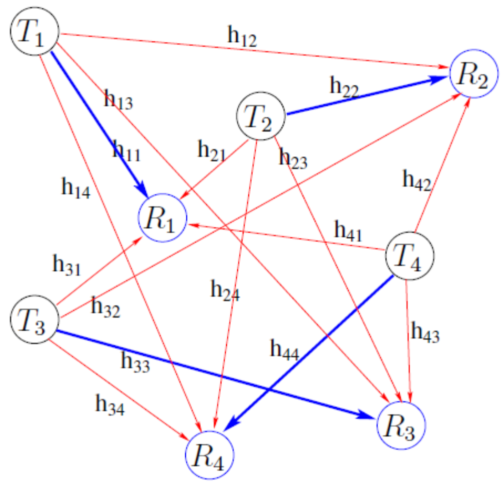


Figure 6.13 Ad hoc network with four pairs of nodes

different regulatory constraints. Note that a special case is when $\check{P}_k^{\varphi(k)} = \hat{P}_k^{\varphi(k)}$, in which case a user always transmits with maximum power on its selected band. This work considers a spread spectrum system, where this power is spread over the entire band and interference from other users in the same band is treated as noise. Each user k's QoS is characterized by a utility function $u_k(\gamma_k^{\varphi(k)})$, which is an increasing and strictly concave function of the received SINR on the chosen channel. The SINR of user k on channel $m \in M$ is

$$\gamma_k^m(p^m) = \frac{p_k^m h_{kk}^m}{n_0 + \sum_{j \neq i} p_j^m h_{jk}^m} \tag{6.56}$$

where $p_m = (p_k^m, k \in \mathcal{K})$ is a vector of the users' transmission powers on channel m. Here a utility function is assumed to be $u_k(\gamma_k^{\varphi(k)}) = \theta_k \log(1 + \gamma_k^{\varphi(k)}$, which is proportional to the Shannon capacity of user k's channel weighted by a user dependent priority parameter, θ_k.

From a network perspective, the objective of this method is to determine each user's channel selection and power allocation to maximize the total utility summed over all users, i.e.,

$$\max_{\{\varphi(k), P_k^{\varphi(k)}\}} u_{\text{tot}}(p) = \sum_{k=1}^{K} u_k(\gamma_k^{\varphi(k)}(p^{\varphi(k)})) \tag{6.57}$$

This is an integer and possibly non-convex optimization problem, which is typically difficult to solve. Moreover, in a spectrum sharing environment it may not be feasible for a single entity to acquire the global information needed to solve this problem.

Single-Channel Asynchronous Distributed Pricing (SC-ADP) Algorithm To solve Eq. 6.57 the SC-ADP algorithm is developed based on a simple, distributed heuristic algorithm. In the SC-ADP algorithm each user $k \in \mathcal{K}$ communicates the negative externality due to interference by announcing an "interference price", $\pi_k^{\varphi(k)}$ for the channel $\varphi(k)$

on which it is currently transmitting. This price is given by

$$\pi_k^{\varphi(k)} = \left| \frac{\partial_{u_k}(\gamma_k^{\varphi(k)}(p^{\varphi(k)}))}{\partial(\sum_{j \neq k} p_j^{\varphi(k)} h_{jk}^{\varphi(k)})} \right| \tag{6.58}$$

which reflects the marginal increase of user k's utility if its received interference is decreased by one unit. Based on the current interference prices and the current level of interference, each user $k \in \mathcal{K}$ selects a channel $\varphi(k)$ and a feasible power allocation $p_k^{\varphi(k)} \in P_k^{\varphi(k)}$ that maximizes its surplus

$$\begin{aligned} & s_i(\varphi(k), p_k^{\varphi(k)}, p_{-k}^{\varphi(k)}, \pi_{-k}^{\varphi(k)}) \\ & = u_k(\gamma_k^{\varphi(k)}(p_k^{\varphi(k)})) - p_k^{\varphi(k)} \sum_{j \neq k} \pi_k^{\varphi(k)} h_{kj}^{\varphi(k)} \end{aligned} \tag{6.59}$$

Here $p_{-k}^{\varphi(k)} = (p_j^{\varphi(k)}, j \in \mathcal{K} \text{ and } j \neq k)$ denotes the vector of powers of every user except user k in channel $\varphi(k)$: $\pi_(^{\varphi(k)} - k)$ is similarly defined. The algorithm progresses by having each user update its price announcement and channel/power allocation according to these rules. In general these updates can be asynchronous across users. For each $k \in \mathcal{K}$, let $\mathcal{T}_k$ be an unbounded set of positive time instances at which user k updates its price and channel/power allocation. The updates at these time instances are specified in the following SC-ADP algorithm.

1. Initialization: For each user $k \in \mathcal{K}$, select an initial channel $\varphi(k) \in \mathcal{M}$ and an initial power allocation $p_(^{\varphi(k)} k) \in P_k^{\varphi(k)}$.
2. At each $t \in \mathcal{T}_k$, user k
   - Selects $\varphi(k) \in \mathcal{M}$ and $p_k^{\varphi(k)} \in P_k^{\varphi(k)}$ to maximize its surplus in Eq. (6.59),
   - Announces price $\pi_k^{\varphi(k)}$ according to Eq. (6.58).

Furthermore, the SC-ADP algorithm can be easily extended to the a multi-mhannel (MC)-ADP algorithm where CR users can transmit over multiple channels and exchange interference prices over each channel. In (MC)-ADP algorithm, each user k distributes power across all of the available M channels to maximize the surplus

$$\sum_{m=1}^{M} u_k(\gamma_k^m(p_k^m)) - \sum_{m=1}^{M} P_k^m \sum_{j \neq k} \pi_j^m h_{kj}^m \tag{6.60}$$

subject to a total power constraint $\sum_{m=1}^{M} p_k^m \leq P_i^{\max}$.

LOCAL BARGAINING

A cooperative local bargaining (LB) scheme is proposed in [36] to provide both spectrum utilization and fairness. The local bargaining framework is formulated based on the framework in [202] and [297]. Local bargaining is performed by constructing local groups according to a poverty line that ensures a minimum spectrum allocation to each user and hence focuses on fairness of users. The evaluations reveal that local bargaining can closely approximate centralized graph coloring approach at a reduced complexity. Moreover, localized operation via grouping provides an efficient operation between a fully distributed and a centralized scheme.

[Local Bargaining Principles]
The approach described in Section 6.3.1 globally optimizes spectrum allocation for a given topology. In a mobile network model, node movements lead to constant changes in network topology. Using the existing approach, we can reapply the spectrum allocation algorithm after each change in the conflict graph. This approach assumes no prior allocation information, and incurs high computation and communication overheads. To reduce these overheads, an adaptive and robust distributed algorithm is proposed, which takes prior allocation into account in new spectrum assignments [36].

An efficient dynamic allocation algorithm can run every time user movement causes a change in the corresponding network conflict graph. Assuming the spectrum allocation was near optimal before the topology change, local bargaining between affected vertices can quickly optimize allocations for utilization and fairness. During local bargaining, sets of neighboring vertices, each of which form a connected component of the conflict graph, self-organize into bargaining groups. In this method, a simple group formation is proposed where a node who wants to improve its spectrum assignment broadcasts a bargaining request to its neighbors. Those neighbors whose are willing to participate reply to the sender and form a bargaining group. For each bargaining group, the requester becomes the group coordinator and performs the bargaining computation. There are two bargaining strategies: *one-to-one bargaining* and *one-buyer-multi-seller bargaining*. In one-to-one bargaining, the node n_1 who initiates the bargaining can choose to bargain with only one neighboring node n_2 at a time. When multiple neighbors e.g. n2 and n3 acknowledge the bargaining request, n1 can sequentially compute assignment assuming bargaining with n2 first, and then with n_3. On the contrary, in one-buyer-multi-seller bargaining, a buyer node n_1 purchases a set of channels M_0, from its neighbors who are currently using any channel in M_0, such that to improve system utility. In this case, the bargaining requires concurrent approval from multiple neighbors.

Once the bargaining groups are organized, the bargaining inside each group should not disturb the spectrum assignment at nodes outside the group. That is, after the bargaining, the modified channel assignment should not lead to any conflict with nodes outside the group. To this end, the members of any two bargaining groups can not be directly connected. An example of isolation between bargaining groups is shown in Figure 6.14. This helps to maintain system stability, so that a bargaining may not invoke a series of reactions due to violations in interference constraints. More importantly, this guarantees that if a bargaining improves the utility in a local area, it also improves the system utility.

[Bargaining Steps]
This work designs a distributed, iterative grouping and local bargaining process is proposed, assuming that nodes periodically broadcast their current channel assignment and interference constraints to their neighbors. Each node has three states: *bargaining*, *disabled* and *enabled*, as shown in Figure 6.15. Only enabled nodes can perform bargaining. The actual bargaining involves the following 4 steps, and repeats until no further bargaining can improve system utility.

1. *Initialize Bargaining Request:* Based on broadcasts of channel assignments and interference constraints from neighbors, an enabled node determines if bargaining

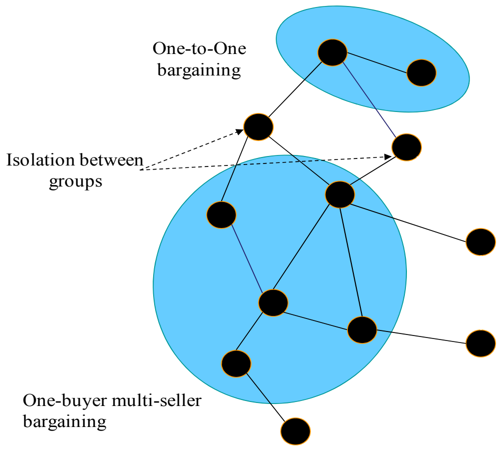


Figure 6.14 Bargaining group

with a neighbor will lead to an improvement in system utility. If such neighbors exist, the node broadcasts a bargain request to the neighbors along with its current channel assignment and interference constraints.

2. *Acknowledge Bargaining Request:* Neighbors who are enabled and willing to bargain reply an ACK message with its current channel assignment and interference constraints. This scheme assumes that nodes are willing to collaborate to improve system utility, and accept requests that improve system utility even if it might degrade their individual channel assignments. If a node receives multiple concurrent requests from its neighbors, it acknowledges the request that leads to the highest bargaining gain as calculated based on information embedded in the request.

3. *Bargain Group Formation:* When the requester receives the replies, it selects the members of the bargaining group, and broadcasts this information along with the proposed modification of the channel assignment to neighbors. Once the bargaining group is set, its members enter bargaining state. They broadcast a DISABLE message with a timer equal to the estimated duration of the bargain process to neighbors not in the bargaining group. Nodes receiving the message enter the disabled state for the duration of the timer. This procedure prevents nodes who are neighbors of existing bargaining group to participate in any future bargaining before the timer expires.

4. *Bargaining:* Once all members acknowledge the changes to the channel assignment, each member updates its local channel assignment. This is straightforward for one-to-one bargaining. For one buyer- multiple-seller bargaining, interactions among members can be coordinated by the bargain requestor. After bargaining, each member enters the enabled state. Figures 6.15 and 6.16 illustrate the node state transition and messages during bargaining. Once the local bargaining procedure is set, the specific bargaining strategy may be customized for different utility functions.

[Fairness in Local Bargaining]
The proposed scheme mainly focus on the local bargaining strategy optimizing for fairness. Based on the definition of proportional fairness, the optimization aims to maximize the total logarithmic user throughput, i.e. the product of user throughput. Therefore, the global

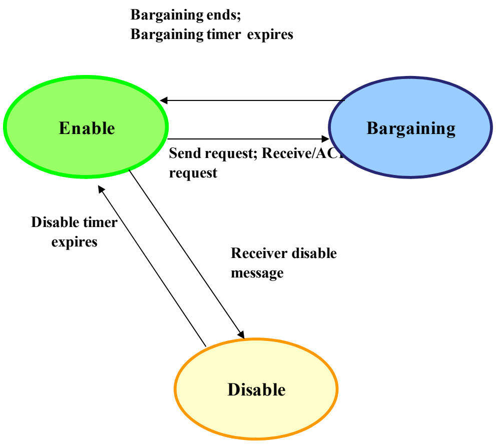


Figure 6.15 Node state and transitions.

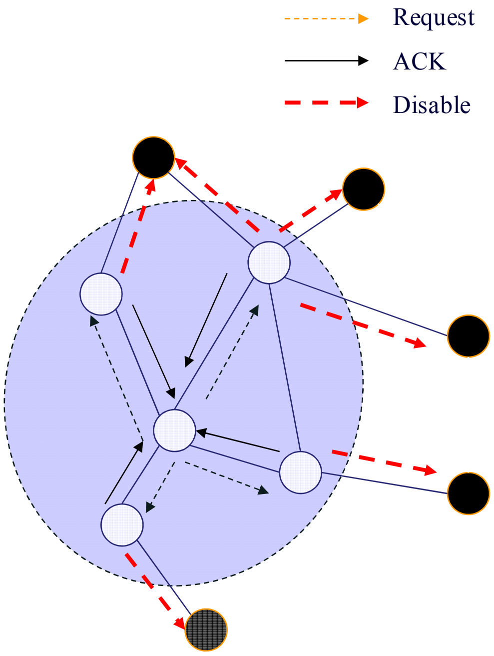


Figure 6.16 Messages exchanged during bargaining.

fairness utility increases if nodes with many assigned channels "give" some channels to nodes with few assigned channels. Let $L(n)$ be the set of channels available at n, $f_A(n)$ be the set of channels assigned to node n under current assignment A, $TP_A(n)$ be the throughput that user n gets under assignment A, and $Nbr(n)$ be the neighbors of n. Then, two fairness bargaining strategies are described as follow:

- *One-to-One Fairness Bargaining:* As described previously, one-to-one bargaining allows two neighboring nodes n_1 and n_2 to exchange channels to improve system utility while complying with conflict constraints from the other neighbors. For n_1

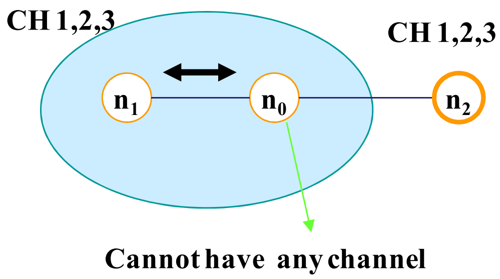


Figure 6.17 An example of Starvation

and n_2 to bargain, they need to first obtain the channels that are bargainable to avoid disturbing other neighbors, referred to as $C_b(n_1, n_2)$. Given $C_b(n_1, n_2)$, the one-to-one bargaining regarding fairness can be performed as follows: For an assignment A, the one-to-one fairness bargaining finds nodes n_1 and n_2, and their bargaining channel set $C_b(n_1, n_2)$, and modifies A to A' related to n_1, n_2 and channels $C_b(n_1, n_2)$, such that

$$TP_{A'}(n_1) \cdot TP_{A'}(n_2) > TP_A(n_1) \cdot TP_A(n_2) : \tag{6.61}$$

The one-to-one Fairness Bargaining increases the product of the bargaining users while other nodes' throughput values remain unaffected. Hence, the system fairness increases with each bargaining. The improvement between each pair of nodes (n_1, n_2) can be calculated as $G(n_1, n_2) = \frac{TP_{A'}(n_1)TP_{A'}(n_2)}{TP_A(n_1)TP_A(n_2)} - 1$. This is used in the bargaining process (in Section III) to determine whether a bargaining can improve system utility. Given (n_1, n_2), assigning channels to n_1 and n_2 to maximize their throughput product is a difficult task. This is because node throughput depends on all channels (including non-bargainable ones) assigned to a node, and the available bandwidth on a channel differs between nodes. The effectiveness of One-to-One bargaining is constrained by the size of $C_b(n_1, n_2)$. In general, due to heavy interference constraints among neighboring nodes, $C_b(\cdot)$ could be very small. Figure 6.17 illustrates an example where the conflict graph is a chain topology consisting of three nodes A,B,C. Node B is not assigned with any channel and the system utility is zero. We refer to this as user starvation. Node a and b cannot bargain due to the constraint from c (i.e. $C_b(a, b) = \emptyset$), while node b and c also cannot bargain due to the constraint from a (i.e. $C_b(b, c) = \emptyset$). Hence, the fairness bargaining is not effective to eliminate user starvation.

- *Feed Poverty Bargaining:* User starvation in most cases is a result of the lack of flexibility in bargaining. As for the example in Figure 6.17, by allowing A and C to give up channel 1 at the same time and feed it to B, we can remove the starvation at B. This is an example of one-buyer-multi-seller bargaining. In this method, a special one-buyer-multi-seller bargaining is proposed, called Feed Poverty where if a node (buyer) has very poor channel assignment, the neighboring nodes can collaborate together to feed it with some channels.

  For an assignment A, a feed poverty bargaining is to find some node n_0 and channel m_0 and modify A to A', such that n_0's neighbors give up channel m_0 and feed it to

n_0 as follow:

$$A'_{m,n} = \begin{cases} 1 & : \quad m = m_0 \text{ and } n = n_0 \\ 0 & : \quad m = m_0 \text{ and } n \in Nbr(n_0) \\ A_{m,n} & : \quad \text{otherwise} \end{cases} \tag{6.62}$$

(intuitively, the assignment let some of n_0's neighbors give up channel m_0 and feed it to n_0) and

$$\begin{aligned} TP_{A'}(n_0) \cdot & \prod_{n \in Nbr(n_0) \wedge A_{m_0,n}==1} TP_{A'}(n) \\ - TP_A(n_0) \cdot & \prod_{n \in Nbr(n_0) \wedge A_{m_0,n}==1} TP_A(n) > 0 \end{aligned} \tag{6.63}$$

This means the product-throughput of the users involved in the bargaining is locally increasing, while the other users' throughput are not affected. So generally the bargaining improves system utility, except that, in case of starvation of other users, the system utility remains $-\infty$. A special case of Feed Poverty is when $A_{m_0,n} = 0$ for all $n \in Nbr(n_0)$. This means none of n_0's neighbors are using channel m_0, and n_0 simply seizes it. When there is no feasible One-to-One Fairness Bargaining, i.e. $|C_b| = \emptyset$, the requestor initializes a Feed Poverty Bargaining on all neighbors who acknowledge the request. The requestor sequentially selects multiple channels to maximize group utility.

In this work, a fairness bargaining mechanism with feed poverty (BF) is proposed, which combines one-to-one fairness bargaining and feed poverty bargaining as follow: 1) Each node who wants to improve its spectrum usage starts with negotiating one-to-one fairness bargaining with its neighbors to improve system utility. 2) If there is no bargainable channels between it and any of its neighbors, a starved node can broadcast a feed-poverty request to its neighbors to initialize feed poverty bargaining. Overall, a channel assignment A is said to be BF-optimal if no further fairness bargaining with feed poverty can be performed on it. In the following, it is shown that when channels are of uniform bandwidth, a theoretical lower bound on the total number of channels or throughput that each user can achieve, can be derived. When the system converges to BF-optimal assignment A, the number of channels a node n obtains is low-bounded. This lower bound, defined as *poverty line*, represents the minimum amount of spectrum a node is entitled to. If a node n has $L(n)$ available channels and $d(n)$ conflicting neighbors, its poverty line is

$$TP(n) \geq \left\lfloor \frac{|L(n)|}{d(n)+1} \right\rfloor = PL(n) \tag{6.64}$$

The detailed proof is found in [36]. The degree of a vertex $d(n)$ is defined as the number of edges it is associated with, a measure of the number of channel sharers in the neighborhood it has to compete with. Eq. (6.64) shows that the proposed Fairness Bargaining with Feed Poverty guarantees a poverty line $PL(n)$ to each vertex n. The poverty line of a vertex, i.e. the throughput a vertex deserves, scales inversely with the number of sharers, which is also the spirit of some greedy allocation algorithms [202] [297]. The poverty line also provides a guideline in bargaining in real systems where a vertex is entitled to request bargaining if its current throughput is below its poverty line. They refer to this as the Poverty guided bargaining.

6.3.3 Non-Cooperative Intra-Network Spectrum Sharing

While cooperative approaches provide better sharing performance, they cause adverse effects on resource-constrained networks due to the additional operations and overhead traffic. On the contrary, non-cooperative spectrum sharing enables CR users to make decisions independently without exchanging information or negotiating with neighbor users, and hence their spectrum sharing strategy should be selfish. While non-cooperative solutions may result in low spectrum utilization, the minimal communication requirements among other users introduce a tradeoff for practical solutions in resource-constrained environments. In the following, we introduce two non-cooperative spectrum sharing schemes.

RULE-BASED CHANNEL ALLOCATION

A potential problem in the solution provided in local bargaining [36] is that a common control channel may not exist in CR networks or can be occupied by a primary user. To address this issue, an opportunistic spectrum management scheme is proposed in [37], where users allocate channels based on their observations of interference patterns and neighbors. In the device centric spectrum management scheme (DCSM), the communication overhead is minimized by providing five different system rules for spectrum allocation.

In the proposed system, nodes observe local conditions and neighbors' actions and independently adapt their spectrum usage. Their behavior is regulated by a set of rules defined by spectrum regulators. Each node n performs spectrum sensing to identify its spectrum usage. Using spectrum rules, each node checks whether it needs to update its channel selections. If an update is required, nodes rely on rules to determine the appropriate channels to use. In contrast to the explicit coordination approach [36], nodes tend to prioritize their own performance with minimal regard to system utility. However, their compliance with the rules promotes efficient and fair spectrum sharing. The comparative analysis of this scheme with the cooperative schemes show that rule-based spectrum access results in slightly worse performance, but the communication overhead is reduced significantly.

The key challenge in this design is how to define the spectrum rules. The rules specify how many and which channels a node should use, such that fairness and utilization can be achieved. The estimation should not be overly aggressive and bring excessive contention, or overly conservative and result in spectrum under-utilization. Further, in a distributed system, each node can only act based on limited local view of the system. In this scheme, five different rules is proposed that tradeoff between performance and signaling complexity for different application scenarios.

- *Rules for Conflict Free Channel Assignment:* In this case, nodes always select idle channels, i.e. channels unclaimed by conflicting peers. To provide fairness, the rules limit the number of channels each node can access. Conflict free channel usage allows for explicit and guaranteed throughput provisioning and control over packet delay.

  – *Rule A (Uniform Idle Preference):* Each node adjusts its spectrum usage to $\Omega = \min_n PL(n)$ number of idle channels. Service providers can optimize the value of Ω for the entire network. However, nodes experiencing intensive interference from legacy nodes (i.e., small $L(n)$) or other peers in a crowded area (i.e., large $d(n)$) can limit the the value of Ω , leading to less than ideal spectrum utilization. Therefore, adapting to each node's interference condition is preferred.

- *Rule B (Poverty Exact Idle Preference):* A node n selects exactly $PL(n) = \lfloor \frac{L(n)}{d(n)+1} \rfloor$ idle channels. If the number of idle channels $< PL(n)$, it "seizes" channels from "richer" nodes without affecting "poor" nodes. A node conflicting with a "poor" node will sense the conflict and give up the channel and switch to other channels following the same procedure. To n, a neighbor is "richer" if it uses more channels than n; otherwise it is "poor". Rule B requires that each node has knowledge of the number of neighbors $d(n)$, and the channel selection of each neighbor in order to identify "richer" nodes. To "grab" non-idle channels, a node n marks the channels occupied by "poor" neighbors as busy, and the rest as idle. node n then selects a set of channels from the "idle" channels until its channel occupancy reaches $PL(n)$. The efficiency of grabbing depends on the set of channels selected.

  A limitation of Rule B is that each node only attempts to use $PL(n)$ channels. Since $PL(n)$ represents a lower bound on spectrum usage derived using a collaboration based approach [36], Rule B could under-utilize available spectrum.

- *Rule C (Poverty Guided Idle Preference):* A node n selects channels from idle channels. Only if there are not enough idle channels to reach $PL(n)$ does node n "grab" channels from "richer" neighbors. The number of channels it can grab from any "richer" node r, is $\max\{0, \min\{C(r).PL(n), PL(n) - C(n)\}\}$ where $C(n)$ and $C(r)$ are the current spectrum usage of node n and r.

  Rule C allows nodes who have attained their poverty line to seize additional idle channels. It still allows nodes below their poverty line to take channels from "richer" neighbors. However, each grabbing can not reduce a "richer node's spectrum below the grabber's poverty line, avoiding cycles of nodes grabbing channels from each other in turn. In particular, a node n can collect all the channels used by its "richer" neighbors but not "poor" neighbors into a channel pool, reserve $PL(n)$ channels for each "richer" neighbor and "grab" from the rest of the pool.

  Rule C does not require each node to have knowledge of its neighbors' poverty line. However, the performance of conflict free channel assignments such as Rules B and C depends on the granularity of spectrum partition, i.e. the number of channels M. When M is small compared to the number of neighbors $d(n)$, some nodes may have a poverty line of zero, and hence no performance guarantee. In this case, the system can increase granularity by partitioning time, e.g. a channel is defined as a frequency band at a particular time slot.

• *Rules for Contention-based Channel Assignment:* Broadcasting spectrum usage to neighbors might be undesirable for a number of reasons, including privacy concerns and protection against jamming from malicious nodes. For these reasons, two more rules are proposed not to require knowledge of neighbors' spectrum usage.

  In this approach, on each channel, nodes follow a set of random access rules such as CSMA to compete fairly for channel access and avoid conflict. Each node performs contention detection, i.e. listens to the channel before initiating any transmission. It initiates the transmission only when the channel is idle for some given time.

Otherwise, it backs off and delays the action for a short period. Because channels have different contention conditions, nodes should invoke independent contention detection and backoff process on each channel. The penalty of such random access is the overhead of contention detection even if there is only one node on the channel.

The following two rules specify the number of channels nodes should use, and how to select these channels. The design of these rule depends on whether nodes have information about their poverty line.

- *Rule D (Selfish Spectrum Contention):* Each node n can use up to the Ψ channels providing the highest throughput. Communication on each channel is through CSMA based time contention. Here the poverty line concept can provide a reference for choosing different value of Ψ for different nodes. Since the poverty line represents throughput attainable from conflict free spectrum usage, Ψ_n should be larger than $PL(n)$ to account for channel contention. Note that in the random access scenario, $PL(n)$ can still be computed using only the number of neighbors, which can be estimated by listening to MAC control packets.
- *Rule E (Poverty Guided Selfish Spectrum Contention):* This rule is the same as Rule D except the number of channels each node n can use is limited by $\Psi_n = \max(KPL(n), 1), \alpha \geq 1$. Both rules encourage nodes to act selfishly. Nodes monitor channel conditions and switch to channels that provide the best throughput, even if such a switch might reduce performance for other neighbors. One question is how to choose the best channels with maximum capacity and minimum contention. Here, the number of competing nodes is used as an indicator of channel quality. Hence, following Rule D or E, nodes always switch to channels with the least number of competing nodes. This also makes both rules efficient.

Based on the definition of each rule, analytical bounds on the performance and complexity of the proposed rules are shown in the following

- *Conflict-Free Rules:*
  1. Rule A guarantees a conflict free spectrum allocation.
  2. Using Rule B or C, the system reaches an equilibrium after an expected number of at most $O(N^2)$ node spectrum modifications. In equilibrium, there is no conflict in spectrum usage, and each node's spectrum usage is no less than its Poverty Line $PL(n)$ (equal to $PL(n)$ for Rule B). Here equilibrium is the state where nodes have no incentive to adjust their spectrum usage.
- *Contention-based Rules:*
  1. Using Rule D or E, the system will reach an equilibrium after at most ΛM node spectrum modifications. Λ is bounded by $O(N^2)$.

     The choice of Ψ and K depends on specific random access mechanisms. To analyze their impact, we use a simple model to characterize channel sharing. A node contending with m other nodes on a channel gets $1/(\lambda \cdot (m+1))$ of the channel throughput, where λ is the contention penalty. We refer to this model as the (λ, m) model.When $\lambda = 1.8$, this model matches the experimental test and analytical results for CSMA-based IEEE 802.11b systems at 11Mbps in [112].

2. Using Rule D and (λ, m) model, a node n's throughput is lower-bounded by

$$LB(n, \Psi) = \begin{cases} \frac{1}{\lambda\{\lfloor \frac{d(n)}{M} \rfloor + 1\}} & \Psi = 1 \\ \frac{\Psi}{\lambda\{\lfloor \frac{\Psi \cdot d(n)}{M} \rfloor + 2\}} & 1 < \Psi < M \\ \frac{M}{\lambda\{d(n)+1\}} & \Psi = M \end{cases} \tag{6.65}$$

NON-COOPERATIVE POWER ALLOCATION GAME

In this work, spectrum sharing for unlicensed band is proposed based on the one-shot normal form game and repeated game [70]. Furthermore, it is shown that orthogonal power allocation, i.e., assigning the channel to only one transmission to avoid co-channel interference with other neighbors, is optimal for maximizing the entire network capacity.

The cooperative spectrum sharing techniques have implicitly assumed that the M systems cooperate to maximize a global utility function by choosing appropriate power allocations. However, in a spectrum sharing scenario where regulations may be lax and systems may be competing with one another to gain access to the shared medium, assuming selfish behavior may be more realistic.

Assume that each system i behaves selfishly and rationally, and are associated with a utility function $U_i(R_i)$, which is concave and increasing in R_i. The systems are selfish in the sense that they only try to maximize their own utility. The rationality assumption means that each system will never choose a strictly dominated strategy. This work analyzes the set of achievable rates in this noncooperative scenario using non-cooperative game theory.

Short Interaction Between Systems: One Shot Game

This work first considers a static game of complete and perfect information, usually known as the Gaussian Interference Game (GIG) [280], and is based on the power allocation model in Section 6.2.4. The game has M players, M systems. The strategy space S_i of system i is the set of power allocations $p_i(f),\ f \in [0, W]$ that satisfy the power constraint. A strategy s_i for user i is the choice of power allocation $p_i(f)$. For a given strategy profile $(s_1, \ldots, s_M)$ the rate of user i is given by Eq. (6.6). The players play simultaneously, and know the utility functions of all the other players ($N_0, \{c_{i,j}\}_{i,j}, \{P_i\}_{i=1}^M, W$ are common knowledge). A strategy profile $\{s_i^*\}_{i=1}^M$ is a Nash Equilibrium (N.E.) of the game if

$$\begin{gathered} R_i(s_1^*, \ldots, s_M^*) \geq R_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_M^*) \\ \text{for all } s_i \in \mathcal{S}_i,\ i = 1, \ldots, M \end{gathered} \tag{6.66}$$

A direct consequence of the flat-fading and white noise assumption is the following fact:

The set of frequency-flat allocations $p_i(f) = P_i/W, f \in [0, W]$ for $i = 1, \ldots, M$ *is a Nash Equilibrium of the GIG.*

This means that the best possible strategy for a given system is to spread its available power over the total bandwidth whenever all the interfering systems are spreading their signals. The above fact can be understood by noting that the best response of a system to a strategy profile of the other systems is to water-fill the available power over the

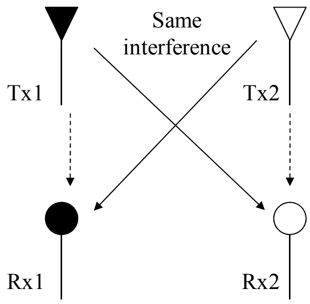


Figure 6.18 Symmetric situations between two systems (scenario I) [70].

noise+interference seen. When all the other systems use flat allocation, the water-filling power allocation is flat, and it follows that flat allocations are best responses to each other.

If the channel gains across systems are sufficiently small the full-spread Nash Equilibrium is the only Nash Equilibrium of the Gaussian game. The following theorem gives a sufficient condition for the uniqueness of the full-spread Nash Equilibrium.

Theorem 1: If $\sum_{j=1,j\neq i}^{M} \frac{c_{j,i}}{c_{i,i}} < 1$ for $i = 1, \ldots, M$ then the full-spread N.E. is the only Nash Equilibrium of the GIG.
The proof of this theorem is found in [70]. However, Theorem 1 does not give us any information about the uniqueness of the Nash equilibrium when the condition $\sum_{j=1,j\neq i}^{M} \frac{c_{j,i}}{c_{i,i}} < 1$ for all i is not met.

In many cases, the set of rates that results from the full-spread Nash Equilibrium is not Pareto efficient (i.e. is not in $\mathcal{R}^*$) so there may be a significant performance loss if M systems operate in this point due to lack of cooperation. And in many cases this inefficient outcome is the only possible outcome of the game.

[Example Scenarios]
Figure 6.18 illustrates a two system scenarios (call it Scenario I) with $c_{1,1} = c_{2,2} = 1$, $c_{2,1} = c_{1,2} = 1/4$, $W = 1$, $N_0 = 1$ and $P_1 = P_2 = P$.

Note that in this case the condition of Theorem 1 is satisfied. If both users spread their signals, they obtain rates $R_1^{\text{FS}} = R_2^{\text{FS}} = \log[1 + P/(1 + P/4)]$ [bits/s/Hz], which tend to $\log(5)$ [bits/s/Hz] as $P \to \infty$. However, if the systems orthogonalize their power allocations using half of the bandwidth each, the resulting rates are $R_1 = R_2 = (1/2)\log(1+2P)$ [bits/s/Hz], which tends to ∞ as $P \to \infty$. The regime in which $P \gg N_0$ corresponds to the high SNR regime. In this regime, when the systems orthogonalize their power allocations they can communicate with an interference free channel, and achieve large data rates. If on the contrary both systems spread their signals, the signal to interference plus noise ratio becomes limited by interference, resulting in a reduced communication rate. This example shows that the inefficiency resulting from choosing the full-spread equilibrium can be arbitrarily large.

Long Term Interactions: A Repeated Game
The two system scenario I mentioned above shows that there are situations in which the only possible outcome of the game is very inefficient, and as a result, there is a large

performance degradation due to lack of cooperation. This negative result can be attributed to the static nature of the game.

Many wireless systems operate and co-exist with the same set of competing systems over a long period of time. In this context, it may be more reasonable to model the scenario as a repeated (or dynamic) game where systems play multiple rounds, remembering the past experience in the choice of the power allocation in the next round. Consider an infinite horizon repeated game, where the GIG is repeated forever. The utility of each player is defined by

$$U_i = (1 - \delta) \sum_{t=0}^{\infty} \delta^t R_i(t) \tag{6.67}$$

where $R_i(t)$ is the utility of user i in the stage game at time t, and $\delta \in (0, 1)$ is a discount factor that accounts for the delay sensitivity of the systems. At the end of each stage, all the players can observe the outcome of the stage-game and can use the complete history of play to decide on the future action. A strategy in the repeated game is a complete plan of action, that defines what the player will do in every possible contingency in which he may need to act.

The sequences of strategy profiles that form a Nash Equilibrium in the stage game, form a Nash Equilibrium in the dynamic game. Furthermore, the dynamic game allows for a much richer set of Nash Equilibrium. This is an advantage from the point of view of policy making or standardization. since having many equilibrium points gives more flexibility in obtaining a fair and efficient resource allocation. The following theorem gives a sufficient condition for the rate vector $(R_1, \ldots, R_M)$ to be achievable as the resulting utilities in a Nash Equilibrium of the repeated game [84] [85].

- *Theorem 2:* Let R_i^{FS} be the rate of system i when all the systems spread their power over the bandwidth W, i.e. the rate obtained in the full-spread Nash Equilibrium. There exists a sub-game perfect Nash Equilibrium of the dynamic GIG with utilities $(U_1, \ldots, U_M) = (R_1, \ldots, R_M)$ whenever $(R_1, \ldots, R_M) \in \mathcal{R}$ and $R_i > R_i^{\mathrm{FS}}$ for $i = 1, \ldots, M$ for a discount factor δ sufficiently close to 1.
- *Theorem 3:* The rate R_i^{FS} is the reservation utility of player i in the GIG. That is, player i can obtain a utility at least as large as R_i^{FS} by using the power allocation $p_i(f) = P_i/W, f \in [0, W]$ regardless of the power allocations used by the other players. Therefore, the rate R_i obtained by user i in any N.E. of the GIG must satisfy $R_i \geq R_i^{\mathrm{FS}}$. The same statement holds for the repeated GIG.

Let $\{p_i(f)\}_{i=1}^{M}$ be the power allocations that result in the rate vector $(R_1, \ldots, R_M)$ (which always exist since $(R_1, \ldots, R_M) \in \mathcal{R}$). The strategy that each system follows to obtain the rate vector $(R_1, \ldots, R_M)$ in Theorem 2 is the following trigger strategy:

- at $t = 1$: use power allocation $p_i(f)$.
- at $t = t_0$: if every user $j \in \{1, \ldots, M\}$ used the power allocation $p_j(f)$ at $t = t_0 - 1$, use $p_i(f)$. Otherwise spread the power over the total band, i.e. use the power allocation P_i/W for $f \in [0, W]$

The idea behind this strategy, is to "cooperate" by using the required power allocation as long as all the other systems cooperated in the previous stages. As soon as at least one system deviates from the "good" behavior, a punishment is triggered where all the other systems spread their powers forever. Since the rates obtained by the systems once the punishment is triggered are lower than those obtained with cooperation, it is in the system's own interest to cooperate. Friedman's analysis shows that if δ is not too small, the above set of strategies forms a sub-game perfect Nash Equilibrium. The sub-game perfection property of the N.E. guarantees that each system will indeed apply the punishment once the punishing situation arises. This property makes the threats believable.

An immediate consequence of Theorems 2 and 3 is that if the desired operating point $(R_1, \ldots, R_M)$ (i.e. the maximizer of a desired global utility) is component-wise greater than the spreading rate vector $(R_1^{\text{FS}}, \ldots, R_M^{\text{FS}})$ there is no performance loss due to lack of cooperation. However, when this condition is not satisfied, the best that one can do is to find the point $(R_1, \ldots, R_M) \in \mathcal{R}^*$ that maximizes the global utility subject to $(R_1, \ldots, R_M) \geq (R_1^{\text{FS}}, \ldots, R_M^{\text{FS}})$.
[Example Scenarios]
Applying these ideas to the two system scenario I, we can define a trigger strategy where system 1 uses the first half of the bandwidth, and system 2 uses the second half, as long as in all the previous stages both systems complied with this frequency allocation. If at some stage any of the systems stops complying, a punishment is triggered where the systems spread their powers forever. For large enough P this pair of strategies forms a N.E. where each system obtains a utility $1/2 \log(1 + 2P)$. This shows how the punishment strategies within the dynamic game formulation allow us to overcome the inefficiency that we observed in the static game.

Consider two other scenarios shown in Figures 6.19 (a) and (b) we assume in both cases that $c_{1,1} = c_{2,2} = 1$, $W = 1$, and $N_0 = 1$. For scenario II, we set $P_1 = 10$, $P_2 = 1$, and $c_{1,2} = c_{2,1} = 1.1$ In scenario III, we set $P_1 = 10$, $P_2 = 1$, $c_{1,2} = 0.5$ and $c2, 1 = 10$.

As shown in Figure 11.1, the optimal sum rate point of scenario II lies within the achievable region in the non-cooperative setting. However, the optimal proportional fair point lies outside of this set and cannot be supported without cooperation. The best that one can do in the noncooperative setting is to operate in the point indicated in the figure. In scenario III both the optimal sum rate and optimal proportional fair rates are achievable in the non-cooperative setting. Note that while in the cooperative case the specific values of the cross gains had no influence on the achievable region (as long as the strong interference condition is satisfied) this is not true in the non-cooperative setting. This is because large cross gains enable the systems to apply punishments, and hence achieve a good N.E. through believable threats. In scenario III, the large value of $c_{2,1}$ allows system 2 to punish system 1 whenever it departs from the proportional fair allocation.

Consider other two user scenario with $c_{1,1} = c_{2,2} = 1$, $W = 1$, $N_0 = 1$, and $P_1 = P_2 = P$ At a given SNR, channel asymmetry can be adjusted through the changes in the cross gains $c_{1,2}$ and $c_{2,1}$. Here, the proportional fair utility U_{PF} is considered to measure the global performance.

For a fixed set of parameters, the power allocations is optimized to maximize the proportional fair metric, obtaining R_1^* and R_2^* as the resulting rates. In the non-cooperative scenario, R_1^* and R_2^* can only be supported by a Nash Equilibrium if $R_1^* \geq R_1^{\text{FS}}$ and $R_2^* \geq R_2^{\text{FS}}$. If these inequalities are not satisfied, the best possible solution is obtained for the non-cooperative case by maximizing $\log(R_1) + \log(R_2)$ subject to the constraint $R_i \geq R_i^{\text{FS}}, i = 1, 2$, being $\tilde{R}_1$ and $\tilde{R}_2$ the corresponding optimal rates. If $R_i^* = \tilde{R}_i$ for

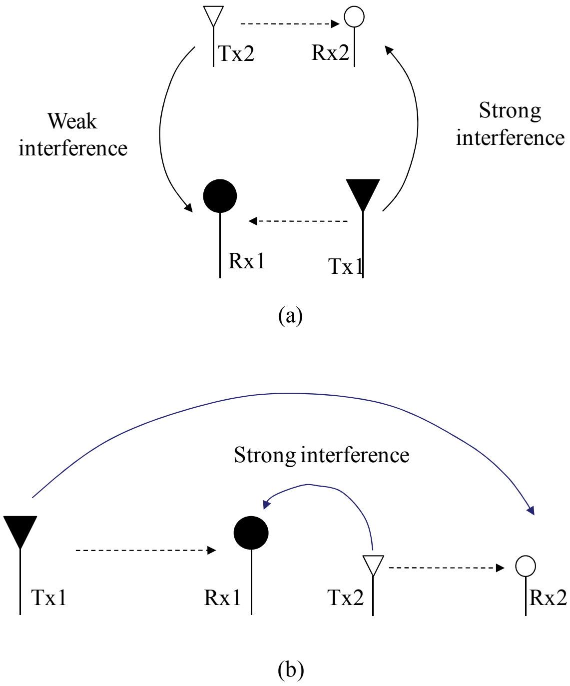


Figure 6.19 Asymmetric situations between two systems: (a) scenarios II: asymmetry in Tx powers, and (b) scenarios III: asymmetry in Tx powers and gains [70].

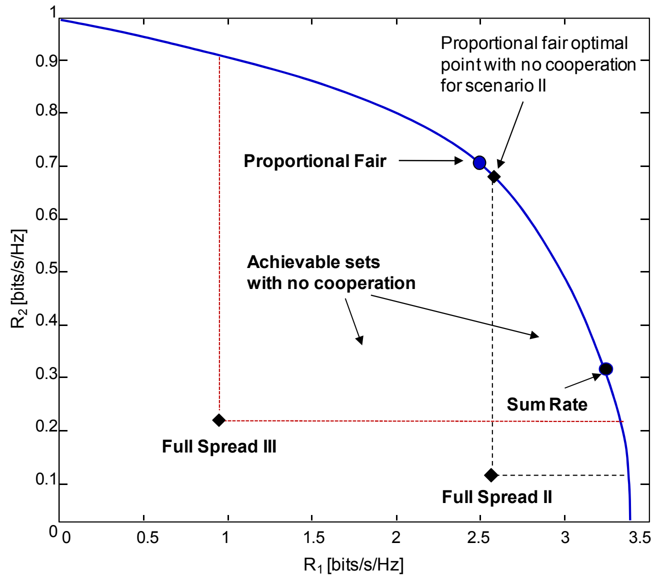
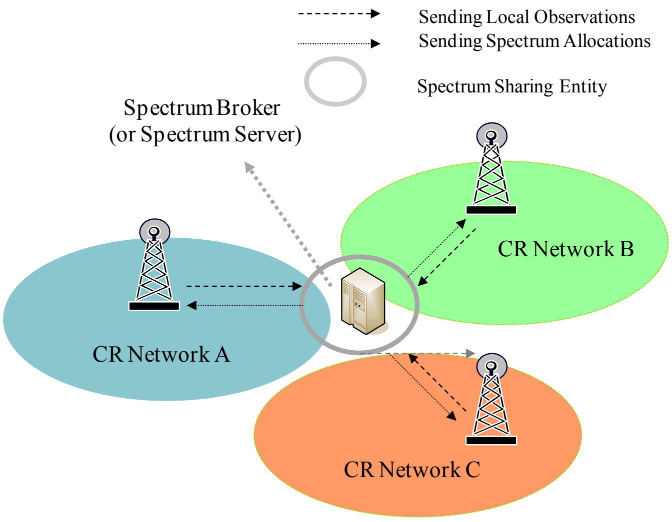
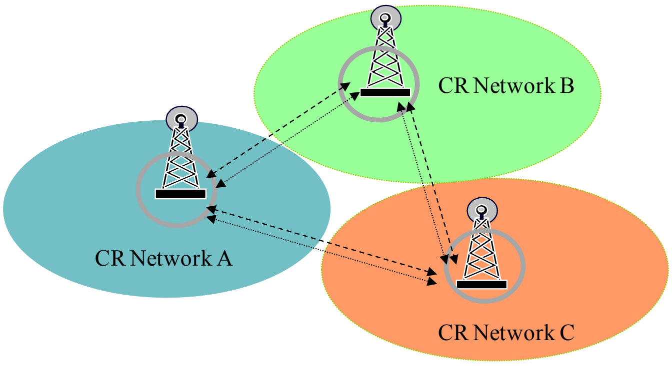


Figure 6.20 Achievable rates with no cooperation for scenarios II and III [70].

$i = 1, 2$, Consequently, there is no loss due to lack of cooperation. If $R_i^* > \tilde{R}_i$ for $i = 1$ or $i = 2$, the loss is measured due to lack of cooperation using $\max_{i \in \{1,2\}} 100(R_i - \tilde{R}_i)/R_i$,



(a)



(b)

Figure 6.21 Inter-network spectrum sharing: (a) centralized spectrum sharing, and (b) distributed spectrum sharing.

i.e. the percentage loss in rate for one of the systems. Note that the other system will have a rate larger than the one obtained with cooperation.

6.4 INTER-NETWORK SPECTRUM SHARING

CR networks are envisioned to provide opportunistic access to the licensed spectrum using unlicensed users. This setting enables multiple systems being deployed in overlapping locations and spectrum as shown in Figure 6.3. Hence, spectrum sharing among these systems is an important research topic in CR networks. Up to date, inter-network spectrum sharing has been regulated via static frequency assignment among different systems or centralized allocations between different access points of a system in cellular networks. In this section, we overview the recent work in this research area.

6.4.1 Centralized Inter-Network Spectrum Sharing

As explained in Chapter 2, the CR network consists of multiple networks or operators that compete the same spectrum resource in a certain area. Furthermore, each network may have different spectrum policies and radio access technologies, making it more complicated to design an efficient spectrum sharing scheme.. The most convenient way to address this problem is to use a spectrum manager that is in charge of resource allocation over multiple operators in a centralized manner. Furthermore, by introducing market mechanisms such as trading or auction, this approach provides economic incentives to both operators and central network entities. In this subsection, we present several centralized approaches for inter-network spectrum sharing.

SPECTRUM POLICY SERVER MODEL

In this work, a central spectrum policy server (SPS) is proposed to coordinate spectrum demands of multiple CR operators [122] [123]. In this scheme, each operator bids for the spectrum indicating the cost it will pay for the duration of the usage. The SPS then allocates the spectrum by maximizing its profit from these bids. The operators also determine an offer for the users and users select which operator to use for a given type of traffic. When compared to a case where each operator is assigned an equal share of the spectrum, the operator bidding scheme achieves higher throughput leading to higher revenue for the SPS, as well as a lower price for the users according to their requirements. This work opens a new perspective by incorporating competition for users as well as the spectrum in CR networks.

[Mathematical Model]

This work considers a limited geographical region for which spectrum management is under the control of a local SPS. Two operators, Operator 1 and Operator 2, provide services to a user within the specified region. Each operator has a number of base stations throughout the region. The SPS keeps track of the vacant spectrum W_A that would be available for usage of Operator 1 or Operator 2. The available bandwidth W_A is assumed to be finite. Here, W_A could be exclusively partially devoted to an operator, for it to offer service to the user. The allocation would be valid for the whole duration of the session established between the specified user and the operator. The proposed spectrum sharing method is based on an exclusive allocation in which no two sessions can occupy the same frequency band in the region. Based on this system model, the user acceptance model and the operator profit model are proposed, which govern the competitive spectrum allocation as follow:

- *Users' Acceptance Model:* In demand responsive pricing [9] [158], it is important to take into account the users' responses to the pricing strategy of the operator. From the user point of view, the service offer made by the operator is acceptable only if the price asked is reasonable. Specifically, this consideration is addressed by introducing an acceptance probability $A(u, P)$ where u is the utility of the user and P is the associated price. Intuitively, the acceptance probability $A(u, P)$ should have the following qualitative properties. It should be an increasing function of u for a fixed P while decreasing in P for fixed u. Mathematically, these properties are formulated as:

$$\frac{\partial A}{\partial U} \geq 0, \quad \frac{\partial A}{\partial P} \leq 0$$
$$\forall P > 0, \quad \lim_{u \to 0} A(u, P) = 0$$
$$\lim_{u \to \infty} A(u, P) = 1 \tag{6.68}$$
$$\forall P > 0, \quad \lim_{u \to 0} A(u, P) = 1$$
$$\lim_{u \to \infty} A(u, P) = 0$$

While there are several candidate choices for the function $A(u, P)$, we will follow [9] [158] and choose

$$A(u, P) = 1 - e^{Cu^{\mu} P^{-\epsilon}} \tag{6.69}$$

where μ is the utility sensitivity of the user, B is the price sensitivity, and C is an appropriate constant. Note that the acceptance probability function can be differentiated among users through the above parameters.

In this work, for simplicity, the role of transmit power in the user utility is ignored, and u is instead parameterized as a function of offered rate R only. The specific model for the utility function chosen here is one that obeys a law of diminishing returns such as [9] [158] [252]:

$$u(R) = \frac{(R/K)^{\zeta}}{1 + (R/K)^{\zeta}} \tag{6.70}$$

where K and ζ are parameters that determine the exact shape of the above sigmoid function. Note that the above expression gives normalized utility values in the interval $[0, 1)$ with the rate $R = K$ yielding a utility of $1/2$.

As $u(R)$ is a function of R, we simplify the notation by representing the acceptance probability as $A(R, P)$ in the rest of the subsection. As expected, the acceptance probability is decreasing with price.

- *Operator Profit Model:*

  From the operator's point of view, the service is worth providing only if the achievable revenue is high enough to compensate for the costs associated with providing the service. The two operators considered in the model, Operator 1 and Operator 2, are able to provide spectral efficiencies of r_i [bps/Hz] to a specified user, where $i \in 1, 2$ is the index denoting the operator. The spectral efficiency may depend on various parameters like the technology used by the operator, the density of the base stations belonging to the operator in the considered geographical region, and the location of the user. For a specific rate offered R_i, the bandwidth required $W_i(R_i)$ by the operator is inversely proportional to the spectrum efficiency: $W_i(R_i) = R_i/r_i$.

  For the offered rate R_i and price P_i, the profit $Q_i(R_i, P_i)$ can be expressed as:

$$Q_i(R_i, P_i) = P_i - F_i - V_i R_i / r_i \tag{6.71}$$

  where F_i [\$] is the fixed cost incurred by the operator, and V_i [\$/Hz] is the price per unit bandwidth that the SPS charges Operator i. The last term denotes the usage-based variable cost for the operator. Note that in most cases, the fixed cost F_i is implicitly related to the efficiency r_i. One would expect operators with higher fixed cost to be able to sustain greater efficiencies resulting from superior infrastructure [134]. Considering the user's acceptance probability, the expected profit for Operator i is

$$Q_i(R_i, P_i) = A(R_i, P_i)Q_i(R_i, P_i) \tag{6.72}$$

Note that for fixed r_i, the acceptance probability $A(R_i, P_i)$ is increasing in R_i and decreasing in Pi while the profit $Q_i(R_i, P_i)$ is decreasing in R_i and increasing in P_i.

Based on these model, operator interactions can be modeled as an noncooperative game. The user response to an offer (R, P) is modelled as in Eq. (6.69). If two offers (R_1, P_1) and (R_2, P_2) are made by Operators 1 and 2 respectively, the offer for which $A(R, P)$ is lower is ignored by the user. The other offer is then accepted with the associated acceptance probability. When the offers made invoke equal acceptance probabilities, $A(R_1, P_1) = A(R_2, P_2)$, we assume that each offer is equally likely to be accepted.

In the context of operators competing for resources and the user preference, the game can be represented by $G = [N, \{S_i\}, \beta_i]$ where $N = 1, 2$ is the index set of the players (operators), S_i is the strategy space available to Operator i, and $\beta_i(\cdot)$ is the resulting expected profit associated with the operator with index i. The strategy space S_i for Operator i consists of all (R, P) pairs which satisfy the bandwidth constraint:

$$S_i = \{\forall(R, P)|F_i + V_i R_i/r_i \le P \le P_{\max}, 0 \le R \le W_A \times r_i\} \tag{6.73}$$

where $P_{\max}$ is the maximum price the operator is permitted to charge.

The resulting expected profit β_i of operator i given the strategy of the opponent operator j is

$$\beta_i(R_i, P_i, R_j, P_j) = \begin{cases} 0 & \text{if } A(R_i, P_i) < A(R_j, P_j) \\ \frac{1}{2}\bar{Q}_i(R_i, P_i) & \text{if } A(R_i, P_i) = A(R_j, P_j) \\ \bar{Q}_i(R_i, P_i) & \text{if } A(R_i, P_i) > A(R_j, P_j) \end{cases} \tag{6.74}$$

The non-cooperative operator game can now be formally stated as

$$\max_{(R_i, P_i)\in S_i} \beta_i(R_i, P_i, R_j, P_j) \;\; i \in \{1, 2\} \tag{6.75}$$

The above game has the following theorem:
Theorem: At any Nash equilibrium for the game G, at least one of the operators has zero expected profit.

Proof: By contradiction: Assume there exist the equilibrium strategies (R_1^*, P_1^*) and (R_2^*, P_2^*) for which $\beta_1(\cdot) > 0$ and $\beta_2(\cdot) > 0$. Considering Eq. (6.74), the only way this can be achieved is to have equality between the achieved acceptance equalities; $A(R_1^*, P_1^*) = A(R_2^*, P_2^*)$. In this situation, in accordance with (7), the corresponding payoffs would be $\beta_1(R_1^*, P_1^*) = \frac{1}{2}\bar{Q}_1(R_1^*, P_1^*)$ and $\beta_2(R_2^*, P_2^*) = \frac{1}{2}\bar{Q}_2(R_2^*, P_2^*)$. Note that the assumption of non-zero profits implies that $\frac{1}{2}\bar{Q}_1(R_1^*, P_1^*) > 0$ and $\frac{1}{2}\bar{Q}_2(R_2^*, P_2^*) > 0$. Consider Operator 1 without loss of generality. If Operator 1 were now to deviate from the strategy (R_1^*, P_1^*) to $(R_1^*, P_1^* - \Delta_P)$ by lowering its price offer by an infinitesimal amount

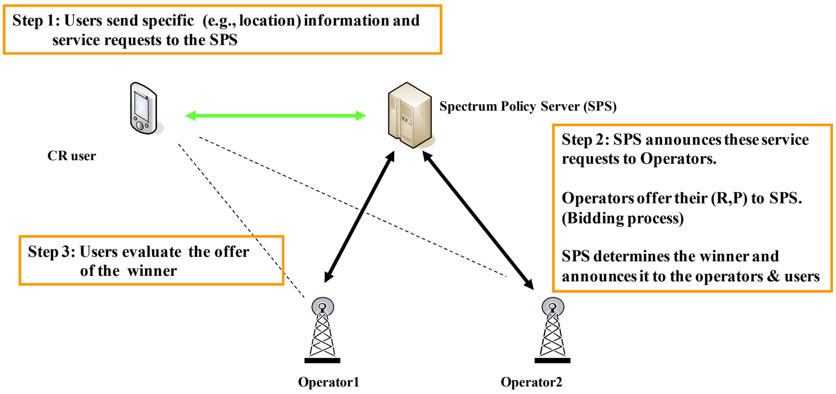


Figure 6.22 Iterative bidding [122].

Δ_P, then it follows that $A(R_1^*, P_1^* - \Delta_P)$ is greater than $A(R_2^*, P_2^*)$.

Further, from Eq. (6.74) it follows that the resulting expected profit for Operator 1 is $\bar{Q}_1(R_1^*, P_1^* - \Delta_P)$. By continuity of the profit function, it follows that $|\bar{Q}_1(R_1^*, P_1^* - \Delta_P) - \bar{Q}_1(R_1^*, P_1^*) < \delta$ for arbitrarily small $\delta > 0$. We can thus bound the change in payoff of Operator 1, i.e., $\bar{Q}_1(R_1^*, P_1^* - \Delta_P) - \frac{1}{2}\bar{Q}_1(R_1^*, P_1^*)$ as $\bar{Q}_1(R_1^*, P_1^*) - \delta \bar{Q}_1(R_1^*, P_1^* - \Delta_P) - \frac{1}{2}\bar{Q}_1(R_1^*, P_1^*) < \frac{1}{2}\bar{Q}_1(R_1^*, P_1^*) + \delta$. Given that $\frac{1}{2}\bar{Q}_1(R_1^*, P_1^*) > 0$ and δ is arbitrarily small, it follows that the change in payoff for Operator 1 is strictly positive. Therefore the strategy (R_1^*, P_1^*) can never be the best response of Operator 1. This contradicts the initial assumption that at equilibrium $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are greater than zero.

[SPS as a Mediator in Iterative Bidding]
In this work, an iterative bidding process is proposed to implement the operator game G. The SPS mediates the bidding process on behalf of the user. Such an SPS based scheme is more practical as it reduces the amount of overhead and control information transmission to and from the user. The scheme is composed of three steps (Figure 6.22):

- Step 1: A new user gets connected to the SPS. User specific information, e.g., $A(R, P), r_1, r_2$, is communicated to the SPS.
- Step 2: The iterative bidding process between the operators is undergone and the winning operator is declared by the SPS.
- Step 3: The winning operator offers the winning bid $(R_{\text{winner}}, P_{\text{winner}})$. The user decides to accept the service with probability $A(R_{\text{winner}}, P_{\text{winner}})$.

Note that at the end of Step 1, the SPS has all the relevant information regarding the user so it can act on the user's behalf. Consequently, in Step 2, during the iterative bidding, only the SPS and the operators are involved. In Step 3, the user makes the final decision whether or not to take the service offer of the winning operator.

In the iterative bidding process, the operators make offers in each iteration. The strategy of each operator is to make the offer such that $A(R, P)$ associated with its offer is greater than the one associated with its opponent's offer while simultaneously maximizing the resulting expected profit. The iterative bidding is initialized by allowing the operators to

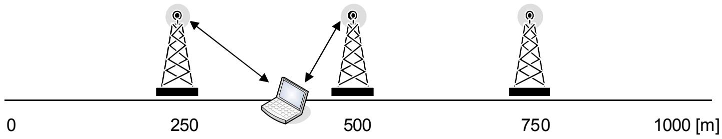


Figure 6.23 Geographical region with a single user and two operator [122].

choose their service offers without consideration of the opponent strategy. It is clear from the structure of $\beta_i(\cdot)$ in Eq. (6.69) that the iteration process is terminated when a zero value for expected profit is declared by at least one operator. The opportunity to offer service to the user is then given to the operator that wins. The winning operator uses its most recent bid $(R_{\text{winner}}, P_{\text{winner}})$, as a service offering to the user. Note that from Theorem 1, the iterative bidding process by definition should converge to a Nash equilibrium of the game G. If both operators declare zero expected profit at the same iteration, both are dismissed. This degenerate situation can happen when both operators have identical fixed costs and the user is located in a geographical location where the spectral efficiencies of both operators are identical. Such an operating point is also a Nash equilibrium. In such a case, we assume that the SPS randomly selects one of them to offer service.

It is assumed that the offers made to any user are final and the operators can not update any offers they have made to a specific user after the competition is over. The winning operator is obliged to provide the transmission rates they have offered. This is considered as part of the regulations enforced on the operators.

[Numerical Results]
Here, we show numerical results that correspond to a linear geographical region with two operators.

- *Single User Systems:* The system and the base station locations are depicted in Figure 6.23. Assume that both operators use the same technology, with the only difference being in the infrastructure density. Operator 1 has two base stations while Operator 2 only one. Consequently, the associated fixed cost for Operator 1 will be twice the fixed cost of Operator 2, i.e., $F_1 = 2F_2$. Note that the fixed cost per base station is the same. It is also assumed that the SPS will be charging both operators at the same variable cost rate V [/Hz]. The spectral efficiency between base station k and the user's mobile terminal is determined as

$$r_k = \log_2 \left[1 + \frac{P_s}{N_0} (\frac{d_k}{L/4})^{-2} \right] \tag{6.76}$$

  where P_s is the signal power, N_o is the AWGN variance, d_k is the distance between the base station k and the terminal, and L is the total length of the linear region in Figure 3 ($L = 1000m$). P_s is set to $2N_o$, which guarantees a SNR = 3 dB at the distance of $L/4 = 250$ m from the base station. Note that Operator 1 always selects a base station that provides higher spectral efficiency to serve the user (i.e., the base station that is closer to the user's mobile terminal).

  The available bandwidth is $W_A = 10$ MHz, and the user acceptance probability and the corresponding parameters are set to $K = 510^6$, $\zeta = 10$, $C = 1$, $B = 4$, $\mu = 4$.

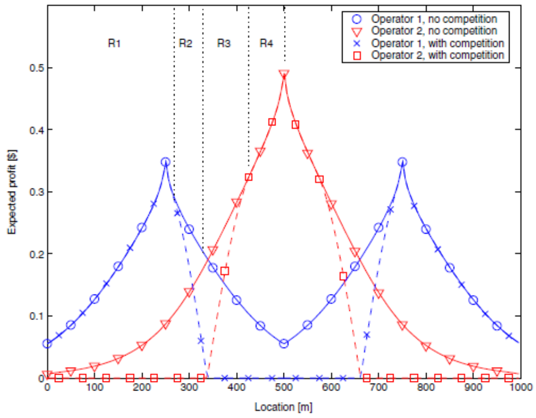


Figure 6.24 The expected profit versus the location of the user within the linear region for $\eta = 1 \times 10.6$, $W_A = 10$ MHz, and $F + VW_A = 2\$$ [122].

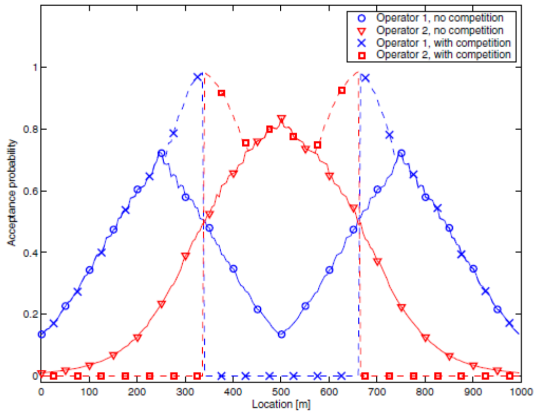


Figure 6.25 The acceptance probability versus the location of the user within the linear region for $\eta = 1 \times 10.6$, $W_A = 10$ MHz, and $F + VW_A = 2\$$ [122].

The cost structure is characterized with the ratio $\eta = V/F$ between the variable cost V [$ /Hz] versus the fixed cost per base station $F = F1/2 = F2$ [$]. Lower values of η correspond to the spectrum being less expensive than the infrastructure. Furthermore the absolute values for F and V are selected such that $F + VW_A = 2$, while $F_1 = 2F$ and $F_2 = F$.

Figure 6.24 shows the expected profit versus the location of the user within the linear region for $\eta = 110^{-6}$. The solid lines correspond to the case when only one of the operators is present. In that case the operator is offering the price and rate such that its expected profit is maximized without a competitor being present. The dashed line corresponds to the case when both operators are present and do compete for the user (as described in the previous sections). From these results, it is shown that the case with no competition provides an upper bound on expected profit for the case with competition. The corresponding acceptance probability is presented in Figure 6.25. Furthermore, depending on the user's location, the following behavior is observed. In Region 1 (denoted as R1 in Figure 6.24), the spectral efficiency of Operator 1 is much higher than that of Operator 2. Consequently Operator 1 is superior in the given region and can drive its expected profit up to the upper bound (case of no competition). In Region 2, the superiority of Operator 1 is diminishing and it is forced to lower its expected profit and out-compete Operator 2. In Region 3, Operator 2 is winning the competition (due to higher spectral efficiency while lowering its expected profit to become more competitive). In Region 4 the user is very close to the Operator 2 base station and its spectral efficiency is much higher than that of Operator 1. Now Operator 2 becomes superior in the given region and can drive its expected profit up to the upper bound (case of no competition). Beyond $L/2$, the situation is symmetric to the discussed regions. Complementary analysis can be presented for the acceptance probability versus the user location.

- *Multiuser System:* In this case, two operators are competing for spectrum and a number N of users with $N > 1$. The operators compete for each user individually. For each user, the SPS mediates the operator competition in accordance with the approach. They assume that as the result of the competition each user chooses a single operator as the service provider and further accepts the service with a certain acceptance probability.

  They further assume that the sessions corresponding to the user-operator pairs are held in nonoverlapping spectrum portions thus leading to interference free transmission. The total available bandwidth W_A is partitioned among these sessions by the SPS as will be described later in the text. It is crucial that the total available bandwidth W_A is sufficient to support all the winning offers. Note that operators, if not assisted by the SPS, can not keep track of the winning bids and their bandwidth requirements for each user, and thus can end up making unrealizable offers.

  Thus, in keeping with the spectrum server role of the SPS, this work proposes an SPS-based resource allocation scheme in which the SPS sets the upper limits on bandwidth usage for each user-operator session. The SPS determines these limits in the context of an optimization problem where it maximizes its expected revenue which is the sum of the expected payments of the operators for their spectrum utilization.

  In the proposed scheme, the SPS maximizes its expected revenue $R_{\text{SPS}}(\cdot, \mathbf{W})$ with respect to bandwidth allocation vector $\mathbf{W} = [W_1, W_2, \ldots, W_N]^T$. The operators competing for user n are subject to the constraint that they must not make offers that require bandwidths greater than W_n. The SPS maximizes its expected revenue subject to the constraint that the total allocated bandwidth does not exceed the total available bandwidth W_A. Consequently, the SPS optimization problem can be expressed as:

$$\begin{aligned} &\max_{W} R_{\text{SPS}}(\cdot, \mathbf{W}) \\ &\text{subject to :} \\ &\sum_{n=1}^{N} W_n \leq W_A \end{aligned} \tag{6.77}$$

  Note that the expected revenue $R_{\text{SPS}}(\cdot, \mathbf{W}))$ is defined as the sum of the expected bandwidth utilizations of the users scaled by the variable cost per bandwidth V [/Hz].

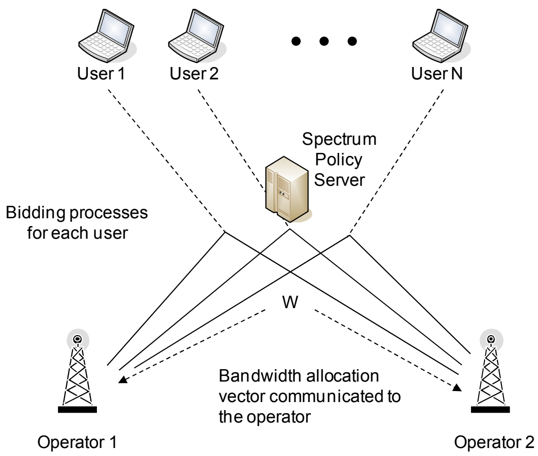


Figure 6.26 SPS mediating iterative bidding processes for N users [122].

In this sense, it is a function of the bandwidth allocation vector W as well as the user locations and the cost parameters in the system:

$$R_{\text{SPS}}(r_1, r_2, F_1, F_2, V, \mathbf{W}) = \sum_{n=1}^{N} V A_n^f(\cdot, \mathbf{W}) W_n^f(\cdot, \mathbf{W}) \tag{6.78}$$

In the above equation, r_1, r_2 are the N dimensional spectral efficiency vectors for Operator 1 and Operator 2, respectively. Each element of the vectors denote the service spectral efficiency the operators enjoy while providing service to a specified user. Note that these vectors depend on the exact locations of the users. F_1 and F_2 denote the fixed costs of Operator 1 and Operator 2, respectively. A_n^f and W_n^f refer to the winning bid acceptance probability and bandwidth usage achieved as a result of the operator competition over user n. W_n^f depends on the winning rate offer and the winning operator's spectral efficiency through the relation $W_n^f = R_{\text{winner}}/r_{\text{winner}}$.

When maximizing the expected revenue over $R_{\text{SPS}}(\cdot, \mathbf{W})$, the SPS performs a centralized optimization whose result is a vector $\mathbf{W}^*$. that maximizes the total expected revenue. In order to determine $\mathbf{W}^*$, the SPS performs an exhaustive search in which it declares all possible $\mathbf{W}_S$ one at a time. For any declared allocation vector $\mathbf{W}$, the operators compete with each other considering the bandwidth constraints imposed by the allocation vector, as illustrated in Figure 6.26. The SPS then computes the resulting $R_{\text{SPS}}(\cdot, \mathbf{W})$ given the competition results for each user. It then selects the vector which achieves the greatest revenue among all $\mathbf{W}_S$ as the optimum allocation vector $\mathbf{W}^*$.

It is interesting to note that the SPS optimization in Eq. (6.77) is equivalent to the maximization of the expected bandwidth utilization $\sum_{n=1}^{N} A_n^f W_n^f$. This demonstrates that a revenue seeking SPS will actually be maximizing the total bandwidth utilization in the system.

Note that, besides the expected SPS revenue $R_{\text{SPS}}(\cdot, \mathbf{W})$, the SPS can have a number of different criteria for determining the optimum bandwidth allocation vector. It could, for instance maximize the average user acceptance probability, as a social goal, or it could try to maximize the total expected profits of the operators.

SPECTRUM ALLOCATION AND SERVICE PRICING THROUGH AUCTIONS

Currently, each provider gets a chunk of the spectrum and has a unique user pool that they cater to. In future, a paradigm shift is very likely to occur where each provider will get a part of the spectrum from the common spectrum pool as and when they need through a spectrum broker. It is also anticipated that the concept of service broker, technically known as Mobile Virtual Network Operators (MVNO), will evolve that will act as an interface between the providers and the users [235]. The users will be able to select their service provider as per their requirements through the service broker. In light of these new developments, it is important to investigate the economic issues that has a profound impact on the service quality and the prices paid by the end users.

The most important factors that the wireless service providers (WSPs) need to consider are the amount of spectrum they need and the price are they willing to pay. In effect, estimation of the demand for bandwidth and expected revenue will drive the provider's strategies. Service pricing by the providers, in turn, will affect the demand for the services by the users, thus resulting in a cyclic dependency in a typical supply-demand scenario. As a result, the relationship between spectrum owner and WSP has a strong correlation with the relationship between WSPs and end-users and must be analyzed together unlike any other industry service model.

In this work two main components of the overall spectrum trading system are introduced from an economic point of view: i) spectrum allocation to WSPs and ii) interaction of end users with the WSPs [234]. For this two-tier trading system, this work presents a winner determining sealed-bid knapsack auction mechanism that dynamically allocates spectrum to the WSPs based on their bids. Furthermore, a dynamic pricing strategy based on game theory is proposed to capture the conflict of interest between WSPs and end users, both of whom try to maximize their respective net utilities.

Auction-based Spectrum Allocation

Spectrum can be allocated from the coordinated access band (CAB) in two ways: asynchronous or synchronous allocation. In asynchronous method, whenever a service provider has a need for spectrum, it makes a request to the spectrum broker. If available, the spectrum broker assigns a chunk of spectrum for the lease period, upon expiry of which, the assigned spectrum is taken back. On the other hand in synchronous allocation, spectrum allocations (and de-allocations) are done in a synchronous manner i.e., providers make requests synchronously. The lease periods can be assumed as discrete unit short span of intervals.

This work focuses on the synchronous allocation. Similarly, pricing can be done in two ways depending on the total demand of spectrum from the service providers. If the total demand of spectrum does not exceed the spectrum available in CAB, then any of the following two pricing models can be adopted:

- *Service provider dominant strategy:* Providers advertise the price they are willing to pay.
- *Spectrum broker dominant strategy:* Spectrum broker advertises a (unit) price and service providers respond by deciding on the amount of spectrum they can acquire.

On the other hand, if the total demand of spectrum exceeds the total spectrum available in the CAB (which will be very often the case and is thus the focus of our research), then one of the strategies for the spectrum broker is to put up the spectrum for bids and decide on the allocation based on the bids i.e., to adopt an auction model. The auction for spectrum can be conducted on a periodic basis and on a small time granularity so that wireless service providers will bid for additional spectrum from CAB synchronously as this would allow the spectrum broker to compare all the requests to maximize the revenue. The assumption in this model is that service providers generates spectrum requests periodically at the beginning of each interval.

A good auction design is important for any type of successful auction and often varies depending on the item on which the auction is held. Unlike classical single-unit auctions, spectrum auctions are multi-unit where bidders bid for a part of the spectrum band, i.e., the bids are for different amounts of bandwidth. Also, multiple winners evolve constituting a winner set. Thus, determination of winner set depends heavily on the auction strategy adopted. The spectrum broker is the seller who owns the coordinated access band and service providers are the buyers/bidders. For In the proposed auction model, three important design issues on auction are considered: 1) how to maximize the revenue generated from bidders, 2) how to entice bidders by increasing their probability of winning, and 3) how to prevent collusion among providers.

[Formulation of Auction Rules]
Consider L WSPs (bidders) who compete for a total spectrum W. All the service providers submit their demands at the same time in a sealed bid manner since sealed bid auction has shown to perform well in all-at-a-time bidding and has a tendency to prevent collusion [224]. Each service provider has knowledge about its own bidding quantity and bidding price but do not have knowledge about other's quantity and price.

This auction procedure is formulated as follows. Let a tuple $q_i = \{w_i, x_i\}$ be the strategy adopted by service provider i where w_i denotes the amount of spectrum requested and xi denotes the corresponding price that the service provider is willing to pay. If the sum of the bidding quantities do not exceed the spectrum available, W, then the requested quantities are allocated. Otherwise, auction is initiated when,

$$\sum_{i=1}^{L} w_i > W \tag{6.79}$$

This work aims at solving the winner-determination problem in such a way so that the spectrum broker maximizes revenue by choosing a bundle of bidders (q_i), subject to condition that the total spectrum allocated does not exceed W, i.e.,

$$\begin{gathered} \text{Maximize } \sum_i x_i \\ \text{subject to} \\ \sum_{i=1}^{L} w_i > W \end{gathered} \tag{6.80}$$

[Bidders' Strategies]
Here, bidders' strategies is investigated for both first and second price bidding schemes under knapsack model. In first price auction, bidder(s) with the winning bid(s) pay their

winning bid(s) while in second price, bidder(s) with the winning bid(s) do not pay their winning bid but pay the second highest bid. Assume that each bidder i submits its demand tuple q_i. Then the optimal allocation of spectrum is done by considering all the demand tuples. This optimal allocation be denoted by M, where M incorporates all the winning demand tuples q_i and is subject to condition given in Eq. (6.80). Assume that bids can take only integer values (as bids in dollar values are always expressed as integer) and number of bidders (providers) is typically of the order of 10. If the number of bidders is large, we use the scaling heuristic. Thus, the winner determination problem can be solved through dynamic programming with reasonably low computation. The aggregate bid can be obtained by summing all the bids from bidder,

$$\sum_{i \in M} x_i \tag{6.81}$$

Consider a particular bidder j who was allocated spectrum and thus belongs to M. Then the aggregate bid generated from the optimal allocation M minus the bid of bidder j is given by

$$\sum_{i \neq j, i \in M, j \in M} x_i \tag{6.82}$$

Now consider that bidder j does not exist and the auction is among the remaining $L - 1$ bidders. Let the optimal allocation be denoted by M^*. The aggregate bid generated in this case is

$$\sum_{i \neq j, i \in M^*, j \notin M^*} x_i \tag{6.83}$$

Therefore, minimum winning bid of bidder j must be at least greater than

$$X_j = \sum_{i \neq j, i \in M^*, j \notin M^*} x_i - \sum_{i \neq j, i \in M, j \in M} x_i \tag{6.84}$$

Thus, bidder j's request is granted if $x_j > X_j$ and not granted if $x_j < X_j$. If $x_j = X_j$, bidder j is indifferent between winning and loosing. Note that, the model under consideration is a non-uniform-price auction and X_j is not generally the same for all bidders. Though Eq. (6.84) gives the winning bid for bidder j, it is not necessary that bidder j will be able to afford it. There exists a price threshold (bidder's reservation price) beyond which a bidder is simply unwilling to pay.

Bidder's reservation price is defined as the most a bidder would be willing to pay. When a service provider buys spectrum from the spectrum broker, the service provider needs to sell that spectrum in form of services to the end users who pay for these services. The revenue thus generated helps the provider to pay for the fixed (static) cost for the statically assigned spectrum and the extra spectrum that the provider might need from the CAB. If the total revenue generated from the users is R and R_{static} goes towards the fixed cost, then the difference, R_{dynamic}, is the maximum amount that the provider can afford for the extra spectrum from CAB i.e.,

$$R_{\text{dynamic}} = R - R_{\text{static}} \tag{6.85}$$

Note, R_{dynamic} is not the bidder's reservation price but is a prime factor that governs this reservation price.
Lemma 1: In the second price knapsack auction, dominant strategy of the bidder is to bid bidder's reservation price.

[Proof:] Let us assume jth bidder has the demand tuple $q_j = \{w_j, x_j\}$ and its reservation price for that amount of spectrum requested be r_j . Now, as shown above in Eq. (6.84), jth bidder's request will be granted and consequently belong to optimal allocation M, only if bid generated by jth bidder is at least X_j . Then according to the second price bidding policy, jth bidder will pay the second price which is X_j in this case. Then the payoff obtained by jth bidder is,

$$E_j = r_j - X_j \tag{6.86}$$

Through proof by contradiction, it is shown that jth bidder's true bid is its reservation price r_j. Assume that jth bidder does not bid its true evaluation of the spectrum requested, i.e., $x_j \neq r_j$. Accordingly bidder j has two options of choosing x_j.

- *Option 1:* Bid is less than the reservation price, i.e., $x_j < r_j$. The values of x_j, r_j and X_j are such that,

    1. $r_j > x_j > X_j$, then bidder j falls inside the optimal allocation M and its request is granted. The expected payoff obtained by jth bidder is still given by: $(r_j.X_j)$.
    2. $r_j > X_j > x_j$, then bidder j loses and its request is not granted. Accordingly, the expected payoff becomes 0.
    3. $X_j > r_j > x_j$, bidder j still loses and the expected payoff is again 0.

- *Option 2:* Bid is more than the reservation price, i.e., $x_j > r_j$. The values of x_j , r_j and X_j are such that,

    1. $x_j > r_j > X_j$, then bidder j falls inside the optimal allocation M and its request is granted. The expected payoff obtained by jth bidder is still given by: $(r_j - X_j)$.
    2. $x_j > X_j > r_j$, though bidder j wins but the expected payoff becomes negative in this case. The expected payoff obtained by jth bidder is given by: $(r_j - X_j) < 0$. Bidder j will not be interested in this scenario.
    3. $X_j > x_j > r_j$, bidder j loses and the expected payoff is again 0.

It is evident that if bidder j wins, then the maximum expected payoff is given by $Ej = r_j - X_j$ and bidding any other price (higher or lower) than its reservation price r_j will not increase payoff. Thus, the dominant strategy of a bidder in second price bidding under knapsack model is to bid its reservation price. □

Lemma 2: In the first price bidding, reservation price is the upper bidding threshold.

[Proof:]Contrary to the Lemma 1, in first price bidding, the expected payoff obtained by jth bidder can be given by, $Ej = r_j - x_j$, as the actual price paid by the bidder is the same as the bid. Then, to increase the expected payoff, i.e., to keep $Ej > 0$, x_j must be less than r_j. Again at the same time, to win, bid x_j must be greater than X_j (Eq. (6)). Thus the dominant strategy for the bidder in first price auction is to bid less than the reservation price. □

[Service Provisioning Using Games]
Here. the most generic abstraction of "always greedy and profit seeking" model is considered, which exists between WSPs and end-users. The WSPs compete among themselves to provide service to a common pool of users. The resource for the WSPs are spectrum

chunks that have been statically allocated. Users on the other hand select service providers depending on the benefit they obtain for the prices they pay. The following are the conflict and decision models that arises between the WSPs and the users.

- *Conflict model:* This work considers the model where any user can access any WSP. The users are the potential buyers who buy services from the WSPs. The selection of a WSP is done on a dynamic basis i.e., a user compares the offerings both in terms of QoS and price for a particular service. Once a service is completed, the user relinquishes the radio resources. As the prices offered are not static, the users do not have any information about other users' strategies i.e., demand for resources or price willingness to pay. In such an incomplete information scenario, the benefit of a user depends not only on its own strategy but also on what others do. Since every user is assumed to be selfish, the problem is modeled as a non-cooperative game.

  Service providers, very much like the users, also act in their self-interest. As a seller of the services, they determine the price for its services depending on the amount of spectrum acquired and the price paid. Similar to the noncooperative incomplete information game among the users, the service providers also do not have any information about other providers' strategies, such as, price assigned for services, allocated resource, remaining resource, existing load, etc. Based on this conflict model, the decisions need to be defined.

- *Decision Model:* As a user, the decision problem is to select the best service provider for the session requested. Now the question arises, how to select the best service provider or rather what criteria determines the best. The quality of service perceived by a user in a network must be considered in this regard. As quality of service depends on the traffic load and the pricing strategies, we must therefore perform a cost benefit analysis to find the best service provider. A natural question that arises in such settings is the existence of an equilibrium where no user will find it beneficial to change the strategy unilaterally. This by definition is known as Nash equilibrium [190]. As a service provider, the decision problem is to advertise a price for a service without knowing what prices are being advertised by its competitors. The optimization is to find a price such that the provider is able to sustain profit in spite of offering a low price i.e., is there any price threshold to reach Nash equilibrium? For finding the existence of Nash equilibrium, the auction model defines the preference of the providers and users, given by their utility functions.

- *Utility Function:*
  An utility function is a mathematical characterization that represents the benefits and cost incurred. Here, the utility functions are defined for both WSP and users. We consider L service providers that cater to a common pool of N users. Let the price per unit of resource advertised by the service provider j, $1 \leq j \leq L$, at time t be $p_j(t)$. Let $b_{ij}(t)$ be the resource consumed by user i, $1 \leq i \leq N$, served by provider j. The total resource (capacity) of provider j is assumed to be C_j . The utility obtained by user i under the provider j can be given by [265]

$$u_{ij}(t) = a_{ij} \log(1 + b_{ij}(t)) \tag{6.87}$$

  where, the coefficient a_{ij} is a positive parameter that indicates the relative importance of benefit and acts as a weightage factor.

This log function reflects the intuition that the initial increase in the perceived throughput is more important to a user. Moreover, log function is analytically convenient, increasing, strictly concave and continuously differentiable.

The cost components incurred by user, is derived as follow: The first cost component is the direct cost paid to the provider for obtaining $b_{ij}(t)$ amount of resource. If $p_j(t)$ is the price per unit of resource, then the direct cost paid to the jth provider is given by $p_j(t)b_{ij}(t)$.

This direct cost component decreases user i's utility. Note that in Eq. (6.87), both price per unit resource and the resource amount requested are variables. The second cost component incurred by the user is the perceived quality of service, one of the manifestations of which is the queuing delay which again depends on the resources consumed by the other users. The queuing process is assumed to be $M/M/1$ at the links. Thus, the delay cost component can be written as

$$\begin{cases} \xi(\frac{1}{C_j - \sum_i^{N_j} b_{ij}(t)}) & \text{if } \sum_i^{N_j} b_{ij}(t) < C_j \\ \infty & \text{if } \sum_i^{N_j} b_{ij}(t) \geq C_j \end{cases} \tag{6.88}$$

where N_j is the number of users currently served by provider j and $\xi(\cdot)$ is a mapping cost function of delay. Combining all the components obtained in Eqs. (6.87) and (6.88)), we get the net utility as

$$U_{ij}(t) = u_{ij}(t).p_j(t)b_{ij}(t) - \xi(\frac{1}{C_j - \sum_i^{N_j} b_{ij}(t)}) \tag{6.89}$$

The utility of service provider j at time t is,

$$V_j(t) = p_j(t) \sum_i^{N_j} b_{ij}(t) - K_j \tag{6.90}$$

where, K_j is the cost incurred to provider j for maintaining network resources. Her this cost is assumed to be constant.

- *Price Threshold:*
  Consider user i has a certain resource demand and wants to connect to a provider at time t. All the providers advertise their price per unit of resource amount and the existing load. As user i wants to maximize its net utility (potential benefit minus cost incurred), it computes the resource vector that would maximize utilities from all the providers and the corresponding maximized utility vector.

User i would then connect to provider j if $U_{ij}(t)$ gives the maximum value in the maximized utility vector, $U_{i1}(t), U_{i2}(t), \ldots, U_{iL}(t)$, and $b_{ij}(t)$ is the requested resource amount from the optimal resource vector, $b_{i1}(t), b_{i2}(t), \ldots, b_{iL}(t)$.

To find the existence of any optimal amount of resource for the users and any pricing bound from the providers that will maximize the users net utility, net utility in

Eq. (6.89) is differentiating Eq. (6.89) with respect to $b_{ij}(t)$ as follow:

$$U'_{ij}(t) = \frac{a_{ij}}{1 + b_{ij}(t)} - pj(t) - \xi'(\frac{1}{C_j - \sum_i^{N_j} b_{ij}(t)}) \tag{6.91}$$

Similarly, the second derivative is

$$U''_{ij}(t) = -\frac{a_{ij}}{1 + b_{ij}(t)}^2 - -\xi''(\frac{1}{C_j - \sum_i^{N_j} b_{ij}(t)}) \tag{6.92}$$

If the auction considers delay and congestion component, such that, $\xi''(\frac{1}{C_j - \sum_i^{N_j} b_{ij}(t)}) > 0$, then, $U''_{ij}(t) < 0$ and it is clear that $U_{ij}(t)$ is strictly concave in the region bounded by $\sum_i^{N_j} b_{ij}(t) = C_j$; and $U_{ij}(t) \to -\infty$ as $\sum_i^{N_j} b_{ij}(t) \to C_j$. Moreover, it can be inferred from Eq. (6.92) that as $U''_{ij}(t) < 0$, $U_{ij}(t)$ contains a unique maximization point.

Thus, equating Eq. (6.91) to 0, and solving for $b_{ij}(t)$ gives the optimal amount of resources needed by the users for a certain price $p_j(t)$ and this resource amount will maximize the utility of the user. From the reverse point of view, it is also clear from the above Eq. (6.91) that there exists a maximum threshold for the price $p_j(t)$.

As the users are homogeneous, to maximize users' utility, first derivative of all the users can be equated to zero,

$$U'_{1j}(t) = U'_{2j}(t) = \ldots = U'_{N_j j}(t) = 0 \tag{6.93}$$

Recall, N_j is the number of users currently served by provider j. Thus Eq. (6.93) reduces to,

$$\frac{a_{1j}}{1 + b_{1j}(t)} = \frac{a_{2j}}{1 + b_{2j}(t)} = \ldots = \frac{a_{N_j j}}{1 + b_{N_j j}(t)} \tag{6.94}$$

If $1 + b_{ij}(t) = m_{ij}(t)$ and with the help of identity, we get,

$$\frac{a_{ij}}{m_{ij}(t)} = \frac{\sum_i^{N_j} a_{ij}}{\sum_i^{N_j} m_{ij}(t)} \tag{6.95}$$

For notational simplicity, we represent $a_{Ij} = \sum_i^{N_j} a_{ij}$ and $m_{Ij}(t) = \sum_i^{N_j} m_{ij}(t)$. Thus, Eq. (6.95) can be written as

$$\frac{a_{ij}}{m_{ij}(t)} = \frac{a_{Ij}}{m_{Ij}(t)} \tag{6.96}$$

Putting Eq. (6.96) into Eq. (6.91)), $U'_{ij}(t)$ can be obtained as

$$U'_{ij}(t) = \frac{a_{Ij}}{m_{Ij}(t)} - p_j(t) - \xi'(\frac{1}{C_j + N_j - m_{Ij}(t)}) \tag{6.97}$$

Note $U'_{ij}(t)$ is strictly decreasing with the values of $m_{Ij}(t)$ lying in the interval $(C_j, C_j + N_j)$. Then for achieving the Nash equilibrium by the providers, the pricing constraint $p_j(t)$ is upper bounded by,

$$\frac{a_{Ij}}{m_{Ij}(t)} - -\xi'(\frac{1}{C_j + N_j - m_{Ij}(t)}) \tag{6.98}$$

This pricing upper bound helps the provider to reach the Nash equilibrium. If all of the other providers and users keep their strategies unchanged, and a provider changes its strategy unilaterally and decides not to maintain its pricing upper bound, then that provider will not be able to maximize its users' utility and thus users will not connect to this provider decreasing provider's revenue.

AUCTION-BASED MODEL

To support real-time dynamic spectrum trading, a computational-efficient auction framework is proposed with simple and effective bidding and fast auction clearing algorithms [222]. Specifically, buyers use a compact and yet expressive bidding format to express their desired spectrum usage and willingness to pay, while sellers execute fast clearing algorithms to derive prices and allocations under different pricing models. In the following, we investigate the proposed bidding formats and the corresponding optimization problems under different pricing models.

[Piecewise Linear Price-Demand (PLPD) Bids]

A good bidding language should provide expressive but concise bids. At the same time, it also needs to be compact, preventing complicated auction-clearing process. We propose to use piecewise linear price demand (PLPD) curves that not only satisfy both requirements, but also lead to low-complexity clearing algorithms.

With PLPD, a bidder i expresses the desired quantity of spectrum f_i at each per-unit price p_i using a continuous concave piecewise linear demand curve. A simple example is linear demand curves

$$p_i(f_i) = -a_i f_i + b_i, a_i \geq 0, b_i > 0 \tag{6.99}$$

where the negative slope represents price sensitivity at buyers - as the per-unit price decreases, demands in general increase. Any PLPD curve can be expressed as a conglomeration of a set of individual linear pieces (see Figure 6.27). For ease of explanation, linear demand curves is used to describe auction problems and solutions. However, algorithms and proofs in [222] easily generalize to concave piecewise linear demand curves.

When $a_i > 0$, the revenue produced by each bidder is a piecewise quadratic function of the price. Figure 6.27 shows the quantity $f_i(p_i)$, and the revenue generated $R_i(p_i)$ as a function of the price p_i:

$$\begin{aligned} f_i(p_i) &= \frac{b_i - p_i}{a_i}, \ 0 \geq p_i \geq b_i \\ R_i(p_i) &= f_i(p_i)p_i = \frac{b_i p_i - p_i^2}{a_i} \end{aligned} \tag{6.100}$$

For linear demand curves, the revenue is a quadratic function of price, with a unique maximum at $p_i = b_i/2$. Further, if $p_i \to 0$, $R_i(p_i) \to 0$ and if $p_i \to b_i$, $R_i(p_i) \to 0$. PLPD has several attractive advantages. First, it is simple and yet highly expressive. PLPD can approximate any arbitrary continuous concave functions, and hence support a broad class of demands. Bidders express their preferences privately, eliminating complex bid signaling and collusive strategies. Second, each single bid covers different pricing options, eliminating the need for auctioneers to collect bids iteratively. Finally, PLPD produces (piecewise) quadratic revenue functions which significantly simplify the auction-clearing problem.

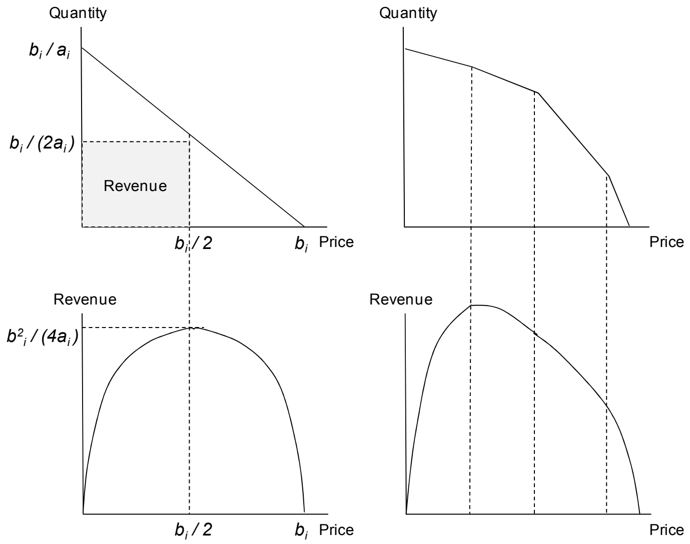


Figure 6.27 On the left, linear demand curve (top) and the corresponding revenue generated (bottom) and on the right a concave piecewise linear demand curve (top) and the corresponding piecewise quadratic revenue function [222].

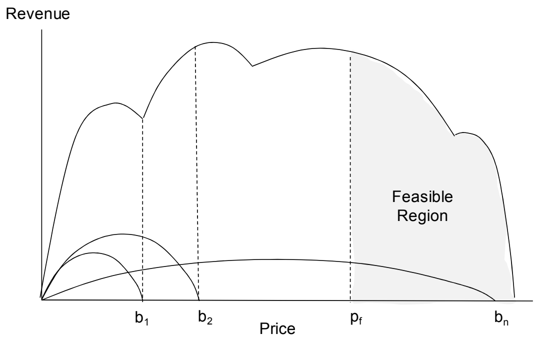


Figure 6.28 The revenue as a function of clearing price p in the uniform pricing model [222].

Although auction revenue and efficiency depend on buyer's social and financial strategy and their PLPD formats, we do not address mechanisms to compute the optimal PLPD curves. Instead, this work assumes that each buyer has its own curve, and focus on how to solve the auction-clearing problem given the bids.

[Pricing Models and Auction-Clearing Problems]
We now describe the auction clearing problem under both uniform and discriminatory pricing models. Note that when $a_i = 0$, the clearing problem becomes a classical weighted throughput maximization problems with good solutions [22] [36] [126]. Hence this work considers the general cases where $a_i > 0$.

- *Uniform pricing:* The auctioneer sets a clearing price p. Each bidder obtains a fraction of spectrum $f_i(p) = (b_i - p)/a_i$ and produces a revenue of $R_i(p) =$

$(b_i p - p^2)/a_i$. Any bidder i with $b_i \leq p$ gets zero assignment. In this case, the optimization problem is to search for the revenue-maximizing price p. Without loss of generality, assume that bidders 1 to n are labeled in increasing order of b_i, i.e. $b_1 \leq b_2 \leq b_3 \leq \ldots \leq b_n$. And $b_0 = 0$. For a given price p, the revenue $R(p)$ is computed as:

$$R(p) = \sum_{i \in |1,n|, b_i > p} R_i(p) = \sum_{i, b_i > p} \frac{b_i p - p^2}{a_i} \tag{6.101}$$

Since each $R_i(p)$ is a quadratic function of p, the total revenue is a continuous piece-wise quadratic function as shown in Figure 6.28. Each of the quadratic piece has a parabolic shape. The overall auction clearing problem becomes

$$\begin{aligned} &\text{Maximize} \sum_{i \in |1,n|, b_i > P} \frac{b_i p - p^2}{a_i} \\ &\text{Subject to} \\ &\qquad \text{Interference Constraints} \\ &\qquad f_i = \frac{b_i - p}{a_i} \end{aligned} \tag{6.102}$$

- *Discriminatory pricing:* This pricing scheme considers the case when the clearing prices are non-uniform and vary across i. Clearly the problem of uniform clearing is a special case. The optimization problem becomes

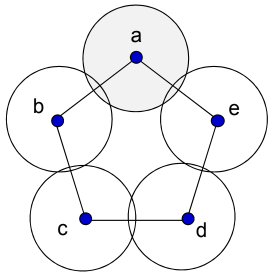
$$\begin{aligned} &\text{Maximize} \sum_{i}^{n} (-a_i f_i^2 + b_i f_i) \\ &\text{Subject to} \\ &\qquad \text{Interference Constraints} \\ &\qquad - a_i f_i + b_i \geq 0,\ f_i \geq 0 \end{aligned} \tag{6.103}$$

[The Optimal Clearing Algorithm]

Both clearing problems are NP-hard. Here, an optimal solution with exponential run time complexity is briefly described. Consider a single channel of the wireless spectrum. If this channel is allocated to any bidder, none of his neighbors in the conflict graph can be allocated this channel. Thus if we consider a maximal independent set of the conflict graph, then all bidders corresponding to the independent set can use the same channel simultaneously. Based upon this observation, Jain et al. [126] proposed an optimal algorithm to resolve interference conflicts: their approach results in a linear programming (LP) problem with an exponentially large number of constraints. Clearly solving such an LP requires exponentially large amount of time and hence not feasible for large number of bidders.

[Linearizing the Interference Constraints]

By judiciously restricting the interference constraints, fast approximations can be developed to the original NP-hard clearing problems in polynomial time. Note that this work assumes the auctioneer has global information on interference constraints and bids.



Node	*NI*	*NLI*	*OPT*
a	$\{1\}$	$\{2,3\}$	$\{1,4\}$
b	$\{2\}$	$\{4,5\}$	$\{2,5\}$
c	$\{3\}$	$\{1,2,3\}$	$\{1,3\}$
d	$\{4\}$	$\{4,5\}$	$\{4,2\}$
e	$\{5\}$	$\{1\}$	$\{3,5\}$

Figure 6.29 Example network, the conflict graph and the channel allocations by NI (Node-Interference), NLI (Node-L-Interference), and OPT (Optimal). There are a total of 5 channels [222].

The auction clearing problem is complex because the discrete interference constraints grow exponentially with the number of buyers. This work proposes to restrict the interference constraints and reduce them into a number of constraints that grow linearly with the number of buyers. The new constraints are stricter and hence lead to a feasible but sub-optimal solution. It is shown that analytically this sub-optimal solution can never be too far off from the optimal one.

To linearize the constraints, the spectrum is supposed to be finely partitioned into a large number of channels. Each buyer i obtains a normalized allocation of $\{f_i : i = 1, 2, \ldots, n\}$ where $f_i \leq 1.0$. For example, a 1MHz spectrum band is divided into 100 channels of 10kHz each. A buyer i with $f_i = 0.143$ will obtain $\lfloor 0.143 \times 100 \rfloor = 14$ channels. In practice this rounding down will lead to some loss of revenue. However, if the number of channels is significantly larger than the highest node degree in the conflict graph, the loss will not lead to undue reduction in revenue. Hence, in the following, f_i behaves as a continuous variable.

In the following, each buyer is referred to as a node in the conflict graph. A neighbor of a node i is defined as any node that interferes with i and hence connects to node i in the conflict graph.

Node-ALL Interference Constraints (NI) The simplest constraint is to restrict i and every neighbor of i to use different spectrum channels, i.e.

$$f_i + \sum_{j \in N(i)} f_j \leq 1, \quad i = 1, 2, \ldots, n \tag{6.104}$$

where $N(i)$ represents the set of neighbors of i and n represents the total number of nodes.

While leading to simple interference free allocations, this constraint is more restrictive than necessary. Using a sample topology, Figure 6.29 illustrates the channel allocation using NI where each node gets only one channel, although node a and d do not conflict with each other and can both use channel 4. Clearly, better approximations is required.

Node-L Interference Constraints (NLI)- A less restrictive constraint is introduced by imposing an order among nodes. By integrating the order in the allocation process, the auction can achieve much more efficient allocations than that using the NI constraints. //

Let two nodes i and j locate at coordinates (x_i, y_i) and (x_j, y_j). Node i is to the left of node j if $x_i < x_j$. If $x_i = x_j$, then the node with the smaller index is considered to be to the node to the left. The constraint becomes: every neighbor of i to the left of i, and i

itself should be assigned with different channels:

$$f_i + \sum_{j \in N_L(i)} f_j \leq 1, \quad i = 1, 2, \ldots, n \tag{6.105}$$

where $N_L(i)$ is the set of neighbors of i lying to its left.

Figure 6.29 compares the allocation results using NLI and NI, and the original constraints (OPT). It is seen that NLI achieves a more efficient channel allocation than NI.

6.4.2 Distributed Inter-Network Spectrum Sharing

Without the complicated coordination of the central network entity, the CR network can coexists with other networks in the same spectrum range. To this end, each CR network has to support the predetermined spectrum etiquettes or the distributed coordination protocols. Similar to the intra-spectrum sharing, distributed inter-spectrum sharing techniques basically require cooperation, which improves spectrum sharing performance, but increases the communication overhead. Several solutions to the distributed inter-network spectrum sharing are introduced in the following.

6.4.3 Etiquette Protocol

As a first step for the coexistence of open spectrum systems, in [133], the common spectrum coordination channel (CSCC) etiquette protocol is proposed for coexistence of IEEE 802.11b and 802.16a networks. The reason we do not consider this work as a complete solution for CR networks is that it necessitates modifications in users using both of the networks. More specifically, each node is assumed to be equipped with a cognitive radio and a low bit-rate, narrow-band control radio. The coexistence is maintained through the coordination of these nodes with each other by broadcasting CSCC messages. Each user determines the channel it can use for data transmission such that interference is avoided. In case channel selection is not sufficient to avoid interference, power adaptation is also deployed. The evaluations reveal that when there is vacant spectrum to use frequency adaptation, CSCC etiquette protocol improves throughput by 35%-160% via both frequency and power adaptation. Another interesting result is that when nodes are clustered around IEEE 802.11b access points, which is a realistic assumption, the throughput improvement of CSCC protocol increases.

[Overview]
There are a number of approaches to improve spectrum sharing. Few notable methods include property rights regimes, spectrum clearinghouse, and unlicensed bands with simple spectrum etiquette, open access, and cognitive radio. Cognitive radio may be viewed as a special case of open access or unlicensed regimes in which radio transceivers are required to meet a relatively high standard of interference avoidance. The cognitive radio principles are under consideration by the FCC, and the research community.

The "agile wideband radio" scheme is the most prevalent concept for cognitive radio in which transmitters scan the channel and choose their frequency band and modulation waveform to meet interference minimization criteria without any protocol-level coordination with neighboring radio nodes. We observe here that agile radios require rapid waveform

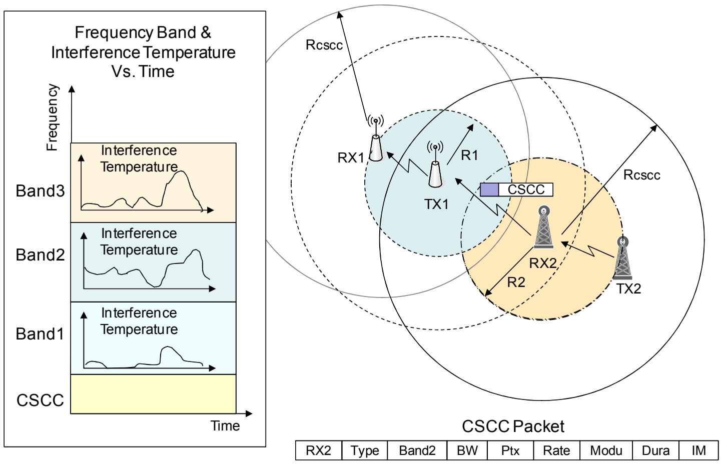


Figure 6.30 Use of CSCC protocol to solve hidden receiver problem [133].

and modulation adaptations which may have a high level of hardware complexity. Without coordination, it suffers from serious limitations due to near-far problems and the hidden receiver problem due to fact that interference is a receiver property while spectrum scanning alone only provides information about transmitters. This problem can be overcome by a small amount of explicit protocol level coordination in which control information is exchanged between transmitters and receivers.

Another simple technique is reactive interference avoidance by control of transmit frequency, rate, power, and/or time occupancy, in which radio nodes do not have any explicit coordination with neighbors but seek equilibrium resource allocation using reactive algorithms to control their parameters. But reactive methods still suffer from the hidden receiver problem since the adaptations are only based on local observations which only provide information about transmitters rather than actual interference experienced by receivers.

With a slightly higher level of protocol complexity, proactive cognitive radio techniques can improve coordination between radio nodes by spectrum etiquette protocols, using either a common spectrum coordination channel (CSCC) at the edge of the shared frequency band or Internet-based spectrum services. Note that the etiquette approach requires some protocol coordination ability including the use of a common control radio for coordination, but may not require full-fledged agile radio capabilities with programmable waveforms. The CSCC protocol considered here achieves the trade-off between the design complexity and the performance improvement, which can help to solve the hidden-receiver problem by explicit announcement of parameters in the CSCC channel.

[The CSCC Protocol]
The CSCC concept is to standardize a common control protocol between different radio systems for spectrum coordination purposes. A simple way to achieve this is to use a Each radio node announces its parameters to neighboring nodes by broadcasting CSCC message. The CSCC message contains information like node ID, center frequency, bandwidth, transmit power, data rate, modulation type, data burst duration, interference margin (IM), service type, etc. The CSCC protocol mechanism is independent of the spectrum

coordination policy, which can be implemented to reflect regional or application-specific requirements.

Since interference needs to be considered at receivers rather than transmitters, CSCC announcements may also be made by receivers involved in active data sessions. CSCC works in a distributed fashion, and the control messages can simply rely on one-hop broadcast and contention can be resolved by periodic repetition with some randomization of transmit time to avoid multiple collisions.

When a node receives a CSCC message, it will know there is a data session going on between neighboring nodes at a specified frequency slot for some duration. It will then coordinate its operations by either switching to other bands with lower interference temperature or limiting its own transmit power to avoid interference with the on-going communications following coordination policies. The interference temperatures varying in time indicate interference power levels in each band (Figure 6.30).

The CSCC protocol can help to solve the hidden-receiver problem, which is illustrated in Figure 6.30. Each node is equipped with a common control radio of range Rcscc. When TX2 initiates a data session to RX2, it first notifies RX2 of the transmit power and the estimated data burst duration T2 by data packet piggybacking. Then RX2 will broadcast a CSCC message in the CSCC channel to claim the current spectrum, e.g., Band2, for T2 time. When TX1 receives the CSCC message from RX2, it will know the spectrum Band2 is taken by RX2 and TX1 can either switch to other available spectrum bands (Band1 or Band3) or coordinate with RX2 in Band2 by reducing its transmit power, i.e., coverage range from R1 to R1'y.

Initially TX1 covers a range of R1, and RX2 covers a range of R2. There is no way for TX1 to notice the existence of RX2 only by reactive scanning or sensing, especially when R2<R1, and therefore the transmission of TX1 will interfere RX2 if they share the spectrum. Assuming TX1/RX1 and TX2/RX2 use different radio technologies for data communication, therefore they require a common spectrum coordination protocol to avoid this problem. TX1 then receives CSCC messages from RX2 which is no longer °hidden– to TX1, and TX1 can switch to a different frequency or reduce its power to avoid interference.

[Spectrum Coordination Policies]
Spectrum coordination policies refer to specific algorithmic procedures used for adaptation of frequency or power based on the in-band interference temperature.

- *Coordination by adaptation in frequency:* When a transmitter initiates data communication with a receiver, the receiver will broadcast its operating parameters in the CSCC channel using the common control radio. Considering the previous example again, when TX1 and RX1 have on-going data communication, RX1 broadcasts a CSCC message in the CSCC channel stating it will take Band2 for some duration, as shown in Figure 6.31. After a while, TX2 notifies RX2 that it has data to send, and then RX2 broadcasts a CSCC message stating it wishes to use Band2 for data transfer. In the event that RX2 has a higher priority, it will take over Band2 and starts communication, while TX1 is forced to change its data channel to a clear channel Band1 and notifies RX1 by either broadcasting a CSCC message or piggybacking in

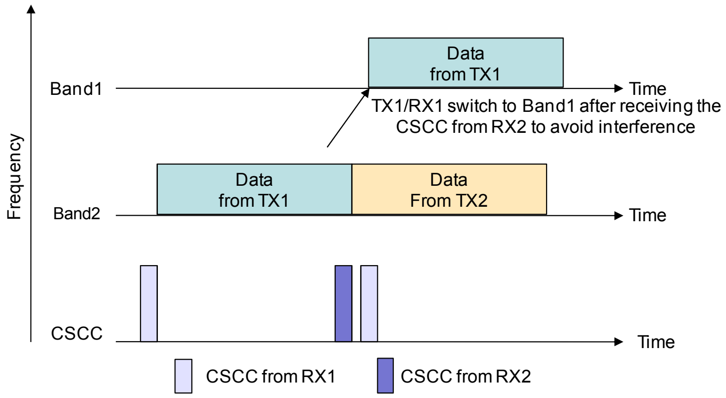


Figure 6.31 Coordination by adaptation in frequency [133].

the data packet. Then RX1 will broadcast a CSCC message to claim Band1.

- *Coordination by adaptation in power:* We consider the case when the spectral band is heavily loaded and frequency selection alone cannot be used to avoid interference between simultaneous users. In such a scenario, adaptation of transmit power is an efficient way to reduce interference. By listening to CSCC messages containing appropriate protocol in this case carries a field called the receiver's interference margin (IM) in the CSCC message. The IM is defined as the maximum interference power a receiver (the one broadcasting the CSCC message) can tolerate without disturbing its ongoing data communication [3].

JOIN SPECTRUM AND POWER ALLOCATION FOR INTER-CELL SPECTRUM SHARING FRAMEWORK

Since the spectrum availability varies over time and space, CR networks are required to have a dynamic spectrum sharing capability. This allows fair resource allocation as well as capacity maximization and avoids the starvation problems seen in the classical spectrum sharing approaches. In this work, a joint spectrum and power allocation framework for inter-cell spectrum sharing is proposed that addresses these concerns by (i) opportunistically negotiating additional spectrum based on the licensed user activity (exclusive allocation), and (ii) having a share of reserved spectrum for each cell (common use sharing). his algorithm consists of inter-cell and intra-cell spectrum sharing schemes, which account for the maximum cell capacity, minimize the interference caused to neighboring cells, and protect the licensed users through a sophisticated power allocation method.

Figure 6.32 shows the proposed framework for spectrum sharing in infrastructure-based CR networks. An *event monitoring* has two different functionalities: spectrum sensing and QoS monitoring. According to the detected event type, the base-station determines the spectrum sharing strategies. An *intra-cell spectrum sharing* enables the base-station to avoid the interference to the primary networks as well as to maintain the QoS of its CR users by allocating spectrum resource adaptively to the event detected inside its coverage. An *inter-cell spectrum sharing* is comprised of two sub-functionalities: *spectrum allocation* and *power allocation*. When the service quality of the cell is below the guaranteed level,

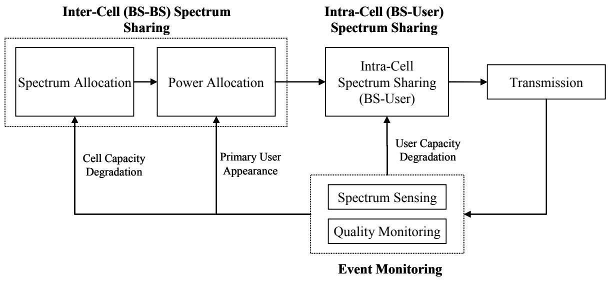


Figure 6.32 Inter-cell spectrum sharing framework

the base-station initiates the inter-cell spectrum sharing and adjusts its spectrum allocation. Based on the spectrum allocation, the base-station determines its transmission power over the allocated spectrum bands [149].

For the inter-cell spectrum sharing, each base-station needs to exchange the following local cell information with its neighbor base-stations:

- *Spectrum Availability / Utilization:* The base-station needs to announce the availability of all spectrum bands as well as its current spectrum utilization to its neighbor cells.

- *Minimum Busy Interference $I_i^{\min}(j)$:* To reduce the communication overhead, we use a single representative information among all sensing results. When the primary user (PU) activity is detected in spectrum i, the base-station j sends the minimum signal strength among all sensing data observed in its users.

- *Maximum Idle Interference $I_i^{\max}(j)$:* If no PU activity is detected at spectrum i, the base-station j sends the maximum value among the interference measured during the transmission period.

[Spectrum Allocation for an Exclusive Model]

Due to the inefficient and unfair spectrum utilization, a classical exclusive approach is not suitable to CR networks. The proposed approach improves the spectrum availability in exclusive allocation by considering the permissible transmission power derived from spatio-temporal characteristics of the PU activity.

- *Cell Characterization:* Even in the same spectrum band, PU activities show different characteristics according to the location. According to the types of cells in the interference range, we classify three different scenarios for exclusive spectrum allocation as shown in Figure 6.33.

    1. *Type I. Same PU Activity in the Interference Range:* If no primary user is detected, the base-station can transmit with the maximum power $P_i^{\max}(j)$. Otherwise the transmission power is zero. Thus, the probabilities $Pr(\cdot)$ of both cases can be derived as follows:

$$\begin{aligned} &Pr(P_i^{\max}(j)) = p_i^{\text{off}}(k') \\ &Pr(0) = 1 - p_i^{\text{off}}(k') \end{aligned} \tag{6.106}$$

where k' is the PU activity region in the interference range, and $p_i^{\text{off}}(k')$ is the idle probability of spectrum i ar region k'.

2. *Type II. Multiple PU Activities in the Interference Range:* The neighbor cells located in different PU activity regions can restrict the transmission power of the current cell even though no PU activity is detected in its transmission range. Let $\mathcal{K}$ be the set of PU activity regions in the interference range, and k' be the region in the transmission range. Here we define the dominant regions $\mathcal{K}_k^*$ as the set of PU activity regions which allow smaller transmission power than the region k when primary users are detected in all regions. Then the probabilities of each permissible transmission power can be determined as follows:

$$\begin{aligned} Pr(P_i^{\max}(j)) &= \prod_{k\in\mathcal{K}} p_i^{\text{off}}(k) \\ Pr(P_i^{\text{pu}}(j,k)) &= p_i^{\text{off}}(k') \cdot \prod_{k^*\in\mathcal{K}_k^*} p_i^{\text{off}}(k^*) \cdot (1 - p_i^{\text{off}}(k)) \quad (k \in \mathcal{K},\ k \neq k') \end{aligned} \tag{6.107}$$

The probability of zero transmission power is the same as that of Type I.

3. *Type III. Multiple PU Activities in the Transmission Range:* The probability of $P_i^{\max}(j)$ is the same as that of Type II. Let the set of primary network regions in its transmission range be $\mathcal{K}'$. Then the probabilities of $P_i^{\text{pu}}(j,k)$ and zero power can be derived as follows:

$$\begin{aligned} Pr(P_i^{\text{pu}}(j,k)) &= \prod_{k'\in\mathcal{K}'} p_i^{\text{off}}(k') \cdot \prod_{k^*\in\mathcal{K}_k^*} p_i^{\text{off}}(k^*) \cdot (1 - p_i^{\text{off}}(k))l \quad (k \in \mathcal{K} - \mathcal{K}') \\ Pr(0) &= 1 - \prod_{k\in\mathcal{K}'} p_i^{\text{off}}(k') \end{aligned} \tag{6.108}$$

- *Permissible Transmission Power:* When no PU activity is detected in any neighbors, the cell j can use the maximum power, $P_i^{\max}(j)$, in spectrum i. Let the power propagation function be $\mathcal{F}(.)$. Then $P_i^{\max}(j)$ can be obtained as follows:

$$I_\Delta(j^*) = P_{\text{temp}} W_i - I_i^{\max}(j^*), \quad j^* \in \mathcal{N}(j) \tag{6.109}$$

$$P_i^{\max}(j) = \min_{j^*\in\mathcal{N}(j)} \mathcal{F}^{-1}(I_\Delta(j^*), D(j,j^*) + R(j^*)) \tag{6.110}$$

where $I_\Delta(j^*)$ is the available power for CR users at the neighbor cell j^* and $\mathcal{N}(j)$ is the set of neighbors of the cell j.

If some neighbors are located in different PU activity regions, the permissible transmission power can be determined according to their locations. Since $I_i^{\max}(\cdot)$ is not available in neighbor cells currently busy, it can be estimated as $I_i^{\min}(j^*) - \gamma \cdot P_{\text{temp}} \cdot W_i$. In this case, the permissible transmission power can be determined so that the received power at the border of neighbor cell nearest from the base-station, does not exceed $I_\Delta(j^*)$. Thus, the restricted transmission power can be obtained as follows:

$$P_i^{\text{pu}}(j,k) = \min_{j^*\in\mathcal{N}_i(j,k)} \mathcal{F}^{-1}(I_\Delta(j,j^*), D(j,j^*) - R_{j^*}) \tag{6.111}$$

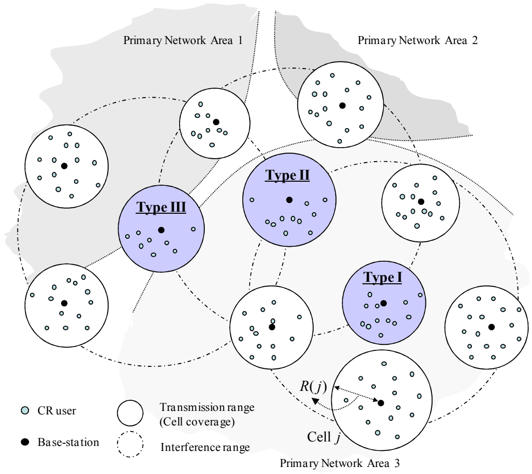


Figure 6.33 Infrastructure-based CR networks and their cell types

where $\mathcal{N}_i(j,k)$ is the set of neighbors of the cell j located at region k of spectrum i.

- *Spectrum Selection: Opportunistic cell capacity* $C_i(j)$, defined as the capacity of spectrum i at the boundary of cell j, can be derived as follows:
  Type I:
  $$C_i(j) = W_i \log_2(1 + \frac{\mathcal{F}(P^{\max}(j), R(j))}{I_i^{\max}(j)}) p_i^{\text{off}} \tag{6.112}$$
  Type II & III:
  $$\begin{aligned} C_i(j) = W_i[&\log_2(1 + \frac{\mathcal{F}(P_i^{\max}(j), R(j))}{I_i^{\max}(j)}) \cdot Pr(P_i^{\max}(j)) \\ &+ \sum_{k \in \mathcal{K}} \log_2(1 + \frac{\mathcal{F}(P_i^{\text{pu}}(j,k), R(j))}{I_i^{\max}(j)}) \cdot Pr(P_i^{\text{pu}}(j,k)] \end{aligned} \tag{6.113}$$
  If the base-station has the multiple available spectrum bands for the exclusive allocation, it selects the one of them with the highest opportunistic capacity.

[Spectrum Allocation for Common Use Model]

In the common use approach, the cell can share the same spectrum with its neighbor cells, which improves fairness but causes capacity degradation due to the inter-cell interference. To mitigate this effect, the following issues should be considered in the common use approach: 1) The common use approach aims at finding a spectrum to enable the cell capacity to be maximized. 2) To reduce the inter-cell interference, CR networks need to find a spectrum to cause less influence on neighbor cells. 3) the uplink transmission is highly probable to interfere with the PU activity detected in its neighbor cells. To address these issues, we propose a two-step spectrum sharing for the common use model.

- *Angle-Based Allocation for Uplink Transmission:* The best way to reduce interference in uplink transmission is to use the spectrum which does not have any PU activities in neighbor cells. If the base-station cannot find this spectrum, alternatively it can exploit the multiple spectrum bands to allow all directions to be covered with their idle regions, referred to as an *angle-based allocation*.

- *Interference-Based Spectrum Allocation:* For the maximum cell capacity, the cell should find the spectrum with the lowest interference in its transmission range. In addition, the cell needs to select the spectrum with lower interference, which shows less influence on the neighbor cells. Furthermore, it is much advantageous for the cell to choose the spectrum with the widest idle angle range. From these observations, we devise the following selection criterion for the common use approach, called an *interference-based spectrum allocation*:

$$i^* = \arg\max_{i \in \mathcal{S}(j)} \frac{\theta_i^{\max}(j)}{2\pi} \cdot \frac{\min_{j \in \mathcal{N}(j)} I_i^{\max}(j^*)}{I_i^{\max}(j)} \tag{6.114}$$

where $\theta_i^{\max}(j)$ is the maximum idle angle in spectrum i at cell j. $\mathcal{S}(j)$ is the set of available spectrum bands at base-station j. Here, in order to consider the effect on all neighbors, the lowest $I_i^{\max}(j^*)$ is chosen among the interference measured in neighbors.

[Joint Spectrum and Power Allocation for Inter-Cell Spectrum Sharing]

- *Power Allocation:* Generally a water filling method is used to optimally allocate the transmission power in the presence of noise. However, unlike the classical water filling, in the inter-cell spectrum sharing, each spectrum has an upper power limit according to the PU activities:

  In case of the exclusive allocation, the upper power limit $P_i^{\text{up}}(j)$ can be obtained as explained in Eq. (6.110). However, $I_i^{\max}(j^*)$ in the common use mode, does not contain the its own signal strength, which needs to be considered in determining available power. In this case, the transmission power of the neighbor cell can be estimated as $\mathcal{F}^{-1}(I_i^{\max}(j), D(j,j^*) - R_j)$. Thus, the available power at the farthest border of the cell j^* from the base-station j can be obtained as follows:

$$I_\Delta(j^*) = P_{\text{temp}} \cdot W_i - [I_i^{\max}(j^*) + \mathcal{F}(\mathcal{F}^{-1}(I_i^{\max}(j), D(j,j^*) - R_j), R_{j^*})] \tag{6.115}$$

  Then $P_i^{\text{up}}(j)$ can be derived using Eq. (6.111).

- *Spectrum Sharing Procedures:* In order to reduce this unnecessary influence on the entire networks, we classify the spectrum as the *assigned* and *unassigned* spectrum bands. The assigned spectrum bands are allowed to be accessed by the current cell. The assigned spectrum can be released to the unassigned one only when it is currently idle and is requested by the the neighbor cells for their exclusive allocation.

  Based on this spectrum classification, the inter-cell spectrum sharing can be performed as follows: If the spectrum sharing event is detected, the base-station initiates the spectrum sharing procedure. If the detected event is related only to the PU activities, the base-station reduces or turns off its transmission power on that spectrum. If the event is a resource shortage for uplink transmission, it selects the spectrum having the proper idle angle through the angle-based allocation. If the base-station detects the quality degradation, it performs the exclusive allocation for the assigned spectrum and then for the unassigned spectrum if necessary. If it cannot find the proper spectrum bands, it turns to the common use sharing and performs the interference-based allocation. Once spectrum is allocated, each base-station allocates the proper transmission power over the assigned spectrum bands.

6.5 SPECTRUM SHARING CHALLENGES

In the previous sections, the theoretical findings and solutions for spectrum sharing in CR networks are investigated. Although there already exists a vast amount of research in spectrum sharing, there are still many open research issues for the realization of efficient and seamless open spectrum operation. In the following, we detail the challenges for spectrum sharing in CR networks along with some possible solutions.

6.5.1 Dynamic Radio Range

Radio range changes with operating frequency due to attenuation variation. In many solutions, a fixed range is assumed to be independent of the operating spectrum [36], [292]. However, in CR networks, where a large portion of the wireless spectrum is considered, the neighbors of a node may change as the operating frequency changes. This effects the interference profile as well as routing decisions. Moreover, due to this property, the choice of a control channel needs to be carefully decided. It would be much efficient to select control channels in the lower portions of the spectrum where the transmission range will be higher and to select data channels in the higher portions of the spectrum where a localized operation can be utilized with minimized interference. So far, there exists no work addressing this important challenge in CR networks and we advocate operation frequency aware spectrum sharing techniques due the direct interdependency between interference and radio range.

6.5.2 Spectrum Unit

Almost all spectrum sharing techniques discussed in the previous sections consider a *channel* as the basic spectrum unit for operation. Many algorithms and methods have been proposed to select the suitable *channel* for efficient operation in CR networks. However, in some work, the channel is vaguely defined as "orthogonal non-interfering" [37], "TDMA, FDMA, CDMA, or a combination of them" [202], or "a physical channel as in IEEE 802.11, or a logical channel associated with a spectrum region or a radio technology" [292]. In other work, the channel is simply defined in the frequency dimension as frequency bands [133], [170], [176], [193], [231]. It is clear that the definition of a channel as a spectrum unit for spectrum sharing is crucial in further developing algorithms. Since a huge portion of the spectrum is of interest, it is clear that properties of a *channel* may not be constant due to the effects of operating frequency. On the contrary, a channel is usually assumed to provide the same bandwidth as other channels [36], [115], [180], [202], [37].

Furthermore, the existence of primary users and the heterogeneity of the networks that are available introduce additional challenges to the choice of a spectrum unit/channel. Hence, different resource allocation units such as CSMA random access, TDMA time slots, CDMA codes, as well as hybrid types can be allocated to the primary users. In order to provide seamless operation, these properties need to be considered in the choice of a spectrum unit. In [114], the necessity of a *spectrum space* for a spectrum unit is also advocated. The possible dimensions of the spectrum space are classified as power, frequency, time, space, and signal. Although not orthogonal, these dimensions can be used to distinguish signals [114].

For this purpose, we describe a three dimensional space model for modeling network resources that has been proposed in [259]. Although this work focuses on heterogeneity in next generation/4th generation (NG/4G) networks, as discussed in [259], it can be easily incorporated into CR networks. Based on this three dimensional *resource-space*, a novel *Virtual Cube* concept has been proposed in order to evaluate the performance of each network. The Virtual Cube concept defines a unit structure based on the resource allocation techniques used in existing networks.

The resource is modeled in a three dimensional *resource-space* with time, rate, and power/code dimensions. The *time dimension* models the time required to transfer information. The *rate dimension* models the data rate of the network. Thus, the capacity of different networks with the same connection durations but different data rates are captured in the rate dimension. Furthermore, in the case of CDMA networks, the bandwidth increase due to the multi-code transmissions is also captured in this dimension. The *power/code dimension* models the energy consumed for transmitting information through the network. Note that, the resource in terms of available bandwidth can be modeled using the time and rate dimensions. However, the cost of accessing different networks vary in terms of the power consumed by the wireless terminal. Hence, a third dimension is required. Each network type requires different power levels for transmission of the MAC frames because of various modulation schemes, error coding and channel coding techniques. As a result, the resource differences in these aspects are captured in the power dimension.

Using this model, heterogeneous access types in existing networks as well as CR network spectrum can be modeled based on a generic spectrum unit. We advocate that determining a common spectrum unit is crucial for efficient utilization of the wireless spectrum and seamless operability with existing primary networks.

6.5.3 Spectrum Access and Coordination

In classical ad hoc networks, the request to send (RTS) and clear to send (CTS) mechanism is used to signal control of the channel and reduce simultaneous transmissions to an extent. In CR networks, however, the available spectrum is dynamic and users may switch the channel after a given communicating pair of nodes have exchanged the channel access signal. Thus, a fresh set of RTS.CTS exchange may need to be undertaken in the new channel to enforce a silence zone among the neighboring CR users in the new spectrum. Moreover, the CR users monitoring the earlier channel are oblivious to the spectrum change on the link. They continue to maintain their timers and wait for the duration needed to complete the entire data transfer before initiating their own transmission. This leads to inefficient spectrum use, and new coordination mechanisms among the CR users is necessary whenever the spectrum access conditions change.

6.5.4 Reactivity to Topology Modifications

The use of non.uniform channels by different CR users makes topology discovery difficult. For example, two CR users A and B experience different PU activity in their respective coverage areas (channels 1 and 2 available for CR A and channel 3 for CR B) and thus may only be allowed to transmit on mutually exclusive channels. The allowed channels for CR

A (1,2) being different from those used by CR B (3) makes it difficult to send out periodic beacons informing the nodes within transmission range of their own ID and other location coordinates needed for networking.

In mobile networks the topology changes quite often, leading to frequent variations of interference profile. In this situation one of the critical points is the fastness of the system to react when the current spectrum sharing configuration does not fulfil anymore the user requirements. The risk is that, if the system needs long time to compute a new spectrum sharing configuration, the solution found could be not suitable anymore looking at the fact that topology could be varied with respect to the one that started the recalculation process. Hence, a technique that leads to a sub.optimal solution in a short amount of time could be preferred, in this context, to another one that requires a longer amount of time, but that finds the optimal solution. More generally a special attention should be put in balancing, according to the scenario, convergence fastness and distance from the optimum.

6.5.5 Reliable Control Channel

To share spectrum resources efficiently, CR transmitter should have feedback information regarding channel condition and QoS status from its receiver. Thus, each CR user necessitates a reliable control channel for exchanging control information. The control channel can be established through either out-of-band or in-band signalling. However, with the in-band signaling, it is not easy to find the neighbor users tuning different spectrum band and exchange information. We may use the dedicated control channel based on out-of-band signalling, which is not reliable due to PU activities. Especially in CR ad hoc networks, asynchronous sensing and transmission schedules make it more difficult to exchange information with its neighbors. As a result, how to reliably obtain the channel and QoS information from the receiver or neighbor users is still unsolved in networks.

Many spectrum sharing solutions, either centralized or distributed, assume a CCC for spectrum sharing [18], [170], [231]. It is clear that a CCC facilitates many spectrum sharing functionalities such as transmitter receiver handshake [170], communication with a central entity [18], or sensing information exchange. However, due to the fact that CR network users are regarded as *visitors* to the spectrum they allocate, when a primary user chooses a channel, this channel has to be vacated without interfering. This is also true for the CCC. As a result, implementation of a fixed CCC is infeasible in CR networks. Moreover, in a network with primary users, a channel common for all users is shown to be highly dependent on the topology, hence, varies over time [292]. Consequently, for protocols requiring a CCC, either a CCC mitigation technique needs to be devised or local CCCs need to be exploited for clusters of nodes [292]. On the other hand when CCC is not used, the transmitter receiver handshake becomes a challenge. For this challenge, receiver driven techniques proposed in [297] may be exploited.

6.5.6 Spectrum Etiquette and Standardization

Spectrum etiquette involves devising protocols that ensure CR devices having different hardware capabilities, carrying traffic with varying QoS requirements, and forming dissimilar connected topologies coexist with fairness in transmission opportunity and end-to-end performance. The problem of identifying a common set of rules becomes more involved in case of CR ad hoc networks belonging to different independent operators that may be present in spatially overlapped regions. There are several forums and committees created both in the user domain and also through government efforts. As an example of the non-profit user domain working group, the IEEE SCC41 P1900.5 group aims to define a policy language along with the consideration of the possible architectures for specifying interoperable, vendor-independent control of networks that are enabled with dynamic spectrum access ability [192]. Similarly, in absence of an appropriate industry organization, the US DARPA Wireless Networking after Next (WNaN) program considers the problem of policy regulation from the viewpoint of software development for the CR radios. However, both these efforts are at a nascent stage and formulation of a set of universally applicable etiquette seems a difficult challenge.

6.5.7 Determination of Channel Structure

Under the *open spectrum sharing* model, the spectrum is made available as a contiguous frequency block, that must be separated into channels for use by the CR users. The number of channels should be such that the CR users have sufficient choice is choosing distinct and non-overlapping channels whenever possible, and at the same time be able to sustain a minimum desired channel throughput. In the absence of a central entity, balancing this tradeoff by creating an optimal number of channel divisions is a challenge.

6.5.8 Detection of Selfish Behavior

In the absence of standard spectrum etiquette, competition based approaches may allow sharing of the spectrum among the CRAHN users, and both cooperative and selfish approaches are discussed in [193]. Cooperation may involve choosing an optimal transmission power, channel bandwidth, transmission rate, among others parameters such that the user's own performance is maximized, along with that of the overall network. In competitive approaches, each user may progressively increase its own usage of the spectrum resource and other communication parameters selfishly till its performance is affected by similar operation of the neighboring users. In this case, the user does not seek to maximize the collective gain, but simply tries to protect its own transmission, thereby settling on a choice of optimal parameters over time [266]. While cooperative strategies are more suited for users belonging to a single operator, the competition based approaches are viable for inter-operator CRAHN coexistence.

As the spectrum is shared by the CR users, they may choose the channel structure independently of the others. Moreover, users belonging to different CR operators may have different channel specifications, such as the amount of allowed spectral leakage in the neighboring channels, transmission masks, channel bandwidth, among others. In such cases, it is important to detect the CR users that exhibit selfish behavior by using the spectrum that exceeds the regulations laid down by the specifications [140]. This may allow some of the CR users to unfairly improve their performance at the cost of the others, making it necessary to devise strategies to detect this selfish behavior.

6.5.9 Penalizing and Regulatory Policing

As CR ad hoc networks do not have a centralized admission control scheme, penalizing the CR users for selfish or malicious behavior is difficult. Moreover, regulatory policing rules must be established for each free spectrum pool, so that CR users can collectively decide on their inclination to forward traffic originating from the node engaging in selfish behavior.

CHAPTER 7

SPECTRUM MOBILITY

CR users are generally regarded as *visitors* to the spectrum. Hence, if the specific portion of the spectrum in use is required by a PU, the communication needs to be continued in another vacant portion of the spectrum. Furthermore, CR networks target to use the spectrum in a dynamic manner by allowing CR users to operate in the best available frequency band. This enables a “Get the Best Available Channel” concept for communication purposes. To realize this concept, the CR user has to capture the best available spectrum. *Spectrum mobility* is defined as the process when an CR user changes its frequency of operation.

CR networks require spectrum mobility functionalities 1) when any PU is detected in the current spectrum, 2) a CR user involving on-going communications moves into an area in which PUs are active, or 3) a current spectrum band cannot satisfy the QoS requirements of CR users. In CR networks, a temporary communication break is inevitable due to the process for discovering a new available spectrum band. Furthermore, since available spectrums are discontiguous and distributed over a wide frequency range, CR users may require the reconfiguration of operation frequency in its RF front-end, leading to significantly longer switching time. The purpose of the spectrum mobility management in CR networks is to ensure smooth and fast transition while minimizing abrupt quality degradation during spectrum switching. In addition, spectrum mobility enables different layering protocols to be transparent to the spectrum switching and the associated latency, by adapting to the channel parameters of the operating frequency. We describe this adaptation in the routing and transport protocols, which are covered in Chapters 9 and 10, respectively.

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

In the following, the main functionalities required for spectrum mobility in the CRAHN are described:

- *Spectrum Handoff:* Spectrum mobility gives rise to a new type of handoff in CR networks, the so-called *spectrum handoff*, in which the CR users transfer their connections to an unused spectrum band when they detect the spectrum mobility event during the transmission.
- *Connection Management:* The CR user sustains the QoS or minimizes quality degradation during the spectrum switching by interacting with each layering protocols.

As stated previously, the spectrum mobility events can be detected as a link failure caused by user mobility as well as PU detection. Furthermore, the quality degradation of the current transmission also initiates spectrum mobility. When these spectrum mobility events are detected through spectrum sensing, neighbor discovery, and routing mechanism, they trigger the spectrum mobility procedures. By collaborating with spectrum sensing and decision, a CR user determines a new spectrum band, and switch its current session to the new spectrum (*spectrum handoff*). During the spectrum handoff, the CR user needs to maintain current transmission not to be interfered by the switching latency (*connection management*).

Figure 7.1 (a) shows spectrum mobility functionalities for infrastructure-based networks. In this architecture, a base-station is responsible for handling all spectrum mobility events on behalf of CR users in its coverage, i.e., once the base-station detects spectrum mobility events based on the observations collected from its CR users, it forces its users to vacate a current spectrum immediately, and to move to a new re-allocated spectrum or to a new base-station if necessary. During spectrum switching, the base-station minimizes the influence on performance of upper-layer protocols, and sustain the level of qualities required by user applications through a connection management function. On the contrary, spectrum mobility in ad hoc networks needs to consider the link failure on the end-to-end route. Compared to the infrastructure-based network, the CR ad hoc network has a more dynamic and complicated topology dependent on both spectrum and user mobilities. Furthermore, as shown in Figure 7.1, spectrum mobility in CR ad hoc networks needs to collaborate with the routing protocol to recover the link failure on its end-to-end route. However, due to the the lack of the central network entity as well as a dynamic network topology, CR ad ho network cannot detect and manage mobility events as efficiently as the infrastructure-based network. For these reasons, it is much more difficult to design spectrum mobility in CR ad hoc networks compared to that in infrastructure-based networks.

In the following subsection, we investigate two main functionalities in spectrum mobility: spectrum handoff and connection management.

7.1 SPECTRUM HANDOFF

The handoff concept has been widely investigated in mobile communications, and is currently provided by all cellular networking technologies. The traditional handoff scheme has the following procedure: When the mobile station or its corresponding base-station detects the link quality deterioration, it scans neighbor base-stations, obtains their information. Accordingly, it determines a target base-station, connects to it, and resumes the

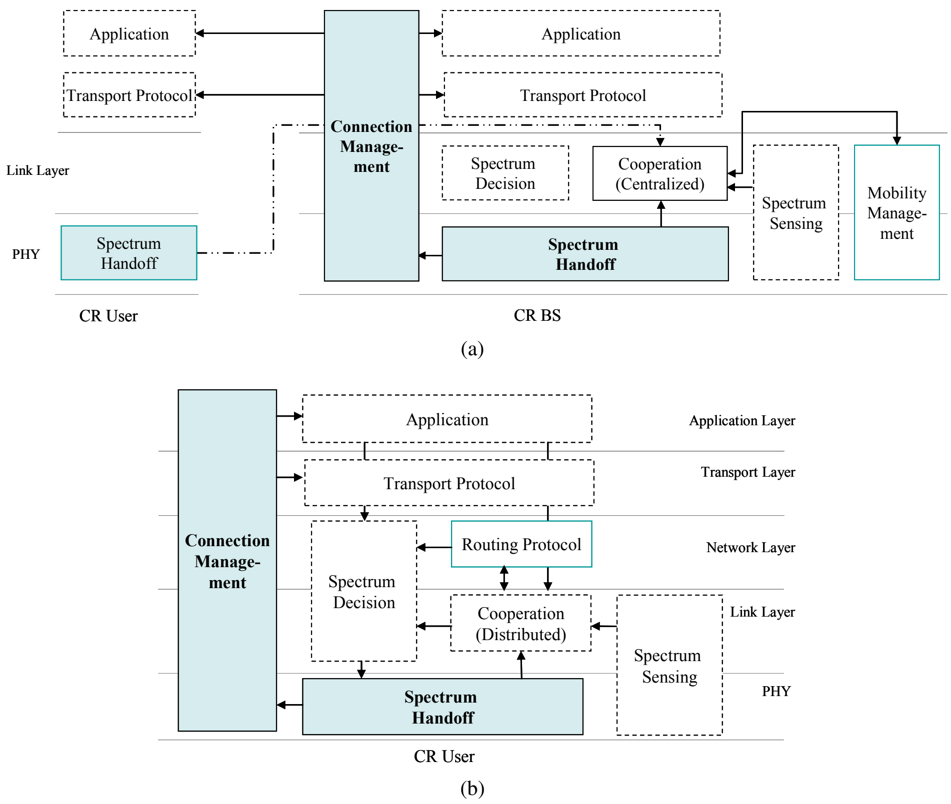


Figure 7.1 Functional block diagram for spectrum mobility: (a) infrastructure-based CR networks, and (b) CR ad hoc networks.

communication.

On the contrary, spectrum handoff is designed for providing seamless frequency switching over a dynamic spectrum environment. Although spectrum handoff also has a switching functionality, it has totally different procedures for the traditional handoff schemes. As mentioned previously, the spectrum handoff can be triggered by various events such as PU user appearance, user mobility, and link quality degradation. Especially, the detection of PU appearance highly depends on in-band spectrum sensing capability. Once the mobility events are detected, the CR user first notifies this detected events to its neighbors. If the CR user receives this notification, it should vacate the current spectrum immediately, and then searches all available spectrum bands through out-of-band spectrum sensing. Among the detected bands, the CR user determines the best one, and accordingly reconfigures its hardware and software optimized to the new spectrum, which are related to spectrum decision. After switching, the CR user resumes its communication by considering the resource contention among other users in the new spectrum.

As seen in the procedures above, spectrum handoff is closely coupled with all other spectrum management functionalities such as spectrum sensing, spectrum decision, and

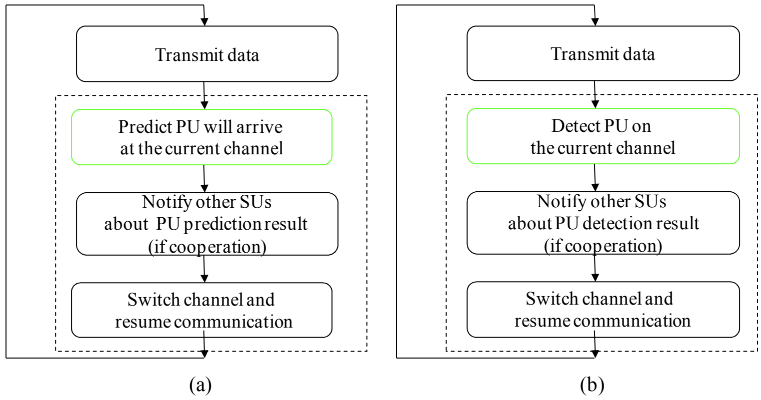


Figure 7.2 Flowchart of spectrum handoff: (a) proactive handoff, and (b) reactive handoff

spectrum sharing. In the following sections, we present the basic functionalities and related work on spectrum handoff in more detail.

7.1.1 Spectrum Handoff Procedures

Spectrum handoff schemes can be classified in two aspects: i.e., according to their handoff strategy, and architectural feature. In the following subsections, we investigate the procedures of each spectrum handoff type.

SPECTRUM HANDOFF STRATEGIES

Spectrum handoff can be implemented based on two different strategies as follow:

- *Reactive spectrum handoff:* CR users perform spectrum switching after detecting link failure due to spectrum mobility,as depicted in Figure 7.2 (a). This method requires immediate spectrum switching without any preparation time, resulting in significant quality degradation in on-going transmissions.

- *Proactive spectrum handoff:* CR users predict future activity in the current link and determine a new spectrum while maintaining the current transmission. They notify their predictions if necessary, and then perform spectrum switching before the link failure happens, as shown in Figure 7.2 (b). Since proactive spectrum handoff can maintain current transmissions while searching a new spectrum band, the spectrum switching is faster but requires more complex algorithms for these concurrent operations.

Figure 7.3 compares the behavior of reactive and proactive spectrum accesses. While the reactive handoff inevitably introduces interference to primary users due to its *listen before talk* strategy, the proactive handoff avoids interruptions by switching the spectrum in advance according to the prediction of primary user activities. However, the effectiveness of proactive access depends heavily on being able to predict spectrum availability accurately. When predictions are imperfect, CR users can make “dumb” switches. Figure 7.4 shows two examples of dumb switching. In Figure 7.4 (a), a CR user falsely interprets spectrum j over i and switches to an occupied band, and thereby suffers from unnecessary interruptions to its communication. In Figure 7.4 (b), a user switches to a band with shorter remaining

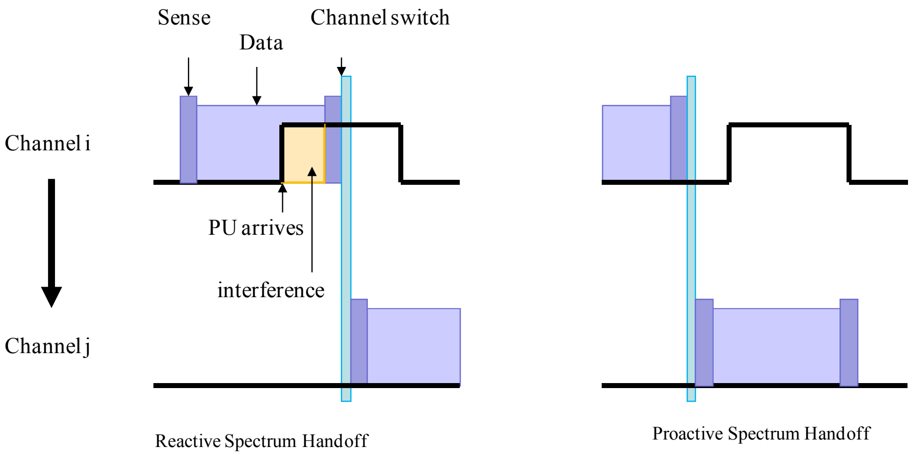


Figure 7.3 Comparison of reactive and proactive handoffs [278]

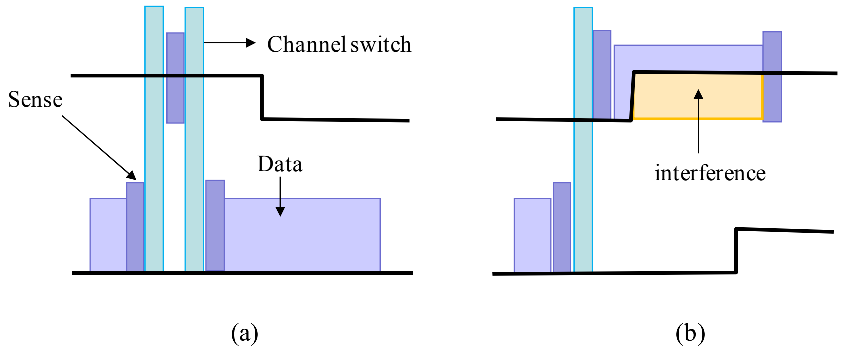


Figure 7.4 Dumb switching cases in proactive handoffs [278]

idle period than the current spectrum, which could reduce its communication period.

As we mentioned above, under reactive spectrum access model, CR users switch channels only after detecting a primary user, causing unavoidable interferences. To solve this problem, a proactive spectrum handoff scheme is proposed to intelligently switch the spectrum by proactively predicting future spectrum availability [278]. The proposed scheme consists of two core modules as follow:

- *Proactive Channel Prediction:* CR users utilize past channel observations to estimate future spectrum availability.
- *Intelligent Channel Switching:* Utilizing prediction results, CR users decide when to exit from a channel and which channel to switch to.

[Proactive Channel Prediction]
The first challenge here is how to use past channel observations to estimate future spectrum availability. Specifically, the proactive handoff is interested in estimating the probability that channel i will be idle in the next time slot, referred to as P_i. This work assumes that each CR user can acquire statistical property of spectrum usage at nearby primary users.

These can be done offline through static traffic analysis, and made available to CR users through online databases. Given primary user's statistical traffic model and parameters, each CR user predicts P_i. In the following, three prediction algorithms are presented:

- *Alternative Exponential Model:* Using renewal theory [64] [143], we can calculate P_i as:

$$P_i = \begin{cases} \frac{\lambda_{Y_i}}{\lambda_{X_i} + \lambda_{Y_i}} + \frac{\lambda_{X_i}}{\lambda_{X_i} + \lambda_{Y_i}} e^{-(\lambda_{X_i} + \lambda_{Y_i})\Delta t_i} & s_i = \text{IDLE} \\ \frac{\lambda_{Y_i}}{\lambda_{X_i} + \lambda_{Y_i}} + \frac{\lambda_{Y_i}}{\lambda_{X_i} + \lambda_{Y_i}} e^{-(\lambda_{X_i} + \lambda_{Y_i})\Delta t_i} & s_i = \text{BUSY} \end{cases} \tag{7.1}$$

  where Δt_i is the time gap from the last history s_i to the next time slot.

- *Periodic Model:* With a sufficient observation time, CR users can always accurately predict the channel availability.

- *Alternative Periodic-Exponential Model:* This model is an intermediate model between two previous extreme cases. The duration of ON (or OFF) periods is fixed to T, and the duration of OFF (or ON) periods is exponential distributed with λ. In this model, P_i can be derived as:

$$P_i = \begin{cases} \frac{1}{T} \int_0^T \sum_{n=0}^{\lfloor \frac{\Delta t_i - x}{T} \rfloor} \frac{\lambda^n (\Delta \hat{t}_i - x)^n}{n!} e^{-\lambda(\Delta \hat{t}_i - x)} dx & \Delta t_i > T \\ \frac{1}{T} \int_0^T e^{-\lambda(\Delta t_i - x)} dx & \Delta t_i < T \end{cases} \tag{7.2}$$

  where $\Delta \hat{t}_i = \Delta t_i - nT$.

In [143], the predicted result P_i is used to determining the scanning order of spectrum bands for out of band sensing. However, this scheme uses these predictions to determine spectrum handoff timing, i.e, to switch the spectrum before the appearance of any primary users, and hence need to continuously update P_i in each time slot.

[Intelligent Channel Switching] Based on observations and predictions, CR users can schedule channel usage to avoid disrupting primary users and maintain reliable communication. The goal of this work is to increase the use of smart switching and avoid dumb switching. The key factor that differentiates smart switching and dumb switching is the accurate prediction of the remaining idle period on each channel. If the remaining idle period in the current channel c is shorter than that in another channel i, then switching from channel c to i is smart. Assuming the traffic of primary users follows alternative exponential model, this method proposes two criteria to plan channel usage:

- *Proactive Planning I:* A user switches to channel i with the largest expected remaining idle period, i.e.

$$i = \arg\max_j \frac{P_j}{\lambda_{X_j}} \tag{7.3}$$

- *Proactive Planning II:* A user switches from channel c to i if with high probability (> 0.5) that the length of the remaining idle period of i is larger than that of c, i.e.:

$$i = \arg\max_j \Pr(T_j > T_c) = \arg\max_j [P_j - \frac{\lambda_{X_i}}{\lambda_{X_i} + \lambda_{Y_i}} P_j P_c \tag{7.4}$$

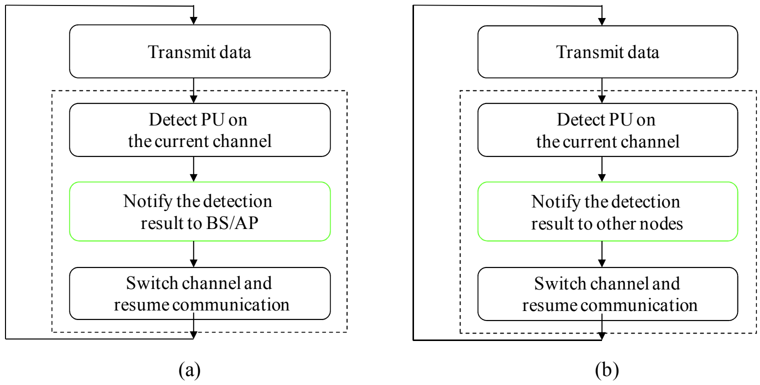


Figure 7.5 Flowchart of spectrum handoff for: (a) infrastructure-based CR network, and (b) CR ad hoc network

NETWORK ARCHITECTURE FOR SPECTRUM HANDOFF

As we mentioned previously, spectrum handoff shows different procedures according to the network architecture. Figure 7.5 depicts the spectrum handoff steps for both centralized and distributed networks. In the infrastructure-based CR network, CR users send their detection results to the central network entity such as a base-station, which is responsible for determining the PU activities, and distributing them back to its users. On the contrary, once CR ad hoc users detect any PU activity, they should promptly notify their observation results fast and reliably to its neighbors in the distributed spectrum handoff.

Thus, how to share the detection results with other neighbors fast and reliably is one of the important issues in distributed spectrum handoff. In [160], a notification protocol based on in-band signaling, named Embedded SpeCtrally Agile radio Protocol for Evacuation (ESCAPE), is proposed to disseminate the evacuation information among all CR users and thus evacuate the licensed spectrum reliably. This protocol uses the spreading code for its transmission, leading to tolerance in interference from both primary and other CR transmissions. Furthermore, due to its flooding-based routing scheme, it requires little prior information on the network topology and density.

The objective of ESCAPE is for multiple cognitive radios to evacuate the channel quickly and reliably. Consider the initiate state when the primary channel is unused. A group of CR users detect the band and start to occupy the channel opportunistically. Later, primary users return, which is detected by one or more CR users. The evacuation step begins. CR users who detect the primary users will transmit a pre-defined warning message declaring "primary-active". The warning message is modulated using a predefined CDMA spreading code. Other CR users hearing the announcement will repeat the same warning message "primary-active". Here, the details of the ESCAPE protocol are as follow:

[Initialization Phase]

A group of CR users operating on a primary band need to agree on a few parameters of the warning message, including the pattern of the warning message, CDMA spreading code to be used, transmission power, and a few other parameters. An evacuation group is a group of connected CR nodes sharing the same spread warning message. One user in the group detecting a primary user(s) will initiate the warning message that evacuates the whole group. The size and membership of the group are determined by the geographic area that need to be evacuated for active primary users. Nodes in a group may belong to different networks (or authority) and do not have regular communications (except for the warning message). A node may also belong to different (partially overlapping) evacuation groups. The purpose of the initialization is to establish and broadcast the spread message. Any communication protocol may be used for such a purpose. A CR user switched to the primary band after the initialization stage would obtain information on the warning message of the area from its neighboring nodes or a central authority.

[Protocol Description]
After the initialization, CR users would sense and utilize the idle primary spectrum. During its normal operational phase, they performs the following procedure individually:

- Step 1: If a CR user has a packet to transmit, it transmits its packet according to its regular access protocol. Then go to Step 2. If a user has no packet to transmit, go directly to Step 3.
- Step 2: Listen to the channel for L_s time unit.
  - If the user notices that the primary is back or it detects the warning signal, go to Step 4.
  - If the user has a packet to transmit or retransmit (e.g., due to a received or missing ACK/NACK or a newly generated packet), go to Step 1.
  - Otherwise, go to Step 3.
- Step 3: Listen to the channel.
- If the user detects a primary or a warning message, go to Step 4.
  - If it has a packet to transmit, go to Step 1.
  - Otherwise, stay in Step 3.
- Step 4: The CR user sends/relays the warning signal at the predetermined power level for N times as shown in Figure 7.6. Go to Step 5.
- Step 5: The user leaves the current band and moves back to the default band.

Step 2 is a required listening phase after each transmission. In current access protocols, a node needs to listen for its ACK/NACK packet after a transmission. On the other hand, in this procedure, L_s can be set larger to enhance the chance of receiving a warning message, as discussed later. Step 3 is the "idle" listening stage where users listen to the channel when they have no packet to transmit. In Step 4, a CR user broadcasts the warning message as node "B" in Figure 7.6 where N is the number of warning message transmissions each user should send, L_t is the maximum transmission time of a regular secondary packet, L_s is the amount of time a CR user listens to the channel, L_p is the prefix transmission time of the warning message, L_w is the transmission time of the warning message, and L_i is the idle interval between two consecutive warning messages from the same CR user. If a CR user is listening to the channel during the transmission of the prefix of the warning message and detects the prefix successfully (known as acquisition), the user will listen to the channel for the following L_w period of time expecting a warning message. If it receives

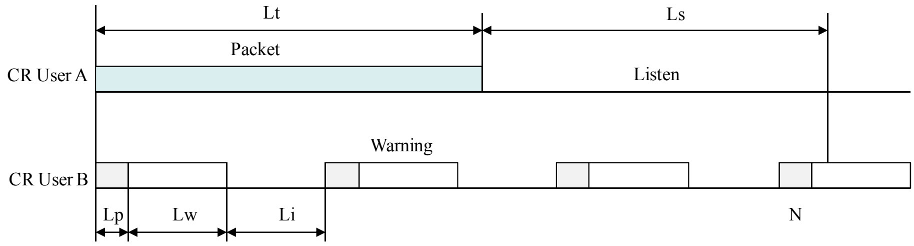


Figure 7.6 Transmission of the warning message..

the warning message successfully, it will broadcast the warning message. Otherwise, the CR user returns to its regular state.

There are two reasons that CR users miss a warning message: 1) it is is transmitting (a regular CR packet) and thus cannot receive the warning message, and 2) the warning message signal is received but cannot be correctly decoded due to signal propagation loss and interference. Such possibilities need to be accommodated in the protocol design. The first issue is addressed by repetition (N) plus an enforced listening window (L_s) and the second by appropriate selection of spreading code and transmission power.

Consider that CR user B is transmitting a warning message and CR user A is a one-hop neighbor of B. We have two options in choosing the values of L_s and N, depending on whether or not a warning message can reach a two-hop neighbor with a high success rate. If so, then the first issue can be solved as follows: Suppose A is transmitting to a one-hop neighbor, R. Because R is within two hops of B, R can receive the warning message. Instead of sending an ACK to A, R will start repeating the warning message. Therefore, A will not receive its ACK but the warning message. In this case, L_s can be set as small as possible. This is the preferred mode because it maximizes the possibility of regular secondary transmissions. In other words, as soon as a node receives its ACK, it can return to the transmission state.

On the other hand, if, due to physical constraints, a warning message may not reach a two-hop neighbor with a high success rate, then L_s and N is set to be large enough to receive the warning message. To elaborate, L_s has to be longer than $(L_p + L_w - \epsilon) + L_i + L_p$, where $\epsilon > 0$. The first term is the delay if "A" just missed the prefix of the warning message, the second one, L_i, is the time between two warning messages, and L_p is the prefix length of the warning message. Thus,L_s is obtained by

$$L_s = 2L_p + L_w + L_i \tag{7.5}$$

In addition, to guarantee that a transmitting CR user has a chance to listen to a warning message, the duty cycle $L_t + L_s$ is defined as

$$L_t + L_s = L_p + (N-1)(L_p + L_w + L_i) + L_p. \tag{7.6}$$

More specifically, A will not start its transmission if it receives the prefix of the warning message sent by B. After transmitting a packet, A will listen to the channel for ACK and for the warning message. Combining Eqs (7.7) and (7.7) L_t is obtained as follow

$$L_t = (N-2)(L_p + L_w + L_i) + L_p \tag{7.7}$$

The above equation has two degrees of freedom, N and L_i, while other parameters are fixed. In the extreme case, we can set $L_i = 0$. In this case, a CR user will send N warning messages back to back. The delay is minimized in such a case. However, there can be (minor) negative impact: 1) More warning message transmissions imply more cumulative interference to primary users; 2) More simultaneous transmissions of the warning message can gradually decrease the reception success probability of the warning message. A similar argument enables the system to support sleep-wake scheduling of users. In particular, a user can sleep up to L_t time units and it has to listen to the channel for L_s time units. In this case, a sleep node can be informed of the evacuation with a high probability. This protocol design guarantees that a CR node has a chance to listen for the warning message, which is improved as N increases.

[MAC and Routing]
The objective of the ESCAPE protocol is to evacuate the spectrum as fast and reliable as possible. More specifically, the warning message "primary-active" needs to bed disseminated to all CR users as soon as possible. The main properties of ESCAPE are that it does not matter where and who start to broadcast the message, and from whom a user gets the message. Thus, the MAC of the warning message is simple: a user transmits the warning message as it wishes, which is similar to the spreading ALOHA protocol where The overlapping transmissions, considered as collision in ALOHA, are handled by spreading codes instead of backoff and retransmission. Since different users exploit the same spreading code, the spreading ALOHA distinguishes different users with asynchronous transmission. If the transmissions of two or more users are synchronized in a chip-level, they cannot be distinguished, resulting in collision However, in ESCAPE, chip-level synchronized transmissions benefit the propagation of warning messages since the signal strength of the received warning message is the summation of multiple copies.

The routing of the warning message is also simple. Basic flooding is assumed. This simplicity again benefits from the auto-correlation property of the spreading code. Flooding also provides reliability enhancement because of its redundancy. If a CR user misses a warning message, it is highly likely that it will receive the message from other neighboring nodes who echo and flood the same warning message. It is possible that more than one user detect the return of primary user and start the warning messages. IA user does not need to resend the warning message if it has already done so.

ESCAPE is designed for the purpose of fast and reliable evacuation. This is different from typical protocol designs where capacity, fairness, and coexistence are critical. From the perspective of primary users, the performance metrics of ESCAPE include evacuation time, peak interference, average interference, and evacuation failure probability. The metrics for CR users include evacuation time and false alarm rate. Thus, design parameters of ESCAPE are spreading code length, transmission power of the warning message, message repetition time (N), and warning message detection threshold given performance constraints of primary and CR users.

7.1.2 Spectrum Handoff in Time and Space Domains

Due to the temporal characteristics of primary user activity, the spectrum opportunity varies over time, which introduces a spectrum handoff in a time domain. Furthermore, as investigated in Chapter 4, spectrum availability also varies from one location to another.

Thus, spectrum handoff can be also initiated by user mobility, i.e., when a CR user moves from the idle region to the one occupied by active primary users, it must switch the spectrum. As a result, both spectrum mobility and user mobility should be jointly considered in designing the spectrum handoff scheme for CR networks, which constitutes an important but unexplored topic in CR networks to date. There are the following challenges [151]:

- *Heterogeneous mobility events:* CR networks are required to provide two different types of handoff schemes: classical inter-cell handoff in infrastructure-based networks (or route recovery in ad hoc networks) due to physical user mobility and spectrum handoff due to spectrum mobility. Thus, it necessitates a unified mobility management framework to exploit different handoff types adaptively to mobility events.

- *Dynamic spectrum availability:* According to the PU activities, spectrum availability varies over time and space in the CR network, making it more difficult to provide seamless and reliable communication to mobile users traversing across multiple PU activity regions. Furthermore, due to this heterogeneity in spectrum availability, the CR user has difficulty in find neighbors, and observing information from them, and hence is not easy to find the best target cell or a next hop neighbor for switching.

- *Broad range of available spectrum:* In CR networks, available spectrum bands are not contiguous and found over a wide frequency range. Thus, when CR users switch their spectrum bands, they need to reconfigure the operating frequency of the radio frequency (RF) front-end so as to tune to a new spectrum band, leading to increase in switching delay. This delay becomes much longer than that in classical wireless networks.

These challenges introduce a unique feature in mobility management for CR networks. Generally, for mobile users, a larger cell coverage in infrastructure-based networks (or a larger transmission range of a mobile user in ad hoc network), is known to be much more advantageous since it reduces the number of classical handoffs [214]. However, in CR networks, a large cell coverage is not always desirable for mobile users. As the cell coverage becomes larger, the PU activity becomes higher since it is more highly probable to include multiple PU activity regions [151].

As shown in Figure 7.7, more PU activity regions can be involved in determining spectrum availability in the extended area, which leads to a higher PU activity. Furthermore, th interference range is generally considered to be larger than its coverage. Thus, even though no primary user is detected within the coverage, interference may occur if the primary user is active in the interference range. Therefore, for the accurate detection, neighbors in the interference range need to be involved in detecting the PU activity with its own detection and false alarm probabilities. Assume that cooperative detection is performed according to the data fusion by OR rule [156]. Then a cooperative detection probability converges to 1 as the number of cells increases [183]. Thus, the detection probability can be ignored when estimating the spectrum availability. On the contrary, the false alarm probability increases as the number of cells increases [183], which influences the spectrum availability significantly in the larger coverage. Even though a spectrum band is idle, it is determined to be unavailable if the false alarm is detected.

Thus, in order to avoid the spectrum handoff, any PU activities or false alarms should not be in detected in the coverage. Here we assume the idle time in the spectrum j at

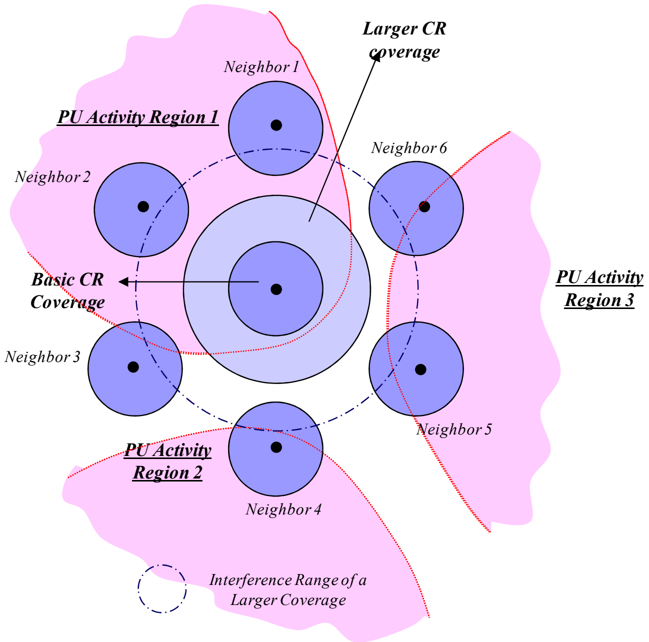


Figure 7.7 The influence of primary user activities in a larger cognitive radio coverage [151].

PU activity region k follows exponential distribution, with mean $1/\beta_i$. Based on these observations, the probability that no primary user can be detected during r sensing periods, can be obtained as follows [151]:

$$P_i^{\mathrm{av}} = \prod_{i' \in \mathcal{N}_i^{\mathrm{E}}} (1 - P_{i'}^{\mathrm{f}})^R \cdot \prod_{k \in \mathcal{A}_i^{\mathrm{E}}(j)} e^{-\beta(j,k)R\Delta t} \tag{7.8}$$

where $R = \lceil T_m/\Delta t \rceil$ where T_m is the expected time of the user m to stay in the coverage. $\mathcal{N}_i^{\mathrm{E}}$ is a set of the neighbors of cell i located in its interference range, and $\mathcal{A}_i^{\mathrm{E}}(j)$ denotes a set of the PU activity regions of spectrum j in the interference range of cell i. The first term represents the probability that all extended neighbors do not generate any false alarms during r sensing periods. This is based on the OR rule in the decision fusion, and will change if other cooperative decision criteria are used. Here, the sensing operation is assumed to be performed in every Δt sensing period. The second term denotes the probability that no PU activity appears in the coverage during r sensing slots. Then the probability of the spectrum handoff due to the PU activity in the CR coverage can be obtained as $1 - P_i^{\mathrm{av}}$. //

Consequently, if the coverage is larger, the number of PU activity regions in the interference range increases, leading to decrease in P_i^{av}.

7.1.3 Theoretical Model for Spectrum Handoff

Spectrum handoff operations are known to be sensitive to the activities of primary user. To investigate this influence theoretically, Markov chain analysis for spectrum handoff in licensed bands presented and corresponding metrics such as forced termination probability,

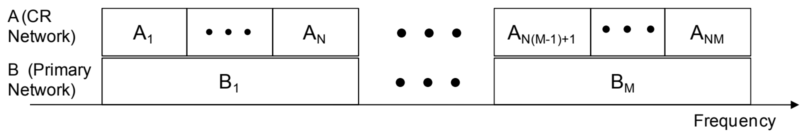


Figure 7.8 Frequency bands used by two types of radio systems.

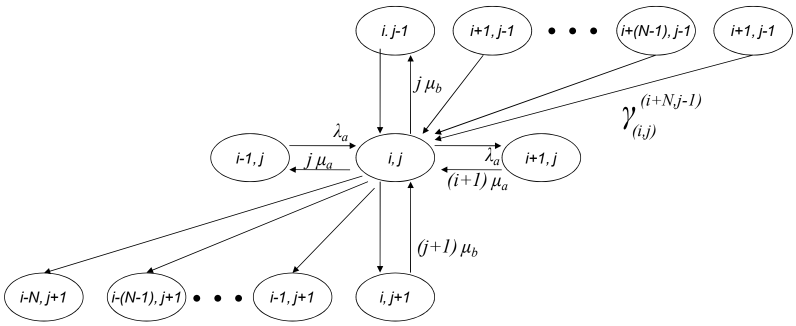


Figure 7.9 Markov chain cognitive radio without spectrum handoff.

blocking probability and traffic throughput, are derived [304]. In addition, a channel reservation scheme for cognitive radio spectrum handoff is proposed. This scheme allows the tradeoffs between forced termination and blocking according to QoS requirements. The proposed scheme can greatly reduce forced termination probability at a slight increase in blocking probability.

SYSTEM MODELING

Let there be two types of radio users, the primary users and CR users, operating in the same spectrum. The spectrum consists of M primary bands and each primary band is divided into N sub-bands. The CR users can use channels A_1 to A_{NM} while the primary users can use channels B_1 through B_M. The A and B channels overlap with each other, as indicated in Figure 7.8. The primary users have the priority to use the spectrum and can reclaim any sub-bands temporarily used by CR users. Therefore, the presence of CR users is entirely transparent to the primary users.

COGNITIVE RADIO WITHOUT SPECTRUM HANDOFF

The process of spectrum occupation is modeled as a continuous time Markov chain. It is characterized by its states and transition rates. The NM sub-bands are shared by the primary users and CR users. In this case states are described by (i, j), where i is the total number of sub-bands used by CR users and j is the total number of primary bands used by the primary users. Assume that the arrivals of CR users and primary users are assumed to be both Poisson processes with arrival rates λ_a and λ_b. The corresponding service times are exponentially distributed with rates μ_a and μ_b. As the primary users have the priority to use the spectrum, the CR users can be preempted by primary users.

Depending on the number of sub-bands occupied by the CR users in the newly preempted primary band, a forced termination in state (i, j) will move the state to one of $(i, j + 1)$, $(i - 1, j + 1)$, $(i - 2, j + 1)$, $\ldots$, $(i - (N - 1), j + 1)$, and $(i - N, j + 1)$ as shown in

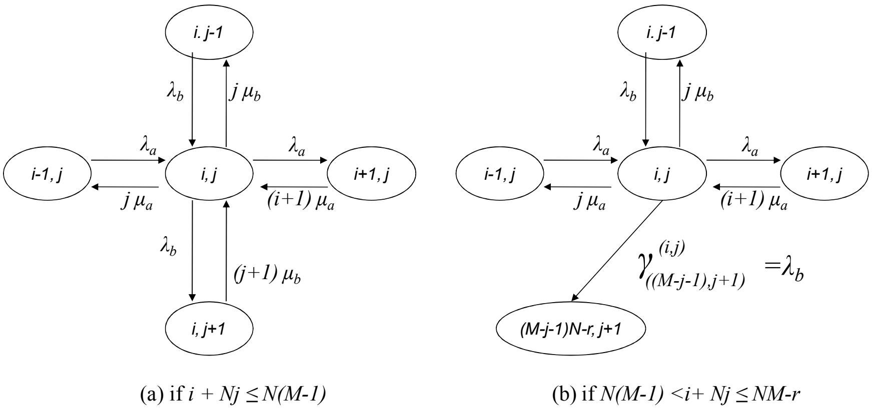


Figure 7.10 Rate diagram of state (i, j) with spectrum handoff and channel reservation [304].

Figure 7.10. The transition rate of a forced termination depends on the number of sub-bands k used by the CR users in this primary band. Let $\gamma^{(i,j)}_{(i-k,j+1)}$ denote the transition rate from state (i, j) to state $(i-k, j+1)$. This transition occurs when the k sub-bands are in the same primary band while the residual $(i-k)$ sub-bands are distributed in the other $(M-j-1)$ primary bands. With the classical partition problem in probability, the transition probability can be obtained by

$$\gamma^{(i,j)}_{(i-k,j+1)} = \frac{\begin{bmatrix} N \\ k \end{bmatrix} \begin{bmatrix} (M-j-1)N \\ i-k \end{bmatrix}}{\begin{bmatrix} (M-j)N \\ i \end{bmatrix}} \lambda_b, \quad k = 1, 2, \ldots, N \tag{7.9}$$

Let Φ denote the set of feasible states of the Markov chain shown in Figure 7.9 and $\phi(i, j)$ be an indicator function of Φ : $\phi(i, j) = 1$ if $(i, j) \in \Phi$, and 0 otherwise. Let $P(i, j)$ be the state probability. From Figure 7.9, the set of balance equations can be written as

$$\begin{aligned}
[j\mu_b + i\mu_a + \lambda_a + \sum_{k=0}^{N} \gamma^{(i,j)}_{(j-,j+1)}]p(i,j)\phi(i,j) = \\
\lambda_a p(i-1,j)\phi(i-1,j) \\
+ (j+1)\mu_a p(i,j+1)\phi(i,j+1) \\
+ (i+1)\mu_a p(i+1,j)\phi(i+1,j) \\
+ \sum_{k=0}^{N} \gamma^{(i,j)}_{(i-k,j+1)} p(i+k,j-1)\phi(i+k,j-1)
\end{aligned} \tag{7.10}$$

$$\sum_{i=0}^{NM} \sum_{j-0}^{M} P(i,j)\phi(i,j) = 1 \tag{7.11}$$

where $i = 0, 1, \ldots, NM$ and $j = 0, 1, \ldots, M$. The total number of states is $1 \sum_{i=1}^{M}(iN + 1) \approx NM^2/2$.

Forced termination represents a disruption of service and should be kept below a tolerable level. When the state transition is from state (i,j) to state $(i-k, j+1)$, k out of i CR users will experience forced termination. Therefore, forced termination probability, P_F , is

$$P_F = \sum_{i=0}^{NM}\sum_{j=0}^{M}\sum_{k=1}^{N} \frac{k}{i} \cdot \gamma_{(i-k,j+1)}^{(i,j)} P(i,j) \tag{7.12}$$

COGNITIVE RADIO WITH SPECTRUM HANDOFF

If spectrum handoff is allowed, the preempted cognitive radio calls will moved immediately to idle sub-bands elsewhere. This re-allocation of band can either be performed by the base station centrally or by the cognitive radios through a suitable distributed protocol. Therefore, as long as there are idle sub-bands around forced termination will not occur. Thus for state (i,j), if $i + jM \leq (N-1)M$, forced termination will not occur with the arrival of a primary user; otherwise, forced termination(s) will move state (i,j) to state $((M-j-1)N, j+1)$ with transition rate

$$\gamma_{(M-j-1)N,j+1)}^{(i,j)} = \lambda_b \tag{7.13}$$

OPTIMAL CHANNEL RESERVATION

If channel reservation is used along with spectrum handoff, the probability of forced termination can be significantly reduced, leading to increase in network capacity. Let r be the number of sub-bands reserved for spectrum handoff. ($r = 0$ corresponds to channel reservation not used) Figure 7.10 shows the transition rate diagram of state (i,j) with r reserved subbands. The blocking of CR users occurs when the current bandwidth occupancy $(i + Nj)$ plus r equals to the total bandwidth NM. Let $P_B(r)$ denote the blocking probability

$$P_B(r) = \sum_{i=0}^{NM-r}\sum_{j=0}^{M} \delta(i + Nj + r - NM)P(i,j) \tag{7.14}$$

With reserved sub-bands r, the forced termination probability becomes

$$P_F(r) = \sum NM - r \sum_{j=0}^{M} \frac{k}{i} \gamma_{(j-k,j+1)}^{(i,j)} \delta(i + Nj > (M-1)N)P(i,j) \tag{7.15}$$

From Eqs. (7.14) and (7.15), $P_B(r)$ and $P_F(r)$ can be tradeoff by adjusting r according to the QoS requirements. A natural way of choosing the optimal r is to maximize the throughput $\rho(r)$ of CR users, where throughput is defined as the average number of service completions per second, namely

$$\rho(r) = \sum_{i=1}^{NM-r}\sum_{j=0}^{M} P(i,j) i\mu_a \tag{7.16}$$

The optimal r can easily be obtained by enumeration.

7.2 CONNECTION MANAGEMENT

When the current operational frequency band becomes busy (this may happen if a licensed user starts to use this frequency) in the middle of a communication by an CR user, then applications running on node have to be transferred to another available frequency band. However, the selection of new operational frequency may take time. The objective of connection management is to provide a seamless communication channel to CR users by either mitigating the delay effect on the on-going transmission during spectrum handoff or minimizing spectrum handoff delay. Furthermore, given a switching delay, how to optimize each upper-layer protocol also is a crucial issue in the connection management functionality.

In the following sections, we focus on connection management functionalities from the perspective of both delay mitigation and delay minimization. Spectrum handoff issues in upper-layer protocols are covered in the chapters of routing and transport protocols (Chapters 9 and 10).

7.2.1 Handoff Delay Minimization

Spectrum handoff delay is the most crucial factor in determining the performance of spectrum mobility. This delay is dependent on the following operations in CR networks: First, the different layers of the protocol stack must adapt to the channel parameters of the operating frequency. Thus, each time a CR user changes its frequency, the network protocols may require modifications on the operation parameters, which may cause protocol reconfiguration delay. Also we need to consider the spectrum and route recovery time and the actual switching time determined by the RF front-end reconfiguration. Furthermore, to find a new spectrum and a new route, CR users need to perform out-of-band sensing and neighbor discovery. This also yield an additional latency for searching proper spectrum bands, but can be significantly reduced by jointly optimizing spectrum sensing and spectrum mobility capabilities.

Thus, recent research has mainly focused on delay minimization in out-of-band sensing through the search-sequence optimization, which is explained in Section 4.4. Furthermore, for fast spectrum discovery, IEEE 802.22 adopts the backup channel lists, which are selected and maintained so as to provide the highest probability of finding an available spectrum band within the shortest time [120]. In [142], an algorithm for updating backup channel lists is proposed to support fast and reliable opportunity discovery with the cooperation of neighbor users, which is investigated in this subsection in more detail.

SPECTRUM HANDOFF BASED ON BACKUP CHANNEL LISTS

To reduce a spectrum handoff delay caused by out of band sensing, IEEE 802.22 proposes channel management scheme based on backup channels. Once CR users detect the transmission of primary users in the operation channel, they should find a new channel for switching promptly. However, it generally takes long time to search all other channels after PU detection. Instead, in IEEE 802.22, CR users (BS and CPE) maintain multiple backup channels, which are obtained in advance through out of band sensing. This operation is performed simultaneously with in-band sensing that protects primary users (incumbents) during their transmission. Backup channels are sensed at every 6s, and is ordered according to their priorities. Whenever a CR user detects PU activity in the operating channel, it can pick a new channel with the highest priority from the backup channel list, which avoids a long channel searching time, and hence minimize the spectrum handoff delay.

In IEEE 802.22, CR users is allowed to access three different types of channels: *operating channel*, *backup channel*, and *candidate channel* [83]. The channels currently in use for CR communications are called *operating channels* , and idle channels other than operating and backup channels are referred to as *candidate channels*. An operating channel is released and added to the backup set by termination of the the transmission on the current operating channel, and the quality of the channel is within the range of the existing members of the backup channel set. If the quality of the channel becomes worse than that of the existing members of the backup channel set, a released operating channel is added to the candidate channel list. If the channel quality of any of the channel in the candidate set is better than that of an existing member of the backup set, it is added to the backup set. In this case, the poorest channel in the backup set becomes a member of the candidate set.

OPTIMIZATION OF BACKUP CHANNEL LISTS

For an efficient spectrum switching, the backup channel list should be properly constructed and maintained. In the following, a lightweight algorithm is presented to obtain optimal backup channel lists, and to maintain them throughout CR operations [142].

[Construction of Initial BCL]

When a backup channel list (BCL) is constructed initially, there could be many candidates for its entries. In IEEE 802.22, for example, there are 68 TV channels (channels 2 to 69) in the VHF/UHF bands (54.806 MHz) [60]. If CR devices are allowed to operate in other licensed spectrum bands, they may have more candidate channels.

Upon selecting N initial backup channels, two conflicting objectives must be met: the BCL should 1) contain as few channels as possible since the cost of channel sequencing grows fast at the rate $O(N^2)$, and 2) have many good channels to increase the chance of finding enough opportunities at the first opportunity-discovery attempt. To achieve both objectives, first, all M licensed channels are ordered according to the sensing-sequence algorithm in Section 4.4.2, and then the initial BCL is constructed by choosing the first N channels of the sequence where N is minimized while achieving a target performance.

The problem of constructing the initial BCL is formulated as follows. Suppose $L_M = \{l_1^*, l_2^*, \ldots, l_M^*\}$ is the (sub)optimally-ordered list of M channels. Also, let $L_N = \{l_1^*, l_2^*, \ldots, l_N^*\}$, $N \leq M$ be a sub-list of L_M with its first N entries. The objective is to find an optimal N such that N channels may contain opportunities more than B_{req} with probability P_{th}, which is a pre-defined threshold. This optimization problem is expressed as follow:

$$\begin{aligned} N^* &= \min\{N \bar{P}_{L_N}^{B_{\text{req}}} \geq P_{\text{th}}\} \\ \bar{P}_{L_N}^{B_{\text{req}}} &= Pr\{\sum_{i \in L_N} C_i I_{\Theta_i} \geq B_{\text{req}}\} \end{aligned} \tag{7.17}$$

where B_{req} is the capacity required to support spectrum demands from a CR user, and $C_{L_N}^{B_{\text{req}}}$ represents the probability that L_N may contain more opportunities than B_{req}. I_i is an indicator function such that $I_i = 1$ if channel i is idle, and 0 otherwise. C_i is the channel capacity, and P_{idle}^i is the idle probability of channel i.

[BCL Update Strategy]

Assume that channels have time-invariant ON/OFF distributions $f_{T_{\text{on}}^i}(t)$ and $f_{T_{\text{off}}^i}(t)$. Then, P_{idle}^i is accurately estimated through the accumulated sensing results extensively collected from M channels, and accordingly an optimal BCL is determined. This optimal BCL list does not need to change over time.

However, since channels are usually time-varying, the above strategy is not effective. To address this issue, the BCL can be updated periodically by obtaining the optimal scanning order of all M channels, causing a large computational overhead. So, an efficient and lightweight BCL update strategy is proposed that sorts BCL or candidate channel list (CCL) separately and only when necessary, with no sampling required on candidate channels. At every Bt_{UPDATE}, $C_{\mathrm{L_N}}^{B_{\mathrm{req}}}$ is calculated with most recent channel estimates, and accordingly BCL is updated where L_N is the current BCL with N backup channels. According to $C_{\mathrm{L_N}}^{B_{\mathrm{req}}}$, one of the following actions is taken:

- *Channel export* $(BCL \rightarrow CCL)$*:* If $C_{L_N}^{B_{\mathrm{req}}} > P_{\mathrm{th}}^{\mathrm{up}}$, we export a certain number of least preferred channels from BCL since it contains more channels than necessary. We use $P_{\mathrm{th}}^{\mathrm{up}} = P_{\mathrm{th}} + \epsilon_1(\epsilon_1 > 0)$ to avoid any impetuous channel export. To export channels, the (sub)optimal sequence of all N (not M!) backup channels is constructed and the optimal BCL size N^* is calculated again. Then, the last $(N - N^*)$ channels in the sequence are exported to CCL.

- *Channel import* $(BCL \leftarrow CCL)$*:* If $C_{L_N}^{B_{\mathrm{req}}} < P_{\mathrm{th}}^{\mathrm{low}}$, a number of candidate channels are imported from CCL to satisfy $C_{L'_N}^{B_{\mathrm{req}}} \geq P_{\mathrm{th}}$, where L'_N is an extended BCL after importing the CCL channels. We use $P_{\mathrm{th}}^{\mathrm{low}} = P_{\mathrm{th}} - \epsilon_2(\epsilon_2 > 0)$ to avoid impetuous channel import. To import channels, candidate channels are sorted in the (sub)optimal order, and are imported to BCL one by one in the order of preference until $C_{L'_N}^{B_{\mathrm{req}}} \geq P_{\mathrm{th}}$ is met.

- *Channel swap* $(BCL \leftrightarrow CCL)$*:* One may want to restrict the size of BCL within some range such as $N^{\mathrm{low}} \geq N^* \geq N^{\mathrm{up}}$. N^{low} helps reserve a minimal number of backup channels so that opportunity-discovery would be successful, and N^{up} upperbounds the computational overhead in sorting backup channels. When N^{low} and N^{up} are used, channel export (or import) cannot be processed if $N^* = N^{\mathrm{low}}$ (or N^{up}). In such a case, we swap the least preferred backup channel with the most preferred candidate channel if the swap helps decrease/increase $C_{L_N}^{B_{\mathrm{req}}}$ as desired.

- *Mandatory channel export* $(BCL \rightarrow CCL)$*:* In the proposed scheme, channels are categorized into two classes: (1) those with long ON/OFF periods (class-L), and (2) those with short ON/OFF periods (class-S). The former includes TV bands where ON/OFF periods are in the order of hours at least, and the latter includes 802.11 channels where ON/OFF periods typically last tens of milliseconds [184].

A mandatory channel export is triggered when a class- L channel (either in-band or backup channel) is sampled to be ‘ON’ (i.e., busy). Such a class-L channel is better to be expelled from BCL since the channel is unlikely to become available soon. Once expelled, the channel is forced to stay in CCL until its timer expires. The timer is a design parameter and can be uniquely determined for each channel. A similar concept was found in IEEE 802.22 [240], where a backup channel detected busy is marked as ‘occupied by PUs’ and never sensed until Non-Occupancy Period (recommended to be 10 minutes) expires. Note that candidate channels with their timer unexpired are not considered for channel import.

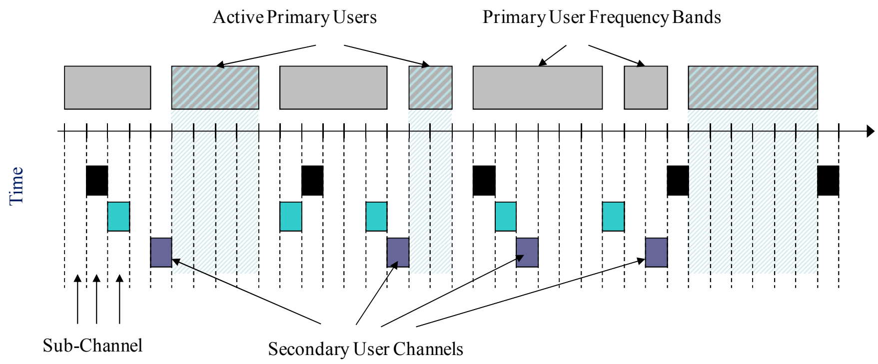


Figure 7.11 Spectrum pooling concept [30].

7.2.2 Handoff Delay Mitigation

Once the handoff delay is determined, how to reduce an adverse influence on the on-going communication of CR users is another important issue in spectrum mobility. When primary users appear in the spectrum band, CR users need to move to a new available band, resulting in a temporary communication break. To solve this problem, multiple non-contiguous spectrum bands can be simultaneously used for the transmission in the CR network. This method can create a signal that is not only capable of high data throughput, but is also immune to the the PU activity. Even if a primary user appears in one of the current spectrum bands, the rest of them will maintain current transmissions [5]. In this subsection, we investigate handoff mitigation techniques mainly focusing on multi-spectrum transmission.

SPECTRUM POOLING STRUCTURE

The most common approach based on this multi-spectrum transmission is a spectrum pooling [268] [30]. A Spectrum pool is defined as a (not necessarily contiguous) frequency range used by CR users. Each spectrum pool is divided into multiple sub-channels. Figure 7.11 shows the principle idea of a spectrum pooling system [30]. Primary users own different parts of the spectrum but may not be active at a certain time. Figure 7.11 shows three different active CR user communications. For each communication a pair of CR users picked a pattern of sub-channels to form a CR user link. The number of sub-channels may vary depending on the quality of the sub-channels, the bandwidth of a single sub-channel and QoS requirement for that connection.

A basic principle of spectrum pooling is that the sub-channels selected to create a CR user link should be scattered over multiple licensed bands, ideally only one sub-channel should be taken out of licensed band. This principle has a double significance. First, it limits the impact of the secondary user on the re-appearing primary user. Second, if a primary user appears during the lifetime of a CR link it would impact very few of the sub-channels used by the CR link. The communication peers using that link would have to immediately clear the affected sub-channel and would start to find a new free sub-channel instead, which mitigates an abrupt quality degradation significantly. This procedure of the CR link reconfiguration is shown in Figure 7.12.

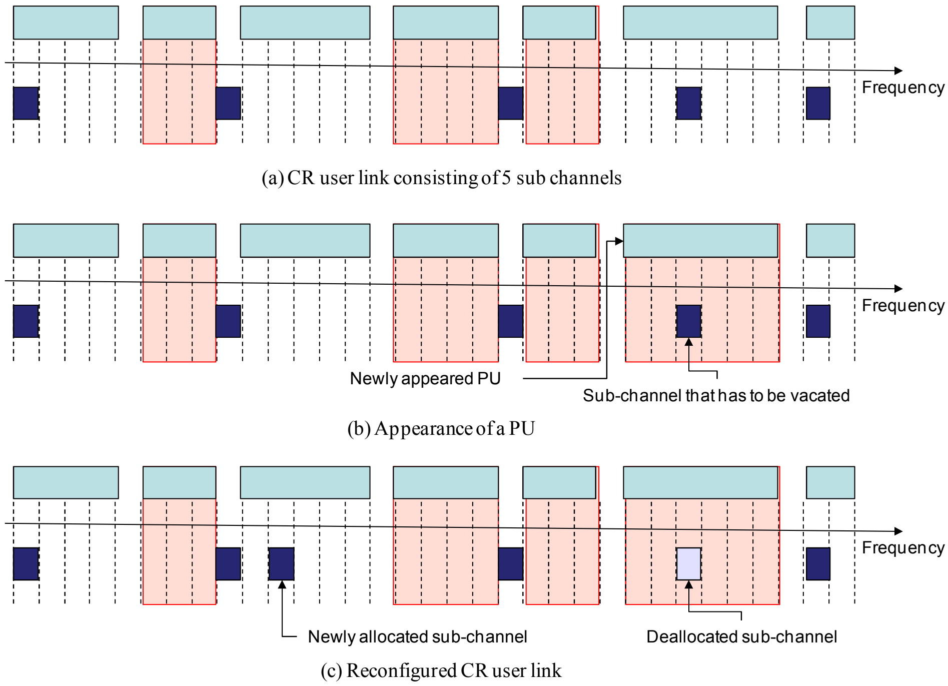


Figure 7.12 Reconfiguration of an SUL in spectrum pooling [30].

LINK MAINTENANCE APPROACH

To keep a continuous QoS in a spectrum pooling system regardless of primary user appearance, CR users should always have a redundant amount of sub-channels for their CR link. In [272], it is shown that there is an optimal number of redundant sub-channels X, i.e. an optimal amount of redundancy for any sub-channel interference probability p, leading to a maximum goodput. Detailed procedures are explained as follow:

Once a PU appears on a used sub-channel it interferes with the SU communication most likely resulting in a corruption of the data sent on this sub-channel. We use redundancy codes to protect data messages from the corruption due to PUI. An appropriate amount of redundancy added to the CR link enables the receiver to reconstruct data messages even if some sub-channels got interfered by a PU and the corresponding data packets got corrupted.

[Link Maintenance Model]

The system model in this method is based on the spectrum pooling architecture [30]. In the system model time is slotted into frames of length t_{frame}. For the general model we use the frame structure as shown in Figure 7.13 (a).

Within t_{frame}, t_{maintain} denotes the time reserved for link maintenance. Each time a sub-channel has to be excluded from the CR link, a new one needs to be acquired through the maintenance frame t_{maintain} to maintain the data-rate. t_{data} is the period of the frame reserved for payload data transmission.

t_{maintain} is further divided into three parts, namely t_{sens}, t_{control} and t_{acquire}. During t_{sens} the whole spectrum pool is scanned in order to detect PU activity. Subsequently .

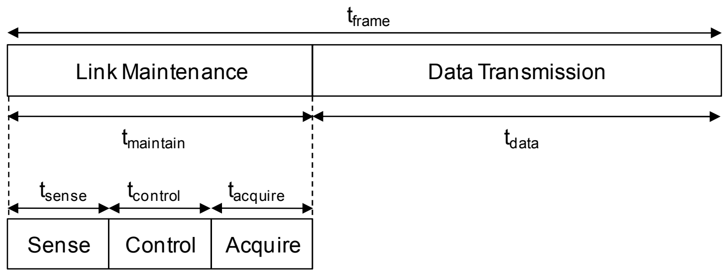


Figure 7.13 Frame structure for link maintenance [272].

within t_{control}. the sensing information has to be exchanged between the communication peers to achieve a consistent view which sub-channels have to be excluded and which can be used for communication. t_{acquire} denotes the time reserved for the acquisition of new sub-channels. Note that t_{acquire} is only necessary if one of the control messages sent during t_{control} indicates that some sub-channel got interfered by a PU. If no sub-channel got lost from the CR link, no new one needs to be acquired and consequently t_{acquire} is not needed. In this case, the next data message can be send right away.

[Link Maintenance Procedures]
Assuming that one encoded symbol(or one packet) is sent per sub-channel, the minimum number of sub-channels needed for the CR link is N, since at least N encoded symbols are required in order to completely retrieve the original data at the receiver. Using only N sub-channels, however, would cause the transmission to fail in case only one single sub-channel is reclaimed by a PU. In order to make the CR link robust against the corruption of data packets due to the appearance of PUs, X redundant sub-channels are added to the CR link, resulting in a total number of $N + X$ sub-channels for an CR link. This means that up to X arbitrary sub-channels can be interfered by a PU without degrading the goodput of the CR link.

The general idea of the proposed approach is shown in Figure 7.14. The transmitter takes a message consisting of $N + X$ packets, where X represents the number of redundant packets. The $N + X$ packets are then sent to the receiver using an CR link with $N + X$ sub-channels. During transmission, some of the packets may get lost due to interference by a primary user on some of the used sub-channels (indicated by the crosses in Figure 7.14). Apart from the appearance of a primary user, there are other reasons to yield the packet loss in CR link such as interference from CR users. However, these factors have been widely investigated in conventional wireless networks, and hence the only source of packet errors considered in this method is the interference from primary user appearance.

In 7.14), the CR link uses five packets of redundancy ($X = 5$) and four packets got lost, so the message can be reconstructed at the receiver out of the arriving packets. After sending of the data, transmitter and receiver both perform spectrum sensing to determine which sub-channels of the CR link need to be dropped due to interference by a PU. Subsequently, transmitter and receiver exchange the sensing results and control messages. The next step is the maintenance of the CR link (t_{acquire}), i.e. the acquisition of new sub-channels. If one of the control messages sent contains at least one sub-channel that has to be excluded from the CR link, a new one needs to be acquired during the t_{acquire} period in order to maintain

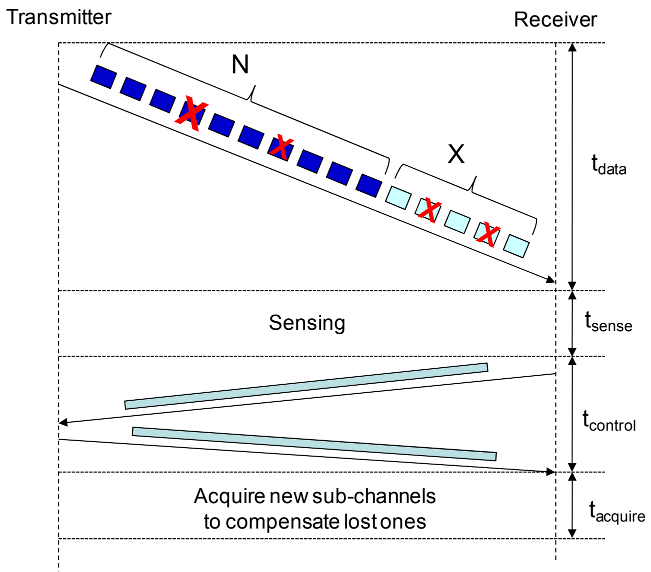


Figure 7.14 Message sequence chart for one frame period [272].

the goodput of the CR link. Only if both sender and receiver did not detect any PUI on one of the used sub-channels, the time needed to acquire new sub-channels (t_{acquire}) can be omitted and the next message can be sent right away.

[Determination of Redundant Sub-channels]

In order to compute the goodput, we need to compute the probability that a message cannot be reconstructed at the receiver. The message can be reconstructed at the receiver, if at least N packets are successfully received. That means, if more than X packets $(X+1, X+2, \ldots, X+N)$ get lost due to PUI, the message cannot be reconstructed and thus has to be sent again. Consequently, the message error probability (P_{err}) computes to

$$P_{\text{err}} = \sum_{i=1}^{N} \binom{N+X}{X+i} p_x^{X+i}(1-p_x)^{N-i} \tag{7.18}$$

Furthermore, in this model, the link maintenance probability P_m, denotes the probability that at least one sub-channel of the CR link cannot be used anymore, and consequently the link has to be maintained by acquiring a new sub-channel. Assume that the CR link consists of N sub-channels and p_x is the probability of a PU appearance in a sub-channel. Thus, each sub-channel has the probability p_x to be excluded from the CR link. Then, the link maintenance probability P_m can be expressed as follow:

$$P_m = 1 - (1-p_x)^N \tag{7.19}$$

Then, the average length of a frame is obtained by

$$t_{\text{frame}} = t_{\text{sens}} + t_{\text{control}} + P_m \cdot t_{\text{acquire}} + t_{\text{data}} \tag{7.20}$$

Based on Eqs (7.18) (7.19) and (7.20), the goodput of the CR link G_{bit} (in bps) can be derived as follow: is as shown in Equation 10.

$$G_{\text{bit}} = \frac{(1 - P_{\text{err}}) \cdot N \cdot b_{\text{sc}} \cdot t_{\text{data}}}{t_{\text{sens}} + t_{\text{control}} + P_m \cdot t_{\text{acquire}} + t_{\text{data}}} \tag{7.21}$$

where b_{sc} is the bitrate of each sub-channel.

SPECTRUM POOLING-BASED CR CELLULAR NETWORKS

In [151], a spectrum pooling based the network architectures is proposed for CR cellular networks. In the classical cellular networks, each cell uses different spectrum bands with its neighbor in order to prevent inter-cell interference. This concept can be also applied to CR cellular networks. Since the spectrum bands in the classical wireless networks such as wireless local area networks (LANs) are contiguous and located in the relatively narrow frequency range, mobile users can switch the spectrum without changing their RF front-ends. On the contrary, CR users need to reconfigure their operating frequency at the RF front-end whenever available spectrum bands changes, which causes significant switching latency.

[System Architecture]

To solve this problem, the spectrum pooling concept is extended to handle both spectrum and user mobilities in multi-cell environment. In the proposed architecture, the spectrum pool is defined as a set of contiguous licensed *spectrum bands*, each of which consists of multiple *channels*. Here spectrum pools are assigned to each cell exclusively with its neighbor cells while maintaining the frequency reuse factor, f. Although this architecture provides the seamless transition between spectrum bands within the pool, it still has difficulty in supporting seamless communication in CR users moving across different cells. To address this problem, in the proposed architecture, each cell has two different cell coverage types: basic area (BA) and extended area (EA). While the basic area is not overlapped with the coverage of its neighbor cells, the extended area has much larger coverage extended to the basic area of its neighbors. As a result, the spectrum pool consists of multiple *basic spectrum* bands that support only the basic area, and a single *extended spectrum* providing both the basic and the extended areas. Due to the extended spectrum, the current cell has another type of neighbors, referred to as *extended neighbors*. The extended neighbors are the cells that have the same spectrum pool within the interference range of the extended spectrum. In this architecture, unlike the basic spectrum, the extended spectrum in the current cell cannot be used in its extended neighbors so as to avoid inter-cell interference.

Mobility management in classical cellular networks is closely related to user mobility. However, CR networks have another unique mobility event, the so-called *spectrum mobility*. By taking into account both mobility events based on the proposed network architecture, we define four different types of handoff schemes as shown in Figure 7.15:

- *Intra-cell/intra-pool handoff:* The CR user moves to the spectrum band in the same spectrum pool without switching a serving BS.
- *Inter-cell/intra-pool handoff:* The CR user switches its serving BS to the neighbor BS without changing the spectrum pool.
- *Inter-cell/inter-pool handoff:* The CR user switches its serving BS to the neighbor BS, which has a different spectrum pool.

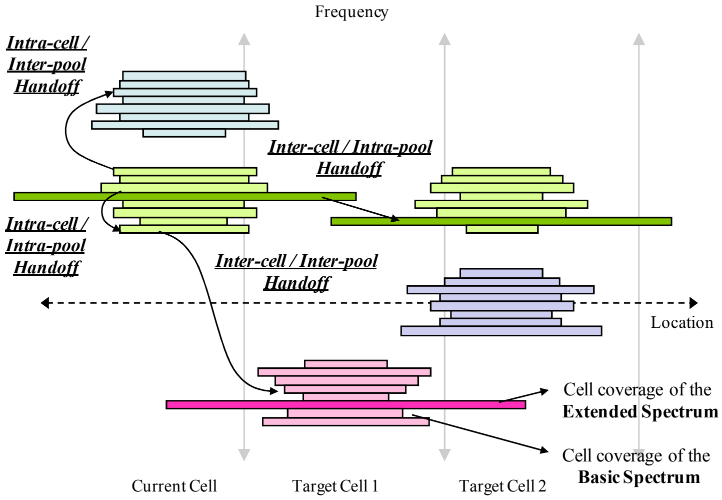


Figure 7.15 Different handoff types in CR networks.

- *Intra-cell/inter-pool handoff:* The CR user changes its spectrum bands from one spectrum pool to another within the current cell.

Each handoff type is related to different mobility event, and its performance is mainly dependent on both network and user conditions, such as resource availability, network capacity, user location, etc. Thus, CR networks require a unified mobility management scheme to exploit different handoff types adaptively to the dynamic nature of underlying spectrum bands.

[Handoff Modeling in CR Cellular Networks]
Based on these strategies, the handoff schemes defined above can be modeled as follows:

- *Intra-Cell/Intra-Pool Handoff:* Intra-cell/intra-pool handoff occurs when primary users are detected in the spectrum. Thus, it is implemented in a reactive approach. First, this handoff approach requires a preparation time to determine the handoff type (d_{prep}). After that, for sensing operations, CR users need to wait for the next sensing cycle, called a sensing synchronization time ($d_{\text{syn}}^{\text{sen}}$). Then, they sense the spectrum bands in the pool (d_{sen}), and determine the proper spectrum (d_{dec}). Finally, CR users move to a new spectrum band and resume transmission after the synchronization to the transmission schedule on that spectrum ($d_{\text{syn}}^{\text{tx}}$). Since spectrum bands in the pool are contiguous, CR users can switch the spectrum without reconfiguring their RF front-ends, and hence the physical spectrum switching delay is negligible. In summary, the latency for intra-cell/intra-pool handoff (Type 1) can be expressed as follows:
$$D_1 = d_{\text{prep}} + d_{\text{syn}}^{\text{sen}} + d_{\text{sen}} + d_{\text{dec}} + d_{\text{syn}}^{\text{tx}} \tag{7.22}$$

- *Intra-Cell/Inter-Pool Handoff :* If CR BSs can exploit multiple spectrum pools, intra-cell/inter-pool handoff may happen in PU activity. If the current spectrum pool does not have enough spectrum resource due to PU activity, CR users detecting PU activities switch to another spectrum pool in the current cell. This is also a reactive handoff. Thus, its handoff latency is similar to that of the intra-cell/intra-pool handoff as follows (Type 2):

$$D_2 = d_{\text{prep}} + d_{\text{recfg}} + d_{\text{syn}}^{\text{sen}} + d_{\text{sen}} + d_{\text{dec}} + d_{\text{syn}}^{\text{tx}} \tag{7.23}$$

However, unlike the intra-cell/intra-pool handoff, this scheme requires the reconfiguration of RF front-end since each spectrum pool is placed in the different frequency range. Usually reconfiguration takes longer time than other delay times.

- *Inter-Cell/Inter-Pool Handoff :*This handoff scheme is similar to that in classical cellular networks, which is required for CR users moving across multiple cells. To determine a target cell, a mobile CR user needs to observe the signals from neighbor cells during its transmission. However, since neighbor cells use different spectrum pools, the mobile CR user needs to stop its transmission and reconfigure its RF front-end in every observation of neighbor cells, which is a tremendous overhead in handoff. Thus, instead of this mobile station-controlled method, a network-controlled approach is more feasible for inter-cell/inter-pool handoff, where the BS determines the target cell based on the stochastic user information. As a result, mobile CR users need a single reconfiguration time. In this case, the BS can prepare the handoff in advance according to user mobility. Thus, this is a proactive handoff and does not requires the handoff preparation time d_{prep} used in previous reactive handoff types as follows (Type 3):

$$D_3 = d_{\text{recfg}} + d_{\text{syn}}^{\text{sen}} + d_{\text{sen}} + d_{\text{dec}} + d_{\text{sych}}^{\text{tx}} \tag{7.24}$$

  Furthermore, PU activities can initiate this handoff scheme in special reactive events. First, when all spectrum pools in the current cell are overloaded due to PU activity, the BS forces CR users to move to neighbor cells. This is exactly the same procedures as the intra-cell/inter-pool handoff, and requires D_2 handoff latency. Second, if a PU activity is detected in the extended spectrum, CR users in the extended spectrum should switch to the neighbor cells. Since there is no other available spectrum in the extended area after PU activity, they lose a control channel as well. To solve this problem, the BS determines handoff information and sends it to a selected target cell. Then, the target cell broadcasts the advertisement message for the CR user through its control channel. In this scenario, CR users need one or more reconfigurations of the RF front-end until it hears the advertisement message. Also in every reconfiguration, CR users monitor the control channel for a certain time (d_{lis}). The latency in this case (Type 4) can be expressed as follows:

$$D_4 = d_{\text{prep}} + \gamma(d_{\text{recfg}} + d_{\text{lis}}) + d_{\text{syn}}^{\text{sen}} + d_{\text{sen}} + d_{\text{dec}} + d_{\text{syn}}^{\text{tx}} \tag{7.25}$$

  Due to multiple reconfigurations, inter-cell/inter-pool handoff in this case shows the worst performance in terms of switching latency. γ is dependent on the searching order of neighbor cells. In this paper, the order is randomly chosen, and hence γ is considered as $(f+1)/2$ on average where f is a frequency reuse factor.

- *Inter-Cell/Intra-Pool Handoff:* This handoff happens when mobile CR users in extended areas successfully switch to the extended neighbors. This is also a proactive handoff. Furthermore, a new target cell is an extended neighbor which uses the same spectrum pool as the current cell, and hence reconfiguration is not required. Therefore, the latency for inter-cell/intra-pool handoff scheme (Type 5) can be expressed as follows:

$$D_5 = d_{\text{syn}}^{\text{sen}} + d_{\text{sen}} + d_{\text{dec}} + d_{\text{syn}}^{\text{tx}} \tag{7.26}$$

In this handoff, the latency is significantly reduced compared to that in other cases. Thus, this type of handoff is more advantageous to mobile CR users, and hence improves mobility in CR networks.

For seamless communication in dynamic radio environments, CR cellular networks should support intelligent connection releasing and re-establishing procedures during spectrum switching. When a CR user is moving, it needs to determine whether it should stay connected to its next hop forwarder through power control or immediately switching to a new neighbor. This has to be undertaken ensuring the network stays connected throughout the handoff procedure.

However, as presented above, there are diverse handoff types in the proposed CR cellular architecture, each of which has different characteristics with regard to total system capacity, and the handoff delay of mobile users. If a mobile user is determined to move an extended (large) area, it has advantages in terms of inter-cell handoff delay. But in this case, the total network capacity will decrease since mobile users require higher power or more bandwidth to support the same service quality as those in the basic (small) area. Furthermore, as explained in Section 7.1.2, mobile users in the extended area experiences a higher PU activity, and hence more spectrum handoffs than those in the basic area. To find a proper handoff type, when the mobility event is detected, first, a CR user estimates the switching cost of all possible handoff types, which includes the handoff delay in the selected handoff type, and the delay of the future expected mobility events. For example, when a CR user is in the boundary of the basic area, it has two possible options, to stay in the extended area of the current cell or to switch to the neighbor cell. Based on the analysis above, the CR user determines the handoff type with the lower expected spectrum cost. //

7.3 RESEARCH CHALLENGES

To the best of our knowledge, there exists few research effort to address the problems of spectrum mobility in CR networks to date. Although the existing work may lay the groundwork in this area, there still exist many open research topics:

- *Switching Delay Management:* The spectrum switching delay is closely related to not only hardware, such as an RF front-end, but also to algorithm development for spectrums sensing, spectrum decision, link layer, and routing. Thus, it is desirable to design spectrum mobility in a cross-layer approach to reduce the operational overhead among each functionalities and to achieve a faster switching time. Furthermore, the estimation of accurate latency in spectrum handoff is essential for reliable connection management.

- *Flexible Spectrum Handoff Framework:* As stated previously, there are two different spectrum handoff strategies: reactive and proactive spectrum handoffs, which show different influence on the communication performance. Furthermore, according to the mobility event, a spectrum switching time will change. For example, since a PU activity region is typically larger than the transmission range of CR users, multiple hops may be influenced by spectrum mobility events at the same time, which makes the recovery time much longer. Furthermore, spectrum handoff should be performed while adapting to the type of applications and network environment. In case of a delay-sensitive application, CR users can use a proactive switching, instead of a reactive switching. In this method, through the prediction of PU activities, CR users switch the spectrum before PUs appear, which helps to reduce the spectrum switching time significantly. On the other hand, energy constrained devices such as sensors need reactive spectrum switching. Thus, we need to develop a flexible spectrum handoff framework to exploit different switching strategies.

- *Optimization of Upper-Layer Protocols:* The purpose of spectrum mobility management in CR networks is to make sure that such transitions are made smoothly and as soon as possible such that the applications running on an CR user perceive minimum performance degradation during a spectrum handoff. Thus, a connection management functionality needs to coordinate the spectrum switching by collaborating with upper-layer protocols. To this end, the connection management protocols is required to be aware of the information about the duration of a spectrum handoff. Once the latency information is available, the CR user can predict the influence of the temporary disconnection on each protocol layer, and accordingly preserve the ongoing communications with only minimum performance degradation through the reconfiguration of each protocol layer and an error control scheme. Consequently, upper layer protocols should be transparent to the spectrum handoff and the associated latency, and multi-layer mobility management functionalities are required to accomplish the spectrum mobility functionalities. These protocols support mobility management adaptive to different types of applications. For example, a transmission control protocol (TCP) connection can be put to a wait state until the spectrum handoff is over. Moreover, since the TCP parameters will change after a spectrum handoff, it is essential to learn the new parameters and ensure that the transition from the old parameters to new parameters are carried out rapidly. This is an important but unexplored topic to date in spectrum mobility.

PART III

PROTOCOL DESIGN CHALLENGES

CHAPTER 8

MEDIA ACCESS CONTROL (MAC) PROTOCOLS FOR COGNITIVE RADIO NETWORKS

The medium access control (MAC) protocol at the link layer plays a fundamental role in CR communication by integrating several spectrum management functions, such as spectrum sensing, decision and sharing, apart from its classical responsibility of resolving channel contention, and providing fair, reliable link layer packet delivery. The CR MAC protocols are differentiated from the classical MAC schemes based on the close coupling with the physical layer and the spectrum usage in the neighborhood. As an example, the carrier sense mechanism at the MAC layer may not reveal complete information regarding the channel availability, owing to its inability to distinguish between the energy radiated by other CR users and the active PUs in the spectrum. If packets collide owing to contention caused by other CR users, it may simply be re-transmitted. However, the transmission must cease immediately if the packet loss is due to PU activity. To differentiate these two cases, the physical layer may provide supporting information to the MAC layer and influence its action by identifying the origin of the radiated power by baseband analysis of the spectrum shape. A general framework of the spectrum management functions and the MAC-layer coupling is shown in Figure 8.1. Based on the radio frequency (RF) stimuli from the physical layer *RF environment*, the *sensing scheduler* at the MAC layer can determine the sensing and transmission times. The availability of the spectrum, whenever a data packet needs to be sent, is coordinated by the *spectrum access* function. The *spectrum sensing* block plays a crucial role, both in terms of long term channel characterization and ensuring that the channel is available at the time of actual data transmission.

In this chapter, we discuss the design of the CR MAC protocols along two directions. First, we describe the efforts for *infrastructure-based* networks, in which a base station

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

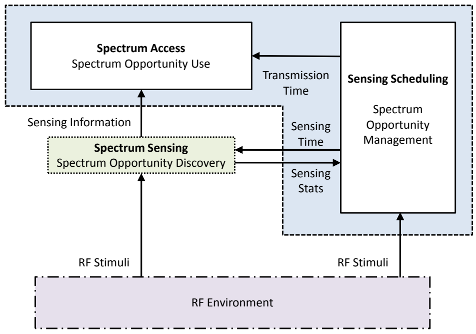


Figure 8.1 Spectrum functions at the MAC layer for CR networks

(BS) has the complete knowledge of the spectrum and CR user environment. Thus, the BS can conveniently regulate the spectrum sharing between the different users, negotiate policies with other CR systems, and provide guarantees on the link layer performance by scheduling transmissions in a globally optimal manner. For *distributed* networks, several scenario-specific MAC protocols that are optimized for a particular type of environment, or user specified application goal, may be used. As an example, nodes in an ad hoc network may exhibit high degrees of mobility with coarse time synchronization that makes sensing coordination difficult. For such cases, the MAC protocol may leverage the mobility to determine which regions (visited by the node during its motion) exhibit high levels of PU activity. The spectrum availability in a region may be *learnt* by the node over time, such that the choice of the transmission spectrum can be improved. Moreover, as the information collected by a node is mostly based on its own observations, the dedicated sensing duration in the initial stage of the learning process may be comparatively higher than the useful data transmission time. While longer sensing intervals provides higher protection to the PUs, it results in reduced throughput that may fail to meet the quality of service (QoS) requirements set by the user. Thus, the MAC protocol operation must be receptive to the application needs, and also support gathering of the necessary spectrum information. These opposing aims make standardization efforts particularly challenging in a distributed environment.

While this chapter is dedicated to the CR MAC protocols at the link layer, it is worth mentioning that several works adopt a cross-layer approach, wherein the channel access and spectrum management functions are indirectly implemented in the operation of a higher layer protocol, such as a routing or transport protocol. As an example, the TCP-based solution proposed in [53] may alter the sensing and transmission cycles to improve the path throughput. Moreover, several routing protocols have joint spectrum selection and path determination. Thus, the idea of the MAC layer being a *stand-alone* entity with a uniquely defined set of spectrum management functions may not always be the best design approach in CR networks. In the following sections, the key design concerns, the state-of-the art in the area, and the future directions of research are covered in detail. This chapter is organized as follows: First, we provide an overview of the component blocks of the CR MAC using Figure 8.1. This is followed by a discussion on the key features that a CR MAC protocol must incorporate. Then, we classify dynamic spectrum access MAC protocols based on their characteristic functionalities from perspective of infrastructure-based and ad hoc CR networks, respectively, respectively, while classifying them based on their characteristic functionalities. The chapter follows closely the survey on MAC protocols by the authors [63].

8.1 EFFECT OF SPECTRUM SENSING ON MAC PROTOCOL

We have covered different spectrum sensing schemes in detail in an earlier chapter, and the following material is intended to stress on some of the important sensing strategies, and their MAC layer implications. The choice of the spectrum sensing time followed by the transmission time deals with two main parameters optimized by the MAC protocol. While higher sensing times (T_s) ensure the correct detection of the spectrum, this may result in a comparatively smaller duration for actual data transmission (T_t) in the total time for which the spectrum may be used (T_f). This effectively lowers the link layer throughput. We discuss these functions further depending on whether they are undertaken independently of each other, or in a joint optimization framework. In addition, when sensing is undertaken over multiple channels in succession, the cumulative delay is significant, and is also covered in this section.

8.1.1 Determining sensing duration only

In this approach, the main objective is to find the sensing that minimizes the *missed detection* probability, i.e., determining the spectrum to be unoccupied when there is an active PU, and conversely, the *false alarm* probability, i.e., incorrectly inferring the presence of a PU in a vacant spectrum band. The sensing time (T_s) optimization problem is studied in [263] while keeping the transmission time (T_t) constant. Here, the channel efficiency is defined as the amount of the time that the idle channel can be utilized by the CR user for data transmission to the total time in a frame (T_f). From this definition, it is clear that the higher channel efficiency (or throughput) obtained by longer transmission time needs to be balanced with the detection accuracy. Towards this aim, the false alarm probability is derived in [263] based on classical detection theories. Moreover, a numerical optimization framework is proposed to solve the sensing time allocation problem, when the detection samples are uncorrelated.

In [97], apart from the sensing time on a single spectrum band, the time for searching multiple spectrum bands is also optimized. The operation is as follows: After transmitting within a certain spectrum band for the duration T_t, the CR user will undertake spectrum sensing. If there is no PU detected, it will continue to transmit in the same spectrum band. If indeed the CR user infers the presence of a PU, it has to search for a new spectrum band. For this, the node tunes its transceiver to another channel and starts spectrum sensing. If the spectrum is detected to be busy, the CR user needs to repeat this procedure on another band till a vacant spectrum band is identified. Thus, this process of sensing multiple spectrum bands in succession introduces significant delay, which must also be incorporated in the optimization program. In [97], an independent and identically distributed ON/OFF random process is assumed for the primary traffic. The search time for the vacant spectrum is minimized, thereby achieving maximum throughput for the CR network.

8.1.2 Determining transmission duration only

The optimal transmission duration (T_t) is derived in [201] while keeping the sensing duration (T_s) constant. The problem is formulated as a collision-throughput tradeoff problem, which finds the optimal value of the frame duration (T_f) for the CR operation. It integrates the minimum desired sensing time requirement and the traffic pattern of the PUs in its transmission time optimization function. The objective of the optimization is to maximize the throughput of the cognitive radio network while keeping the packet collision probability for the primary network under a certain threshold. For this, the authors assume exponential on-off traffic model for PUs but present a simplified treatment for the optimal frame duration.

8.1.3 Determining sensing and transmission durations jointly

In [150], a theoretical framework is proposed for jointly optimizing the sensing and transmission parameters in order to maximize the spectrum efficiency subject to interference avoidance constraints. With the goal of exploiting multiple spectrum bands, a spectrum selection and scheduling method is proposed, where the best spectrum bands for sensing are selected to maximize the sensing capacity. An adaptive and cooperative spectrum sensing method is also proposed that considers the number of cooperating users in a multi-user and multi-spectrum environment. In particular, sensing and transmission are performed in a periodic manner with separate observation and transmission periods. We believe that this approach is best suited for CR networks, as it balances the trade-off between sensing accuracy and the spectrum utilization efficiency. However, this approach is inherently more complex, and issues such as obtaining a solution to the optimization problem in real-time and the computational overhead must be considered.

8.1.4 Multi-spectrum Sensing Delay

The order in which the spectrum bands are chosen for sensing for the presence of PUs, called as the spectrum search sequence, and it determines the overall time used for searching the vacant spectrum. The performance of several such spectrum search schemes are investigated in [168]. An interesting approach is the consideration of correlated spectrum band occupancy models, in which it is more likely to detect a PU transmission in the neighboring spectrum of a band that is already known to be occupied. In addition, PUs may use several spectrum bands at a time, depending upon the way these bands are structured for use by the licensed users. The authors conclude their work with a general n-step serial search [168].

In [141], both sensing duration and spectrum search sequence optimization problems are jointly studied. The aim here is to discover as many spectrum opportunities as possible in advance, while minimizing the average time taken to detect a vacant spectrum band. The authors assume a semi-Markov traffic model for the PU spectrum usage and propose an estimation technique to learn the traffic pattern exhibited by the PUs. Moreover, the problem of deciding on an on-demand sensing schedule, as opposed to using periodic sensing, is investigated.

8.1.5 Research Challenges

The spectrum sensing involves several research challenges at the MAC layer that are described below:

- *Enforcing silence zones:* For accurate sensing, the measurements on the channel must be undertaken during *quiet* periods, when the other CR users in the neighborhood are silenced. This ensures correctly attributing the measured power to the PU signal alone. However, apart from the overhead of coordinating the silence durations, the lack of time synchronization in an ad hoc network makes it difficult to achieve these silence zones.

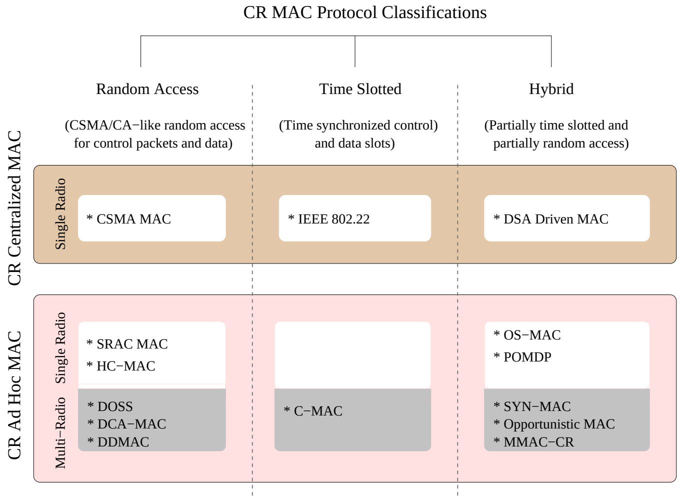


Figure 8.2 Classification of CR MAC protocols

- *Cooperation overhead:* Each CR user obtains a local estimation of the spectrum usage during the sensing intervals. By sharing these measurements with neighboring nodes, the total time for sensing in the same geographical area may be reduced. This allows for increased link throughput, though the benefit of cooperation must be carefully evaluated against the overhead of transmitting spectrum measurements.

- *Sensing and transmission coexistence:* The MAC protocol may specifically switch to pre-determined packet lengths, or artificially introduce inter-packet transmission delays to allow different CR users for sampling the channel free from intra-CR interference. This aid to sensing may be undertaken every fixed intervals of time, and there is a need for new analytical models that capture MAC protocol performance under these mixed-mode operation.

After identifying the available spectrum resource through spectrum sensing, the MAC protocol must now determine the spectrum access scheme. We next investigate this in the next section.

8.2 SPECTRUM ACCESS

Spectrum access enables multiple CR users to share the spectrum resource by determining who will access the channel, and for the duration for which it is reserved. In classical ad hoc networks, the request to send (RTS) and clear to send (CTS) mechanism is used to signal control of the channel and reduce simultaneous transmissions to an extent. In CR networks, however, the available spectrum is dynamic and users may switch the channel after a given communicating pair of nodes have exchanged the channel access signal. Thus, a fresh set of RTS-CTS exchange may need to be undertaken in the new channel. Moreover, the CR users monitoring the earlier channel are oblivious to the spectrum change on the link. They continue to maintain their timers and wait for the duration needed to complete the entire data transfer before initiating their own transmission. This leads to inefficient spectrum use, and new coordination mechanisms among the CR users is necessary whenever the spectrum

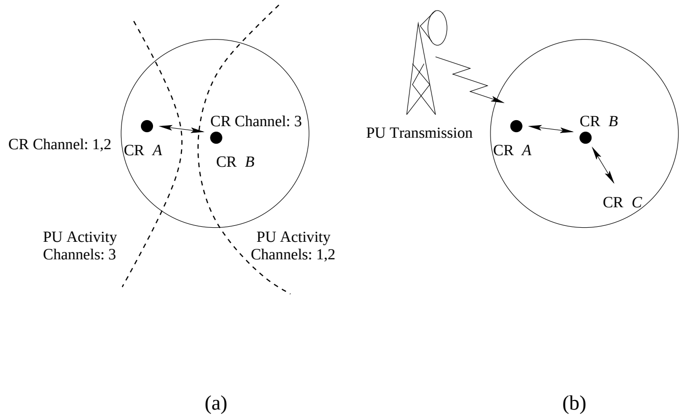


Figure 8.3 MAC layer spectrum access issues for CR ad hoc networks.

access conditions change. In Figure 8.3(b), the CR user C observes that the spectrum is currently being used by the CR users A and B. During the ongoing transfer, CR user A may detect a PU arrival, causing the spectrum on the link A-B to be changed. As this spectrum change occurs after the RTS-CTS control message exchange, user C continues to remain silent for the duration of the transfer specified earlier. This leads to lost spectrum opportunity as the PU detected by user A does not affect transmission by CR user C.

While both time slotted and random access schemes may be used in infrastructure-based networks, the difficulty in maintaining network-wide time synchronization in mobile ad hoc networks makes it infeasible to adopt completely slotted protocols.

In this chapter, we provide a thorough description of MAC protocols for both infrastructure-based and ad hoc CR networks. We classify the existing approaches into (i) random access protocols, (ii) time slotted protocols, and (iii) hybrid protocols, as shown in Figure 8.2. In addition, the number of radio transceivers also decides the working of the MAC protocol. We explain the classification as follows:

- **Random Access Protocols:** The MAC protocols in this class do not need time synchronization, and are generally based on the collision sense multiple access with collision avoidance (CSMA/CA) principle. Here, the CR user monitors the spectrum band to detect when there is no transmission from the other CR users and transmits after a backoff duration to prevent simultaneous transmissions.

- **Time Slotted Protocols:** These MAC protocols need network-wide synchronization, where time is divided into slots for both the control channel and the data transmission.

- **Hybrid Protocols:** These protocols use a partially slotted transmission, in which the control signaling generally occurs over synchronized time slots. However, the following data transmission may have random channel access schemes, without time synchronization. In a different approach, the durations for control and data transfer may have predefined durations constituting a super-frame that is common to all the users in the network. Within each control or data duration, the access to the channel may be completely random.

The main research challenges for spectrum access are described next.

8.2.1 Research Challenges

The challenges for efficient spectrum access are as follows:

- *Getting Neighborhood Information:* The use of non-uniform channels by different CR users makes topology discovery difficult. From Figure 8.3(b), we see that the CR users A and B experience different PU activity in their respective coverage areas and thus may only be allowed to transmit on mutually exclusive channels. The allowed channels for CR A $(1, 2)$ being different from those used by CR B (3) makes it difficult to send out periodic beacons informing the nodes within transmission range of their own ID and other location coordinates needed for networking.

- *Adapting to PU transmission*: Some PUs have specific transmission patterns, such as pre-determined spectrum usage times and durations, such as television broadcast stations, or may have occasional random access to the channel, such as public service agencies. At these times, the CR MAC protocol may infer the nature of the PU and adapt its own transmission to avoid both interference to itself and also prevent conflict with the PUs. Towards this aim, dynamic power control and transmission scheduling schemes need to be devised.

We note that spectrum access and spectrum handoff (moving seamlessly from one spectrum to another) are important features that play a role in the design of the MAC protocol. However, as these concepts are covered in great detail in the first few chapters of this book, we mainly focus on the complete MAC protocols that implement these functionalities.

We next describe in detail the existing CR MAC protocol implementations, beginning with the infrastructure-based networks.

8.3 MAC PROTOCOLS FOR CR INFRASTRUCTURE-BASED NETWORKS

These protocols need a central entity, such as a base station, that manages network activities, synchronizes and coordinates operations among nodes. However, the central entity is static and generally forms a single hop link with the mobile CR users that are within its coverage area. This architecture helps in the coordination among the CR users for collecting the information about the network environment, and allows the spectrum decisions to be localized.

We classify the existing works for such infrastructure-based or centralized networks based on the random access of the channel, time slotted behavior, and a hybrid approach that partially combines both of the previous access schemes, as shown in Figure 8.2.

8.3.1 Random Access Protocols

In [157], both the CR and PU network have their own BSs, possibly with overlapping coverage areas. The CR BS adaptively controls the transmission power and the transmission rate of the CR network, thereby ensuring minimized interference to the PU network. The MAC protocol is efficient in the sense that it requires a single transceiver, and in-band signaling. The CR users and the PUs establish direct single-hop connections with their respective base stations. The proposed MAC protocol allows simultaneous transmission of the CR users even when the PUs are detected, as long as the interference caused to them is contained within a pre-decided threshold. The operation of the protocol is as follows: The

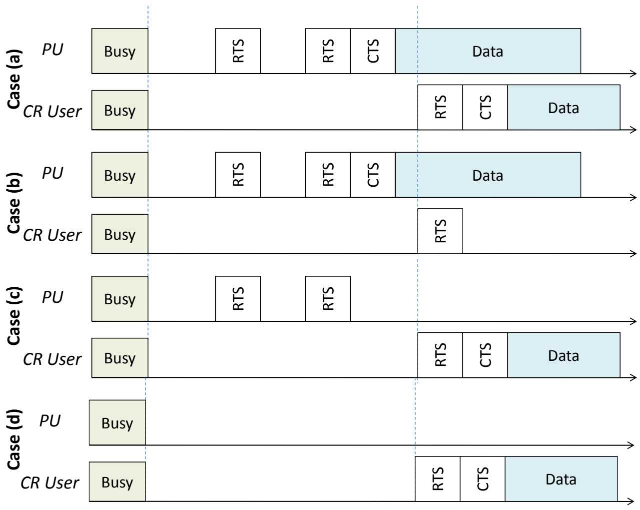


Figure 8.4 CSMA based protocol with four-way handshaking procedure

primary network follows classical CSMA, in which the PU undertakes carrier sensing for period τ_p before sending a request-to-send (RTS) packet to its base station. The primary base station may reply with the clear-to-send (CTS) if it is available for the data transaction. However, the CR users have a longer carrier sensing time (τ_s, where $\tau_s >> \tau_p$) so that priority of spectrum access is given to the PUs. Based on the (i) distance of the CR users from the CR base station, and the (ii) noise power, the base station decides the transmission parameters, namely the transmit power and data rate, for the current transfer. The CR user is allowed to send just one packet in one round of this negotiation in order to minimize the risk of interference to the other PUs.

Figure 8.4 shows the detailed protocol behavior in four different cases (a-d) plotted against a horizontal time axis:

- Case (a): Here, the PU gains the access to the channel after carrier sensing and backoff (or by retransmission following a prior collision), and sends its data. The CR user senses the channel for a period τ_s, and on finding the channel vacant (i.e. assuming the transmitting PU and the CR user are separated by a large distance), the CR user contends to gain the access to the channel through the RTS-CTS handshake. It then starts transmitting data with the power and rate suggested by the base station so that the concurrently occuring PU transmission is unaffected;

- Case (b): In this case, the RTS packet sent by the CR user experiences collision. The user must now wait for the next transmission opportunity after repeating the previous sensing process;

- Case (c): The PU sends repeated RTS packets but incurs collision each time. Here, the CR user can start transmission independently of the primary network, i.e. without adjusting its power and rate;

- Case (d): PU has no packet to send, thus the channel stays idle during the CR user's sensing period. Similar to the previous case, the CR user can start transmission without considering the primary network.

While coexistence is important, a significant interaction between the CR and the primary networks is implicitly assumed. The CR base station and users cannot determine if the PUs experience multiple failed transmission attempts (Case (c) above) without feedback from the primary network. Moreover, the transmission power for the CR users is only partitioned into two discrete levels (low or high) that does not reliably protect the PUs for all possible topologies. Moreover, there is no detailed explanation of how the transmit power, coding scheme, transmission rate are assigned to the CR users, especially considering the interdependencies that exist in these parameters.

8.3.2 Time Slotted Protocols

The MAC function provided in the IEEE 802.22 standard is an example of this class of protocols [59, 56] with fixed point-to-multipoint access. A complete discussion on the IEEE 802.22 standard has been included as a separate chapter, and we merely summarize the key features of the MAC protocol. A BS for such a network may have data rates in the range of $18 - 24\,\mathrm{Mbps}$, and coverage of $33 - 100\,\mathrm{km}$, partly due to the fact that this standard uses the TV channels that have better propagation characteristics.

The MAC protocol in this standard is time slotted. Here, the concept of a *superframe* is introduced that is composed of multiple smaller frames. A key novelty here is that the frame spans not only a finite duration of time, but also several channels on the frequency scale [61]. This allows for better diversity and increased system capacity. Moreover, the concept of a transmission "channel" is flexible, and may imply of a combination of the individual TV channels. To identify which specific channels are to be used, the BS sends a preamble at the start of each superframe that covers all the acceptable (or vacant) channels for communication. A special feature called as distributed sensing that allows the nodes to inform the BS of their local spectrum observations allows the BS to maintain an updated PU activity information throughout its large coverage region.

The main drawback of this protocol is that the control header exchange is extensive, which may result in lower data throughout or reduced channel utilization. Moreover, the time synchronization is difficult to maintain between the different CR base stations, we well as CR users in a given cell.

8.3.3 Hybrid Protocols

A game theoretic dynamic spectrum access (DSA) is proposed in [298]. The data transfer occurs in pre-determined time slots, while the control signaling uses random access scheme, making it a hybrid protocol. Moreover, this MAC is cluster based and the game policy in each cluster is managed by a central entity within the cluster. The proposed MAC protocol has high spectrum utilization, collision free spectrum access with QoS and fairness guarantees.

Four integral components can be recognized in the DSA-Driven MAC framework, as shown in Figure 8.5: (i) DSA algorithm, (ii) clustering algorithm, (iii) negotiation mechanism, and (iv) collision avoidance mechanism. Each of these functions is described in detail as follows:

8.3.3.1 DSA Agorithm The game theoretic DSA algorithm aims at pursuing a global optimization solution by reaching the Nash Equilibrium. In particular, the CR user behavior can be modeled as a repeated game model $\Gamma = \langle N, S_i, u_i, T \rangle$, where N is the set of players, S_i is the strategy of player i, u_i is the local utility function of player i, and T is the decision

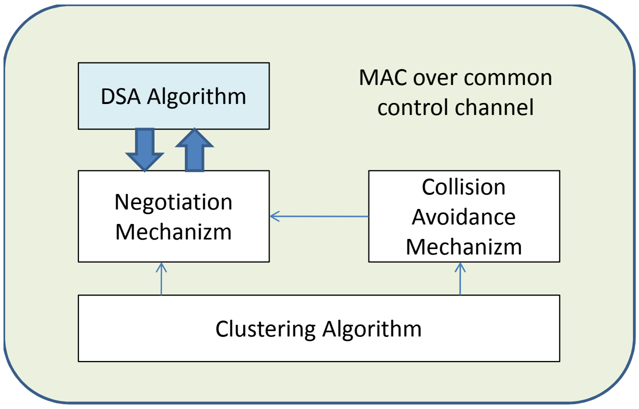


Figure 8.5 DSA-Driven MAC framework [298]

timing for the game. Therefore, each player keeps updating its strategy in order to maximize its own local utility function until the game converges to the Nash Equilibrium, after which a collision free channel access can be experienced. The utility function is composed of two components: the payoff or the gain obtained from the choice of the strategy, and the price the player should pay to the others for its strategy. The utility function may also take into account QoS and fairness requirement. From the networking viewpoint, the Nash equilibrium represents the assignment of spectrum access opportunities to all the CR users.

8.3.3.2 Clustering Agorithm For simplicity, a hexagonal cluster instead of a circular one is assumed. All the nodes within the hexagonal area are part of the cluster. The identity of each cluster is exclusively given by its position. When a node is added to the network, it can chose independently which cluster to join, based on the smallest distance from the cluster center. After joining a cluster, the node broadcasts with maximum power its coordinates and the cluster ID, so that all the other nodes within other clusters are aware of topology changes. The concept of virtual header (VH) is used, which is a packet unique to the cluster that also carries a token. The token contains the updated player list, i.e. accounts for the nodes joining/leaving the cluster. The beginning and termination of the VH propagation reflects the start and the end of one round of the game, respectively. The cluster head is the node to whom the VH is granted in that round.

8.3.3.3 Negotiation Mechanism The negotiation mechanism illustrated in Figure 8.6 deals with the control message exchange and coordination of the actions of the CR users. This negotiation occurs over a CCC and is composed of two phases: (i) inquiry stage and (ii) formal negotiation stage. The aim of the inquiry stage is identify the nodes that wish to start data communication. After this stage, the nodes that have packets to transmit will then become quasi-game players and will be considered in the formal negotiation stage.

When a node wants to start a new transmission, it sends a report packet to the VH node, thereby entering the inquiry stage. During this stage, all the cluster members build their player set and game strategy. Then, the VH is passed to the first player in the player set and the formal negotiation stage can start. The VH now carries the negotiation (NG) token which contains the dynamic game information required by the game players. In this way, the game players can update their local strategy and the related information in the NG

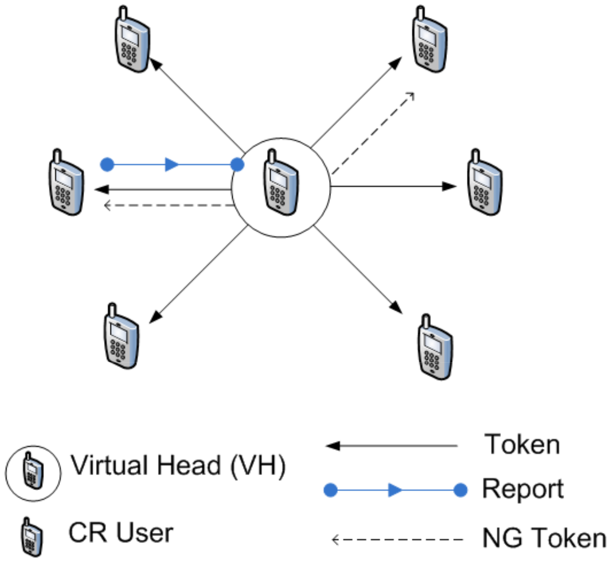


Figure 8.6 Negotiation process in the DSA-MAC [298]

token. The NG token is then passed to the next player in the list. The formal negotiation stage ends when convergence to the Nash equilibrium is achieved.

An example summarizing the entire negotiation mechanism is shown in Figure 8.6. Node 2 wants to start a transmission and reports the request to the VH node. During the inquiry stage a token inquires all the cluster members but only nodes 2 and 4 become quasi-game players, as they have packet to transmit. Then, the formal negotiation stage is carried out in order to coordinate nodes 2 and 4 to process the formal game.

8.3.3.4 Collision Avoidance Mechanism This mechanism aims at avoiding collisions during negotiations in different clusters and relies on out-of-band busy tones. Two different types of busy tones are exploited: *inside-cluster* and *outside-cluster* busy tones. Inside-cluster busy tone is set up by a node receiving a message to prevent other nodes external to its cluster from interfering with the ongoing negotiation. Outside-cluster busy tone is set up by a node overhearing messages from other clusters in order to avoid initiating a new round of negotiation within the cluster, as it may result in interference.

The time and the number of iterations taken to converge to the Nash equilibrium maybe prohibitively large. Moreover, the proposed scheme does not provide sensing support but assumes them to readily available. Enforcing the quiet period necessary for sensing is difficult in such conditions. Another drawback of the above protocol is the low scalability as the negotiation delay increases with the number of players. The difficulty in maintaining synchronization and possible collisions in the game information packets are some of the other factors that affect the protocol performance.

8.4 MAC PROTOCOLS FOR CR AD HOC NETWORKS

These protocols do not have a central entity for the operation of the network. Though the resulting architecture is scalable and has flexible deployment, the distributed spectrum sensing, sharing and access necessitate increased cooperation with the neighboring nodes. Maintaining time synchronization throughout the network and obtaining the information from surrounding nodes with minimum overhead are some of the factors that must be considered in the protocol design. We describe the existing works next based on the classification given in Figure 8.2.

8.4.1 Random Access Protocols

This class of protocols is specially suited for ad hoc networks. Some of these protocols have support for multiple radio transceivers [171] [198], while others use a single radio [172] [132].

8.4.1.1 **Dynamic Open spectrum Sharing (DOSS) MAC** Most of the works assume that a set of fixed non-overlapping spectrum bands are given, and a node can use only one of them at a time. However, if nodes are allowed to dynamically combine the available bands, it will result in better network performance. The Dynamic Open Spectrum Sharing (DOSS) MAC protocol provides an innovative solution to address the hidden node and exposed node problem [171]. Three radios are assigned distinctly to the control, data and busy-tone band, respectively. The spectrum bands used for data transfer are mapped to the frequencies in the busy tone band. Thus, whenever a node transmits or receives data on a given channel, it also emits a busy signal in the corresponding busy tone band.

The DOSS MAC protocol consists of the following steps.

- *PU Detection*: Since the CR users can use the spectrum only when the PUs are absent, the CR node continuously monitors the spectrum in its vicinity.
- *Set-up of Three Operational Frequency Bands*: First, the bands used for data transmission are determined, which could be a set of non-contiguous frequencies. Then, an out-of-band CCC is chosen for control signaling. The protocol proposes (i) traffic limiting, (ii) bandwidth ratio setting, and (iii) control channel migration techniques to alleviate the control channel saturation problem.
  - Traffic limiting: traffic which goes over the control channel is kept under a certain threshold ;
  - Bandwidth ratio setting: the bandwidth ratio of the control channel and the data channel is adjusted to avoid the control channel to become a bottleneck;
  - Control channel migration: after fixing the initial control channel during the network setup, it can slowly migrate towards a better channel.

  Thirdly, a band is exclusively reserved for busy tones, the frequencies in which have a direct one-to-one mapping with the data transmission bands.
- *Spectrum Mapping*: The mapping of the busy tones with the data transmission bands are done as follows. The busy tone band has a frequency range $[f_l, f_u]$, and the data band is contained in $[F_l, F_u]$. The data spectrum is considerably larger, i.e. $F_u - F_l >> f_u - f_l$ and this linear mapping allows neighboring nodes to realize that

the spectrum is actually used by another CR user by observing the corresponding busy tone.

- *Spectrum Negotiation*: The sender sends a request (REQ) packet with the available spectrum bands at its end. The receiver then replies with a REQ_ACK packet containing the information of a mutually acceptable spectrum band. According to the spectrum mapping, it also issues the corresponding busy tone, telling its neighbors not to transmit on the chosen data spectrum band. Upon receiving the REQ_ACK, the sender tunes its data transmitter to the negotiated band.
- *Data Transfer*: If a packet is correctly received, the receiver replies with DATA_ACK packet and turns off the busy tone. By receiving the DATA_ACK packet, the sender knows the transmission is successful. Otherwise, after a timeout, it will retransmit the data packet.

Apart from avoiding intra-CR network interference, we believe that this solution can also be applied to coordinate the MAC layer sensing. A node may sense on the channel which does not have a corresponding busy tone, thereby ensuring that the transmission of the other CR users are not mistaken for the PU activity.

The main drawback of this protocol is the use of separate and out-of-band spectrum for issuing the busy tones and for the CCC. Thus, the spectrum is not efficiently utilized. Moreover, the need for multiple transceivers is not justified as two of them are not used for data communication at all.

8.4.1.2 Distance-dependent MAC (DDMAC) The effect of transmission power is considered in the DDMAC protocol [227]. . The main contribution of this work is a probabilistic channel assignment mechanism that exploits the dependence between the signalŠs attenuation model and the transmission distance, while incorporating the application-specific needs. In particular, DDMAC assigns channels with lower average SINR to shorter transmission distances and those applications that require infrequent use, and vice versa. The key insight used in this work is that the path loss $\mathcal{F}$ is inversely proportional to the square of the transmission frequency f. Note that this characteristic of $L \propto \frac{1}{f^2}$ has also exploited in other works [50, 51]

The first variant of the proposed algorithm assumes that the traffic patterns of all nodes are known in advance. Hence, an optimization is easily constructed that aims to maximize the total number of simultaneous transmissions in a given neighborhood, even if this comes at the cost of individual gain of the node. The classical constraints include the consideration that similar channels cannot be assigned in overlapping areas, a bound on the maximum number of channels for a given transmission, and satisfying a minimum threshold SINR. However, this is an NP-hard problem, and it is unclear how a CR node is able to receive updated knowledge of the neighboring nodes (over multiple hops) to arrive at the best solution.

In the second variant, the CR users form concentric transmission circles around their location, each circle representing a feasible SINR value. The nodes listen to the control messages transmitted over a control channel to calculate how many transmitter-receiver pairs are likely to be active, and the overlap of these active data transfer regions with its own transmission circles. Thus, a node calculate probabilistically, the extent of spectrum usage in each of its transmission circles, and is able to assign channels based on the external activity, and the level of desired SINR. The resulting MAC protocol is a simple CSMA/CA scheme with a few new packets introduced for exchanging power and channel information among the nodes.

This work has the following limitations. First, the calculation of distance based on a the known transmission power of a single (RTS) packet is prone to errors in an uncertain channel. The possibility of not overhearing the control packets (and hence, miscalculating the spectrum usage) is non-negligible. Moreover, this work has ideal assumptions of perfect circular transmission regions, and well defined SINR gradation with distance that limit its use in practical settings.

8.4.1.3 **Distributed Channel Assignment (DCA) based MAC** A simple extension of the IEEE 802.11 CSMA/CA protocol using distributed channel assignment (DCA) is proposed in [198]. It uses multiple transceivers, with a dedicated out-of-band CCC for signaling. In addition, the proposed protocol also utilizes spectrum pooling which helps to enhance spectral efficiency by reliably detecting the primary network activity, thus serving as physical layer signalling.

The operation of the protocol is as follows:

- *Maintaining Spectrum Information:* Each mobile host maintains two data structures called the (i) current usage list (CUL), and (ii) the free channel list (FCL). Each node's CUL list records information of its neighbors including their addresses and the corresponding data channels utilized by them as well as the expected time of use. The FCL can not be derived from the CUL and continuously updated to determine the available spectrum opportunity.

- *Data Transfer:* This process is similar to the data transfer stage of [171], where the FCL is matched at both the sender and receiver ends using the RTS-CTS handshake. These messages also serve to silence the neighboring CR users, as seen in the classical IEEE 802.11 operation.

The use of a separate CCC results in wastage of the spectrum and may also become the bottleneck on the link. Moreover, there is no specific support for spectrum sensing or PU related adaptation that is required for CR networks. The protocol proposed in [243] has similar functioning and drawbacks, but uses a single transceiver that alternates between monitoring the CCC and the data spectrum bands.

8.4.1.4 **Single-Radio Adaptive Channel MAC (SRAC) Protocol** The single-radio adaptive channel (SRAC) algorithm is proposed in [172] that adaptively combines spectrum bands based on the CR user requirement, called as *dynamic channelization*. In addition, it uses a frequency division multiplexing (FDM)-like scheme, called as *cross-channel communication*, in which a CR user may transmit packets on one spectrum band but receive messages on another. These two features are described as follows:

- *Dynamic Channelization:* First, the basic spectrum unit (say, b) is decided, and the actual spectrum used is considered as an odd multiple of this unit (say, mb). Thus the number of possible transmission bands is much larger than the actual spectrum bands present, as the latter can be grouped differently by varying the multiplier m. Based on the spectrum demand, the usable transmission spectrum can be adaptively changed. Moreover, the spectrum bands are characterized based on the observed load and the usage by the PUs.

- *Cross-channel Communication:* In order to avoid frequency jamming and PU activity, a CR user may use different transmission spectrums for sending and receiving. This also allows for reserving larger spectrum for sending data, while the return

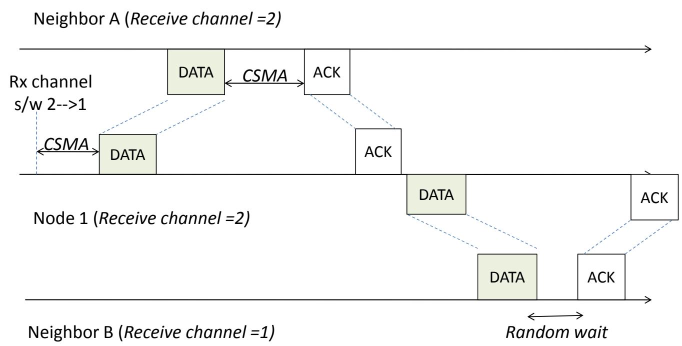


Figure 8.7 The cross-channel communication in the SRAC-MAC [172]

acknowledgments may be received over smaller spectrum bands for efficient utilization of the spectrum. Each node maintains the list of spectrum bands used by the neighbors for receiving. Figure 8.7 shows the case where node 1 has two neighbors A and B. Their respective receive channels are 2, 2, and 1 respectively. If the node 1's receive spectrum needs to be changed, it sends a notification packet in the receive bands of its neighbors (A and B) and immediately switches the spectrum. If the nodes hears acknowledgements (ACKs) from all its neighbors in the new spectrum band, the notification is completed. On the other hand, it continues to broadcast the new spectrum in the receive bands of the neighbor nodes (say node B that has yet not replied) till all the ACKs are received, or a retry limit is reached.The channel access is assumed to be CSMA with random wait

However, this work does not completely address the means to detect the presence of a jammer and distinguish malicious activity from legitimate network conditions. Though this approach uses a single radio, it will result in significant *deaf* periods, where control messages not sent on the receive spectrum band of the node will not be monitored. Moreover, the signaling overhead for maintaining updated receive spectrum bands of all the neighbors continuously adds to the traffic.

8.4.1.5 **Hardware Constrained MAC (HC-MAC)** The Hardware-Constrained MAC [132] protocol aims at efficient spectrum sensing and spectrum access by considering the hardware constraints, such as, the operational limitations of a single radio, partial spectrum sensing, and spectrum aggregation limits. It uses a CCC, but also has a single radio that simplifies the hardware requirements.

Hardware constraints can be divided into two classes given by (i) sensing constraints and (ii) transmission constraints. The sensing constraints concern the tradeoff between time taken for sensing and the resulting accuracy. As an example for *fine* sensing, a larger proportion of time needs to be allocated per channel, and hence a limited portion of the spectrum may be scanned. On the other hand, the transmission constraints are related to the limitations posed by the orthogonal frequency division multiplexing (OFDM) that decides the bandwidth range, as well as the maximum allowed number of the subcarriers. The distinct contributions made this protocol are as follows:

- *Sensing Decision:* In order to determine how many channels should be sensed, a *stopping* rule for successive channel sensing must be decided. By choosing a greater number of channels, the available bandwidth increases, leading to a higher data rate or reward. However, the cost of sensing, especially if the channel is found to be occupied and unavailable for use, must also be considered. The proposed finite horizon stopping rule chooses a time to stop channel searching such that the expected reward is maximized. The choice of how many channels to sense is also determined by the maximum allowed spectrum bandwidth that can be access by the transceiver at a given time, and also by the maximum number of permissible subcarriers that can be used from the available channels in this range. The authors propose backward induction to solve this problem, and complexity reduction techniques are given that reduce the computational time, especially if the number of channels is large.

- *Protocol Operation:* The MAC protocol is constituted by the operations of (i) contention, (ii) sensing, and (iii) transmission. In the contention phase, the C-RTS and the C-CTS packets sent over the CCC are used for gaining access to the channel. The transmission pair that wins the contention then exchange S-RTS and S-CTS packets for each channel that is sensed. At the end of each sensing round, the decision is made on whether to initiate the sensing on a new channel, based on the stopping rule. After the channels are decided by the node pair, the data transmission begins and multiple channels found during the sensing may be used. Finally, the T-RTS and T-CTS packets are exchanged on the CCC signaling the end of this transfer and releasing the channel for other users.

A key difference of this protocol as against the previous work is that the sensing at either ends of the link is initiated *after* a pair of CR users win the contention on the dedicated CCC. However, the control messages used for channel negotiation may not be received by the neighboring nodes if they are engaged in their own data transfers. Moreover, the number of control messages are significant and may saturation the control channel earlier than classical single channel RTS-CTS based MAC protocols.

8.4.2 Time Slotted Protocols

In this section we look at protocols that have set durations, and intervals for transmission, exhibiting time slotted behavior.

8.4.2.1 **Cognitive MAC (C-MAC)** The synchronized and time slotted cognitive MAC (C-MAC) [57] protocol is aimed at higher aggregate link throughput and robustness to spectrum change using multiple transceivers. C-MAC includes two key concepts: the rendezvous channel (RC), and the backup channel (BC). The RC is selected as the channel that can be used for the longest time throughout the network, without interruption among all other available choices. It is used for node coordination, PU detection, as well as multi-channel resource reservation. The BC, determined by out-of-band measurements, is used to immediately provide a choice of alternate spectrum bands in case of the appearance of a PU.

In C-MAC, each spectrum band has recurring *superframes* composed of a beacon period (BP) and a data transfer period (DTP), as shown in Figure 8.8. The RC is used on a network-wide communication, neighbor discovery, and sharing of load information for each band. Moreover, this is also used to exchange the schedules for the BP, so that the beacons are not simultaneously sent over all the spectrum bands. Upon power-up, each CR user scans all the available spectrum bands to determine the vacant spectrum resource. In these bands, if

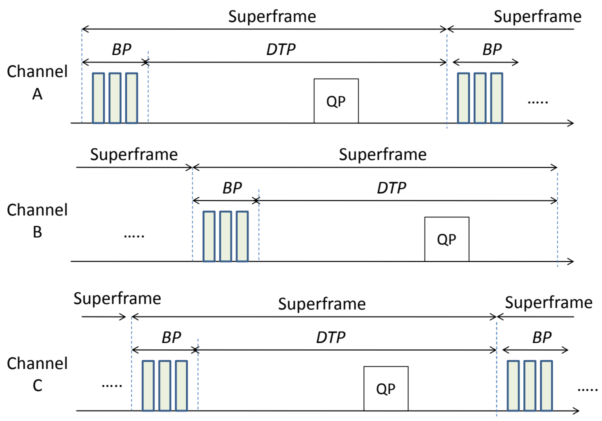


Figure 8.8 Multi-channel superframe structure in C-MAC (each channel is structured in the form of superframes whose beacon periods (BPs) are non-overlapping across channels) [57]

it hears a beacon, then it may choose to join that specific band and also set the global RC to the band specified in the beacon.

The working of the protocol is described in the following steps:

- *Distributed Beaconing:* Each BP is further time slotted so that the individual CR users may issue their beacons without interference. By re-broadcasting the received beacon information in its own beacon slot, a CR user helps to inform its neighbors of the other devices that are present at a distance greater than the transmission range.
- *Inter-Channel Coordination:* The CR users periodically tune to the RC and transmit their beacons. If they need to establish a new data spectrum band, this is communicated over these beacons. Any spectrum change that occurs in C-MAC must first be announced by the CR users over the RC, before using that band. In addition, the periodic tuning to the RC allows the CR users to re-synchronize and obtain the recent neighborhood topology information. The superframe structure (Figure 8.8) is now used in the new spectrum band.
- *Coexistence:* The time-slotted nature of the protocol allows the establishment of non-overlapping quiet periods (QP) for each of the spectrum bands (Figure 8.8). This ensures that the PUs are differentiated from the CR users and correctly detected. Moreover, the beacons are transmitted with the most robust modulation and coding so that these packets that signal the presence of the PUs are reliably received. At this time, one of the spectrum bands from the BC is chosen, similar to the IEEE 802.22 protocol described in Section 8.3.2.
- *Load Balancing:* The load balancing mechanism in C-MAC is achieved by accumulating the load statistics from the analysis of beacons, which carry the node traffic reservation information for the current superframe.

The main drawbacks of C-MAC are the following: All the beacons sent by the CR users must be accommodated in the BP of a superframe, which results in low scalability. Moreover, it is expected that the RC converges to a constant spectrum band over time, which cannot be guaranteed in distributed networks. Moreover, the spectrum switching is not instantaneous - the information must first be disseminated to the other CR users in the beacon period of the RC. It is unclear how the non-overlapping nature of the BPs and the quiet periods are enforced without the presence of a central entity.

The limits associated with the use of a RC are circumvented by the distributed slotted protocol proposed in [294], which provides in-band signaling through a dedicated control window in addition to the beacon and the data transfer periods. Furthermore, during this window, the bridge nodes are allowed to use multiple channels, i.e., to access more than one coordination group in each superframe for optimizing the performance.

8.4.3 Hybrid Protocols

In this section we describe MAC protocols that have both a time slotted and random access component.

8.4.3.1 **Opportunistic Spectrum MAC (OS-MAC)** The OS-MAC protocol uses predetermined window periods for coordinating the choice of spectrum among the CR users and exchanging control information to separate the latter intro groups [109]. However, within each window, the spectrum access is random, and hence this a hybrid protocol.

The spectrum bands used for data communication are considered to be non-overlapping and a separate CCC is assumed for exchanging control packets between users on different bands. It uses a single radio that needs to switch between the data band the CCC. The protocol operation is described as follows:

- *Network Initialization Phase:* Here the CR users form clusters, such that all the members of the same cluster wish to communicate with each other. The new user has an option of either forming its own cluster or joining one of the existing ones. During this entire stage of forming the cluster membership, the CR user keeps its radio tuned to the CCC. At any given moment, only one CR user is active in a cluster called the delegate.

- *Session Initialization Phase:* Here, the active delegate chooses a spectrum band for the group and communicates this to all the members of the cluster.

- *Data Communication Phase:* The members of the cluster use IEEE 802.11 DCF for accessing the spectrum band. At the same time, the active delegate monitors the CCC for collecting information of the spectrum environment. It then informs its own cluster members of a change in the spectrum band, if needed.

- *Update Phase:* Each cluster delegate now sends the traffic information of its own cluster to the other delegates over the CCC, and returns back to the currently used spectrum band at the end of this transfer.

- *Select Phase:* On learning of the spectrum usage statistics of the neighboring clusters, the cluster delegate may initiate changing of the spectrum used in the cluster. This is done by using a smaller wait duration between consecutive packets, so that the delegate wins the contention and transmits the new spectrum choice with higher priority.

- *Delegate Phase:* The role of the delegate is now passed onto another CR user in the same cluster for the next round of the protocol operation.

All these phases occur sequentially and have window durations determined by their respective timers. Moreover, these durations are flexible, and can be chosen so that each CR user in the cluster can access the spectrum band in a fair manner.

The OS-MAC protocol has several drawbacks. The membership of the CR users to the clusters is based on the assumption that each user already knows which cluster to join. As the clusters are formed based on group-communication needs, this is infeasible without exchanging detailed cluster information. Moreover, as the CR delegate does not coordinate with the other clusters for efficient spectrum sensing, as each cluster operates independently without enforcing silent periods. Moreover, there is no consideration of protection to the PUs either by adapting transmission, power control, among others.

8.4.3.2 **Multichannel MAC Protocol for CR networks (MMAC-CR** The MMAC-CR protocol incorporates a power saving mode, similar to the one present in the IEEE 802.11 distributed coordination function (DCF) [250]. The reason we include this protocol under the hybrid category is that it relies on fixed-length time intervals, and each such interval begins with a reserved time frame called at ad hoc traffic indication message (ATIM) window. The ATIM window is followed by the DATA window for the remainder of the current interval. Similar use of the ATIM window has been described earlier in [188, 207, 281].

The CR users use an out-of-band CCC for the exchange of spectrum information through the scan result packet (SRP), and periodic beacons. They learn the network-wide spectral opportunities by listening to the information carried in $\mathcal{C}$ minislots of the SRP, where $\mathcal{C}$ is the total number of PU channels. Thus, each minislot is unique mapped to a channel, and the presence of a signal, or its absence in that particular minislot, indicates the availability for the corresponding channel. As the data is not transmitted during the ATIM window, a node can randomly select any licensed channel for spectrum sensing in this duration. Thus, though this window lowers the overall throughput, it enables the creation of a *silence* zone for effective distributed sensing. A CR user pair may also negotiate the transmission channel in the ATIM window, through a unicast ATIM message and an ATIM-ACK. These users now need to remain active for the subsequent DATA interval, while the others can go to a low-power state.

The authors use two different sensing types. The fast sensing (e.g., energy detection) is used within the ATIM window for all the channels. In the subsequent DATA window (i.e., after the ATIM window but before the beacon period is completed), a node may either transmit packets or undertake fine spectrum sensing (e.g., cyclostationary detection). The work in [250] further describes the conditions under which one of the above sensing methods are adopted, and an analytical expression for throughput for a single-hop CR network is derived.

In this work, the authors assume that a CR user is able to continuously monitor the control channel as well perform data transmission in the license channel. Thus, two radios are required. Moreover, the ATIM window for several licensed channels may remain unused as nodes may not choose these channels for sensing, thus making sub-optimal use of the available spectrum resource.

8.4.3.3 **Partially Observable Markov Decision Process (POMDP) based MAC** A partially slotted single-radio MAC protocol based on the theory of partially observable markov decision process (POMDP) is proposed in [295]. A similar approach is also used in the cognitive radio access scheme in [96], where limited sensing capabilities of the cognitive radio imply that only one channel can be sensed at a time. In this case the system is also classified as partially observable and the analysis becomes involved.

The approach adopted in [295] integrates the design of spectrum access protocols at the MAC layer with spectrum sensing at the physical layer and trafÞc statistics determined by the application layer. The two main issues addressed are: (i) joint consideration of the spectrum sensing and spectrum access issues, and (ii) transmitter-receiver synchronization, i.e. ensuring that both the transmitter and receiver hop in the spectrum together without additional control overhead. The time is divided into slots, and in each slot the spectrum access follows a sensing-RTS-CTS-DATA-ACK schedule.

The POMDP is a generalization of a markov decision process and is addressed as partial because the network state cannot be fully observed due to partial spectrum sensing or due to sensing error. Here, time is divided into slots, and at the start of each slot, the protocol decides a set of spectrum bands for sensing, and another set of bands for transmission. These decisions are made with the aim of maximizing the throughput of the CR user while limiting the interference to the PUs and exploiting the past history of the spectrum band. During transmission, classical CSMA is assumed.

This work has the following novel approaches:

- *Performance Metric:* As the protocol provides a decision on which spectrum bands to transmit based on the sensing, it proposes a new metric to measure the *reward* for the action. It is defined as the number of bits delivered when the user senses the spectrum bands during the sensing interval, and transmits in those bands that are deemed to be free of PU activity. The framework also integrates the cases where the sensing is error-prone and thus is a realistic representation. The above performance metric depends upon the network state and the reward is continuously added over time, and compared with the maximum value that could be accrued for perfect decisions.

- *Learning Support:* Unlike several previous works, the *cognitive* feature is fully integrated in the working of this protocol. The proposed MAC protocol in [295] accumulates the spectrum band history and learns which of these bands are best suited for long term use. No prior statistical traffic information for the PUs is known, and the probabilistic spectrum selection process converges to a value bounded within a constant error of the optimal solution, when observed over a sufficiently long time.

- *Synchronized Spectrum Switch:* For a given transceiver pair, the probability of choosing the spectrum for transmission is the same, as it is assumed that the spectrum environment seen over either ends of the link is similar. Thus, without an additional CCC, both the sender and receiver synchronously change the spectrum band, which is an important issue in CR networks.

The theoretical basis for the proposed MAC protocol assumes that the spectrum usage statistics remain unchanged for several time slots. As a result of this, the PU activity *pattern* is learnt over time and the protocol is strongly affected with frequent and random spectrum changes. Moreover, the optimal result is reached after very large time durations, and the protocol does not perform well in the initial stage.

8.4.3.4 **Synchronized MAC (SYN-MAC)** The SYN-MAC protocol proposed in [146] does not need a CCC but has a dedicated radio for listening on the channel for control messages. A second transceiver is used for data traffic.

The main idea of the protocol is the following: Time is divided into time slots and each slot represents a particular data channel. The control signal exchange occurs in the channels represented by the slots while the data transfer can occur in any channel that is found suitable between a given node pair. Thus, the control signaling is similar to slow

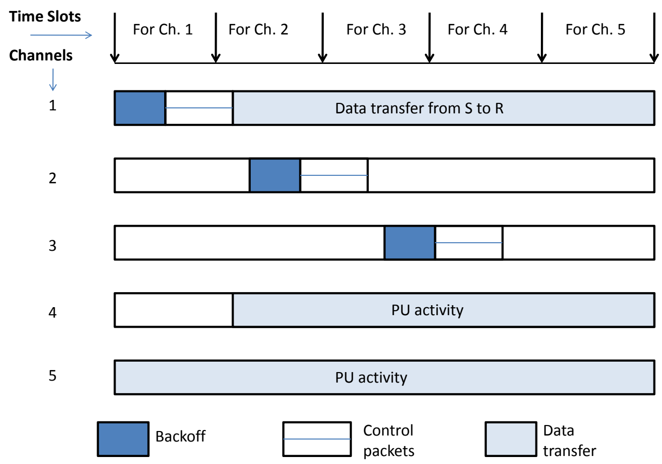


Figure 8.9 Illustration of control and data packet exchange in SYN-MAC [146]

frequency hopping, in which the channel is switched periodically. At the beginning of each time slot, the CR users tune their dedicated control radios to the channel specified by it, and the users that wish to initiate a data transfer send out a beacon at this time. Interested neighbors respond with their their own list of available channels, and further communication is carried out in one of those selected channels.

The protocol is explained in detail through the example shown in Figure 8.9. There are five time slots, each representing one of the five channels. Consider two CR users S and R that wish to communicate, and have the free channel sets $\{1, 2, 5\}$ and $\{1, 3, 5\}$, respectively. Node S chooses channel 1 and waits for the beginning of the related time slot to tune the listening radio to that channel. After a backoff period, S contends for the channel using as access scheme similar to the IEEE 802.11 distributed coordination function (DCF). If the contention is successful, it starts the data transfer.

Consider another example, in which four nodes A, B, C, and D form a linear topology. Their respective available channels are given by the sets $\{2, 4\}$, $\{2, 4\}$, $\{1, 3, 4\}$, and $\{1, 3, 4\}$, respectively. If node B detects PU activity on channel 4, it has to send a control packet to its neighbors informing of a change in its channel set. User B waits for the time slot dedicated to channel 2, the channel common with the neighbor (A), and sends this information after a backoff. A similar behavior is shown by node C, which has channel 3 in common with neighbor D. It waits for the related time slot to communicate its new channel set after detecting PU activity on channel 4.

The above protocol has the advantage of not using a dedicated CCC, and the dedicated listening also addresses the multichannel hidden terminal problem. However, this approach does not guarantee protection to the PUs, as their arrivals are notified only in specific time slots to the neighbors. In addition, the channel may not be utilized efficiently, as it can be used only once in a given cycle.

8.4.3.5 **Opportunistic MAC** The opportunistic cognitive MAC protocol proposed in [242] uses two transceivers, one for a dedicated CCC, and the other that can be dynamically tuned to any chosen spectrum. As shown in Figure 8.10, the time is slotted for the data transfer over the licensed channels, while the CCC operation is partly slotted, followed by a random access negotiation phase. Thus, it is a hybrid protocol.

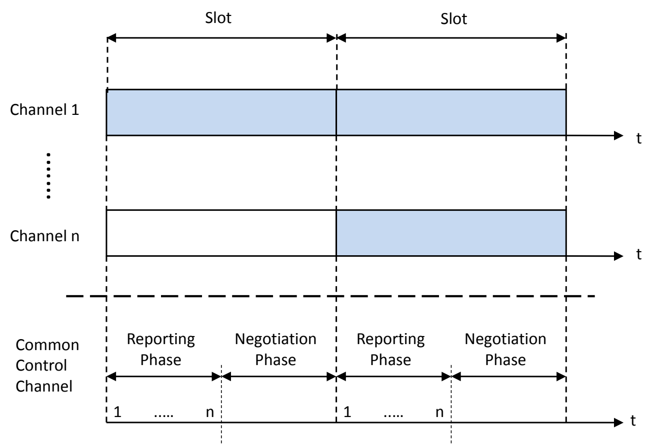


Figure 8.10 Working principle of opportunistic MAC [242]

The detailed explanation of the working of the MAC protocol is described below with reference to Figure 8.10. The CCC has the following two phases:

- *Reporting Phase:* The reporting phase is further divided into n mini-slots, where n is the number of channels. At the beginning of each time slot, the cognitive user senses one of the channels. If the i_{th} channel is perceived to be idle, it sends a beacon over the CCC during the i_{th} mini-slot of the reporting phase. No beacons are sent if no PU is detected. These beacons serve to inform the neighbors of the PU activity.
- *Negotiation Phase:* During the negotiation phase, the CR users negotiate via contention-based algorithms, such as those based on the IEEE 802.11 and p-persistent carrier sense multiple access.

To ensure that all the channels are sensed, each CR user independently chooses a channel with equal probability. If sufficient number of CR users are present, then all the channels can be covered with high probability. Moreover, the authors provide a detailed analytical treatment of the average number of channels available to the CR users, and the upper bound on their throughput.

Apart from the overhead of maintaining the time synchronization and the need of multiple transceivers, this work does not specify the exact link layer interactions between the nodes. As an example, multiple transmissions may be possible at the same time between different node-pairs, that may affect the sensing results. As the channel for sensing is randomly chosen, the neighboring nodes do not have a priori knowledge of this event and do not silence their own transmissions to improve the sensing accuracy.

8.5 CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, we present an overview of the state of the art for medium access protocols in cognitive radio networks. The existing works in the two main functions of the MAC protocol, namely the spectrum sensing and spectrum access were discussed. With respect to spectrum sensing, we believe that there is further work needed in devising accurate models that account for false alarm and missed detection probabilities in one framework.

For this, the simplified ON/OFF PU traffic model may not be suitable in a practical environment where the licensed users may be cellular, contention based, or have other possible access technologies. Regard the existing CR MAC solutions, several open issues remain that must be addressed. Firstly, the information from multiple layers must be seamlessly integrated in the working of the MAC protocol. As an example, the results of channel sensing and interference detection obtained from PHY layer can be used by MAC layer to build the channel occupancy history over time. Most of the existing works do not completely integrate the sensing function. Hence, the sensing accuracy may be affected due to concurrent packet transmissions. There is also significant scope for devising protocols that adapt the CR transmissions based on the type of the interferer. As an example, the CR users may store packets to be transmitted during the off durations of duty cycled PUs. Newer performance metrics that capture the CR specific improvements should be devised and used for evaluating the different MAC protocols. Thus, we believe that MAC protocol design for cognitive radio is an open area of research and will be of interest to both the industry and the academia as this technology matures in the next few years.

CHAPTER 9

ROUTING PROTOCOLS FOR COGNITIVE RADIO NETWORKS

9.1 INTRODUCTION

The Network Layer is responsible for the task of path selection between the source and destination. In the general case, such a path is formed by nodes forwarding packets generated by the source node towards the destination node. In a CR network, the temporal and spatial variation in the spectrum availability poses a non-trivial task in determining the best selection of packet forwarding nodes. The challenge in such a dynamic spectrum environment lies in ensuring that the process of packet forwarding is not periodically interrupted, and is undertaken within permissible end-to-end delay bounds. While both centralized and distributed architectures require novel routing solutions, we begin our discussion by drawing from a general example of multihop CR ad hoc networks. The issues specific to centralized routing can then be easily understood once the main challenges in this general multihop scenario are identified and visualized. We present an example drawn from commonly used technology, namely Wireless LANs (WLANs) and ZigBee.

Consider a PU that is transmitting in the licensed spectrum band. Depending upon the channel definition for the PU, the spectral shape of the transmitted signal can cover a large range of frequencies. Figure 9.1 displays the shape of the received power spectrum observed in a typical wireless LAN (WLAN) transmission using the IEEE 802.11b standard in the $2.4\,\mathrm{GHz}$ ISM band. Observe that the spectrum may span several channels defined by a different standard, e.g., channels $17-21$ specified by the IEEE $802.15.4$ standard. In this example, devices that operate in accordance with the $802.15.4$ standard, such as ZigBee nodes, may be unable to use these channels at the same time as the WLAN. The

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury

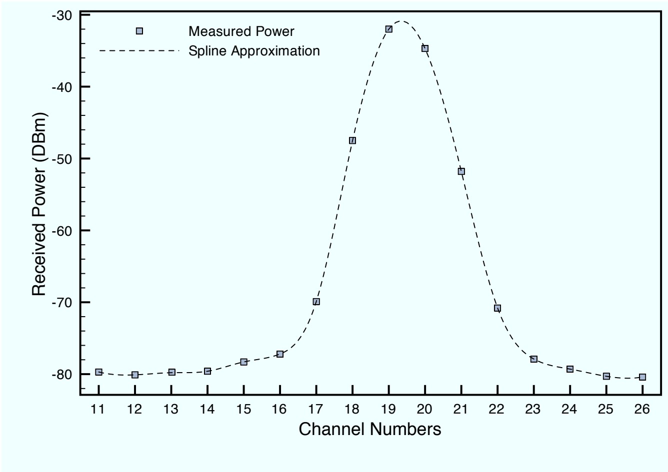


Figure 9.1 The spectrum shape of a typical transmitter.

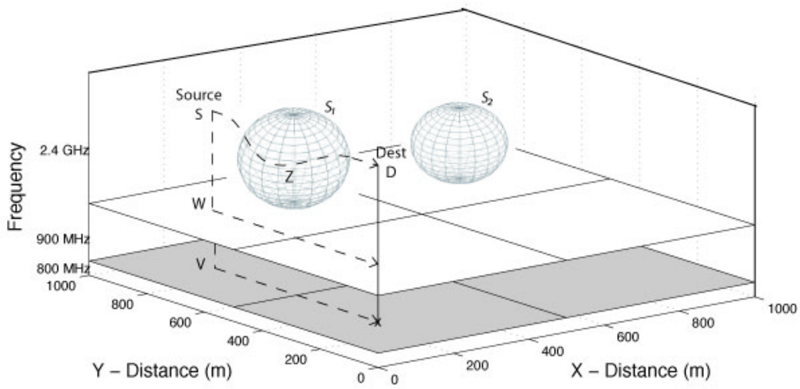


Figure 9.2 The routing problem in CR networks.

transmission power follows a bell-shaped distribution curve around the center-frequency, requiring paths formed by ZigBee devices to either avoid regions of active WLAN activity altogether, or to choose a frequency as far as possible from the center frequency of the WLAN.

Figure 9.2 shows how the wireless spectrum may be occupied at different locations, and to varying extents on the frequency scale. Assume two PUs are placed at the center of the spherical regions, S_1 and S_2. These spheres represent the following important features: (i) the range of the frequencies occupied by the PU (along the frequency or z-axis), and (ii) the geographical extent of the region (along the cartesian or $x-y$ plane) which is rendered unusable for the CR user owing to the PU transmission. From Figure 9.1, we observe that power levels of the transmitted signal vary on the frequency scale (recall the bell-shaped curve of the WLAN spectrum), and hence, the coverage regions for each of these individual frequencies, approximated as circles for simplicity, are also of unequal radii.

Consider a CR user called the *source* at location S (note: S, W, and V have the same coordinates in the $x - y$ plane) wishing to find a route to another CR user *destination* at location D. There are the following three options in the path selection:

- A path that uses completely different channels from the PU, i.e. above or below the spherical region S_1 and shown by the path $S - W - D$. This incurs a higher channel switching delay, but gives the shortest path along the straight connecting line between *source* and *destination*.

- A path that uses the same channel as the PU, but avoids the coverage regions by circumventing these areas, i.e. around the spherical region S_1 and maintaining the same frequency plane. This is shown by path $S - Z - D$.

- A path that switches the spectrum band altogether, and shown by $S - V - D$. The difference between this case and the earlier path of $S - W - D$ is that the spectrum band switching results in a drastic change in the transmission frequency. A lower spectrum band allows for higher propagation distances for the transmitted signal, and hence, require fewer hops. This effect is almost negligible during channel switching within the same spectrum band.

For single hop ad hoc networks and infrastructure-based networks with a BS directly communicating to a CR user, the *destination* can provide direct feedback to the *source*, and thus, the spectrum decisions can be made at the link layer itself. Instead, and in the remainder of this chapter, we focus on multihop networks, as the above spectrum and next hop selection decisions are non-trivial, and even the simplest case is NP complete. Thus, routing is an active area of research, and still in a nascent stage compared to the more evolved techniques for spectrum management, i.e., spectrum sensing, sharing, and decision covered in the previous chapters. In this chapter, we first highlight the main design approaches for CR routing, and discuss existing schemes for different distributed network architectures.

9.2 FACTORS INFLUENCING ROUTING PROTOCOL DESIGN

Though the route formation is undertaken at the Network Layer, several factors that influence the selection of the path rely on information availability from the lower layers of the protocol stack concerned with the spectrum management functions. Thus, necessarily in CR ad hoc networks, and unlike their classical wireless counterpart, routing decisions involve cross-layer information. We identify the following unique metrics that must be considered in CR protocol design:

- *Effect of spectrum band switching on path delay:* Since a cognitive radio user can detect a number of spectrum bands for communication, there exist several choices for the selection of spectrum band along a path. Spectrum band switching at a node results in non-zero delay. The time required for band switching is significant when the two frequency bands are far away on the radio spectrum. In Figure 9.2, the switching time for the path $S - W - D$ (different channels within the same spectrum band) is lesser than that for the path $S - V - D$ (different spectrum band altogether). This delay has to be taken into account in the routing algorithm, and efforts must be made to minimize frequent switching once the route is operational. Statistical information about the long term availability of the spectrum is often leveraged in this design.

- *Relationship between spectrum choice and path delay:* The choice of frequency decides the rate of exponential decay of the wireless signal, with lower spectrum bands allowing greater propagation for the same transmission power. Thus, to increase the per-hop progress towards the destination and incur fewer hops in total, the routing algorithm must be linked closely with the spectrum decision process.

- *Route quality determination:* While classical metrics of end-to-end throughput, delay, energy, and fairness may be used in CR networks to determine the route quality, other unique factors, such as, PU activity on the channel, the least (or bottleneck) bandwidth of the spectrum on the links forming the path, must also be incorporated. Thus, new route quality determination metrics must be identified that are more suited to capturing the dynamic nature of the CR networks.

- *Need for route maintenance:* The sudden appearance of a PU in a given location may render a given channel unusable in a given area, thus resulting in unpredictable route failures, which may require frequent path re-routing either in terms of node selection or the chosen channels at the nodes. In this scenario, effective signaling procedures are required to restore the disconnected paths with minimal effect to the end user.

- *Coupled path and spectrum selection:* This kind of design is a cross-layer approach that jointly considers route and spectrum selection. In particular, each source node makes decision on both route and channel selection - the decision includes not only the selected route, but also the channel to be used by each link on the route, and possibly, the time schedule of the channel usage.

9.3 OVERVIEW OF THE CLASSIFICATION OF ROUTING PROTOCOLS

After the routing algorithm is designed, considering the factors outlined in Section 9.2, the CR users rely on the spectrum information to make their routing decisions. Depending on the network architecture assumptions, this awareness of the spectrum environment varies to different extents. In Figure 9.3 we give a classification of the different routing approaches, in which the *centralized* and *distributed* protocols are placed at the top of the hierarchy. In the former case, all the nodes are aware of the exact spectrum availability throughout the area under study. Recently, the FCC has encouraged the formation of spectrum maps that indicate time and space the channel availabilities [76] in the spectrum below 900MHz and around 3 GHz. Industry efforts led by Google, Motorola, and Dell have resulted in the formation of the Whitespace Database group, which promises to expedite the spectrum map information access to the general user. In this chapter we follow the broadly the classification offered by [39].

For *centralized* protocols, the complete network knowledge results in knowing the instantaneous spectrum availability between any two pairs of nodes. Thus, by modeling the network as a connected graph, the edges can be weighted to reflect the spectrum information for the node pair. This greatly aids in the use of various graph theoretic tools to determine the shortest path between nodes, often with deterministic guarantees of end-to-end performance. Moreover, global optimization approaches that can sift through a large set

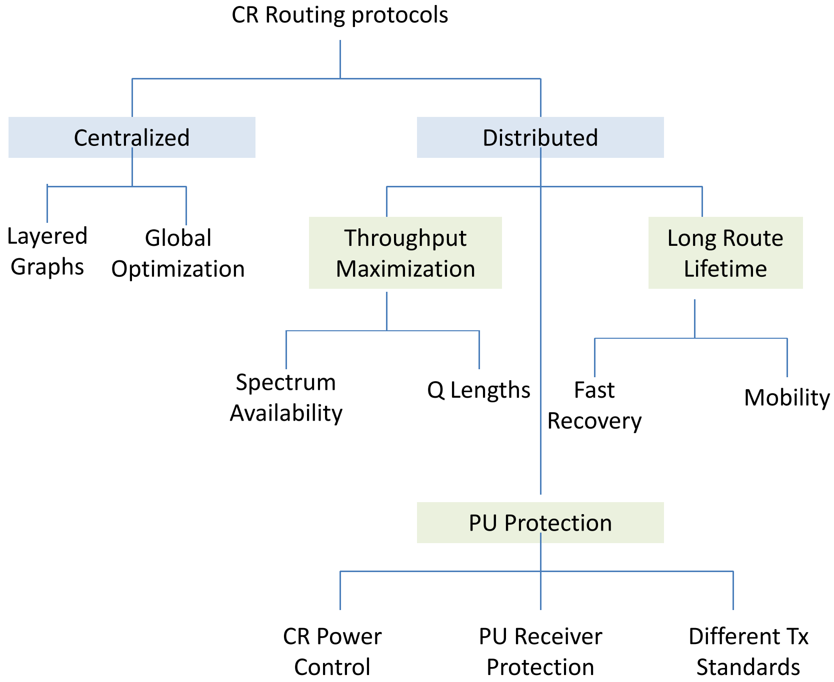


Figure 9.3 Classification of different CR routing algorithms.

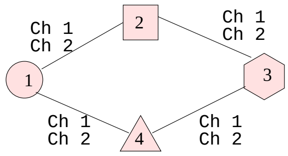
of constraint equations for optimizing the route can also be developed. While completely centralized protocols have limited relevance in practical deployment scenarios, they are included in our study as they serve as an upper bound for the purpose of comparison.

When nodes rely on only the local information for route formation, the resulting *distributed* protocols are generally of a heuristic nature, and performance bounds for their operation are difficult to derive. For such protocols, the neighborhood information between the nodes must be exchanged through a common control channel (CCC), and different approaches target one or more of the routing design principles from Section 9.2. As shown in Figure 9.3, we can broadly classify the *distributed* routing protocols under protocols that aim for: (i) PU interference protection, (ii) CR end-to-end delay minimization, (iii) CR end-to-end throughput maximization, and (iv) Identifying routes with long lifetime.

We discuss these different classes of routing protocols in detail in the subsequent sections.

9.4 CENTRALIZED PROTOCOLS

This class of protocols rely on a central decision making entity, which generally has a global knowledge of the network environment. The routing strategy is hence, near optimal if the network conditions are static. The overhead involved in disseminating local knowledge to the central entity (if it exists) is often prohibitive for large networks.



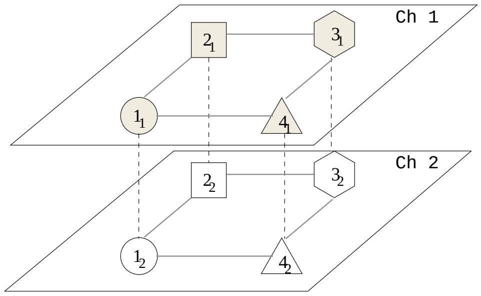


Figure 9.4 An example four node network and its layered graph model.

9.4.1 Graph-theory based protocols

Traditionally, a network is modeled as a directed graph $G = (N, L)$, where N represents the set of all nodes or *vertices* of the graph, and L is the set of data links or *edges* connecting the nodes. Moreover, in a CR network, each node can individually detect a number of spectrum opportunities (channels) for communication, and thus the edge L needs to be qualified further to represent a specific channel. Let L_{ij}^k denote the channel k used on the link (i, j). It exists if and only if $j \in S_i$, $k \in SC_{ij}$, where S_i is the set of all nodes that can be reached by node i, $SC_{ij}$is the set of all possible spectrum/channel options over link (i, j). Each data channel in each link is unidirectional. Each node can communicate with a subset of other nodes in the network via these channels. There are two stages while using graph-theory based tools: First, the actual network environment has to be mapped to the theoretical graph representation, and second, a connected path is constructed using the graph that optimizes the edge weights between the source and destination.

9.4.1.1 Routing through layered-graphs In the layered graph, each layer represents one of the available channels [26][27]. If node A has n channels for use, it has corresponding n subnodes $A_1, A_2, \ldots, A_n$, with each subnode residing in a given layer. A subnode is active when an interface is assigned to this channel, otherwise it is inactive. The model defines four types of edges: *access* edge, *horizontal* edge, *vertical* edge and *internal* edge. These edges have the following functionalities: The *access* edges connect a node to its subnodes. The *horizontal* edges within a given layer denote the links can be established on the corresponding channel between nodes. Thus, at each layer i, if there is a channel available between two potential neighboring nodes, 1 and 2, then let the (horizontal) edge $(1_i, 2_i) \in L$. Figure 9.4 shows a topology with four nodes and two channels, ch_1 and ch_2, with the corresponding layered graph representation. In the layered graph, the subscript notation for a node, say node 2_i indicates that the subnode corresponds to channel i.

After the layered graph G' is constructed, finding a routing path between two nodes translates to finding the shortest path in G'. Though this approach is application for both centralized and distributed scenarios, the need for dynamically updating the graph when the information changes for all the nodes of the network makes it suitable for the former case. In the works of [26][27], the PU dynamics are assumed to be low enough such that the channel assignment and the routing among the CR users can be statically designed. The authors further focus on the case where cognitive devices are equipped with a single half-duplex cognitive radio transceiver, which can be tuned to M available spectrum bands or channels.

9.4.2 Global optimization based protocols

Mixed integer non-linear programming (MINLP) is a powerful mathematical technique that is often used to identify the optimal solution among several complex system governing constrains. In the general representation, MINLPs are composed of an objective function, say $f(x, y)$ that needs to be minimized, subject to m constraints given by $g^j(x, y) \geq 0$, where $j = 1, \ldots, m$. Moreover, x and y must belong to the set of real numbers $\mathbb{R}$ and y must be a member of the set of integers $\mathbb{I}$. Hence, formally,

$$\min\{f(x, y) : g^j(x, y) \geq 0\},\ j = [1, m],\ x \in \mathbb{R}^{n_1},\ y \in \mathbb{I}^{n_2}, \tag{9.1}$$

where n_1 and n_2 are the number of real continuous and integer constrained variables, respectively.

The MINLP technique is leveraged in [276][277] where the authors specify the constraints ($g^j(x, y)$, $m = 3$) of (i) *link capacity* being greater than the aggregate traffic flow requirements, (ii) limiting *interference* caused by selection of the same channel by neighboring nodes, and (iii) *flow balancing in the route* for each of the intermediate nodes from source to destination. The overall aim of the optimization is to minimize ($f(x, y)$) the overall bandwidth consumption of the network by considering the link effects of all the nodes forming the route.

To ensure a tractable solution, the authors solve the above problem by relaxing the constraints that allow users to transmit over a give channel to linear values, thereby transforming the problem to a linear programming (LP) case. Several techniques (such as the classical Simplex method) may be employed to solve this LP problem. Moreover, the authors provide other heuristic techniques when such a solver is unavailable or infeasible owing to computation reasons.

While this method has the benefit of obtaining the optimal solution for the joint scheduling and routing problem, the complete knowledge of the network topology is a strong requirement. Moreover, interference at any given point is cumulative, and the effect from multiple different sources should be considered. The authors simplify this calculation using an *interference range* which is a binary result, instead of a gradual summation of individual contributions of the received power by neighboring nodes.

9.5 DISTRIBUTED PROTOCOLS

Distributed protocols generally have function with limited spectrum knowledge of the environment, and need frequent control message exchange with nodes to maintain updated information about the spectrum state. Such protocols typically find application in CR ad hoc networks with multiple hops between the source and destination.

9.5.1 PU interference protection

In this section, we review work that is specifically focused on the key consideration of safeguarding the transmission of the PUs, over intra-CR network performance improvement. This is a challenging task, given the dynamically changing network environment that the

CR users must adapt to, but important for the feasibility of the opportunistic transmission in the licensed bands.

9.5.1.1 Routing protocols with power control

CR users periodically sense the spectrum and decide on the spectrum availability. Typically, the CR users that are located in regions with fewer cases of positive PU transmission detections may be preferred for routing. This can be achieved by ensuring that the total CR transmission power obtained by summation of their individual signal powers is contained within a threshold. Thus, routes that incorporate interference awareness through power control help to prevent any adverse effect on the PU performance.

The interference and connectivity tradeoff is evaluated in [275] for single-hop and multi-hop CR transmissions, when the locations of the PUs and the CR users is known. This proceeds by outlining two conditions:

- *Transparency:* The CR user transmissions does not lower the QoS below the threshold for the PUs
- *Reliability:* The CR user at the destination is able so successfully decode the data.

The authors propose a geometric analysis under Rayleigh fading conditions that helps decide when should a direct transmission be made to the destination, and when it should be broken into multi-hop relaying. Two routing methods, called as nearest-neighbor routing (NNR) and farthest-neighbor routing (FNR) respectively, are proposed in this work. In the NNR method, the source node i attempts to find the nearest neighbor inside the circular region of radius D^i_{max}, which the maximum distance derived under the *transparency* and *reliability* constraints. Similarly, FNR finds the relay at the farthest distance from the source, but within the limit of D^i_{max}. Results show that FNR is better for the end-to-end channel utilization and reliability, while NNR gives better energy efficiency.

9.5.1.2 Routing with multiple different transmission standards

The minimum weight routing protocol (MWRP) looks at architectures where each CR user may be equipped with transceivers for different wireless technologies, such as cellular (TDMA / FDMA / CSMA) and also 802.11 b/g cards [25]. Each transmission technology, also called as wireless system (WS), has a different communication range that depends primarily upon the transmission power (i.e., allowed by the standard) for that particular WS. A routing weight is assigned to each WS depending upon the ability to reach a point closest to the destination. As an example, the routing weight for the WS given by cellular TDMA is much higher than the corresponding weight for the 802.11 b transmission, meaning that the node should chose the formed preferentially while selecting the next hop. The proposed routing protocol locally finds the path to minimize the routing weight between a source and a destination. The route discovery procedure is very similar to link state routing algorithms but with the link weights represented by the above weights specific to each WS. To ensure neighbor discovery and to understand which radio interfaces of the nodes are within range, a CCC is used. The node advertises its reachability information with the maximum range on the CCC to ensure that the knowledge of its connectivity is adequately disseminated in its neighborhood.

Though the choice of the communication technology could depend upon the allowed coverage range for the CR user (if there is a presence of a PU in its vicinity), the algorithm neither performs channel selection nor considers the intra-CRAHN interference caused by the transceiver selection.

9.5.1.3 Routing with explicit PU receiver protection: The approach of interference control assured protection to the PU transmitters that are within the range of the CR devices, as only these signals are detected by the CR user. For certain PU applications such as television broadcast, the transmission is uni-directional, and the PU transmitters do not suffer from CR network interference. Rather, transmission by neighboring CR users may affect the PU receivers that cannot be detected easily (no transmission, low leakage power from the reception circuitry). The CR routing protocol must provide protection to these PU receivers by avoiding entire regions where such devices may possibly be present. The cognitive routing protocol (CRP) has been proposed to address this concern [51]. A unique contribution of this work deals with the formulation of routing classes: *Class I* assigns higher significance to end-to-end latency while meeting minimum PU interference avoidance. As opposed to this, *class II* routes prioritize the PU protection at a higher level by allowing a permissible performance degradation to the CR operation.

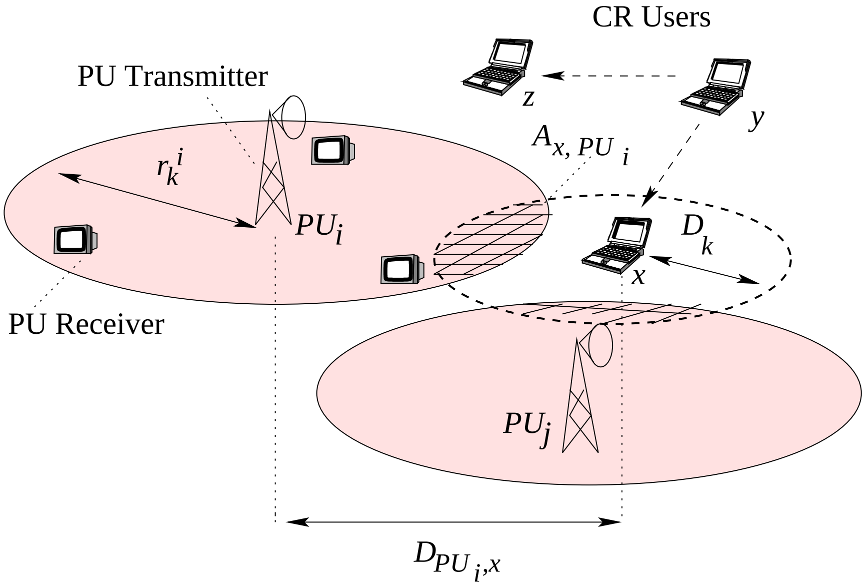


Figure 9.5 Route establishment in CRP.

The route-setup in the CRP protocol is composed of two stages - (i) the *spectrum selection* stage, and the (ii) *next hop selection* stage. The source node broadcasts the RREQ over the control channel, and this packet is propagated to the destination. Each intermediate forwarder identifies the best possible spectrum band, and the preferred channels within that band during *spectrum selection*. To enable this, we have proposed several unique CR metrics that are weighted appropriately in an optimization framework for choosing the spectrum. Moreover the function is cast differently for each Class of CR route. As an example, for the *class I* route, the CR network end-to-end latency is the key consideration. Here, the spectrum chosen by a given candidate forwarding node must (i) support the highest propagation distance, with the (ii) longest allowed duration for transmission given the sensing schedules of the neighboring nodes. Consequently, the optimization function for *class I* route attempts to maximize these two factors during the *spectrum selection* stage.

The next stage is the *next hop selection* stage, where the candidate CR users rank themselves depending on the choice of the spectrum and the local network and physical environmental conditions. These ranks determine which CR users take the initiative in the subsequent route formation. As an example, in Figure 9.5, the PU transmitters i and j are separated by distance $D_{PU\,\mathrm{i}}x$. The shaded circles indicate their coverage ranges in which PU receivers may be present, though their locations are unknown to the CR users *x*, *y*, and *z*. Note that user *x* has greater overlap of its own transmission radius, given by D_k with the coverage regions of the PU transmitters, which implies higher possibility of interference with the PU receivers. Consequently, it also has a lower *initiative* than z for forwarding packets. Assume that the RREQ is broadcast by CR user *y*, and received by both users *x* and *z*. As CR user z has higher *initiative* as compard to *x*, it also has a lower forwarding delay. Hence, it transmits the RREQ earlier than user *x*. The arrival times of the RREQs at the destination (over several, possibly disjoint paths) is dependent on these forwarding delays. Hence, the earlier arriving RREQs also represent paths that pass through regions preferred for CR operation. This method reduces the overhead of forming routes in all possible channels over several different spectrum bands. It also tries to map the spectrum characteristics at the intermediate hops to the RREQ arrival times, thereby reducing both the need for transferring large volumes of node information over the RREQ packet, and the resulting computational complexity at the destination.

In this work, we assume the network architecture is composed of stationary PU transmitters with known locations and maximum coverage ranges, as seen in the case for television broadcast towers. In Figure 9.5, this implies the locations of PU i and j are known, and the range r_k^i to be fixed. The CR users are mobile, location-aware, and have no knowledge of the PU receivers. Additionally, the statistical knowledge of the channel availability is assumed for the different spectrum bands.

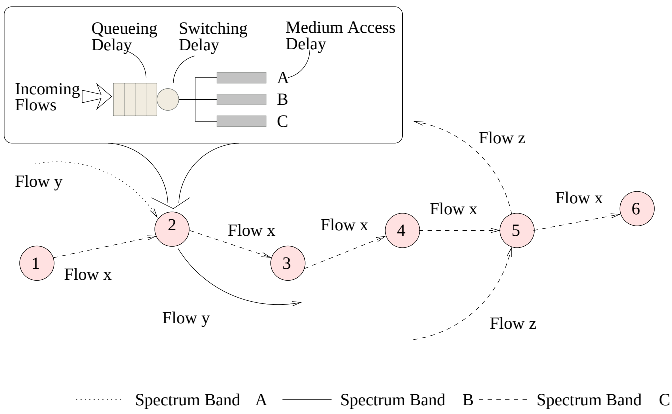


Figure 9.6 Switching, MAC layer channel access, and queueing delays.

9.5.2 CR end-to-end delay minimization and throughput maximization

The path delay in a CR ad hoc network is a summation of the individual link delays caused by the variance of the spectrum availability with time. When the PU activity is high, the nodes need to frequently switch the transmission channels, leading to *switching delay*. Different frequency bands may experience different levels of congestion. Hence the links of a given path may experience different MAC layer *backoff* durations during channel contention. The effect of this *MAC delay* when considering all the links of the route cumulatively impacts the end-to-end performance. Similarly, if a CR user has multiple radios, each radio servicing a given channel, then the backlog of the packets on each of these network interfaces results in a *queueing delay*.

Figure 9.6 shows how these three delay components affect the operation of a CR node when there are three different spectrums, A, B, and C served by three different radio interfaces. These radios are fed by a common queue that contains the incoming packets for all the flows. For each packet, one of these three radios must be chosen. Consider the operation of the two nodes 2 and 5. For node 2, there are two flows that pass through it. Flow x uses the same spectrum C throughout the path while flow y switches spectrum A to B. At node 5, the two intersecting flows x and z use the same spectrum C. Thus, *the switching delay* is the dominant effect at node 2, while the MAC access delay is dominant at node 5. All the nodes $1, \dots, 6$ along the path have a queuing delay that is a function of the MAC layer operation.

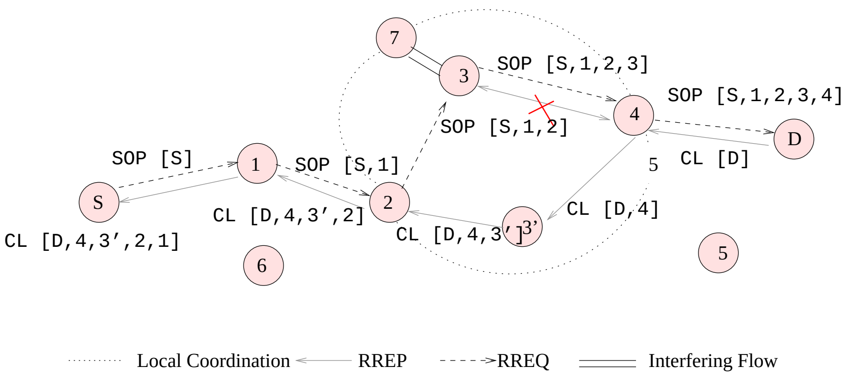


Figure 9.7 Route establishment in [279]

9.5.2.1 Throughput maximization based on spectrum availability

Need to insert details about [237] and reference to Fig. 9.8.

The Spectrum Aware Mesh Routing (SAMER) accounts for long-term and short-term spectrum availability [200]. SAMER captures both the above two types of availability by defining a path spectrum availability (PSA) metric by collecting the spectrum information (bandwidth, loss rate of the link) from the neighborhood periodically. The packets are delivered opportunistically along the path with the highest PSA value, thereby probabilistically choosing regions with the best spectrum availability.

The PSA is expressed as the throughput between a pair of nodes (i, j) across a spectrum block b as:

$$Thr_{(i,j),b} = T_{f,b} \cdot B_{w,b} \cdot (1 - p_{loss,b}), \tag{9.2}$$

where $B_{w,b}$ is the bandwidth and $p_{loss,b}$ the loss probability of the spectrum b, which is a function of the loss rate of broadcast packets between pairs of neighboring nodes. $T_{f,b}$ is the minimum time that a node pair (i, j) can communicate over the spectrum b. The aggregate throughput $Thr(i, j)$ between these nodes. In the SAMER protocol, a weighted average of the throughput is undertaken to capture both the current view and the spectrum availability trend over time.

SAMER is a specialized protocol for mesh networks, and has an associated overhead to maintain the mesh topology. Moreover, the network cost for obtaining the loss probability

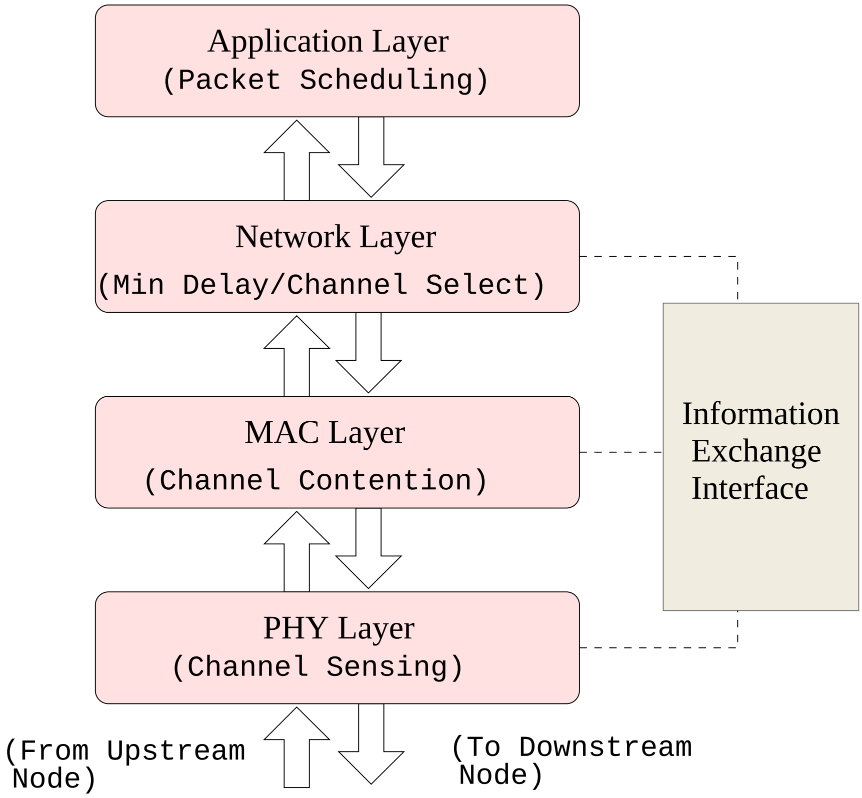


Figure 9.8 Cross-layer architecture in [237].

for all the neighbors, the accurate bandwidth estimation, and the effect of lost packets containing this information on the protocol performance have not been evaluated.

A similar approach is presented in [108], wherein the link weights are first calculated probabilistically based on interference from the PUs, the received signal strength, and the PU occupancy rate on all the channels at the given link. The CR users calculate an expected delay from themselves to the possible destinations, and run a classical distance-vector algorithm, such as Bellman Ford or Dijkstra, to decide on the optimal path at each hop.

9.5.2.2 Throughput maximization based on queue lengths Several previous works [48] [279] have incorporated queuing delay in the cost function C_i (cost to reach the *destination* from node i) represented by,

$$C_i = \sum_{n=i}^{n=destination} \left\{ D^i_{switching} + D^i_{queueing} + D^i_{backoff} \right\} + DP_i, \quad (9.3)$$

where DP_i is the increase in path delay considering the nodes of the chosen route. This delay is added to the existing path cost given by the summation in the first term of the equation 9.3. The terms $D^i_{switching}$, $D^i_{queueing}$, and $D^i_{backoff}$ in this summation represent the measured delays associated with channel switching, packet queueing backlogs, and MAC layer backoff during channel contention. The calculation of the term DP_i relies on the following assumptions that limit the validity of the framework: The authors assume that the switching delay is a function of the separation of the channels on the frequency scale, whose impact is negligible using the digital switching that are typically used at the radio front-end. The event of changing the transmission frequency is simply, altering the frequency of the incoming carrier signal used at the *mixer* stage of the front-end, and thus, independent of the magnitude of the change. Moreover, a priori knowledge of the probability of link layer packet collision is expected to be known, which cannot be trivially derived. Finally, the proposed work does not map well to the CR domain with PU consideration as there is really no integrated protection to the PUs or awareness of of the interference caused to them. It should be noted that the approach in [48] [279] is itself an extension of previous works [28] [48] that do not consider the queueing delay, but the rest of the framework remains the same.

The underlying routing protocol in the above approaches is AODV. Similar to classical AODV, the RREQ control message is propagated towards the destination, with an additional cost function C_i inserted at each node i. A unique contribution of the above works is that the RREP is not always forwarded along the return path that was used for the RREQ traversal. This phenomenon is described in Figure 9.7, where node 3 does not return the RREP to the downstream node directly. Instead, it observes that another neighboring node $3'$ is better suited for the path, and informs the upstream node 4. The latter now re-transmits the RREP to $3'$ which is then routed back to the source node.

The ROSA protocol is a distributed cross-layer control scheme that allows secondary users to jointly control the routing, spectrum, and power allocation functionalities to maximize the global network throughput [68]. This work addresses a variety of different issues, including consideration of power control and interference to PUs. However, we include this work in this section given the main aim of throughput maximization of the protocol.

First, the CR nodes ensure that the interference generated owing to the transmission of a given node can be tolerated by the neighbors with the lowest receiver sensitivity. Towards this aim, each CR user chooses its transmission power that provides the minimum bit error rate (BER) at the CR receiver, and at the same time avoids interference to neighbors.

When a given CR node i has packets queued, it evaluates a spectrum utility for link (i, j), where j is a potential next hop forwarder. The utility function depends upon the queue size, with the nodes with smaller backlogged packets have a higher probability of being selected as next hop. ROSA incorporates a distributed algorithm that i) selects the **best next hop** j, ii) performs **spectrum allocation** (which set of minibands F to to use, and at what power, at the transmitter) based on local queue and spectrum occupancy information collected from the neighboring nodes. The selection of the above is obtained by solving an optimization problem that maximizes the spectrum utility. U_{ij}^s:

$$(\mathbf{F}, \mathbf{P}, j, r) = \arg\max\{U_{ij}^s = c_{ij}(\mathbf{F}, \mathbf{P}, r)(Q_i^{s_{ij}^*} - Q_j^{s_{ij}^*})\}, \tag{9.4}$$

where s^* is the session with maximal differential backlog on link (i, j), and $c_{ij}(t)$ represents the achievable capacity for link based on the classical Shannon limit. The transmission rate on the link is given by r. The utility function is basically the product between the achievable data rate (capacity) and the differential backlog of the session. By maximizing the expression above based on the current dynamic spectrum, queueing, and channel fading conditions, the node chooses jointly the i) next hop, and ii) spectrum allocation, i.e., power, and frequencies to be used to maximize the spectrum utility.

While ROSA achieves true cross-layer performance, the overhead associated with assimilating the significant amount of neighborhood information at each node remains to be evaluated. Moreover, the success of the protocol relies on the presence of a CCC, and the distributed method needs to be analytically compared in terms of performance bounds with the centralized benchmark scheme, also proposed by the authors.

A different approach using AODV-like route formation is proposed in Spectrum aware routing protocol (SPEAR) [229]. However, unlike AODV, multiple routes are formed between the source-destination pair. The authors discuss different approaches by which the destination may select the best path, and implement a maximum throughput based solution. Here, the route is selected by sorting on the maximum possible throughput, and then using the hop count metric to break ties.

9.5.3 Identifying routes with long lifetime

In this section, we discuss two protocols that allow for long route lifetime by either attempting to maintain the entire route constant by moving all nodes on the route to a new spectrum (fast recovery in STOD-RP), or by estimating node mobility proactively (SEARCH protocol).

9.5.3.1 Routes with fast route recovery Proposed for wireless mesh networks, STOD-RP has two key components: The first is proactive routing, and the second is on-demand route discovery [303]. The basic idea of this work is that a routing tree is formed in each spectrum band, centered on a root node within the network. The root node for the tree is selected in a distributed manner by flooding the network with a *root request* message containing the number of available spectrum bands, and the duration for which these bands are available.

This work also defines a new link quality metric for a given link l_i that is used during the route formation:

$$C_i = \left[O_{ca} + O_p + \frac{P_{kt}}{r_i} \cdot \frac{1}{1 - e_{pti}} \cdot \frac{1}{T_{l_i}} \right], \tag{9.5}$$

where, the O_{ca}, O_p, and P_{kt} are constants for specific access technologies for channel access overhead, protocol overhead, and packet size, respectively. The terms r_i, e_{pti}, and T_{l_i} represent the link rate, packet error rate, and the spectrum availability duration, respectively for the link. This metric is used to measure the quality of a route by a the summation of the terms C_i for all the k links that are present in the route. The final end-to-end route quality metric also incorporates the number of times the spectrum is switched (M), and also the time delay of each switching instance (D_{sw}), as shown below:

$$C = \sum_{i=1}^{k} C_i + M \cdot D_{sw} \tag{9.6}$$

The operation of the STOD-RP protocol is explained in detail below:

- *Proactive Routing:* This stage is similar to classical AODV, with a few subtle differences. The route metric defined earlier in (9.5) and (9.6) are used. All packets area sent by the CR users to the root node, which must then determine if the destination is within the same spectrum tree, or in a different one. This leads to intra-tree and inter-tree routing cases, and each time the root node forwards the packet to either the next hop within the same tree, or another node that exists in both the source and destination spectrum trees.

- *Route recovery:* The tree formation allows quick reconfiguration in case the PU reclaims the spectrum. The root node informs all the other nodes of the tree of a new spectrum, and thus, all the nodes transition to this spectrum without any change in the tree-topology. The spectrum change is hence transparent to the ongoing routing function. In the event that no new spectrum is available, then the nodes break away from the tree and attach themselves to the

One of the problems of this scheme is the large-scale flooding during the root selection stage. Every node broadcasts a packet which considerably adds to the network overhead. Moreover, tree formation can be guaranteed only for quasi-stationary topologies.

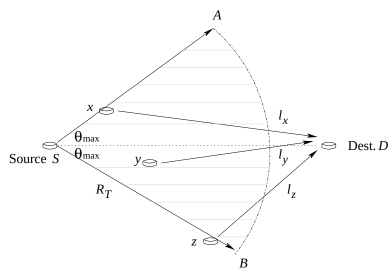


Figure 9.9 The greedy forwarding and focus regions

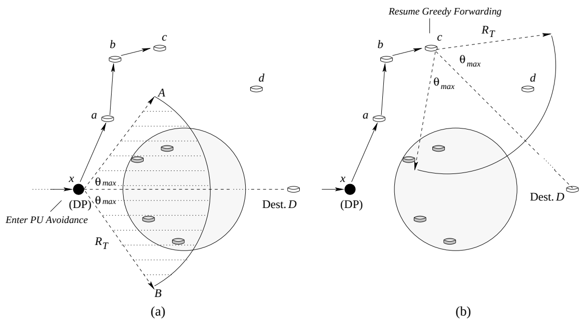


Figure 9.10 The PU avoidance phase

9.5.3.2 Routing protocol incorporating node mobility The SEARCH protocol attempts to find the length of the shortest path based on greedy advancement that may be traversed on a combination of channels to the destination [52]. The key functionality in our proposed approach is evaluating when the coverage region of the PU should be circumvented, and when changing the channel is a preferred option. It also has a route maintenance function that deals with route outages due to PU arrival and node mobility.

SEARCH operates in two modes - *greedy forwarding* and *PU avoidance*, depending on whether the RREQ is propagating along the greedy shortest path to the destination or needs to circumvent a region of PU activity, respectively.

In the greedy forwarding phase, the RREQ is sent out on each of the available channels by the source. However, only the nodes that are not currently in a PU coverage region may forward it. Moreover, the chosen forwarders must lie in a specific region around the current hop, called as the *focus region*, and the authors provide a set of performance evaluation results with different extents of the focus region. Thus, a node with a lesser advance towards the destination but within the focus region is chosen over another node closer to the destination that lies outside this area. These two aspects differentiate SEARCH from the classical ad-hoc protocols like [138]. In Figure 9.9, the source S has the nodes x, y and z within its transmission radius R_T. These nodes are at a straight line distance of l_x, l_y and l_z, respectively, from the destination D, where $l_x > l_y > l_z$. The focus region for S is shown by the sector $S - AB$ and extends to an angle of θ_{max} from the line SD. Classical geographic routing protocols like GPSR would have chosen node z at this stage, while SEARCH chooses the node with the greatest advance *within* its focus region, i.e. node y. If no such node exists, then SEARCH switches from the greedy forwarding phase to the PU avoidance phase.

SEARCH defines a node that lies in the focus region of the previous hop along the path, but does not find a forwarding node for the RREQ in its own focus region, as the decision point. Figure 9.10(a) shows the shaded circular area under the influence of a PU on the channel being used for forwarding the RREQ. In addition, the focus region for node x on this channel is given by the sector $x - AB$ with the maximum angle of $2 \cdot \theta_{max}$. Some of the nodes that sense the PUs and do not participate in the forwarding of the RREQ, lie in the focus region of the node x. Through the periodic beacon update, these affected nodes inform their one-hop neighbors, including node x, of the current state of the channel environment. Thus, node x is aware that the closest node to the destination that can forward the RREQ (node a) lies outside its focus region. Also, node x concludes that it is a DP. The DP marks the point from which the route must circumvent the region of PU activity on the given channel. There may be several such DPs in the path to the destination and this information is collected by the RREQ as it traverses through the network. When a node is reached that has a candidate forwarder in its focus region, in cancels the PU avoidance phase. In the example shown in Figure 9.10(b), the RREQ traverses the node a, b and finally reaches node c. The latter has a candidate forwarder, node d, that lies in its focus region. At this point, i.e. at node d, the PU avoidance phase is completed, and the greedy forwarding is resumed.

The route management phase recognizes that nodes may be mobile, and each node predicts the future location of the next hop downstream nodes by the Kalman filtering method [139]. Here, the signal strength of the periodic beacons transmitted by the neighbors are used to estimate the distance between two adjacent nodes on the routing path, and subsequently, the connectivity of the route. If a given next hop node is predicted to move out of range, another node closest to the original location used during the route formation stage is selected. If no such node is found, a new route may be formed.

A slightly different approach is undertaken for non-geographic routing in [104], where the authors derive probabilistically the connectivity of the link ($L(\widehat{T_p})$) as a function of the required time ($\widehat{T_p}$) for which the link needs to be active. This probability incorporates both the effect of user mobility that causes the nodes to go out of transmission range, as well as the effect of moving into a region of PU activity (and still remain connected to the previous hop). In either case, the channel can no longer be used. Analytically, this probability is derived as:

$$L(\widehat{T_p}) \approx e^{-\lambda \widehat{T_p} e^{\lambda\tau}} + \zeta(1 - e^{-\lambda \widehat{T_p}}), \tag{9.7}$$

where λ is the initial separation distance, τ and ζ are constants derived through measurements. Once the suitable link availability metric is obtained, based on (9.7), the routing path can be constructed. The authors use the distributed localized Dijkstra topology control

(LDTC) algorithm, which runs the classical Dijkstra algorithm over a neighborhood graph. Hence, this approach is distributed, and requires only local topology information.

9.6 ROUTING CHALLENGES

There are several major challenges and open research topics that has not been addressed in the current state-of-the-art. Below, we summarize the open research issues for routing in CR networks:

- *Need for analytical modeling:* Owing to the several complex interactions between the spectrum management and network layer routing, it is challenge to provide steady-state performance guarantees in CR routing protocols. From the practical implementation standpoint, it is important to develop theoretical end-to-end models that incorporate a large number of variables, topology of the CR network, PU locations, their respective receiver sensitivity thresholds, maximum transmission power, statistical PU activity, among others. Moreover, the models must yield a solution within reasonable time, using available computational tools. Analytical models rely on simplifying assumptions of the network operation. Thus, we believe that future modeling efforts shall rely on carefully choosing which factors are critical for the CR operation, and laying greater emphasis on them in the derived models. As an example, with improvements in radio technology, spectrum switching can be undertaken in few μs. However, PU "on" durations may last several hours, and is hence, more important to capture this feature in the model to ensure it is practicable, and yet tractable.

- *Route management:* When a PU reclaims the channel, the CR user is unable to transmit for the duration in which the PU is active. When the typical PU “on” times are short, this disruption is for a minimal time. An interesting area for future work is deciding when should a fresh route formation be signaled, and under which conditions the current route is kept alive, albeit in a dormant state. When the route failure message is aggressively generated, it results in a considerable network overhead to form fresh routes. However, a prolonged delay will degrade the end-user performance, and an optimal balance must be carefully derived based on PU activity models. Moreover, in case of route failure, protocols that focus on local route repair by modifying portions of the existing route (through new next hop or spectrum selection) must be preferred over those that simply delegate the responsibility of route repair to the source node.

- *Need for incorporating existing whitespace databases:* To date, routing protocols base their next hop selection on either the past spectrum history communicated by the neighboring nodes, or on the basis of the statistical knowledge of the PU activity. A different, and perhaps the most reliable approach is obtaining the spectrum usability data directly from publicly available spectrum databases that are being constructed. In the US, the Whitespace Database group has been formed to create such a repository of spectrum information. In Europe, few institutions have undertaken such measurements [269], and this database will greatly help researchers. In the ideal case, these databases shall have multiple access points and will be updated in real time. Each node can now ensure that accurate region-specific spectrum information is available selecting the routing paths.

CHAPTER 10

TRANSPORT LAYER PROTOCOLS FOR COGNITIVE RADIO NETWORKS

10.1 INTRODUCTION

The transport layer has the main responsibilities of congestion control and reliable delivery. As it runs on the end devices, the transport protocol is expected to estimate the state of the network through the clues provided by packet arrival delays and packet losses. While this *pure* end-to-end paradigm works well for wired devices, owing to the wireless channel uncertainties, transport protocols often require additional information about the state of the node. In CR networks, we believe that participation of the intermediate hops is critical to the success of the transport protocol. Though the rate of sending packets may still be controlled at the source, *when* and by *how much* are hard questions to address in a dynamic spectrum environment, without the nodes on the chosen path reporting their sensed spectrum information to the source.

At the transport layer in classical wireless ad hoc networks, the main challenge lies in distinguishing (i) congestion, (ii) channel-induced, and (iii) mobility-based packet losses. In the first case, the packet experiences greater queueing delay in the buffers of the intermediate routes, thereby also increasing the round trip time (RTT) resulting in TCP timeout events. In the second case, occasional channel related losses, such as those caused by fading or shadowing may cause a packet drop, and this is often mistaken by the source as a congestion event. Mobility related losses are mostly permanent, and if the sender already has a large number of in-flight packets, then all of them are likely to be lost. Though these loss-inducing factors are also applicable to CRAHNs, an increase in the observed RTT may be caused if an intermediate node on the route is engaged in spectrum sensing, and hence,

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

unable to forward packets. Also, the sudden appearance of a primary user may force the CR nodes in its vicinity to limit their transmission, leading to an increase in the RTT. In such cases, the network is partitioned until a new channel is identified and coordinated with the nodes on the path. As an example, in Figure 10.1, consider a chain topology formed by the source S, destination D, and intermediate forwarding nodes. If the node 2 is performing spectrum sensing, then for that duration, it is unable to send or receive packets, resulting in a virtual disconnection of the path. Consequently, the data packets in node 1 and moving towards D, and acknowledgements (ACKs) in node 3 for the source S, both experience greater queueing delays. If a timeout indeed occurs, the source is immediately penalized and the rate of sending data is drastically reduced. Similarly, consider the case in which the spectrum used by node 4 is reclaimed by the licensed or primary user (PU), and it must immediate cease transmission. There is a finite time duration in which node 4 must identify a new spectrum, switch its transceivers, and coordinate this choice with its neighbors. Thus, in both the above cases of spectrum sensing and switching, the source may mistake the increased RTT (or timeouts caused by this increase) for congestion. One possible approach to address this issue is to involve the intermediate nodes to a greater extent, allowing them to give periodic feedback to the source. As an example, the in TP-CRAHN protocol that extends TCP for CRAHNs [53], intermediate nodes periodically piggyback their spectrum information on the ACKs, or in times of a sudden event like a PU arrival, explicitly notify the source. While several works have focussed on spectrum sensing algorithms in the last few years [6], the integration of the channel information collected at the nodes and the performance study of these approaches from the viewpoint of an end-to-end protocol remains an open challenge.

The local spectrum decisions undertaken by a node strongly influence the end-to-end performance. As an example, if the spectrum sensing duration is large, then the node can better detect the PU activity in its local area. However, this also results in lower end-to-end throughput [238]. Thus, the optimal balance between protection to the PUs (higher sensing time) with the increase in CR network throughput (lower sensing time) must also be decided.

The frequent spectrum changes by the nodes within the CRAHN may result in a significant change in bandwidth of the affected link. Here, the number of packets that can be supported by the network in a unit time can suddenly increase, especially if the earlier spectrum was the *bottleneck* spectrum allowing very low date rates. A possible solution to this problem is the artificial scaling of the TCP congestion window (*cwnd*) to respond quickly to the change in the environment. As shown in Figure 10.1 (b), when node 4 switches the spectrum, choosing a higher capacity channel for the link $4-5$ (node 5 is not shown) the corresponding *cwnd* is increased immediately from its normal linear trajectory at B to a new value B' that allows the source to fully utilize the spectrum. This is especially important as spectrum is available for limited durations, and the CR user must make the most efficient use of it.

In addition to the above spectrum-related considerations in a CRAHN, there are several concerns of classical wireless ad hoc networks, such as the effect of mobility that must be accounted for in a dynamic spectrum environment. Individual nodes are likely to be engaged in frequent spectrum sensing and switching effects, and these events generally occur asynchronously along the path. Hence, any mobility-based updates to the source may be severely delayed due to the repeated interruptions to end-to-end performance.

The remainder of this chapter is dedicated to exploring these unique characteristics of transport protocols in CRAHNs, through simulation studies and the description of the main contributions in this area. At this point, we would also like to highlight the need for further research for TCP and TCP-friendly protocol design. TCP, in general, is a well researched area and several theoretical models exist that explain and predict its behavior in wireless networks [274]. It is also implemented at the transport layer for commercially available devices. In addition, the ad-hoc network may ferry user traffic to and from the external infrastructure network, receiving configuration commands from remote stations. TCP is the de-facto standard in the wired world and a measure of compatibility is useful from the network management perspective. Hence, the goal of a TCP-friendly protocol is to retain the window-based approach of the classical TCP, allow TCP-based streams to fairly utilize the network resources, and at the same time, introduce novel changes to the classical TCP design that allow its applicability in CRAHNs.

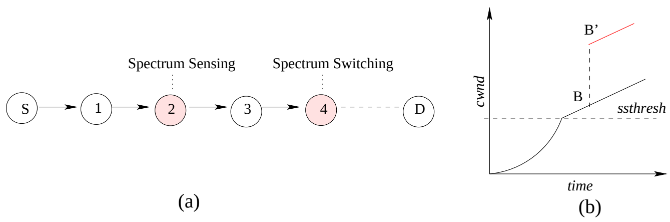


Figure 10.1 A multi-hop CR ad-hoc network (a) and the forced *cwnd* scaling (b).

10.2 CURRENT STATE-OF-THE-ART

As the transport protocol usually runs at the end nodes (source and destination), it has limited knowledge of the conditions of the intermediate nodes. Classical TCP suffers from some of the issues outlined in Section 10.1, and efforts have been made to address them for wireless scenarios, as described next.

10.2.1 TCP Adaptation in Classical Wireless Ad Hoc Networks

[159][244]. However, these protocols for classical wireless ad-hoc networks do not consider the cases that may arise in CR ad-hoc networks. As an example, in a classical wireless ad-hoc network, packets may incur a longer round trip time (RTT) owing to network congestion or due to a temporary route outage. In CRAHNs, a similar effect on the packet RTT may be caused if an intermediate node on the route is engaged in spectrum sensing and hence, unable to forward packets. Also, the sudden appearance of a primary user may force the CR nodes in its vicinity to limit their transmission leading to an increase in the RTT. In such cases, the network is partitioned until a new channel is identified and coordinated with the nodes on the path. The duration of the periodic spectrum sensing decides, in part, the end-to-end performance - a shorter sensing time may result in higher throughput but may affect the transport layer severely if a PU is mis-detected. While several works have focussed on spectrum sensing algorithms in the last few years [6], the integration of the channel information collected at the nodes and the performance study of these approaches from the viewpoint of an end-to-end protocol remains an open challenge.

Since its original proposal in 1974, several versions of TCP have been proposed for wired networks. All of them provide congestion and source-rate control, by means of a congestion window (CW) which limits the total number of un-acknowledged packets which can be in transit end-to-end. Modern implementations of TCP operate in four different protocol phases: slow start, congestion avoidance, fast retransmit and fast recovery. In the slow start phase, the CW size grows exponentially by one segment for each TCP-ACK received. In the congestion avoidance phase, the CW size is increased by one segment per round-trip-time (RTT), till a packet loss occurs. Packet losses are used as an indicator of congestion. In such a case, the rate is decreased by reducing the CW size at the sender side. Different versions of TCP differ in the way they detect and react to packet losses. TCP New-Reno [82], which is probably the most commonly used TCP on Internet, reduces the CW size when (i) three duplicate TCP-ACK packets are received (ii) or when the Retransmission Time Out (RTO) of a segment expires before receiving the corresponding TCP-ACK. Other modern variants of TCP are Vegas [17] and TCP-SACK [178]. TCP Vegas adapts the current CW size at the sender side, so that the number of queued packets in the network is always between a minimum and maximum threshold value. TCP-SACK implements a selective retransmission scheme, so that only packets actually missing are retransmitted by the sender node. While all these variants work well over stable wired connections, many recent papers have investigated and provided evidence that TCP performs poorly on wireless environments, and on mobile ad hoc networks in particular. Simulation studies for wireless networks have investigated the impact of hidden terminal, wireless channel errors and node mobility on TCP performance [113]. Moreover, many TCP modifications have been proposed for wireless ad hoc networks. Ad hoc TCP (ATCP) [159] utilizes network feedback to detect packet losses caused by mobility or by channel errors rather than by network congestion. In ATCP, the standard TCP is not modified but a thin layer between IP and transport layer is added to filter network feedback and to adapt the CW accordingly. A different approach is exploited in [244], where a completely new protocol (called ATP) is developed for wireless ad hoc networks. ATP decouples congestion control from reliability mechanisms, and exploits feedback from the intermediate nodes traversed by the connection to adapt the sending rate.

While TCP over traditional ad hoc networks constitutes a well investigated research area, there is a lack of papers addressing transport protocols for CRAHNs. The existing works applicable for CRAHNs are described next.

10.2.2 TCP Adaptation in CRAHNs

In [166][165], the authors propose to improve the TCP end-to-end performance over CR links through an adaptive spectrum selection scheme. Their cross-layered approach is shown in Figure 10.2, which optimizes the throughput based on channel selection, sensing, channel access decision, selection of the modulation and coding scheme, and the frame size. The authors do not propose a new TCP protocol, or make any changes to the existing TCP newReno. Instead, they describe a learning framework that decides the above parameters using rewards measured on the basis of observed TCP throughput. As the TCP functioning itself is not modified, and the decisions can be made locally at the nodes, the search space for the optimal parameters for the variables defined in a CRAHN is extremely large. Moreover, even if the system converges to the optimal set after several trials, the dynamic behavior of the network may not allow the system to remain in this equilibrium state for extended durations of time.

In [232], the authors propose a cross-layered architecture as shown in Figure 10.3. The authors consider a separate knowledge module (KM) is for storing information or knowledge about the (i) applicationŠs needs and the (ii) state of the network environment. As this knowledge is relevant to several different layers of the protocol stack, the KM is spread over the stack. The cognitive module (CM) stores the (i) algorithms and CR

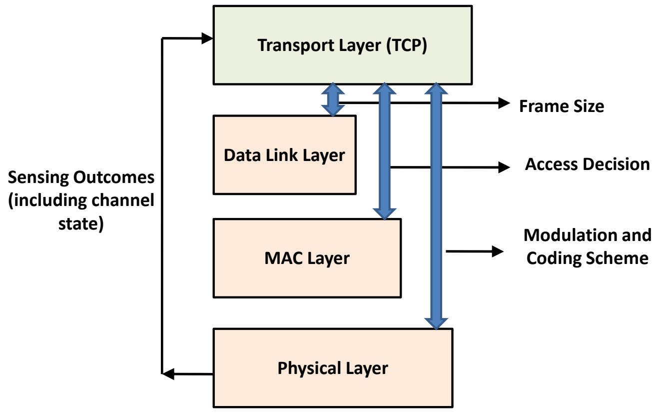


Figure 10.2 The cross-layered architecture of the transport protocols proposed in [165].

policies, as well as the (ii) control signals that use the KM to manage the operation of the transport layer. Using this framework, a family of TCP variants for CR networks, which implement adaptive scaling of the congestion window based on the available bandwidth estimation. The authors propose a new congestion window calculation technique based on the classical TCP Westwood [177]. The latter alters then congestion window calculation of TCP newReno on the basis of the minimum RTT (RTT_{min}) and the estimated available bandwidth (EBW), as follows:

$$W_{i+1} = RTTmin \cdot EBW_i, \tag{10.1}$$

where $EBW_i = \alpha EBW_{i-1} + (1 - \alpha_i)OBW_i$. The coefficient α can be static or dynamic, and OBW_i is the observed bandwidth. The authors in [232] suggest that the expected bandwidth be unchanged (i.e., $EBW_i = EBW_{i-1}$) if the current value of the RTT is close to RTT_{min}. This prevents the TCPE bandwidth estimation to incorrectly reduce during spectrum sensing. However, this work does not consider the impact of sensing activity, the durations for spectrum switching, and network mobility on the rate and congestion control algorithms of TCP. We next explore in detail the specific problems faced by TCP through a quantitative analysis, followed by a novel solution called as TP-CRAHN.

10.3 CR ENVIRONMENT CHALLENGES AT THE TRANSPORT LAYER

In this section, we discuss the problems with the existing implementations of transport protocols based on TCP NewReno in CR ad-hoc networks, in which, nodes are equipped with a single radio transceiver. The features of the CR network that we study are: (i) spectrum sensing (ii) effect of primary user (PU) activity, and (iii) spectrum change. On any given channel, the PU may be modeled as Poisson arrivals, with an "*on*" time ($\frac{1}{\alpha}$) and "*off*" time ($\frac{1}{\beta}$). The reader should note, however, that despite the popularity of this model, extensive measurement studies in the TV channels point to a more long-tailed nature of the PU activity model [41]. Thus, the PU "on" (or "off') time distribution is exponential-like, but with a long-term tapering probability of being active (or inactive) at a given time.

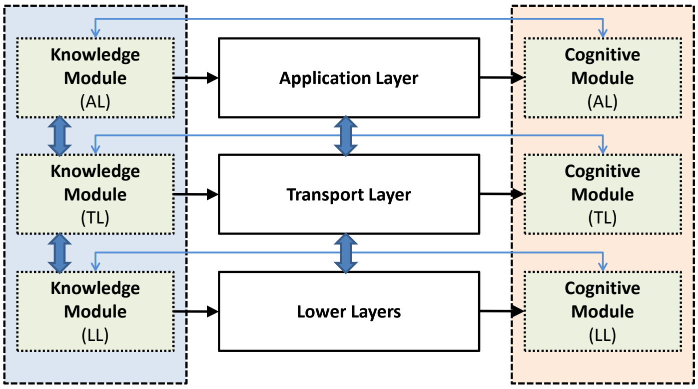


Figure 10.3 The cross-layered architecture of the transport protocols proposed in [232].

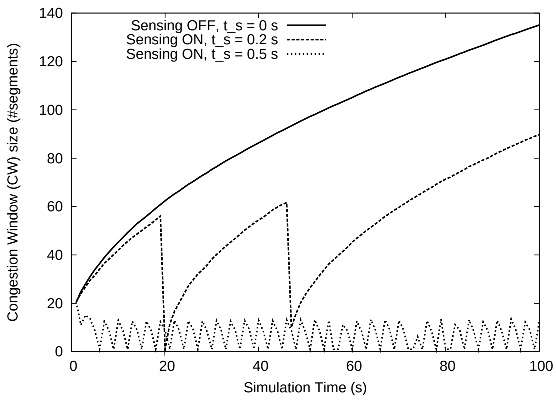


Figure 10.4 A study of the *cwnd* size as a function of the varying sensing time.

10.3.1 Spectrum Sensing State

CR users periodically monitor the current channel over a pre-decided sensing duration for the occurrence of PUs before using it for transmission. During this interval, the nodes are not actively involved in transmitting data packets, and the multi-hop network is virtually disconnected at the node performing spectrum sensing.

In Figure 10.4 we show the impact of sensing-induced delay on TCP NewReno performance, when there is no PU activity on the current channel. We analyze the behavior of the Congestion Window (CW) size under three different configurations of the sensing time t^s, i.e. 0s (sensing disabled), 0.2s and 0.5s. When sensing is disabled, we observe that the CW keeps increasing till the capacity of the channel is reached. When sensing is enabled, DATA and ACK packets experience an extra-delay which triggers timeout events at TCP sender side. As a result, TCP reduces the CW to 1 segment, and resets to the slow-start state. When t^s is equal to 0.5, the sensing delay is comparable with the maximum RTO timer value, and thus, frequent RTO events are triggered, degrading the end-to-end performance. This analysis is also in accordance with results shown in [232]. In [80] we also showed that the duration of t^s can play a critical role in deciding the optimal end-to-end throughput, because it constitutes a trade-off between (*i*) accurate PU detection and (*ii*) efficient channel utilization. Thus, it is responsibility of the transport layer to adapt the current rate during the sensing state, and to decide the optimal setting of t^s so that the throughput is maintained at the desired level while the interference on PUs is minimized.

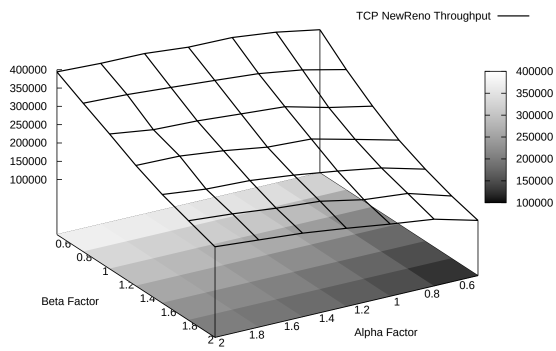


Figure 10.5 The impact of different PU activity on TCP throughput.

10.3.2 Effect of PU Activity

On detecting the presence of a PU, either during spectrum sensing or an ongoing data transfer, the CR users cease their operation on the affected channel and search for a different vacant portion of the spectrum. While the spectrum sensing on the current channel is periodic and has a well defined interval, the time taken to (*i*) search for a set of available channels on different spectrum bands, and (*ii*) coordinate with the next hop neighbors to find a mutually acceptable channel in this set, is generally uncertain. Moreover, the path to the destination is disconnected until the new channel is successfully found, the time for which is not known to the source in advance. Thus, the transport protocol needs to differentiate this state from other causes of route disconnections with the help of an explicit feedback from the nodes affected by the PU activity.

In Figure 10.5 we show the impact of PU activity in terms of average "*on*"-time (x-axis) and "*off*"-time (y-axis) on the TCP NewReno throughout (z-axis). Based on the values of α and β, it is possible to distinguish among 4 different patterns of PU activity: High-Activity Region ($\frac{1}{\alpha} \leq 1, \frac{1}{\beta} > 1$), Low-Activity Region ($\frac{1}{\alpha} > 1, \frac{1}{\beta} \leq 1$), Short-Term Activity region ($\frac{1}{\alpha} > 1, \frac{1}{\beta} > 1$) and Long-Term Activity region ($\frac{1}{\alpha} \leq 1, \frac{1}{\beta} \leq 1$). Not surprisingly, TCP performance are maximized when the CRs have more possibility to access the licensed spectrum without interfering with the PU activity, (i.e. in the Low Activity Region) and minimized when the PUs are more active on the current channel (i.e. High Activity Region). At the same time, results shown in Figure 10.4 and discussed in [79] demonstrate that TCP suffers of performance decrease when there are frequent "*on*"-"*off*" switches (i.e. in the Short Term Activity Region) due to the fact that the CW can not increase because of frequent PU arrivals on the current channel. As a result, we believe that transport layer should be informed of spectrum handoff operations occurring at the lower layer, in order to distinguish packet losses caused by congestion or by PU interference.

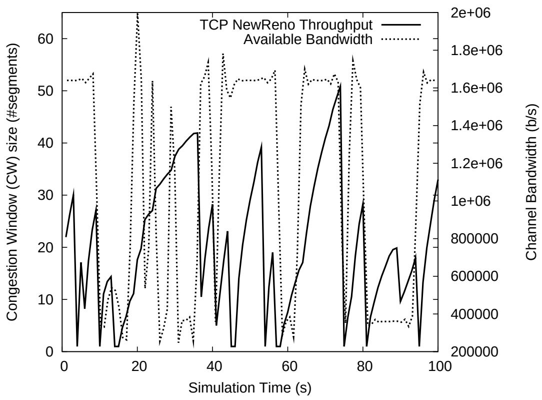


Figure 10.6 The effect of changing channel bandwidth on *cwnd*.

10.3.3 Spectrum Change State

A key concern in CR networks is the efficient utilization of the spectrum resource, as the opportunity for transmission in the licensed bands is available for a limited time. The licensed channels may have a large variation in bandwidth, especially as nodes switch from one spectrum band to the other. In Figure 10.6, we study through simulation how classical TCP increases the *cwnd* as it probes for the additional bandwidth available on a single link. There are three different channel bandwidths possible- $2/3\,\mathrm{Mbps}$, $4/3\,\mathrm{Mbps}$, and $2\,\mathrm{Mbps}$. The vertical bars denote the bandwidth available to the node and at any given time, this is the upper limit that can be utilized by the TCP connection. This gives three distinct levels of bandwidth availability with time. On each channel, the PU is modeled as a Poisson arrival, with an "*on*" time ($\frac{1}{\alpha} = 4$s) and "*off*" time ($\frac{1}{\beta} = 5$s). When the PU arrives, the CR user switches to a different channel, and consequently TCP must adjust to the new available bandwidth. From the figure, we observe that the *cwnd* is unable to correctly track the available bandwidth. Moreover, the spectrum opportunity is often lost before the *cwnd* has increased to half the segments that may be supported on the new channel. A similar conclusion is drawn in [238], where TCP cannot effectively adapt to brief reductions in capacity, if the end-to-end delay is large. We believe that the *cwnd* in TCP must be scaled appropriately to meet the new channel conditions, as shown in the transition from the operating point B to the point B' in Figure 10.1 (b). Estimating this new operating point is a challenge and link layer metrics, that determine the effective bandwidth, must also be considered apart from the raw bandwidth. Bandwidth estimation techniques have been proposed in [38][177], that do not require information from the intermediate nodes, but also do not respond immediately to the available spectrum.

10.4 TP-CRAHN OVERVIEW

To address the above issues, TP-CRAHN is proposed that extends the classical TCP, with the finite state machine of the protocol shown in Figure 10.7. Our protocol comprises of the following 6 states [53]: (i) Connection Establishment, (ii) Normal, (iii) Spectrum Sensing, (iv) Spectrum Change, (v) Mobility Predicted, and (vi) Route Failure. Based on the feedback received from the destination and limited network knowledge collected from the underlying network, link and physical layers of the intermediate nodes, it enters into one of these states, as described next.

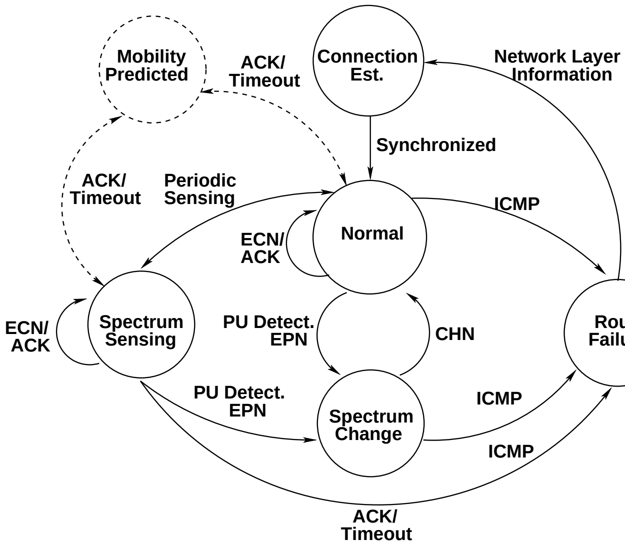


Figure 10.7 Finite state machine model of our transport protocol.

10.4.1 Connection Establishment:

TP-CRAHN modifies the three-way handshake in TCP newReno so that the source can obtain the sensing schedules of the nodes in the routing path. First, the source sends out a synchronization (SYN) packet to the destination. An intermediate node, say i, in the routing path appends the following information to the SYN packet: (i) its ID, (ii) a timestamp, and (iii) the tuple $\{t_i^1, t_i^2, t_i^s\}$. Here, t_i^1 is the time left before the node starts the next round of spectrum sensing, measured from the timestamp. t_i^2 is the constant duration between two successive spectrum sensing events, and t_i^s is the time taken to complete the sensing in the current cycle. On receiving the SYN packet, the receiver sends a SYN-ACK message to the source. The sensing information collected for each node is piggybacked over the SYN-ACK and thus, the source knows when a node in the path shall undertake spectrum sensing and its duration. The final ACK is then sent by the source to the destination completing the handshake.

We note that the calculation of the sensing time t_i^s by a node i is undertaken locally. Based on the bandwidth of the channel (W), the external signal to noise ratio (γ), and the probabilities of the on period (P_{on}) and the off period (P_{off}), a framework to calculate this time is given as follows [150],

$$t_i^s = \frac{1}{W \cdot \gamma^2}[Q^{-1}(P_f) + (\gamma + 1)Q^{-1}(\frac{P_{off}P_f}{P_{on}})]^2 \tag{10.2}$$

Equation (10.2) gives the sensing time t_i^s that minimizes the probability of missed primary user detection P_f, i.e., incorrectly stating the channel is vacant when indeed there is an active PU and Q is the standard Q function. The sensing times collected from the nodes are the preliminary values which may be dynamically updated.

10.4.2 Normal State:

This state is similar to the classical TCP newReno where the *cwnd* is increased based on the incoming ACKs. The congestion is notified to the source through the Explicit Congestion Notification (ECN) message, the following information from the intermediate nodes is piggybacked over the ACK- *Residual buffer space (B_i^f):* Consider a node i that has B_i^u bytes currently of unoccupied buffer space. Let the number of flows passing through it be n_i^f. The fair share of the residual buffer space per flow is, $B_i^f = \frac{B_i^u}{n_i^f}$. *Observed link bandwidth ($W_{i,i+1}$):* Each node i maintains a weighted average of the observed bandwidth on the link formed with its next hop, i.e. $\{i, i+1\}$, during the *normal* state. This is obtained from the link layer as the ratio of the acknowledged data bits to the time taken for this transfer between the nodes i and $i+1$. *Total link latency ($L_{i,i+1}^T$):* Let $L_{i,i+1}$ be the sum of the (i) time taken by a packet of the current flow to move to the head of the queue (ii) the time for contending the access to the channel and finally, (iii) the transmission time measured at node i with respect to the next hop $i+1$. The total link latency is now defined considering the bidirectional link latencies, $L_{i,i+1}^T = L_{i,i+1} + L_{i+1,i}$.

10.4.3 Spectrum Sensing State:

In this state, our proposed approach ensures that the intermediate nodes do not suffer from buffer overflow, then the path gets disconnected due to spectrum sensing by a node i. When the sensing time of the closest node is completed, the buffer space of node $j-1$ is used in the *ewnd* computations. Knowing the receiver advertised window *rwnd*, effective window, *ewnd* at the sender is modified to include an estimate of the free buffer space, B_{i-1}^f, at the previous hop node $i-1$ as $ewnd = \min\{cwnd, rwnd, B_{i-1}^f\}$. As the packets fill up the buffer in the node $i-1$, the remaining free buffer space needs to be progressively reduced, so that the effective window can be computed. Thus, the space available B_{i-1}^f in the node $i-1$ is decremented at intervals of $L_{i-2,i-1}$, when node i is engaged in sensing, and $L_{i-2,i-1}$ is the link latency made known to the sender in the *normal state*. We note that if the buffer at node $i-1$ is reaching the overflow limit, the congestion condition will be signaled and the *cwnd* will be reduced to 1 at the source as a response.

From equation 10.2, we observe that the sensing time is a function of the target missed detection probability. When a node is in a region that is not subject to frequent spectrum changes due to PU activity, or the target detection probability is low, the sensing duration can be reduced. This allows the CR users to accrue greater end-to-end throughput. In order to identify a specific node i for adjusting the sensing time, our approach ranks the nodes in the path based on the number of times the operational channel was changed due to PU activity by keeping a count of the CHN messages. Intuitively, the node that generated the highest proportion of the CHN message also experienced the maximum number of PU detection events and thus, must be located in a region of frequent PU activity. Such a node needs to retain a higher sensing duration.

10.4.4 Spectrum Change State:

Consider three nodes given by $i-1$, i and $i+1$ on the current path and the channels used by the links $\{i-1, i\}$ and $\{i, i+1\}$ be $c_{i-1,i}$ and $c_{i,i+1}$, respectively. If the PU is on the channel ξ_p^x and either $c_{i-1,i} = \xi_p^x$ or $c_{i,i+1} = \xi_p^x$, the node i must search for a new channel to prevent interference to itself and to the PU, respectively. At this stage, it sends an explicit pause notification (EPN) to the source, which in turn, freezes the protocol state and waits for a new channel CHN message to resume the transmission. The CHN message contains the estimated link bandwidth $W_{i,i-1}$ calculated using the observed link latencies $L_{i,i-1}$ and $L_{i-1,i}$, as follows:

$$W_{i,i-1} = \frac{P_{probe} + P_{ACK}}{L_{i,i-1} + L_{i-1,i}}, \tag{10.3}$$

where the probe and ACK packets exchanged over the link are of the size P_{probe} and P_{ACK}, respectively.

On receiving the CHN message, the source first estimates the new RTT using (i) the earlier observed RTT' during the last *normal* state of the protocol and (ii) adjusting for the new bidirectional link delay, $L^T_{i,i-1}$ as $RTT = RTT' + L^T_{i,i-1} - {L'}^T_{i,i-1}$. For the given path of n nodes, let W'_b be the old observed bottleneck bandwidth, before the channel change. After the channel change, the new bottleneck bandwidth is identified as W_b. The updated estimate of the bandwidth $W_{i,i-1}$ is used in this calculation from equation (10.3). If the ratio of the old bottleneck bandwidth to the new is within a permissible range, then the congestion window remains the same, or else it is scaled by a factor α.

10.4.5 Mobility Predicted State:

The nodes of the path monitor the connectivity to their next hop downstream node by measuring the RSS of the ACKs and the periodic beacon messages. At each epoch, the prediction value is compared with the minimum RSS required for receiver operation. If the condition of possible link failure is predicted in the next epoch, the destination is informed, which then sets the *mobility flag* (MF) in the outgoing ACKs. The source responds to this by limiting the *cwnd* to the $ssthresh$ and the congestion avoidance phase is never initiated. The aim of this adjustment, *cwnd* $\leq ssthresh$, is to limit the number of packets injected into the route which has a possibility of an outage. If no ICMP message is received at the source subsequently, the protocol reverts back to the *normal* state, where the *cwnd* is no longer bounded.

10.4.6 Route Failure State:

This is the terminal state of the current cycle and a fresh TCP connection must be established when a new route is formed. The protocol enters this state on receiving the route failure message (ICMP) and the source must cease transmission immediately.

10.5 OPEN RESEARCH CHALLENGES

- *Reduced intermediate feedback:* TP-CRAHN's approach relies on intermediate nodes of the path reporting back local information. However, in the true spirit of end-to-end reliable delivery, the transport layer should be able to infer these conditions merely from the observations at the end points of the connection. Thus, spectrum estimation techniques must be developed that allows TCP to regulate the sender rate without explicitly probing the nodes for their link latency, observed link bandwidth, among others.
- *Coexistence:* Coexistence of different flavors of TCP protocol needs to be analyzed. Questions of exhibition of aggressive behavior, compatibility with existing TCP flavors, and comparisons with rate-based schemes needs to be addressed.

- *Congestion window scaling:* During spectrum change, TP-CRAHN evaluates the ratio of the old bottleneck bandwidth (W_b) to the new bottleneck bandwidth (W_b') to check if it lies in the range $[1-\varpi, 1+\varpi]$, i.e. $\frac{W_b'}{W_b} \in [1-\varpi, 1+\varpi]$. If so, then no scaling of the earlier *cwnd* is needed. There are two important questions that must be investigated further: (i) What should the allowed range determined by ϖ be, and (ii) if the *cwnd* should be scaled to a value $\alpha_c \cdot W_b \cdot RTT$, then how should the scaling factor α be chosen?

- *Error Modeling:* To combat wireless interference losses with the changing environment and the PU activity, based on loss rate, novel compensation mechanisms need to be introduced at the source. Hence, modeling of spatially varying error rate due to licensed user activity may be needed.

- *RTT Estimation:* The classical mechanism of RTT estimation needs to be refined to correctly incorporate the effects of the spectrum-related functions of the cognitive cycle. This will allow correct determination of the end-to-end delay, and the possibility of dissociating the response of the protocol to the effects of PU activity and network congestion.

CHAPTER 11

SECURITY IN COGNITIVE RADIO NETWORKS

11.1 BASICS IN SECURITY

The purpose of this section is to provide the reader with the definitions of the features that a secure communication system should have, according to [179], where is also shown how existing wireless systems provide them. Sometimes, two or more features are given at the same time by a specific technology, but they could still be identified. Their meaning are also characterized to the CR context

- Availability: Availability is the capability to assure to the users the access to the network. In the general wireless context, it is referred both to the availability of the network and to the availability of the communication medium. If a network, or its communication medium, is not available, the aim of having the network itself cease to exist.

  In IEEE 802.11, the access to the wireless medium is assured to every user by collision avoidance mechanisms. In the CR context, interferences from CR users to PUs should be avoided even in those scenarios where PUs and CR users can coexist. Where PUs are absent, selfish behaviors of some CR users should be avoided too, making fair the spectrum access.

- Integrity: Integrity means the validity of the transmitted data. Any node in the network, even if trusted until a certain moment, could switch its behavior to malicious and modify the data flowing through it. Lack of integrity in the exchanged data

978-0470-68852-6

between two communicating parties can nullify the communication itself. However, at same time, some entity in-between the two communicating parties can need to legitimately modify some data, e.g., some flag in the header of a packet of a transport layer protocol.

In the classical wireless approach, one makes use of cryptography to provide Message Authentication Codes (MACs) and verify the sender identity.

- Identification: Identification is to map a user with a user identity uniquely. In the Global System for Mobile communications (GSM) the International Mobile Equipment Identity (IMEI) is used to verify which devices are allowed in the network and the Subscriber Identity Module (SIM) contains a cryptographic key to identify the subscriber (i.e., the user) on the network. A Public Key Infrastructure (PKI) can be used with CR.

- Authentication: The Authentication feature is needed to assure that a specific user in a network is indeed the one that he claims to be. Even a naive attacker could easily enter a wireless network because of the intrinsic flaw of the wireless medium, but he could not enter any provider system since it would be never authenticated as a legitimate user.

  In the CR context a PKI could be used, with a centralized Certification Authority (CA) granting credentials for the users, but this could be a very expensive infrastructure to be realized.

- Authorization: Authorization deals with granting different levels of privileges to the users. A user without any privileges can not even enter a network while a user with too much privileges could use them for selfish purposes.

  An authorization mechanism in a CR network should take into account the different kinds of users and be aware of the environment.

- Confidentiality: Confidentiality ensures that exchanged data are understandable only to those who have access to them and it is usually achieved by means of cryptography. IEEE 802.11 employs the Advanced Encryption Standard (AES) to provide confidentiality, as well as it could be done in a CR network.

- Non-repudiation: Non-repudiation means that a sender/receiver can be sure that the sent/received data are genuine, enabling them to prove that they have in fact sent/received those data.

11.2 COGNITIVE RADIO SECURITY THREATS

In this section, the security threats inherent to CR are discussed at each protocol layer, as proposed in [179]. Each threat can be either found in traditional wireless networks or specific for CR networks. Application Layer has been intentionally neglected.

Table 11.1. provides a prospect of these threats.

11.2.1 Physical Layer

The physical layer is the base of the protocol stack. It has direct access to the transmission medium and looks after of transmitting an information bit stream. Bit rate, channel capac-

Table 11.1. Security threats in the protocol stack as classified in [179]

Stack Layer	Security Threat
Transport	Key Depletion
	Jellyfish
Network	Network Endo-Parasite
	Channel Ecto-Parassite
	Routing Information Jamming
	Low Cost Ripple Effect
Link	Biased Utility
	Asynchronous Sensing
	Routing Information Jamming
	False Feedback
Physical	Intentional Jamming
	Primary Receiver Jamming
	Sensitivity Amplifying
	Overlapping Secondary User

ity, bandwidth, maximum achievable throughput and transmitting power depends directly on this layer.

The traditional wireless physical layer deals with a fixed frequency and has the main task of achieving the maximum throughput with the minimum transmit power, while peculiar to the CR physical layer is its parameters reconfigurability. According to the interference avoidance to the PUs, it should be able to transmit at different frequency bands of the spectrum pool and to accomplish at the same time to the traditional features of a physical layer, even during a frequency switching.

Among the threats to which the physical layer can be exposed are the following.

1. *Intentional Jamming:* In the most basic scenario, this is the attack performed by a malicious CR user who jams PUs and other CR users by intentionally and continuously transmitting in a licensed band. By increasing the transmit power and the number of jammed spectral bands, the attack can be more severe. Such attack can be easily detected with energy-based techniques, but the time needed to locate the malicious CR user, with triangulation techniques, can be extremely fatal for the network operation. The attacker cannot be even caught if would moving fast enough from a geographical area to another.

2. *Primary Receiver Jamming:* In a collaborative environment, a malicious CR can attack a Primary Receiver (PR) when he is about him, causing interference to his communication. The attacker continuously requests transmissions to himself from other CR users and, increasing the interference temperature, he prevents the PU from listening to primary transmissions, though the interference temperature itself is kept below a specified threshold at some other point in space.

3. *Sensitivity Amplifying:* In order to avoid interference, some PU detection techniques are performed with a higher sensitivity towards primary transmissions. This increases

the false detection probability and, consequently, the number of missed opportunities for the CR users. A malicious CR can amplify that sensitivity and, by replaying to the primary transmissions, increase the number of missed opportunities for the CR users. This attack can be realized also with a low power transmitter and the attacker can still cause missed opportunities for the CR users.

4. *Overlapping CR User:* In the scenario with multiple secondary networks coexisting over the same geographical area of a primary one, a malicious CR can transmit in a network interfering with the PUs and CR users of that one. This type of attack is hard to prevent because the attacker may not be under the direct control of the victim network.

11.2.2 Link Layer

Going back up through the protocol stack we find the Link Layer. Tasks of this layer are granting the access to the physical resources, data fragmentation, error correction and modulation. Channels assignment is due to the Medium Access Control (MAC) sub-layer.

Channels are not fixed for CR. They vary in the whole spectrum pool and the transmission might be simultaneously over multiple channels, in order to achieve a higher throughput. Besides the Signal-to-Noise Ratio (SNR), other parameters, such as the sensing time, the transmitting time, the interference temperature and the delay introduced by the cognitive cycle, should be taken into account for a genuine and fair channel usage.

1. Biased Utility: This is the selfish behavior of a malicious CR user, depriving other CR users of the transmission medium to increase his own bandwidth. If the CR users and/or base stations area unable to detect such anomalous behavior, some CR users may not even get to transmit.

2. Asynchronous Sensing: This is the case when a malicious CR user transmits asynchronously during other CR users sensing operations. If the base station or other CR users consider this as a transmission from a PU, then this could result in missed opportunities. This attack can be made more efficient by transmitting only during sensing periods.

3. False Feedback: For protocols that rely on secondary users exchanging information, false feedback from one or a group of malicious users could make other secondary users take inappropriate actions and violate the goals of the protocol.

11.2.3 Network Layer

Routing is performed in the network layer with flow control and Quality-of-Service (QoS) assurance. Besides the classical metrics, calculating the routing paths in a CR network concern also with the dynamic, and sometimes even fleeting, spectrum usage by the CR users. Awareness of the environment for each CR is essential.

1. Network Endo-Parasite: In the Network Endo-Parasite Attack (NEPA) there is at least one compromised or malicious node in the network and its aim is to increase the interference at heavily loaded high priority channels, affecting those links which are along the routing path through the malicious nodes towards the wired gateway.

Because of such behavior [179], they call this an endo-parasite attack. Normally, a node assigns the least loaded channels to its interfaces and updates with this information its environment. In case of NEPA, the malicious node assigns to its interfaces the high priority channels, without broadcasting this information to its neighbors, hence the network remains unaware. As result, there will be in the network a hidden usage of heavily loaded channels, hence the links using those channels experience interference, decrease in available bandwidth and continuous degraded performance.

2. Channel Ecto-Parasite: As reported in [179], "a compromised node launches CEPA by switching all its interfaces to the channel that is being used by the highest priority link".

11.2.4 Transport Layer

Flow control, error control and congestion control are performed at the transport layer. Dealing with CR networks, Round Trip Time (RTT) and packet loss probability should be reconsidered in light of the dynamic and opportunist spectrum usage.

Key Depletion: High RTTs and frequently occurring retransmissions at this layer will result in large number of sessions being initiated for any given application, hence the number of key establishments will increase the probability of using the same key twice. Key repetitions can be exploited to break the underlying cipher system.

11.2.5 Cross-Layer

We refer to cross-layer attacks as them which performed at a lower layer, will flaw security in an higher one.

1. Routing Information Jamming: This attack is proposed as a novel cross-layer attack in [179]. According to them, a malicious CR user jams the exchange of routing information among neighbors, exploiting the lack of a Common Control Channel (CCC) and the spectrum handoff delay, resulting in incorrect routing through the CR network. Right before the victim is exchanging routing information, the attacker causes spectrum handoff to his victim, who will stops all ongoing communication, to proceed to spectrum handoff. The victim then will release his current channel and select a new one, going through the cognitive cycle. During this time, the neighboring nodes will use a stale path towards the victim node. Performed at the link layer, this attack will affect the network layer routing information.

2. Jellyfish: According to [179], there are four variants of this attack. They are performed at the network layer to affect the transport layer and aim to reduce the throughput of the TCP protocol.

   The first is the misreordering attack. A malicious CR node intentionally and periodically reorders packets passing through it, causing TCP retransmissions and lowering the throughput. In the the packet dropping attack, a malicious CR node periodically and intelligently drops packets passing through it. When the number of dropped packets coincides with the TCP transmission window, throughput can be even dropped to zero.

In the third variant, which is called delay variance attack, the packets are randomly delayed by a malicious CR node when passing through it. “This cause TCP timers to be invalid which results in congestion inferences”.

A fourth variant is proposed in [179]. A malicious CR node leads the CR victim to a spectrum handoff, causing a considerable delay in the network and transport layers until pushing RTT to round trip timeout (RTO). RTOs result in retransmissions and hence drastic degradation of TCP throughput.

11.3 ATTACK TECHNIQUES

11.3.1 Primary User Emulation

As mentioned previously, CR users use the spectrum opportunities in an opportunistic way, without creating interference to the PUs. A Primary User Emulation (PUE) attack involves the sensing phase of the cognitive cycle. It is based on the emulation of a primary user, therefore the same emulation prevents the utilization of the spectrum by a secondary user. Indeed if the PUE attack occurs, a secondary user has to move from that channel because of the higher priority of the PUs, even if who is occupying the same channel is an unlicensed user. An attacker has a couple of reasons to emulate a primary user: due to a selfish behavior, that is to obtain an higher throughput through a maximization of the spectrum usage; or due to a malicious behavior, that is to block the legitimate traffic of the CR users.

Here we will present some mechanisms to detect PUE attacks, one of these is comparing the signal shape: if a secondary user detects a recognizable signal, it assumes that who is transmitting is a secondary user, on the contrary it determines that who is transmitting is a primary user. This simple method does not cover neither a malicious nor a selfish behavior, given that an attacker can easily exploit the spectrum sensing process. Indeed an attacker can find an unrecognizable signal shape and it would be considered a primary user. Further smarter approaches to discover a PUE attack are using matched filters or cyclostationary detectors. Hence, PUs can be recognized through intrinsic features, such as number of accesses to the MAC layer, or channel usage percentage, or signal strength level.

Unfortunately, even if these cross checks, it is still possible to emulate a primary user, simply emulating its cyclic spectral characteristics and signal shape. Substantially different from the previous strategies, the scheme presented in [44] bases is countermeasure approach on three steps: verification of signal characteristics, measurement of received signal energy level, and localization of the signal source (Figure 11.1). The same approach leads to a better identification of PUE attacks, even in case of multiple attackers.

11.3.2 Spectrum Sensing Data Falsification

This attack has been proposed in [43]. Assuming that a secondary users bases own decisions on the gathered sensed data, a Spectrum Sensing Data Falsification Attack aims to let the secondary user process the sensing data so as to lead it to improperly believe that a channel is occupied while it is idle or that a channel is idle while someone is transmitting

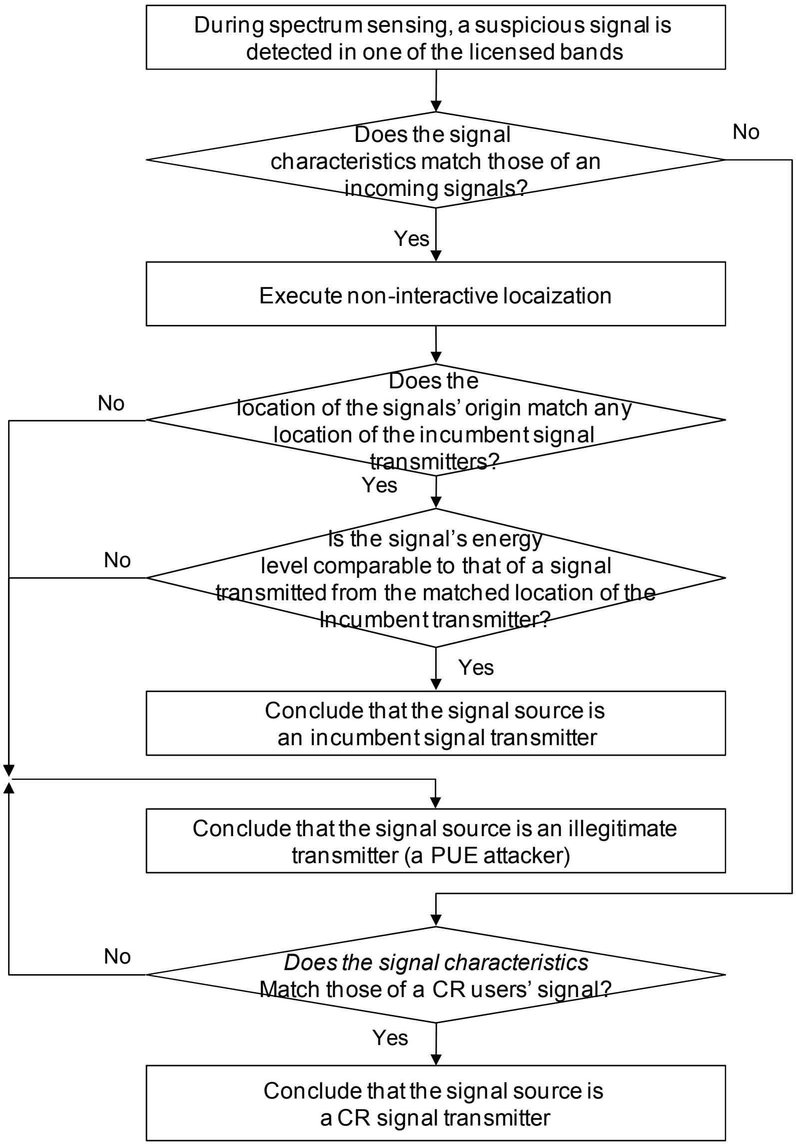


Figure 11.1 Flowchart of the transmitter verification scheme [44]

on the same channel.

The CR users implement different strategies to achieve the best decisions. One of those is the Decision Fusion: it is based on collecting, processing and comparing the sensed data to a threshold value which stands for the decision policy over the attack. An other strategy is the Bayesian Detection: it requires to know the a priori probability of a certain event. It makes easier an attacker detection, because of the statistics and events comparison. The last strategy is the Neyman-Pearson test: it relies on the maximum acceptable probability of false alarm and the maximum acceptable probability of miss detection taking into account an optimality problem. Although all those solutions, the best strategy would be the ideal one that accepts true data and drops tampered ones.

11.3.3 Denial-of-Service

A Denial-of-Service (DoS) attack is an attempt to prevent an user from the regular utilization of resources, therefore from regular operations and functions. We know that a secondary user uses the spectrum in an opportunistic way, transmitting and receiving both on licensed and unlicensed bands.

Considering a PHY layer DoS attack, it can be realized through jamming activity. For instance, overwhelming the channel with packets would compromise the integrity of the traffic, corrupting the same. In such case, the disadvantage of a cognitive radio network is that spreading the signal, one useful countermeasure against jamming attacks, is limited by the availability of the bands due to primary activities. On the contrary one robustness of the cognitive radio network is that it can operates in a large spectrum and avoid the jamming activity moving from a jamming affected channel to an idle one. Therefore, an effective jamming attack against a cognitive radio network has to jam many frequencies. How much is it feasible?

Considering a PHY layer DoS attack, one feature of the cognitive radio networks is that the transmission phase depends on the information gathered and processed during the sensing phase. While a traditional jamming attack affects only the receiver, a cognitive radio jamming attack can be effective either affecting the transmitter or the receiver. Further, an attacker should be closed to the target in order to be effective. Tampering the EM environment, a jammer can lead either the cognitive transmitter to do not transmit on a certain frequency or the cognitive receiver to be not able to receive the transmitted information. Indeed, a jammer can implement a DoS attack against a cognitive radio communication only positioning in a preferred position and tampering the EM environment tampering the sensing phase in transmission and corrupting the received packets in reception. DoS attack can be also performed through the PUE. Indeed, one principle of the cognitive radio network is to avoid interference with licensed users. Therefore an attacker can emulate a primary user transmission and do not let the cognitive users to use that channel.

DoS attack can represent a further threat for a cognitive radio network, if no authentication is done within the same network. Let we assume that a command control channel is used in order to manage the communications between the cognitive radio users. A non authenticated user, say an attacker, can easily send unicast or multicast control packets obtaining the network's block.

Summarizing [19], a DoS attack can be made achieving different states for the cognitive radio users that make impossible the radio cognitive activity: PUE's occupy all the channels, compromised sensed information does not allow the transmission, location information is not available, there is no connection to the receiver, all the policy is tampered through bad control packets. Security against DoS attacks can be achieved caring about the security issues such as confidentiality, authentication, authorization, non-repudiation. All these security policies are harder for cognitive radio networks than for traditional ones. Difficulties are particularly due to the opportunistic nature of the cognitive radio networks, which makes them very flexible in channel utilization, but let them to face the complexity that leads to vulnerabilities as: the sensing phase, that allows to jam the transmitter, the impossibility to have a granted band, that does not grant not even the availability of a command control channel, or the lack of security protocols for a cognitive radio scenario.

11.4 COUNTER MEASURES

There are several countermeasures that can be implemented in order to contrast the attacks described in the previous section. Many of them still present some weaknesses, because of the early research in the CR network field. Indeed, only once the CR technology will be widely used, attacks and respective countermeasures will be better addressed.

The simpler countermeasure against PHY layer attack is distinguishing between a natural RF signal (noise) and a man-made one (jamming activity). In order to improve such detection, sensors must be more reliable. Also cooperative sensing can actually improve such jamming activity detection. Countermeasures which only involve individual CR users are:

- Matching the CR users' signal shape with the ones received let the CR users know if that signal has been transmitted by a PU. Such an overly simplistic transmitter verification scheme, could let an attacker to easily hide as PU, only using a different signal shape from that of CR users.
- Implementing a comparison method it is possible to introduce a "common sense" to prevent from attacks. The same common sense policy must be protected from eavesdropping, on the contrary the attacker has the possibility to perform an attack respecting the same policies.
- Geolocating primary users could distinguish a primary user from a jammer, even if location is hard to perform.
- Updating learned behaviors such that a compromised behavior cannot last for a long time decreases the performance, but does not allow that a malicious attack propagates in time. On the other hand, countermeasures that involve Cognitive radio network, therefore single decisions are taken with information gathered and exchanged by many cognitive devices through a CCC, are described as follows:
- Taking a common decision instead of an individual one is an advantage, in case of the reliability of CR users is evaluated. Possible metrics are the proximity to the source or the sensors' quality.
- Using matched filters or cyclostationary detectors can improve the simple matching detection countermeasure. Indeed, PUs can be recognized through intrinsic features
- Any combination of these countermeasures.

In this chapter, security issues on CR networks have been presented. Security threats have been pointed out at each protocol layer and their causes have been identified among the new CR features. Together with possible counter measures, new attack techniques have been shown. As long as some of these main CR security issues will not be addressed, the standardization process will not develop in the right direction and either a market could never start up. CR could not accomplish what is promising and people could not take advantage of such a novel technology.

CHAPTER 12

STANDARDS FOR COGNITIVE RADIO NETWORKS

Till date, wide strides have been taken in the different aspects of CR research without special consideration towards inter-operability between different CR networks, repeatability of experiments, and guarantees for PU protection. Thus, the solutions proposed so far may have different assumptions on the level of interference tolerance to the PUs, the permissible transmission power in a given spectrum, etiquettes for spectrum sharing, among others that limit the deployment of CR networks in a practical setting. Moreover, different CR network operators may choose their own network parameters and raising questions of fairness and co-existence. These issues can be addressed by laying down standards for communication, such as the IEEE 802.15.4 for low-power bluetooth-based devices, IEEE 802.11 for different classes of WLANs, among others. In this chapter we shall focus on two different efforts. The first is the IEEE 802.22 for centralized networks, and the other is the IEEE P1900 working group, now called Standards Coordinating Committee 41 (SCC 41).

12.1 IEEE 802.22

IEEE 802.22 is a standard for Wireless Regional Area Network (WRAN) using the available bands in the TV frequency spectrum [56]. The development of the IEEE 802.22 WRAN standard is aimed at using cognitive radio techniques to allow the sharing of geographically unused spectrum allocated to the Television Broadcast Service, to bring broadband access to hard-to-reach, low population density areas, typical of rural environments, and is therefore timely and has the potential for a wide applicability worldwide. The IEEE

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury

978-0470-68852-6

802.22 working group on Wireless Regional Area Networks was formed in October 2004. Its project, formally called Standard for Wireless Regional Area Networks (WRAN) focuses on constructing a consistent, national fixed point-to-multipoint WRAN that will use UHF/VHF TV bands between $54 - 862\,\mathrm{MHz}$. Specific TV channels as well as the guard bands of these channels are planned to be used for communication in IEEE 802.22.

12.1.1 IEEE 802.22 Overview

A centralized approach for available spectrum discovery is proposed in this standard. Specifically, each base station (BS) in the network would be provided with a GPS receiver which would allow its position to be reported. This information would be sent back to centralized servers, which would respond with the information about available free TV channels and guard bands in the area of the BS. Other proposals would allow local spectrum sensing only, where the BS would decide by itself which channels are available for communication. A combination of these two approaches is also envisioned. Devices which would operate within this standard are of two types: fixed devices and Personal/Portable devices. The Fixed devices would have geo-location capability with embedded GPS device. Fixed devices also communicate with central database to identify other transmitters in the area operating in the aforementioned band. More interesting, the additional measures suggested by the FCC and IEEE to avoid interference relies on dynamic spectrum sensing and dynamic power control.

12.1.2 IEEE 802.22 Operation Details

Owing to the centralized nature of operation, the IEEE 802.22 uses BSs for spectrum access and sharing [59, 56] The BS manages its own cell and all associated consumer premise equipments (CPE) or CR users in this case. In the downstream (DS) direction, 802.22 MAC protocol uses Time Division Multiplexing, while in the upstream (US) direction, demand assigned TDMA is utilized.

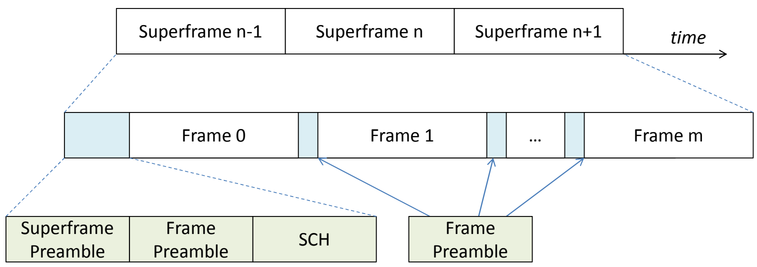


Figure 12.1 Superframe Structure in IEEE 802.22 [56].

The standard specifies time-slotted operation, with the frame hierarchy is shown in Figure 12.1. At the apex, a *superframe* is defined, each of which is composed of multiple MAC *frames* preceded by the frame preamble. At the start of each superframe, there is also a superframe control header (SCH) that is used to inform the CR users of the current available channels, different bandwidths supported, future spectrum access time, among

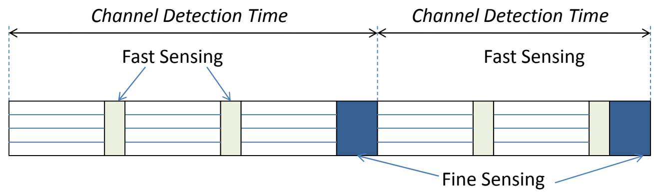


Figure 12.2 Two stage sensing (TSS) mechanism in IEEE 802.22 [56].

others. The MAC frame is formed by two parts in the frame structure, called as the DS subframe and US subframe. The DS subframe contains a single packet burst from a given CPE, while the US subframe has multiple packet bursts, each transmitted from different CPEs. The different fields in these two subframes are as follows: In the DS subframe, the preamble deals with synchronization and channel estimation, the frame control header (FCH) contains the size of the DS- and US-MAP fields together with channel descriptors, and the DS/US-MAPs give the scheduling information for user bursts. In the US subframe, the Urgent Coexistence Situation (UCS) notification informs of the incumbent licensees that have just been detected, while the other fields are used to derive the distance from the base station (ranging), and the individual bandwidth (BW) requests.

12.1.3 Features of the IEEE 802.22 standard

The key features of the IEEE 802.22 standard are (i) extensive support for spectrum sensing, (ii) spectrum recovery, and (iii) coexistence of the different users [56].

12.1.3.1 Spectrum Sensing Support The IEEE 802.22 protocol has a two-stage sensing (TSS) mechanism as shown in Figure 12.2. The transmission durations are shown by the rectangular packets shaded with horizontal lines, followed by the *fine* or the *fast* sensing times.

To reliably attribute the source of the received power to the PUs, the standard enforces quiet periods throughout the CR network called as *channel detection time*. The TSS consists of two stages which have different durations and goals:

- *Fast Sensing:* This is done at the rate of 1 ms/channel, and the sensing results are used to decide is a subsequent fine sensing stage is needed. The sensing is completed quickly though the accuracy is low.

- *Fine Sensing:* Fine sensing is performed on-demand, which allows CR networks to meet the strict quality of service (QoS) requirements by decreasing the rate of false alarms. The duration for this is much larger than the fast sensing, and gives a tradeoff between improving the sensing accuracy at the cost of transmission time.

12.1.3.2 Spectrum Recovery When a licensed user is detected, the incumbent detection recovery protocol (IDRP) is used, that enables the network to restore normal operation with minimal performance degradation. In IRDP, backup channels are used that allow to restore communication in case a channel needs to be vacated after PU appearance. These backup channels are kept in a priority list and are used whenever a CPE looks for a BS during the recovery procedure. This makes the protocol very efficient as both CPE and

BS know in advance on which channels to recover the transmission when a PU is detected in their channel of operation.

12.1.3.3 Coexistence with Users The intra-network coexistence for CR network is achieved by the coexistence beacon protocol (CBP). The CR network may be composed of multiple base stations, which must regulate their transmission parameters based on the actions taken by each other. This requires communication between the cells, which is undertaken by the CBP beacons [56]. These beacons carry information about the cells and the DS/US bandwidth allocations for the users. CBP packets are allowed to be transmitted during specially marked periods called as a self-coexistence window. During this window period, contention-based scheme is used to access the spectrum band. As this protocol is also used for inter-base station communication, the latter have a higher priority than CPEs for spectrum access in this window period. This scheme allows the base stations to exchange information in priority over the general data traffic of the CR users.

The main drawback of this protocol is that the control header exchange is extensive, which may result in lower data throughout or reduced channel utilization. Moreover, the time synchronization is difficult to maintain between the different CR base stations, we well as CR users in a given cell.

12.1.4 Working groups within the 802.22

The IEEE 802.22 working group has been in operation for over five years at the time of this writing. Its current efforts are focused on these three main issues:

- **IEEE P802.22:** the main standard specification and policy, currently being addressed by IEEE 802.22 working group (WG).
- **IEEE P802.22.1:** a standard being developed to enhance harmful interference protection for low power licensed devices operating in TV Broadcast Bands currently being studied by the IEEE 802.22 Task Group 1 (TG1).
- **IEEE P802.22.2:** a recommended practice for the installation and deployment of IEEE 802.22 Systems, developed by the IEEE 802.22 Task Group 2 (TG2).

12.2 IEEE P1900 - STANDARDS COORDINATING COMMITTEE 41 (SCC 41)

The IEEE P1900 Standards Committee [192] addresses a wider class of CR issues, by systematically classifying the terminology, spectrum access types, recofigurable software, and network architectures. It provides a way in which different operators following their own radio access technologies (RATs) can coexist in the same frequency space. The IEEE Standards Board reorganized the IEEE 1900 effort as the Standards Coordinating Committee 41 (SCC 41), Dynamic Spectrum Access Networks (DySPAN). The IEEE SCC 41 is divided in the following groups:

- **IEEE 1900.1 - Working Group on Terminology and Concepts for Next Generation Radio Systems and Spectrum Management:** aimed to provide technically precise definitions and explanations of key concepts in the fields of spectrum management, cognitive radio and related technologies from different perspectives. The standard also seeks to describe how these technologies can be used in a wide variety of communication service environments to achieve new capabilities while at the same time providing mechanisms supportive of new spectrum management paradigms and spectrum access.

- **IEEE 1900.2 - Working Group on Recommended Practice for Interference and Coexistence Analysis:** provides a model that facilitates the analysis of coexistence/ interference between CR users and PUs operating in the same frequency band or between different frequency bands. The model also provides guidance for estimating the co-channel, adjacent channel and out ofband interference under a variety of scenarios. It also analyses how factors such as directional antennas, power control, and licensed channel avoidance strategies affect the aggregate interference.
- **IEEE 1900.3 - Working Group on Recommended Practice for Conformance Evaluation of Software Defined Radio (SDR) Software Modules:** provides technical guidelines for analyzing Software Defined Radio(SDR) software modules to ensure compliance with regulatory and operational requirements. Compliance with requirements for spectrum use is tested using formal mathematical concepts and methods.
- **IEEE 1900.4 - Working Group on Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks:** aimed to increase the overall system utilization of reconfigurable terminals while increasing the perceived QoS. It does this by defining the overall system architecture in such a way so as to split functionality between terminals and the network and also the information exchange between coordinating entities.
- **IEEE 1900.5 - Working Group on Policy Language and Policy Architectures for Managing Cognitive Radio for Dynamic Spectrum Access Applications:** defines a set of policy languages, and their relation to policy architectures, for managing the features of cognitive radios for dynamic spectrum access applications.
- **IEEE 1900.6 - Working Group on Spectrum Sensing Interfaces and Data Structures for Dynamic Spectrum Access and other Advanced Radio Communication Systems:** aimed to develop a standard that will define the interfaces and data structures required for exchange of sensing related information. The resulting standard will provide a formal definition of data structures and interfaces for exchange of sensing related information.
- **IEEE 1900.A - Study Group on Dependability and Evaluation of Regulatory Compliance for Radio Systems with Dynamic Spectrum Access:** aimed to specify techniques for testing and analysis to be used during regulatory compliance evaluation of dynamic spectrum access radio systems. The methods include recommended radio system design features that simplify the assessment challenge in addition to test and analysis procedures.

We would like to draw the reader's attention to the efforts pertaining to IEEE 1900.4 group. This group allows CR networks composed of reconfigurable BSs to share the spectrum. By reconfiguration, we mean that the BS could choose any RAT from say, GSM, WCDMA, and WiMAX either by having dedicated hardware installed, or through software modification. They may hence operate anywhere in a wide frequency space between 400 MHz and 6 GHz [81].

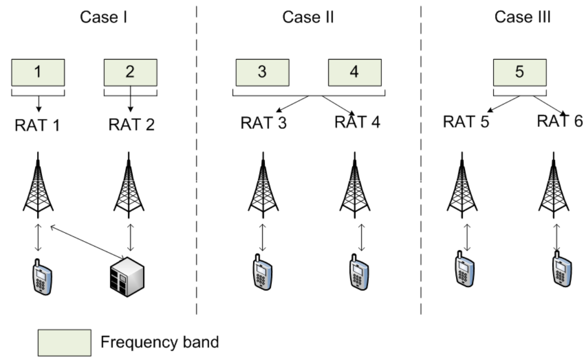


Figure 12.3 The operation of the P1900.4 [81].

In Fig. 12.3, we describe the following three use-cases that are considered in the P1900.4 standard. The frequency bands are numbered for $1-6$, each occupying a distinct range on the frequency scale.

- **Case I - Distributed radio resource usage optimization:** Here, frequency bands assigned to different radio access networks (RANs) are fixed, and each RAN may further use its own RAT. Both legacy terminals that interface with only a single RAT, as well as re-configurable terminals that link to more than one (such multiple connections to two RATs are shown in Fig. 12.3).
- **Case II - Dynamic spectrum assignment:** In this case, to improve radio resource usage, the operators may freely use different portions of the frequency band opportunistically. As an example, a single operator with multiple RANs may select spectrum for each of its RANs flexibly.
- **Case III - Dynamic spectrum assignment:** Here, several RANs using same or different RATs can share the same frequency band. Thus, P1900.4 should provide safeguards during operation within the unlicensed spectrum, or in the use of TV whitespaces.

Despite the increasing efforts on standardization of future CR networks, this ambitious effort is still in a nascent stage. However, we believe that this standard shall play a crucial role in shaping the CR networks of tomorrow. On December 2010, the SCC 41 was transferred under the IEEE Communication Society.

The success of a standard relies squarely on the interaction between the standards bodies for laying down the rules, the academia of leading future research under those set of rules, and participation by industry so that real and implementable prototypes that operate under the standards guidelines can be made available. Thus, a good standard may be critical to the success of CR technology in the future.

REFERENCES

1. 3GPP. Physical layer-general description. *3GPP TS 25.201*.
2. C. da Silva A. Malady. Clustering methods for distributed spectrum sensing in cognitive radio systems. In *Proc. of IEEE MILCOM 2008*, 2008.
3. I. F. Akyildiz, W. Y. Lee, and K. Chowdhury. CRAHNs: Cognitive radio ad hoc networks. *Ad Hoc Network Journal (Elsevier)*, 7(5):810–836, July 2009.
4. I. F. Akyildiz, W. Y. Lee, and K. Chowdhury. Spectrum management in cognitive radio ad hoc networks. *IEEE Network*, 23(4):6–12, July 2009.
5. I. F. Akyildiz, W. Y. Lee, M. Vuran, and S. Mohanty. A Survey on Spectrum Management in Cognitive Radio Networks. *IEEE Communication Magazine*, 46(4):40–48, 2008.
6. 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 Journal (Elsevier)*, 50:2127–2159, September 2006.
7. I. F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. *Computer Networks Journal (Elsevier)*, 47(4):445–487, November 2005.
8. S. Appadwedula, V. Veeravalli, and D. Jones. Robust and locally-optimum decentralized detection with censoring sensors. In *Proc. of Information Fusion*, volume 1, pages 56–63, 2002.
9. L. Badia, M. Lindstrom, J. Zander, and M. Zorzi. Demand and pricing effects on the radio resource allocation of multimedia communication systems. In *Proc. of IEEE Globecom 2003*, December 2003.
10. J. Bazerque and G. Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, March 2010.
11. L. Berlemann, S. Mangold, and B.H.Walke. Policy-based reasoning for spectrum sharing in cognitive radio networks. In *Proc. of IEEE DySPAN 2005*, November 2005.

Cognitive Radio Networks, First edition. By Ian F. Akyildiz, Won-Yeol Lee, and Kaushik R. Chowdhury
978-0470-68852-6

12. D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, second edition edition, 2001.

13. R. Bhatti, K. Chugg, and J. Draper. Standard cell based pseudo-random clock generator for statistical random sampling of digital signals. In *Proc. of 50th Midwest Symposium on Circuits and Systems (MWSCAS) 2007*, 2007.

14. R. Blum, S. Kassam, and H. Poor. Distributed detection with multiple sensors I. Advanced topics. *Proceedings of the IEEE*, 85(1):64–79, January 1997.

15. S. Borst and P. Whiting. Dynamic rate control algorithms for HDR throughput maximization. In *Proc. of INFOCOM 2001*, pages 976–985, April 2001.

16. S. Boyd and L. Vandenberghe. *Convex Optimization*. Camb, Cambridge, MA, 2004.

17. L. Brakmo and L. Peterson. TCP Vegas: end to end congestion avoidance on a global Internet. *IEEE Journal on Sel. Areas of Comm.*, 13(8):1465–1480, 1995.

18. V. Brik, E. Rozner, S. Banarjee, and P. Bahl. DSAP: A protocol for coordinated spectrum access. In *Proc. of IEEE DySPAN 2005*, pages 611–614, November 2005.

19. T. X. Brown and A. Sethi. Potential cognitive radio denial-of-service vulnerabilities and protection countermeasures: a multi-dimensional analysis and assessment. *MONET*, 13(5):516–532, 2008.

20. Brown2005. An analysis of unlicensed device operation in licensed broadcast service bands. In *Proc. of IEEE DySPAN 2005*, pages 11–29, November 2005.

21. M. M. Buddhikot, P. Kolody, S. Miller, K. Ryan, and J. Evans. DIMSUMNet: New directions in wireless networking using coordinated dynamic spectrum access. In *Proc. of IEEE WoWMoM 2005*, pages 78–85, June 2005.

22. C. Buragohain, S. Suri, C. Toth, and Y. Zhou. Improved throughput bounds for interference-aware routing in wireless networks. *UCSB Technical Report*, (2006-13):2006, 2006.

23. and P. Willett C. Rago and Y. Bar-Shalom. Censoring sensors: a lowcommunication-rate scheme for distributed detection. *IEEE Transactions on Aerospace and Electronic Systems*, 32(2):2, February 1996‘.

24. and W. Zhang C. Sun and K. Ben. Cluster-based cooperative spectrum sensing in cognitive radio systems. In *Proc. of IEEE ICC 2007*, pages 2511–2515, 2007.

25. M. Hasegawa C. W. Pyo. Minimum weight routing based on a common link control radio for cognitive wireless ad hoc networks. In *Proceedings of the international conference on Wireless communications and mobile computing, IWCMC'07*, pages 399–404, 2007.

26. C.-C. Shen C. Xin, B. Xie. A novel layered graph model for topology formation and routing in dynamic spectrum access networks. In *Proceedings of First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN'05*, pages 308–317, 2005.

27. C.-C. Shen C. Xin, L. Ma. A path-centric channel assignment framework for cognitive radio wireless networks. *Mobile Networks and Applications*, 13(5):463–476, 2008.

28. C.-C. Shen C. Xin, L. Ma. Spectrum aware routing for multi-hop cognitive radio networks with a single transceiver. In *Proceedings of 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom'08*, pages 1–6, 2008.

29. D. Cabric and R. W. Brodersen. Physical layer design issues unique to cognitive radio systems. In *Proc. of IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 2, pages 759–763, September 2005.

30. D. Cabric, S. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz. A cognitive radio approach for usage of virtual unlicensed spectrum. In *Proc. of 14th IST Mobile and Wireless Communications Summit*, June 2005.

31. D. Cabric, S. M. Mishra, and R. W. Brodersen. Implementation issues in spectrum sensing for cognitive radios. *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 36:772–776, Nov 2004.

32. D. Cabric, A. Tkachenko, and R. W. Brodersen. Spectrum sensing measurements of pilot, energy, and collaborative detection. In *Proc.of IEEE Military Communications Conference (MILCOM)*, October 2006.

33. B. Canberk, I. F. Akyildiz, and S. Oktug. Primary user activity modeling using first-difference filter clustering and correlation in cognitive radio networks. *IEEE Transactions on Networking*, 2011. to appear.

34. E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.

35. E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

36. L. Cao and H. Zheng. Distributed spectrum allocation via local bargaining. In *Proc. of IEEE Sensor and Ad Hoc Communications and Networks (SECON)*, pages 475–486, September 2005.

37. L. Cao and H. Zheng. Distributed rule-regulated spectrum sharing. *IEEE JSAC*, 26(1):130–145, January 2008.

38. A. Capone, L. Fratta, and F. Martignon. Bandwidth Estimation Schemes for TCP over Wireless Networks. *IEEE Trans. on Mobile Computing*, 3(2):129–143, 2004.

39. M. Cesana, F. Cuomo, and E. Ekici. Routing in Cognitive Radio Networks: Challenges and Solutions. *Ad Hoc Networks Journal (Elsevier), to appear*, 2010.

40. M. R. Chari, F. Ling, A. Mantravadi, R. Krishnamoorthi, R. Vijayan, G. K. Walker, and R. Chandhok. FLO physical layer: an overview. *IEEE Trans. Broadcasting*, 53(1):145–159, March 2007.

41. D. Chen, S. Yin, Q. Zhang, M. Liu, and S. Li. Mining Spectrum Usage Data: a Large-scale Spectrum Measurement Study. In *Proc. of ACM MOBICOM*, September 2009.

42. R. Chen, J.-M. Park, and K. Bian. Robust distributed spectrum sensing in cognitive radio networks. In *Proc. of IEEE Infocom 2008*, 2008.

43. R. Chen, J.-M. Park, Y. Hou, and J. Reed. Toward secure distributed spectrum sensing in cognitive radio networks,. *IEEE Communications Magazine*, 46(4):50–55, April 2008.

44. R. Chen, J.-M. Park, and J. Reed. Defense against primary user emulation attacks in cognitive radio networks,. *IEEE Journal on Selected Areas in Communications*, 26(1):25–37, January 2008.

45. T. Chen, H. Zhang, G.M. Maggio, and I. Chlamtac. CogMesh: A cluster-based cognitive radio network. In *Proc. of IEEE DySPAN 2007*, pages 168–178, April 2007.

46. Y. Chen, Q. Zhao, and A. Swami. Joint design and separation principle for opportunistic spectrum access. In *Proc. IEEE Asilomar Conference on Signals, Systems and Computers 2006*, October 2006.

47. Y. Chen, Q. Zhao, and A. Swami. Distributed spectrum sensing and access in cognitive radio networks with energy constraint. *IEEE Transactions on Signal Processing*, 57(2):783–797, February 2009.

48. G. Cheng, W. Liu, Y. Li, and W. Cheng. Joint On-Demand Routing and Spectrum Assignment in Cognitive Radio Networks. In *Proceedings of IEEE International Conference on Communications, ICC'07*, pages 6499–6503, 2007.

49. Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin Company, 1971.

50. K. Chowdhury and I. Akyildiz. Cognitive wireless mesh networks with dynamic spectrum access. *IEEE Journal on Selected Areas in Communications*, 26(1):168–181, 2008.

51. K. Chowdhury and I. Akyildiz. CRP: A Routing Protocol for Cognitive Radio Ad Hoc Networks. *IEEE Journal on Selected Areas in Communications, to appear*, 2011.

52. K. Chowdhury and M. Felice. Search: A routing protocol for mobile cognitive radio ad-hoc networks. *Computer Communications*, 32(18):1983–1997, 2009.

53. K. R. Chowdhury, M. D. Felice, and I. F. Akyildiz. TP-CRAHN: A Transport Protocol for Cognitive Radio Ad-hoc Networks. pages 2482–2490, Apr 2009.

54. T. Clancy. Formalizing the interference temperature model. *Wiley Journal on Wireless Communications and Mobile Computing*, 7(9):1077–1086, November 2007.

55. E. Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, 1971.

56. IEEE 802 LAN/MAN Standards Committee. IEEE 802.22 Standard. [Online]. Available: http://www.ieee802.org/22/.

57. C. Cordeiro and K. Challapali. C-MAC: A Cognitive MAC Protocol for MultiChannel Wireless Networks. *Proceedings of IEEE DySPAN*, 36:147–157, Apr 2007.

58. C. Cordeiro, K. Challapali, and D. Birru. IEEE 802.22: An introduction to the first wireless standard based on cognitive radios. *Journal of communications*, 1(1), April 2006.

59. C. Cordeiro, K. Challapali, D. Birru, and S. Shankar N. IEEE 802.22: the first worldwide wireless standard based on cognitive radios. pages 328–337, 2005.

60. C. Cordeiro, K. Challapali, and M. Ghosh. Cognitive PHY and MAC layers for dynamic spectrum access and sharing of TV bands. In *Proc. of ACM TAPAS*, August 2006.

61. C. Cordeiro, K. Challapali, and M. Ghosh. Cognitive PHY and MAC layers for dynamic spectrum access and sharing of TV bands. In *TAPAS '06: Proceedings of the first international workshop on Technology and policy for accessing spectrum*, Aug 2006.

62. P. Cordier, P. Houze, S.B. Jemaa, O. Simon, D. Bourse, D. Grandblaise, K. Moessner, J. Luo, Clemens Kloeck, N. Olaziregi K Tsagkaris, R. Agusti, Z. Boufidis, E. Buracchini, P. Goria, and A. Trogolo. E2R cognitive pilot channel concept. In *Proc. of 15th IST Mobile and Wireless Communications Summit 2006*, June 2006.

63. C. Cormio and K. R. Chowdhury. A survey on MAC protocols for cognitive radio networks. *Ad Hoc Networks*, 7(7):1315–1329, 2009.

64. D. Cox. *Renewal Theory*. Spottiswoode Ballantyne and Co. Ltd, 1962.

65. DARPA XG WG. The XG architectural framework v1.0. 2003.

66. DARPA XG WG. The XG vision RFC V1.0. 2003.

67. F. F. Digham, M. S. Alouini, and M. K. Simon. On the energy detection of unknown signals over fading channels. In *Proc. IEEE ICC 2003*, pages 3575–3579, May 2003.

68. L. Ding, T. Melodia, S. Batalama, J. Matyjas, and M. Medley. Cross-layer Routing and Dynamic Spectrum Allocation in Cognitive Radio Ad Hoc Networks. *IEEE Transactions on Vehicular Technology, to appear*, 2010.

69. D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

70. Raul Etkin, Abhay Parekh, , and David Tse. Spectrum sharing for unlicensed bands. *IEEE JSAC*, 25(3):517–526, April 2007.

71. FCC. Fcc-07-78a1. page 2007, May.

72. FCC. Et docket no 03-222. *Notice of Proposed Rule Making and Order*, December 2003.

73. FCC. Et docket no 03-289. *Notice of Inquiry and Notice of Proposed Rulemaking*, November 2003.

74. FCC. Part 15 - radio frequency devices. pages 1–155, 2006.

75. FCC. Unlicensed operation in the tv bands, motorola ex-parte filing. *FCC Docket ET 04-186*, 2007 2007.

76. FCC. Docket 08-260, November 2008. Unlicensed Operation in the TV Broadcast Bands.

77. FCC. Second memorandum opinion and order. *ET Docket No. 10-174*, September 2010.

78. A. Fehske, J. D. Gaeddert, and J. H. Reed. A new approach to signal classification using spectral correlation and neural networks. In *Proc. IEEE DySPAN 2005*, pages 144–150, November 2005.

79. M. Di Felice, K. R. Chowdhury, and L. Bononi. Modeling and Performance Evaluation of Transmission Control Protocol over Cognitive Radio Ad Hoc Networks. In *ACM MSWIM*, pages 4–12, 2009.

80. M. Di Felice, K. R. Chowdhury, W. Kim, A. J. Kassler, and L. Bononi. End-to-end protocols for Cognitive Radio Ad Hoc Networks: An Evaluation Study. In *Performance Evaliuation Journal, to appear*, 2010.

81. S. Filin, H. Harada S, H. Murakami, K. Ishizu, and G. Miyamoto. IEEE draft standards P1900.4.1 and P1900.4a for heterogeneous type and spectrum sharing type cognitive radio systems. pages 1–6, Sept. 2009.

82. S. Floyd and T. Henderson. The NewReno Modification to TCP? Fast Recovery Algorithm, Apr 1999. Internet Engeneering Task Force, Request for Comments (Experimental) 2582.

83. G. Ko A. A. Franklin, S You, J Pak, M. Song;, and C. Kim. Channel management in ieee 802.22 wran systems. *IEEE Communication Magazine*, 48(9):88–94, September 2010.

84. J. Friedman. A non-cooperative equilibrium for super games. *Review of Economic Studies*, 38(113):1–12, 1971.

85. J. Friedman. *Oligopoly and the theory of games*. Amsterdam North-Holland, 1977.

86. D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

87. G. Ganesan and Y. G. Li. Cooperative spectrum sensing in cognitive radio -part I: two user network. *IEEE Transactions on Wireless Communications*, 6(6):22042213., June 2007.

88. G. Ganesan and Y. G. Li. Cooperative spectrum sensing in cognitive radio part II: multiuser networks. *IEEE Transactions on Wireless Communications*, 6(6):22042213., June 2007.

89. W. A. Gardner. Structural characterization of locally optimum detectors in terms of locally optimum estimators and correlators. *IEEE Transactions on Information Theory*, 28:924–932, November 1982.

90. W. A. Gardner. *Introduction to Random Processes with Applications to Signals and Systems*. MacMillan, 1985.

91. W. A. Gardner. Signal interception: a unifying theoretical framework for feature detection. *IEEE Transactions on Communications*, 36(8):897–906, August 1988.

92. W. A. Gardner. Exploitation of spectral redundancy in cyclostationary signals. *IEEE Signal Processing Magazine*, 8(2):14–36, August 1991.

93. W. A. Gardner. *Cyclostationarity in Communications and Signal Processing*. IEEE Press, 1994.

94. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1990.

95. S. Geirhofer, L. Tong, and B. M. Sadler. Dynamic spectrum access in the time domain: modeling and exploiting white space. *IEEE Communications Magazine*, 45(5):66–72, May 2007.

96. S. Geirhofer, L. Tong, and B. M. Sadler. Cognitive Medium Access: Constraining Interference Based on Experimental Models. *IEEE Journal on Selected Areas in Communications*, 26(1):95–105, Jan 2008.

97. A. Ghasemi and E. S. Sousa. Optimization of spectrum sensing for opportunistic spectrum access in cognitive radio networks. pages 1022–1026, 2007.

98. A. Ghasemi and E.S. Sousa. Collaborative spectrum sensing for opportunistic access in fading environment. In *Proc. of IEEE DySPAN 2005*, pages 131–136, November 2005.

99. A. Ghasemi and E.S. Sousa. Optimization of spectrum sensing for opportunistic spectrum access in cognitive radio networks. In *Proc. of IEEE Consumer Communications and Networking Conference (CCNC) 2007*, pages 1022–1026, january 2007.

100. S. Gjerstad and J. Dickhaut. Price formation in double auctions. *Games and Economic Behavior*, 22:1–29, 1998.

101. L.P. Goh, Z. Lei, and F. Chin. DVB detector for cognitive radio. In *Proc. of the IEEE ICC 2007*, pages 6460–6465, June 2007.

102. X. Gong, W. Yuan, W. Liu, W. Cheng, and S. Wang. A cooperative relay scheme for secondary communication in cognitive radio networks. In *Proc. IEEE Globecom 2008*, November 2008.

103. T. Groves. Incentive in terms. *Econometrica*, 41:617–631, 1973.

104. Q. Guan, F. R. Yu, S. Jiang, and G. Wei. Prediction-Based Topology Control and Routing in Cognitive Radio Mobile Ad Hoc Networks. *IEEE Transactions on Vehicular Technology*, 59(9):4443 –4452, 2010.

105. M. Gudmundson. Correlation model for shadow fading in mobile radio systems. *Electronics Letters*, 27(23):2145–2146, 1991.

106. C. Guo, T. Peng, S. Xu, and H.Wang andW.Wang. Cooperative spectrum sensing with cluster-based architecture in cognitive radio networks. In *Proc of IEEE VTC Spring 2009*, 2009.

107. D. Gurney, G. Buchwald, L. Ecklund, S. Kuffner, and J. Grosspietsch. Geo-location database techniques for incumbent protection in the TV white space. In *Proc. of IEEE DySPAN 2008*, October 2008.

108. N. Malouch H. Khalife and S. Fdida. Multihop cognitive radio networks: To route or not to route. *IEEE Network Magazine*, 23(4):20–25, 2009.

109. B. Hamdaoui and K. G. Shin. Os-mac: An efficient mac protocol for spectrum-agile wireless networks. *IEEE Transactions on Mobile Computing*, 7(8):915–930, Aug 2008.

110. N. Han, S. Shon, J.H. Chung, and J.M. Kim. Spectral correlation based signal detection method for spectrum sensing in IEEE 802.22 WRAN systems. In *Proc. of ICACT 2006*, pages 1765–1770, February 2006.

111. S. Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23(2):201–220, February 2005.

112. M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda. Performance anomaly of 802.11b. In *Proc. of INFOCOM 2003*, March 2003.

113. G. Holland and N. H. Vaidya. Analysis of tcp performance over mobile ad hoc networks. In *Proc. of ACM MOBICOM*, pages 219–230, August 2009.

114. W. D. Horne. Adaptive spectrum access: Using the full spectrum space. In *Proc. of Telecommunications Policy Research Conference (TPRC)*, September 2003.

115. J. Huang, R. A. Berry, and M. L. Honig. Spectrum sharing with distributed interference compensation. In *Proc. of IEEE DySPAN 2005*, pages 88–93, November 2005.

116. J. Huang, R. A. Berry, and M. L. Honig. Spectrum sharing with distributed interference compensation. In *Proc. of IEEE DySPAN 2005*, pages 88–93, November 2005.

117. J. Huang, R. A. Berry, and M. L. Honig. Auction-based spectrum sharing. *Mobile Networks and Applications*, 11, 3 2006.

118. X.L. Huang and B. Bensaou. On max-min fairness and scheduling in wireless ad-hoc networks: Analytical framework and implementation. In *Proc. of MobiHoc 2001*, October 2001.

119. L. Hurwicz, R. Radner, and S. Reiter. A stochastic decentralized resource allocation process: Part I. *Econometrica*, 43:363–393, 1975.

120. IEEE WRAN WG. Standard for wireless regional area networks (WRAN) - specific requirements - part 22: Cognitive wireless RAN medium access control (MAC) and physical layer (PHY) specifications: Policies and procedures for operation in the TV bands. 2009.

121. Y. Altunbasak I.F. Akyildiz and, F. Fekri, and R. Sivakumar. AdaptNet: Adaptive protocol suite for next generation wireless internet. *IEEE Communications Magazine*, 42(3):128–138, 2004.

122. O. Ileri, D. Samardzija, and N. B. Mandayam. Demand responsive pricing and competitive spectrum allocation via a spectrum server. In *Proc. of IEEE DySPAN 2005*, November 2005.

123. O. Ileri, D. Samardzija, and N. B. Mandayam. Dynamic property rights spectrum access: Flexible ownership based spectrum management. In *Proc. of IEEE DySPAN 2007*, 2007.

124. J. Mitola III. Cognitive radio for flexible mobile multimedia communication. In *Proc. IEEE Int'l Workshop on Mobile Multimedia Communications (MoMuC) 1999*, pages 3–10, November 1999.

125. J. Mitola III. *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio*. PhD thesis, KTH Royal Institute of Technology, 2000.

126. K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu. Impact of interference on multi-hop wireless network performance. In *Proc. of MobiCom 2003*, Semtember 2003.

127. P. Jallon. A spreading signals detection algorithm based on the 2nd order statistics in semiblind contexts. In *Proc. of CROWNCOM 2008*, May 2008.

128. Z. Ji and K. J. R. Liu. Dynamic spectrum sharing: A dynamic spectrum sharing: a game theoretical overview. *IEEE Communications Magazine*, 45:88–94, 2007.

129. Z. Ji, W. Yu, and K. J. R. Liu. An optimal dynamic pricing framework for autonomous mobile ad hoc networks. In *Proc. of IEEE INFOCOM 2006*, April 2006.

130. J. Jia, J. Zhang, and Q. Zhang. Cooperative relay for cognitive radio networks. In *Proc. IEEE INFOCOM 2009*, pages 2304–2312, 2009.

131. J. Jia, Q. Zhang, and X. Shen. HC-MAC: a hardware-constrained cognitive mac for efficient spectrum management. *IEEE Journal on Selected Areas in Communications*, 26(1):106–117, January 2008.

132. J. Jia, Q. Zhang, and X. Shen. Hc-mac: A hardware-constrained cognitive mac for efficient spectrum management. *IEEE Journal on Selected Areas in Communications*, 26(1):106–117, Jan 2008.

133. X. Jing and D. Raychaudhuri. Spectrum co-existence of IEEE 802.11b and 802.16a networks using CSCC etiquette protocol. In *Proc. of IEEE DySPAN 2005*, pages 243–250, November 2005.

134. K. Johansson, A. Furuskar, P. Karlsson, and J. Zander. Relation between cost structure and base station characteristics in cellular systems. In *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2005*, September 2005.

135. R. Johari and J.N. Tsitsiklis. Efficiency loss in a network resource allocation game. *Mathematics of Operations Research*, 29(3):407–435, 2004.

136. F. K. Jondral. Software-defined radio- basic and evolution to cognitive radio. *EURASIP Journal on Wireless Communication and Networking 2005*, 2005.

137. J.Wei and X. Zhang. Energy-efficient distributed spectrum sensing for wireless cognitive radio networks. In *Proc of INFOCOM IEEE Conference on Computer Communications Workshops*, 2010.

138. B. Karp and H. T. Kung. GPSR: Greedy Perimeter Stateless Routing for Wireless Networks. In *Proc. of ACM MobiCom*, 2000.

139. S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc, 1993.

140. H. Kim and J. M. Peha. Detecting selfish behavior in a cooperative commons. In *Proc. of IEEE DySPAN 2008*, Chicago, IL, USA,, October 2008.

141. H. Kim and K. G. Shin. Efficient discovery of spectrum opportunities with mac-layer sensing in cognitive radio networks. *IEEE Transactions on Mobile Computing*, 7:533–545, May 2008.

142. H. Kim and K. G. Shin. Fast discovery of spectrum opportunities in cognitive radio networks. In *Proc. of IEEE DySPAN 2008*, October 2008.

143. H. Kim and K.G. Shin. Efficient discovery of spectrum opportunities with mac-layer sensing in cognitive radio networks. *IEEE Transactions on Mobile Computing*, 7(5):533–545, May 2008.

144. S.-J. Kim, E. Dall'Anese, and G. B. Giannakis. Sparsity-aware cooperative cognitive radio sensing using channel gain maps. In *Proc. of Asilomar Conference on Signals, Systems and Computers 2009*, pages 518–522, 2009.

145. S.-J. Kim, E. Dall'Anese, and G. B. Giannakis. Cooperative spectrum sensing for cognitive radios using Kriged Kalman filtering. *IEEE Journal of Selected Topics in Signal Processing*, 2011. To appear.

146. Y. R. Kondareddy and P. Agrawal. Synchronized mac protocol for multi-hop cognitive radio networks. In *Proceedings of IEEE International Conference on Communications (ICC)*, pages 3198–3202, May 2008.

147. V. Krishna. *Auction Theory*. Academic Press, 2002.

148. P. Kyasanur, X. Yang, and N.H. Vaidya. Mesh networking protocols to exploit physical layer capabilities. In *Proc. of IEEE Workshop on Wireless Mesh Networks (WiMesh),*, September 2005.

149. W. Y. Lee and I. F. Akyildiz. Joint spectrum and power allocation for inter-cell spectrum sharing in cognitive radio networks. In *IEEE DySPAN 2008, Chicago, IL, USA*, October 2008.

150. W. Y. Lee and I. F. Akyildiz. Optimal spectrum sensing framework for cognitive radio networks. *IEEE Transactions on Wireless Communications*, 7(10):3845–3857, October 2008.

151. W. Y. Lee and I. F. Akyildiz. Spectrum-aware mobility management in cognitive radio cellular networks. *submitted for journal publication*, 2010.

152. W. Y. Lee and I. F. Akyildiz. A spectrum decision framework for cognitive radio networks. *to appear in IEEE Transactions on Mobile Computing*, 2010.

153. D. Lehmann, L. O'callaghan, and Y. Shoham. Truth revelation in approximately efficient combinatorial auctions. *Journal of the ACM*, 49(5):577–602, September 2002.

154. K. B. Letaief and W. Zhang. Cooperative communications for cognitive radio networks. *Proceedings of the IEEE*, 97(5):878–893, May 2009.

155. Z. Li, F. Yu, and M. Huang. A cooperative spectrum sensing consensus scheme in cognitive radios. In *Proc. of IEEE Infocom 2009*, page 25462550, 2009.

156. Y. C. Liang, Y. Zeng, E.C.Y. Peh, and A. T. Hoang. Sensing-throughput tradeoff for cognitive radio networks. *IEEE Transactions on Wireless Communications*, 7(4):1326–1337, April 2008.

157. S. Y. Lien, C. C. Tseng, and K. C. Chen. Carrier Sensing based multiple access protocols for cognitive radio networks. In *Proceedings of IEEE International Conference on Communications (ICC)*, pages 3208–3214, May 2008.

158. M. Lindstrom. *Demand responsive resource management for cellular networks*. PhD thesis, KTH Royal Institute of Technology, 2005.

159. J. Liu and S. Singh. ATCP: TCP for mobile ad hoc networks. *IEEE Journal on Sel. Areas of Comm.*, 19(7):1300–1315, 2001.

160. X. Liu and Z. Ding. ESCAPE: a channel evacuation protocol for spectrum-agile networks. In *Proc. of IEEE DySPAN 2007*, April 2007.

161. B. F. Lo and I. F. Akyildiz. Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks. In *Proc. of IEEE PIMRC 2010*, 2010.

162. I. F. Akyildiz.and B. F. Lo and R. Balakrishnan. Cooperative spectrum sensing in cognitive radio networks: A survey. *Physical Communication (Elsevier) Journal*, 2011. To appear.

163. J. Lunden, V. Koivunen, A. Huttunen, and H. Poor. Censoring for collaborative spectrum sensing in cognitive radios. In *Proc. of Forty-First Asilomar Conference on Signals, Systems and Computers (ACSSC) 2007*, pages 772–776, 2007.

164. J. Lunden, V. Koivunen, A. Huttunen, and H. Poor. Collaborative cyclostationary spectrum sensing for cognitive radio systems. *IEEE Transactions on Signal Processing,*, 57(11):4182–4195, November 2009.

165. C. Luo, F. R. Yu, H. Ji, and V. C. M. Leung. Cross-Layer Design for TCP Performance Improvement in Cognitive Radio Networks. *IEEE Transactions on Vehicular Technology*, 59(5):2485 –2495, June 2010.

166. C. Luo, F. R. Yu, H. Ji, and V. C. M. Leung. Optimal Channel Access for TCP Performance Improvement in Cognitive Radio Networks. *Springer Wireless Networks*, 1(16):1–14, 2010.

167. H. Luo, S. Lu, and V. Bharghavan. A new model for packet scheduling in multihop wireless networks. In *Proc. of MobiCom 2006*, August 2006.

168. L. Luo and S. Roy. Analysis of Search Schemes in Cognitive Radio. In *Proceedings of IEEE Workshop on Networking Technologies for Software Defined Radio Networks*, pages 647–654, 2007.

169. J. Ma, G. Zhao, and Y. Li. Soft combination and detection for cooperative spectrum sensing in cognitive radio networks. *IEEE Transactions on Wireless Communications*, 7:4502–4507, 2008.

170. L. Ma, X. Han, and C. C. Shen. Dynamic open spectrum sharing MAC protocol for wireless ad hoc network. In *roc. of IEEE DySPAN 2005*, pages 203–213, November 2005.

171. L. Ma, X. Han, and C.C. Shen. Dynamic Open Spectrum Sharing for Wireless Ad Hoc Networks. *Proceedings of IEEE Dynamic Spectrum Access Networks (DySPAN)*, 36:203–213, Nov 2005.

172. L. Ma, C. C. Shen, and B. Ryu. Single-Radio Adaptive Channel Algorithm for Spectrum Agile Wireless Ad Hoc Networks. *Proceedings of IEEE DySPAN*, 36:547–558, Apr 2007.

173. R. Maheswaran and T. Basar. Nash equilibrium and decentralized negotiation in auctioning divisible resources. *Group Decision and Negotiation*, 12(5):361–395, 2007.

174. D. Maldonado, B. Lie, A. Hugine, T.W. Rondeau, and C.W. Bostian. Cognitive radio applications to dynamic spectrum allocation. In *Proc. of IEEE DySPAN 2005*, number November, page 597600., 2005.

175. S. Maleki, A. Pandharipande, and G. Leus. Energy-efficient distributed spectrum sensing for cognitive sensor networks. *IEEE Sensor Journal*, 2011. to appear.

176. G. Marias. Spectrum scheduling and brokering based on qos demands of competing wisps. In *Proc. of IEEE DySPAN 2005*, pages 684–687, November 2005.

177. S. Mascolo, C. Casetti, M. Gerla, M. Y. Sanadidi, and R. Wang. TCP westwood: Bandwidth Estimation for Enhanced Transport over Wireless Links. In *Proc. of ACM MOBICOM*, pages 287–297, 2001.

178. M. Mathis, J. Mahadavi, S. Floyd, and A. Romanow. TCP selective acknowledgment options, Oct 2006. Internet Engeneering Task Force, Request for Comments (Proposed Standard) 2018.

179. C. N. Mathur and K. P. Subbalakshmi. *Cognitive Networks: Towards Self-Aware Networks*, chapter Security Issues in Cognitive Radio Networks. John Wiley & Sons, Ltd, 2007.

180. R. Menon, R. M. Buehrer, and J. H. Reed. Outage probability based comparison of underlay and overlay spectrum sharing techniques,. In *Proc. of IEEE DySPAN 2005*, pages 101–109, November 2005.

181. J. Mietzner, L. Lampe, and R. Schober. Distributed transmit power allocation for relay-assisted cognitive-radio systems. In *Proc. of Asilomar Conf. on Signals, Systems, and Comuputers, November 2007*, pages 792–796, 2007.

182. A. Min and K. Shin. An optimal sensing framework based on spatial rssprofile in cognitive radio networks. In *Proc. of IEEE SECON 2009*, 2009.

183. S. M. Mishra, A. Sahai, and R. W. Brodersen. Cooperative sensing among cognitive radios. In *Proc. of IEEE ICC 2006, Istanbul, Turkey*, volume 4, pages 1658–1663, June 2006.

184. A. Motamedi and A. Bahai. MAC protocol design for spectrum-agile wireless networks: Stochastic control approach. In *Proc. of IEEE DySPAN 2007*, pages 448–451, April 2007.

185. A. Mu'alem and N. Nisan. Truthful approximation mechanisms for restricted combinatorial auctions: extended abstract. In *Proc. of 18th national conference on Artificial intelligence 2002*, pages 379–384, 2001.

186. R. Murty. Software-defined reconfigurability radios: smart, agile, cognitive, and interoperable. *Technology@Intel Magazine*, July 2003.

187. R. B. Myerson. *Game Theory: Analysis of Conflict*. Harward University Press, Cambridge, MA, 1991.

188. H. Nan, T. Hyon, and S. Yoo. Distributed Coordinated Spectrum Sharing MAC Protocol for Cognitive Radio. In *Proceedings of IEEE DySPAN*, page 240?49, 2007.

189. T. Nandagopal, T. Kim, X. Gao, and V. Bharghavan. Achieving MAC layer fairness in wireless packet networks. In *Proc. of MobiCom 2000*, August 2000.

190. J. F. Nash. Equilibrium points in n-person games. In *Proc. of theNational Academy of Sciences*, volume 36, pages 48–49, 1950.

191. J. Neel. *Analysis and design of cognitive radio networks and distributed radio resource management algorithms*. PhD thesis, Virginia Polytechnic Institute and State University, September 2006.

192. IEEE Standards Coordinating Committee 41 (Dynamic Spectrum Access Networks. IEEE 1900 Standard. [Online]. Available: http://www.scc41.org/crinfo.

193. N. Nie and C. Comaniciu. Adaptive channel allocation spectrum etiquette for cognitive radio networks. In *Proc. IEEE DySPAN 2005*, pages 269–278, November 2005.

194. M. Oner and F. Jondral. yclostationarity based air interface recognition for software radio systems. In *Proc. of Radio and Wireless Conference*, pages 263–266, September 2004.

195. M. Oner and F. Jondral. On the extraction of the channel allocation information in spectrum pooling systems. *IEEE Journal on Selected Areas in Communications*, 25(3):558–565, April 2007.

196. ORACLE Project. Oracle project deliverable number d2.4 - sensing algorithms part ii. Technical report, 2007.

197. M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

198. P. Pawelczak, R. Venkatesha Prasad, L. Xia, and G. M. M. Ignas Niemegeers. Cognitive Radio Emergency Networks - Requirements and Design. *Proceedings of IEEE DySPAN*, 36:601–606, Nov 2005.

199. V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

200. I. Pefkianakis, S. Wong, and S. Lu. Samer: Spectrum aware mesh routing in cognitive radio networks. In *Proceedings of 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySpan'08*, pages 1–5, 2008.

201. Y. Pei, A. T. Hoang, and Y.-C. Liang. Sensing throughput tradeoff in cognitive radio networks: how frequently should spectrum sensing be carried out? In *Proceedings of 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 5330–5335, 2007.

202. C. Peng, H. Zheng, and B. Y. Zhao. Utilization and fairness in spectrum assignment for opportunistic spectrum access. *ACM Mobile Networks and Applications (MONET)*, 2006.

203. H. N. Pham, Y. Zhang, P. Engelstad, T. Skeie, and F. Eliassen. Energy minimization approach for optimal cooperative spectrum sensing in sensor-aided cognitive radio networks. In *Proc. of ICST WiCON 2010*, March 2010.

204. J. G. Proakis. *Digital communications*. McGRAW-Hill, 2001.

205. FARAMIR Project. Flexible and spectrumaware radio access through measurements and modelling in cognitive radio systems - state of the art review. *Document Number D2.1*, April 2010.

206. M.B. Pursley and T.C. Royster IV. Low-complexity adaptive transmission for cognitive radios in dynamic spectrum access networks. *IEEE Journal on Selected Areas in Communications*, 26(1):83–94, January 2008.

207. H. Zheng Q. Wang. Route and spectrum selection in dynamic spectrum networks. In *Proceedings of 3rd IEEE Consumer Communications and Networking Conference, CCNC'06*, volume 1, pages 625–629, 2006.

208. Z. Quan, S. Cui, A. Sayed, and H. Poor. Optimal multiband joint detection for spectrum sensing in cognitive radio networks. *IEEE Transactions on Signal Processing*, 57(3):11281140., March 2009.

209. Z. Quan, S. Cui, and A. H. Sayed. Optimal linear cooperation for spectrum sensing in cognitive radio networks. *IEEE Journal on Selected Topics in Signal Processing*, 2(1):28–40, February 2008.

210. L. A. DaSilva R. W. Thomas and A. B. MacKenzie. Cognitive networks. In *Proc. IEEE DySPAN 2005*, pages 352–360, November 2005.

211. T. Ragheb, S. Kirolos, J. Laska, A. Gilbert, M. Strauss, R. Baraniuk, and Y. Massoud. Implementation models for analog-to-information conversion via random sampling. In *Proc. of 50th Midwest Symposium on Circuits and Systems (MWSCAS) 2007*, pages 325–328, 2007.

212. C. Raman, R.D. Yates, and N.B. Mandayam. Scheduling variable rate links via a spectrum server. In *Proc. IEEE DySPAN 2005*, pages 110–118, November 2005.

213. S. Ramanathan. A unified framework and algorithm for channel assignment in wireless networks. *Wireless Networks*, 5(2):81–94, March 1999.

214. P. A. Ramsdale and W. B. Harrold. Techniques for cellular networks incorporating microcells. In *Proc. of IEEE PIMRC 1992*, pages 169–173, Boston, MA, USA, October 1992.

215. B. Razavi. *RF Microelectronics*. Prentice Hall, 1997.

216. M. Di Renzo, W.Wang, H. L. Imbriglio, F. Graziosi, and F. Santucci. Cooperative spectrum sensing over correlated log-normal sensing and reporting channels. In *Proc. of IEEE GLOBECOM 2009*, 2009.

217. M. Di Renzo, W.Wang, H. L. Imbriglio, F. Graziosi, and F. Santucci. Distributed data fusion over correlated log-normal sensing and reporting channels: Application to cognitive radio networks. *IEEE Transactions on Wireless Communications*, 8:5813–5821, 2009.

218. J. Riihijarvi, P. Mahonen, M. Petrova, and V. Kolar. Enhancing cognitive radios with spatial statistics: From radio environment maps to topology engine. In *Proc. of 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM) 2009*, Hannover, Germany, June 2009.

219. J. Riihijarvi, P. Mahonen, M. Wellens, and M. Gordziel. Characterization and modelling of spectrum for dynamic spectrum access with spatial statistics and random fields. In *Proc. of the First International Workshop on Cognitive Radios and Networks (CRNETS) 2008*, Cannes, France, 2008.

220. T. W. Rondeau, D. Maldonado B. Le, D. Scaperoth, and C. W. Bostian. Cognitive radio formulation and implementation. In *Proc. of CROWN 2006*, 2006.

221. S. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Pearson Education, Inc, Upper Saddle River, New Jersey, 2003.

222. S. S. Gandhi, L. Cao C. Buragohain, H. Zheng, and S. Suri. A general framework for wireless spectrum auctions. In *Proc. of IEEE DySPAN*, 2007.

223. and Y. Liu S. Xie, Y. Zhang, and R. Yu. A parallel cooperative spectrum sensing in cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 59:4079–4092, 2010.

224. A.-R. Sadeghi, M. Schunter, and S. Steinbrecher. Private auctions with multiple rounds and multiple items. In *Proc. of 13th Intl. Workshop on Database and Expert Systems Applications, 2002*, pages 423–427, 2002.

225. A. Sahai and D. Cabric. Cyclostationary feature detection. Tutorial presented at the IEEE DySPAN 2005, Baltimore, MD, November 2005. (http://www.eecs.berkeley.edu/ sahai/Presentations/DySPAN05 part2.ppt.).

226. A. Sahai, N. Hoven, and R. Tandra. Some fundamental limits on cognitive radio. In *Proc. of Allerton Conference on Communication, Control, and Computing 2004*, October 2004.

227. H. A. Bany Salameh, M. Krunz, and O. Younis. Cooperative Adaptive Spectrum Sharing in Cognitive Radio Networks. *IEEE/ACM Transactions on Networking*, 18(4):1181 –1194, 2010.

228. T. Salonidis and L. Tassiulas. Distributed on-line schedule adaptation for balanced slot allocation in wireless ad hoc networks. In *Proc. of IWQoS 2004*, 2004.

229. A. Sampath, L. Yang, L. Cao, H. Zheng, and B. Y. Zhao. High throughput spectrum-aware routing for cognitive radio based ad-hoc networks. In *in: 3th International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrwownCom 2008*, 2008.

230. T. Sandholm and S. Suri. Market clearability. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI) 2001*, 2001.

231. S. Sankaranarayanan, P. Papadimitratos, A. Mishra, and S. Hershey. A band- width sharing approach to improve licensed spectrum utilization. In *Proc. of IEEE DySPAN 2005*, pages 279–288, November 2005.

232. D. Sarkar and H. Narayan. Transport Layer Protocols for Cognitive Networks. In *INFOCOM IEEE Conference on Computer Communications Workshops*, pages 1–6, 2010.

233. Y. Selen, H. Tullberg, and J. Kronander. Sensor selection for cooperative spectrum sensing. In *Proc. of IEEE DySPAN 2008*, 2008.

234. S. Sengupta, M. Chatterjee, and S. Ganguly. An economic framework for spectrum allocation and service pricing with competitive wireless service providers. In *Proc. of DySpan 2007*, 2007.

235. P. J. Seok and K.S. Rye. Developing MVNO market scenarios and strategies through a scenario planning approach. In *Proc. of 7th International Conference on Advanced Communication Technology, (ICACT) 2005,*, volume 1, pages 137–142, 2005.

236. S. Shenker. Fundamental design issues for the future internet. *IEEE Journal of Selected Areas in Communications*, 13(7):1176–1188, September 1995.

237. H.-P. Shiang and M. van der Schaar. Distributed resource management in multihop cognitive radio networks for delay-sensitive transmission. *IEEE Transactions on Vehicular Technology*, 58(2):941–953, 2009.

238. A. M. R. Slingerland, P. Pawelczak, R. V. Prasad, A. Lo, and R. Hekmat. Performance of transport control protocol over dynamic spectrum access links. In *Proc. of IEEE DySPAN*, pages 486–495, April 2007.

239. C. Song and Q. Zhang. Sliding-window algorithm for asynchronous cooperative sensing in wireless cognitive networks. In *Proc. of IEEE ICC 2008*, 2008.

240. C. R. Stevenson, C. Cordeiro, E. Sofer, and G. Chouinard. Functional requirements for the 802.22 WRAN standard, September 2006.

241. J. A. Stine. Spectrum management: the killer application of ad hoc and mesh networking. In *Proc. of IEEE DySPAN 2005*, 2005.

242. H. Su and X. Zhang. Opportunistic MAC Protocols for Cognitive Radio Based Wireless Networks. In *Proceedings of Annual Conference on Information Sciences and Systems*, pages 363–368, Mar 2007.

243. H. Su and X. Zhang. CREAM-MAC: An Efficient Cognitive Radio-EnAbled multi-Channel MAC Protocol for wireless Networks. *IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 36:1–8, 2008.

244. K. Sundaresan, V. Anantharaman, H. Y. Hsieh, and R. Sivakumar. ATP: A reliable transport protocol for ad hoc networks. *IEEE Trans. on Mobile Computing.*, 4(6):588–603, November 2005.

245. R. Tandra and Sahai A. Snr walls for feature detectors. In *IEEE DySPAN 2007*, pages 559–570, Dublin, Ireland, April 2007.

246. R. Tandra and A. Sahai. Snr walls for signal detection. *IEEE Journal of Selected Topics in Signal Processing*, 2(1):4–17, February 2008.

247. H. Tang. Some physical layer issues of wide-band cognitive radio system. In *Proc. IEEE DySPAN 2005*, pages 151–159, November 2005.

248. Z. Tian. Compressed wideband sensing in cooperative cognitive radio network. In *Proc. of IEEE Globecom 2008*, November 2008.

249. Z. Tian and G. Giannakis. Compressed sensing for wideband cognitive radios. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1357–1360, April 2007.

250. M. Timmers, S. Pollin, A. Dejonghe, L. Van der Perre, and F. Catthoor. A Distributed Multichannel MAC Protocol for Multihop Cognitive Radio Networks. *IEEE Transactions on Vehicular Technology*, 59(1):446 –459, 2010.

251. J. Unnikrishnan and V. V. Veeravalli. Cooperative sensing for primary detection in cognitive radio. *IEEE Journal of Selected Topics in Signal Processing*, 2(1):18–27, January 2008.

252. H. R. Varian. *Intermediate Microeconomics: A Modern Approach*. W. W. Norton and Company, New York, US, 1987.

253. P. K. Varshney. *Distributed Detection and Data Fusion*. Springer-Verlag, New York, US, 1997.

254. W. Vickrey. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, 16:8–37, 1961.

255. E. Visotsky, S. Kuffner, and R. Peterson. On collaborative detection of TV transmissions in support of dynamic spectrum sharing. In *Proc. of IEEE DySPAN 2005*, page 338345., 2005.

256. F. Visser, G. Janssen, and P. Pawelczak. Multinode spectrum sensing based on energy detection for dynamic spectrum access. In *Proc. of IEEE VTC2008-Spring*, pages 1394–1398, 2008.

257. D.T.P. Viswanath and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48(6):1277–1294, June 2002.

258. S. De Vries and R. V. Vohra. Combinatorial auctions: A survey. *INFORMS Journal on Computing*, 3:284–309, 2003.

259. M. C. Vuran and I. F. Akyildiz. AMAC: Adaptive medium access control for next generation wireless terminals. *IEEE/ACM Transactions on Networking*, 15(3):574–587, June 2007.

260. and Z. Han W. Saad, M. Debbah, A. Hjorungnes, and T. Basar. Coalitional games for distributed collaborative spectrum sensing in cognitive radio networks. In *Proc. of IEEE INFOCOM*, pages 2114–2122, 2009.

261. A. Wald. *Sequential Analysis*. Wiley, New York, US, 1947.

262. B. Wang, K. Ray Liu, and T. Clancy. Evolutionary cooperative spectrum sensing game: how to collaborate? *IEEE Transactions on Communications*, 58(3):890–900, March 2010.

263. P. Wang, L. Xiao, and S. Zhou. Optimization of detection time for channel efficiency in cognitive radio system. pages 111–115, Mar 2007.

264. P. Wang, L. Xiao, S. Zhou, and J. Wang. Optimization of detection time for channel efficiency in cognitive radio systems. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, pages 111–115, March 2007.

265. W. Wang and B. Li. Market-driven bandwidth allocation in selfish overlay networks. In *Proc. of IEEE INFOCOM 2005*, pages 2578–2589, 2005.

266. X. Wang and Q. Zhu. Power control for cognitive radio based on game theory. In *in Proc. of IEEE Wireless Communications, Networking and Mobile Computing (WiCom) 2007*, pages 1256–1259, September 2007.

267. T. A. Weiss, J. Hillenbrand, A. Krohn, and F. K. Jondral. Efficient signaling of spectral resources in spectrum pooling systems. In *Proc. 10th Symposium on Communications and Vehicular Technology (SCVT) 2003*, November 2003.

268. T. A. Weiss and F. K. Jondral. Spectrum pooling: an innovative strategy for the enhancement of spectrum efficiency. *IEEE Communications Magazine*, 42(3):8–14, March 2004.

269. M. Wellens and P. Mahonen. Lessons Learned from an Extensive Spectrum Occupancy Measurement Campaign and a Stochastic Duty Cycle Model. *Proc. of International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TridentCom)*, Apr 2009.

270. M. Wellens, J. Riihijarvi, and P. Mahonen. Empirical time and frequency domain models of spectrum use. *Physical Communication Journal (Elsevier)*, 2(1):10–32, May 2009.

271. B. Wild and K. Ramchandran. Detecting primary receivers for cognitive radio applications. In *Proc. of IEEE DySPAN 2005*, November 2005.

272. D. Willkomm and A. Wolisz J. Gross. Reliable link maintenance in cognitive radio systems. In *Proc. of IEEE DySPAN 2005*, pages 371–378, November 2005.

273. D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary users in cellular networks: a large-scale measurement study. In *Proc. of IEEE DySPAN 2008*, Chicago, IL, USA, October 2008.

274. H. Xiao, K. C. Chua, J. A. Malcolm, and Y. Zhang. Theoretical Analysis of TCP Throughput in Adhoc Wireless Networks. In *Proc. of IEEE GLOBECOM*, pages 2714–2719, Nov-Dec 2005.

275. M .Xie, W. Zhang, and K.-K. Wong. A geometric approach to improve spectrum efficiency for cognitive relay networks. *IEEE Transactions on Wireless Communications*, 9(1):268–281, 2010.

276. H. Sherali Y. Hou, Y. Shi. Optimal spectrum sharing for multi-hop software defined radio networks. In *Proceedings of 26th IEEE International Conference on Computer Communications, INFOCOM'07*, pages 1–9, 2007.

277. H. Sherali Y. Hou, Y. Shi. Spectrum sharing for multi-hop networking with cognitive radios. *IEEE Journal on Selected Areas in Communications*, 26(1):146–155, 2008.

278. L. Yang, L. Cao, and H. Zheng. Proactive channel access in dynamic spectrum networks. *Physical Communication (Elsevier)*, 1(2):103–111, June 2008.

279. Z. Yang, G. Cheng, W. Liu, W. Yuan, and W. Cheng. Local coordination based routing and spectrum assignment in multi-hop cognitive radio networks. *Mobile Networks and Applications*, 13(1-2):67–81, 2008.

280. W. Yiu, G. Ginis, and J. Cioffi. Distributed multiuser power control for digital subscriber lines. *IEEE Journal on Selected Areas in Communications*, 20(5):1105–1115, June 2002.

281. S. Yoo, H. Nan, , and T. Hyon. *DCR-MAC: Distributed Cognitive Radio MAC Protocol for Wireless Ad Hoc Networks.* , 9(5):631?53, May 2009.

282. T. Yucek and H. Arslan. A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys and Tutorials*, 11(1):116–130, First Quater 2009.

283. G. Giannakis Z. Tian. A wavelet approach to wideband spectrum sensing for cognitive radios. In *Proc. of 3rd International Conference on Cognitive Radio Oriented Wireless Networks Communications (CROWNCOM) 2006*, June 2006.

284. S. A. Zekavat and X. Li. User-central wireless system: Ultimate dynamic channel allocation. In *Proc. of IEEE DySPAN 2005*, pages 82–87, November 2005.

285. F. Zeng, C. Li, and Z. Tian. Distributed compressive spectrum sensing in cooperative multihop cognitive networks. *IEEE Journal of Selected Topics in Signal Processing.*, 2011. to appear.

286. F. Zeng, Z. Tian, and C. Li. Distributed compressive wideband spectrum sensing in cooperative multi-hop cognitive networks. In *Proc. of IEEE ICC 2010*, 2010.

287. Y. Zeng and Y.-C. Liang. Covariance based signal detections for cognitive radio. In *Proc. of IEEE DySPAN 2007*, pages 202–207, Dublin, Ireland, April 2007.

288. Y. Zeng and Y.-C. Liang. Eigenvalue-based spectrum sensing algorithms for cognitive radio. *IEEE Transactions on Communications*, 56(6):1784–1793, June 2009.

289. J. Zhang and Q. Zhang. Stackelberg game for utility-based cooperative cognitive radio networks. In *Proc. of ACM MobiHoc 2009*, 2009.

290. Q. Zhang, J. Jia, and J. Zhang. Cooperative relay to improve diversity in cognitive radio networks. *IEEE Communications Magazine*, 47(2):111–117, February 2009.

291. W. Zhang and K. Letaief. Cooperative spectrum sensing with transmit and relay diversity in cognitive radio networks. *IEEE Transactions on Wireless Communications*, 7(12):4761–4766, December 2008.

292. J. Zhao, H. Zheng, and G.-H. Yang. Distributed coordination in dynamic spectrum allocation networks. In *Proc. IEEE DySPAN 2005*, pages 259–268, November 2005.

293. J. Zhao, H. Zheng, and G.-H. Yang. Spectrum agile radios: Utilization and sensing architecture. In *Proc. of IEEE International Symposium on Dynamic Spectrum Access Networks*, pages 259–268, November 2005.

294. J. Zhao, H. Zheng, and G. H. Yang. Spectrum sharing through distributed coordination in dynamic spectrum access networks. *Wireless Communications and Mobile Computing*, 7(9):1061–1075, Nov 2007.

295. Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. *IEEE Journal on Selected Areas in Communications*, 25(3):589–600, Apr 2007.

296. Y. Zhao, J. Gaeddert, K. K. Bae, and J. H. Reed. Radio environment map enabled situation-aware cognitive radio learning algorithms. In *Proc. of Software Defined Radio (SDR) Technical Conference 2006*, 2006.

297. H. Zheng and C. Peng. Collaboration and fairness in opportunistic spectrum access. In *Proc. of 40th annual IEEE International Conference on Communications 2005*, June 2005.

298. C. Zhou and C. Chigan. A Game Theoretic DSA-Driven MAC Framework for Cognitive Radio Networks. *IEEE International Conference on Communications (ICC)*, 36:4165–4169, May 2008.

299. X. Zhou, S. Gandhi, S. Suri, and H. Zheng. eBay in the sky: strategy-proof wireless spectrum auctions. In *Proc. of Mobicom 2008*, 2008.

300. X. Zhou, J. Ma, G. Li, Y. Kwon, and A. Soong. Probability-based combination for cooperative spectrum sensing. *IEEE Transactions on Communications*, 58(2):463–466, February 2010.

301. X. Zhou, J. Ma, G. Y. Li, Y. H. Kwon, and A. C. K. Soong. Probability-based combination for cooperative spectrum sensing in cognitive radio networks. In *Proc. of IEEE ICC 2009*, 2009.

302. X. Zhou, J. Ma, G. Y. Li, Y. H. Kwon, and A. C. K. Soong. Probability-based combination for cooperative spectrum sensing. *IEEE Transactions on Communications*, 58(2):463–466, February 2010.

303. G.-M. Zhu, I. Akyildiz, and G.-S. Kuo. STOD-RP: A Spectrum-Tree Based On-Demand Routing Protocol for Multi-Hop Cognitive Radio Networks. In *Proceedings ofIEEE Global Telecommunications Conference, GLOBECOM'08*, pages 1–5, 2008.

304. X. Zhu, L. Shen, and T.-S. P. Yum. Analysis of cognitive radio spectrum access with optimal channel reservation. *IEEE Communications Letters*, 11(4):304–306, April 2007.

305. Q. Zou, S. Zheng, and A. Sayed. Cooperative spectrum sensing via sequential detection for cognitive radio networks. In *Proc. of IEEE 10th Workshop on Signal Processing Advances inWireless Communications 2009*, page 121 125., 2009.

306. Q. Zou, S. Zheng, and A. Sayed. Cooperative sensing via sequential detection. *IEEE Transactions on Signal Processing.*, 2011. to appear.