行政院國家科學委員會補助專題研究計畫 ■ 成果報告
□期中進度報告

廣義的機會式通訊：
無線行動網路中之競爭、合作與感知--子計畫四:
感知無線行動網路之協力式媒體存取控制協定設計與用戶/基地台選取研究

計畫類別：□ 個別型計畫 ■ 整合型計畫
計畫編號：NSC 96-2628-E-009-004-MY3
執行期間： 96 年 8 月 1 日至 99 年 7 月 31 日

計畫主持人：王蒞君 教授
共同主持人：
計畫參與人員： 王中瑋

成果報告類型(依經費核定清單規定繳交)：□精簡報告 ■完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告零份
□赴大陸地區出差或研習心得報告零份
□出席國際學術會議心得報告及發表之論文各三份
□國際合作研究計畫國外研究報告書零份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、
列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學電信工程研究所

中 華 民 國 99 年 9 月 24 日

摘要

本報告探討感知無線網路的頻譜管理問題。在此網路中，來自主要使用者的『多次中斷』將大大地影響次要使用者的通訊效能。每當次要使用者被主要使用者中斷時，次要使用者必須選擇一個適合的通道進行頻譜切換，以便繼續未完成的傳輸。很明顯地，『多次中斷』將造成多次的頻譜切換，並且增加次要使用者連線的傳輸延遲。為了從一個宏觀的角度來分析感知無線網路下『多次中斷』行為對『次要使用者連線』所造成的傳輸延遲，本報告提出一個優先權排隊理論的分析模型替感知無線網路的頻譜使用行為進行建模。藉由此模型，我們分析次要使用者的一個重要服務品質參數：『完整系統時間』。

在此報告中，基於排隊理論分析模型，我們發展具服務品質考量的頻譜管理機制，其中包括（1）頻譜選擇機制、(2)頻譜切換機制、和（3）頻譜分享機制的設計與討論。針對這些機制的具體研究成果敘述如下：(1)針對頻譜選擇問題，我們提出一個具有負載平衡功效的頻譜選擇機制來優化次要使用者的『完整系統時間』；(2)針對頻譜切換問題，我們量化在多通道下多次頻譜切換對次要使用者所造成的『完整系統時間』增加量；(3)針對頻譜分享問題，我們提出一個允入控制機制來避免主要使用者被次要使用者干擾並優化次要使用者的『完整系統時間』。此外，我們也分析不同的媒介存取協定所造成的傳輸延遲，並提出兩種媒介存取協定來改善網路效能。我們完整探討這三種頻譜管理機制對次要使用者所造成的傳輸延遲。基於這些分析結果，在不同資料到達率與服務時間分佈下，我們可以設計相對應的頻譜管理機制來增強次要使用者連線的傳輸品質。

總而言之，本報告的主要貢獻是提出一個以排隊理論為基礎的分析模型並用多樣化的角度與觀點來對感知無線網路效能進行分析。本報告所建議之模型可以提供一個很好的感知無線網路效能之分析架構。

Abstract

In this report, we investigate spectrum management techniques in cognitive radio (CR) networks with quality of service (QoS) provisioning. One fundamental issue in enhancing QoS performance for the secondary users is the multiple interruptions from the primary users during each secondary user's connection. These interruptions from the primary users result in the phenomenon of multiple spectrum handoffs within one secondary user's connection. Thus, a set of target channels for spectrum handoffs are needed to be selected sequentially. In order to characterize the general channel usage behaviors with multiple handoffs from a macroscopic viewpoint, an analytical framework based on the preemptive resumption priority (PRP) M/G/1 queueing theory is introduced. Based on the PRP M/G/1 queueing network model, we can evaluate the effects of multiple handoffs on the overall system time, which is an important QoS performance measure for the secondary connections in CR networks.

The proposed analytical framework can provide important insights into the design of spectrum management techniques in CR networks. In order to demonstrate the effectiveness of this analytical model, we discuss various spectrum management techniques, consisting of spectrum decision, spectrum sharing, and spectrum mobility. For the *spectrum decision* issue, we show how to determine which channels are required to probe and transmit. For the *spectrum mobility* issue, we illustrate how to characterize the effects of multiple handoffs, where the secondary users can have different operating channels before and after spectrum handoff. For the *spectrum sharing* issue, we explore how to determine the optimal admission probability to avoid the interference between primary and secondary users in the presence of false alarm and missed detection. Furthermore, the latency performance of

various MAC protocols is analyzed and two MAC protocols are proposed to enhance the QoS performance of CR networks. From numerical results, we can develop traffic-adaptive spectrum management policies to enhance the QoS performance of the secondary users in CR networks with various traffic arrival rates and service distributions.

To summarize, the main contribution of this report is to investigate the modeling techniques for CR networks from a macroscopic viewpoint based on the queueing theory. The proposed analytical framework can help analyze the performances of CR networks and provide important insights into the design of various spectrum management techniques with enhanced QoS performances.

Contents

List of Tables

List of Figures

Glossary of Symbols

- type-i secondary connection: the secondary user's connection that has experienced i interruptions.
- $\lambda_p^{(k)}$: the arrival rate of the primary users' connections whose default channels are channel k.
- λ_s: the arrival rate of the secondary users' connections.
- $\lambda_s^{(k)}$: the initial arrival rate of the secondary users' connections whose initial channels are channel k.
- $\omega_i^{(k)}$: the arrival rate of the type-i secondary connections at channel k.
- $X_p^{(k)}$: the service time of the primary users' connections whose default channels are channel k.
- X_s: the service time of the secondary users' connections.
- $X_s^{(k)}$: the service time of the secondary users' connections whose default channels are channel k.
- $\Phi_i^{(k)}$ is the effective service time for the i^{th} interruption at channel k. It is the transmission duration of a secondary connection between the i^{th} and the $(i+1)^{th}$ interruptions at channel k.

- $\widetilde{X}_s$: the actual service time of the secondary users' connections when the effects of sensing errors are considered.
- $f_p^{(k)}(x)$: probability density function of $X_p^{(k)}$.
- $f_s(x)$: probability density function of X_s.
- $f_s^{(k)}(x)$: probability density function of $X_s^{(k)}$.
- $f_i^{(k)}(\phi)$: probability density function of $\Phi_i^{(k)}$.
- P_M: missed detection probability.
- P_F: false alarm probability.
- π_p: outage probability for the primary connections.
- π_s: outage probability for the secondary connections.

Chapter 1

Introduction

Recent measurements show that the licensed spectrum is under-utilized [1]. In order to solve this spectrum waste issue, many technologies have been proposed. Cognitive radio (CR) is one of the promising approaches to improve spectrum utilization [2–8]. A CR network consists of the primary and the secondary networks as shown in Fig. 1.1. The primary networks are defined as the systems that own the licensed spectrum, such as the cellular mobile networks or the TV broadcast networks. By contrast, the secondary networks do not have any licensed frequency. By allowing the secondary users to temporarily access the primary user's under-utilized licensed spectrum, CR can significantly improve spectrum efficiency and enhance the quality of service (QoS) performance of the secondary users.

One fundamental issue for enhancing QoS performance of the secondary users in CR networks is the spectrum handoff issue. When the high-priority primary user appears at its licensed channel being occupied by the low-priority secondary users, these secondary users must vacate the occupied channel. In order to vacate the occupied channel to the primary user and discover the suitable target channel to resume the unfinished transmission,

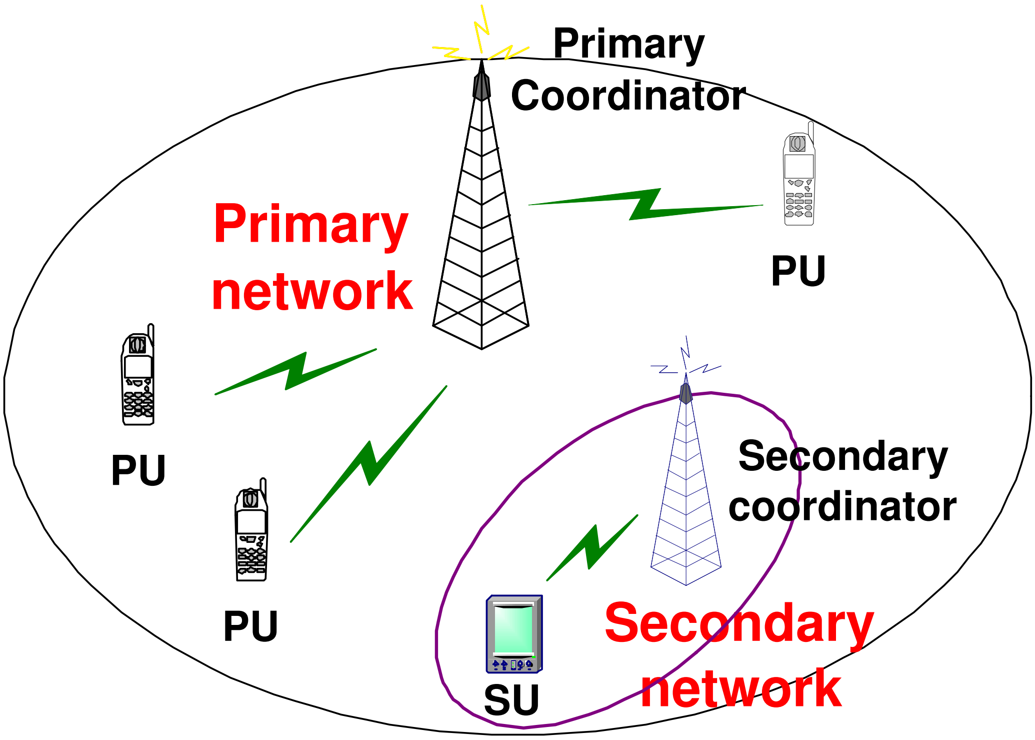


Figure 1.1: An illustrative example of CR network, which consists a primary network and a secondary network. There are three primary users (PUs) and one secondary user (SU) in the primary and secondary networks, respectively.

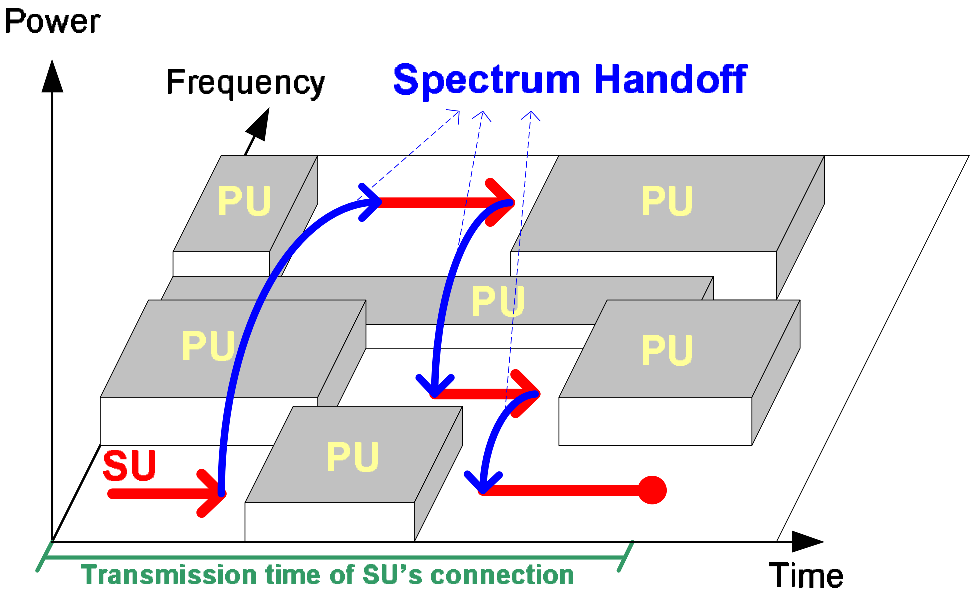


Figure 1.2: During the transmission period of secondary user (SU), it experiences multiple handoffs.

the *spectrum handoff* procedures are initiated for the secondary users [9, 10][1]. During the transmission period of a secondary connection, multiple interruptions from the primary users result in multiple spectrum handoffs as show in Fig. 1.2. These spectrum handoffs will degrade the QoS performance of the secondary users.

In order to overcome the performance degradation issue due to multiple

[1] Spectrum handoff in CR networks is different from the conventional handoff mechanisms in cellular mobile networks. Spectrum handoff considers two types of users with different priorities, where the high-priority primary users have the right to interrupt the transmission of the low-priority secondary users. When the interruption event occurs, the secondary user must stop using the current channel even though the received signal strength is still acceptable. In contrast, all users in the conventional handoff mechanisms have the same priority to access channels and they change their operating channels mainly due to deterioration of signal quality [11].

spectrum handoffs for the secondary users, various spectrum management techniques in CR networks are re-examined from a link connection quality perspective. There are four spectrum management functionalities in CR networks [12]:

1. Spectrum sensing: The secondary users should monitor all channels in order to capture channel characteristic and detect spectrum holes. Based on sensing results, the secondary users can find some candidate channels to transmit data. In this report, we consider a fully-connected CR network. Hence, the transmitter and receiver of a secondary connection can have the same consensus on sensing results.

2. Spectrum decision: The secondary users can select the best channel from many candidate channels to transmit data. This decision should take the traffic statistics of the primary users as well as the secondary users into account.

3. Spectrum mobility: The secondary users must vacate the occupied channel when the primary user appears because the primary users have the preemptive priority to access channels. In order to return the occupied channel to the primary users and resume the unfinished transmission at the suitable channel, the *spectrum handoff* procedures are initiated for the interrupted secondary user.

4. Spectrum sharing: The secondary users must coordinate their transmissions and avoid interfering with transmission of the primary users.

Referring to [13], the relationships of these four spectrum management functionalities are shown in Fig. 1.3. In the beginning, the traffic requests of secondary users arrive at the CR network. With the spectrum decision

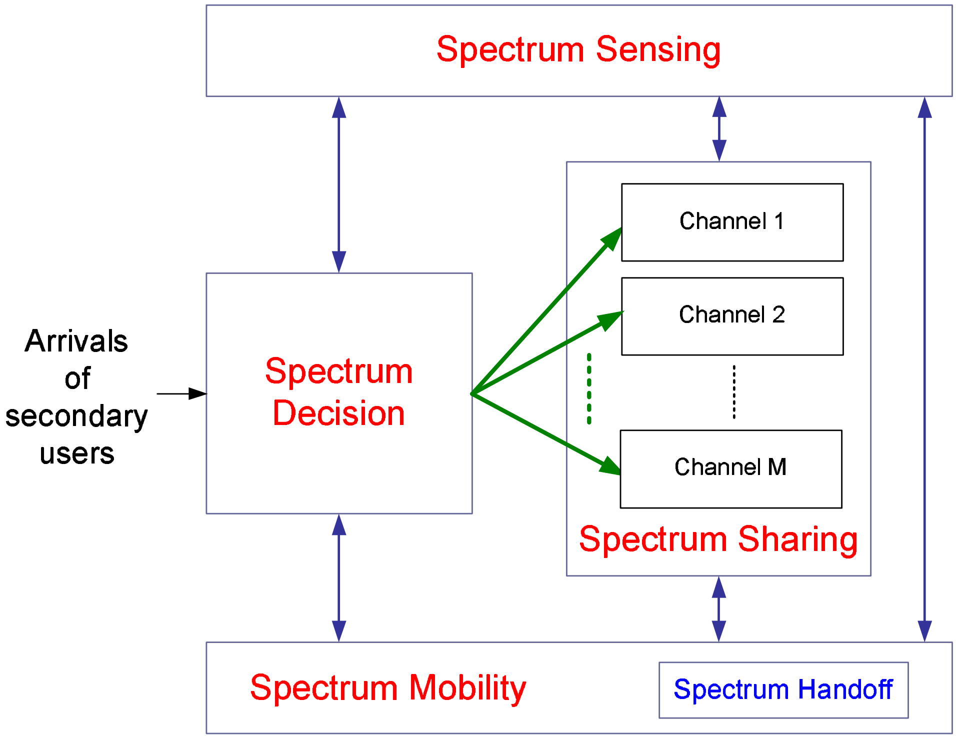


Figure 1.3: Relationships between spectrum sensing, spectrum decision, spectrum mobility, and spectrum sharing functionalities.

functionality, they can determine their initial operating channels from all M candidate channels, which can be found by the spectrum sensing functionality. In order to alleviative the channel contention when multiple secondary users select the same channel and the interference on the primary users when missed detection occur, the spectrum sharing functionality must be implemented. Furthermore, if the primary user appears at the occupied channel, the *spectrum handoff* procedures in the spectrum mobility functionality must be initiated. Based on this dynamic spectrum management, spectrum efficiency can be improved. Note that the multiple handoff issue should be taken into account when designing these spectrum management functionalities.

In this report, we focus on the spectrum decision, spectrum mobility, and spectrum sharing issues. In order to evaluate the system performance of the proposed spectrum management techniques, an analytical framework based on the preemptive resumption priority (PRP) M/G/1 queueing theory is developed to characterize the connection-based channel usage behaviors with multiple handoffs. We investigate the effects of multiple handoffs on the QoS performance and study the performance limitation of various spectrum management techniques in different traffic loads. Different from the traditional work which investigated the effects of multiple handoffs on the network throughput, this report concentrates on the effects of latency performance of the secondary users. Based on the proposed analytical framework, some useful insights into the design of the spectrum management techniques can be provided and the traffic-adaptive spectrum management schemes can be developed according to traffic conditions such as traffic arrival rates and service time distributions.

1.1 Problems and Solutions

In this section, we will briefly describe our problem formulations and the corresponding solutions, including modeling technique for CR networks, traffic-adaptive spectrum mobility issues, load-balancing spectrum decision issues, and interference-avoiding spectrum sharing issues.

1.1.1 Modeling Techniques for Cognitive Radio Networks

In this part, we outline the fundamental modeling issues of opportunistic spectrum access in cognitive radio (CR) networks. In particular, we identify the effects of the general behaviors for the connection-based channel usage on the quality of service (QoS) performances of spectrum management techniques. During the transmission period of a secondary user's connection, the phenomenon of multiple spectrum handoffs resulted from the interruptions of the primary users arises quite often. In addition to multiple interruptions, the connection-based channel usage behaviors are also affected by other factors, including spectrum sensing time, channel switching between different channels, generally distributed service time, and channel contention between multiple secondary users. An analytical framework based on the preemptive resumption priority M/G/1 queueing theory is introduced to characterize these effects simultaneously. The proposed analytical framework can provide important insights into the design of spectrum management techniques in CR networks and can be adapted more flexibly for various traffic arrival rates and service time distributions.

1.1.2 Load-Balancing Spectrum Decision

In this part, we present an analytical framework to design system parameters for load-balancing multiuser spectrum decision schemes in cognitive radio (CR) networks. Unlike the non-load-balancing methods that multiple secondary users may contend for the same channel, the considered load-balancing schemes can evenly distribute the traffic loads of secondary users to multiple channels. Based on the preemptive resume priority (PRP) M/G/1 queueing theory, a spectrum decision analytical model is proposed to evaluate the effects of multiple interruptions from the primary user during each link connection and the sensing errors of the secondary users. With the objective of minimizing the overall system time (i.e., waiting time plus data delivery time) of the secondary users, we derive the optimal number of candidate channels and the optimal channel selection probability for the sensing-based and the probability-based spectrum decision schemes, respectively. We find that the probability-based scheme can yield a shorter overall system time compared to the sensing-based scheme when the traffic loads of the secondary users is light, whereas the sensing-based scheme performs better in the condition of heavy traffic loads. If the secondary users can intelligently adopt the best spectrum decision scheme according to sensing time and traffic parameters, the overall system time can be improved by 50% compared to the existing methods. Furthermore, the proposed analytical model also takes into account of the probability of missed detection and false alarm for the appearance of the primary users, and can help evaluate the impacts of imperfect sensing on the spectrum decision schemes for CR networks.

1.1.3 Proactive Spectrum Handoff

In this part, we present an analytical framework to evaluate the latency performance of connection-based spectrum handoffs in cognitive radio (CR) networks. During the transmission period of a secondary connection, multiple interruptions from the primary users result in multiple spectrum handoffs and the need of predetermining a set of target channels for spectrum handoffs. To quantify the effects of channel obsolete issue on the target channel predetermination, we should consider the three key design features: (1) generally distributed service time of the primary and secondary connections; (2) different operating channels in multiple handoffs; and (3) queueing delay due to channel contention from multiple interrupted secondary connections. To this end, we propose the preemptive resume priority (PRP) M/G/1 queueing network model to characterize the spectrum usage behaviors with all the three design features. This model aims to analyze the extended data delivery time of the secondary connections with proactively designed target channel sequences under various traffic arrival rates and service time distributions. These analytical results are applied to evaluate the latency performance of the connection-based spectrum handoff based on target channel sequences used in the IEEE 802.22 wireless regional area networks standard. Then, to reduce the extended data delivery time, a traffic-adaptive spectrum handoff is proposed, which changes the target channel sequence of spectrum handoffs based on traffic conditions. Compared to the existing target channel selection methods, this traffic-adaptive target channel selection approach can reduce the extended data transmission time by 35%, especially for the heavy traffic loads of the primary users.

1.1.4 Optimal Proactive Spectrum Handoff

In this part, we investigate how to determine an optimal target channel sequence for multiple spectrum handoffs with the minimum cumulative handoff delay for the secondary users in cognitive radio networks. In addition to multiple interruptions from the high-priority primary users, the optimal sequence for spectrum handoffs incorporates the effects of various traffic statistics of both the primary and the secondary users. Compared to the exhaustive search with time complexity of $O(M^L)$, where L is the total number of elements in the target channel sequence and M is the total number of candidate channels for spectrum handoffs, a dynamic programming algorithm with the complexity of $O(LM^2)$ is proposed to determine the optimal target channel sequence for spectrum handoffs. Furthermore, we propose a greedy algorithm with time complexity of $O(M)$ for spectrum handoffs and prove that it only requires to compare six permutations of the target channel sequences. Numerical results show that the cumulative handoff delay of the low-complexity greedy algorithm can approach that of the optimal solution.

1.1.5 Reactive Spectrum Handoff

Spectrum handoff is an important functionality in cognitive radio (CR) networks. Whenever a primary user appears, transmission of the secondary users must be interrupted. In this case, spectrum handoff procedures are initiated for the secondary users in order to search the idle channel to resume the unfinished transmission. Although this dynamic spectrum access scheme can enhance channel utilization, multiple interruptions from the primary users will result in multiple handoffs and thereby increase the transmission latency of the secondary users. Hence, two fundamental issues in CR networks are how much channel utilization can be improved and how long transmission latency is extended for the secondary users due to multiple spectrum handoffs. To solve the first problem, we introduce the preemptive resume priority (PRP) M/G/1 queueing network to characterize the channel usage behaviors of CR networks. Based on this queueing network, channel utilization under various traffic arrival rates and service time distributions can be evaluated. Furthermore, on top of the proposed queueing network, a state diagram is developed to characterize the effects of multiple handoff delay on the transmission latency of the secondary users. The analytical results can provide a helpful insight to study the effects of traffic arrival rates and service time on the transmission latency and then facilitate the designs of admission control rules for the secondary users subject to their latency requirements.

1.1.6 Interference-Avoiding Spectrum Sharing

In this part, we present an analytical framework to design key system parameters for an interference-avoiding admission control mechanism to enhance channel utilization, while maintaining the quality of service (QoS) requirements for both the primary users and secondary users in cognitive radio (CR) networks. Intuitively, a larger admission probability for the secondary users can increase channel utilization, but it leads to more contention between the secondary users and thus affects the latency performance of the secondary users. More importantly, if the missed detection for the presence of the primary user happen, the larger the admission probability of the secondary user, the more the interference to the primary user. In order to find the optimal traffic admission probability, a cross-layer optimization problem is formulated. Our cross-layer design can incorporate the following effects: (1) false alarm and missed detection, power outage in the physical layer; (2) admission probability in the medium access control (MAC) layer; and (3) the traffic statistics as well as the QoS constraints of both the primary and the secondary users in the application layer. The analytical results proposed in this part can calculate the optimal traffic admission probability under various cross-layer parameters and provide useful insights into the tradeoff design of channel utilization and the QoS performance for both the primary and the secondary users.

1.1.7 Latency Analysis for Spectrum Sharing

Spectrum sharing is the key feature of cognitive radio (CR) networks, but it also poses many new challenges on the medium access control (MAC) design. One of key challenges is the fact that the secondary CR users can only borrow the licensed spectrum from the primary users for a short period of time. Hence, unlike many available multi-channel MAC protocols for ad hoc networks where throughput is the main performance issue, the MAC protocols in CR networks shall place more emphases on the access latency. Hence, one fundamental issue arises: how can the spectrum be dimensioned for control channels in order to minimize the access delay of MAC protocol in CR networks? In this paper, we provide a comparative study in an analytical manner on the latency performance of two MAC protocols: 1) dedicated control channel, and 2) embedded control channel approaches. Our results show that an optimal ratio of the control channel bandwidth over the total channel bandwidth can be found to minimize the latency of MAC with dedicated control channels. However, the delay performance of MAC with dedicated control channels is more sensitive to the variations of the data lengths than that of MAC with embedded control channels. Hence, we conclude that the way of dimensioning the spectrum for control frames for MAC in CR networks should be adaptive to the variations of the traffic characteristics and the number of users.

1.1.8 Location-Aware Concurrent Transmission for Spectrum Sharing

In traditional way, the spectrum sensing over a wide frequency range involves the sophisticated time and energy-consuming signal processing [14]. Instead of developing another efficient spectrum sensing technique, in this part, we discuss a challenging but fundamental issue – Can CR devices effectively identify the available spectrum holes without wideband spectrum sensing? Intuitively, when the secondary CR users are far away from the primary user of the legacy system, both CR and the primary users can concurrently transmit their data without causing interference. If a CR device knows the region where it can concurrently transmit with the primary user, the user does not need to rely on the time- and energy-consuming wide-band spectrum scanning to detect the spectrum holes. In addition, it is clear that concurrent transmission can enhance the overall throughput. In this sense, identifying concurrent transmission opportunity shall be given a higher priority over spectrum sensing for a CR user.

The next important issue is how to identify the concurrent transmission region where CR users will not cause any interference to the legacy wireless systems. In this part, we propose to utilize the location awareness techniques to help CR users to identify the concurrent transmission opportunity. Our specific goal is to dimension the concurrent transmission region where CR devices can establish an overlaying ad hoc network on top of an infrastructure-based legacy system. The overlaying ad hoc network can be considered an important application for CR devices because it can reuse the underutilized spectrum and significantly improve the efficiency of the frequency band. It is assumed that the location information of other nodes can be obtained from the upper layer, like the method in [15]. We will also investigate the throughput improvement resulting from concurrent transmissions based on the carrier sense multiple access with collision avoidance (CSMA/CA) MAC protocol.

1.1.9 Neighbor-Aware Cognitive MAC Protocol for Spectrum Sharing

We propose an enhanced CSMA/CA cognitive MAC protocol. The CSMA/CA MAC is a well known MAC protocol for resolving contentions in many existing wireless networks, such as the IEEE 802.11 wireless local area network (WLAN) [16]. Since the CSMA/CA MAC protocol requests a station to sense the channel usage before transmissions, this kind of MAC protocol has potential and is natural to become a strong candidate for the overlaying cognitive ad hoc networks. Thus, we are motivated to ask two fundamental questions:

Can the CSMA/CA MAC protocol be directly adopted for the overlaying ad hoc networks without changes? If not, how should the CSMA/CA MAC protocol be modified to become an appropriate cognitive MAC protocol for the overlaying ad hoc networks in the presence of primary users?

To this end, we propose a MAC protocol with QoS provisioning for overlaying single-hop ad hoc networks on the legacy system, which is assumed to adopt the time-division multiple access (TDMA) MAC protocol for packet transmissions to guarantee QoS requirements, like GSM and WiMax. The proposed MAC protocol coincides with the four stages of the cognition cycle, as shown in Fig. 1.4 [17–19]. Corresponding to the cognition cycle, the neighbor list establishment mechanism maps to the ***observe*** stage, which can help a secondary user to recognize the spectrum usage around its neighborhood. In the ***plan*** stage, the improved contention resolution mechanism can enhance the overall performance in terms of throughput, access delay and fairness. In the ***decide*** stage, the newly proposed reservation scheme ensures the established ad hoc link with satisfactory QoS requirements. The distributed frame synchronization mechanism then coordinates the transmissions among stations in the ***act*** stage. With the helps of these QoS enhanced techniques, an ad hoc network based on the proposed improved CSMA/CA MAC protocol can coexist with a legacy TDMA system.

1.2 Report Outlines

This report consists of four themes as shown in Fig. 1.5. The first part is to outlines the fundamental modeling issues of various spectrum management techniques in CR networks. Then, an analytical framework based on the preemptive resumption priority M/G/1 queueing theory is introduced to

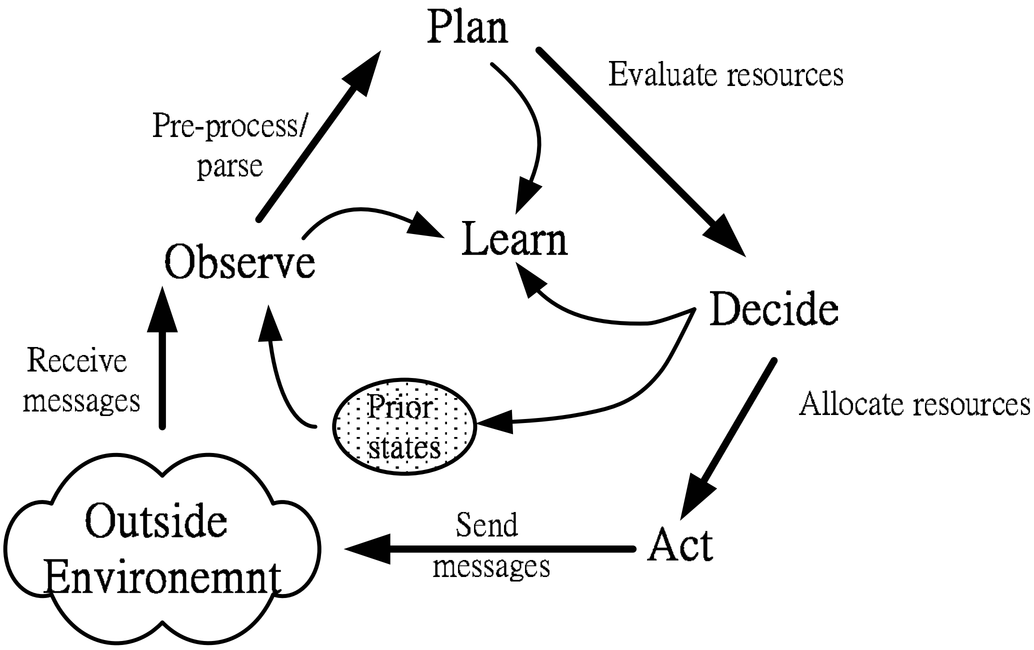


Figure 1.4: The main functions of a cognition cycle for CR devices.

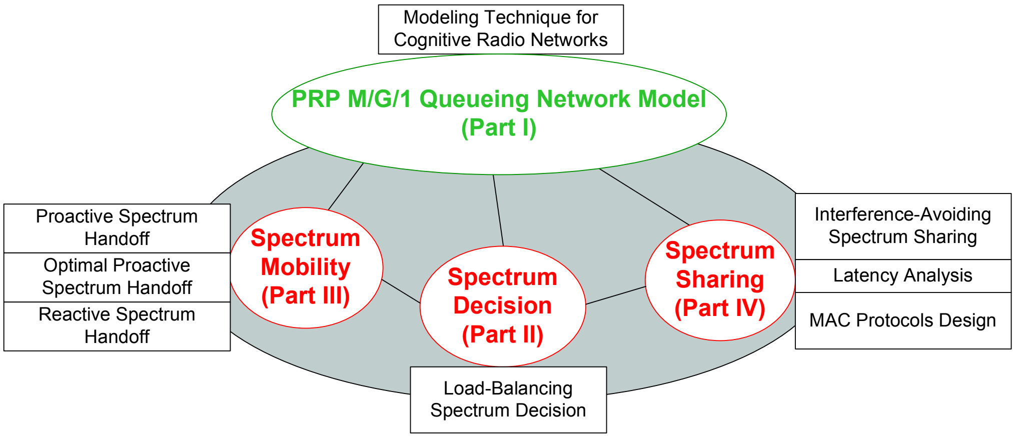

Modeling Technique for Cognitive Radio Networks
PRP M/G/1 Queueing Network Model (Part I)
Proactive Spectrum Handoff
Optimal Proactive Spectrum Handoff
Reactive Spectrum Handoff
Spectrum Mobility (Part III)
Spectrum Decision (Part II)
Spectrum Sharing (Part IV)
Interference-Avoiding Spectrum Sharing
Latency Analysis
MAC Protocols Design
Load-Balancing Spectrum Decision

Figure 1.5: Outline of this report.

characterize these modeling issues simultaneously. In order to demonstrate the effectiveness of this model, three illustrative examples are presented in the second. third, and fourth parts as follows. The second part investigates the spectrum decision issue. We determine which channels to probe and transmit in a load-balancing manner. The third part focuses on the spectrum mobility issue. We illustrate how to model the effects of multiple handoffs, where the secondary users can have different operating channels before and after spectrum handoff. The final part considers the spectrum sharing issue. The optimal admission probability for the secondary users is determined to satisfy the interference constraint to the primary users. Furthermore, the latency performance of various MAC protocols is analyzed and two MAC protocols are proposed to enhanced the QoS performance of CR networks.

The remaining chapters of this report are organized as follows. In Chapter 2, we first give a literature survey of the state-of-the-art techniques. Chapter 3 provides an analytical framework to characterize the general channel usage behaviors with multiple handoffs from a macroscopic viewpoint. Based on the proposed analytical framework, Chapter 4 designs system parameters for load-balancing multiuser spectrum decision schemes in CR networks. Furthermore, Chapters 5 and 6 evaluate the latency performance and determine the optimal target channel sequence for the proactive spectrum handoff, respectively. Next, the effect of sensing time on the latency performance of the reactive spectrum handoff is investigated in Chapter 7. In Chapter 8, an admission control mechanism for the secondary users' spectrum sharing is discussed. Chapter 9 analyzes the latency performance of various MAC protocols. Furthermore, two MAC protocols are proposed to enhanced the QoS performance of CR networks in Chapters 10 and 11. Finally, the concluding remarks and some suggestions for future research topics are given in

Chapter 12.

Chapter 2

Background and Literature Survey

In this chapter, we firstly survey related work to the modeling techniques of the connection-based channel usage behaviors with multiple handoffs. Then, the existing spectrum management techniques, which consist of spectrum decision, spectrum mobility, and spectrum sharing, also are discussed.

2.1 Modeling Techniques for Cognitive Radio Networks

Most of the modeling techniques of channel usage behaviors in CR networks can be classified into three categories: the partially observable Markov decision process (POMDP), the two-dimensional Markov chain (TDMC), and the PRP M/G/1 queueing model (QM). However, these models have not simultaneously considered all of the five design features. Table 2.1 classifies the existing modeling techniques, where the signs “ $\circ$ ” and “ $\times$ ” indicate that the proposed model “does” and “does not” consider the corresponding

Table 2.1: Comparison of Various Analytical Models for CR Networks.

Model Name	Multiple Spectrum Handoff	Spectrum Sensing Time	Various Operating Channels	General Service Time	Multiple Secondary Connections
POMDP [20]	○	×	○	×	×
TDMC [21]	○	×	○	×	○
QM [22–24]	○	×	×	○	○
Proposed Model	○	○	○	○	○

feature, respectively.

In [20], the evolutions of the channel usage of the primary network is characterized by a discrete-time Markov chain which has two occupancy states (the busy and the idle states). The framework of partially observable Markov decision process (POMDP) was developed to preselect the best action (target channel) to maximize the immediate reward (expected per-slot throughput) of the decision maker (secondary user) at the next time slot [20]. Unlike [20] considered only the effects of the traffic loads of the primary network, the authors in [21] considered the effects of the traffic loads of both the primary and the secondary users on the statistics of channel occupancy. In [21], the channel usage behaviors of a CR network is modeled by a two-dimensional Markov chain where the two dimensions represent the total numbers of the primary and the secondary users in a CR network, respectively. When the secondary users are interrupted, it is assumed that they can immediately find the idle channel if at least one idle channel exists. Hence, the spectrum sensing time is neglected in this model. Note that the two Markov chain models are suitable for the exponentially distributed service time, and how to extend them to the case with generally distributed service time is not clear.

Some researchers used the PRP M/G/1 queueing model to characterize the spectrum usage behaviors of each channel. For example, the effects of multi-user contention and multiple interruptions on the latency performance of the secondary users' connections were studied in [22–24]. This PRP M/G/1 queueing model assumed that the secondary user must stay on its current operating channel to resume its unfinished transmission when it is interrupted. That is, there is one candidate channel for spectrum handofff and thus the sensing time issue has not been addressed.

2.2 Load-Balancing Spectrum Decision

The load-balancing spectrum decision schemes can be categorized into two methods: the sensing-based spectrum decision and the probability-based spectrum decision. Table 2.2 compares the existing load-balancing spectrum decision schemes. In the following, we discuss the features of these spectrum decision methods in more details.

2.2.1 Probability-based Spectrum Decision

In the literature, many probability-based spectrum decision schemes were proposed to balance the traffic loads of secondary users in multi-channel CR networks, which can be categorized into three types: (1) packet-wise probabilistic (PP) approach [25–29]; (2) game-theoretic (GT) approach [30–32]; and (3) learning automata (LA) approach [33].

- Packet-wise probabilistic spectrum decision approaches [25–29] aim at maximizing the expected throughput of the secondary users at each slot by determining the probability of selecting each channel from the

Table 2.2: Comparison of Various Load-balancing Spectrum Decision Schemes for Cognitive Radio Networks, where PP, GT, and LA stand for the packet-wise probabilistic, game-theoretic, and learning automata approaches, respectively.

	Channel Occupancy Model of a Primary Network		Multiple Interruptions	Sensing Errors
Probability-based Methods	PP	Bernoulli Process [25]	×	×
		Bernoulli Process [26–29]	×	○
	GT	Deterministic Process [30]	×	×
		M/M/1 [31] or M/G/1 [32]	○	×
	LA	General Distribution [33]	○	×
Sensing-based Methods	Deterministic Process [34]		×	×
	Two-state Markov Chain [35]		×	×
	Bernoulli Process [36–38]		×	×
Proposed Model	M/G/1		○	○

pool of candidate channel. Based on busy probability and capacity of each channel, [25] suggested a method to determine the probability for selecting channels on top of p-persistent carrier sense multiple access (CSMA) medium access control (MAC) protocol in a decentralized manner. They claimed that their proposed sub-optimal channel probability assignment can achieve the Nash equilibrium as the number of secondary users tends to infinity. Furthermore, [26, 27] considered the effects of sensing errors in terms of false alarm and missed detection on the throughput of the secondary users in a two-channel system, and proposed a probabilistic channel selection approach to maximize the throughput of the secondary users in each slot while maintaining the latency constraint of the primary users. Moreover, [28] formulated an optimization problem for channel selection probability to maximize the throughput of the secondary users in each slot while maintaining the interference constraint of the primary users when the primary and secondary networks are asynchronous. Unlike [26–28] considered only the case that one single secondary user can select the channel at each time instant, [29] further extended the probabilistic channel selection approach of [26, 27] to the case that multiple users can simultaneously select their operating channels from the pool of candidate channels, and analyzed the throughput of the secondary users based on the probabilistic channel selection approach taking into account of the effects of channel contention as well as sensing errors. Note that the packet-wise probabilistic spectrum decision approaches in [25–29] were executed in a slot-by-slot manner, which may lead to many channel-switching behaviors during each secondary user's link connection. Moreover, it is assumed that the traffic loads of the secondary users are saturated.

Further, the channel occupancy model of a primary network is modeled as a Bernoulli process and thus the length of busy and idle periods are exponentially distributed.

- Game-theoretic approaches were proposed to solve the spectrum decision problem in CR networks [30–32]. Based on the game theory model, each player (secondary user) can decide the best strategy (channel selection probability) to maximize its utility function. [30] proposed a game-theoretic load-balancing approach to find a set of channel selection probabilities so that no secondary user has incentive to unilaterally change his/her action. To converge to such the Nash equilibrium, a best-reply algorithm was designed for each user to calculate each channel's selection probability as well as its transmission duration based on a utility function related to the load-balancing channel selection. Beside the load-balancing issue, [31] suggested that the utility function in the game-theoretic spectrum decision should also incorporate the channel bandwidth and its idle period as well as the cost of spectrum handoff because the spectrum decision procedure must be executed many times due to multiple interruptions. They emphasized that the channel selection game shall be repeated many times to capture the scenario when primary users stochastically activate or deactivate at each epoch. Unlike the pervious work that considered the homogeneous secondary users, [32] assumed that the secondary users can have different priorities. They proposed a dynamic strategy learning algorithm to determine the channel selection strategies that can converge to the Nash equilibrium. Noteworthily, the Nash equilibrium solution of the game-theoretic approach is not necessary the globally optimal solution from the viewpoint of the overall network [39].

- In [33], a learning automata (LA) approach was suggested to determine the channel selection probabilities by exploring the uncertainty of traffic patterns in CR networks. After a huge number of trials, the secondary users can estimate the optimal channel selection probability. However, the problem for this method is its converging speed, especially for a large number of users.

2.2.2 Sensing-based Spectrum Decision

The sensing-based spectrum decision scheme requires scanning all the candidate channels to determine the most suitable operating channel. Thus, the total number of candidate channels significantly affects the overall system time in the sensing-based spectrum decision scheme. In [34–36], the optimal number of candidate channels to maximize the spectrum accessibility and the procedures to determine the optimal set of candidate channels were investigated. Furthermore, the authors in [37, 38] formulated the sequential channel sensing problem as an optimal stopping problem with the objective of maximizing the throughput of the secondary users. They studied when the secondary users shall stop sensing and start transmitting data. Nevertheless, the effects of multiple interruptions from the primary user and the sensing errors for the primary user's occurrence on the overall system time of the secondary users in the CR networks have not been addressed in these existing sensing-based spectrum decision methods.

2.3 Proactive Spectrum Handoff

In order to characterize the multiple handoff behaviors in CR networks, we should consider the three key design features, consisting of (1) generally

Table 2.3: Comparison of Various Proactive Handoff Models.

Model Name	General Service Time	Various Operating Channels	Multiple Secondary Connections
TMC [35, 40–45]	×	○	×
OORP [46–48]	○	○	×
BRP [49]	×	○	×
MMC [50]	×	×	○
QM [22–24, 32, 51–55]	○	×	○
Proposed Model	○	○	○

distributed service time; (2) various operating channels; and (3) queueing delay due to channel contention from multiple secondary connections. Based on these three features, Table 2.3 classifies the existing modeling techniques for the proactive spectrum handoff. In the literature, the modeling techniques for spectrum handoff behaviors can be categorized into the following five types: (1) the two-state Markov chain; (2) the Bernoulli random process; (3) the arbitrary ON/OFF random process; (4) the birth-death process with multi-dimensional Markov chain; and (5) the PRP M/G/1 queueing model. One can observe that the current modeling techniques have not considered all the aforementioned three design features. In the following, we briefly discuss the features of these analytical models for spectrum handoff behaviors.

- **Two-state Markov chain (TMC)**: In [35, 40–45], the evolutions of the channel usage of the primary networks at each channel were characterized by a discrete-time Markov chain which has two occupancy states: busy (ON) and idle (OFF) states. The idle (OFF) state can be regarded as a potential spectrum opportunity for the secondary users.

Note that the Markov chain model is suitable for the exponentially distributed service time, and it is not clear how to extend it to the case with generally distributed service time. In this model, the target channel selection problem in every time slot is modeled as a Markov decision process. According to the channel occupancy state at the current time slot, a decision maker (secondary user) can preselect the best action (target channel) to maximize its immediate reward at the next time slot such as expected per-slot throughput [35, 40–43], expected idle period [44], or expected waiting time [45]. Note that this model belongs to the slot-based modeling technique because the secondary user shall decide its target channel at each time slot. In this scheme, even though the primary users do not appear at the current operating channel, the secondary user may still need to change its target channel, resulting in frequent spectrum handoffs.

- **Arbitrary ON/OFF random process (OORP)**: Unlike [35, 40–45] assumed that the channel usage behaviors of the primary networks have the Markov property, the authors in [46–48] used the ON/OFF random process with arbitrary distributed ON/OFF period to characterize the channel usage behaviors of the primary networks at each channel. It was assumed that the secondary user can estimate the distributions of the ON period and the OFF period based on long-term observations. In each time slot, the secondary user must calculate the expected remaining idle periods of all channels and then will immediately switch to the channel with the longest remaining idle period. This model also belongs to the slot-based modeling technique because the target channel is decided in each time slot.

- **Bernoulli random process (BRP)**: The authors in [49] examined the effects of multiple interruptions from the primary users on the connection maintenance probability in a connection-based environment, where the spectrum usage behaviors of the primary networks on each channel were characterized by a Bernoulli random process. Because both the busy and idle periods of the considered primary networks follow the geometrical distributions, it is more difficult to extend this modeling technique to the cases with other generally distributed service time.

- **Multi-dimensional Markov chain (MMC)**: In [50], the spectrum usage behaviors of both the primary and secondary networks were modeled by the multi-dimensional Markov chain. Each state in the Markov chain indicates the identity number for the serving users and the waiting users for the channel. It was assumed that the secondary user must stay on its current operating channel after the primary user's interruption. This analytical model is suitable for the single channel network, and the issue of different operating channels in multiple handoffs has not been addressed.

- **M/G/1 queueing model (QM)**: Some researchers used the preemptive resume priority (PRP) M/G/1 queueing model to characterize the spectrum usage behaviors in a single-channel CR network. The effects of multi-user sharing and multiple interruptions on the extended data delivery time of the secondary users were studied in [22–24, 32, 51–55]. Note that the authors in [22–24, 32, 51–55] also assumed that the secondary users must stay on the current operating channel to resume their unfinished transmissions when they are interrupted.

To summarize, the first three analytical models, two-state Markov chain, arbitrary ON/OFF random process, and bernoulli random process, did not incorporate the effects of the traffic loads of the secondary users on the statistics of channel occupancy. How to extend these models to consider the queueing delay due to channel contention from multiple secondary connections is unclear. The last two models, multi-dimensional Markov chain and M/G/1 queueing model, can characterize the effects of spectrum sharing between multiple secondary users. However, these two models assumed that the interrupted secondary user must stay on the current operating channel. and have not dealt with the handoff interaction issue among different channels.

2.4 Optimal Proactive Spectrum Handoff

In the literature, some predetermined target channel selection methods for spectrum handoffs have been proposed and can be categorized into two kinds: probability-based channel selection methods and Markov decision process.

- In [26,27,29], the probability-based channel selection methods were developed to predetermine the probability that each channel is selected to the target channel. Based on the predetermine probabilities, the optimal channel hopping sequence can be decided in packet-by-packet or slot-by-slot manners. The work in [26, 27] designed of the optimal channel hopping sequence in the single-user case, while [29] extended the similar problem to the multiple user case. The above approaches for channel hopping sequence design are optimal in the sense of maximizing the per-slot throughput. However, the latency issue in spectrum handoff has not been considered yet. Clearly, the cumulative delay in one connection due to multiple spectrum handoffs is an important QoS performance measure for CR networks.

- Besides the probability-based channel selection methods, another kind of target channel selection approach is to apply the theory of Markov decision process. In [35, 40–45], the target channel selection problem in every time slot is modeled as a Markov decision process. According to the channel occupancy state at the current time slot, a decision maker (secondary user) can preselect the best action (target channel) to maximize its immediate reward at the next time slot. The considered reward or objective function includes the expected per-slot throughput [35, 40–43], expected idle period [44], and expected waiting time [45]. However, only the effects of channel usage behaviors of the primary users are considered on the channel occupancy. In fact, the traffic loads of the secondary users are also needed to be considered in channel selection.

2.5 Reactive Spectrum Handoff

A key property of reactive spectrum handoff is that the interrupted secondary user can actually find the idle if at least one idle channel exists at the moment of link transition. In order to characterize the channel usage behaviors with this property, we should consider the three key design features, consisting of (1) heterogeneous arrival rates of the primary users (PUs); (2) various arrival rates of the secondary users (SUs); (3) handoff processing time. Based on these three features, Table 2.4 classifies the existing modeling techniques for the reactive spectrum handoff. In the literature, the modeling techniques for spectrum handoff behaviors can be categorized into the following four types: (1) ON/OFF random process; (2) M/M/m queueing Model; (3) two-

Table 2.4: Comparison of Various Channel Usage Models.

Model Name	Heterogeneous Arrival Rates of PUs	Various Arrival Rates of SUs	Handoff Processing Time
OORP [56, 57]	×	×	×
OORP [49]	×	×	○
M/M/m [58]	×	×	×
MDMC [21, 59–72]	×	×	×
M/G/1 [73, 74]	○	○	×
Proposed Unifying Model	○	○	○

dimensional Markov chain; and (4) M/G/1 queueing model. One can observe that the current modeling techniques have not considered all the aforementioned three design features. In the following, we briefly discuss the features of these analytical models for spectrum handoff behaviors.

- **ON/OFF random process (OORP)**: In [49, 56, 57], the ON/OFF random process was used to characterize the channel usage behaviors of the primary networks at each channel, where the distributions of ON (busy) period- and OFF (idle) period at each channel are geometrical distributed. The OFF state can be regarded as a potential spectrum opportunity for the secondary users. The authors in [56] and [57] investigated the channel utilization factors and the extended data delivery time of the secondary users, respectively. Unlike [57] that did not address the effects of spectrum sensing time, the authors in [49] examined the effects of spectrum sensing time on the extended data delivery time of the secondary users. However, [49] assumed that at least one channel is certainly available after spectrum sensing, and the case that all channels are busy after spectrum sensing did not been considered.

- **M/M/m queueing model**: In [58], the channel usage behaviors of the primary users are characterized by the M/M/m queueing model, where m is the total number of channels in the CR network. The author in [58] calculated the handoff delay of the secondary users. However, it is assumed that the handoff delay only results from the waiting time which is the duration from the instant that interruption event occurs until the instant that one idle channel is found. The sensing time had not been considered when calculating handoff delay.

- **Multiple-dimensional Markov chain (MDMC)**: In [21,59–69], the spectrum usage behaviors of both the primary and secondary networks were modeled by a two-dimensional Markov chain, where the two dimensions represent the total numbers of the primary and the secondary users in a CR network, respectively. The blocking probability and forced termination probability for the secondary users' connections in the CR network without and with queue are studied in [21, 59–64] and [65–67], respectively. Different from [21, 59–67] that considered infinite user population, [68, 69] derived the blocking probability in a CR network with finite user population. Furthermore, the authors in [70–72] further extended the two-dimensional Markov chain model to the multiple-dimensional Markov chain, where the new dimension is used to describe the channel state or queue length. Note that these analytical models are suitable for the CR network with homogeneous traffic loads, and the issues of heterogeneous arrival rates of the primary and the secondary users has not been addressed.

- **M/G/1 queueing model**: [73,74] used the M/G/1 queueing model to characterize the channel usage behaviors of a secondary network, where each secondary user can simultaneously use all idle channels to transmit its data. Because the total number of idle channels depends on how many channels are occupied by the primary users, the service rates of the secondary users are related to the traffic statistics of the primary users, which results in a non-trivially distributed service time. Thus, the authors suggested using the M/G/1 queueing system to characterize this system. However, authors did not show how to obtain this non-trivial service time distribution.

2.6 Interference-Avoiding Spectrum Sharing

In order to determine the optimal admission probability, we should consider four key design features: (1) interference on the primary users (PUs), where the transmission of the primary users may be stained by the secondary users due to missed detection; (2) channel contention between multiple secondary users (SUs), where channel contention will increase waiting time of the secondary users; (3) multiple handoffs, a secondary user may have multiple handoffs due to multiple interruptions from the primary users during its transmission period; and (4) generally distributed service time, where the probability mass functions (pmfs) of service time of the primary and secondary connections can be any distributions. Based on these four design features, Table 2.5 classifies the existing admission control techniques.

2.6.1 Admission Control with Perfect Sensing

- **Network-throughput-oriented approach**: The authors in [6,25,50] determined the optimal admission probability to maximize the overall

Table 2.5: Comparison of Various Objective Functions.

Objective Function	Interference on PUs	Channel Contention between SUs	Multiple Spectrum Handoffs	General Traffic Loads
Network Throughput [6,25,50]	×	○	×	×
User Throughput [55]	×	×	○	○
Dropping Probability [63]	×	○	×	×
Transmission Latency [70]	×	○	×	×
Network Throughput [26,27]	○	×	×	×
User Throughput [75]	○	×	×	×
The Proposed Model	○	○	○	○

throughput of a secondary network. They found that a larger admission probability can enhance the per-secondary-user throughput. However, it results in more competition between the secondary users, thereby degrading the overall network throughput. In [6,25], the authors provided the analytical approaches to determine the admission probability p on top of p-persistence carrier sense multiple access (CSMA) medium access control (MAC) protocol. They assumed that the secondary users have the saturated traffic load, i.e., the secondary users always have data to send. Unlike [6,25] that assumed only one secondary user can transmit data at one slot, [50] considered the case that multiple secondary users can transmit data simultaneously. The interference from the coexistence of multiple secondary users is incorporated in the proposed analytical model, where each secondary user's service time is exponentially distributed.

- **User-throughput-oriented approach**: [55] determined the optimal payload length of the secondary users to maximize the throughput of each secondary user. They assumed that the interrupted secondary users must retransmit whole data connection instead of resume the unfinished transmission. They found that the longer payload length can increase per-secondary-user throughput when header length is a constant, but it also increase the interrupted probability, thereby degrading per-secondary-user throughput. Then, a preemptive repeat priority queueing model was proposed to solve this tradeoff issue.

- **Dropping-probability-oriented approach**: In [63], the optimal arrival rate (admission probability) is determined to minimize the dropping (or forced termination) probability of the secondary users by formulating it as a nonlinear optimization problem with the constraint of given the blocking probability for the secondary users. In order to derive the closed-form expressions for the dropping and blocking probabilities in terms of arrival rate, a two-dimensional Markov chain is used to characterize the channel usage behaviors of a CR network, where the two dimensions represent the total numbers of the primary and the secondary users, respectively. Note that this Markov chain model is suitable for the exponentially distributed service time, and how to extend it to the case with generally distributed service time it is not clear.

- **Latency-oriented approach**: In [70], authors proposed a Markov chain model to characterize channel usage behavior of both the primary users and the secondary users. Based on this model, the transmission latency of the secondary users under various arrival rates of the primary users and the secondary users can be evaluated. Hence, when the arrival rate of the primary users is given, the maximum arrival rate of the secondary users can be determined to satisfy the transmission latency requirement of the secondary users.

2.6.2 Admission Control without Perfect Sensing

- **Network-throughput-oriented approach**: [26, 27] considered the effects of sensing errors in terms of false alarm and missed detection on the throughput of the secondary users in a two-channel system, and determined the optimal channel access probability for each channel to maximize the throughput of the secondary users in each slot while maintaining the latency constraint of the primary users. [26,27] considered only the case that single secondary user can select the channel at each time instant and assumed that the secondary users have infinite amount of data to transmit.

- **User-throughput-oriented approach**: [75] further extended the analytical model of [26,27] to determine the optimal false alarm probability to maximize the throughput of each secondary users, while maintaining the latency constraint of the primary users. They found that a lower false alarm probability can enhance per-secondary-user throughput. However, a lower false alarm probability results in higher missed detection probability and thus increasing interference to the primary users. Hence, an optimal false alarm probability exists.

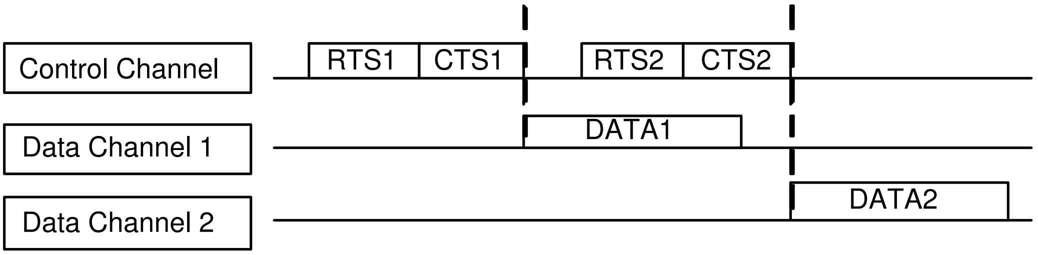


Figure 2.1: Dedicated control channel approach

2.7 Latency Analysis for Spectrum Sharing

We summarize some current multi-channel MAC protocols. Then, we discuss their operation process and drawbacks.

2.7.1 Dedicated Control Channels Method

The typical operation process for the multi-channel MAC protocols with dedicated control channels [76–81] is shown as Fig.2.1. All nodes have assigned a common (or some) dedicated control channel(s) beforehand. Thus, all nodes can exchange the spectrum usage situation and control packets (request-to-send (RTS) and clear-to-send (CTS)) on common spectrums. Then, after exchanging some control messages, transmitter and receiver can know which channel is the desired data channel and can communicate on it.

Assigning a dedicated common control channel can give each node a common communication platform and decrease the complexity of spectrum access. However, this operation process may cause the control channel saturation problem (we will discuss in next section) and increase the access latency. Furthermore, this operation process is suitable for CR environment. Thus, this operation can be considered when design a CR MAC protocol.

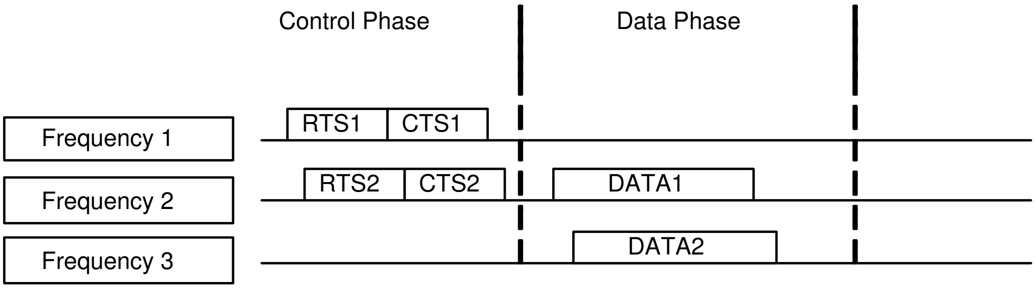


Figure 2.2: Split phase approach

2.7.2 Common Operation Period Method

Another kind of MAC protocols with embedded control channels is shown in [82–85]. The basic idea of this method is to divide total time period into two operation phases: control phase and data phase as shown in Fig. 2.2. In the control phase, all nodes can exchange control messages on all channels. In data phase, all nodes can transmit and receive the data frames on the desired data channels.

The authors in [86] propose a multi-channel MAC with hybrid control channels. In this protocol, the control frames (e.q., RTS,CTS) can be transmitted both on control channel and data channels. Furthermore, the control frames can be exchanged not only on control phase, but also on data phase. Thus, this method can increase the channel utilization.

The advantage for this method is no control channel saturation problem. However, this method need all nodes to be synchronized. Furthermore, this method needs to have the operation schedule beforehand. However, this is hard to achieve in CR networks because the spectrum availability is changing from time to time. Thus, nodes are hard to have a transmission schedule beforehand.

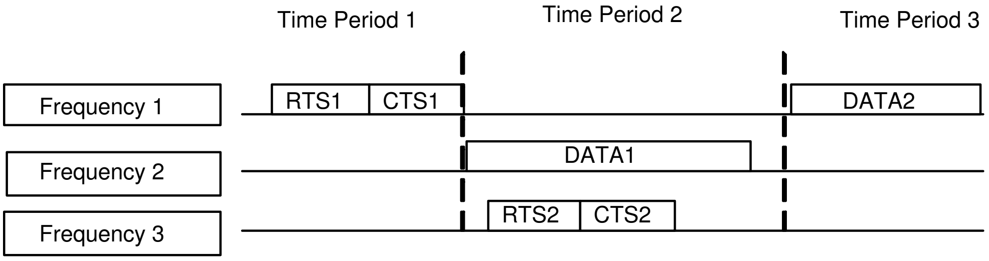


Figure 2.3: Common hopping approach

2.7.3 Common Hopping Sequence Method

Another kind of multi-channel MAC protocol is shown in [87–89]. The basic idea for this method is to use a common hopping sequences to do channel switch as shown in Fig.2.3. Each node uses a hopping sequence to do channel hopping and can communicate with another node once switch to the same frequency band.

This operation process can ignore the control channel saturation problem. However, this method needs all nodes to be tightly synchronized. Furthermore, it needs to have a schedule beforehand and is not suitable for the time-varying CR networks.

2.7.4 Embedded Control Channels Method

We introduce a CR MAC with the embedded control channels in Fig.2.4. There are two candidate data channels CH 1, CH 2 in this system and CH 2 has the largest idle period.

1. At T1 - In the beginning of the procedure, SU 2 doesn't have data to transmit and its default channel CH 3 is busy. Thus, SU 2 switches to the idle channel CH 1 even though the CH 2 has the longest idle period.

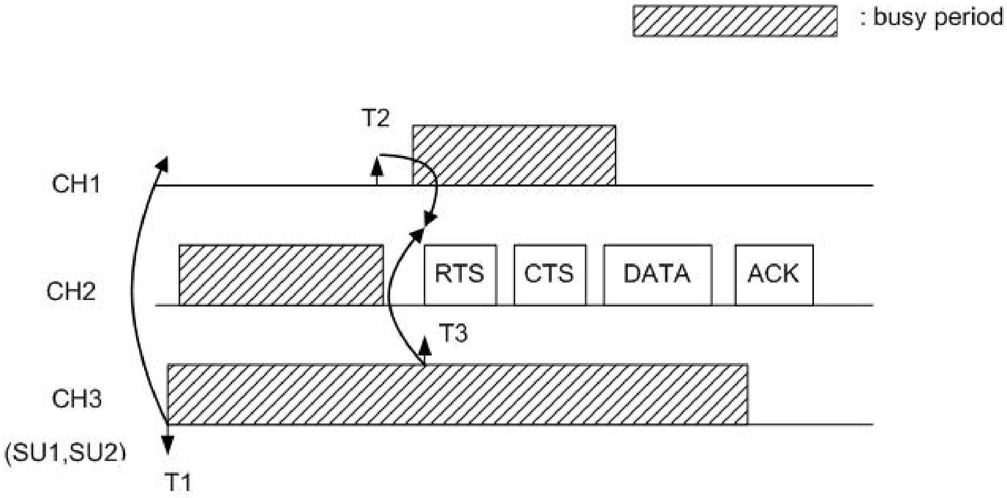


Figure 2.4: A scenario uses embedded control channel

2. At T2 - The channel with the longest idle period CH 2 becomes idle. Thus, SU 2 switches to CH 2 according to the rule of choosing the channel with the longest idle period.

3. At T3 - SU 1 has data to transmit to SU 2 at this moment. Firstly, SU 1 switches to SU 2's default channel CH 3 and finds that CH 3 is busy. Thus, SU 1 switches to the channel with the longest idle period CH 2. Then, SU 1 transmits RTS to SU 2 on CH 2 and SU 2 replies CTS to SU 1. Finally, SU 1 can transmit data to SU 2 successfully.

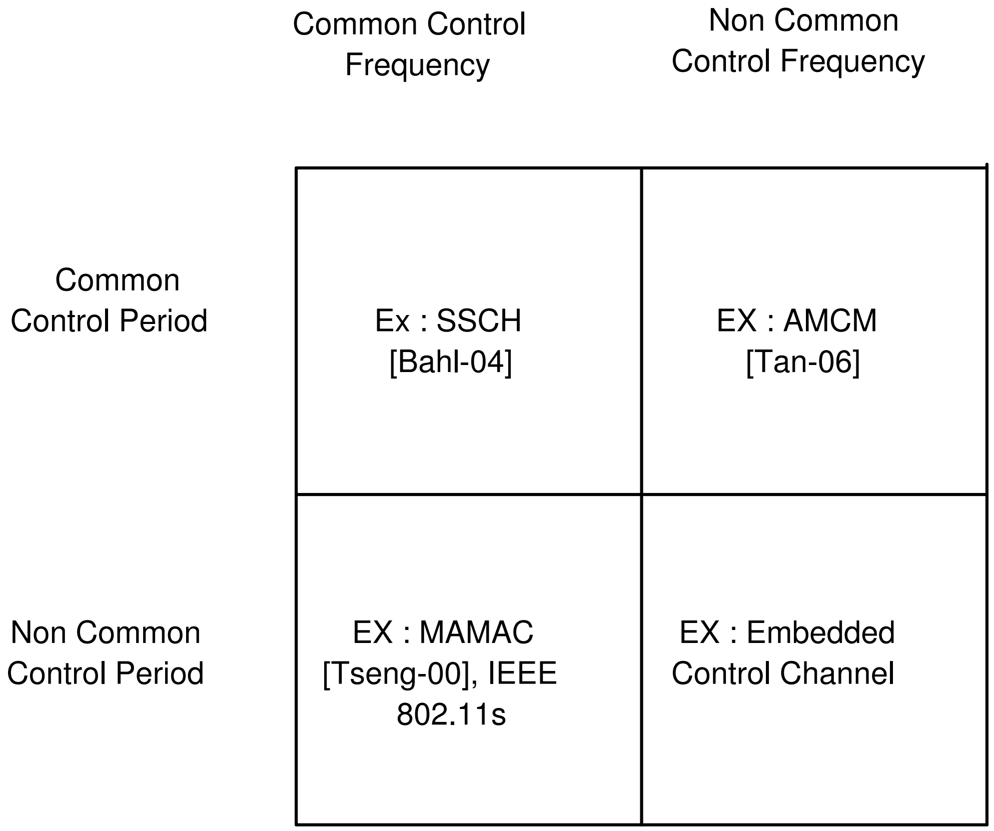
The advantage for this method is without control channel saturation problem. However, this method may have some problems (e.q, channel mismatch problem that we will be discussed in the next section) and increase the access delay. Furthermore, this method can be used in CR networks because of no transmission schedule beforehand.

2.7.5 Summary

Basically, the design of control channels in multi-channel networks can be categorized into four kinds as shown in Figure 2.5: (I) common control frequency, non common control period (ex: dedicated control channel [76, 90–92]); (II) non common control frequency, common control period (ex: dedicated control period [82, 93, 94]); (III) common control frequency, common control period. (ex: common hopping sequence [87, 95]); (IV) non common control frequency, non common control period. (ex: embedded control channel). Because we focus on CR networks, Types II, III are not discussed in this thesis. Thus, we only focus on Type I (called the dedicated control channel) and Type IV (called the embedded control channel) in this thesis.

2.8 Concurrent Transmission for Spectrum Sharing

In general, the goal of concurrent transmission can be achieved by the physical-layer or the MAC-layer techniques [96]. From the aspect of physical layer, the power control mechanism has been used to avoid interference between the concurrent transmission links [8,97–100]. In addition, [24,101,102] tuned the physical carrier sensing threshold to increase the opportunity of spatial reuse. Next, the rate adaptation approach has been proposed to improve the signal to interference and noise ratio (SINR) of communication links [103,104]. Thus, more links can concurrently transmit. Furthermore, the multiple-input and multiple-output (MIMO) and the directional antenna techniques have been used to cancel interference from other active links [105] and avoid interference to other active links [106], respectively.



	Common Control Frequency	Non Common Control Frequency
Common Control Period	Ex : SSCH [Bahl-04]	EX : AMCM [Tan-06]
Non Common Control Period	EX : MAMAC [Tseng-00], IEEE 802.11s	EX : Embedded Control Channel

Figure 2.5: Summary table for multi-channel MAC protocols

2.9 Cognitive MAC Protocol for Spectrum Sharing

The carrier sense multiple access with collision avoidance (CSMA/CA) MAC protocol is a well known MAC protocol for resolving the contention issue in many existing wireless networks, such as the IEEE 802.11 wireless local area network (WLAN) [16]. Since the CSMA/CA MAC protocol requests a station to sense the channel usage before transmissions, this kind of MAC protocol has potential and is natural to become a strong candidate for the CR networks. Thus, we are motivated to ask two natural and fundamental questions: Can the CSMA/CA MAC protocol be directly adopted for CR networks without changes? If not, how should the CSMA/CA MAC protocol be modified to become an appropriate cognitive MAC protocol?

To answer the above questions, we examine the CSMA/CA MAC protocol in the four stages of the cognition cycle one by one. First, from the viewpoint of the ***observe*** stage, the cognitive MAC protocol is required to record the spectrum usage time of the primary users and to collect the traffic characteristics, such as the delay-sensitive or non-real-time data traffic. For the CSMA/CA MAC protocol, most research results currently focus on either sensing the carrier transmission in the surrounding environment or avoiding interference [16, 107–109], instead of identifying the interference. Thus, the functions of recording the spectrum usage and traffic characteristics are not fully considered in the current CSMA/CA MAC protocols.

Second, in the ***plan*** stage of the cognition cycle, the cognitive MAC protocol shall determine whether the requested frame transmission from a CR device will interfere the primary user's connection. Because the cognitive MAC protocol will only permit a CR device to utilize the spectrum of the existing legacy system in the spare time of the primary user, the channel usage for the CR users shall be more efficient and effective than for the primary users. Furthermore, a cognitive MAC protocol shall differentiate the priority for various traffic types with QoS provisioning. In [109–113], the authors suggested adjusting the transmission probability, e.g. by different contention window (CW) sizes and different length of black bursts, to differentiate the traffic types. However, the induced fairness problem may not be well handled [114–117].

Third, in the ***decide*** stage, the cognitive MAC protocol needs to schedule the frame transmissions for the CR users while satisfying the QoS requirements of delay-sensitive traffic. In the previous works [118–123], the authors suggested certain mechanisms to reserve time slots for the delay-sensitive frames. However, such a reservation method requires a polling process or handshaking procedure to coordinate frame transmissions. Both methods consume battery energy and waste the valuable bandwidth in sending management frames. How to design a *distributed* mechanism to reserve the transmissions for high priority frames becomes an important issue for the CR networks.

At last, in the ***act*** stage of the cognition cycle, the cognitive MAC protocol needs to synchronize the stations and execute the transmission at the specified time. To synchronize the clock of each station, the methods designating a centralized controller to broadcast "*beacon*" signals or utilizing the global clock provided by Global Positioning System (GPS) were suggested in [16, 122, 123]. However, both methods are complex and require additional devices, which may not be appropriate for the CR networks.

To our knowledge, we have not seen too many papers in the literature to discuss how to design a complete cognitive MAC protocols based on the specific functions in the four stages of the cognition cycle . In this paper, we propose such a cognitive MAC protocol with QoS provisioning for the CR networks. The proposed MAC protocol is extended from our previous works [124,125] and coincides with the four stages of the cognition cycle. The neighbor list establishment mechanism maps to the ***observe*** stage, which can help a CR user recognize the channel usage around its neighborhood. Corresponding to the ***plan*** stage, the improved contention resolution mechanism can enhance the overall performance in terms of throughput, access delay and fairness over the legacy CSMA/CA MAC protocol. In the ***decide*** stage, the newly proposed reservation scheme ensures the established ad hoc link with satisfactory QoS requirements. The novel distributed frame synchronization mechanism coordinates the transmissions among stations in the ***act*** stage. In the following sections, we detail our proposed cognitive MAC protocol.

Chapter 3

Queueing-Theoretical Modeling Techniques for Cognitive Radio Networks

Basically, according to the principle of selecting the target channel for spectrum handoff, the operating mode of the secondary networks can be categorized as either the non-hopping mode or the hopping mode. The secondary user always stays on its current operating channel when it is interrupted in the non-hopping mode, which is the basic mode of IEEE 802.22 systems [126]. In the hopping mode, the interrupted secondary user can either stay on its current operating channel or change to another channel, which is determined according to traffic statistics. An example of hopping mode is the phase-shifting hopping method of the IEEE 802.22 systems [126]. As a result, the non-hopping mode is a special case of the hopping mode. Note that the secondary users' connection may execute multiple handoffs during its transmission period due to the interruptions from the primary users [127]. Clearly, these handoffs will degrade the quality of service (QoS) performances of the latency-sensitive traffic for the secondary users.

In this chapter, in order to evaluate the QoS performances of spectrum management techniques in the non-hopping or the hopping modes, an analytical framework based on the preemptive resumption priority (PRP) M/G/1 queueing theory is developed. The proposed analytical framework can provide important insights into the system parameter design of spectrum management techniques and can be adapted more flexibly for various traffic arrival rates and service time distributions. Finally, we also provide some new research directions and open issues on top of this model.

3.1 Motivation

Although it is conceptually simple, the proposed PRP M/G/1 queueing network model faces the new challenges. Specifically, it is needed to consider the general behaviors of the connection-based channel usage, including the following key design features: (1) multiple interruptions and handoffs; (2) spectrum sensing time; (3) various operating channels, the operating channels before and after spectrum handoff can be different; (4) generally distributed service time, where the probability mass functions (pmfs) of service time of the primary and the secondary connections can be any distributions; and (5) channel waiting time due to multiple secondary connections' contention.

How to find a unifying analytical model to characterize the effects of the five key design features is important. Because there are many approaches for target channel selection in the hopping model, it is challenging to find a unifying model to facilitate the QoS performance evaluation for various spectrum management techniques. Based on the proposed analytical framework, we provide a systematic approach to catch the randomness property of the target channel selection and can evaluate its effects on system performance metrics such as transmission latency and channel utilization. Many analytical models have been proposed to characterize the channel usage behaviors in the CR networks [20–24]. However, these five key design features have not been considered simultaneously. To our knowledge, an analytical model which is integrated with these design features has rarely been discussed in the literature.

3.2 Transmission Processes with Multiple Handoffs for the Secondary Users' Connections

Figure 3.1 illustrates the transmission processes of a secondary connection in a two-channel CR network. The procedures consist of the following steps:

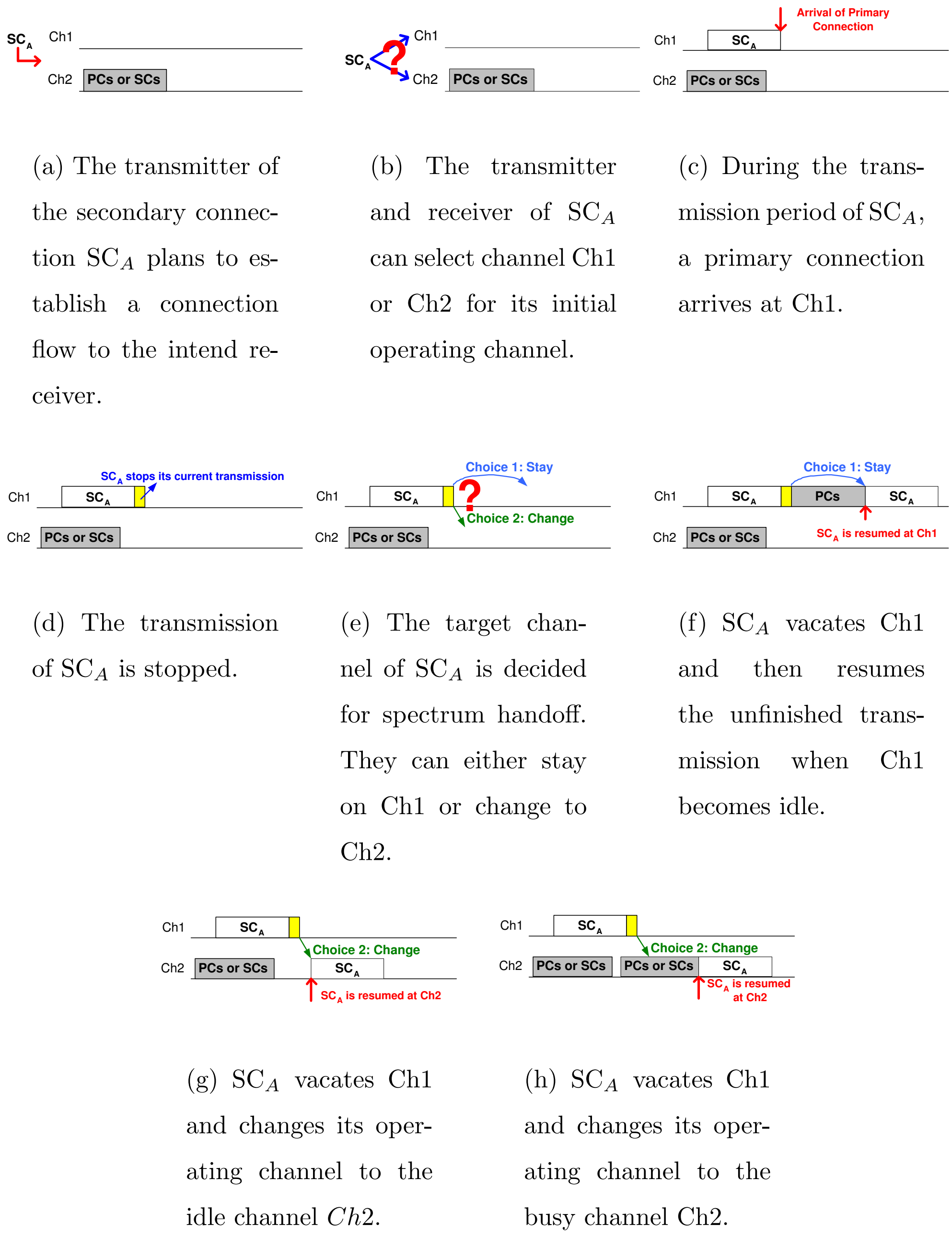
1. In Fig. 3.1(a), a secondary user plans to establish a new connection flow SC$_A$ to its intend receiver.

2. Next, in Fig. 3.1(b), the transmitter and receiver of SC$_A$ select the initial operating channel. In this example, they can select channel Ch1 or Ch2.

3. In Fig. 3.1(c), SC$_A$ is established at Ch1. During the transmission period of SC$_A$, a request of primary connection may arrive at Ch1.

4. Next, in Fig. 3.1(d), SC$_A$ detects the appearance of the primary user[1].

[1] In this report, we assume that the considered CR network is a time-slotted system. In order to detect the presence of primary users, each secondary user must perform spectrum sensing at the beginning of each time slot. If the current operating channel is idle, the secondary user can transmit one slot-sized frame in this time slot. Otherwise, the secondary user must perform spectrum handoff procedures to resume its unfinished transmission at



(a) The transmitter of the secondary connection SC_A plans to establish a connection flow to the intend receiver.

(b) The transmitter and receiver of SC_A can select channel Ch1 or Ch2 for its initial operating channel.

(c) During the transmission period of SC_A, a primary connection arrives at Ch1.

(d) The transmission of SC_A is stopped.

(e) The target channel of SC_A is decided for spectrum handoff. They can either stay on Ch1 or change to Ch2.

(f) SC_A vacates Ch1 and then resumes the unfinished transmission when Ch1 becomes idle.

(g) SC_A vacates Ch1 and changes its operating channel to the idle channel $Ch2$.

(h) SC_A vacates Ch1 and changes its operating channel to the busy channel Ch2.

Figure 3.1: Illustration of transmission procedures in a two-channel system. The gray areas indicate that the channels are occupied by the existing primary users' connections (PCs) or the other secondary users' connections (SCs).

Then, the spectrum handoff procedures are initiated to vacate Ch1 and discover the suitable target channel to resume the unfinished transmission.

5. Then, in Fig. 3.1(e), the target channel of SC$_A$ must be decided for spectrum handoff. If the non-hopping mode is adopted, the operating channel of SC$_A$ cannot be changed and thus SC$_A$ must select Ch1 to be its target channel. However, SC$_A$ can select Ch1 or Ch2 for its target channel when the hopping mode is adopted. There are many methods to select the target channel. For example, the target channel can be searched by instantaneous spectrum sensing at this moment of interruption. In this case, the effect of spectrum sensing time τ on the latency performance of SC$_A$ must be considered.

6. Finally, if SC$_A$ chooses to stay on Ch1, its remaining transmission will be resumed after all traffic loads of the primary users at Ch1 have been served as shown in Fig.3.1(f). On the other hand, if the decision is change to Ch2, there are two possible situations. If Ch2 is idle, SC$_A$ can transmit remaining data immediately as shown in Fig.3.1(g). Otherwise, SC$_A$ must wait at the queue until all secondary users in the present queue of Ch2 are served as shown in Fig.3.1(h).

7. Note that the similar spectrum handoff behaviors may be executed many times because a secondary connection may experience multiple interruptions from the primary users during its transmission period. Hence, the procedures from Figs. 3.1(c) to 3.1(h) will be executed

the target channel. Furthermore, The secondary user can differentiate the appearance of the primary user or the secondary user by exiting spectrum sensing techniques such as feature detection.

repeatedly. In this case, a set of target channels will be selected sequentially, called the *target channel sequence* in this report.

3.3 Queueing Theoretical Framework for Spectrum Management

3.3.1 Assumptions

Assume that the considered CR network is a time-slotted system as [40,128–131]. In order to detect the presence of primary users, each secondary user must perform spectrum sensing at the beginning of each time slot. If the current operating channel is idle, the secondary user can transmit one slot-sized frame in this time slot. Otherwise, the secondary user must perform spectrum handoff procedures to resume its unfinished transmission at the preselected target channel. This kind of listen-before-talk channel access scheme is implemented in many wireless techniques, such as the quiet period of the IEEE 802.22 standard [132] and the clear channel assessment (CCA) of the IEEE 802.11 standard [133].

3.3.2 Overview of the PRP M/G/1 Queueing Network Model

Now we propose a preemptive resume priority (PRP) M/G/1 queueing network model to characterize the connection-based spectrum usage behaviors in CR networks. This queueing network analytical framework is quite general and can be easily adjusted to evaluate the performance of various spectrum management techniques under different traffic conditions. Furthermore, it can also be applied to general CR network architectures, including ad hoc

CR network and centralized CR networks such as the IEEE 802.22 standard. Key features of the proposed PRP M/G/1 queueing network model are listed below:

- Each server (channel) has two types of customers (connections). Before transmitting data, the traffic of the primary and the secondary users enter to the high-priority queue and the low-priority queue[2], respectively. Then, according to the traffic arrival time at queues, the *primary connections* and the *secondary connections* can be established without any collision. Here, we assume that the connections with the same priority follow the first-come-first-served (FCFS) scheduling discipline[3].
- The primary users have the preemptive priority to interrupt the transmission of the secondary users. The interrupted secondary user can resume the unfinished transmission on the selected target channel, instead of retransmitting the whole data. Note that the target channel of an interrupted secondary connection can be different from its current operating channel. This is a key difference from the spectrum usage model based on the conventional PRP M/G/1 queueing theory [22–24].
- A secondary connection may experience multiple interruptions from the primary users during its transmission period. This model can characterize the effects of multiple spectrum handoffs.

Note that this model can be also extended to characterize the effects of sensing errors (i.e., missed detection and false alarm) and the heterogeneous

[2] Note that we assume the considered two queues have an infinite length.

[3] In fact, the analytical results of mean values obtained based the proposed framework can be applied to other scheduling discipline which is independent of the service time of the primary and secondary connections because the averages of system performance metrics will be invariant to the order of service in this case (see page 113 in [134]).

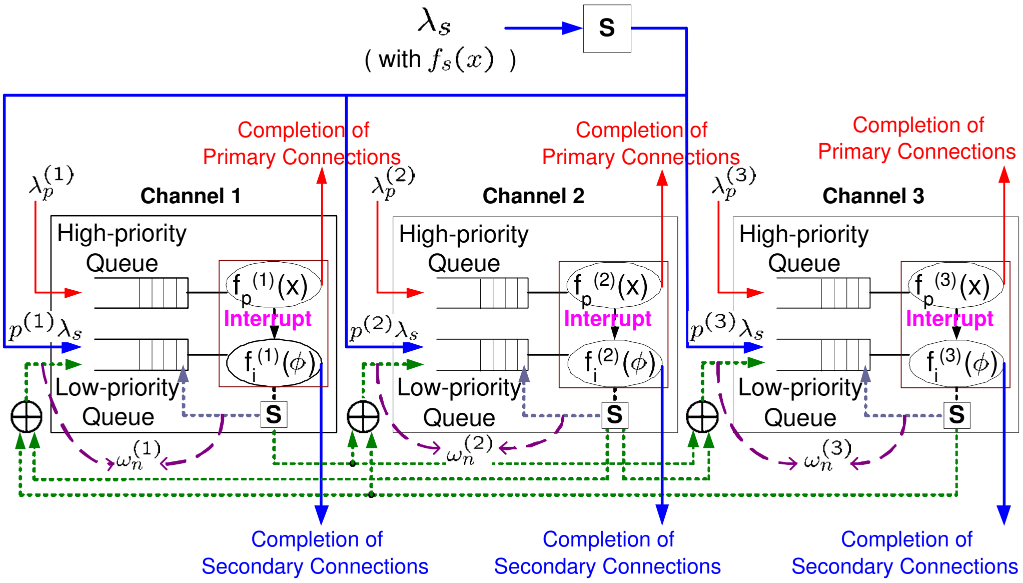


Figure 3.2: The PRP M/G/1 queueing network model with three channels. $\lambda_p^{(k)}$, $\lambda_s^{(k)}$, and $\omega_n^{(k)}$ are the arrival rates of the primary connections, the secondary connections, and the type-n secondary connections ($n \geq 1$) at channel k. Note that $\omega_0^{(k)} = \lambda_s^{(k)}$. Furthermore, $f_p^{(k)}(x)$ and $f_i^{(k)}(\phi)$ are the pmfs of $X_p^{(k)}$ and $\Phi_i^{(k)}$, respectively.

channel bandwidth [135]. Some assumptions are adopted for ease of analysis.

- The arrival processes of the primary and the secondary connections are Poisson.
- Only one user can transmit on each channel at any time instant.
- The secondary transmitter can notifies its corresponding receiver of the interruption event by certain spectrum handoff protocols [136].

Figure 3.2 shows an example of the PRP M/G/1 queueing network model with three channels. Let λ_s (arrivals/slot) be the arrival rates of the secondary connections in CR network. When a secondary connection arrives at CR network, it can select its initial operating channel from one of three channels. Let $p^{(k)}$ be the probability that it selects channel k for its initial operating channel. Thus, the effective arrival rate of the secondary connection at channel k is $\lambda_s^{(k)} = p^{(k)}\lambda_s$. Note that various spectrum decision algorithms will yield different values of $p^{(k)}$.

When a newly arriving secondary connection is connected to the low-priority of its initial operating channel, it can be transmitted immediately if the selected channel is idle. Otherwise, it must wait until this channel becomes idle. Furthermore, when a secondary connection is transmitting at channel k, it will be interrupted if a primary user appears at channel k. In this case, the secondary connection can either stay on the current operating channel or change to another channel through different feedback paths. The decision depends on which operating mode and spectrum handoff scheme are adopted. If the secondary connection chooses to stay on its current operating channel, the remaining data of the interrupted secondary connection must wait at the head of the low-priority queue of the current operating channel. If the decision is to change its operating channel, its remaining data will be connected to the tail of the low-priority queue of another channel. Note that $\oplus$ represents that the traffic of the interrupted secondary connection is merged. Furthermore, when the interrupted secondary connection transmits its remaining data on the selected target channel, it may be interrupted again. Hence, this model can describe the effects of multiple handoffs.

In Fig. 3.2, $\boxed{\text{S}}$ represents the channel selection point, where the newly arriving secondary connection must select its initial operating channel or the interrupted secondary connection must select its target channel for spectrum handoff. There are many methods to select these channels. For example, the secondary connection can decide its initial operating channel or target channel according to the predetermined probability or the outcomes from instantaneous spectrum sensing. If the spectrum sensing is executed to search the idle channels, $\boxed{\text{S}}$ can be regarded as a tapped delay line or a server with constant service time, which related to sensing time. Hence, the effect of spectrum sensing time on the latency performance of the secondary connections can be characterized.

3.3.3 Modeling of the Connection-based Channel Usage Behaviors

Now, we explain why the proposed model can characterize the connection-based channel usage behaviors in a CR network. In order to accurately characterize the transmission processes of a secondary connection, we must take the seven events as discussed in Section 3.2 into account.

1. Secondary connection arrival event as shown in Fig. 3.1(a): We assume that the arrival process of the secondary connections is Poisson. Let X_s be the service time of the secondary connections and $f_s(x)$ be the probability mass function (pmf) of X_s.

2. Initial channel selection event of the secondary connections as shown in Fig. 3.1(b): We use $p^{(k)}$ to represent the probability that the secondary connection selects channel k for its initial operating channel. Furthermore, if the spectrum sensing is executed to decide the initial operating channel, the effect of sensing time can be modeled by $\boxed{\text{S}}$.

3. Primary connection arrival event as shown in Fig. 3.1(c): We assume that the arrival process of the primary connections is Poisson. Denote $\lambda_p^{(k)}$ as the arrival rate of the primary connections whose default channels are channel k. Furthermore, let $X_p^{(k)}$ be the service time of the

primary connections whose default channels are channel k and $f_p^{(k)}(x)$ be the pmf of $X_p^{(k)}$.

4. Interruption event as shown in Fig. 3.1(d): In the PRP M/G/1 queueing network model, the primary users have the preemptive priority and thus they can interrupt transmission of the secondary users. In other words, the secondary users must vacate the occupied channel when the primary users appear.

5. Target channel selection event as shown in Fig. 3.1(e): An interrupted secondary connection can either stay on its current channel or change to another channel. To this end, its remaining transmission must be connected to the low-priority queue of current channel or another channel through different feedback paths. Furthermore, if the spectrum sensing is executed to search the target channel, the effect of sensing time can be modeled by $\boxed{\text{S}}$.

6. Resumption event as shown in Figs. 3.1(f)-(h): The interrupted secondary connection can resume its unfinished transmission on the target channel, instead of retransmitting the whole data.

7. Multiple handoff events: Two auxiliary parameters ($\omega_i^{(k)}$ and $\Phi_i^{(k)}$) are suggested to characterize the traffic flows of the interrupted secondary connections.

3.3.4 Two Auxiliary Parameters: $\omega_i^{(k)}$ and $\Phi_i^{(k)}$

In Fig. 3.2, we use two auxiliary parameters to characterize the traffic flows of the interrupted secondary connections. We call the secondary connections which have experienced i interruptions the type-i secondary connections where $i \geq 0$. At channel k, denote $\omega_i^{(k)}$ as the arrival rate of traffic flows redirected from the type-$(i-1)$ secondary connections. That is, $\omega_i^{(k)}$ is the arrival rate of the type-i secondary connections at channel k. Note that $\omega_0^{(k)} = \lambda_s^{(k)}$. Furthermore, let $\Phi_i^{(k)}$ be the transmission duration of a secondary connection between the i^{th} and the $(i+1)^{th}$ interruptions at channel k and $f_i^{(k)}(\phi)$ be the pmf of $\Phi_i^{(k)}$. That is, $\Phi_i^{(k)}$ is the effective service time of the type-i secondary connections at channel k.

Figure 3.3 illustrates the physical meaning of random variable $\Phi_i^{(k)}$. Recall that X_s is the service time of the secondary connections. We generate X_s five times in Fig. 3.3. The five realizations are divided into many segments due to multiple primary users' interruptions. For example, the first secondary connection (realization) is divided into four segments because it experiences three interruptions in total. The first, second, third, and fourth segments are transmitted at channels 1, 1, 1, and 2, respectively. Thus, this secondary connection's initial operating channel is Ch1 and its target channel sequence is (Ch1,Ch1,Ch2). In Fig. 3.3, random variable $\Phi_2^{(1)}$ is one of the gray regions, representing the transmission duration of a secondary connection between the 2^{nd} and the 3^{rd} interruptions at Ch1. That is, $\Phi_2^{(1)}$ is one of the third segments of the first, the third, and the fourth secondary connections in Fig. 3.3. Note that the fifth secondary connection in Fig. 3.3 does not have the third segment because it is interrupted only once.

In the hopping mode, it is quite complex to find the probability mass function of the effective service time of each segment because the effective service time is dependent on the traffic statistics of the primary and other secondary users of each channels and the operating channels for these segments can be different. Fortunately, based on the proposed analytical framework, we provide a systematic approach to study the effects of various system parameters on the effective service time and then can derive the closed-form expression for the probability mass function of the effective service time of each segment.

3.3.5 Constraint

Finally, we denote $\rho^{(k)}$ as the busy probability of channel k. In an M-channel network, the following constraint shall be satisfied:

$$\rho^{(k)} \triangleq \lambda_p^{(k)}\mathbf{E}[X_p^{(k)}] + \sum_{i=0}^{\infty} \omega_i^{(k)}\mathbf{E}[\Phi_i^{(k)}] < 1 \ , \tag{3.1}$$

Note that $\rho^{(k)}$ can be also interpreted as the utilization factor of channel k.

3.4 Summary

In the following chapters, we will discuss various spectrum management techniques to demonstrate the effectiveness of this analytical model. For the *spectrum decision* issue, we show how to determine which channels are required to probe and transmit. For the *spectrum mobility* issue, we illustrate how to characterize the effects of multiple handoffs, where the secondary users can have different operating channels before and after spectrum handoff. For the *spectrum sharing* issue, we explore how to determine the optimal admission probability to avoid the interference between primary and secondary users in the presence of false alarm and missed detection.

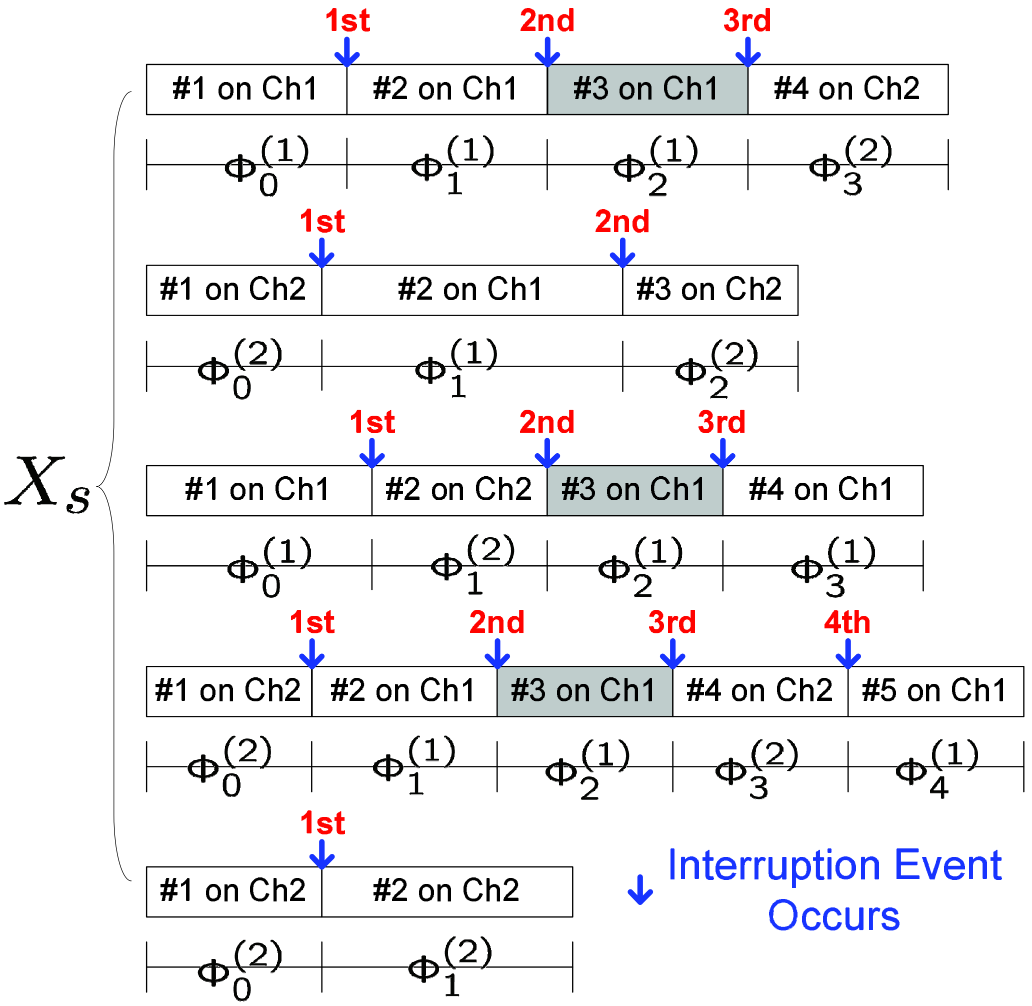


Figure 3.3: Illustration of the physical meaning of random variable $\Phi_i^{(k)}$. For example, $\Phi_2^{(1)}$ is one of the third segments (gray areas) of the first, the third, and the fourth secondary connections.

Chapter 4

Load-Balancing Spectrum Decision

Spectrum decision is a crucial process in CR networks [13], which helps the secondary user select the best channel to transmit data from candidate channels. In order to distribute the traffic loads of the secondary users evenly to these candidate channels, an effective spectrum decision scheme should take the traffic statistics of the primary users as well as the secondary users into account. In this chapter, we introduce a performance measure for evaluating various spectrum decision schemes – the overall system time of the secondary connection, which is defined as the duration from the instant that data arrives at system until the instant of finishing the whole transmission.

In this chapter, we investigate how to evaluate the overall system time for the sensing-based and the probability-based spectrum decision schemes in the CR network when multiple interruptions from the primary user and sensing errors are taken into account. To this end, we design our multiuser spectrum decision schemes on top of the preemptive resume priority (PRP) M/G/1 queueing model. Based on the proposed analysis-based framework, we can design the suitable parameters to shorten the overall system time. Unlike the non-load-balancing methods that multiple secondary users may contend for the same channel, the channel selection schemes based on the designed parameters of the proposed analytical model can evenly distribute the traffic loads of secondary users to multiple channels, thereby reducing the average overall system time. The major contributions of this chapter are summarized in the following:

- Derive the optimal selection probability for the probability-based channel selection scheme.
- Develop a method to determine the optimal number of candidate channels for the sensing-based channel selection scheme.
- Compare the sensing-based and the probability-based channel selection methods and suggest which spectrum decision scheme can result in shorter overall system time with various sensing error probabilities and traffic parameters.
- Characterize the effects of sensing errors on the spectrum decision schemes of CR networks in terms of the overall system time of the primary and the secondary connections.

4.1 Motivation

The overall system time of the secondary users' connections is affected by the multiple interruptions from the primary users and the sensing errors like missed detection and false alarm for the primary users. Within the transmission period of the secondary users' connection, it is likely to have multiple spectrum handoffs due to the interruptions from the primary users. Clearly, multiple spectrum handoffs will increase the overall system time [127]. In the meanwhile, false alarm occurs when the detector mistakenly reports the presence of a primary user. In this situation, the overall system time of the secondary user's connections becomes longer because the secondary users cannot transmit data even with an idle channel. When the detection of a primary user is missed, data collision of both the primary user and the secondary user occurs, resulting in retransmitting and prolonging the overall system time of the secondary users' connections. Hence, it is crucial is incorporating the effects of multiple handoffs and the sensing errors of false alarm and missed detection in spectrum decision methods for CR networks.

In this chapter, two kinds of spectrum decision schemes are considered: (1) the sensing-based spectrum decision scheme; and (2) the probability-based channel selection scheme. For the sensing-based spectrum decision method, a secondary user selects its operating channel according to the *instantaneous* sensing results from scanning the wideband spectrum. For the probability-based spectrum decision method, the operating channel is selected based on the predetermined probabilities which are determined according to traffic statistics from the *long-term* observation. Note that the sensing outcomes in both the methods are related to the traffic statistics of both the primary users and the secondary users. The two considered spectrum decision schemes have different design issues. For the sensing-based spectrum decision scheme, the total number of candidate channels for channel selection significantly affects the overall system time because this scheme requires scanning all the candidate channels. Intuitively, a narrowband sensing (or a smaller number of candidate channels) can reduce the total sensing time. However, it is difficult to find one idle channel from a small number of candidate channels. Hence, one challenge is to determine the optimal number of candidate channels to minimize the overall system time. On the other hand, the probability-based spectrum decision scheme needs to prevent the secondary users from selecting a busy channel. Hence, the most important issue is to determine the optimal channel selection probability to minimize the overall system time.

4.2 System Model

4.2.1 Assumptions

In practice, many reasons may lead to an error on sensing the presence of the primary users. If such an sensing error occurs, not only the primary user's connection will be stained, but the secondary user's transmission will be affected. There are two types of sensing errors regarding the detection of the primary users: false alarm and missed detection. False alarm occurs when the detector reports the presence of a primary user while it is absent, while missed detection occurs when the detector reports the absence of a primary user while it is present. In this chapter, the effects of false alarm and missed detection on CR network performance are discussed in Section 4.5.

4.2.2 Spectrum Decision Behavior Model

Fig. 4.1 illustrates the spectrum decision behavior model, which will be used to evaluate the overall system time of a secondary user's connection for different channel selection schemes. We assume that the arrival processes of the primary and the secondary connections[1] are Poisson. Let $\lambda_p^{(k)}$ (arrivals/slot)

[1] When a secondary transmitter has data to send, how to establish a secondary connection to its intended receiver has been investigated in [137].

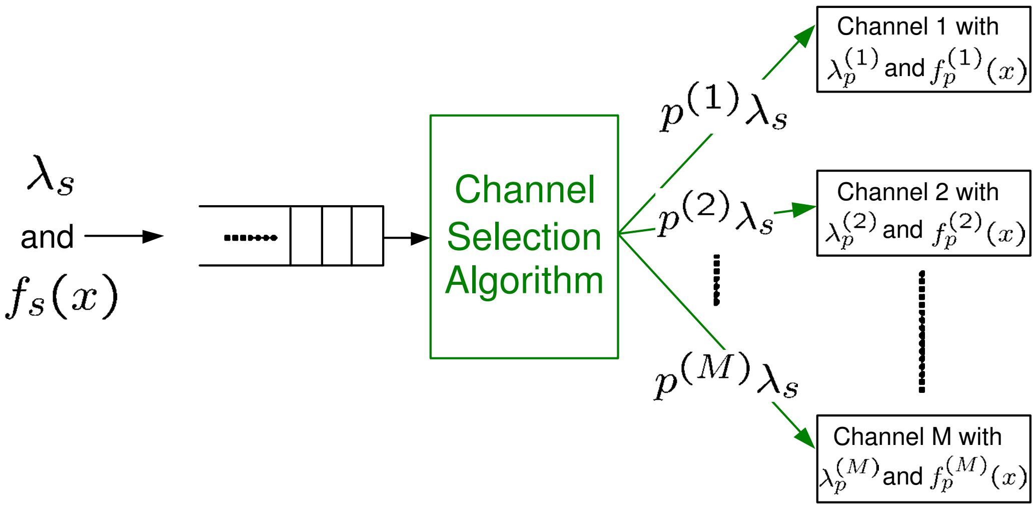

λ_s
and
$f_s(x)$
Channel Selection Algorithm
$p^{(1)}\lambda_s$
$p^{(2)}\lambda_s$
$p^{(M)}\lambda_s$
Channel 1 with $\lambda_p^{(1)}$ and $f_p^{(1)}(x)$
Channel 2 with $\lambda_p^{(2)}$ and $f_p^{(2)}(x)$
Channel M with $\lambda_p^{(M)}$ and $f_p^{(M)}(x)$

Figure 4.1: Spectrum decision behavior model.

and λ_s (arrivals/slot) be the average arrival rates of the primary connections at channel k and the secondary connections of CR network, respectively. Also, denote $X_p^{(k)}$ (slots/arrival) and X_s (slots/arrival) the service time of the primary connections of channel k and the secondary connections, respectively; and let $f_p^{(k)}(x)$ and $f_s(x)$ be the probability mass functions (pmf) of $X_p^{(k)}$ and X_s, respectively. It is assumed that $\lambda_p^{(k)}$, λ_s, $f_p^{(k)}(x)$, and $f_s(x)$, which can be estimated by the existing methods [138], are known to all the secondary users.

As shown in Fig. 4.1, each secondary connection can select one of M candidate channels for its operating channel. Based on our proposed analytical framework, which will be discussed in more detail later, all the secondary users can dynamically select their operating channels with suitable probability that can balance the traffic loads of secondary users in multiple channels. The distribution probability vector (denoted by $\boldsymbol{p} = (p^{(1)}, p^{(2)}, \cdots, p^{(M)})$) represents the set of probabilities for selecting all the candidate channels, in which $p^{(k)}$ denotes the probability of a secondary connection selecting channel

k for its operating channel. Thus, the effective arrival rate of the secondary connection at channel k is $\lambda_s^{(k)} = p^{(k)}\lambda_s$. Note that various channel selection algorithms yield different distribution probability vectors.

4.3 Problem Formulation

4.3.1 Performance Metric: Overall System Time

The overall system time (denoted by S) is an important quality of service (QoS) metric for the connection-based service of the secondary users. It consists of the waiting time (denoted by W) and the extended data delivery time (denoted by T) as shown in Fig. 4.2. Hence, we have

$$\mathbf{E}[S] = \mathbf{E}[W] + \mathbf{E}[T] \ , \tag{4.1}$$

where $\mathbf{E}[\cdot]$ is the expectation function. Here, the waiting time is defined as the duration from the instant that a data transmission request arrives at the system until the instant of starting transmitting data. The duration of waiting time depends on the channel selection scheme that the secondary users adopt. Furthermore, the extended data delivery time is defined as the duration from the beginning of transmitting the data in the first time slot until the completion of the data in the last time slot. Clearly, multiple handoff behaviors significantly affect the extended data delivery time.

4.3.2 Overall System Time Minimization Problem for Probability-based Channel Selection Scheme

For the probability-based channel selection method, each secondary user selects its operating channel from all the M candidate channels based on a

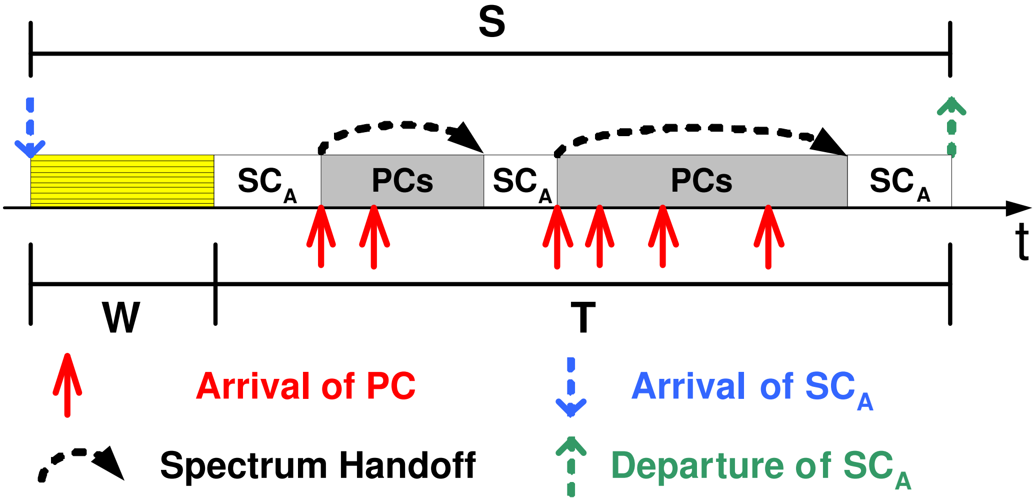


Figure 4.2: Example of the overall system time of the secondary connection SC$_A$. The white areas indicate that channel is occupied by SC$_A$. Furthermore, the gray areas indicate that channel is occupied by the primary connections (PCs) and its duration is the busy period resulting from transmissions of the primary connections. Here, SC$_A$ encounters two interruptions from the primary connections during its transmission period.

predetermined distribution probability vector $\boldsymbol{p}_{pb}$. In this case, an **Overall System Time Minimization Problem for Probability-based Channel Selection Scheme** can be formulated as follows. Given the set of candidate channels $\Omega = \{1, 2, \ldots, M\}$, we aim to find the optimal distribution probability vector (denoted by $\boldsymbol{p}^*$) to minimize the average overall system time of the secondary connections (denoted by $\mathbf{E}[S_{pb}]$). Formally,

$$\boldsymbol{p}^* = \underset{\forall\, \boldsymbol{p}_{pb}}{\arg\min}\, \mathbf{E}[S_{pb}(\boldsymbol{p}_{pb})] \ \ , \tag{4.2}$$

subject to:

$$0 \leq p_{pb}^{(k)} \leq 1, \quad \forall\ k \in \Omega \ \ , \tag{4.3}$$

$$\sum_{k \in \Omega} p_{pb}^{(k)} = \sum_{k=1}^{M} p_{pb}^{(k)} = 1 \ \ . \tag{4.4}$$

and

$$\rho^{(k)} = \rho_p^{(k)} + \rho_s^{(k)} < 1 \ \ , \tag{4.5}$$

where $\rho^{(k)}$ is the busy probability of channel k. Furthermore, $\rho_p^{(k)}$ and $\rho_s^{(k)}$ are the busy probabilities resulting from the primary and the secondary connections at channel k when sensing errors are considered, respectively. In Section 4.4, we will derive the closed-form expressions for $\rho_p^{(k)}$ and $\rho_s^{(k)}$.

4.3.3 Overall System Time Minimization Problem for Sensing-based Channel Selection Scheme

For the sensing-based channel selection scheme, the secondary users perform wideband sensing to find an idle channel from all the candidate channels. If more than one idle channel is found, the secondary user randomly selects one channel from the idle channels for its operating channel. Furthermore, if all the candidate channels are busy, the secondary user still randomly select one channel from all the candidate channels and wait for the available time slot of the selected channel.

In order to decrease the total sensing time, the secondary users shall reduce the number of candidate channels by sensing only the best n channels among M channels. Without loss of generality, we assume that the channel preference of the secondary users follows the lexicographic order. That is, channel i is not better than channel j if $i > j$. Note that the ordering issue for channel preference has been discussed in [139]. Let Ω be the set of candidate channels. Then, we can have $\Omega = \{1, 2, \ldots, n\}$, where $n = |\Omega| \leq M$. Next, we formulate an **Overall System Time Minimization Problem for Sensing-based Channel Selection Scheme** as follows. Given the total number of channels M, we aim to find the optimal number of candidate channels (denoted by n^*) to minimize the average overall system time of the secondary connections (denoted by $\mathbf{E}[S_{sb}]$). Formally,

$$n^* = \arg\min_{1 \leq n \leq M} \mathbf{E}[S_{sb}(n)] \quad . \tag{4.6}$$

4.3.4 Performance Model

In order to calculate the overall system time of various spectrum decision schemes, we extend the general model in Fig. 4.1 to characterize the sensing- and the probability-based channel selection schemes. Fig. 4.3 shows the performance model for the probability-based scheme. When the traffic of the secondary user (i.e., the secondary connection) arrives at the system, it can be directly connected to the selected channel based on the predetermined distribution probability vector. On the other hand, Fig. 4.4 shows the performance model of the sensing-based scheme. When the traffic of a secondary user arrives at the system, the secondary user performs spectrum sensing to find idle channels. The total sensing time can be modeled by a tapped delay

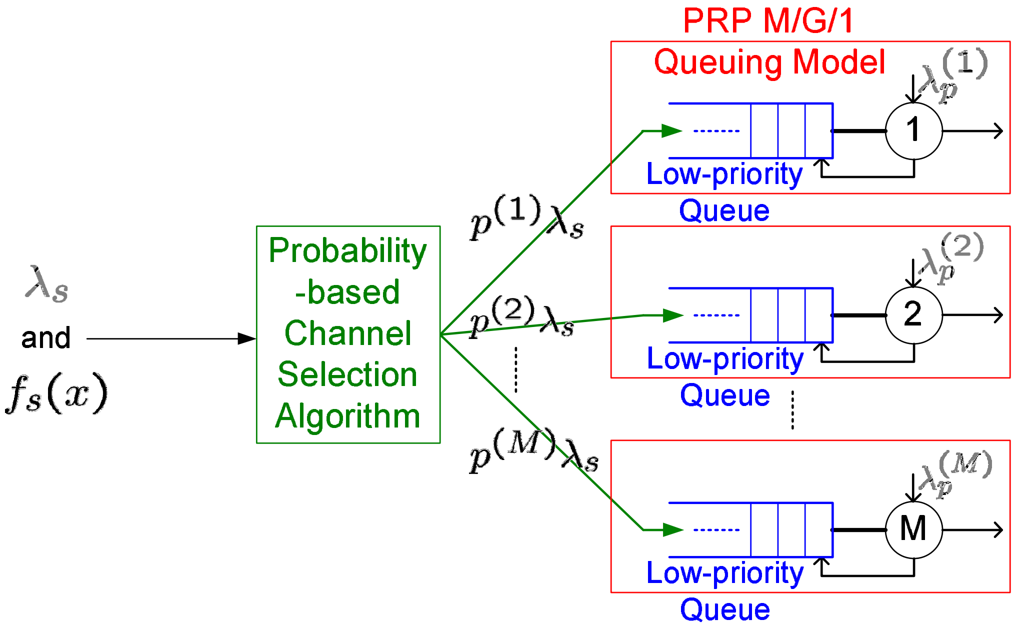


Figure 4.3: Performance model for the probability-based channel selection scheme where the channel usage behaviors are characterized by the PRP M/G/1 queueing systems.

line $\boxed{\text{S}}$. In this case, $\boxed{\text{S}}$ can be regarded as a server with constant service time, which equals to sensing time. If an idle channel can be found, the secondary connection can be served immediately. Finally, in Figs. 4.3 and 4.4, the channel usage behaviors of each channel is characterized by a PRP M/G/1 queueing model, which had been presented in Chapter 3. Here, we assume that the non-hopping mode is adopted. Hence, the secondary user must stay on its current channel when it is interrupted.

Based on the proposed performance models, we can analytically compare the overall system time resulting from both the spectrum decision schemes for various sensing time and traffic parameters. Then, each secondary user can intelligently adopt the best channel selection scheme to minimize its overall

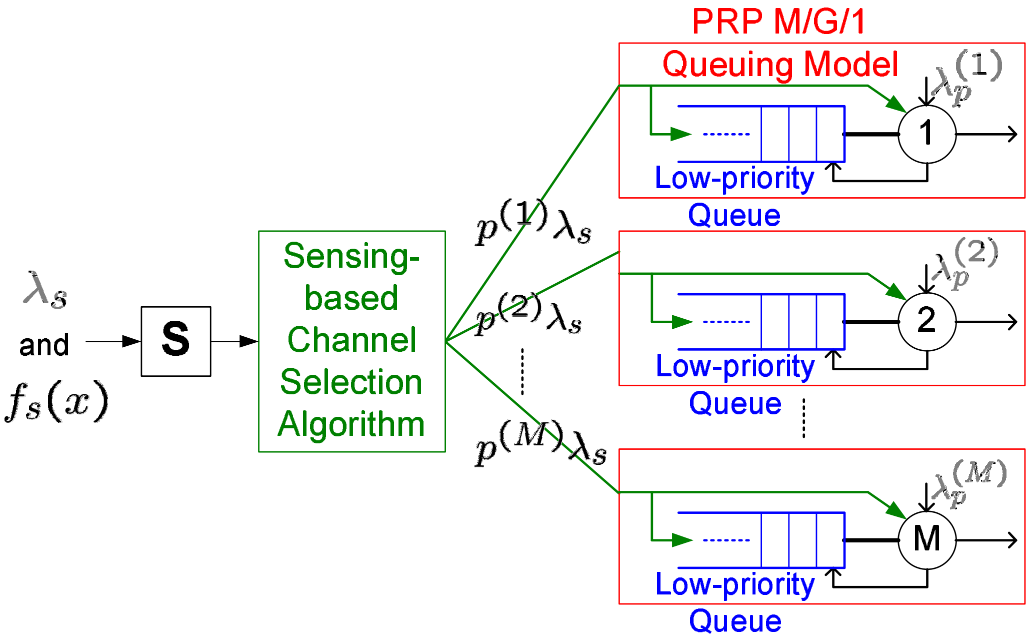


Figure 4.4: Performance model for the sensing-based channel selection scheme where the channel usage behaviors are characterized by the PRP M/G/1 queueing systems.

system time. Thus, the optimal overall system time (denoted by S^*) can be expressed as follows:

$$S^* = \min\left(\mathbf{E}[S_{pb}], \mathbf{E}[S_{sb}]\right) \ . \tag{4.7}$$

In the next section, we will show how to derive $\mathbf{E}[S_{pb}]$ and $\mathbf{E}[S_{sb}]$.

4.4 Analysis of Overall System Time

As discussed in Section 4.3.1, the overall system time consists of the waiting time and the extended data delivery time. Let $\mathbf{E}[T_{pb}]$ and $\mathbf{E}[T_{sb}]$ be the average data delivery time for the probability- and sensing-based spectrum decision methods, respectively. Furthermore, denote $\mathbf{E}[W_{pb}]$ and $\mathbf{E}[W_{sb}]$ as the average waiting time for the probability- and sensing-based spectrum decision methods, respectively. Then, we can have

$$\mathbf{E}[S_{pb}] = \mathbf{E}[W_{pb}] + \mathbf{E}[T_{pb}] \ , \tag{4.8}$$

and

$$\mathbf{E}[S_{sb}] = \mathbf{E}[W_{sb}] + \mathbf{E}[T_{sb}] \ . \tag{4.9}$$

In the following, we will investigate how to obtain the average extended data delivery time and the average waiting time.

4.4.1 Extended Data Delivery Time

First, we investigate the effects of multiple interruptions on the extended data delivery time. Within the transmission period of a secondary connection, it is likely to have multiple spectrum handoffs due to the interruptions from the primary users. The spectrum handoff procedure helps the secondary users vacate the occupied channel and then resume the unfinished transmission when this channel becomes idle. Clearly, multiple spectrum handoffs will increase the extended data delivery time and degrade the QoS for the latency-sensitive traffic of the secondary users [140].

Based on the PRP M/G/1 queueing model, we can derive the extended data delivery time of the secondary connections as follows. Let $N^{(k)}$ be the total number of interruptions for a secondary connection at channel k. Furthermore, denote $Y_p^{(k)}$ as the duration from the time instant that channel k is occupied by the primary connections until the time instant that the high-priority queue becomes empty. This duration is called the *busy period* resulting from transmissions of multiple primary connections at channel k. When a secondary connection is interrupted by primary users, it must stop transmitting on the current operating channel until all the primary connections in the high-priority queue have been served. In this case, the secondary connections of channel k must wait for the duration of $\mathbf{E}[Y_p^{(k)}]$ on average after the interruption event occurs. Denote $\widetilde{X}_s$ as the actual service time of the secondary connections when the effects of sensing errors are considered[2] and $T^{(k)}$ as the extended data delivery time of the secondary connections at channel k. We can have

$$\mathbf{E}[T^{(k)}] = \mathbf{E}[\widetilde{X}_s] + \mathbf{E}[N^{(k)}]\mathbf{E}[Y_p^{(k)}] \ . \tag{4.10}$$

Let $\widetilde{X}_p^{(k)}$ be the actual service time of the primary connections at channel k when the effects of sensing errors are considered. One can obtain $\mathbf{E}[N^{(k)}] =$

[2] Although this chapter assumes that all M channels have the same data transmission rate (or equivalently service rate), the proposed model can be applied to the CR network where all channels have different data rates. In the CR network with heterogeneous data rates, the secondary connections at different channels have different average service time. Hence, they will have different average actual service time. In this case, the notation $\widetilde{X}_s$ in (4.10) should be replaced by the notation $\widetilde{X}_s^{(k)}$, which is the actual service time of the secondary connections at channel k. More discussions had been shown in [135].

$\lambda_p^{(k)}\mathbf{E}[\widetilde{X}_s]$ and $\mathbf{E}[Y_p^{(k)}] = \frac{\mathbf{E}[\widetilde{X}_p^{(k)}]}{1-\lambda_p^{(k)}\mathbf{E}[\widetilde{X}_p^{(k)}]}$ according to to [141]. Note that $\mathbf{E}[\widetilde{X}_s]$ and $\mathbf{E}[\widetilde{X}_p^{(k)}]$ will be derived in Section 4.5.

Finally, the average extended data delivery time for the probability- and sensing-based channel selection methods can be expressed as follows:

$$\mathbf{E}[T_{pb}] = \sum_{k=1}^{M} p_{pb}^{(k)}\mathbf{E}[T^{(k)}] \ , \tag{4.11}$$

and

$$\mathbf{E}[T_{sb}] = \sum_{k=1}^{n} p_{sb}^{(k)}\mathbf{E}[T^{(k)}] \ . \tag{4.12}$$

For various channel selection algorithms, we use different methods to evaluate the corresponding distribution probability vectors $\boldsymbol{p}$. For the probability-based scheme, the distribution probability vector $\boldsymbol{p}_{pb}$ can be designed by solving the **Overall System Time Minimization Problem for Probability-based Channel Selection Scheme** in (4.2). For the sensing-based scheme, the distribution probability vector $\boldsymbol{p}_{sb}$ is determined inherently based on the given traffic patterns. Intuitively, a channel with larger idle probability will be selected more frequently through spectrum sensing. How to derive $\boldsymbol{p}_{sb}$ from the given traffic parameters will be discussed in Appendix A.

4.4.2 Waiting Time

Next, we focus on the derivations of the average waiting time for the probability-based and sensing-based channel selection schemes.

Probability-based Channel Selection Scheme

For the probability-based channel selection scheme, a secondary connection selects its operating channel based on the predetermined probability. Then, it is directly connected to the low-priority queue of the selected channel. It cannot be served until all the primary and the secondary connections in the high-priority queue and the present low-priority queue of the selected channel have been served. Hence, the waiting time is the required duration from the time instant that a secondary connection arrives at the low-priority queue of the selected channel until the time instant that the selected channel becomes idle. That is, the waiting time is the duration spent in the waiting queue by a secondary connection. Hence, $\mathbf{E}[W_{pb}]$ can be expressed as follows:

$$\mathbf{E}[W_{pb}] = \sum_{k=1}^{M} p_{pb}^{(k)} \mathbf{E}[W_{pb}^{(k)}] \ , \tag{4.13}$$

where $W_{pb}^{(k)}$ is the waiting time of the secondary connections at channel k for the probability-based channel selection scheme. Applying the PRP M/G/1 queueing theory [142], one can obtain

$$\mathbf{E}[W_{pb}^{(k)}] = \frac{\mathbf{E}[R^{(k)}]}{(1-\rho_p^{(k)})(1-\rho_p^{(k)}-\rho_s^{(k)})} \ , \tag{4.14}$$

where $\rho_p^{(k)}$ and $\rho_s^{(k)}$ are the busy probabilities resulting from the primary and the secondary connections at channel k when sensing errors are considered, respectively. Hence, we can have $\rho_p^{(k)} = \lambda_p^{(k)} \mathbf{E}[\widetilde{X}_p^{(k)}]$ and $\rho_s^{(k)} = \lambda_s^{(k)} \mathbf{E}[\widetilde{X}_s]$. Furthermore, $\mathbf{E}[R^{(k)}]$ is the average remaining time to complete the service of the connection being served at channel k. Referring to [142], we have

$$\mathbf{E}[R^{(k)}] = \frac{1}{2}\lambda_p^{(k)} \mathbf{E}[(\widetilde{X}_p^{(k)})^2] + \frac{1}{2} p_{pb}^{(k)} \lambda_s \mathbf{E}[(\widetilde{X}_s)^2] \ . \tag{4.15}$$

Then, substituting (4.14) and (4.15) into (4.13), we can obtain the closed-form expression for $\mathbf{E}[W_{pb}]$.

Finally, substituting (4.11) and (4.13) into (4.8), we can obtain the relationship between the average overall system time and the distribution probability vector $\boldsymbol{p}_{pb}$ for the probability-based channel selection scheme. Then, the optimal distribution probability vector $\boldsymbol{p}^*$ can be determined by solving the **Overall System Time Minimization Problem for Probability-based Channel Selection** in (4.2).

Sensing-based Channel Selection Scheme

The waiting time W_{sb} for the sensing-based channel selection method consists of the total sensing time and the queueing time (denoted by W'_{sb}). Let τ be the sensing time for scanning one candidate channel. Hence, $n\tau$ is the total sensing time for scanning all the n candidate channels. After wideband sensing, the secondary user can decide channel availability and then transmits data at one of the idle channels. Moreover, if the idle channel cannot be found, the secondary user cannot transmit immediately. In this case, the secondary user's connection will be put into the low-priority queue of the randomly selected channel. Hence, we can have

$$\mathbf{E}[W_{sb}] = n\tau + \mathbf{Pr}(\mathcal{E}) \times 0 + \mathbf{Pr}(\mathcal{E}^c) \times \mathbf{E}[W'_{sb}] \ , \tag{4.16}$$

where $\mathcal{E}$ is the event that at least one idle channel can be found after sensing, and $\mathcal{E}^c$ is the compliment of $\mathcal{E}$.

Next, the closed-form expressions for $\mathbf{Pr}(\mathcal{E})$ and $\mathbf{Pr}(\mathcal{E}^c)$ can be derived by the following two observations. First, a channel is called actual idle if and only if (1) this channel is not occupied by the primary connections and (2) the low-priority queue of this channel is empty. Note that the second condition should be contained because the FCFS scheduling discipline is adopted. Secondly, an actual idle channel is assessed as idle through spectrum sensing if and only if false alarm does not occur. Hence, we can have

$$\begin{aligned}\mathbf{Pr}(\mathcal{E}) &= \sum_{k=1}^{n} [\mathbf{Pr}(\mathcal{E}|\text{k channels are actually idle}) \times \mathbf{Pr}(\text{k channels are actually idle})] \\ &= \sum_{k=1}^{n} \left[[1-(P_F)^k] \times \sum_{\Im \subseteq \Omega, |\Im|=k} \left[\prod_{i\in\Im}(1-\rho^{(i)}) \prod_{j\in\Omega-\Im} \rho^{(j)} \right] \right] \ , \end{aligned} \tag{4.17}$$

where $\rho^{(k)} = \rho_p^{(k)} + \rho_s^{(k)}$ and P_F is the false alarm probability. On the other hand, $\mathcal{E}^c$ is the compliment of $\mathcal{E}$. That is,

$$\mathbf{Pr}(\mathcal{E}^c) = 1 - \mathbf{Pr}(\mathcal{E}) \ . \tag{4.18}$$

Moreover, when all channels are assessed as busy, each channel is selected by the secondary users with probability $1/n$. Hence, in this case, one can derive the average queueing time based on the PRP M/G/1 queueing theory as follows [142] :

$$\mathbf{E}[W'_{sb}] = \sum_{k=1}^{n} \left[\frac{1}{n} \cdot \frac{\mathbf{E}[R^{(k)}]}{(1-\rho_p^{(k)})(1-\rho_p^{(k)}-\rho_s^{(k)})} \right] \ . \tag{4.19}$$

Finally, substituting (4.12) and (4.16) into (4.9), we can obtain the relationship between the average overall system time and the number of candidate channels n for the sensing-based channel selection scheme.

Determining the optimal number of candidate channels (denoted by n^*) is the key issue for sensing-based spectrum decision scheme. Intuitively, a small number of candidate channels can reduce the total sensing time $n\tau$ in (4.16). However, it is harder to find one idle channel from fewer candidate channels, resulting in a larger value of $\mathbf{Pr}(\mathcal{E}^c)$ in (4.16) and thus increasing the overall system time. The optimal number of candidate channels n^* can be determined by solving the **Overall System Time Minimization Problem for Sensing-based Channel Selection** in (4.6).

4.5 Effects of Sensing Errors

Sensing errors such as false alarm and missed detection will degrade the performance of the secondary users and the primary users[3]. This section in-

[3] The relationship between the missed detection probability P_M and the false alarm probability P_F can be characterized by the receiver operating characteristic curve [143].

vestigates the effects of false alarm and missed detection on the transmission time of the secondary and the primary connections. Specifically, we will show how to derive the first and the second moments of $\widetilde{X}_s$ and $\widetilde{X}_p^{(k)}$.

4.5.1 False Alarm

First, we study the effect of false alarm on the actual service time of the secondary connections. When a false alarm occurs, a secondary user cannot transmit data even with an idle channel. Hence, the actual service time of a secondary connection will be extended to $\widetilde{X}_s$ (slots/arrival) from X_s (slots/arrival). The first and the second moments of $\widetilde{X}_s$ can be expressed as follows:

$$\mathbf{E}[\widetilde{X}_s] = \sum_{x=1}^{\infty} \mathbf{E}[\widetilde{X}_s | X_s = x] \mathbf{Pr}(X_s = x) \ , \tag{4.20}$$

and

$$\mathbf{E}[(\widetilde{X}_s)^2] = \sum_{x=1}^{\infty} \mathbf{E}[(\widetilde{X}_s)^2 | X_s = x] \mathbf{Pr}(X_s = x) \ . \tag{4.21}$$

Note that because the false-alarm slot cannot be exploited by any secondary or primary connections, it can be regarded as a busy slot. Hence, we can have $\rho_s^{(k)} = \lambda_s^{(k)} \mathbf{E}[\widetilde{X}_s]$.

When a false alarm occurs, the data transmission is postponed to the next slot. Hence, for a connection with x slots, its actual service time will be extended to $x+i$ slots if and only if false alarms occur in i slots out of the first $x+i-1$ slots and false alarms do not occur at the $(x+i)^{th}$ slot. Thus, the conditional expectation of the actual service time follows the negative binomial distribution with parameter P_F. That is,

$$\mathbf{E}[\widetilde{X}_s | X_s = x] = \sum_{i=0}^{\infty} (x+i) \binom{x+i-1}{i} (1-P_F)^x (P_F)^i \ , \tag{4.22}$$

and

$$\mathbf{E}[(\widetilde{X}_s)^2|X_s = x] = \sum_{i=0}^{\infty}(x+i)^2\binom{x+i-1}{i}(1-P_F)^x(P_F)^i \ , \tag{4.23}$$

where P_F is the false alarm probability. Because $\mathbf{Pr}(X_s = x)$ is given by $f_s(x)$, we can obtain $\mathbf{E}[\widetilde{X}_s]$ and $\mathbf{E}[(\widetilde{X}_s)^2]$ by substituting (4.22) and (4.23) into (4.20) and (4.21), respectively. For example, if $f_s(x)$ is the geometric distribution, i.e.,

$$f_s(x) = (1 - \frac{1}{\mathbf{E}[X_s]})^{x-1}(\frac{1}{\mathbf{E}[X_s]}) \ , \tag{4.24}$$

we can have

$$\mathbf{E}[\widetilde{X}_s] = \frac{\mathbf{E}[X_s]}{1-P_F} \ , \tag{4.25}$$

and

$$\mathbf{E}[(\widetilde{X}_s)^2] = \frac{\mathbf{E}[X_s](2\mathbf{E}[X_s]-1+P_F)}{(1-P_F)^2} \ . \tag{4.26}$$

4.5.2 Missed Detection

The data frame of the primary connection will be stained by the secondary connection when a missed detection occurs. Thus the primary user will request to retransmit this stained data frame in the next slot. Hence, the actual service time of a primary connection will be extended from $X_p^{(k)}$ (slots/arrival) to $\widetilde{X}_p^{(k)}$ (slots/arrival). The first and the second moments of $\widetilde{X}_p^{(k)}$ can be expressed as follows:

$$\mathbf{E}[\widetilde{X}_p^{(k)}] = \sum_{x=1}^{\infty}\mathbf{E}[\widetilde{X}_p^{(k)}|X_p^{(k)} = x]\mathbf{Pr}(X_p^{(k)} = x) \ , \tag{4.27}$$

and

$$\mathbf{E}[(\widetilde{X}_p^{(k)})^2] = \sum_{x=1}^{\infty}\mathbf{E}[(\widetilde{X}_p^{(k)})^2|X_p^{(k)} = x]\mathbf{Pr}(X_p^{(k)} = x) \ . \tag{4.28}$$

Basically, there are two types of missed detections in CR networks [64,66]. Firstly, when a primary user transmits data, a newly arriving secondary connection may incorrectly determine that this specific channel is available in its first sensing phase. We call this situation the class-A missed detection. After a secondary user arrives at a CR network for a while, it may also fail to detect the presence of primary users. In this case, the class-B missed detection occurs. The authors in [64,66] found that the class-B missed detection is small because the sensing results at the first sensing phase can be employed to improve the accuracy of the sensing results at the following sensing phases.

Next, we explain the effect of class-A missed detection on the actual service time of the primary connection at channel k. We consider a transmission slot of this primary connection. During this slot, more than one arrival of the secondary connection appears with probability $1 - e^{-\lambda_s^{(k)}\Delta}$, where Δ is the slot duration. For these arrivals of secondary connections, each of them will assess this busy slot as idle if and only if (1) a missed detection occurs and (2) the low-priority queue of channel k is empty. Let $Q_s^{(k)}$ be the length of the low-priority queue at channel k. Hence, the first arrival at the considered slot will make an error channel assessment with probability $P_M\mathbf{Pr}\{Q_s^{(k)} = 0\}$, where P_M is the missed detection probability for spectrum sensing and $\mathbf{Pr}\{Q_s^{(k)} = 0\}$ has been derived in [144]. However, for the remaining arrivals in the considered slot, we have $\mathbf{Pr}\{Q_s^{(k)} = 0\} = 0$ because the first arrival has been put into the low-priority queue of channel k. Thus, the remaining arrivals do not make the error channel assessment. From above observations, we can conclude that a primary connection's transmission slot is stained by the arrivals of the secondary connections with probability

$$P_I^{(k)} = (1 - e^{-\lambda_s^{(k)}\Delta})P_M\mathbf{Pr}\{Q_s^{(k)} = 0\} \ . \tag{4.29}$$

Similar to the case of missed detection, we find that the random variables $\widetilde{X}_p^{(k)}$ and $(\widetilde{X}_p^{(k)})^2$ follows the negative binomial distribution with parameter $P_I^{(k)}$ when $X_p^{(k)} = x$. Then, because $\mathbf{Pr}(X_p^{(k)} = x)$ can be determined by $f_p^{(k)}(l)$, we can calculate the values of $\mathbf{E}[\widetilde{X}_p^{(k)}]$ and $\mathbf{E}[(\widetilde{X}_p^{(k)})^2]$ in (4.27) and

(4.28), respectively. For example, if $f_p^{(k)}(l)$ is the geometric distribution, i.e.,

$$f_p^{(k)}(x) = (1 - \frac{1}{\mathbf{E}[X_p^{(k)}]})^{x-1}(\frac{1}{\mathbf{E}[X_p^{(k)}]}) \ , \tag{4.30}$$

we can have

$$\mathbf{E}[\widetilde{X}_p^{(k)}] = \frac{\mathbf{E}[X_p^{(k)}]}{1 - P_I^{(k)}} \ , \tag{4.31}$$

and

$$\mathbf{E}[(\widetilde{X}_p^{(k)})^2] = \frac{\mathbf{E}[X_p^{(k)}](2\mathbf{E}[X_p^{(k)}] - 1 + P_I^{(k)})}{(1 - P_I^{(k)})^2} \ . \tag{4.32}$$

4.6 Numerical Results

In this section, numerical results are presented to show how to design the system parameters for the load-balancing spectrum decision methods, including the probability-based and the sensing-based spectrum decision schemes. We adopt the system parameters in the IEEE 802.22 standard in our simulation [145], where the time slot duration is 10 msec, $P_M = 0.1$, and $P_F = 0.1$. Because this report focuses on the latency-sensitive traffic, we can assume that the service time distributions of primary and secondary connections are geometrically distributed (see page 135 in [142]). Note that we only use the geometric distribution as an example here. Indeed, the proposed analytical framework can be applied to any distributions. It only requires the knowledge of the first and the second moments of the service time distributions for the primary and the secondary connections.

4.6.1 Probability-based Spectrum Decision Scheme

Figure 4.5 shows the effect of various arrival rates of the secondary connections on the optimal distribution probability vector, where the distribution probability vector is plotted in each bar and the summation of all probabilities in each bar is 1. In the figure, we consider a four-channel system with the following traffic parameters: $\lambda_p^{(1)} = 0.01$, $\lambda_p^{(2)} = 0.01$, $\lambda_p^{(3)} = 0.02$, and $\lambda_p^{(4)} = 0.02$ as well as $\mathbf{E}[X_p^{(1)}] = 20$, $\mathbf{E}[X_p^{(2)}] = 30$, $\mathbf{E}[X_p^{(3)}] = 20$, and $\mathbf{E}[X_p^{(4)}] = 25$. When $\lambda_s = 0.01$, all the secondary users prefer selecting channel 1 to be their operating channels because channel 1 has the lightest traffic loads. Furthermore, as λ_s increases, some secondary users tend to select other channels to transmit data in order to balance the traffic loads in each channel. For example, when $\lambda_s = 0.1$, the optimal distribution probability vector is $(0.4142, 0.2784, 0.2131, 0.0943)$. Inevitably, channel 1 is still selected to be the operating channel with the largest probability.

Furthermore, Fig. 4.6 shows the channel busy probability under various arrival rates of the secondary connections. In the beginning, channel 1 has the lowest busy probability. However, when $\lambda_s \geq 0.05$, channel 1 has the highest busy probability because most secondary users prefer to select channel 1 to transmit data. Although channel 1 has the highest busy probability in this case, one can find that the secondary users still favor channel 1 from the viewpoint of the overall system time. The performance advantages of the choosing the probability vector based on the proposed analytical framework over the traditional channel selection methods will be illustrated in Fig. 4.11 from the perspective of the overall system time.

Figure 4.7 shows that most secondary connections prefer selecting a channel with the largest arrival rate and the shortest service time of the primary connections even though all the channels have the same busy probability of the primary connections. Here, we consider the following traffic parameters: $\lambda_p^{(1)} = 0.01$, $\lambda_p^{(2)} = 0.02$, $\lambda_p^{(3)} = 0.04$, and $\lambda_p^{(4)} = 0.08$ as well as $\mathbf{E}[X_p^{(1)}] = 40$, $\mathbf{E}[X_p^{(2)}] = 20$, $\mathbf{E}[X_p^{(3)}] = 10$, and $\mathbf{E}[X_p^{(4)}] = 5$. Hence, all channels have

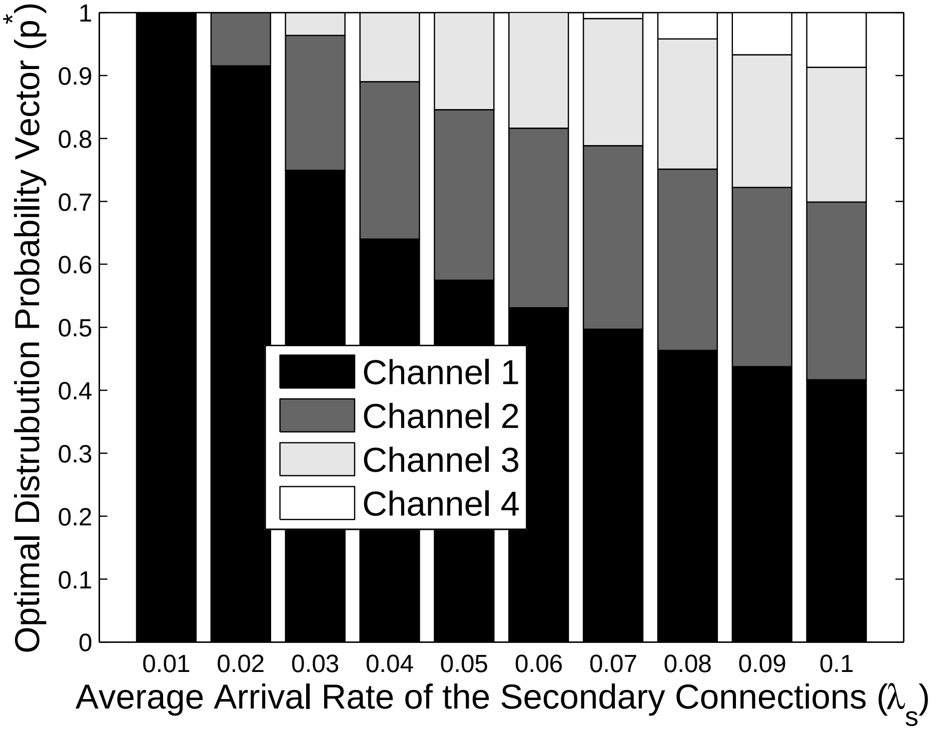


Figure 4.5: Optimal distribution probability vector for the probability-based spectrum decision with various arrival rates of the secondary connections, where $P_F = 0.1$, $P_M = 0.1$, and $\mathbf{E}[X_s] = 10$.

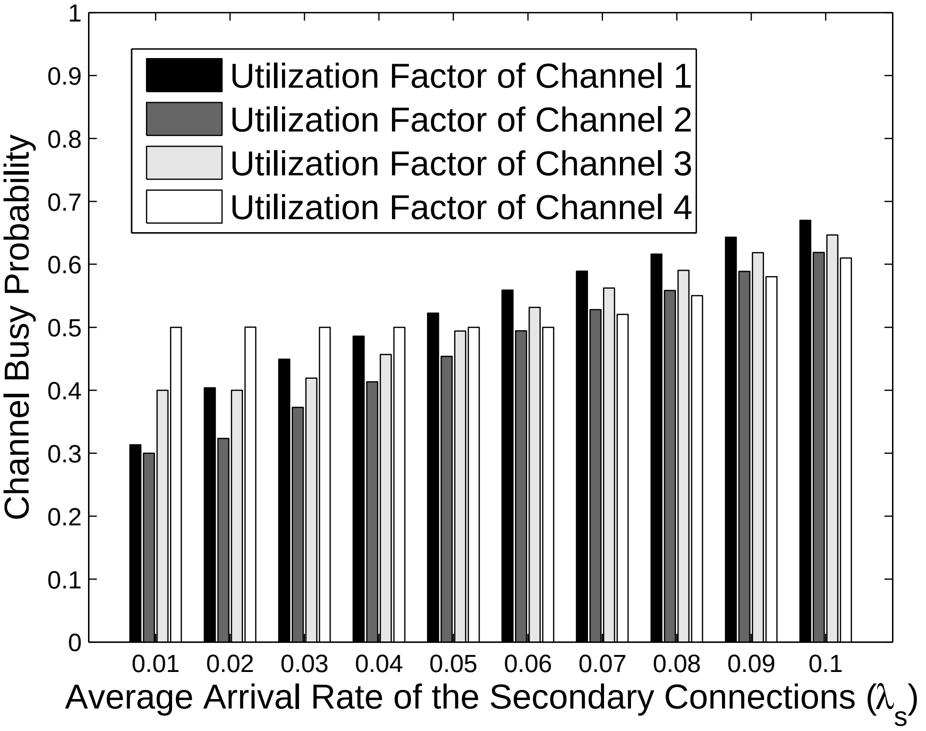


Figure 4.6: Channel busy probability for the probability-based spectrum decision with various arrival rates of the secondary connections, where $P_F = 0.1$, $P_M = 0.1$, and $\mathbf{E}[X_s] = 10$.

the same busy probability, which is equal to 0.4, when $\lambda_s = 0$. According to (4.13) and (4.14), we know that selecting channel 4 can result in shorter average waiting time ($\mathbf{E}[W_{pb}]$) because channel 4 has the smallest value of $\mathbf{E}[R^{(k)}]$. Consequently, most secondary connections prefer selecting channel 4 and thus it has the highest busy probability when $\lambda_s > 0$.

Figure 4.8 shows the effects of false alarms on the optimal distribution probability vector. When $P_F = 0.05$, only three channels can be the candidate channels. However, all the four channels can be the candidate channels when $P_F \geq 0.1$. This phenomenon can be interpreted as follows. When P_F becomes higher, $\mathbf{E}[\widetilde{X}_s]$ increases due to more false alarms. Hence, the actual traffic loads ($\rho_s = \lambda_s \mathbf{E}[\widetilde{X}_s]$) of the secondary connections become heavy. Then, the secondary connections must distribute overall traffic loads to more channels in order to prevent channel contention.

4.6.2 Sensing-based Spectrum Decision Scheme

Figures 4.9 and 4.10 show the effects of $\mathbf{E}[X_s]$ and P_F on the optimal number of candidate channels n^*, respectively. Here, we consider a four-channel system with the following traffic parameters: $(\lambda_p^{(1)}, \lambda_p^{(2)}, \lambda_p^{(3)}, \lambda_p^{(4)}) = (0.01, 0.015, 0.02, 0.025)$, $\lambda_s = 0.02$ and $\tau = 2$. Moreover, $\mathbf{E}[X_p^{(k)}] = 20$ for any k. From Fig. 4.9, one can see that when $P_F = 0.1$, $n^* = 1$ and 2 for $\mathbf{E}[X_s] = 5$ and 10, respectively. In Fig. 4.10, we see that $n^* = 1$ and 2 for $P_F = 0.1$ and 0.5 when $\mathbf{E}[X_s] = 5$. It is observed that the optimal value of n monotonically increases as $\mathbf{E}[X_s]$ or P_F increases. This is because a larger value of $\mathbf{E}[X_s]$ or P_F can lead to a larger value of $\mathbf{E}[\widetilde{X}_s]$ according to (4.25). From (4.19) one can expect that the queueing time will become longer for a larger value of $\mathbf{E}[\widetilde{X}_s]$. In this case, the secondary users shall sense more channels to increase the probability of finding idle channels $\mathbf{Pr}(\mathcal{E})$, which will reduce the waiting time.

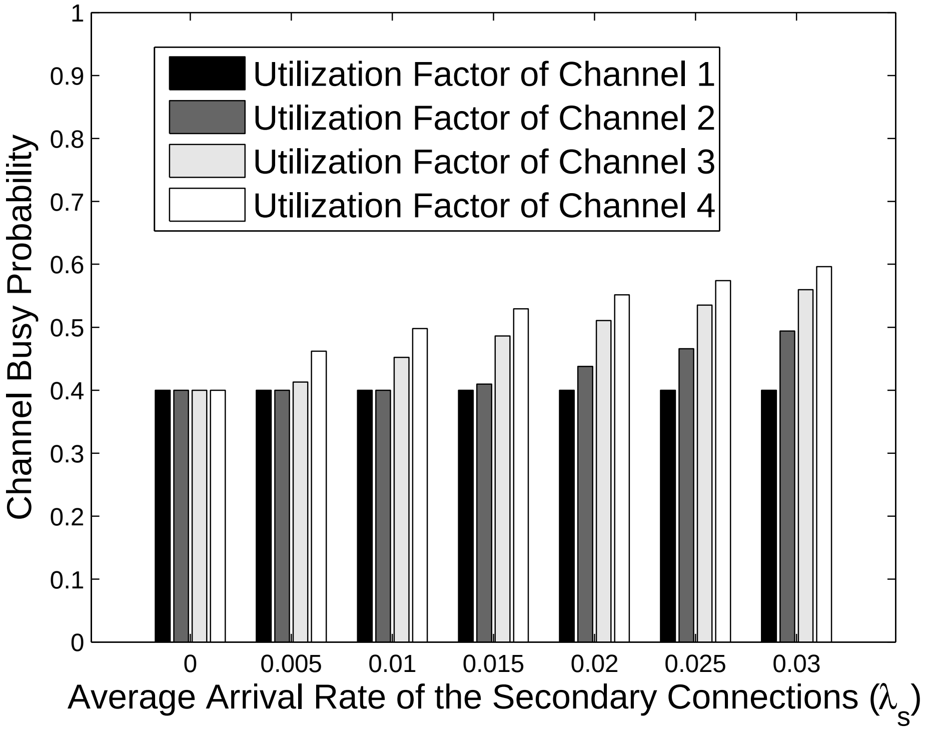


Figure 4.7: Channel busy probability for the probability-based spectrum decision with various arrival rates of the secondary connections, where $P_F = 0$, $P_M = 0$, and $\mathbf{E}[X_s] = 15$.

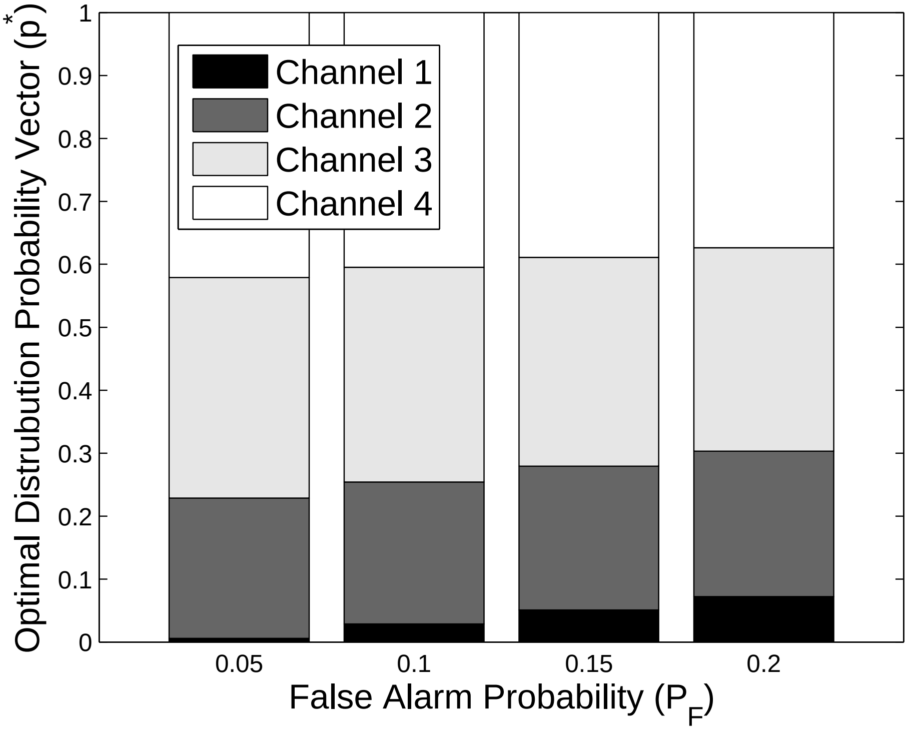


Figure 4.8: Optimal distribution probability vector for the probability-based spectrum decision with various arrival rates of the secondary connections, where $P_M = 0.1$, $\lambda_s = 0.03$, and $\mathbf{E}[X_s] = 15$.

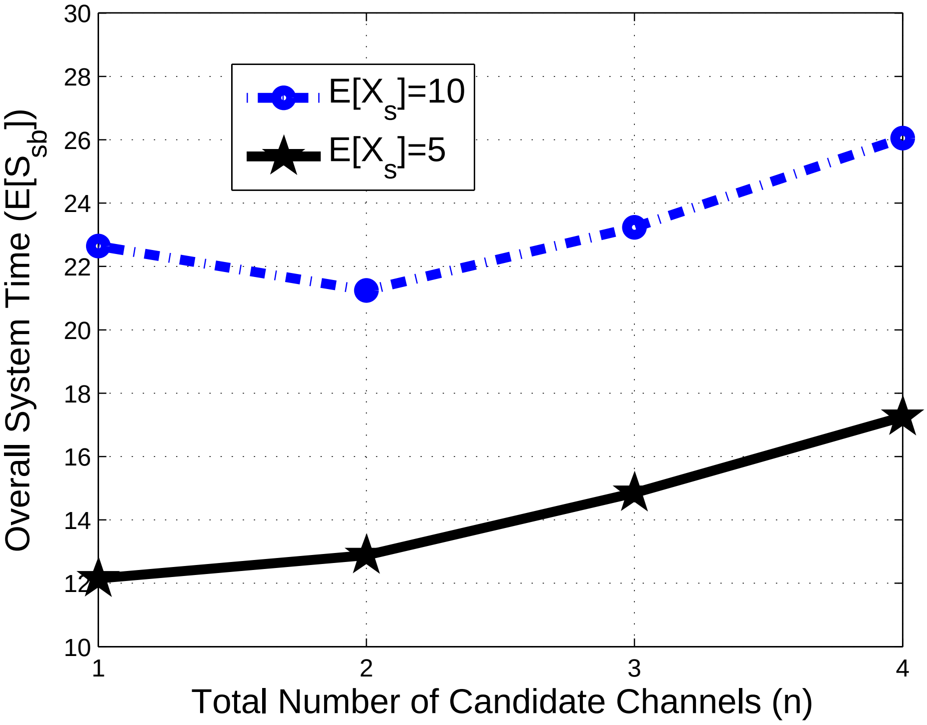


Figure 4.9: Overall system time for the sensing-based spectrum decision with various numbers of candidate channels n, where $P_F = 0.1$, $P_M = 0.1$, $\tau = 2$, and $\mathbf{E}[X_p] = 20$.

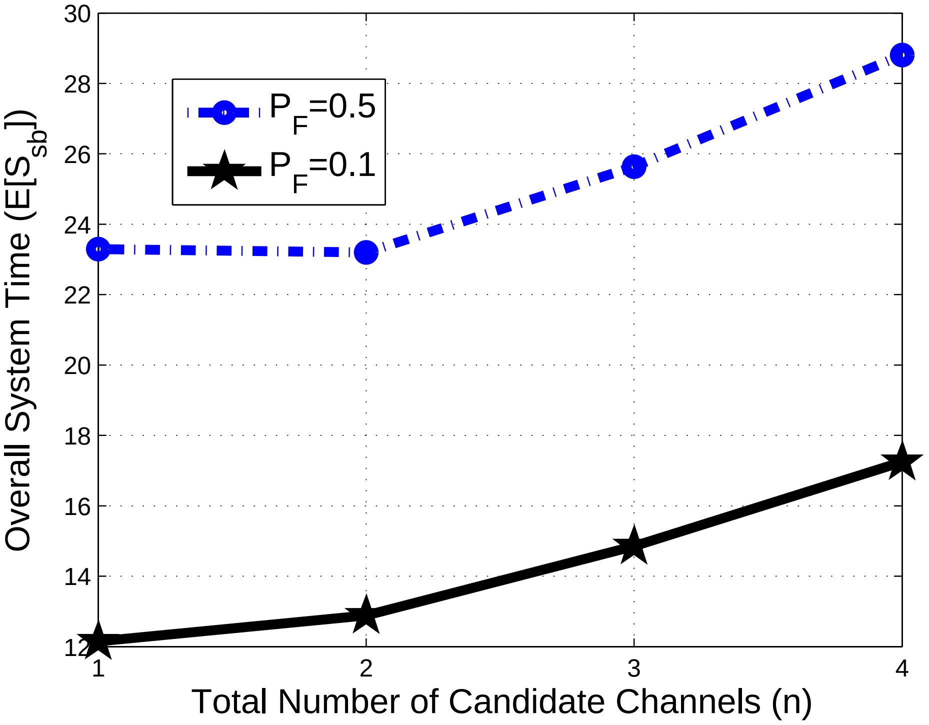


Figure 4.10: Overall system time for the sensing-based spectrum decision with various numbers of candidate channels n, where $P_M = 0.1$, $\tau = 2$, $\mathbf{E}[X_p] = 20$, and $\mathbf{E}[X_s] = 5$.

4.6.3 Comparison between Different Spectrum Decision Schemes

Figure 4.11 shows the effects of λ_s on the average overall system time for three different channel selection schemes: (1) sensing-based method; (2) probability-based method; and (3) non-load-balancing method. Consider a three-channel system with the following traffic parameters: $(\lambda_p^{(1)}, \lambda_p^{(2)}, \lambda_p^{(3)}) = (0.02, 0.02, 0.03)$, $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}]) = (20, 25, 20)$, and $\mathbf{E}[X_s] = 10$. The overall system time of the probability-based and sensing-based channel selection schemes are calculated from (4.8) and (4.9), respectively. For the non-load-balancing method, all the secondary connections will select channel 1 to be their operating channels because channel 1 has the lowest busy probability. One can find that both the load-balancing channel selection schemes can significantly reduce the average overall system time compared to the non-load-balancing scheme, especially for larger λ_s. When τ is small (e.g. 5 slots), the sensing-based spectrum decision scheme can result in the shortest overall system time. As τ increases, the improvement of the sensing-based spectrum decision over other schemes decreases. In addition, we also observe that when $\tau = 17$ and $\lambda_s < 0.026$, the probability-based scheme has better overall system time performance than the sensing-based scheme. This is because the probability-based spectrum decision scheme can select the channels with lower interrupted probability. By contrast, if $\lambda_s > 0.026$, the sensing-based scheme can result in shorter overall system time because the sensing-based scheme can significantly reduce waiting time through wideband sensing. Based on (4.7), each secondary user can intelligently adopt the best channel selection scheme to minimize its overall system time. The two considered load-balancing spectrum decision methods can reduce the overall system time by over 50% compared to the existing non-load-balancing

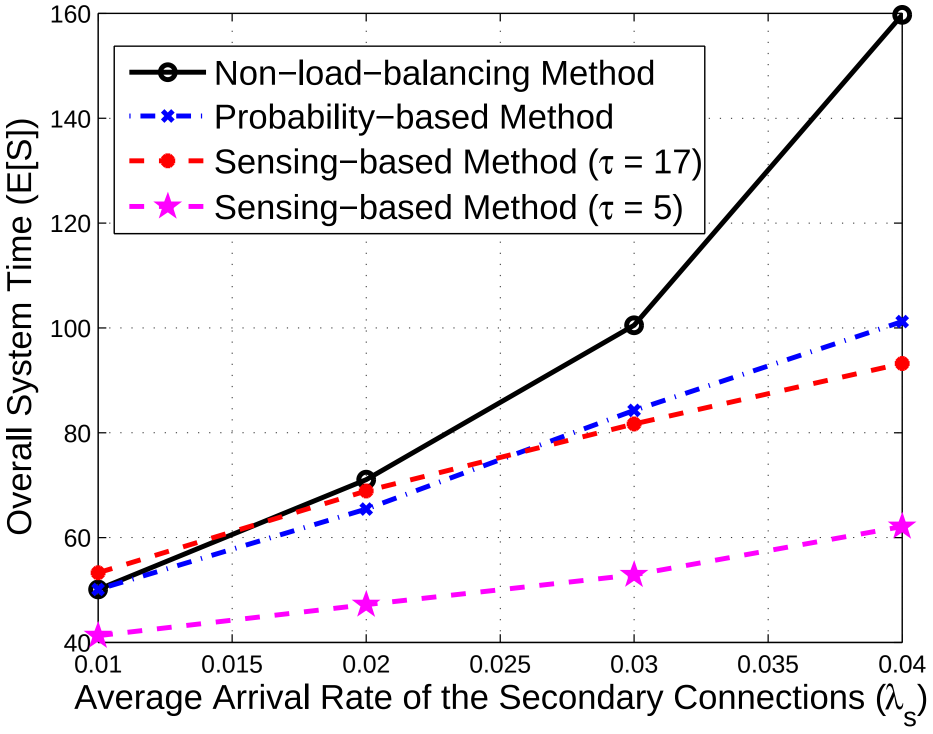


Figure 4.11: Comparison of the overall system time for three considered spectrum decision schemes, where $P_F = 0.1$, $P_M = 0.1$, and $\mathbf{E}[X_s] = 10$.

method when $\lambda_s = 0.04$.

Chapter 5

Proactive Spectrum Handoff

Spectrum handoff mechanisms can be generally categorized into two kinds according to the decision timing of selecting target channels [141]. The first kind is called the proactive spectrum handoff[1], which decides the target channels for future spectrum handoffs based on the long-term traffic statistics before data connection is established [128, 146, 147]. The second kind is called the reactive spectrum handoff scheme [148]. For this scheme, the target channel is searched in an on-demand manner [149, 150]. After a spectrum handoff is requested, spectrum sensing is performed to help the secondary users find idle channels to resume their unfinished data transmission. Both spectrum handoff schemes have their own advantages and disadvantages. A quantitative comparison of the two spectrum handoff schemes was provided in [151].

[1] In this report, we assume that spectrum handoff request is initiated only when the primary user appears as discussed in the IEEE 802.22 wireless regional area networks (WRAN) standard. In this scheme, the proactive spectrum handoff represents the spectrum handoff scheme with the proactively designed target channel sequences. It is different from the proactive spectrum handoff in [35, 40–48] that assumes spectrum handoff can be performed before the appearance of the primary users.

In this chapter, we focus on the modeling technique and performance analysis for the proactive spectrum handoff scheme, while leave the related studies on the reactive spectrum handoff in Chapter 6. Compared to the reactive spectrum handoff scheme, the proactive spectrum handoff is easier to achieve a consensus on their target channels between the transmitter and its intended receiver because both the transmitter and receiver can know their target channel sequence for future spectrum handoffs before data transmission. Furthermore, the change switching delay of the proactive spectrum handoff is shorter than that of the reactive spectrum handoff because scanning wide spectrum to determine the target channel is not necessary at the moment of link transition. Nevertheless, the proactive spectrum handoff scheme shall resolve the obsolescent channel issue because the predetermined target channel may not be available any more when a spectrum handoff is requested.

The contribution of this chapter is to propose a preemptive resume priority (PRP) M/G/1 queueing network model to characterize the spectrum usage behaviors of the connection-based multiple-channel spectrum handoffs. Based on the proposed model, we derive the closed-form expression for the extended data delivery time of different proactively designed target channel sequences under various traffic arrival rates and service time distributions. We apply the developed analytical method to analyze the latency performance of spectrum handoffs based on the target channel sequences specified in the IEEE 802.22 wireless regional area networks (WRAN). We also suggest a traffic-adaptive target channel selection principle for spectrum handoffs under different traffic conditions.

5.1 Motivation

To characterize the channel obsolescence effects and the spectrum usage behaviors with a series of interruptions in the secondary connections, we suggest a new performance metric - the extended data delivery time of the secondary connections. It is defined as the duration from the instant of starting transmitting data until the instant of finishing the whole connection, during which multiple interruptions from the primary users may occur. In the context of the connection-based spectrum handoffs, how to analyze the extended data delivery time is challenging because three key design features must be taken into account: (1) generally distributed service time, where the probability density functions (pdfs) of service time of the primary and secondary connections can be any distributions; (2) different operating channels before and after spectrum handoff; and (3) queueing delay due to channel contention from multiple secondary connections. To the best of our knowledge, an analytical model for characterizing all these three features for multiple handoffs has rarely been seen in the literature.

5.2 System Model

5.2.1 Assumptions

In this chapter, we make the following assumptions:

- A default channel is preassigned to each secondary user through spectrum decision algorithms in order to balance the overall traffic loads of the secondary users to all the channels [135]. When a secondary transmitter has data, it can transmit handshaking signal at the default channel of the intended receiver to establish a secondary connection [152].

If the corresponding receiver's default channel is busy, the secondary transmitter must wait at this channel until it becomes available [40].

- Each primary connection is assigned with a default or licensed channel.
- Each secondary user can detect the presence of the primary user. In fact, this model can be also extended to consider the effects of false alarm and missed detection as discussed in Chapter 7.
- Any time only one user can transmit data at one channel.

5.2.2 Illustrative Example of Proactive Multiple Handoffs with Multiple Interruptions

A secondary connection may encounter multiple interruption requests during its transmission period. Because spectrum handoff procedures must be performed whenever a primary user appears, a set of target channels will be sequentially selected, called the *target channel sequence* in this report. Fig. 5.1 shows an example that three spectrum handoff requests occur during the transmission period of the secondary connection SC$_A$. In this example, SC$_A$'s initial (default) channel is Ch1 and its *target channel sequence* for spectrum handoffs is (Ch2, Ch2, Ch3, $\cdots$). The extended data delivery time of SC$_A$ is denoted by T. Furthermore, D_i is the handoff delay of the i^{th} interruption. Here, the handoff delay is the duration from the instant when the transmission is interrupted until the instant when the unfinished transmission is resumed.

We assume that the transmitter of SC$_A$ plans to establish a connection flow consisting of the 28 slot-sized frames to the corresponding receiver. Then, the transmission process with multiple handoffs is described as follows:

1. In the beginning, SC$_A$ is established at its default channel Ch1. When an interruption event occurs, SC$_A$ decides its target channel according to the predetermined target channel sequence.

2. At the first interruption, SC$_A$ changes its operating channel to the idle channel Ch2 from Ch1 because the first predetermined target channel is Ch2. In this case, the handoff delay D_1 is the channel switching time (denoted by t_s).

3. At the second interruption, SC$_A$ stays on its current operating channel Ch2 because the second target channel is Ch2. SC$_A$ cannot be resumed until all the high-priority primary connections finish their transmissions at Ch2. In this case, the handoff delay D_2 is the duration from the time instant that Ch2 is used by the primary connections until the time instant that the high-priority queue becomes empty. This duration (denoted by $Y_p^{(2)}$) is called the *busy period* resulting from the transmissions of multiple primary connections at Ch2.

4. At the third interruption, SC$_A$ changes its operating channel to Ch3 because the third target channel is Ch3. In this example, because Ch3 is busy, SC$_A$ must wait in the low-priority queue until all the data in the present high-priority and low-priority queues of Ch3 are served[2].

[2] Here, the 1-persistent waiting policy is adopted. That is, the interrupted secondary user must stay on the selected target channel even though the selected channel is busy and then transmit unfinished data when channel becomes idle. Another possible approach is to reselect a new channel at the next time slot when a busy channel is selected. However, this approach is more impractical because it will lead to many channel-switching behaviors during a secondary connection.

Hence, the handoff delay D_3 is the sum of this waiting time and the channel switching time t_s.

5. Finally, SC$_A$ is completed on Ch3.

When a secondary connection changes its operating channel from channel k to k' where $k' \neq k$, the expected handoff delay is the sum of the channel switching time t_s and the average waiting time of channel k' (denoted by $\mathbf{E}[W_s^{(k')}]$) for the secondary connections. Note that this waiting time $W_s^{(k')}$ is the duration from the time instant that a secondary connection enters the low-priority queue of channel k' until it gets a chance to transmit at channel k'. After the secondary connection's operating channel is changed to channel k', one of two situations will occur. If channel k' is idle as the first interruption in Fig. 5.1, the expected handoff delay is t_s since $\mathbf{E}[W_s^{(k')}|\text{channel } k' \text{ is idle}] = 0$. On the other hand, the expected handoff delay is $t_s + \mathbf{E}[W_s^{(k')}|\text{channel } k' \text{ is busy}]$ if channel k' is busy as the third interruption in Fig. 5.1.

5.3 Analytical Model

We use the PRP M/G/1 queueing network model proposed in Chapter 3 to characterize the channel usage behaviors of a CR network. Let $X_s^{(\eta)}$ (slots/arrival) be the service time of the secondary connections whose default channels are channel η and let $f_s^{(\eta)}(x)$ be the pdf of $X_s^{(\eta)}$. Figure 5.2 shows an example of the PRP M/G/1 queueing network model with three channels, in which the traffic flows of the primary connections and the secondary connections are directly connected to the high-priority queue and the low-priority queue, respectively. When a primary connection appears at the channel being occupied by the secondary connection, the interruption event occurs. The interrupted secondary connection decides its target chan-

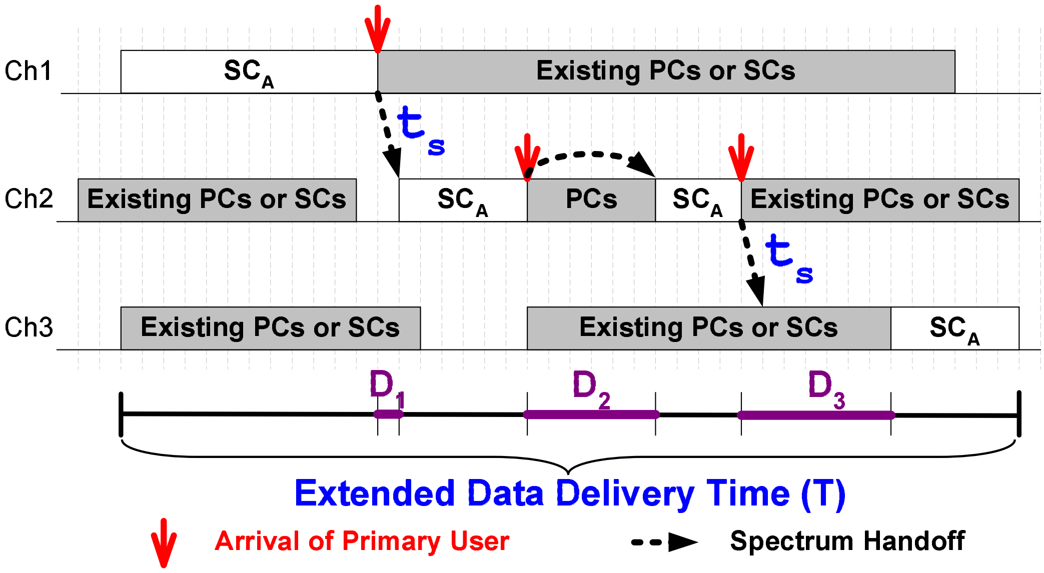


Figure 5.1: An example of transmission process for the secondary connection SC$_A$, where t_s is the channel switching time, T is the extended data delivery time of SC$_A$, and D_i is the handoff delay of the i^{th} interruption. The gray areas indicate that the channels are occupied by the existing primary connections (PCs) or secondary connections (SCs). Because SC$_A$ is interrupted three times in total, the overall data connection is divided into four segments.

nel for spectrum handoff according to the target channel predetermination algorithm which is implemented in the channel selection point $\boxed{\text{S}}$. In our queueing network model, the interrupted secondary connection can either stay on its current channel or change to another channel through different feedback paths. If a secondary connection chooses to stay on its current operating channel, its remaining data will be connected to the head of the low-priority queue of its current operating channel. On the other hand, if the decision is to change its operating channel, the remaining data of the interrupted secondary connection will be connected to the tail of the low-priority queue of the selected channel after channel switching time t_s. In order to characterize the handoff delay from channel switching time t_s, $\boxed{\text{S}}$ must be regarded as a server with constant service time t_s. Note that $\oplus$ in the figure represents that the traffic of the interrupted secondary connection is merged. Furthermore, when the interrupted secondary connection transmits the remaining data on the target channel, it may be interrupted again. Hence, this model can incorporate the effects of multiple interruptions in multi-channel spectrum handoffs.

5.4 Analysis of Extended Data Delivery Time

Based on the proposed PRP M/G/1 queuing network model, we can evaluate many performance metrics of the secondary connections with various target channel sequences. In this chapter, we focus on the analysis of the extended data delivery time, which is an important performance measure for the latency-sensitive traffic of the secondary connections.

A secondary connection may encounter many interruptions during its transmission period. Without loss of generality, we consider a secondary

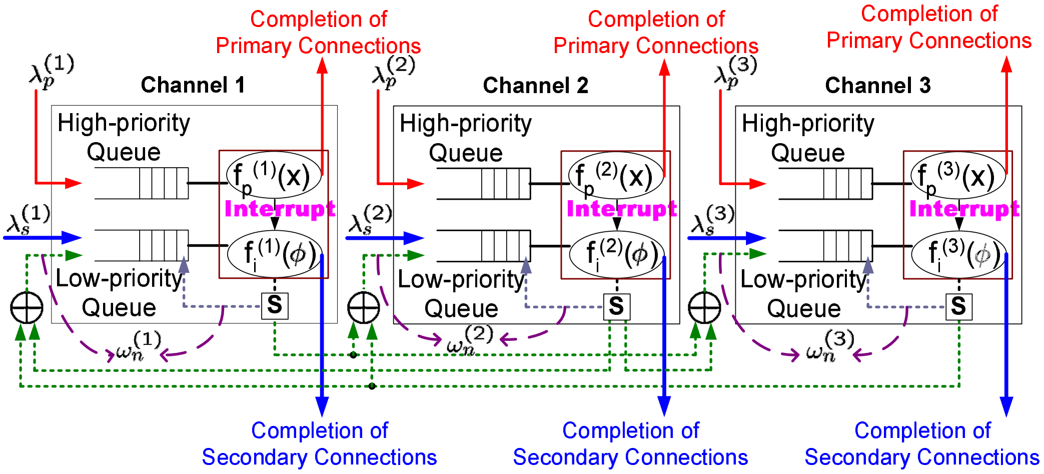


Figure 5.2: The PRP M/G/1 queueing network model with three channels where $\lambda_p^{(k)}$, $\lambda_s^{(k)}$, and $\omega_n^{(k)}$ are the arrival rates of the primary connections, the secondary connections, and the type-n secondary connections ($n \geq 1$) at channel k. Note that $\omega_0^{(k)} = \lambda_s^{(k)}$. Furthermore, $f_p^{(k)}(x)$ and $f_i^{(k)}(\phi)$ are the pdfs of $X_p^{(k)}$ and $\Phi_i^{(k)}$, respectively.

connection whose default channel is channel η in the following discussions. Let N be the total number of interruptions of this secondary connection. Then, the average extended data delivery time of this secondary connection can be expressed as

$$\mathbf{E}[T] = \sum_{n=1}^{\infty} \mathbf{E}[T|N = n]\mathbf{Pr}(N = n) \ . \tag{5.1}$$

First, we show how to derive the value of $\mathbf{E}[T|N = n]$ of (5.1). The considered secondary connection can be divided into many segments due to multiple interruptions as discussed in Fig. 5.1. Hence, the extended data delivery time of this secondary connection consists of the original service time and the cumulative delay resulting from multiple handoffs. Let D_i be the handoff delay of the considered secondary connection for the i^{th} interruption. When $N = n$, we have $D_i = 0$ if $i \geq n+1$. Then, the conditional expectation of the extended data delivery time of the considered secondary connection given the event $N = n$ can be derived as

$$\mathbf{E}[T|N = n] = \mathbf{E}[X_s^{(\eta)}] + \sum_{i=1}^{n} \mathbf{E}[D_i] \ . \tag{5.2}$$

Next, we investigate how to derive the value of $\mathbf{Pr}(N = n)$ of (5.1). For the considered secondary connection, denote $s_{0,\eta}$ and $s_{i,\eta}$ as its default channel and its target channel at the i^{th} interruption, respectively. Thus, we have $s_{0,\eta} = \eta$ and this secondary connection's target channel sequence can be expressed as $(s_{1,\eta}, s_{2,\eta}, s_{3,\eta}, \cdots)$. Let $p_i^{(s_{i,\eta})}$ be the probability that the considered secondary connection is interrupted again at channel $s_{i,\eta}$ when it has experienced i interruption. Then, the probability that the considered secondary connection is interrupted exactly n times can be expressed as

$$\mathbf{Pr}(N = n) = (1 - p_n^{(s_{n,\eta})}) \prod_{i=0}^{n-1} p_i^{(s_{i,\eta})} \ . \tag{5.3}$$

Finally, substituting (5.2) an (5.3) into (5.1) yields

$$\mathbf{E}[T] = \mathbf{E}[X_s^{(\eta)}] + \sum_{n=1}^{\infty} \left[\left(\sum_{i=1}^{n} \mathbf{E}[D_i] \right) (1 - p_n^{(s_{n,\eta})}) \prod_{i=0}^{n-1} p_i^{(s_{i,\eta})} \right] \quad , \tag{5.4}$$

where the values of $\mathbf{E}[D_i]$ and $p_i^{(k)}$ can be obtained from the Propositions 1 and 2, respectively.

Proposition 1.

$$\boldsymbol{E}[D_i] = \begin{cases} \boldsymbol{E}[Y_p^{(s_{i,\eta})}] & , \quad s_{i-1,\eta} = s_{i,\eta} \\ \boldsymbol{E}[W_s^{(s_{i,\eta})}] + t_s & , \quad s_{i-1,\eta} \neq s_{i,\eta} \end{cases} \quad , \tag{5.5}$$

where

$$\boldsymbol{E}[Y_p^{(k)}] = \frac{\boldsymbol{E}[X_p^{(k)}]}{1 - \rho_p^{(k)}} = \frac{\boldsymbol{E}[X_p^{(k)}]}{1 - \lambda_p^{(k)} \boldsymbol{E}[X_p^{(k)}]} \quad , \tag{5.6}$$

and

$$\boldsymbol{E}[W_s^{(k)}] = \frac{\lambda_p^{(k)} \boldsymbol{E}[(X_p^{(k)})^2] + \sum_{i=0}^{\infty} \omega_i^{(k)} \boldsymbol{E}[(\Phi_i^{(k)})^2] + \frac{(\lambda_p^{(k)})^2 \boldsymbol{E}[(X_p^{(k)})^2]}{1 - \lambda_p^{(k)} \boldsymbol{E}[X_p^{(k)}]} \boldsymbol{E}[X_p^{(k)}]}{2(1 - \lambda_p^{(k)} \boldsymbol{E}[X_p^{(k)}] - \sum_{i=0}^{\infty} \omega_i^{(k)} \boldsymbol{E}[\Phi_i^{(k)}])} \quad . \tag{5.7}$$

Proof. The handoff delay $\mathbf{E}[D_i]$ depends on which channel is selected for the target channel at the i^{th} interruption. For the secondary connection with $(i-1)$ interruptions, its current operating channel is $s_{i-1,\eta}$. When it is interrupted again, its new operating channel is $s_{i,\eta}$. When $s_{i-1,\eta} = s_{i,\eta}$, it means that the considered secondary connection will stay on the current channel. When $s_{i-1,\eta} \neq s_{i,\eta}$, it represents that the considered secondary connection will change its operating channel to another channel. Both cases are discussed as follows.

(1) Staying case: When the considered secondary connection stays on its current operating channel $s_{i,\eta} = k$, it cannot be resumed until all the high-priority primary connections of channel k finish their transmissions. Hence, the handoff delay is the busy period resulting from multiple primary connections of channel k (denoted by $Y_p^{(k)}$) as discussed in Section 5.2.2. That is, we can have $\mathbf{E}[D_i] = \mathbf{E}[Y_p^{(k)}]$.

The value of $\mathbf{E}[Y_p^{(k)}]$ can be derived as follows. Denote I_p as the idle period resulting from the primary connections. This idle period is the duration from the termination of the busy period to the arrival of the next primary connection. Because of the memoryless property, the idle period follows the exponential distribution with rate $\lambda_p^{(k)}$. Hence, we have

$$\mathbf{E}[I_p^{(k)}] = \frac{1}{\lambda_p^{(k)}} \ . \tag{5.8}$$

Next, according to the definition of the utilization factor at channel k, we have

$$\rho_p^{(k)} = \lambda_p^{(k)} \mathbf{E}[X_p^{(k)}] \ . \tag{5.9}$$

Because $\rho_p^{(k)}$ is also the busy probability resulting from the primary connections of channel k, we have

$$\rho_p^{(k)} = \frac{\mathbf{E}[Y_p^{(k)}]}{\mathbf{E}[Y_p^{(k)}] + \mathbf{E}[I_p^{(k)}]} \ . \tag{5.10}$$

Then, substituting (5.8) and (5.9) into (5.10), we can obtain (5.6).

(2) Changing case: In this case, the considered secondary connection will change to channel $s_{i,\eta} = k'$. After switching channel from channel k to k', it must wait in the low-priority queue of channel k' until all the traffic in the high-priority and the present low-priority queues of channel k' are served as discussed in Section 5.2.2. Denote $W_s^{(k')}$ as this waiting time for the secondary connections at channel k'[3]. Hence, we have $\mathbf{E}[D_i] = \mathbf{E}[W_s^{(k')}] + t_s$.

[3] A secondary connection needs to change its operating channel only when a primary connection appears. Because the arrivals of the primary connections follow Poisson dis-

The value of $\mathbf{E}[W_s^{(k')}]$ can be derived as follows. Let $\mathbf{E}[Q_p^{(k')}]$ be the average number of the primary connections which are waiting in the high-priority queue of channel k' and $\mathbf{E}[Q_i^{(k')}]$ be the average number of the type-i secondary connections which are waiting in the low-priority queue of channel k'. Because the newly arriving secondary connections cannot be established until all the secondary connections in the low-priority queue and the primary connections in the high-priority queue have been served, the average waiting time of channel k' is expressed as

$$\mathbf{E}[W_s^{(k')}] = \mathbf{E}[R_s^{(k')}]+\mathbf{E}[Q_p^{(k')}]\mathbf{E}[X_p^{(k')}]+\sum_{i=0}^{\infty}\mathbf{E}[Q_i^{(k')}]\mathbf{E}[\Phi_i^{(k')}]+\lambda_p^{(k')}\mathbf{E}[W_s^{(k')}]\mathbf{E}[X_p^{(k')}] \ , \tag{5.11}$$

where $\mathbf{E}[R_s^{(k')}]$ is the average residual effective service time of channel k'. That is, $\mathbf{E}[R_s^{(k')}]$ is the remaining time to complete the service of the connection being served at channel k'. This connection being served can be the primary connection or the type-i secondary connection. Furthermore, $\mathbf{E}[Q_p^{(k')}]\mathbf{E}[X_p^{(k')}]$ and $\sum_{i=0}^{\infty}\mathbf{E}[Q_i^{(k')}]\mathbf{E}[\Phi_i^{(k')}]$ in (5.11) are the cumulative workload resulting from the primary connections and the secondary connections in the present queues of channel k', respectively. Moreover, the fourth term ($\lambda_p^{(k')}\mathbf{E}[W_s^{(k')}]\mathbf{E}[X_p^{(k')}]$) in (5.11) is the cumulative workload resulting from the arrivals of the primary connections during $W_s^{(k')}$.

In (5.11), the closed-form expression for $\mathbf{E}[\Phi_i^{(k')}]$ is derived in Appendix C. Next, we will derive $\mathbf{E}[R_s^{(k')}]$, $\mathbf{E}[Q_p^{(k')}]$, and $\mathbf{E}[Q_i^{(k')}]$. Firstly, according to

tribution, the arrivals of the interrupted secondary connections at channel k' also follow Poisson distribution. Applying the property of Poisson arrivals see time average (PASTA) on the arrivals of the interrupted secondary connections at channel k' [153], all of them must spend time duration $\mathbf{E}[W_s^{(k')}]$ on average to wait for an idle channel k'. This waiting time is uncorrelated to the number of interruptions.

the definition of residual time in [154], we have

$$\mathbf{E}[R_s^{(k')}] = \frac{1}{2}\lambda_p^{(k')}\mathbf{E}[(X_p^{(k')})^2] + \frac{1}{2}\sum_{i=0}^{\infty}\omega_i^{(k')}\mathbf{E}[(\Phi_i^{(k')})^2] \ , \tag{5.12}$$

where $\omega_i^{(k')}$ is derived in Appendix B. Secondly, according to Little's formula, it follows that

$$\mathbf{E}[Q_p^{(k')}] = \lambda_p^{(k')}\mathbf{E}[W_p^{(k')}] \ , \tag{5.13}$$

where $\mathbf{E}[W_p^{(k')}]$ is the average waiting time of the primary connections at channel k'. It is the duration from the time instant that a primary connection enters the high-priority queue of channel k' until it gets a chance to transmit at channel k'. Hence, it follows that

$$\mathbf{E}[W_p^{(k')}] = \mathbf{E}[R_p^{(k')}] + \mathbf{E}[Q_p^{(k')}]\mathbf{E}[X_p^{(k')}] \ , \tag{5.14}$$

where $\mathbf{E}[R_p^{(k')}]$ is the average residual service time resulting from only the primary connections of channel k' and $\mathbf{E}[Q_p^{(k')}]\mathbf{E}[X_p^{(k')}]$ is the total workload of the primary connections in the present high-priority queue of channel k'. According to [154], we have $\mathbf{E}[R_p^{(k')}] = \frac{1}{2}\lambda_p^{(k')}\mathbf{E}[(X_p^{(k')})^2]$. Then, solving (5.13) and (5.14) simultaneously yields

$$\mathbf{E}[W_p^{(k')}] = \frac{\mathbf{E}[R_p^{(k')}]}{1-\rho_p^{(k')}} = \frac{\lambda_p^{(k')}\mathbf{E}[(X_p^{(k')})^2]}{2(1-\lambda_p^{(k')}\mathbf{E}[X_p^{(k')}])} \ , \tag{5.15}$$

and

$$\mathbf{E}[Q_p^{(k')}] = \frac{\lambda_p^{(k')}\mathbf{E}[R_p^{(k')}]}{1-\rho_p^{(k')}} = \frac{(\lambda_p^{(k')})^2\mathbf{E}[(X_p^{(k')})^2]}{2(1-\lambda_p^{(k')}\mathbf{E}[X_p^{(k')}])} \ . \tag{5.16}$$

Next, according to Little's formula, we can obtain

$$\mathbf{E}[Q_i^{(k')}] = \omega_i^{(k')}\mathbf{E}[W_s^{(k')}] \ . \tag{5.17}$$

Finally, substituting (5.12), (5.16), and (5.17) into (5.11), we can obtain (5.7).

□

Proposition 2.

$$p_i^{(k)} = \begin{cases} \lambda_p^{(k)} \boldsymbol{E}[\Phi_i^{(k)}] & , \quad k = s_{i,\eta} \\ 0 & , \quad k \neq s_{i,\eta} \end{cases} . \tag{5.18}$$

Proof. The value of $p_i^{(k)}$ can be evaluated as follows. Because the considered secondary connection will operate at channel $s_{i,\eta}$ after i^{th} interruption, we have $p_i^{(k)} = 0$ when $k \neq s_{i,\eta}$. Furthermore, for the case that $k = s_{i,\eta}$, we consider the time interval $[0, t]$ at channel k. Total $\lambda_p^{(k)} t$ primary connections and $\omega_i^{(k)} t$ type-i secondary connections arrive at channel k during this interval. Hence, there are total $\omega_i^{(k)} t p_i^{(k)}$ type-i secondary connections will be interrupted on average during this interval. Furthermore, applying the property of Poisson arrivals see time average (PASTA) on the arrivals of the primary connections [153], we can obtain the probability of a primary connection finding channel k being occupied by the type-i secondary connections is $\rho_i^{(k)}$. Thus, during this interval, the total $\lambda_p^{(k)} t \rho_i^{(k)}$ primary connections can see a busy channel being occupied by the type-i secondary connections. For each primary connection, it can interrupt only one secondary connection when it arrives at a busy channel being occupied by the secondary connection because only one secondary user can transmit at any instant of time. Thus, the total number of the interrupted secondary connections at channel k is also $\lambda_p^{(k)} t \rho_i^{(k)}$. Hence, we have $\omega_i^{(k)} t p_i^{(k)} = \lambda_p^{(k)} t \rho_i^{(k)}$. That is,

$$\rho_i^{(k)} = \frac{\omega_i^{(k)}}{\lambda_p^{(k)}} p_i^{(k)} . \tag{5.19}$$

Next, we consider a type-i secondary connection at channel k. Before the $(i+1)^{th}$ interruption event occurs, its effective service time is $\mathbf{E}[\Phi_i^{(k)}]$. Thus, from queueing theory, we can have

$$\rho_i^{(k)} = \omega_i^{(k)} \mathbf{E}[\Phi_i^{(k)}] . \tag{5.20}$$

Comparing (5.19) and (5.20), we can obtain (5.18). □

5.5 Applications to Performance Analysis in IEEE 802.22

To demonstrate the usefulness of the developed analytical method, we apply these analytical results in Section 5.4 on two typical target channel sequences used in the IEEE 802.22 WRAN standard[4]. Specifically, we consider the *always-staying* and *always-changing* spectrum handoff sequences, which are respectively introduced in the non-hopping mode and the phase-shifting hopping mode of the IEEE 802.22 standard [126]. From the analytical results, an adaptive target channel selection approach can be provided.

5.5.1 Derivation of Extended Data Delivery Time

For the *always-staying* sequence, a secondary connection always stays on its default channel η when it is interrupted. That is, its target channel sequence can be expressed as (Chη, Chη, Chη, $\cdots$) and thus $s_{i,\eta} = \eta$ for each i. Hence, we can have $\mathbf{E}[D_i] = \mathbf{E}[Y_p^{(\eta)}]$ for each i in (5.4). Then, the average extended data delivery time of the secondary connections for the always-staying sequence can be expressed as follows:

$$\mathbf{E}[T_{stay}] = \mathbf{E}[X_s^{(\eta)}] + \sum_{n=1}^{\infty} \left(\sum_{i=1}^{n} \mathbf{E}[Y_p^{(\eta)}] \right) (1 - p_i^{(\eta)}) \prod_{i=0}^{n-1} p_i^{(\eta)} \ . \qquad (5.21)$$

Next, we consider the *always-changing* sequence. In this case, the secondary connection sequentially changes its operating channel to the next

[4] In fact, the analytical results can help the secondary users find the best target channel sequence to minimize its extended data delivery time by comparing the extended data delivery time resulting from all possible target channel sequences.

neighboring channel. Without loss of generality, its corresponding target channel sequence can be expressed as (Ch$\eta+1$, Ch$\eta+2$, $\cdots$, ChM, Ch1, Ch2, $\cdots$, Chη, Ch$\eta+1$, $\cdots$), where channel η is the default channel of the secondary connection. That is, at the i^{th} interruption, the target channel of the interrupted secondary connection is channel $s_{i,\eta} \equiv \mathcal{MOD}(i+\eta, M)$ where $\mathcal{MOD}(a,b)$ is the Modulus function and it returns the remainder resulting when a is divided by b. Hence, we have $\mathbf{E}[D_i] = \mathbf{E}[W_s^{(s_{i,\eta})}] + t_s$ for each i in (5.4). Thus, the average extended data delivery time of the secondary connections for the always-changing sequence can be expressed as follows:

$$\mathbf{E}[T_{change}] = \mathbf{E}[X_s^{(\eta)}] + \sum_{n=1}^{\infty}\left[\left(\sum_{i=1}^{n}(\mathbf{E}[W_s^{(s_{i,\eta})}] + t_s)\right)(1-p_i^{(s_{n,\eta})})\prod_{i=0}^{n-1}p_i^{(s_{i,\eta})}\right] \ . \tag{5.22}$$

Based on the analytical results, the secondary connection can adaptively adopt the better target channel sequence to reduce its extended data delivery time. Thus, the average extended data delivery time with this adaptive channel selection principle (denoted by $\mathbf{E}[T^*]$) can be expressed as follows:

$$\mathbf{E}[T^*] = \min\left(\mathbf{E}[T_{stay}], \mathbf{E}[T_{change}]\right) \ . \tag{5.23}$$

5.5.2 An Example for Homogeneous Traffic Loads

Now, we give an example to explain how to apply our analytical results to find the better target channel sequence when traffic parameters are given. We consider a special case that the primary and the secondary connections have the same traffic parameters in a three-channel system (i.e., $\lambda_p^{(1)} = \lambda_p^{(2)} = \lambda_p^{(3)} \equiv \lambda_p$, $\lambda_s^{(1)} = \lambda_s^{(2)} \equiv \lambda_s$, and $\mathbf{E}[X_p^{(1)}] = \mathbf{E}[X_p^{(2)}] = \mathbf{E}[X_p^{(3)}] \equiv \mathbf{E}[X_p]$). Because the three channels are identical, three channels have the same performance metrics. Thus, the superscript (k) can be dropped to ease the notations. Furthermore, we assume that the service time of the secondary connections follows the same exponential distribution, i.e., $f_s^{(1)}(x) = f_s^{(2)}(x) = f_s^{(3)}(x) \equiv f_s(x) = \mu_s e^{-\mu_s x}$. Hence, we have $\mathbf{E}[X_s^{(1)}] = \mathbf{E}[X_s^{(2)}] = \mathbf{E}[X_s^{(3)}] \equiv \mathbf{E}[X_s] = \frac{1}{\mu_s}$.

Derivation of $p_i^{(\eta)}$ and $\mathbf{E}[Y_p^{(\eta)}]$ in (5.21)

First, according to Appendix C, we can derive $\mathbf{E}[\Phi_i^{(\eta)}]$ as follows:

$$\mathbf{E}[\Phi_i^{(\eta)}] = \mathbf{E}[\Phi_i] = \frac{1}{\lambda_p + \mu_s} \ . \tag{5.24}$$

Then, the value of $p_i^{(\eta)}$ can be derived from (5.18) as follows:

$$p_i^{(\eta)} = \lambda_p^{(\eta)} \mathbf{E}[\Phi_i^{(\eta)}] = \frac{\lambda_p}{\lambda_p + \mu_s} \equiv p_i \ . \tag{5.25}$$

Next, referring to (5.6), it follows that

$$\mathbf{E}[Y_p^{(\eta)}] = \mathbf{E}[Y_p] = \frac{\mathbf{E}[X_p]}{1 - \lambda_p \mathbf{E}[X_p]} \ . \tag{5.26}$$

Finally, substituting (5.25) and (5.26) into (5.21), we can obtain the closed-form expression for the extended data delivery time with the always-staying target channel sequence.

Derivation of $\mathbf{E}[W_s^{(s_{i,\eta})}]$ and $p_i^{(s_{i,\eta})}$ in (5.22)

Referring to Appendixes B and C, we can have

$$\omega_i^{(s_{i,\eta})} = \omega_i = \lambda_s (\frac{\lambda_p}{\lambda_p + \mu_s})^i \ , \tag{5.27}$$

and

$$\mathbf{E}[(\Phi_i^{(s_{i,\eta})})^2] = \mathbf{E}[(\Phi_i)^2] = \frac{2}{(\lambda_p + \mu_s)^2} \ . \tag{5.28}$$

Next, substituting (5.24), (5.27), and (5.28) into (5.7), we can have

$$\mathbf{E}[W_s^{(s_{i,\eta})}] = \mathbf{E}[W_s] = \frac{\lambda_p \mathbf{E}[(X_p)^2] + \frac{2\lambda_s \mathbf{E}[X_s]}{(\lambda_p + \mu_s)} + \frac{(\lambda_p)^2 \mathbf{E}[(X_p)^2]}{1 - \lambda_p \mathbf{E}[X_p]} \mathbf{E}[X_p]}{2(1 - \lambda_p \mathbf{E}[X_p] - \lambda_s \mathbf{E}[X_s])} \ . \tag{5.29}$$

Then, referring to (5.18), it follows that

$$p_i^{(s_{i,\eta})} = p_i = \frac{\lambda_p}{\lambda_p + \mu_s} \ . \tag{5.30}$$

Finally, substituting (5.29) and (5.30) into (5.22), we can obtain the closed-form expression for the extended data delivery time with the always-changing target channel sequence. Note that this closed-form expression for p_i in this special case had been discussed in [127]. However, [127] cannot extend to the case with the generally distributed service time.

In summary, the average extended data delivery time with our adaptive target channel selection approach can be expressed as follows:

$$\mathbf{E}[T^*] = \begin{cases} \mathbf{E}[T_{stay}] & , \quad \mathbf{E}[Y_p] \leq \mathbf{E}[W_s] + t_s \\ \mathbf{E}[T_{change}] & , \quad \mathbf{E}[Y_p] \geq \mathbf{E}[W_s] + t_s \end{cases} . \tag{5.31}$$

Note that the always-staying and the always-changing sequences have the same extended data delivery time when $\mathbf{E}[Y_p] = \mathbf{E}[W_s] + t_s$.

5.6 Numerical Results

We show numerical results to reveal the importance of the three key design features for modeling spectrum handoffs as discussed in Section 5.1, which consist of (1) generally distributed service time; (2) various operating channels; and (3) queueing behaviors of multiple secondary connections.

5.6.1 Simulation Setup

In order to validate the proposed analytical model, we perform simulations in non-slot-based (continuous-time) cognitive radio systems, where the inter-arrival time and service time can be the duration of non-integer time slots. We consider a three-channel CR system with Poisson arrival processes of rates

λ_p and λ_s for the high-priority primary connections and the low-priority secondary connections, respectively. The high-priority connections can interrupt the transmissions of the low-priority connections, and the connections with the same priority follow the first-come-first-served (FCFS) scheduling discipline[5]. Referring to the IEEE 802.22 standard, we adopt time slot duration of 10 msec in our simulations [145].

5.6.2 Effects of Various Service Time Distributions for Primary Connections

Firstly, we investigate the effects of various service time distributions for primary connections on the extended data delivery time of the secondary connections. The truncated Pareto distribution and the exponential distribution are considered in our simulations. Referring to [142], these two distributions match the actual data and voice traffic measurements very well, respectively. The truncated Pareto distribution is expressed as follows:

$$f_X(x) = \begin{cases} \alpha\frac{K^\alpha}{x^{\alpha+1}} & , \quad K \leq x \leq m \\ \frac{K^\alpha}{m^\alpha} & , \quad x = m \end{cases} . \tag{5.32}$$

According to [155], the traffic shaping parameter $\alpha = 1.1$ and the scale parameter $K = 81.5$, and the truncated upper bound $m = 66666$ bytes in (5.32). Then, the average connection length is 480 bytes for the primary connections. If the exponentially distributed primary connections are considered, the average connection length is also 480 bytes. Moreover, we assume that $\mathbf{E}[X_s^{(1)}] = \mathbf{E}[X_s^{(2)}] = \mathbf{E}[X_s^{(3)}] \equiv \mathbf{E}[X_s] = 10$ (slots/arrival),

[5] In fact, the analytical results of mean values obtained in this report can be applied to other scheduling discipline which is independent of the service time of the primary and secondary connections because the averages of system performance metrics will be invariant to the order of service in this case (see page 113 in [134]).

and $\mathbf{E}[X_p^{(1)}] = \mathbf{E}[X_p^{(2)}] = \mathbf{E}[X_p^{(3)}] \equiv \mathbf{E}[X_p]$. When the data rate of the primary connections is 19.2 Kbps, we have $\mathbf{E}[X_p] = \frac{480\times 8 \text{ bits}}{19.2 \text{ Kbps}} \div \frac{10 \text{ msec}}{\text{slot}} = 20$ (slots/arrival) for the Pareto and the exponential distributions. Furthermore, we consider that $\lambda_s = 0.01$ (arrivals/slot). Recall that ρ_p is the channel busy probability resulting from the transmissions of the primary connections. We only consider the case that $0 \leq \rho_p < 1 - \lambda_s \mathbf{E}[X_s] = 0.9$ in the following numerical results. When $\rho_p + \lambda_s \mathbf{E}[X_s] \geq 1$ (or equivalently $\lambda_p \geq \frac{\lambda_s \mathbf{E}[X_s]}{\mathbf{E}[X_p]} = 0.045$ (arrivals/slot)), the secondary connections will encounter the infinite extended data delivery time on average.

Figure 5.3 compares the effects of Pareto and exponential service time distributions for primary connections when the always-changing spectrum handoff sequence is adopted. First, we find that the simulation results match the analytical results quite well, which can validate the slot-based assumption used in our analysis. Next, compared to the exponentially distributed service time for primary connections, the Pareto distributed service time results in longer average extended data delivery time in the secondary connections. This phenomenon can be interpreted as follows. Because of the heavy tail property of Pareto distribution, the second moment $\mathbf{E}[(X_p)^2]$ of service time with Pareto distribution is larger than that with exponential distribution. According to (5.29) and (5.22), an interrupted secondary connection will encounter longer waiting time and extended data delivery time when the primary connections' service time distribution is Pareto. For example, when $\rho_p = 0.44$ or equivalently $\lambda_p = \frac{\rho_p}{\mathbf{E}[X_p]} = 0.022$ (arrivals/slot), the average extended data delivery time with the Pareto-typed primary connection service time is four times longer than that with the exponential-typed primary connection service time. Because the developed analytical framework can characterize the effects of generally distributed service time, it is quite useful.

When the always-staying spectrum handoff sequence is adopted, Fig. 5.4 shows the average extended data delivery time of the secondary connections. According to (5.21), the extended data delivery time in this case is related to the average busy period $\mathrm{E}[Y_p]$ for the primary connections. Because the considered Pareto and exponential distributions have the same average service time, these two distributions result in the same average busy period $\mathrm{E}[Y_p]$ for the primary connections according to (5.26), resulting in the same average extended data delivery time as well.

5.6.3 Traffic-adaptive Target Channel Selection Principle

Figure 5.5 compares the extended data delivery time of the always-staying and the always-changing spectrum handoff sequences when the service time of the primary connections is exponentially distributed. Based on (5.31), the traffic-adaptive channel selection approach can appropriately change to better target channel sequence according to traffic conditions. We can see that both the always-staying and the always-changing sequences result in the same extended data delivery time when $\rho_p = 0.44$ or equivalently $\lambda_p = \frac{\rho_p}{\mathbf{E}[X_p]} = 0.022$ (arrivals/slot). When $\rho_p > 0.44$, the interrupted user prefers the always-staying sequence. This phenomenon can be interpreted as follows. A larger value of ρ_p (or equivalently a larger value of λ_p) will increase the probability that an interrupted secondary user experiences long waiting time when it changes its operating channel. As a result, the average handoff delay for changing operating channel (i.e., $\mathbf{E}[W_s] + t_s$) will be extended. Then, the average extended data delivery time will be also prolonged. In our case, the secondary user prefers staying on the current

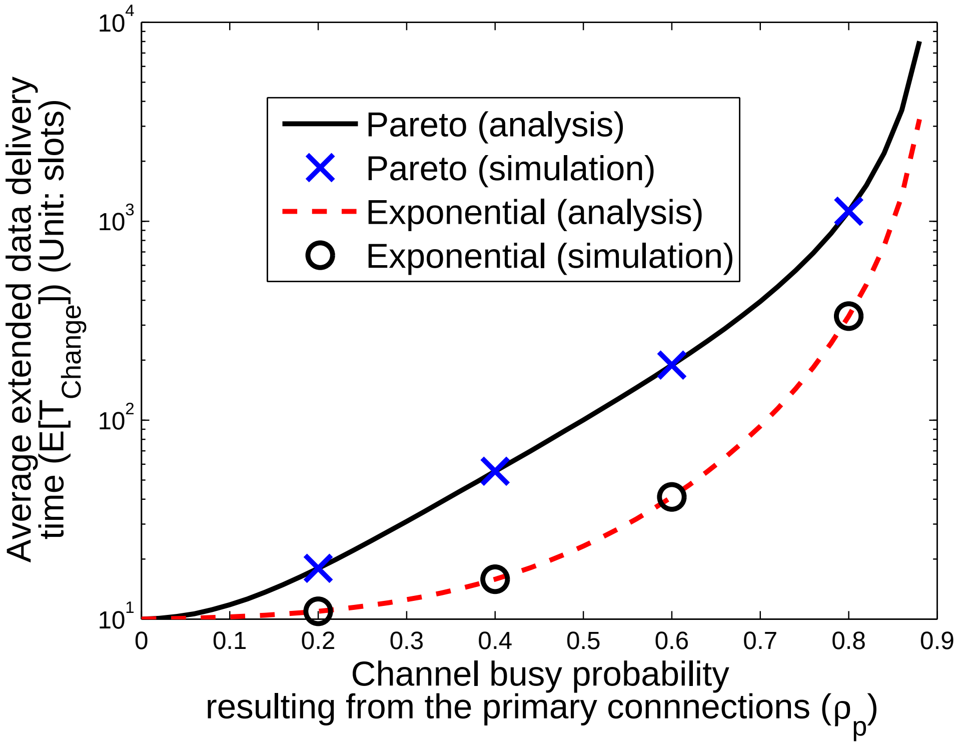


Figure 5.3: Effects of Pareto and exponential service time distributions for primary connections on the extended data delivery time ($\mathbf{E}[T_{change}]$) of the secondary connections when the **always-changing** spectrum handoff sequence is adopted, where $t_s = 1$ (slot), $\lambda_s = 0.01$ (arrivals/slot), $\mathbf{E}[X_s] = 10$ (slots/arrival), and $\mathbf{E}[X_p] = 20$ (slots/arrival).

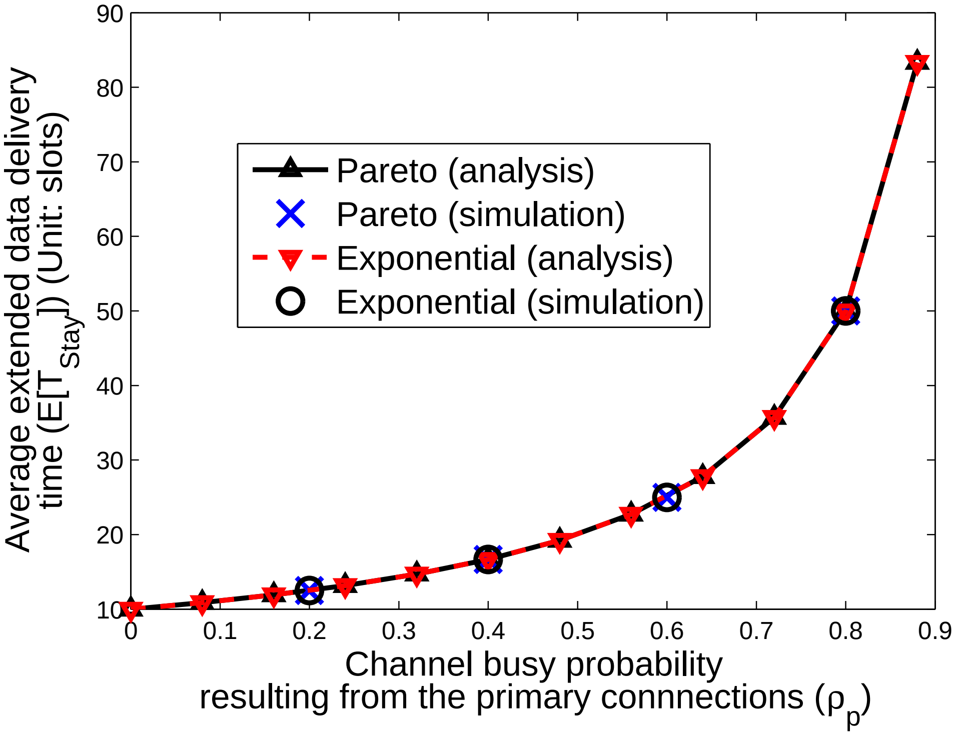


Figure 5.4: Effects of Pareto and exponential service time distributions for primary connections on the extended data delivery time ($\mathbf{E}[T_{stay}]$) of the secondary connections when the **always-staying** spectrum handoff sequence is adopted, where $t_s = 1$ (slot), $\lambda_s = 0.01$ (arrivals/slot), $\mathbf{E}[X_s] = 10$ (slots/arrival), and $\mathbf{E}[X_p] = 20$ (slots/arrival).

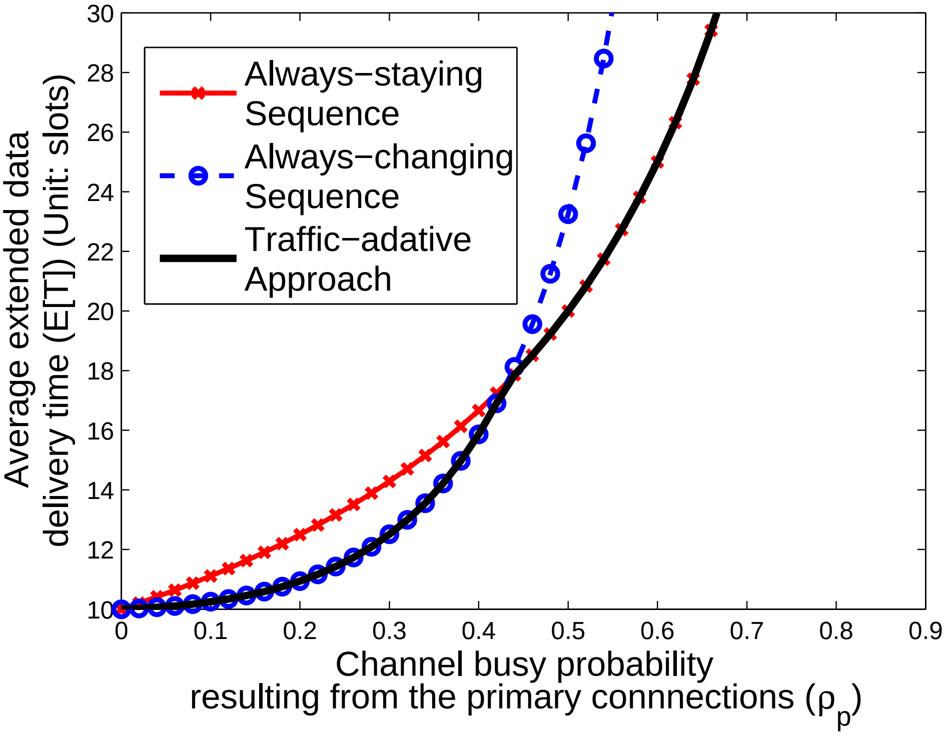


Figure 5.5: Comparison of the extended data delivery time for the always-staying and always-changing spectrum handoff sequences as well as the traffic-adaptive channel selection approach, where $t_s = 1$ (slot), $\lambda_s = 0.01$ (arrivals/slot), $\mathbf{E}[X_p] = 20$ (slots/arrival), and $\mathbf{E}[X_s] = 10$ (slots/arrival).

operating channel when $\rho_p > 0.44$. By contrast, when $\rho_p < 0.44$, the traffic-adaptive channel selection approach can improve latency performance by changing to the always-changing sequence. For example, when $\rho_p = 0.2$, the traffic-adaptive approach can improve the extended data delivery time by 15% compared to the always-staying sequence. Compared to the single-channel spectrum handoff model [22–24, 32, 51–55], the developed analytical framework for multi-channel spectrum handoff is more general because it can incorporate the effects of changing operating channels.

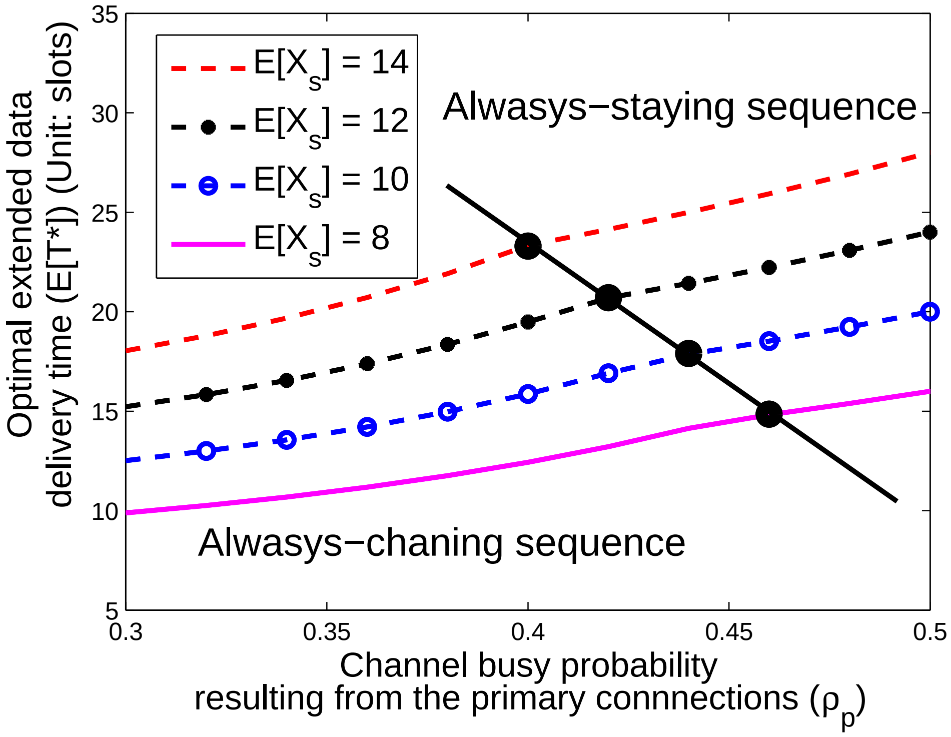


Figure 5.6: Effects of secondary connections' service time $\mathbf{E}[X_s]$ on the cross-point for the traffic-adaptive channel selection approach, where $t_s = 1$ (slot), $\mathbf{E}[X_p] = 20$ (slots/arrival), and $\lambda_s = 0.01$ (arrivals/slot).

Figure 5.6 shows the effect of secondary connections' service time $\mathbf{E}[X_s]$ on the cross-point for traffic-adaptive channel selection approach. According to (5.31), for a lager value of $\mathbf{E}[X_s]$, the interrupted secondary connection prefers staying on the current channel because the average handoff delay for changing its operating channel is longer than that for staying on the current channel. Thus, the cross-point of "always-staying" and "always-changing" sequences moves toward left-hand side as $\mathbf{E}[X_s]$ increases as seen in the figure.

The analytical results developed in this chapter can be used to design the admission control rule for the arriving secondary users subject to their latency requirement. Fig. 5.7 shows the admissible region for the normalized traffic workloads (or channel utilities) (ρ_p, ρ_s)[6] for the Voice over IP (VoIP) services. The maximum allowable average cumulative delay resulting from multiple handoffs is 20 ms for the VoIP traffic [156]. Assume $\mathbf{E}[X_p] = 20$ (slots/arrival) and $\mathbf{E}[X_s] = 10$ (slots/arrival). The admission control policy can be designed according to this figure. When $\rho_p < 0.166$, a CR network can accept all arrival requests from the secondary users until the CR network is saturated, i.e., $\rho_p + \rho_s \simeq 1$. Furthermore, when $0.166 < \rho_p < 0.312$, a part of traffic workloads of the secondary users must be rejected in order to satisfy the delay constraint for the secondary users. In this case, $0.31 < \rho_p + \rho_s < 0.645$. For example, when $\rho_p = 0.25$, a CR network can support at most 0.214 workload for the secondary users. That is, a CR network can accept at most $\lambda_s = 0.0214$ (arrivals/slot) based on the results shown in the figure when $\lambda_p = 0.0125$ (arrivals/slot). In order to design the most allowable λ_s to achieve this arrival rate upper bound for the secondary connections, many arrival-rate control methods can be considered, such as the p-persistent carrier sense multiple access (CSMA) protocol in [25] and the call admission control mechanisms in [50, 63, 157]. Finally, when $\rho_p > 0.312$, no secondary user can be accepted.

5.6.4 Performance Comparison between Different Channel Selection Methods

Now we compare the extended data delivery time of the following three schemes: (1) the slot-based target channel selection scheme; (2) the random-based target channel selection scheme; and (3) the traffic-adaptive target channel selection scheme. We consider a three-channel network with various

[6] $\rho_p = \lambda_p \mathbf{E}[X_p]$ and $\rho_s = \lambda_s \mathbf{E}[X_s]$.

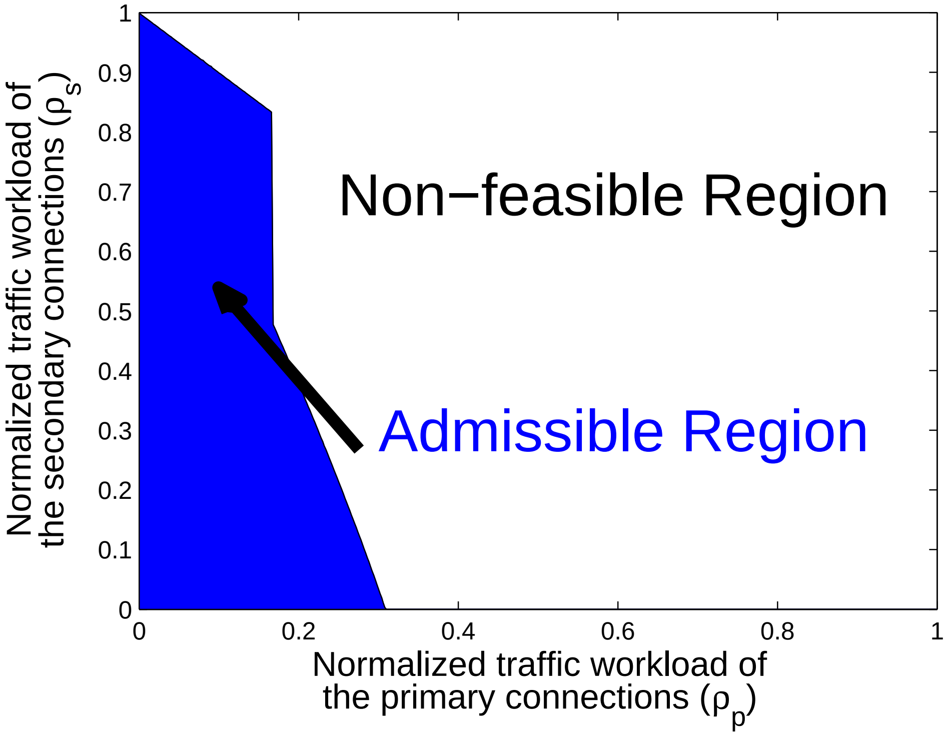


Figure 5.7: Admissible region for the normalized traffic workloads (ρ_p, ρ_s), where the average cumulative delay constraint can be satisfied when $t_s = 0$ (slot), $\mathbf{E}[X_p] = 20$ (slots/arrival) and $\mathbf{E}[X_s] = 10$ (slots/arrival).

traffic loads, where $\lambda_p^{(1)} = \lambda_p^{(2)} = \lambda_p^{(3)} \equiv \lambda_p$, $\lambda_s^{(1)} = \lambda_s^{(2)} = \lambda_s^{(3)} \equiv 0.01$ (arrivals/slot), $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}]) = (5, 15, 25)$ (slots/arrival), and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}]) = (15, 15, 15)$ (slots/arrival). For the slot-based scheme, the secondary connections prefer selecting the channel which has the lowest busy probability resulting from the primary connections in each time slot. That is, when handoff procedures are initiated in the beginning of each time slot, all the secondary connections will select channel 1 to be their target channels. Furthermore, the random-based scheme selects one channel out of all the three channels for the target channel. Hence, each channel is selected with probability 1/3. Moreover, based on the considered traffic parameters, the traffic-adaptive scheme will adopt the always-changing sequence and the always-staying sequence when $\lambda_p \leq 0.018$ (arrivals/slot) and $\lambda_p \geq 0.018$ (arrivals/slot), respectively. The three target channel selection schemes result in various target channel sequences. Based on the proposed analytical model, we can evaluate the average extended data delivery time resulting from these target channel sequences.

Figure 5.8 compares the extended data delivery time of the three target channel selection methods. We have the following three important observations. First, we consider $\lambda_p < 0.018$ (arrivals/slot). Because the probability of changing operating channel is higher than that of staying on the current operating channel for the interrupted secondary user in the random-based scheme, we can find that the average extended data delivery time for the random-based target channel selection scheme is similar to that for the traffic-adaptive target channel selection scheme, which adopts the always-changing sequence. Secondly, when $\lambda_p > 0.018$ (arrivals/slot), the traffic-adaptive scheme can shorten the average extended data delivery time because it adopts the always-staying sequence. For a larger value of λ_p, the

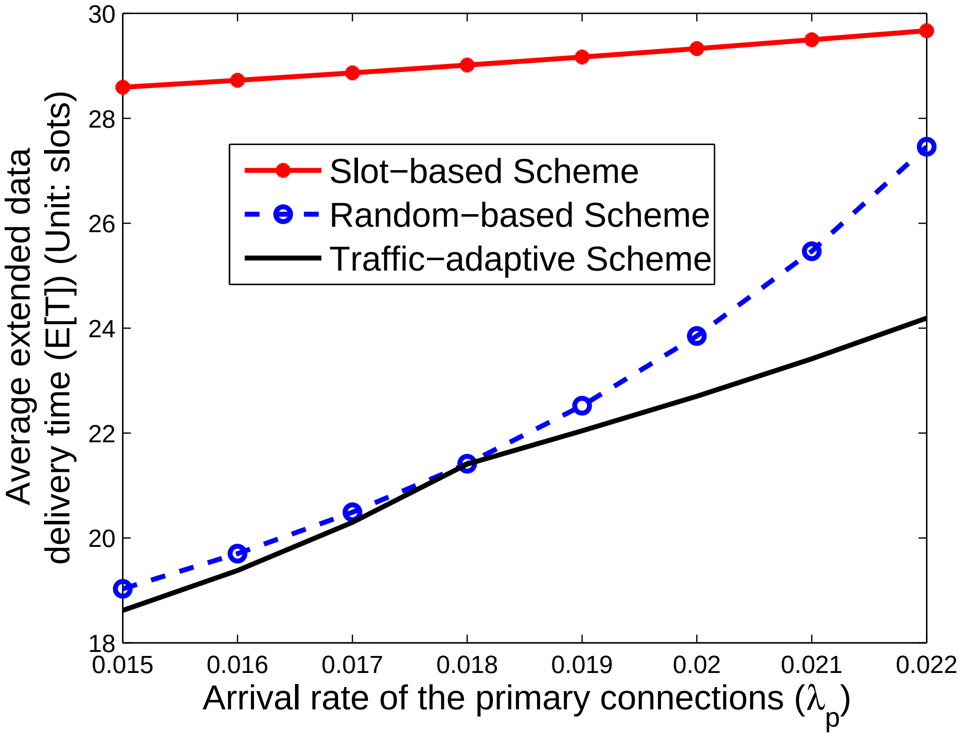


Figure 5.8: Comparison of average extended data delivery time for different target channel selection sequences.

traffic-adaptive scheme can improve the extended data delivery time more significantly. Thirdly, it is shown that the random-based and traffic-adaptive schemes can result in shorter extended data delivery time compared to the slot-based scheme. For example, when $\lambda_p = 0.018$, the random-based and traffic-adaptive schemes can improve the extended data delivery time by 35% compared to the slot-based scheme. This is because the slot-based scheme ignores the queueing behaviors of the secondary connections.

Chapter 6

Optimal Proactive Spectrum Handoff

Extended to the discussions of the proactive spectrum handoff in Chapter 5, we further investigate how to predetermine the optimal target channel sequence for future handoffs. We incorporate two important features in the design of spectrum handoff to ensure the quality of service (QoS) for the secondary users. First, due to multiple interruptions from the primary users in each secondary user's connection, a series of spectrum handoffs are considered in our model. Secondly, we consider the impacts of the traffic statistics of both the primary and secondary users on the handoff delay.

In this chapter, we formulate an optimization problem of finding a target channel sequence for multiple handoffs with the objective of minimizing the cumulative delay per connection for a newly arriving secondary user. We will simultaneously consider two design features in spectrum handoffs: (1) multiple spectrum handoffs and (2) various service time of the primary and secondary users. The contributions of this chapter can be summarized in the following:

- We propose a dynamic-programming-based algorithm with time complexity of $O(LM^2)$ to find an optimal target channel sequence with minimum cumulative spectrum handoff delay, where L and M are the length of the target channel sequence and the total number of candidate channels for spectrum handoffs, respectively.

- Furthermore, a low-complexity greedy algorithm is proposed to find the suboptimal solution with time complexity of $O(M)$. We prove that only six permutations of the target channel sequences are required to be compared, and demonstrate that it can approach the optimal solution.

6.1 Problem Formulation

The extended data delivery time is an important QoS performance metric for secondary users from a connection viewpoint. The extended data delivery time per connection consists of the service time of one connection and the cumulative handoff delay resulting from multiple handoffs. Because the cumulative handoff delay depends on which channels are selected when the primary users' interruptions occur, one of important issue for the secondary users is to search the best *target channel sequence*.

We consider a CR network $\mathcal{G}$ with M independent channels, where the target channel sequence for future spectrum handoffs is determined proactively for each newly arriving secondary user. For a secondary user with default channel $s_0 \triangleq \eta$, we denote its target channel sequence as $\boldsymbol{s}(\eta) \triangleq (s_1, s_2, s_3, \cdots)$ where $s_{i,\eta}$ is the target channel for spectrum handoff at the i^{th} interruption. Next, we formulate a **Cumulative Handoff Delay Minimization Problem** for multiple spectrum handoffs. Given a set of candidate channels $\Omega = \{1, 2, \ldots, M\}$ and the required length L of the target channel sequence for L spectrum handoffs, we aim to determine a target channel sequence (denoted by $\boldsymbol{s}(\eta)^*$) to minimize the average cumulative handoff delay $\mathbf{E}[D(\boldsymbol{s}(\eta))]$ for a newly arriving secondary user's connection. Formally, we have

$$\boldsymbol{s}(\eta)^* = \underset{\forall \boldsymbol{S}(\eta)\in\Omega^L}{\arg\min}\, \mathbf{E}[D(\boldsymbol{s}(\eta))] \ , \tag{6.1}$$

where $\mathbf{E}[\cdot]$ is the expectation function. In the next section, the closed-form expression for $\mathbf{E}[D(\boldsymbol{s}(\eta))]$ will be derived given the arrival rates and service time distributions of both primary and secondary users.

6.2 Cumulative Handoff Delay Analysis

In this section, we derive the closed-form expression for the average cumulative handoff delay with different target channel sequence $\boldsymbol{s}(\eta)$ of the newly arriving secondary user's connection. To ease notation, we denote $\boldsymbol{s}$ for $\boldsymbol{s}(\eta)$ in the rest of this chapter. Let N be the total number of interruptions in the considered connection. According to the total probability principle, it follows that

$$\begin{aligned} \mathbf{E}[D(\boldsymbol{s})] &= \sum_{n=1}^{L} \mathbf{Pr}\{N=n\}\mathbf{E}[D(\boldsymbol{s})|N=n] \\ &= \sum_{n=1}^{L}\left[\mathbf{Pr}\{N=n\}\sum_{i=1}^{n}\mathbf{E}[d(s_{i-1},s_i)]\right] \ , \end{aligned} \tag{6.2}$$

where $d(s_{i-1}, s_i)$ is the handoff delay when the interrupted secondary users change their operating channel from channel s_{i-1} to s_i.

Firstly, we evaluate $\mathbf{E}[d(s_{i-1}, s_i)]$ in (6.2). When a primary user's connection appears at the channel being occupied by the newly arriving secondary user's connection, an interruption event occurs. The spectrum handoff delay depends on which channel is selected for the target channel. The interrupted secondary users can either stay on the current channel or change to another channel. If the considered secondary user's connection chooses to stay on its current operating channel (i.e., $s_{i-1} = s_i$), the expected handoff delay is the duration from the time instant that current operating channel is used by the primary users' connections until this channel becomes idle. This duration is called the *busy period* (denoted by $Y_p^{(s_i)}$) resulting from the transmissions of multiple primary users' connections at channel s_{i-1}. In the other case, the considered secondary user's connection may change its operating channel (i.e., $s_{i-1} \neq s_i$). After switching channel from channel s_{i-1} to s_i, the considered secondary user's connection cannot be resumed until all the present primary and secondary users' connections at channel s_i are served. Let $\mathbf{E}[W_s^{(s_i)}]$ be the waiting time[1] for the secondary users' connections at channel s_i. Then, the expected handoff delay is the sum of $\mathbf{E}[W_s^{(s_i)}]$ and the channel switching time t_s. Thus,

$$\mathbf{E}[d(s_{i-1}, s_i)] = \begin{cases} \mathbf{E}[Y_p^{(s_i)}] & , \quad s_{i-1} = s_i \\ \mathbf{E}[W_s^{(s_i)}] + t_s & , \quad s_{i-1} \neq s_i \end{cases} . \tag{6.3}$$

Next, we evaluate $\mathbf{Pr}\{N = n\}$ in (6.2). Denote $p_i^{(s_i)}$ as the probability that a secondary user's connection is interrupted by the arrival of primary user's connection again at channel s_i after i interruptions. Then, the probability that the considered secondary user's connection is interrupted exactly

[1] A secondary user's connection needs to change its operating channel only when a primary user's connection appears. Because the arrivals of the primary users' connections follow Poisson distribution, the arrivals of the interrupted secondary users' connections at channel s_i also follow Poisson distribution. Applying the property of Poisson arrivals see time average (PASTA) on the arrivals of the interrupted secondary users' connections at channel s_i [153], all of them must spend time duration $\mathbf{E}[W_s^{(s_i)}]$ on average to wait for an idle channel s_i. This waiting time is uncorrelated to the number of interruptions.

n times can be expressed as

$$\mathbf{Pr}\{N = n\} = (1 - p_n^{(s_n)}) \prod_{i=0}^{n-1} p_i^{(s_i)} \ . \tag{6.4}$$

Now, we apply the proposed preemptive resume priority (PRP) M/G/1 queueing network model in Chapters 3 and 4 [158] to derive the closed-form expressions for $\mathbf{E}[d(s_{i-1}, s_i)]$ and $\mathbf{Pr}\{N = n\}$. Let $\lambda_p^{(\eta)}$ (arrivals/slot) and $\lambda_s^{(\eta)}$ (arrivals/slot) be the initial arrival rates of the primary users' and secondary users' connections at channel η in $\mathcal{G}$, respectively, and $X_p^{(\eta)}$ (slots/arrival) and $X_s^{(\eta)}$ (slots/arrival) be their corresponding service time, respectively. Furthermore, we assume that the existing secondary users' connections in $\mathcal{G}$ must stay on the current operating channel when they are interrupted. Then, referring to [135], a newly arriving secondary user's connection will experience the following performance measures:

$$\mathbf{E}[Y_p^{(s_i)}] = \frac{\mathbf{E}[X_p^{(s_i)}]}{1 - \lambda_p^{(s_i)}\mathbf{E}[X_p^{(s_i)}]} \ , \tag{6.5}$$

$$\mathbf{E}[W_s^{(s_i)}] = \frac{\lambda_p^{(s_i)}\mathbf{E}[X_p^{(s_i)}]^2 + \lambda_s^{(s_i)}\mathbf{E}[X_s^{(s_i)}]^2}{(1 - \lambda_p^{(s_i)}\mathbf{E}[X_p^{(s_i)}])(1 - \lambda_p^{(s_i)}\mathbf{E}[X_p^{(s_i)}] - \lambda_s^{(s_i)}\mathbf{E}[X_s^{(s_i)}])} \ , \tag{6.6}$$

and

$$p_i^{(s_i)} = \lambda_p^{(s_i)}\mathbf{E}[\Phi_i^{(s_i)}] \ , \tag{6.7}$$

where $\mathbf{E}[\Phi_i^{(k)}]$ is the considered newly arriving secondary connection's transmission duration between the i^{th} and the $(i+1)^{th}$ interruptions at channel k. When the service time (denoted by χ_s) of the considered newly arriving secondary connection is given, we can derive the closed-form expression for $\mathbf{E}[\Phi_i^{(k)}]$ according to [159] and thus $p_i^{(s_i)}$ can be evaluated. For example, when χ_s is geometrically distributed, we can have

$$p_i^{(s_i)} = \frac{\lambda_p^{(s_i)}\mathbf{E}[\chi_s]}{\lambda_p^{(s_i)}\mathbf{E}[\chi_s] + 1} \ . \tag{6.8}$$

Note that the distributions of service time χ_s for the newly arriving secondary user's connection and $X_s^{(\eta)}$ for the existing secondary users' connections can be different in our model. Finally, we can obtain the values of $\mathbf{E}[d(s_{i-1}, s_i)]$ and $\mathbf{Pr}\{N = n\}$ by substituting (6.5) and (6.6) into (6.3) and (6.7) into (6.4), respectively.

The **Cumulative Handoff Delay Minimization Problem** can be solved by exhaustively searching all the possible target channel sequences. Because this brute force must enumerate all M^L possible permutations of the target channel sequences and compute how long each permutation will take to find a target channel sequence with the minimum cumulative handoff delay, the exhaustive search method has the time complexity of $O(M^L)$ and it is infeasible when M is large. Hence, it is critical to design a low-complexity algorithm to solve this optimization problem.

6.3 An Optimal Dynamical Programming Algorithm

In this section, we propose a low-complexity algorithm to solve the **Cumulative Handoff Delay Minimization Problem**. First, we develop a state diagram to characterize the evolution of target channel sequences and their corresponding cumulative handoff delay. We observe that the considered optimization problem has the optimal substructure property, and thus propose a dynamic programming algorithm for the **Cumulative Handoff Delay Minimization Problem** with a time complexity of $O(LM^2)$.

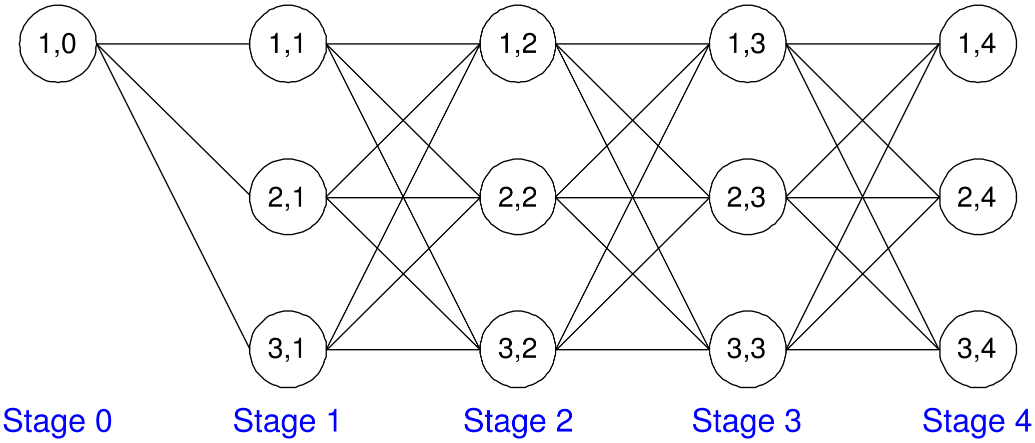


Figure 6.1: An example of state diagram of the target channel sequences for a newly arriving secondary user, where the default channel $\eta = 1$, the number of total channels $M = 3$, and the required length of the target channel sequence $L = 4$. Furthermore, (k, i) stands for the state of operating at the channel k with the i^{th} interruption.

6.3.1 State Diagram for Target Channel Sequences

The proposed state diagram is a two-dimensional chain where state (k, i) represents that channel k is selected for the target channel at the i^{th} interruption. Because the default channel is channel η, the initial state of this state diagram model is $(\eta, 0)$. Furthermore, the set of all possible states for the i^{th} interruption is called stage i. The state transitions occur only at the states between the adjacent stages. Specifically, a transition link from (k, i) to (k', i') exists if $i' = i + 1$, and vice versa. An example of state diagram is shown in Fig. 6.1, where $\eta = 1$, $M = 3$, and $L = 4$.

The cost of state transition shall be proportional to the handoff delay of the interrupted secondary user's connection. For example, the transition from states $(k, i - 1)$ to (k, i) represents the situation that the considered secondary user's connection stays on the current channel k when the i^{th}

interruption event occurs. Hence, in this case, the transition cost shall be proportional to $\mathbf{E}[d(k,k)] = \mathbf{E}[Y_p^{(k)}]$. Furthermore, the transition from states $(k, i-1)$ to (k', i) represents the situation that the considered secondary user's connection changes its operating channel from channel k to k' when the i^{th} interruption event occurs. Hence, the transition cost shall be proportional to $\mathbf{E}[d(k,k')] = \mathbf{E}[W_s^{(k')}] + t_s$. Let $w(k; k', i)$ be the transition cost from state $(k, i-1)$ to state (k', i). Because transition cost is proportional to the handoff delay of the interrupted secondary user's connections, it follows that

$$w(k; k', i) = \nu_i \cdot \mathbf{E}[d(k,k')] \ , \tag{6.9}$$

where ν_i is a normalized factor. It can be obtained by Propositions 3.

Proposition 3. *For the considered secondary user's connections, given the total number of interruptions N and the interrupted probability $p_i^{(s_i)}$ at channel s_i after i interruptions. It follows that*

$$\nu_i = \boldsymbol{Pr}\{N \geq n\} = \prod_{i=0}^{n-1} p_i^{(s_i)} \ . \tag{6.10}$$

Proof. Recall that $\mathbf{E}[D^{(\boldsymbol{S})}]$ is defined as the average cumulative handoff delay of the considered newly arriving secondary user's connection with the target channel sequence $\boldsymbol{s}$. $\mathbf{E}[D^{(\boldsymbol{S})}]$ can be also interpreted as the cumulative cost for the state transition path $(s_0 \longrightarrow s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow \cdots \longrightarrow s_L)$. Hence, it follows that

$$\begin{aligned} \mathbf{E}[D(\boldsymbol{s})] &= \sum_{i=1}^{L} w(s_{i-1}; s_i, i) \\ &= \sum_{i=1}^{L} [\mathbf{E}[d(s_{i-1}, s_i) \cdot \nu_i]] \ . \end{aligned} \tag{6.11}$$

Furthermore, from (6.2), we can have

$$\begin{aligned}\mathbf{E}[D(\boldsymbol{s})] &= \sum_{i=1}^{L} [\mathbf{E}[d(s_{i-1}, s_i)] \cdot \mathbf{Pr}\{N \geq n\}] \\ &= \sum_{i=1}^{L} \left[\mathbf{E}[d(s_{i-1}, s_i)] \cdot \prod_{i=0}^{n-1} p_i^{(s_i)} \right] \ . \qquad (6.12)\end{aligned}$$

Comparing (6.11) and (6.12), we can obtain (6.10). □

6.3.2 Optimal Substructure Property

Next, we show that this optimization problem has the optimal substructure property based on the proposed state diagram. Let $m(k', i)$ be the cumulative cost of the minimum cost path from the initial state $(\eta, 0)$ to the state (k', i) where $i \geq 1$. Then, we have the following recursive relationship:

$$m(k', i+1) = \min_{k \in \Omega} \{m(k, i) + w(k; k', i+1)\} \ , \qquad (6.13)$$

where

$$\begin{aligned} m(k', 1) &= w(\eta; k', 1) \\ &= \begin{cases} \mathbf{E}[Y_p^{(k')}] p_0^{(\eta)} & , \quad k' = \eta \\ (\mathbf{E}[W_s^{(k')}] + t_s) p_0^{(\eta)} & , \quad k' \neq \eta \end{cases} \ . \qquad (6.14)\end{aligned}$$

Based on this optimal substructure, we can build an optimal solution to the considered optimization problem from the optimal solutions to the subproblems. Then, the shortest cumulative handoff delay (denoted by m^*) can be obtained as follows:

$$m^* = \min_{k' \in \Omega} m(k', L) \ . \qquad (6.15)$$

6.3.3 Dynamic-Programming-Based Target Channel Selection Algorithm

Based on the optimal substructure, we propose a dynamic programming algorithm with time complexity of $O(LM^2)$ to search the minimum cost path in the state diagram to minimize the cumulative handoff delay of the considered secondary user's connection. The detail is shown in Algorithm 1.

Algorithm 1: Dynamic Programming Algorithm

```
Input: $M$,$L$, $\eta$, and $w$
Output: $m(k', L)$
for $k'$ = 1 : $M$ do
    $m(k', 1) = w(\eta; k', 1)$ ;
end
for $i$ = 2 : $L$ do
    for $k'$ = 1 : $M$ do
        $m(k', i) = \infty$ ;
        for $k$ = 1 : $M$ do
            $m'(k', i) = m(k, i-1) + w(k; k', i)$ ;
            if $m(k', i) > m'(k', i)$ then
                $m(k', i) = m'(k', i)$ ;
            end
        end
    end
end
```

Our ultimate goal is to find an optimal target channel sequence to minimize the average cumulative handoff delay (or equivalently to find a minimum cost path). To this end, when evaluating the cumulative cost $m(k, i)$, we must keep track of how to construct an optimal solution to find the corresponding minimum cost path.

6.4 A Suboptimal Low-Complexity Greedy Algorithm

In the section we present a suboptimal greedy algorithm to further reduce the time complexity of solving the **Cumulative Handoff Delay Minimization Problem**. Based on the suggested greedy target channel selection strategy, the proposed greedy algorithm has time complexity of $O(M)$.

6.4.1 Greedy Target Channel Selection Strategy

In each spectrum handoff, the greedy target channel selection strategy is suggested to select the channel with *the shortest expected handoff delay*. Some permutations of the target channel sequences will never occur when this shortest-handoff-delay strategy is adopted. Here, we give such an example. Consider a secondary user's connection whose default channel is channel 1 ($Ch1$). In a two-channel system, it can select either channel 1 or channel 2 ($Ch2$) for its target channel when an interruption event occurs. Now, we assume that the average busy period of $Ch1$ (denoted by $\mathbf{E}[Y_p^{(1)}]$) is shorter than the sum of channel switching time (denoted by t_s) and the average waiting time of $Ch2$ (denoted by $\mathbf{E}[W_s^{(2)}]$). Hence, when the first interruption occurs and the greedy target channel selection strategy is adopted, the secondary user selects $Ch1$ as its target channel for spectrum handoff. The similar argument can be held again for all the upcoming interruptions. That is, the target channel sequence will be $(Ch1, Ch1, Ch1, Ch1, Ch1, \cdots)$. In this case, some permutations of the target channel sequences, such as

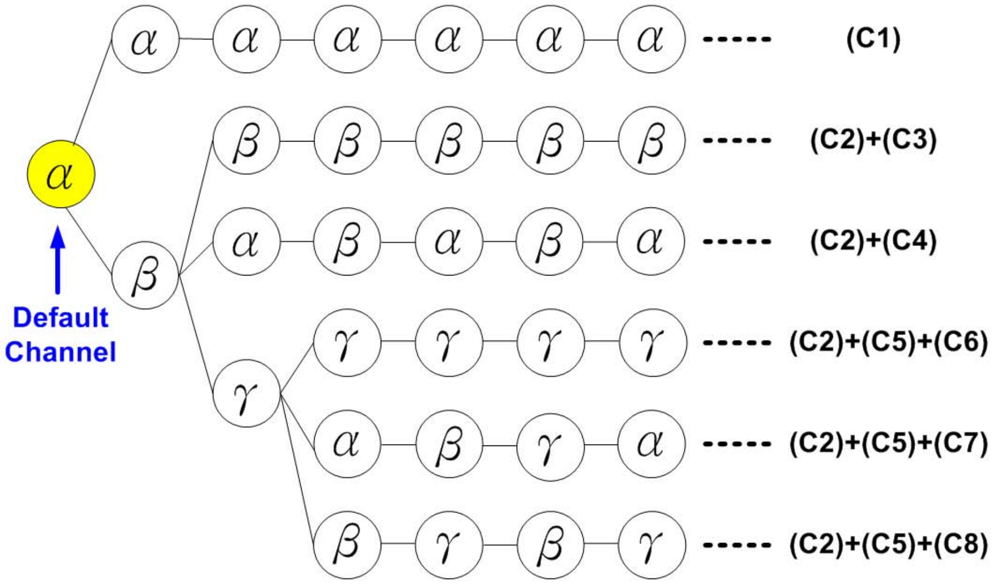


Figure 6.2: Six kinds of candidate sequences for the **Cumulative Handoff Delay Minimization Problem** when the greedy shortest-handoff-delay target channel selection strategy is adopted.

$(Ch1, Ch2, Ch2, Ch2, Ch2, \cdots)$ or $(Ch1, Ch2, Ch1, Ch1, Ch1, \cdots)$, will never occur. In Theorem 4, we prove that only six permutations of the target channel sequences are required to be compared when the greedy shortest-handoff-delay target channel selection strategy is employed.

Theorem 4. *The shortest-handoff-delay target channel selection strategy only requires to compare six permutations of the target channel sequences, as shown in Fig. 6.2.*

Proof. Consider a secondary user's connection whose default channel is α ($\alpha \in \Omega$). If the strategy is to select a channel with the shortest handoff delay when an interruption event occurs, the secondary user will compare the expected handoff delay of staying on the same channel and that of changing to a new channel. Now, we discuss what conditions will cause target channel sequences not to be considered by the greedy target channel selection strategy.

(1) At the first interruption: The secondary user can select channel α or channel k ($k \in \Omega/\{\alpha\}$) for the target channel. If staying on channel α, the expected delay for the non-hopping spectrum handoff equals the average busy period of the primary users' connections at channel α (denoted by $\mathbf{E}[Y_p^{(\alpha)}]$). If changing its operating channel to channel k, the secondary user will experience the delay of the hopping spectrum handoff, which is equal to the sum of the channel switching time (denoted by t_s) and the average waiting time on channel k (denoted by $\mathbf{E}[W_s^{(k)}]$). On the one hand, if the following condition **(C1)** is satisfied,

$$\mathbf{(C1)} : \mathbf{E}[Y_p^{(\alpha)}] \leq \min_{\forall k \in \Omega/\{\alpha\}} \{\mathbf{E}[W_s^{(k)}] + t_s\} \ ,$$

channel α is the first element in the target channel sequence. This implies that the interrupted secondary user must wait until all the primary users' connections leave channel α. When the traffic statistics of all channels are stationary, the interrupted secondary user will always stay on channel α because **(C1)** always can be satisfied for all the upcoming interruptions. That is, the target channel sequence becomes $(\alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \cdots)$, as shown in Fig. 6.2. On the other hand, if the condition

$$\mathbf{(C2)} : \begin{cases} \beta = \arg\min_{\forall k \in \Omega/\{\alpha\}} \mathbf{E}[W_s^{(k)}] \\ \mathbf{E}[W_s^{(\beta)}] + t_s < \mathbf{E}[Y_p^{(\alpha)}] \end{cases}$$

is satisfied, the first element in the target channel sequence is channel β. Clearly, **(C2)** is not sufficient to determine the remaining elements in the target channel sequence.

(2) At the second interruption: When **(C2)** is satisfied, the secondary user will encounter one of the following three conditions at the second interruption. Denote $\mathbf{E}[Y_p^{(\beta)}]$, $\mathbf{E}[W_s^{(\gamma)}]+t_s$, and $\mathbf{E}[W_s^{(\alpha)}]+t_s$ as the average handoff delay for staying on channel β, that for changing to channel γ ($\gamma \neq \alpha$ and β), and that for switching back to channel α, respectively.

- First, we consider the case

  $$\textbf{(C3)} : \mathbf{E}[Y_p^{(\beta)}] \leq \min_{\forall k \in \Omega/\{\beta\}} \{\mathbf{E}[W_s^{(k)}] + t_s\} \ .$$

  Similar to **(C1)**, the interrupted secondary user prefers staying on channel β when **(C3)** is satisfied. Thus, **(C2)** and **(C3)** lead to the target channel sequence $(\beta, \beta, \beta, \cdots)$.

- Next, we consider the condition

  $$\textbf{(C4)} : \mathbf{E}[W_s^{(\alpha)}] + t_s < \min\{\min_{\forall k \in \Omega/\{\alpha,\beta\}} \{\mathbf{E}[W_s^{(k)}] + t_s\}, \mathbf{E}[Y_p^{(\beta)}]\} \ .$$

  When **(C4)** is satisfied, the interrupted secondary user will switch back to channel α. After that, the third interruption event may occur. If so, this interrupted secondary user will switch back to channel β due to **(C2)**. Hence, **(C2)** and **(C4)** yields the target channel sequence $(\beta, \alpha, \beta, \alpha, \beta, \alpha, \cdots)$.

- When **(C3)** and **(C4)** are not satisfied, it implies that there exists channel γ ($\gamma \neq \alpha$) such that

  $$\textbf{(C5)} : \begin{cases} \gamma = \arg\min_{\forall k \in \Omega/\{\alpha,\beta\}} \mathbf{E}[W_s^{(k)}] \\ \mathbf{E}[W_s^{(\gamma)}] + t_s < \mathbf{E}[Y_p^{(\beta)}] \\ \mathbf{E}[W_s^{(\gamma)}] < \mathbf{E}[W_s^{(\alpha)}] \end{cases} \ .$$

  Then, the second element in the target channel sequence is channel γ. Since **(C2)** and **(C5)** are not sufficient to determine the remaining elements in the target channel sequence, the third interruption event will be considered.

(3) At the third interruption: In this case, we need to compare the expected handoff delay of staying on channel γ and that of switching back to channels α and β, i.e., $\mathbf{E}[Y_p^{(\gamma)}]$, $\mathbf{E}[W_s^{(\alpha)}] + t_s$, and $\mathbf{E}[W_s^{(\beta)}] + t_s$, respectively. Given **(C2)** and **(C5)**, three different situations exist.

- First of all, if the condition

$$\textbf{(C6)} : \mathbf{E}[Y_p^{(\gamma)}] \leq \min_{\forall k \in \Omega / \{\gamma\}} \{\mathbf{E}[W_s^{(k)}] + t_s\}$$

  is satisfied, the interrupted secondary user prefers staying on channel γ. Hence, **(C2)**, **(C5)**, and **(C6)** result in the target channel sequence $(\beta, \gamma, \gamma, \gamma, \cdots)$.

- Furthermore, if the condition

$$\textbf{(C7)} : \mathbf{E}[W_s^{(\alpha)}] + t_s < \min\{\min_{\forall k \in \Omega / \{\alpha, \gamma\}} \{\mathbf{E}[W_s^{(k)}] + t_s\}, \mathbf{E}[Y_p^{(\gamma)}]\}$$

  is satisfied, the interrupted secondary user switches back to channel α. Then, **(C2)** and **(C5)** will make this secondary user switches to channel β and channel γ at the fourth and fifth interruptions, respectively. Thus, when **(C2)**, **(C5)** and **(C7)** are satisfied, the target channel sequence becomes $(\beta, \gamma, \alpha, \beta, \gamma, \alpha, \beta, \gamma, \alpha, \cdots)$.

- Similarly, when

$$\textbf{(C8)} : \mathbf{E}[W_s^{(\beta)}] + t_s < \min\{\min_{\forall k \in \Omega / \{\beta, \gamma\}} \{\mathbf{E}[W_s^{(k)}] + t_s\}, \mathbf{E}[Y_p^{(\gamma)}]\}$$

  is satisfied, one can show that the target channel sequence is $(\beta, \gamma, \beta, \gamma, \beta, \gamma, \cdots)$.

Now, we will prove that it is not necessary to consider other channels except for channels α, β, and γ when the third interruption event occurs. Assume that there exists channel $\xi = \underset{\forall k \in \Omega/\{\alpha,\beta,\gamma\}}{\arg\min} \mathbf{E}[W_s^{(k)}]$ such that

$$\begin{cases} \mathbf{E}[W_s^{(\xi)}] + t_s < \mathbf{E}[Y_p^{(\gamma)}] \\ \mathbf{E}[W_s^{(\xi)}] < \mathbf{E}[W_s^{(\alpha)}] \\ \mathbf{E}[W_s^{(\xi)}] < \mathbf{E}[W_s^{(\beta)}] \end{cases} .$$

Then, it follows that

$$\mathbf{E}[W_s^{(\xi)}] + t_s < \mathbf{E}[W_s^{(k)}] + t_s, \quad \forall\ k \neq \xi, \gamma\ . \tag{6.16}$$

However, from **(C2)**, we know that

$$\mathbf{E}[W_s^{(\beta)}] + t_s < \mathbf{E}[W_s^{(k)}] + t_s, \quad \forall\ k \neq \alpha, \beta\ . \tag{6.17}$$

Since (6.16) yields $\mathbf{E}[W_s^{(\xi)}] < \mathbf{E}[W_s^{(\beta)}]$, but (6.17) yields $\mathbf{E}[W_s^{(\beta)}] < \mathbf{E}[W_s^{(\xi)}]$, it leads to a contradiction and proves that no other channels need to be considered further.

From the above discussions, **(C1)-(C8)** have considered all the conditions between $\mathbf{E}[W_s^{(k)}]$ and $\mathbf{E}[Y_p^{(k)}]$. Hence, we can conclude that the greedy shortest-handoff-delay target channel selection strategy results in only six permutations of the target channel sequences that are needed to be further compared in the **Cumulative Handoff Delay Minimization Problem**, are shown in Fig. 6.2. □

Based on Theorem 4, the transmitter and receiver need to consider only three channels for spectrum handoff as long as the greedy shortest-handoff-delay target channel selection strategy is considered. Thus, it relieves channel consensus issue in CR networks. Theorem 4 can be extended to other greedy strategies for the target channel selection based on various criteria, such as the longest expected remaining idle period.

6.4.2 Greedy Target Channel Selection Algorithm

The shortest-handoff-delay strategy is adopted to select the target channel in the proposed greedy algorithm, as shown in Algorithm 2. Hence, this Algorithm 2 has time complexity of $O(M)$.

```
Algorithm 2: Suboptimal Greedy Algorithm
    Input: $M$,$L$, $\eta$, $\mathbf{E}[W_s^{(k)}]$, and $\mathbf{E}[Y_p^{(k)}]$
    Output: $m(k', L)$
    for $j = 1 : 8$ do
        Checking whether the condition (Cj) can be satisfied by comparing
        the values of $\mathbf{E}[W_s^{(k)}]$ and $\mathbf{E}[Y_p^{(k)}]$ for any $k$, where $1 \leq k \leq M$.
    end
    According to Fig. 6.2, determining the target channel sequence.
```

6.5 Numerical Results

In this section, by applying the proposed analytical models to the environments with various statistics of service time distributions for both primary and secondary users, we show the cumulative spectrum handoff delay performance for the proposed target channel sequence design approaches subject to the effects of multiple handoffs. Five target channel selection schemes are compared, which consist of (1) the random selection strategy; (2) the throughput-orientated strategy; (3) the greedy shortest-handoff-delay target channel selection strategy; (4) the dynamic programming (DP)-based solution; and (5) the exhaustive search (ES)-based solution. For the random selection strategy, the secondary user randomly selects one channel for its target channel when an interruption event occurs. Furthermore, the throughput-orientated strategy selects the channel that is accessed by the primary users with the lowest probability. Moreover, the suggested greedy strategy selects the channel with the shortest handoff delay for its target channel.

6.5.1 Effects of Traffic Statistics for Arriving Secondary User's Service Time

Firstly, we investigate the effects of the newly arriving secondary user's average service time ($\mathbf{E}[\chi_s]$) on its cumulative handoff delay. Figure 6.3 compares the cumulative handoff delay of the five considered target channel selection schemes in a four-channel CR network with $\lambda_p^{(k)} = 0.02$ and $\lambda_s^{(k)} = 0.01$ for $1 \leq k \leq 4$. We have the following observations:

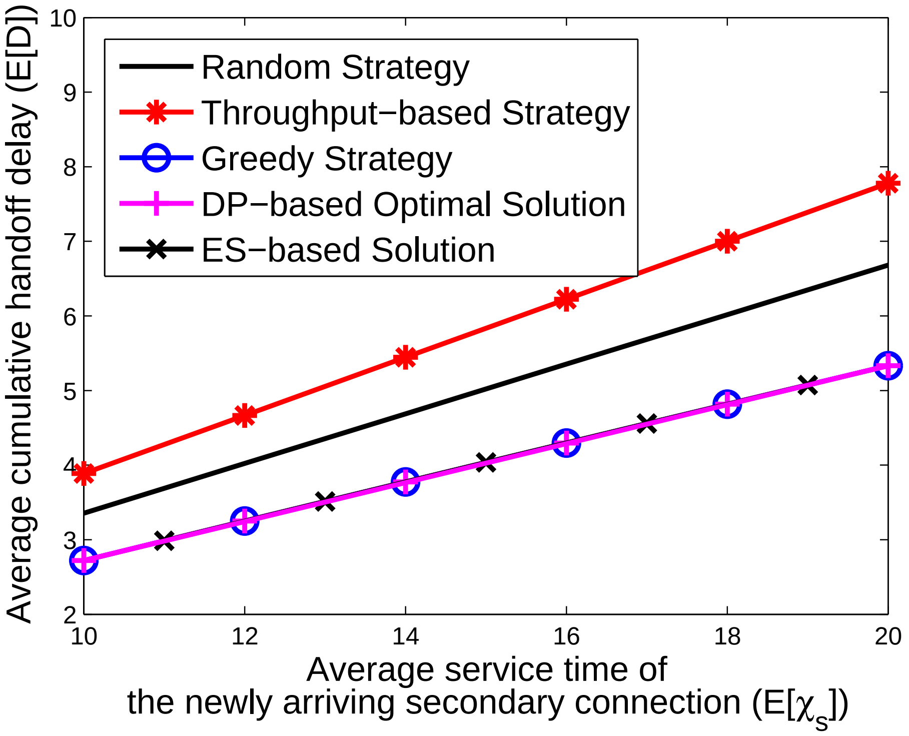
- When the average service time $\mathbf{E}[\chi_s]$ increases, a secondary user experiences more interruptions on average and thus the cumulative handoff delay increases as shown in the figure.
- The dynamic programming approach yields almost the same result as the exhaustive search approach.
- Compared to the random strategy, the greedy strategy and the optimal solution can significantly reduce the cumulative handoff delay because it takes a priori traffic statistics into account when determining the target channel for spectrum handoff. For example, the greedy strategy and the optimal solution can reduce the cumulative handoff delay by over 20% and 58% compared to the random strategy when $\mathbf{E}[\chi_s] = 20$ in Figs. 6.3(a) and 6.3(b), respectively.
- Figure 6.3(a) shows that the cumulative handoff delay of the low-complexity greedy target channel selection strategy is the same as the

optimal solution when the primary user service time has similar distributions in different channels, where $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}], \mathbf{E}[X_p^{(4)}])$ $= (14, 15, 15, 15)$, and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}], \mathbf{E}[X_s^{(4)}]) = (10, 12, 14, 16)$. However, if the primary user's service time distributions of each channel are different, the cumulative handoff delay resulting from the suggested greedy strategy is slightly larger than the optimal solution. For example, when $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}], \mathbf{E}[X_p^{(4)}]) = (10, 15, 20, 25)$ and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}], \mathbf{E}[X_s^{(4)}]) = (10, 10, 10, 10)$, Fig. 6.3(b) shows that the cumulative handoff delay of the greedy strategy is 9% higher than that of the optimal solution at $\mathbf{E}[\chi_s] = 20$.

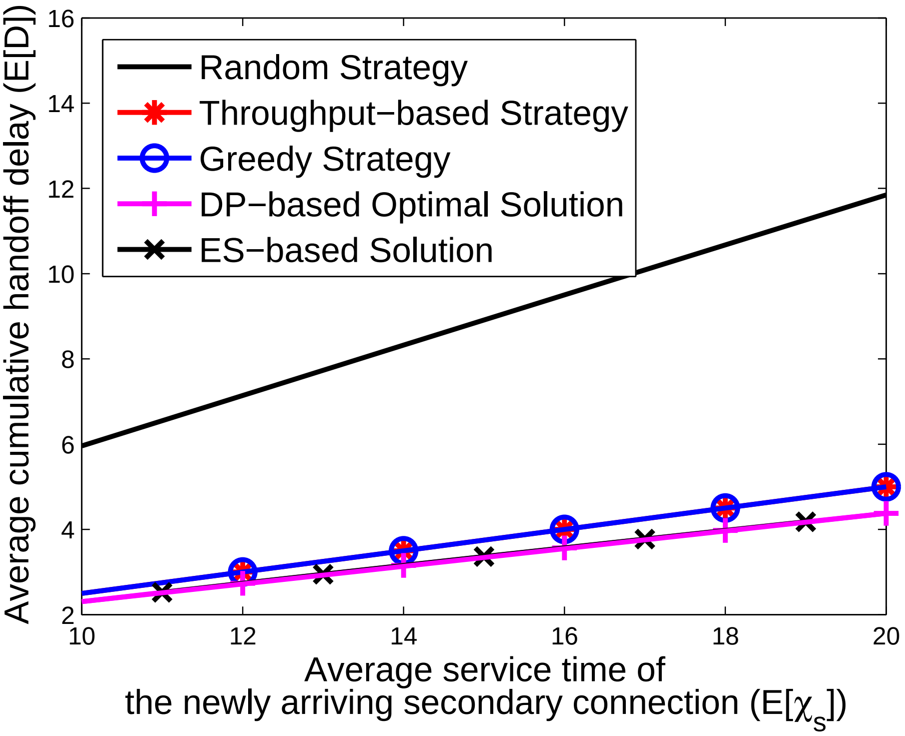
- When the arrival rates and service time of the secondary users' connections are the same at all the four channels as shown in Fig. 6.3(b), the throughput-orientated strategy and the suggested greedy strategy result in the same cumulative handoff delay. However, when the secondary users' connections have different traffic statistics as shown in Fig. 6.3(a), it is shown that at $\mathbf{E}[\chi_s] = 20$ the greedy strategy improves the cumulative handoff delay of the throughput-orientated strategy by 46%. This is because the suggested greedy strategy takes into account of the traffic statistics of both the primary and secondary users when determining the target channel.

6.5.2 Effects of Traffic Statistics of Existing Secondary Users' Connections

Figure 6.4 shows how the existing secondary connections' traffic statistics, including the average service time $\mathbf{E}[X_s]$ and the arrival rate λ_s, affect the cumulative handoff delay of the newly arriving secondary user's connection.



(a) $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}], \mathbf{E}[X_p^{(4)}]) = (14, 15, 15, 15)$, and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}], \mathbf{E}[X_s^{(4)}]) = (10, 12, 14, 16)$.



(b) $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}], \mathbf{E}[X_p^{(4)}]) = (10, 15, 20, 25)$, and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}], \mathbf{E}[X_s^{(4)}]) = (10, 10, 10, 10)$.

Figure 6.3: Effects of the newly arriving secondary user's average service time $\mathbf{E}[\chi_s]$ on the cumulative handoff delay for $\lambda_p^{(k)} = 0.02$ and $\lambda_s^{(k)} = 0.01$ when $1 \leq k \leq 4$.

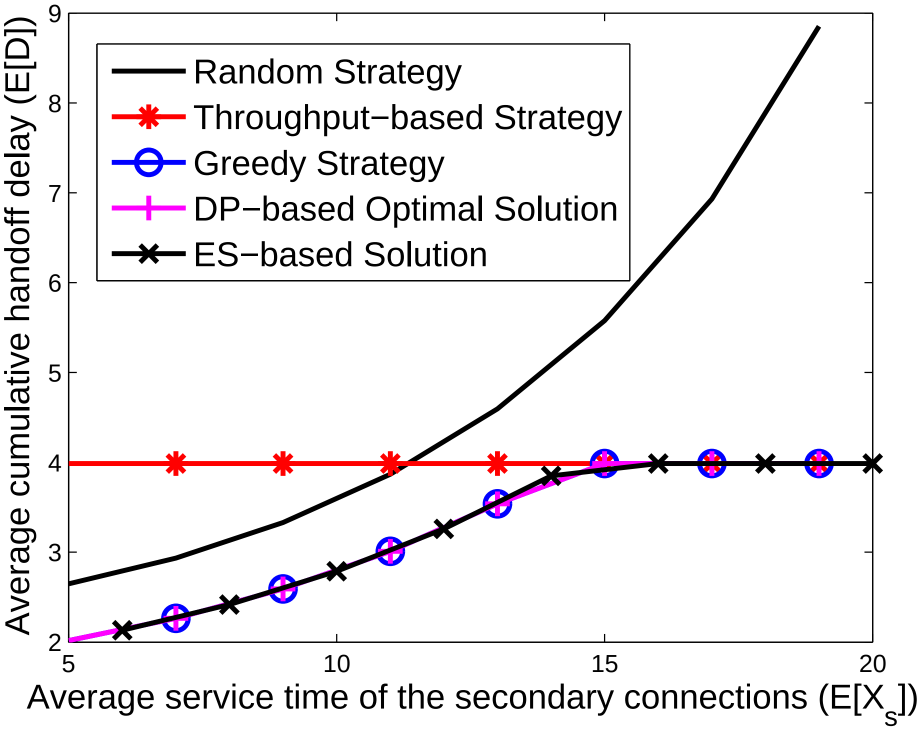
Assume that the service time χ_s of the newly arriving secondary user's connection is geometrically distributed with $\mathbf{E}[\chi_s] = 10$, and the primary users have the similar traffic parameters in different channels. From Fig. 6.4(a) where $\mathbf{E}[X_s^{(k)}] = \mathbf{E}[X_s]$ for $1 \leq k \leq 4$, we observe the following:

- In the range of small $\mathbf{E}[X_s]$ (e.g., $\mathbf{E}[X_s] < 15$), the cumulative handoff delay increases as $\mathbf{E}[X_s]$ increases for the random selection strategy, the greedy strategy, and the optimal solution.

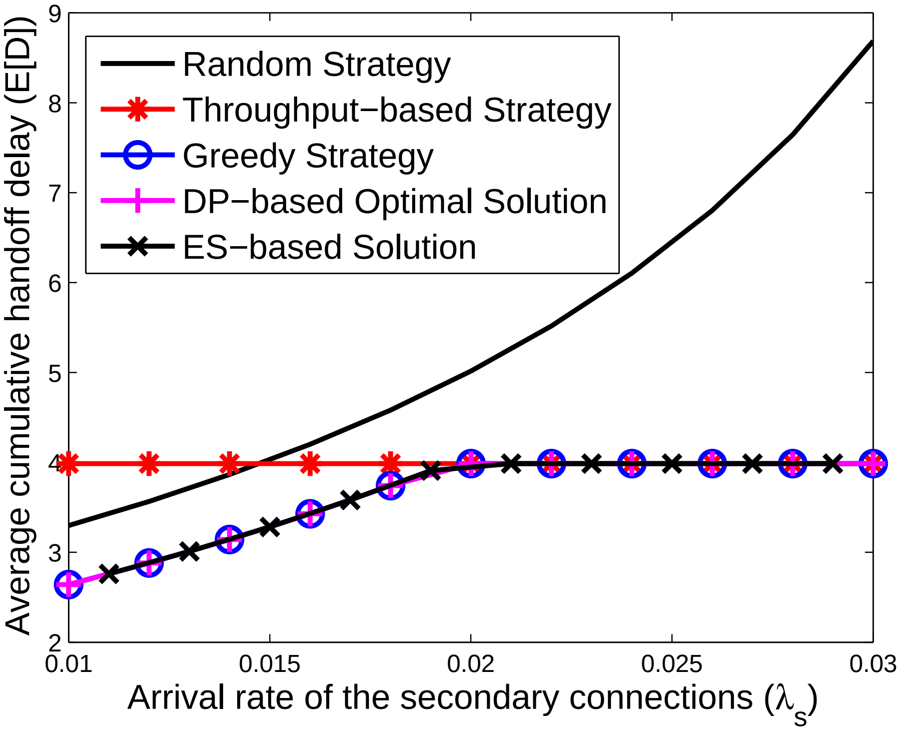
- In the range of large $\mathbf{E}[X_s]$ (e.g., $\mathbf{E}[X_s] \geq 15$), the secondary user will experience long waiting time when it changes its operating channel according to (6.6). Hence, the greedy strategy and the optimal solution prefer staying on the current operating channel when interruptions occur to reduce handoff delay. In this case, their average handoff delay for each handoff is a constant $\mathbf{E}[Y_p]$. Thus, the average cumulative handoff delay is also a constant. However, the random strategy still selects to change channel for spectrum handoff sometimes. Hence, the cumulative handoff delay of the random strategy still increases as $\mathbf{E}[X_s]$ increases.

- Because the throughput-orientated strategy always selects channel 1 for the target channel, the corresponding average handoff delay is a constant $\mathbf{E}[Y_p]$. Hence, the throughput-orientated strategy results in the same average cumulative handoff delay for various $\mathbf{E}[X_s]$.

Note that the similar observations can be also found in Fig. 6.4(b), where $\lambda_s^{(k)} = \lambda_s$ for $1 \leq k \leq 4$. When $\lambda_s \geq 0.02$, the interrupted secondary users will always stay on the current operating channel for the greedy strategy and the optimal solution.



(a) Effect of the average service time $\mathbf{E}[X_s]$ on the cumulative handoff delay $\mathbf{E}[D]$, where $(\lambda_s^{(1)}, \lambda_s^{(2)}, \lambda_s^{(3)}, \lambda_s^{(4)}) = (0.01, 0.015, 0.02, 0.025)$.



(b) Effect of the arrival rate λ_s on the cumulative handoff delay $\mathbf{E}[D]$, where $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}], \mathbf{E}[X_s^{(3)}], \mathbf{E}[X_s^{(4)}]) = (10, 12, 14, 16)$.

Figure 6.4: Effect of the average service time $\mathbf{E}[X_s]$ and the arrival rate λ_s of the secondary users' connections on the cumulative handoff delay of the newly arriving secondary user's connection for $(\lambda_p^{(1)}, \lambda_p^{(2)}, \lambda_p^{(3)}, \lambda_p^{(4)}) = (0.019, 0.02, 0.02, 0.02)$ and $\mathbf{E}[X_p^{(k)}] = 15$ when $1 \leq k \leq 4$.

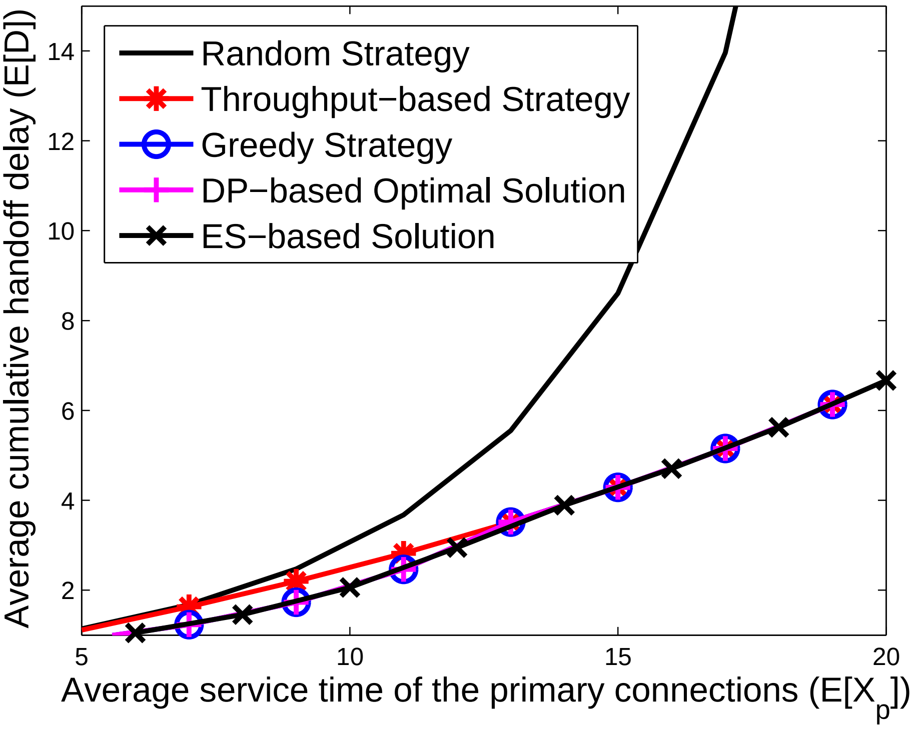
6.5.3 Effects of Traffic Statistics of Existing Primary Users' Connections

Figure 6.5 shows the effects of the average service time $\mathbf{E}[X_p]$ and the arrival rate λ_p of the primary users' connections on the cumulative handoff delay of the newly arriving secondary user's connection. We consider that $\lambda_s^{(k)} = \lambda_s$ and $\mathbf{E}[X_s^{(k)}] = \mathbf{E}[X_s]$ for $1 \leq k \leq 4$ as well as the service time χ_s is geometrically distributed and $\mathbf{E}[\chi_s] = 10$. In Fig. 6.5(a), we assume that $\mathbf{E}[X_p^{(k)}] = \mathbf{E}[X_p]$ for $1 \leq k \leq 4$. We can find the following:

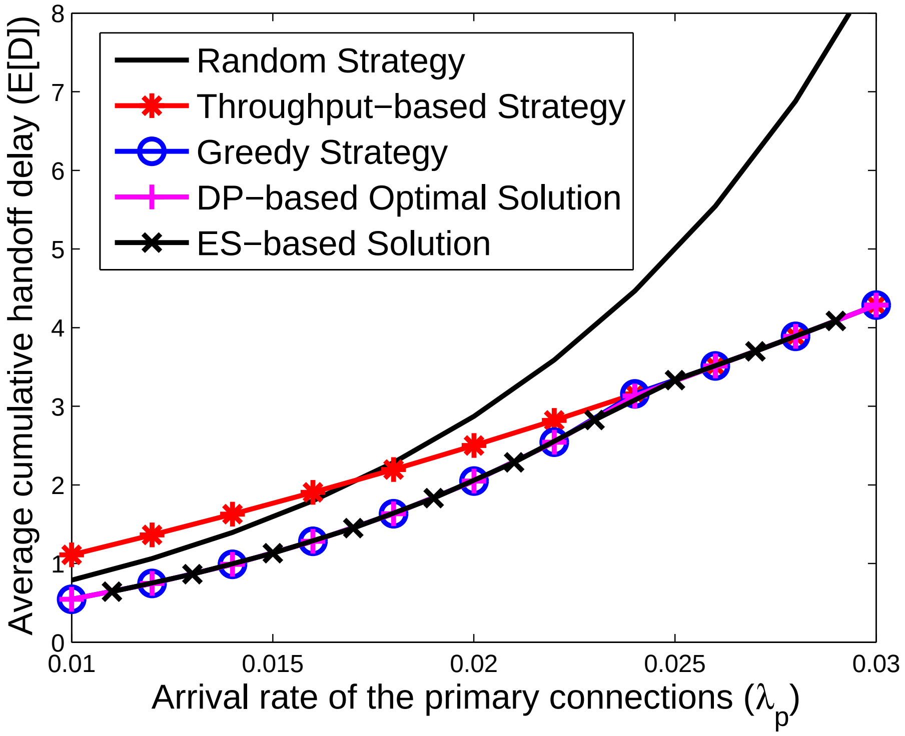
- For all methods, the cumulative handoff delay increases as $\mathbf{E}[X_p]$ increases because a larger value of $\mathbf{E}[X_p]$ results in heavier traffic load.

- For the throughput-orientated strategy, the greedy strategy, and the optimal solution, their cumulative handoff delay at various $\mathbf{E}[X_p]$ will ultimately converge to the same value as shown in the region of $\mathbf{E}[X_p] \geq 13$ in Fig. 6.5(a). In the region, the handoff delay is only related to the busy period $\mathbf{E}[Y_p]$ and uncorrelated to the value of $\mathbf{E}[X_s]$ because the interrupted secondary users always stay on the current operating channel when $\mathbf{E}[X_p] \geq 13$.

Note that we can have the similar conclusions in Fig. 6.5(b), where $\lambda_p^{(k)} = \lambda_p$ for $1 \leq k \leq 4$.



(a) Effect of the average service time $\mathbf{E}[X_p]$ on the cumulative handoff delay $\mathbf{E}[D]$, where $(\lambda_p^{(1)}, \lambda_p^{(2)}, \lambda_p^{(3)}, \lambda_p^{(4)}) = (0.02, 0.025, 0.03, 0.035)$.



(b) Effect of the arrival rate λ_p on the cumulative handoff delay $\mathbf{E}[D]$, where $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}], \mathbf{E}[X_p^{(3)}], \mathbf{E}[X_p^{(4)}]) = (10, 12, 14, 16)$.

Figure 6.5: Effect of the average service time $\mathbf{E}[X_p]$ and the arrival rate λ_p of the primary users' connections on the cumulative handoff delay of the newly arriving secondary user's connection for $\lambda_s^{(k)} = 0.01$ and $\mathbf{E}[X_s^{(k)}] = 15$ when $1 \leq k \leq 4$.

Chapter 7

Reactive Spectrum Handoff

As discussed in Chapter 5, spectrum handoff mechanisms can be categorized as either the proactive spectrum handoff or the reactive spectrum handoff schemes. In this chapter, we focus on the modeling technique and performance analysis for the reactive spectrum handoff scheme. Compared to the proactive spectrum handoff scheme that the preselected target channel may no longer be available at the instant that spectrum handoff procedures are initiated, the reactive spectrum handoff may have shorter handoff delay because it can reliably find an idle channel through spectrum sensing. Nevertheless, the reactive spectrum handoff scheme needs the sensing time to search the idle channels. In addition, it also needs the handshaking time to achieve a consensus on the target channel between the transmitter and receiver of a secondary connection. Hence, one important issue for the reactive spectrum handoff scheme is to characterize the effects of the sensing time and the handshaking time on the handoff delay. Obviously, when the sensing time and the handshaking time is too large, the reactive spectrum handoff is worse than the proactive spectrum handoff in terms of the extended data delivery time.

The goal of this chapter is to investigate the effects of spectrum handoffs on the channel utilization and the extended data delivery time of the secondary users' connections with various traffic arrival rates and service time distributions. We consider the three key design features for spectrum handoff, consisting of (1) heterogeneous arrival rates of the primary users at different channels, where various channels have different traffic arrival rates of the primary users because these channels may belong to different primary system operators; (2) various arrival rates of the secondary users at different channels, where the arrival rates can be determined by the initial operating channel selection mechanisms [135]; and (3) handoff processing time, resulting from the sensing time, the handshaking time, and the channel switching time. How to model the channel utilization at each channel and the extended data delivery time in the context of multiple handoffs is challenging since the operating channels for multiple handoffs are selected according to the channel occupancy states at the moments of link transitions. To the best of our knowledge, an analytical model for characterizing all the three features for multiple handoffs has rarely been seen in the literature. The contributions of this chapter are summarized in the following:

- First, The preemptive resume priority (PRP) M/G/1 queueing network model is proposed to characterize the channel usage behaviors of CR networks. Based on this queueing model, we can evaluate the channel utilizations of different channels under various traffic arrival rates and service time distributions.

- Next, a state diagram is developed to characterize the effect of multiple handoff delay on the extended data delivery time of the secondary connections. Then, we can evaluate how long the extended data delivery time is prolonged due to multiple spectrum hanodffs.

7.1 System Model

7.1.1 Assumptions

In this chapter, we consider the spectrum handoff protocol presented in [151]. When the spectrum handoff procedures are initiated, the secondary users must spend τ slots on spectrum sensing to find the idle channels. Note that if more than one channel is assessed as idle, the interrupted secondary user will randomly select one idle channel from all idle channels to be its target channel for spectrum handoff. Here, we assume that this random selection follows the uniform distribution. Furthermore, the interrupted secondary user will stay on the current operating channel if all channels are busy. Next, the handshaking time of t_h slots is spent in order to achieve a consensus on the target channel between the transmitter and the receiver of a secondary connection. Hence, when a secondary user changes its operating channel to another channel, the total processing time for executing spectrum handoff procedures is $\delta_c \triangleq \tau + t_h + t_s$ where t_s (slots) is the channel switching time. On the other hand, if the secondary user stays on the current operating channel, the total processing time is $\delta_s \triangleq \tau + t_h$.

7.1.2 Illustrative Example of Reactive Multiple Handoffs with Multiple Interruptions

A secondary user's connection may experience multiple interruption requests from the primary users during its transmission period. Because these interruptions result in multiple handoffs, a series of target channels is selected by spectrum sensing, called the *target channel sequence* in this report. Fig. 7.1 shows an example that three spectrum handoffs occur during the transmission period of the secondary connection SC$_A$, where SC$_A$'s initial (default) channel is Ch3. We assume that the transmitter of SC$_A$ wants to establish a connection flow with 30 slots to the corresponding receiver. Its extended data delivery time is denoted by T. Furthermore, D_i is the handoff delay of the i^{th} interruption. Here, the handoff delay is defined as the duration from the instant that transmission is interrupted until the instant that the unfinished transmission is resumed. Then, the transmission process with multiple handoffs is described as follows:

1. In the beginning, SC$_A$ is established at its default channel Ch3. When an interruption event occurs, SC$_A$ performs spectrum sensing to search the idle channel for spectrum handoff.

2. At the first interruption, SC$_A$ changes its operating channel to the idle channel Ch2 from Ch3. Thus, the handoff delay D_1 is δ_c.

3. At the second interruption, SC$_A$ stays on its current operating channel Ch2 because all channels are busy. SC$_A$ cannot be resumed until all the high-priority primary connections at Ch2 finish their transmissions. In this case, the handoff delay D_2 is the sum of δ_s and the duration from the time instant that Ch2 is used by the primary users' connections until the time instant that the high-priority queue becomes empty. This duration (denoted by $Y_p^{(2)}$) is called the *busy period* resulting from the transmissions of multiple primary users' connections at Ch2.

4. At the third interruption, SC$_A$ finds both Ch1 and Ch3 are idle. Then, SC$_A$ randomly selects one channel to be the target channel. In this example, SC$_A$ selects Ch1 to be its target channel. Note that the handoff delay D_3 in this case is δ_c.

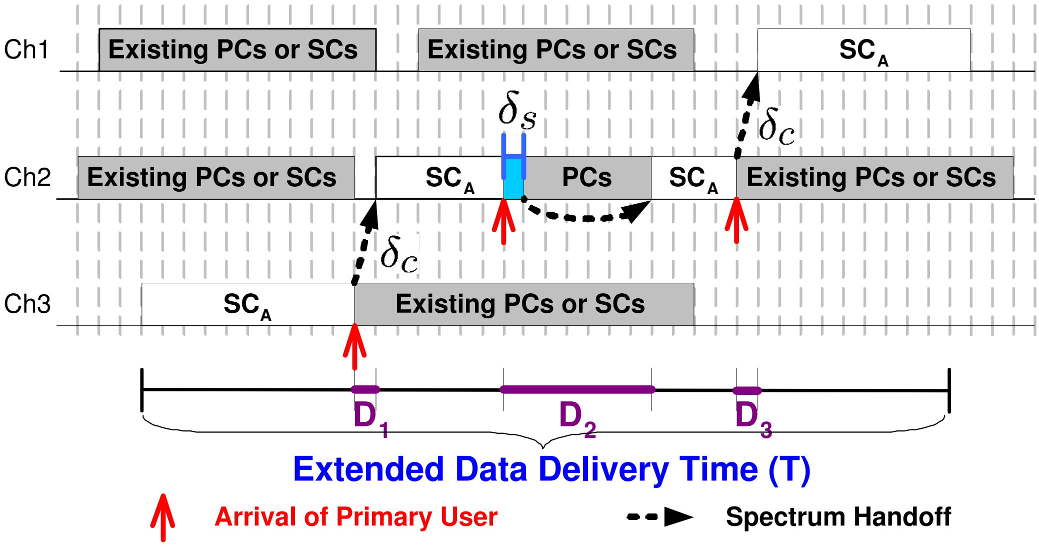


Figure 7.1: An example of transmission process for the secondary connection SC$_A$, where T is the extended data delivery time of SC$_A$ and D_i is the handoff delay of the i^{th} interruption. The gray areas indicate that the channels are occupied by the existing primary connections (PCs) or secondary connections (SCs). Because SC$_A$ is interrupted three times in total, the overall data connection is divided into four segments. Note that $D_1 = \delta_c$, $D_2 = \delta_s$, and $D_3 = \delta_c$.

5. Finally, SC$_A$ is completed on Ch3.

Hence, SC$_A$'s *target channel sequence* is (Ch2, Ch2, Ch1) in this example.

7.2 Analytical Model

We use the PRP M/G/1 queueing network model proposed in Chapter 3 to characterize the channel usage behaviors of a CR network. Let $X_s^{(\eta)}$ (slots/arrival) be the service time of the secondary connections whose default channels are channel η and let $f_s^{(\eta)}(x)$ be the probability density function (pdf) of $X_s^{(\eta)}$. Figure 7.2 shows an example of the PRP M/G/1 queueing network model with three channels. The traffic flows of the primary connections and the secondary connections are directly connected to the high-priority queue and the low-priority queue, respectively. A secondary connection will be interrupted when the primary connection appears. The interrupted secondary connection can decide its target channel for spectrum handoff according to the instantaneous spectrum sensing outcomes. Note that the required time for spectrum sensing is modeled in $\boxed{\text{S}}$. It can be regarded as a tapped delay line or a server with constant service time, which equals to the handoff processing time. In the proposed queueing network, the interrupted secondary connection can either stay on its current channel or change to another channel through different feedback paths. If the interrupted secondary connection chooses to stay on its current operating channel, its remaining data will be connected to the head of the low-priority queue of its current operating channel. On the other hand, if the decision is to change its operating channel, the remaining data of the interrupted secondary connection can be connected to the empty low-priority queue of the selected channel. In the figure, $\oplus$ represents that the traffic of the interrupted secondary connection is merged. Furthermore, when the interrupted secondary connection transmits the remaining data on the target channel, it may be interrupted again. Hence, this model can incorporate the effects of multiple interruptions.

7.2.1 Notations

Now, we define some notations as follows. We call the secondary connection that has experienced i interruptions, where $i \geq 0$, the type-i secondary connection. Firstly, we consider the type-i secondary connections whose default

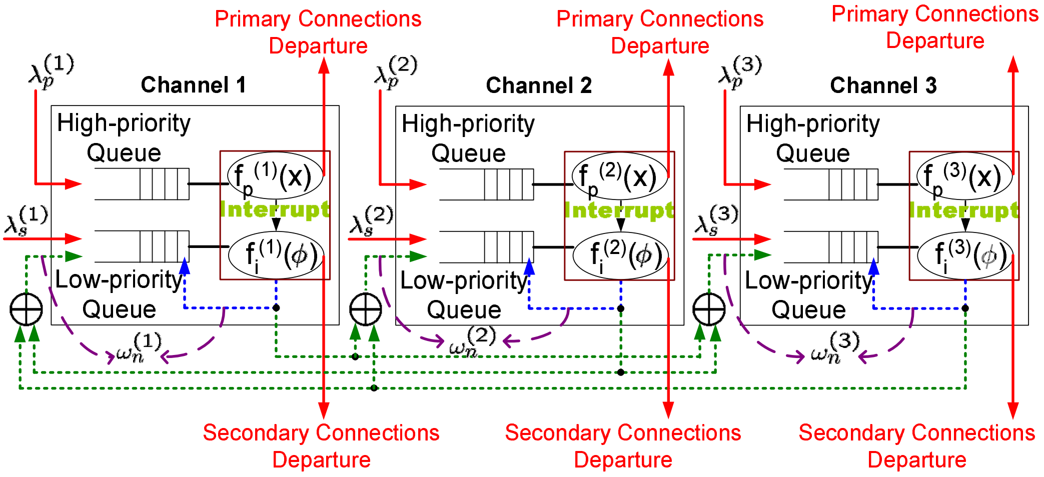


Figure 7.2: The PRP M/G/1 queueing network model with three channels where $\lambda_p^{(k)}$, $\lambda_s^{(k)}$, and $\omega_n^{(k)}$ are the arrival rates of the primary connections, the secondary connections, and the type-n secondary connections ($n \geq 1$) at channel k. Note that $\lambda_s^{(k)} = \omega_0^{(k)}$. Furthermore, $f_p^{(k)}(x)$ and $f_i^{(k)}(\phi)$ are the probability density functions (pdfs) of $X_p^{(k)}$ and $\Phi_i^{(k)}$, respectively.

channels are the channel η. Two more important system parameters $\omega_{i,\eta}^{(k)}$ and $\Phi_{i,\eta}^{(k)}$ are defined as follows:

- $\omega_{i,\eta}^{(k)}$ is the arrival rate of the considered secondary connections at channel k. Note that $\omega_{0,\eta}^{(\eta)} = \lambda_s^{(\eta)}$. How to derive $\omega_{i,\eta}^{(k)}$ from the four traffic parameters is discussed in Section 7.3.

- $\Phi_{i,\eta}^{(k)}$ is the effective service time of the considered secondary connections at channel k. That is, $\Phi_{i,\eta}^{(k)}$ is the considered secondary connection's transmission duration between the i^{th} and the $(i+1)^{th}$ interruptions at channel k. In Section 7.3, we will discuss how to derive $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$ from the four traffic parameters.

Next, let $\omega_i^{(k)}$ and $\Phi_i^{(k)}$ be the arrival rate and the effective service time of the type-i secondary connections at channel k, respectively. We can have

$$\omega_i^{(k)} = \sum_{\eta=1}^{M} \omega_{i,\eta}^{(k)} \ , \tag{7.1}$$

and

$$\mathbf{E}[\Phi_i^{(k)}] = \sum_{\eta=1}^{M} \frac{\omega_{i,\eta}^{(k)}}{\omega_i^{(k)}} \mathbf{E}[\Phi_{i,\eta}^{(k)}] \ , \tag{7.2}$$

respectively.

Finally, we denote $\rho_p^{(k)}$ and $\rho_i^{(k)}$ as the channel busy probabilities resulting from the transmissions of the primary connections and the type-i secondary connections whose current operating channels are the channel k, respectively. Then, in an M-channel network, we can have

$$\rho_p^{(k)} = \lambda_p^{(k)} \mathbf{E}[X_p^{(k)}] \ , \tag{7.3}$$

and

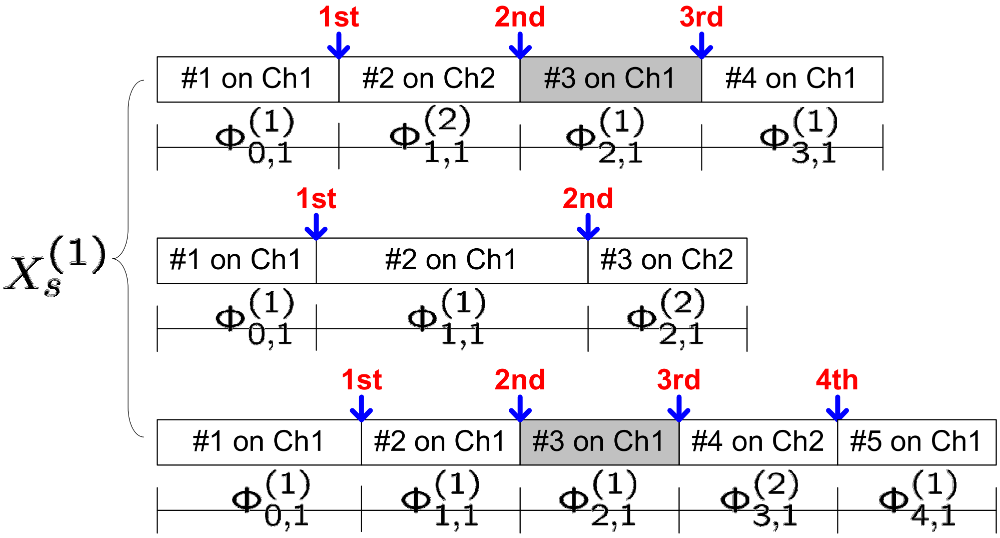
$$\rho_i^{(k)} = \omega_i^{(k)} \mathbf{E}[\Phi_i^{(k)}] = \sum_{\eta=1}^{M} \omega_{i,\eta}^{(k)} \mathbf{E}[\Phi_{i,\eta}^{(k)}] \ , \tag{7.4}$$

respectively. Furthermore, the busy probability of channel k (denoted by $\rho^{(k)}$) shall satisfy the following constraint:

$$\rho^{(k)} \equiv \rho_p^{(k)} + \sum_{i=0}^{\infty} \rho_i^{(k)} < 1 \ , \tag{7.5}$$

where $1 \leq k \leq M$. Note that $\rho^{(k)}$ can be also interpreted as the utilization factor of channel k.

Figure 7.3 illustrates the physical meaning of random variable $\Phi_{i,\eta}^{(k)}$. Consider a two-channel network with the service time of the secondary connections $X_s^{(1)}$ and $X_s^{(2)}$ at the channels 1 and 2, respectively. In the channel 1, random variable $X_s^{(1)}$ are generated three times in Fig. 7.3(a). Similarly, Fig. 7.3(b) shows the three realizations of $X_s^{(2)}$ for the channel 2. Each secondary connection is divided into many segments due to multiple primary users' interruptions. For example, the first secondary connection in Fig. 7.3(a) is divided into four segments because it encounters three interruptions in total. The first, second, third, and fourth segments are transmitted at channels 1, 2, 1, and 1, respectively. Thus, this secondary connection's default channel is Ch1 and its target channel sequence is (Ch2, Ch1, Ch1). In Fig. 7.3(a), random variables $\Phi_{2,1}^{(1)}$, one of the gray regions, represents the transmission duration between the 2^{nd} and the 3^{rd} interruptions at the channel 1 for the secondary connection whose default channel is the channel 1. Similarly, random variables $\Phi_{2,2}^{(1)}$, one of the gray regions in Fig. 7.3(b), represents the transmission duration between the 2^{nd} and the 3^{rd} interruptions at the channel 1 for the secondary connection whose default channel is the channel 2. Furthermore, random variable $\Phi_2^{(1)}$, one of the gray regions in Fig. 7.3, represents the transmission duration of a secondary connection between the 2^{nd} and the 3^{rd} interruptions at channel 1. That is, $\Phi_2^{(1)}$ is one of the third segments of the first and the third secondary connections in Fig. 7.3(a) as well



(a) Three realizations of $X_s^{(1)}$.

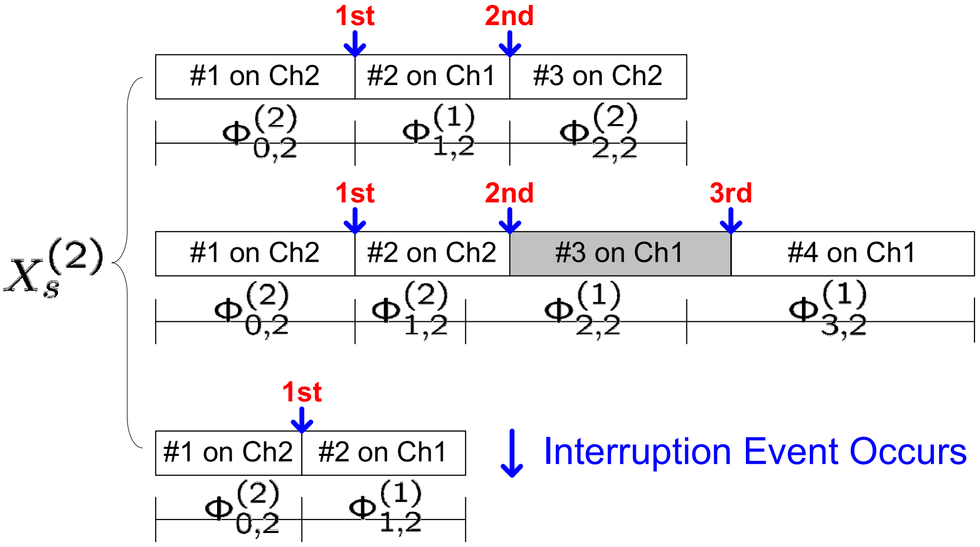
as the second secondary connection in Fig. 7.3(b).

7.3 Analysis of Channel Utilization Factor

Based on the proposed PRP M/G/1 queueing network model, we can evaluate many performance measures of CR networks with various traffic parameters. In this section, we show how to evaluate the channel utilization factor $\rho^{(k)}$. Referring to (7.5), for each channel k $(1 \leq k \leq M)$, it follows that

$$\rho^{(k)} = \lambda_p^{(k)}\mathbf{E}[X_p^{(k)}] + \sum_{i=0}^{\infty}\left[\sum_{\eta=1}^{M}\omega_{i,\eta}^{(k)}\mathbf{E}[\Phi_{i,\eta}^{(k)}]\right] . \tag{7.6}$$

Note that $\rho^{(k)}$ is unrelated to channel sensing time τ, channel notification time t_n, and channel switching time t_s. In (7.6), $\lambda_p^{(k)}$ and $\mathbf{E}[X_p^{(k)}]$ are system parameters and can be known in advance. In the following, we will show how to derive $\omega_{i,\eta}^{(k)}$ and $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$.



(b) Three realizations of $X_s^{(2)}$.

Figure 7.3: Illustration of the physical meaning of random variable $\Phi_i^{(k)}$. For example, $\Phi_2^{(1)}$ is one of the third segments (gray areas) of the first and the third secondary connections in (a) as well as the second secondary connection in (b). Note that the third secondary connection in (b) does not have the third segment because it is interrupted only once.

7.3.1 Derivations of $\omega_{i,\eta}^{(k)}$ and $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$

Without loss of generality, we consider a secondary connection whose default channel is channel $\eta \triangleq s_0$ in the following discussions. Its target channel sequence is denoted by $\mathbf{S}^{(\eta)} \triangleq (S_{1,\eta}, S_{2,\eta}, S_{3,\eta}, \cdots)$, where $S_{i,\eta}$ is the target channel at the i^{th} interruption. Note that $S_{i,\eta}$ is a random variable for each $i \geq 1$. It is decided according to the instantaneous sensing results after the i^{th} interruption event occurs. Thus, $\mathbf{S}^{(\eta)}$ is a random sequence. Based on the definitions of these notations, $\omega_{i,\eta}^{(k)}$ and $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$ can be derived from the Propositions 8 and 9, respectively. Then, the channel utilization $\rho^{(k)}$ can be obtained.

Lemma 5. *Let $p_{i,\eta}^{(k)}$ be the probability that the considered type-i secondary connection is interrupted again at channel k. It follows that*

$$p_{i,\eta}^{(k)} = \lambda_p^{(k)} \boldsymbol{E}[\Phi_{i,\eta}^{(k)}] \ . \tag{7.7}$$

Proof of this lemma can be found in [159].

Definition 6. *Let $\boldsymbol{s}_n \triangleq (s_1, s_2, s_3, \cdots, s_n)$ be any target channel sequence which has n elements. That is, $\boldsymbol{s}_n \in \Omega^n$, where $\Omega = \{1, 2, \cdots, M\}$.*

Claim 7. *Denote $\boldsymbol{Pr}[S_{i,\eta} = s_i | S_{i-1,\eta} = s_{i-1}]$ as the probability that the considered secondary connection will select the channel s_i to be its target channel when an interruption event occurs at the channel s_{i-1}. Then, we have*

$$\begin{aligned} &\boldsymbol{Pr}[S_{i,\eta} = s_i | S_{i-1,\eta} = s_{i-1}] \\ = &\begin{cases} \prod\limits_{1 \leq j \leq M, j \neq s_{i-1}} \rho^{(j)} & , \quad s_i = s_{i-1} \\ (1-\rho^{(s_i)}) \sum\limits_{\forall \mathbb{A} \subseteq \Omega / \{s_{i-1}, s_i\}} \left[\frac{1}{1+|\mathbb{A}|} \prod\limits_{\forall v \in \mathbb{A}} (1-\rho^{(v)}) \prod\limits_{\forall v' \notin \mathbb{A}} \rho^{(v')} \right] & , \quad s_i \neq s_{i-1} \end{cases}, \end{aligned} \tag{7.8}$$

where $1 \leq i \leq N$ and N is the total number of interruptions for the considered secondary connection during its transmission period.

Proof. When an interruption event occurs at channel s_{i-1}, the type-$(i-1)$ secondary connection must search its target channel s_i for spectrum handoff through spectrum sensing. The probability that one channel is selected to be the target channel is related to the channel busy probabilities of all channels. If all channels are busy, the type-$(i-1)$ secondary connection will stay on its current operating channel (i.e., $s_i = s_{i-1}$). On the other hand, if these exists one idle channel, the type-$(i-1)$ secondary connection will change to this idle channel from channel s_i. Note that this type-$(i-1)$ secondary connection will randomly and uniformly select one channel from all idle channels to be its target channel if more than one channel is idle. From these observations, we can have (7.8). □

Proposition 8. *At channel k', denote $\omega_{i,\eta}^{(k\to k')}$ as the arrival rate of the type-i secondary connections which be redirected from channel k. Then, we have*

$$\omega_{i+1,\eta}^{(k')} = \sum_{k=1}^{M} \omega_{i+1,\eta}^{(k\to k')} \quad , \tag{7.9}$$

where

$$\omega_{i+1,\eta}^{(k\to k')} = \omega_{i,\eta}^{(k)} \cdot p_{i,\eta}^{(k)} \boldsymbol{Pr}[S_{i+1,\eta} = k' | S_{i,\eta} = k] \ . \tag{7.10}$$

Proof. When a type-i secondary connection is interrupted at channel k, it will turn into a new arrival of the type-$(i+1)$ secondary connection at channel k'. That is, the traffic loads of the type-$(i+1)$ secondary connections at channel k' can come from the remaining traffic loads of the type-i secondary connections at any one of M channels. Thus, the arrival rate of the type-$(i+1)$ secondary connections at channel k' can be expressed as (7.9).

The values of $\omega_{i+1,\eta}^{(k \to k')}$ can be evaluated as follows. For the type-i secondary connection at the channel k, it will be interrupted again with probability $p_{i,\eta}^{(k)}$. When an interruption event occurs at the channel k, the type-i secondary connection must search its target channel for spectrum handoff through spectrum sensing. Without loss of generality, we assume that the channel k' is selected to be the target channel. This situation occurs with probability $\mathbf{Pr}[S_{i+1,\eta} = k' | S_{i,\eta} = k]$. When channel k' is selected, the type-i secondary connection will turn into a new arrival of the type-$(i+1)$ secondary connection at channel k'. Hence, we can have (7.10). □

Proposition 9. *Based on the proposed PRP M/G/1 queueing network model, we can derive the closed-form expression for* $\boldsymbol{E}[\Phi_{i,\eta}^{(k)}]$.

Proof. According to the total probability principle, we have

$$\mathbf{E}[\Phi_{i,\eta}^{(k)}] = \sum_{N=1}^{L} \sum_{\forall \boldsymbol{s}_N \in \Omega^N} \mathbf{Pr}[\mathbf{S}^{(\eta)} = \boldsymbol{s}_N] \mathbf{E}[\Phi_{i,\eta}^{(k)} | \mathbf{S}^{(\eta)} = \boldsymbol{s}_N] \ , \tag{7.11}$$

where L is the maximum number of interruptions among all secondary users' connections, i.e., the maximum length of the target channel sequence. Based on the proposed queueing network, we can derive $\mathbf{E}[\Phi_{i,\eta}^{(k)} | \mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$ for any given $\boldsymbol{s}_N$ can been derived when $\lambda_p^{(k)}$, $\lambda_s^{(k)}$, $f_p^{(k)}(x)$, and $f_s^{(k)}(x)$ are given. The derivation detail can be found in [159]. As to $\mathbf{Pr}[\mathbf{S}^{(\eta)}$, it can be expressed as follows:

$$\mathbf{Pr}[\mathbf{S}^{(\eta)} = \boldsymbol{s}_N] = (1 - p_{N,\eta}^{(s_N)}) \prod_{i=1}^{N} \mathbf{Pr}[S_{i,\eta} = s_i | S_{i-1,\eta} = s_{i-1}] \ , \tag{7.12}$$

where $S_{0,\eta}$ is the default channel $s_0 = \eta$. By substituting (7.7) and (7.8) into (7.12), the value of $\mathbf{Pr}[\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$ can be obtained. □

7.3.2 An Example for the Exponentially Distributed Service Time

Now, we show how to derive channel utilization factor in the following special case. We assume that all secondary users' connections have the same service time distribution. Hence, we have $f_s^{(k)}(x) = f_s(x)$ and $\mathbf{E}[X_s^{(k)}] = \mathbf{E}[X_s]$, where $1 \leq k \leq M$. Furthermore, because this report focuses on the latency-sensitive traffic for the secondary users, it is reasonable to assume that the service time X_s is exponentially distributed (page 135 in [142]). Hence, we have $f_s(x) = \mu_s e^{-\mu_s x}$, where $\mu_s = \frac{1}{\mathbf{E}[X_s]}$.

Based on these traffic parameters, we derive $\rho^{(k)}$ as follows. Firstly, referring to [159] and the Proposition 9, we can have

$$\mathbf{E}[\Phi_{i,\eta}^{(k)}] = \frac{1}{\lambda_p^{(k)} + \mu_s} \ . \tag{7.13}$$

Secondly, from (7.7), it follows that

$$p_{i,\eta}^{(k)} = \lambda_p^{(k)} \mathbf{E}[\Phi_{i,\eta}^{(k)}] = \frac{\lambda_p^{(k)}}{\lambda_p^{(k)} + \mu_s} \ . \tag{7.14}$$

Then, according to the Proposition 8, we can derive $\omega_{i,\eta}^{(k')}$ as follows:

$$\omega_{i,\eta}^{(k')} = \sum_{k=1}^{M} \frac{\lambda_p^{(k)}}{\lambda_p^{(k)} + \mu_s} \cdot \omega_{i-1,\eta}^{(k)} \cdot \mathbf{Pr}[S_{i,\eta} = k' | S_{i-1,\eta} = k] \ . \tag{7.15}$$

Note that $\omega_{i,\eta}^{(k')}$ is a function of $\rho^{(k)}$ because $\mathbf{Pr}[S_{i,\eta} = k' | S_{i-1,\eta} = k]$ is a function of $\rho^{(k)}$. Furthermore, according to (7.6), we can find that $\rho^{(k)}$ is a function of $\omega_{i,\eta}^{(k)}$. Hence, we can determine $\omega_{i,\eta}^{(k')}$ and $\rho^{(k)}$ by solving (7.6) and (7.15) iteratively.

7.4 Analysis of Extended Data Delivery Time

In this section, we show how to evaluate the extended data delivery time, which is an important performance measure for the latency-sensitive traffic of the secondary connections. Without loss of generality, we consider the secondary connection whose default channel is the channel η in the following discussions. It extended data delivery time consists of the original service time $X_s^{(\eta)}$ and the cumulative delay resulting from multiple handoffs $\mathbf{E}[D^{(\eta)}]$. Let $\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$ be its cumulative handoff delay when its target channel is $\boldsymbol{s}_N$. Then, its average extended data delivery time (denoted by $\mathbf{E}[T^{(\eta)}]$) can be expressed as

$$\begin{aligned}\mathbf{E}[T^{(\eta)}] &= \mathbf{E}[X_s^{(\eta)}] + \mathbf{E}[D^{(\eta)}] \\ &= \mathbf{E}[X_s^{(\eta)}] + \sum_{N=0}^{L} \sum_{\forall \boldsymbol{s}_N \in \Omega^N} \mathbf{Pr}\{\mathbf{S}^{(\eta)} = \boldsymbol{s}_N\} \mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N] \quad (7.16)\end{aligned}$$

Hence, in order to calculate $\mathbf{E}[T^{(\eta)}]$, we will show how to evaluate $\mathbf{Pr}\{\mathbf{S}^{(\eta)} = \boldsymbol{s}_N\}$ and $\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$ in the following.

7.4.1 Derivations of $\mathbf{Pr}\{\mathbf{S}^{(\eta)} = \boldsymbol{s}_N\}$ and $\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$

In order to evaluate $\mathbf{Pr}\{\mathbf{S}^{(\eta)} = \boldsymbol{s}_N\}$ and $\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$, a state diagram is developed by characterizing the evolutions of the target channel sequence and the corresponding cumulative handoff delay for the secondary connection. The proposed state diagram is a two-dimensional chain. Because the considered secondary's default channel is η, the initial state of this state diagram is $(\eta, 0)$. Next, the state (k, i), where $1 \leq k \leq M$, represents that the channel k is selected for the target channel at the i^{th} interruption. The state $(M+1, i)$ represents that the secondary user can finish its transmission after the $(i-1)^{th}$ interruption, and thus the state $(M+1, i)$ is the ending of state

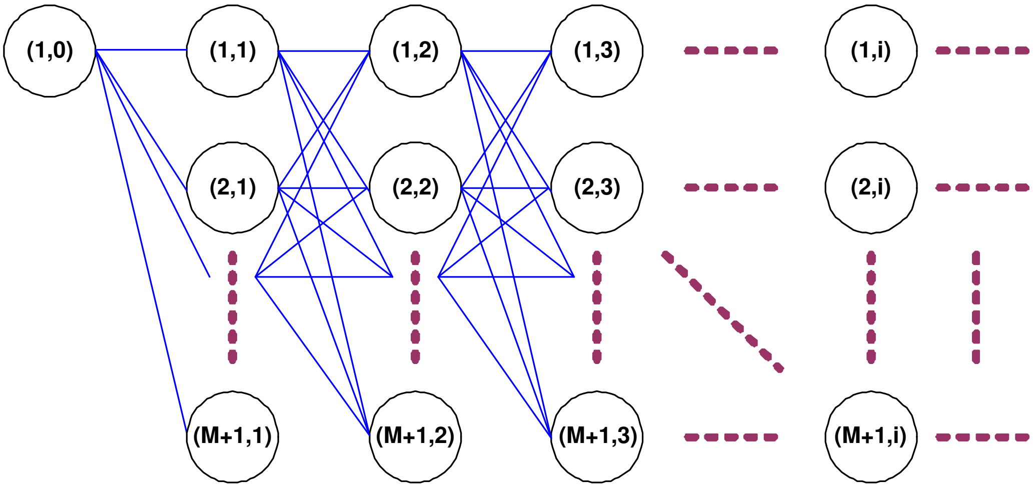


Figure 7.4: State diagram of target channel sequence for a secondary connection, where default channel $\eta = 1$.

transition. Note that the state transition occurs only at two adjacent states. Specifically, a transition link from (k, i) to (k', i') exists if $i' = i + 1$, and vice versa. An example of state diagram is shown in Fig. 7.4, where $\eta = 1$.

In this state diagram, the state transition path can be regarded as a target channel sequence. For example, for a target channel sequence $\boldsymbol{s}_n \triangleq (s_1, s_2, s_3, \cdots, s_n)$, the corresponding state transition path $(\eta, 0) \rightarrow (s_1, 1) \rightarrow (s_2, 2) \rightarrow (s_3, 3) \rightarrow \cdots \rightarrow (s_n, n) \rightarrow (M + 1, n + 1)$. Hence, calculating the average cumulative handoff delay over all possible target channel sequences can be regarded as calculating the cumulative transition cost over all possible state transition paths. In the following, we show how to design the state transition probability and cost in the developed state diagram.

State Transition Probability

When an interruption event occurs, the interrupted secondary connection must search its target channel for spectrum handoff through spectrum sensing. The probability that each channel is selected to be the target channel is related to the statistics of channel occupancy. Let $P[(k',i)|(k,i-1)]$ be the transition probability from states $(k,i-1)$ to (k',i). At channel k, the considered secondary connection may do not experience interruption again and can finish its transmission at channel k. In this case, the transition from states $(k,i-1)$ to $(M+1,i)$ will occur. On the other hand, the transition from states $(k,i-1)$ to (k',i) where $1 \leq k' \leq M$ will occur when the considered secondary connection is interrupted again at channel k. Thus, $P[(k',i)|(k,i-1)]$ can be expressed as follows:

$$P[(k',i)|(k,i-1)] = \begin{cases} 1 - p_{i-1,\eta}^{(k)} & , \quad k' = M+1 \\ p_{i-1,\eta}^{(k)} \cdot \mathbf{Pr}[S_{i,\eta} = k' | S_{i-1,\eta} = k] & , \quad k' \neq M+1 \end{cases} . \tag{7.17}$$

State Transition Cost

The cost of state transition is defined as the handoff delay of the interrupted secondary connection. The handoff delay from channels k to k' depends on the state of channel occupancy. Recall that δ_s and δ_c are the total processing time for executing spectrum handoff procedure when the secondary users stay on the current channel and change to another channel, respectively. If one idle channel exists after spectrum sensing, the interrupted secondary connection will change to this idle channel. Hence, the handoff delay in this case is δ_c. Furthermore, if all channels are busy, the interrupted secondary connection will stay on its current operating channel (i.e., $k = k'$). Hence, the expected handoff delay is the sum of δ_s and the duration from the time instant that channel k is used by the primary connections until the time instant that channel k becomes idle. This duration is called the *busy period* resulting from the transmissions of multiple primary connections at channel

k and denoted by $Y_p^{(k)}$. Let $C[k'|k]$ be the transition cost from states k to k' in the state diagram. Then, we can have

$$C[k'|k] = \begin{cases} 0 & , \quad k' = M+1 \\ \delta_s + \mathbf{E}[Y_p^{(k)}] & , \quad k' = k \\ \delta_c & , \quad \text{others} \end{cases} \quad . \tag{7.18}$$

Note that, referring to [159], we can have

$$\mathbf{E}[Y_p^{(k)}] = \frac{\mathbf{E}[X_p^{(k)}]}{1 - \lambda_p^{(k)} \mathbf{E}[X_p^{(k)}]} \quad . \tag{7.19}$$

From this developed state diagram, $\mathbf{Pr}\{\mathbf{S}_N^{(\eta)} = \boldsymbol{s}_N\}$ and $\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N]$ can be expressed as follows:

$$\mathbf{Pr}\{\mathbf{S}_N^{(\eta)} = \boldsymbol{s}_N\} = P[(M+1, N+1)|(s_N, N)] \prod_{i=0}^{N-1} P[(s_{i+1}, i+1)|(s_i, i)] \ , \tag{7.20}$$

and

$$\mathbf{E}[D|\mathbf{S}^{(\eta)} = \boldsymbol{s}_N] = \sum_{i=0}^{N} C[s_{i+1}|s_i] \ . \tag{7.21}$$

Note that (7.20) is equivalent to (7.12). Finally, substituting (7.20) and (7.21) into (7.16), we can obtain the closed-form expression for the extended data delivery time $\mathbf{E}[T^{(\eta)}]$ of the secondary connections with any service time distribution $f_s^{(\eta)}$ based on this developed state diagram.

7.4.2 An Example for the Exponentially Distributed Service Time

Now, we investigates how to derive the cumulative handoff delay when the secondary connection's service time is exponentially distributed as adopted in Section 7.3.2. Intuitively, we can evaluate the cumulative handoff delay by examining all possible transition paths in the state diagram, which is quite complex. Fortunately, the derivations of the cumulative handoff delay can be simplified due to the memoryless property of the exponential distribution.

Without loss of generality, we consider the secondary connection whose default channel is the channel k. It cumulative handoff delay $\mathbf{E}[D^{(k)}]$ in (7.16) can be derived as follows. Because the considered secondary connection's service time distribution is the exponential distribution, its remaining service time after an interruption event occurs also follows the identical exponential distribution. Hence, for the secondary connections at state (k, i) and (k', i'), they will experience the same cumulative handoff delay and interrupted probability in their remaining transmissions if $k = k'$, $k \neq M + 1$, and $k' \neq M + 1$.

From the aforementioned discussions, we can re-plot the state diagram expression for the target channel selection as a tree-structured representation as shown in Fig. 7.5, where Chk represents that channel k is selected for the target channel and the "grounding symbols" represent the endings of state transition. Note that at the second stage of Fig. 7.5, the average cumulative handoff delay of the type-1 secondary connection is equal to $\mathbf{E}[D^{(k)}]$ when this type-1 secondary connection's current operating channel is the channel k. Furthermore, because the state transition probability is independent of the number of interruptions for the secondary connections due to memoryless property, we can have $P[(k', i + 1)|(k, i)] = P[k'|k]$ for each $i \geq 0$. Hence, it follows that

$$\mathbf{E}[D^{(k)}] = P[M+1|k] \cdot C[M+1|k] + \sum_{k'=1}^{M} P[k'|k] \cdot (C[k'|k] + \mathbf{E}[D^{(k')}]) \ , \quad (7.22)$$

for any k where $1 \leq k \leq M$. Finally, substituting (7.17) and (7.18) into (7.22), we can obtain M independent equations. Hence, the closed-form expressions for the cumulative handoff delay $\mathbf{E}[D^{(k)}]$ can be derived by solving these simultaneous equations iteratively.

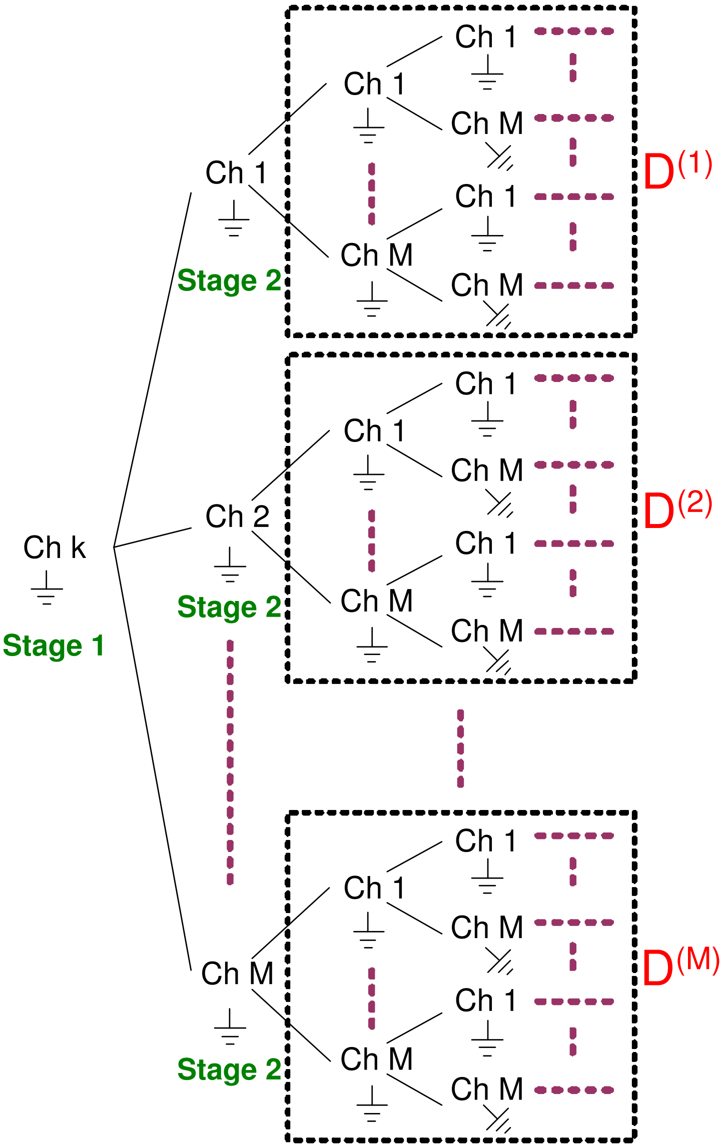


Figure 7.5: Tree-structured representations of the proposed state diagram where the grounding symbols represent the ending of state transition. Note that this figure considers the secondary connections whose default channels are Chk.

7.5 Numerical Results

We show numerical results to reveal the importance of the three key design features for modeling spectrum handoffs, which consist of (1) various arrival rates of the secondary users' connections; (2) heterogeneous arrival rates of the primary users' connections; and (3) the handoff processing time.

7.5.1 Simulation Setting

In order to validate the proposed analytical model, we perform simulations based on the Monte-carlo method in non-slot-based (continuous-time) cognitive radio systems, where the inter-arrival time and service time can be the duration of non-integer time slots. We consider a two-channel CR system with Poisson arrival processes of rates λ_p and λ_s for the high-priority primary connections and the low-priority secondary connections, respectively. The high-priority connections can interrupt the transmissions of the low-priority connections, and the connections with the same priority follow the first-come-first-served (FCFS) scheduling discipline[1]. Referring to the IEEE 802.22 standard, we adopt time slot duration of 10 msec in our simulations [145].

7.5.2 Effects of Various Arrival Rates for the Secondary Users' Connections

Firstly, we investigate the effects of various arrival rate for secondary users' connections on the channel utilization and the extended data delivery time

[1] In fact, the analytical results of mean values obtained in this report can be applied to other scheduling discipline which is independent of the service time of the primary and secondary connections because the averages of system performance metrics will be invariant to the order of service in this case (see page 113 in [134]).

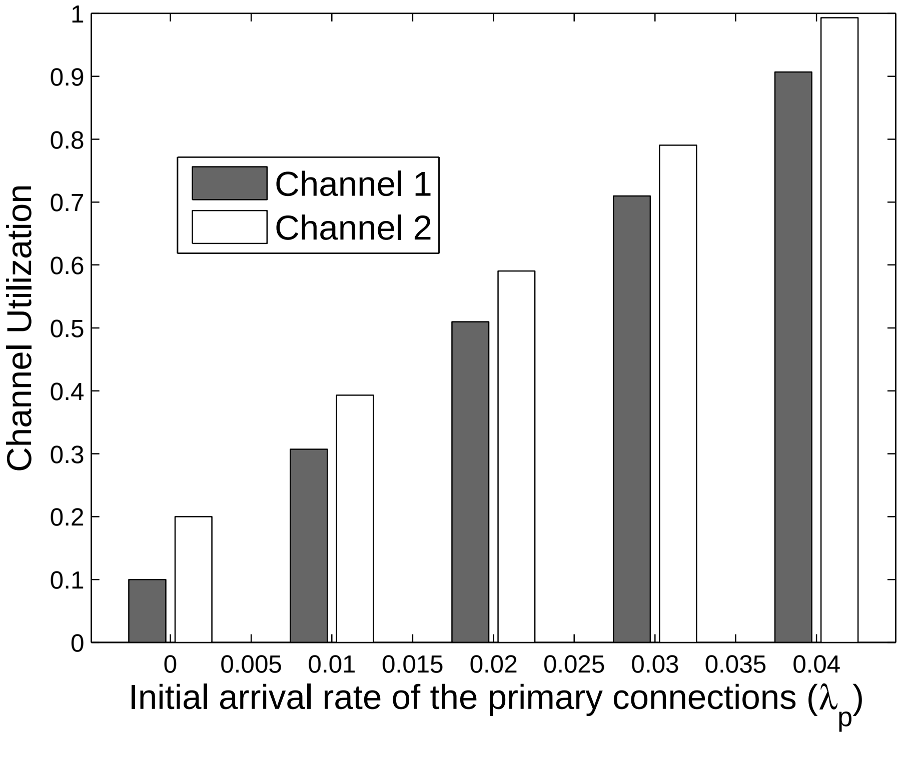


Figure 7.6: Effects of the arrival rate of the primary connections (λ_p) on the channel utilizations at the channels 1 and 2, where $\delta_s = 1$ and $\delta_c = 2$.

of the secondary connections. We consider a two-channel CR network, where $\lambda_p^{(1)} = \lambda_p^{(2)} \triangleq \lambda_p$, $(\lambda_s^{(1)}, \lambda_s^{(2)}) = (0.01, 0.02)$ (arrivals/slot), $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}]) = (20, 20)$ (slots/arrival), and $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}]) = (10, 10)$ (slots/arrival). We only consider the case that $0 \leq \lambda_p \leq 0.04$ (arrivals/slot) in the following numerical results. When $\lambda_p \geq 0.05$ (arrivals/slot), the overall normalized traffic workloads in the considered CR network will be saturated because $\lambda_p\mathbf{E}[X_p^{(1)}] + \lambda_p\mathbf{E}[X_p^{(2)}] + \lambda_s^{(1)}\mathbf{E}[X_s^{(1)}] + \lambda_s^{(1)}\mathbf{E}[X_s^{(1)}] > 2$.

Figure 7.6 shows the effects of the arrival rate of the primary connections (λ_s) on channel utilizations of the channels 1 and 2. As λ_p increases, the channel utilizations of the two channels also increase. Because the nor-

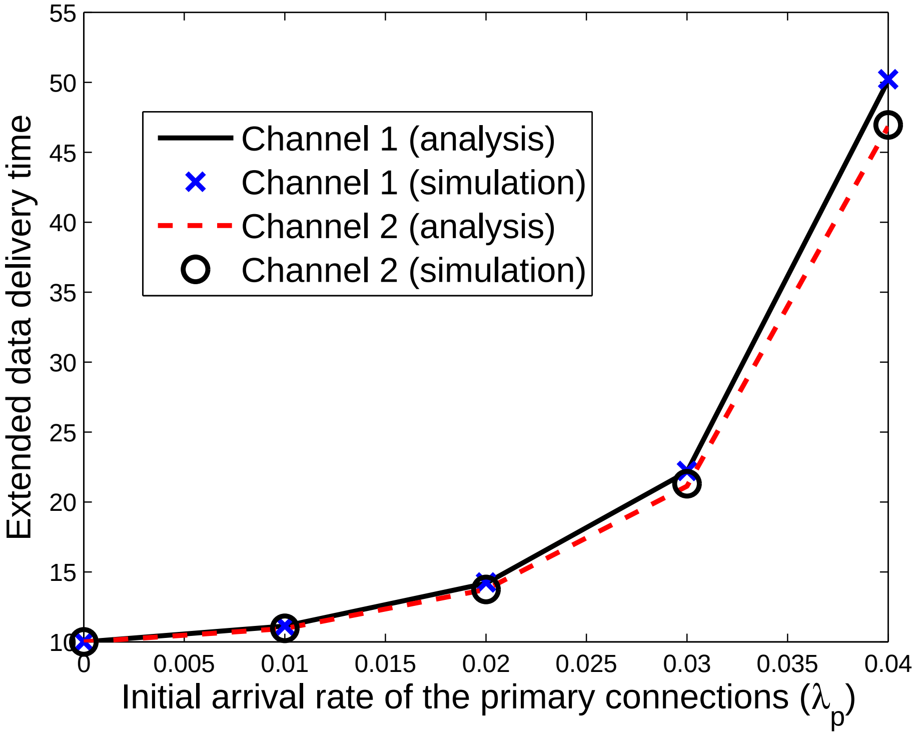


Figure 7.7: Effects of the arrival rate of the primary connections (λ_p) on the extended data delivery time of the secondary connections whose default channels are channels 1 and 2, where $\delta_s = 1$ and $\delta_c = 2$.

malized traffic workloads of the secondary connections are constant at each channel, the difference between channel utilizations of the two channels is also a constant, which equals to 0.1.

Figure 7.7 shows the effects of the arrival rate of the primary connections (λ_p) on the extended data delivery time of the secondary connections whose initial default channels are the channels 1 and 2. We have three important observations. First of all, because the channels 1 and 2 have the same arrival rate of the primary connections, the secondary connections at the two channels will encounter same interrupted probability according to (7.14). Hence, the secondary connections whose default channels are the channels 1 and 2 have similar extended data delivery time even though the channels 1 and 2 have different channel utilization. Next, it is shown that the extended data delivery time of the secondary connections increases as λ_p increases because a larger λ_s will lead to higher channel busy probabilities and longer average handoff delay. More importantly, we find that the simulation results match the analytical results quite well, which can validate the slot-based assumption used in our analysis.

7.5.3 Effects of Heterogeneous Arrival Rates for the Primary Users' Connections

Secondly, we demonstrate the effects of various arrival rate for primary users' connections on the channel utilization and the extended data delivery time of the secondary connections. We also consider a two-channel network, where $\lambda_s^{(1)} = \lambda_s^{(2)} \triangleq \lambda_s$, $(\lambda_p^{(1)}, \lambda_p^{(2)}) = (0.03, 0.01)$ (arrivals/slot), $(\mathbf{E}[X_s^{(1)}], \mathbf{E}[X_s^{(2)}]) = (20, 20)$ (slots/arrival), and $(\mathbf{E}[X_p^{(1)}], \mathbf{E}[X_p^{(2)}]) = (10, 30)$ (slots/arrival). Note that the two channels have the same channel utilizations resulting form the primary users' connections. Specifically, we have $\rho_p^{(1)} = \rho_p^{(2)} = 0.3$ (arrivals/slot). We only consider the case that $0 \leq \lambda_s \leq 0.03$ (arrivals/slot) in the following numerical results. When $\lambda_s \geq 0.04$ (arrivals/slot), the overall normalized traffic workloads in the considered CR network will be saturated because $\lambda_p^{(1)}\mathbf{E}[X_p^{(1)}] + \lambda_p^{(1)}\mathbf{E}[X_p^{(2)}] + \lambda_s\mathbf{E}[X_s^{(1)}] + \lambda_s\mathbf{E}[X_s^{(1)}] > 2$.

The effects of the initial arrival rate of the secondary connections (λ_s) on the channel utilizations of the channels 1 and 2 is shown in Fig. 7.8. When $\lambda_s = 0$, the two channels have the same channel utilizations of 0.3. As λ_s increases, the channel utilizations of the two channels also increase. However, the increases of the two channels are different even though the two chan-

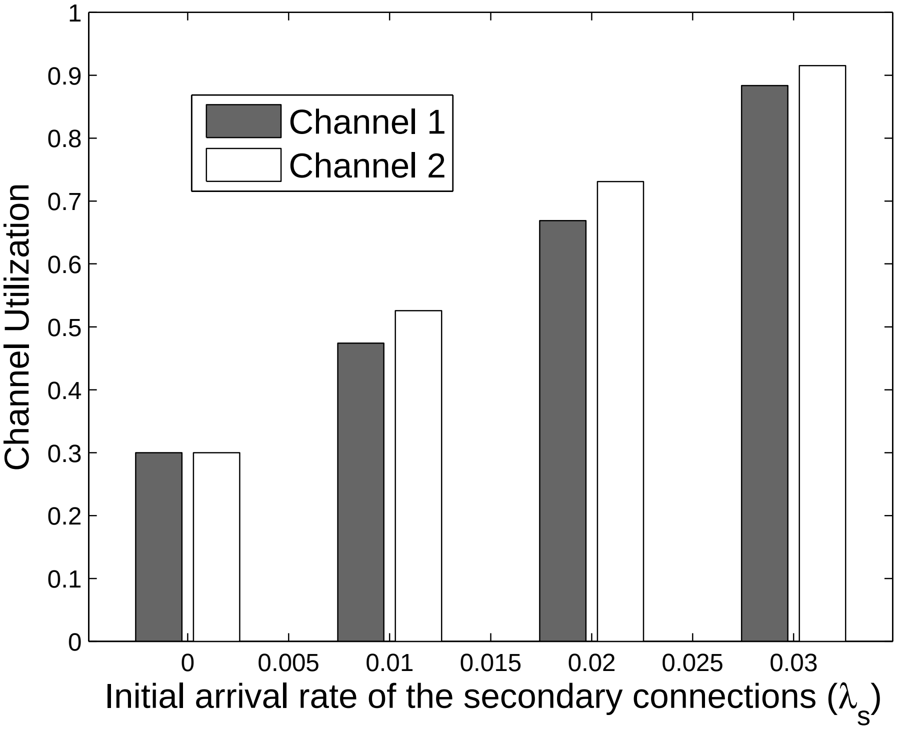


Figure 7.8: Effects of the initial arrival rate of the secondary connections (λ_s) on the channel utilizations at the channels 1 and 2, where $\delta_s = 1$ and $\delta_c = 2$.

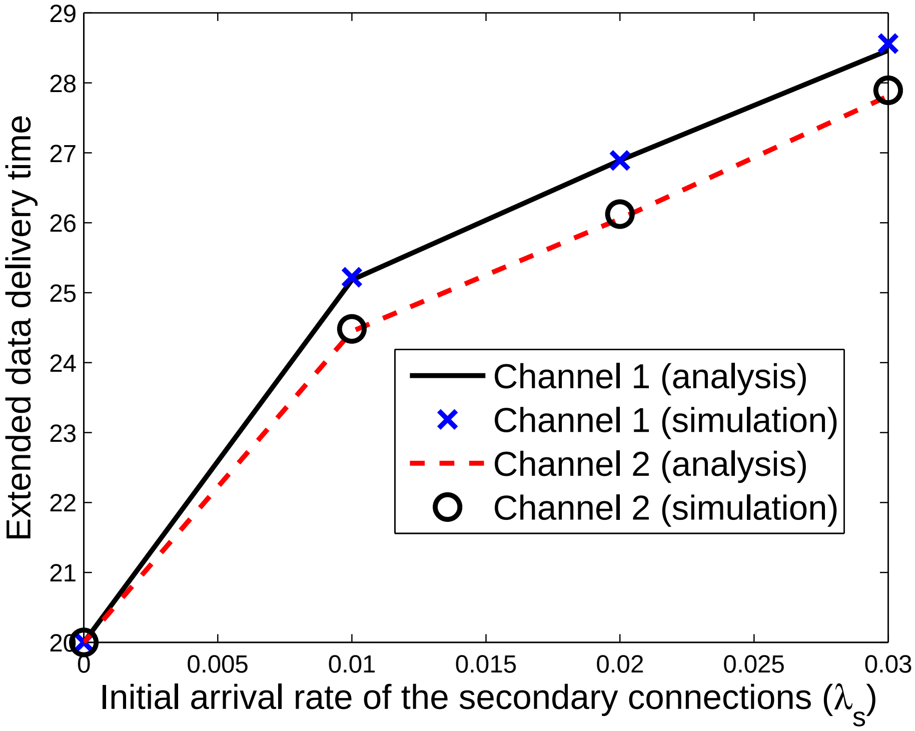


Figure 7.9: Effects of the initial arrival rate of the secondary connections (λ_s) on the extended data delivery time of the secondary connections whose default channels are the channels 1 and 2, where $\delta_s = 1$ and $\delta_c = 2$.

nels have the same busy probability resulting form the primary connections. Compared to the secondary connections at the channel 2, the secondary connections at the channel 1 will encounter higher interrupted probability because the channel 1 has larger arrival rate of the primary connections. Thus, the time that the secondary connections can use channel 1 is shorter than the time that the secondary connections can use channel 2. Hence, the increase of channel utilization at channel 2 is larger than that at channel 1.

Figure 7.9 shows the effects of the initial arrival rate of the secondary connections (λ_s) on the extended data delivery time of the secondary connections whose initial channels are the channels 1 and 2. We can find that the extended data delivery time of the secondary connections increases as λ_s increases because a larger λ_s will lead to a higher channel busy probabilities as shown in Fig. 7.8. Furthermore, the secondary connections whose initial channel is channel 2 has shorter extended data delivery time compared to the secondary connections whose initial channel is channel 1. This is because the secondary connections whose initial channel is channel 2 can have lower interrupted probability and the smaller number of interruptions during their transmission period.

7.5.4 Effects of Handoff Processing Time

Finally, we discuss the effects of handoff processing time. A two-channel CR network is considered with the following parameters: $t_n = 0$ (slot), $t_s = 1$ (slot), $\lambda_p^{(1)} = \lambda_p^{(2)} \triangleq \lambda_p$, $\lambda_s^{(1)} = \lambda_s^{(2)} = 0.02$ (arrivals/slot), $\mathbf{E}[X_p^{(1)}] = \mathbf{E}[X_p^{(2)}] = 5$ (slots/arrival) and $\mathbf{E}[X_s^{(1)}] = \mathbf{E}[X_s^{(2)}] = 10$ (slots/arrival). Then, based on the proposed analytical model, we can evaluate the average extended data delivery time and then design the admission control rule for the secondary users as shown in Figs. 7.10 and 7.11.

Figure 7.10 compares the cumulative handoff delay of the following three target channel selection schemes: (1) the always-staying strategy; (2) the random selection strategy; and (3) the reactive selection strategy. For the always-staying approach, the interrupted secondary user always stays on its default channel to resume its unfinished data transmission. The method is one kind of proactive spectrum handoff [141] because the target channels are predetermined and it is similar to the non-hopping mode of IEEE 802.22 [126]. In the random selection approach, the interrupted user randomly selects a target channel from all channels. From this figure, we have

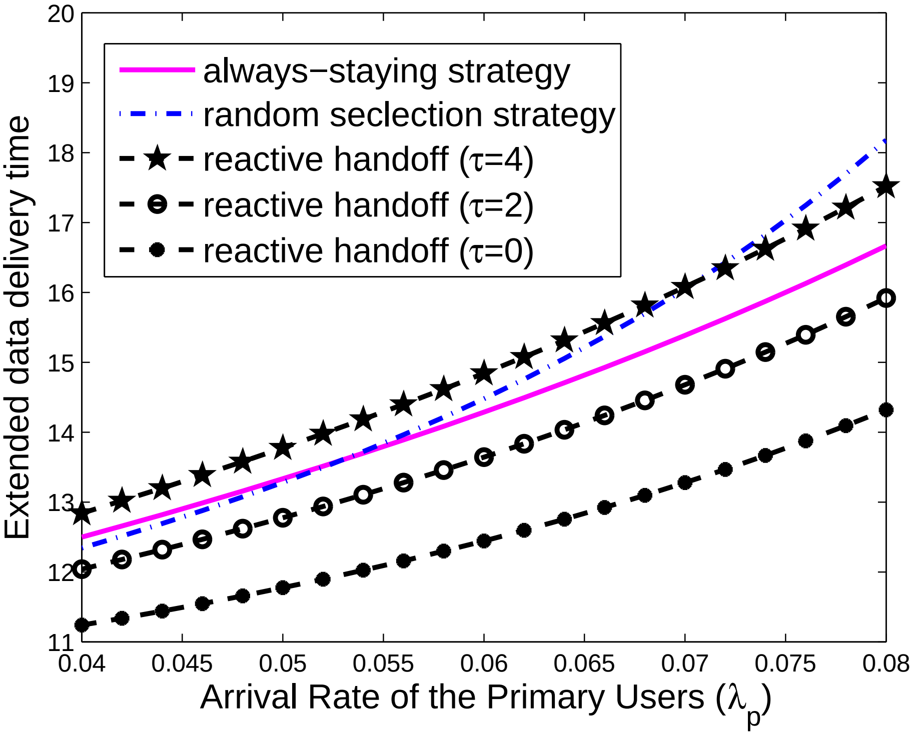


Figure 7.10: Comparison of average extended data delivery time for different target channel selection schemes.

the following three important observations. First, we find that the cumulative handoff delay resulting from the random selection method is longer than that resulting from the always-staying strategy when $\lambda_p \geq 0.052$. For a larger value of λ_p, the interrupted secondary users with the random selection method must spend much more time to wait when it changes its operating channel because this selected target channel is likely busy. Thus, the handoff delay of the random selection method becomes longer in this case. Next, it is shown that the reactive spectrum handoff can result in the shortest cumulative handoff delay in the ideal sensing case (i.e., $\tau = 0$) because it can reliably find idle channels by performing spectrum sensing. In this case, the cumulative handoff delay can be shortened around 40% compared to the other approaches under various arrival rates of the primary users' connections. However, when the spectrum sensing time becomes longer (e.g., $\tau = 4$), the reactive spectrum handoff is worse than the random selection method in terms of cumulative handoff delay when $\lambda_p \leq 0.071$. Finally, we find that the sensing technology can effectively shorten the average cumulative handoff delay only when $\tau \leq 2$ compared to the always-staying strategy.

The analytical results developed in this chapter can be used to design the admission control rule for the arriving secondary users subject to their latency requirement. Fig. 7.11 shows the admissible region for the normalized traffic workloads (or channel utilities) (ρ_p, ρ_s)[2] for the Voice over IP (VoIP) services when $\tau = 0$ (slot). The maximum allowable average cumulative delay resulting from multiple handoffs is 20 ms for the VoIP traffic [156]. The admission control policy can be designed according to this figure. When $\rho_p < 0.1667$, a CR network can accept all arrival requests from the secondary users until the CR network is saturated, i.e., $\rho_p + \rho_s \simeq 1$. Furthermore, when

[2] $\rho_p = \lambda_p \mathbf{E}[X_p]$ and $\rho_s = \lambda_s \mathbf{E}[X_s]$.

$0.1667 < \rho_p < 0.3397$, a part of traffic workloads of the secondary users must be rejected in order to satisfy the delay constraint for the secondary users. For example, when $\rho_p = 0.25$, a CR network can support at most 0.279 workload for the secondary users. That is, a CR network can accept at most $\lambda_s = 0.0279$ (arrivals/slot) based on the results shown in the figure when $\lambda_p = 0.05$ (arrivals/slot). In order to design the most allowable λ_s to achieve this arrival rate upper bound for the secondary connections, many arrival-rate control methods can be considered, such as the p-persistent carrier sense multiple access (CSMA) protocol in [25] and the call admission control mechanisms in [50,63,157]. Finally, when $\rho_p > 0.3397$, no secondary user can be accepted. Note that the size of the admissible region decreases as τ increases.

7.5.5 Comparison between Proactive and Reactive Spectrum Handoff Scheme

Figure 7.12 compares the extended data delivery time for the proactive and the reactive spectrum handoff schemes. Here, we consider a two-channel system with the following traffic parameters: $\lambda_s^{(1)} = \lambda_s^{(2)} = 0.01$, $\mathbf{E}[X_p^{(1)}] = \mathbf{E}[X_p^{(2)}] = 10$, and $\lambda_p^{(1)} = \lambda_p^{(2)} = \lambda_p$. From this figure, we have the following important observations. First, the extended data delivery time of the reactive spectrum handoff has a singular point at $\lambda_p = 0.043$. This is because the two different predetermined target channel sequences are adopted in the cases of $\lambda_p < 0.043$ and $\lambda_p > 0.043$. Based on the proposed model, the traffic-adaptive proactive spectrum handoff scheme can be designed to appropriately change to better target channel sequence according to traffic conditions in order to reduce the extended data delivery time of the secondary connections. Next, we focus on the reactive spectrum handoff scheme. In the ideal case that spectrum sensing time (denoted by τ) is 0 slot, the extended data

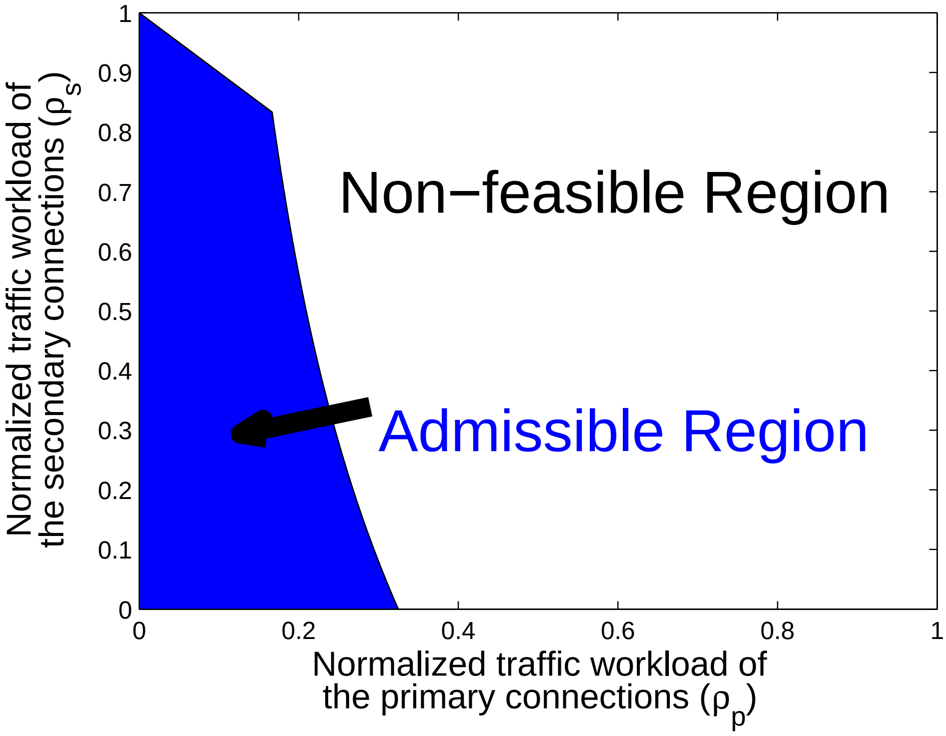


Figure 7.11: Admissible region (λ_p, λ_s), where the average extended data delivery time constraint can be satisfied when $\tau = 0$.

delivery time can be shortened around 7% $\sim$ 20%, compared to the proactive spectrum handoff over various arrival rates of the primary connections, because the reactive spectrum handoff scheme can perform spectrum sensing to find the idle channels. Furthermore, the extended data delivery time of the reactive spectrum handoff scheme increases as sensing time increases. When $\tau = 5$ (slots), the reactive handoff scheme is not always better than the proactive handoff scheme. As shown in this figure, when $\lambda_p < 0.037$, the proactive handoff scheme can result in shorter extended data delivery time. In this case, we can conclude that the proactive handoff scheme can yield shorter extended data delivery time compared to the reactive handoff scheme when the traffic loads of the primary users is light, whereas the reactive scheme performs better in the condition of heavy traffic loads. Finally, the reactive spectrum handoff scheme will result in the longest extended data delivery time when $\tau = 10$ (slots). Based on the proposed model, we can provide a principle to determine which spectrum handoff scheme should be adopted in CR networks for various sensing time and traffic parameters.

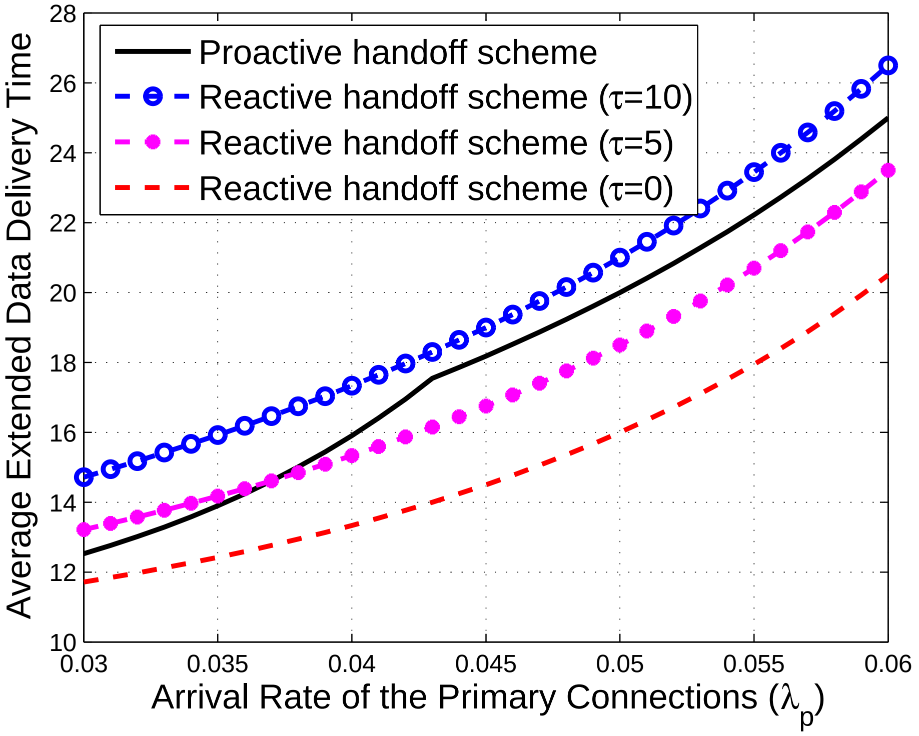


Figure 7.12: Comparison of the average extended data delivery time for different spectrum handoff schemes, where $\mathbf{E}[X_s] = 10$, $t_s = 0$, and $t_h = 0$

Chapter 8

Interference-Avoiding Spectrum Sharing

The goal of this part is to develop an analytical mechanism to design the admission control rule for the arriving secondary users subject to the following two quality of service (QoS) constraints:

- Interference avoidance for the primary users: Because the primary users have the highest priority to access channel, how to prevent the secondary users from interfering with the transmission of the primary users is the most important issue.
- Latency guarantee for the secondary users: The transmission latency of the secondary users is stained by many factors such as power outage due to fading channel, multiple handoffs due to interruptions from the primary users, and waiting time due to multiple secondary users.

In this chapter, we suggest a cross-layer approach to find the optimal traffic admission probability with the objective of maximizing channel utilization and maintaining the QoS requirements of the primary users and the secondary users. Our cross-layer design can incorporate the effects in the physical, medium access control (MAC), and application layers. In the physical layer, we incorporate the sensing errors for missed detection as well as false alarm, and power outage. In the MAC layer, traffic admission probability for the secondary users is considered. In the application layer, we consider the traffic statistics and QoS constraints of both the primary and the secondary users. The proposed analytical approach can calculate the optimal traffic admission probability under various cross-layer parameters. Furthermore, it also provides useful insights into the tradeoff design between channel utilization and the QoS performances for the primary as well as secondary users. To our knowledge, such a PHY/MAC/APP cross-layer analytical approach to determine the optimal traffic admission probability for CR networks has rarely been seen in the literature.

8.1 Motivation

In order to satisfy the two QoS constraints, the secondary users must be able to accurately sense the presence of the primary users. However, missed detection and false alarm may occur because the perfect sensing is impossible. Missed detection occurs when the detector reports the absence of a primary user while it is present. In this case, the transmission of the primary users will be affected by the secondary users. On the contrary, false alarm occurs when the detector mistakenly reports the presence of a primary user. In this situation, the secondary users cannot transmit data even though channel is indeed idle, which cause the transmission latency of the secondary users. Basically, a smaller missed detection probability implies a larger false alarm probability, and vice versa, which is a performance tradeoff design issue [143].

In the literature, many methods have been proposed to improve the issues of interference avoidance and latency guarantee simultaneously for the CR network under the imperfect sensing situation, such as the optimal sensing parameters and the optimal admission control. On one hand, the authors in [129, 130, 160] found that the secondary users can decrease the missed detection and the false alarm probabilities by increasing spectrum sensing time. On the other hand, [161] suggested to control the arrival rate of the secondary users to maintain two aforementioned QoS constraints. A lower arrival rate can decrease the interference on the primary users and reduce the waiting time due to multiple secondary users' contention.

In this chapter, we focus on designing an admission control mechanism to adaptively control the arrival rate of the secondary users by adjusting their traffic admission probability α in order to maintain the interference constraint of the primary users and the latency requirement of the secondary users. When the traffic of the secondary users arrives at system, it will be accepted with probability α and be dropped with probability $1-\alpha$. Intuitively, a larger traffic admission probability for the secondary users can increase channel utilization. However, a larger traffic admission probability degrades the QoS performance of the primary users due to much more interference from the secondary users if missed detection happens. Furthermore, a larger traffic admission probability also degrades the QoS performance of the secondary users due to more contention between the secondary users. Hence, there exists an optimal tradeoff between channel utilization in the system-level performance measure and interference ratio as well as transmission latency in the user-level performance measures.

8.2 System Model

8.2.1 Assumptions

In CR networks, there are four key system parameters. Assume that the arrival processes of the primary and the secondary users' connections are Poisson. Denote λ_p (arrivals/slot) and λ_s (arrivals/slot) as the traffic arrival rates of the primary and the secondary users' connections, respectively. Furthermore, let X_p (slots/arrival) and X_s (slots/arrival) be the service time of the primary users' connections and the secondary users' connections, respectively. Then, the probability density functions of X_p and X_s are denoted by $f_p(x)$ and $f_s(x)$, respectively.

The service time of the primary and secondary connections will be extended due to imperfect sensing and power outage. Denote P_M and P_F as missed detection and false alarm probabilities, respectively. Their relationship can be characterized by the receiver operating characteristic (ROC) curve [143]. When missed detections occur, the primary user must retransmit these stained data frames in the next slots. Thus, the service time of a primary connection will be extended from X_p (slots/arrival) to $\widetilde{X}_p$ (slots/arrival). Furthermore, a secondary user cannot transmit data even with an idle channel when a false alarm occurs. Hence, a secondary user needs to spend more time to complete its connection transmission. Then, the service time of a secondary connection will be extended to $\widetilde{X}_s$ (slots/arrival) from X_s (slots/arrival). Furthermore, power outage will extend the actual service time to Z_p (slots/arrival) and Z_s (slots/arrival) from $\widetilde{X}_p$ (slots/arrival) and $\widetilde{X}_s$ (slots/arrival), respectively. In the remaining part of this chapter, we call $\widetilde{X}_p$ and $\widetilde{X}_s$ the actual service time of the primary and secondary connections as well as Z_p and Z_s the actual service time of the primary and secondary connections in the physical channel.

8.2.2 Admission Control Mechanism

An arrival rate control mechanism for the secondary connections can be used to satisfy the interference constraint of the primary users and the latency requirement of the secondary users. When a new traffic request arrives at system, it will be dropped with probability $1-\alpha$ and be accepted with probability α. If this request is accepted, the secondary user will perform spectrum sensing to identify channel availability and then decide whether a secondary connection can be established. By contrary if this request is dropped, it will be retransmitted later by the upper layer protocol such as automatic repeat-request (ARQ) protocol. Here, α is called the traffic admission probability of the secondary connections. Because $0 \geq \alpha \geq 1$, the effective arrival rate $\alpha\lambda_s$ is not higher than the original arrival rate. This kind of arrival rate control mechanism is very similar to the concept of p-persistent carrier sense multiple access (CSMA) protocol as discussed in [25].

8.3 Problem Formulation and Analytical Model

8.3.1 Problem Formulation

In order to maximize channel utilization while maintaining the interference and latency requirements, we formulate the **Utilization Maximization Problem** for the secondary users as follows. Given the maximum allowable interference ratio Θ_{max} on the primary connections and the longest allowable overall system time S_{max} of the secondary connections, we aim to find the optimal traffic admission probability (denoted by α^*) to maximize the channel utilization (denoted by U). Formally,

$$\alpha^* = \arg\max_{0<\alpha\leq 1} U(\alpha) \ , \tag{8.1}$$

subject to

$$\Theta_p(\alpha) \triangleq \frac{\mathbf{E}[Z_p(\alpha)]}{\mathbf{E}[X_p]} \leq \Theta_{max} \ , \tag{8.2}$$

and

$$\mathbf{E}[S_s(\alpha)] \leq S_{max} \ , \tag{8.3}$$

where $\mathbf{E}[\cdot]$ is the expectation function and $S_s(\alpha)$ is the average overall system time of the secondary connections, which is defined as the duration from the instant that data arrives at system until the instant of finishing the whole transmission. Note that from queueing theory, it follows that

$$U(\alpha) = \lambda_p \mathbf{E}[X_p] + \alpha\lambda_s \mathbf{E}[X_s] \ . \tag{8.4}$$

From (8.4), we can found that $U(\alpha)$ is a strictly increasing function of α. Hence, our optimization problem can be solved by maximizing α while maintaining the constraints (8.2) and (8.3).

In the **Utilization Maximization Problem**, (8.2) and (8.3) represent the interference and latency constraints of the primary and secondary connections in the application layer, respectively. Obviously, U, $\mathbf{E}[Z_p]$ and $\mathbf{E}[S_s]$ are related to not only α in the MAC layer and traffic statistics of the primary and secondary users in the application layer but also missed detection probability P_M, false alarm probability P_F, outage probability π_p for the primary users, and outage probability π_s for the secondary users in the physical layer. To solve this optimization problem, we use the preemptive resume priority (PRP) M/G/1 queueing model to evaluate these unknown system performance measures. We will detail this queueing mode in the following.

8.3.2 Analytical Model

In this chapter, we derive the closed-form expressions for $\mathbf{E}[Z_p]$ and $\mathbf{E}[S_s]$ based on the preemptive resume priority (PRP) M/G/1 queueing model [135]. Some important properties for the PRP M/G/1 queueing model are listed below:

- Each server (channel) has two types of customers (connections). The connections of the primary and secondary users are connected to the high-priority queue and the low-priority queue, respectively.
- The primary users have the preemptive priority to interrupt the transmission of the secondary users. The remaining transmission of the interrupted secondary user will be put into the head of the low-priority queue of the current operating channel. Furthermore, the interrupted secondary user can resume the unfinished transmission when the current channel becomes idle, instead of retransmitting the whole data.
- A secondary connection may encounter multiple interruptions from the primary connections during its transmission period. This model can characterize the effects of multiple spectrum handoffs.

Here, we assume that connections which have the same priority access channels with the first-come-first-served (FCFS) scheduling discipline. Based on this model, when the four traffic parameters λ_p, λ_s, $f_p(x)$, and $f_s(x)$ are known, $\mathbf{E}[Z_p]$ and $\mathbf{E}[S_s]$ can be evaluated analytically in the next section.

8.4 Analysis of Constraint Functions in the Utilization Maximization Problem

This section show how to derive the closed-form expressions for $\mathbf{E}[Z_p]$ and $\mathbf{E}[S_s]$ in (8.2) and (8.3).

8.4.1 Analysis of Actual Service Time of the Primary Connection in the Physical Channel

Effect of Missed Detection

Basically, missed detections in CR networks can be categorized into two kinds [64, 66]. Firstly, when a primary user is transmitting data, a newly arriving secondary connection may incorrectly assess that this specific channel is available in its first sensing phase. In this case, the class-A missed detection occurs. Next, a secondary user may also fail to detect the presence of primary users after it arrives at a CR network for a while. This situation is called the class-B missed detection. The authors in [64, 66] found that the class-B missed detection is small because the sensing results at the first sensing phase can be employed to improve the accuracy of the sensing results at the following sensing phases.

Next, we explain the effect of class-A missed detection on the actual service time of the primary connections. We consider a transmission slot of this primary connection. During this slot, more than one arrival of the secondary connection appears with probability $1 - e^{-\lambda_s \Delta}$, where Δ is the slot duration. For these arrivals of secondary connections, each of them will assess this busy slot as idle if and only if (1) a missed detection occurs and (2) the low-priority queue of the considered channel is empty. Let Q_s be the length of the low-priority queue. Hence, the first arrival at the considered slot will make an error channel assessment with probability $P_M \mathbf{Pr}\{Q_s = 0\}$, where P_M is the missed detection probability for spectrum sensing and $\mathbf{Pr}\{Q_s = 0\}$ has been derived in [144]. However, for the remaining arrivals in the considered slot, we have $\mathbf{Pr}\{Q_s = 0\} = 0$ because the first arrival has been put into the low-priority queue. Thus, the remaining arrivals do not make the error channel assessment. From above observations, we can conclude that a primary connection's transmission slot is stained by the arrivals of the secondary connections with probability

$$P_I = (1 - e^{-\lambda_s \Delta}) P_M \mathbf{Pr}\{Q_s = 0\} \ . \tag{8.5}$$

Then, we consider an observation period with $I + B$ slots, where I and B are the total numbers of idle and busy slots resulting from the primary connections when missed detections do not occur. Hence, we have $\frac{B}{I+B} = \lambda_p \mathbf{E}[X_p]$. For the primary connections, a total of BP_I slots out of B slots must be retransmitted at the next slot due to missed detection. Furthermore, when these retransmitted data are stained again, they must be retransmitted. On average, a total of $B(P_I)^2$ slots out of BP_I slots must be retransmitted. The similar arguments can be applied for all the upcoming retransmissions. Hence, it follows that

$$\rho_p = \frac{B \sum_{i=0}^{\infty} (P_I)^i}{I + B} = \frac{\lambda_p \mathbf{E}[X_p]}{1 - P_I} \ . \tag{8.6}$$

Finally, because $\rho_p = \lambda_p \mathbf{E}[\widetilde{X}_p]$ in (8.6), the expected actual service time $\mathbf{E}[\widetilde{X}_p]$ can be written as

$$\mathbf{E}[\widetilde{X}_p] = \frac{\mathbf{E}[X_p]}{1 - P_I} \ . \tag{8.7}$$

Effect of Power Outage

When outage occurs, the users must retransmit the failed slot. Hence, the actual transmission time of the primary connections in the physical channel will be extended from $\widetilde{X}_p$ to Z_p. The first and the second moments of Z_p can be expressed as follows:

$$\mathbf{E}[Z_p] = \sum_{\widetilde{x}=1}^{\infty} \mathbf{E}[(Z_p)|\widetilde{X}_p = \widetilde{x}]\mathbf{Pr}(\widetilde{X}_p = \widetilde{x}) \ , \tag{8.8}$$

and

$$\mathbf{E}[(Z_p)^2] = \sum_{\widetilde{x}=1}^{\infty} \mathbf{E}[(Z_p)^2|\widetilde{X}_p = \widetilde{x}]\mathbf{Pr}(\widetilde{X}_p = \widetilde{x}) \ . \tag{8.9}$$

When outage occurs, the failed slots must be retransmitted in the next slot. Hence, for a connection with transmission duration $\widetilde{x}$, its actual service time Z_p in the physical channel will be extended to $z+i$ if and only if outages occur in i slots of the first $\widetilde{x}+i-1$ slots and outage does not occur at the $(z+i)^{th}$ slot. Hence, the conditional expectation of the actual service time in the physical channel follows the negative binomial distribution. That is,

$$\mathbf{E}[Z_p|\widetilde{X}_p = \widetilde{x}] = \sum_{i=0}^{\infty} (\widetilde{x}+i)\binom{\widetilde{x}+i-1}{i}(1-\pi_p)^{\widetilde{x}}\pi_p^i \ , \tag{8.10}$$

and

$$\mathbf{E}[Z_p^2|\widetilde{X}_p = \widetilde{x}] = \sum_{i=0}^{\infty} (\widetilde{x}+i)^2\binom{\widetilde{x}+i-1}{i}(1-\pi_p)^{\widetilde{x}}\pi_p^i \ . \tag{8.11}$$

where π_p is the outage probability for the primary connections.

Finally, when $\mathbf{Pr}(\widetilde{X}_p = \widetilde{x})$ is given, we can obtain $\mathbf{E}[Z_p]$ and $\mathbf{E}[(Z_p)^2]$ by substituting (8.10) into (8.8) and (8.11) into (8.9), respectively. Note that how to derive $\mathbf{Pr}(\widetilde{X}_p = x)$ from $f_p(x)$ has been discussed in Appendix D. For example, if $f_p(x)$ is the geometric distribution, i.e.,

$$f_p(x) = (1 - \frac{1}{\mathbf{E}[X_p]})^{x-1}(\frac{1}{\mathbf{E}[X_p]}) \ , \tag{8.12}$$

we can have

$$\mathbf{E}[Z_p] = \frac{\mathbf{E}[\widetilde{X}_p]}{1-\pi_p} = \frac{\mathbf{E}[X_p]}{(1-P_I)(1-\pi_p)} \ , \tag{8.13}$$

and

$$\mathbf{E}[(Z_p)^2] = \frac{\mathbf{E}[X_p](2\mathbf{E}[X_p]-1+P_I+\pi_p-P_I\pi_p)}{((1-P_I)(1-\pi_p))^2} \ , \tag{8.14}$$

where π_p is the outage probability of the primary connections.

8.4.2 Analysis of Overall System Time of the Secondary Connections

The overall system time (denoted by S_s) is an important quality of service (QoS) metric for the secondary users' connections. It consists of the waiting time (denoted by W) and the extended data delivery time (denoted by T). Hence, we have

$$\mathbf{E}[S_s] = \mathbf{E}[W] + \mathbf{E}[T] \ , \tag{8.15}$$

Here, the waiting time is defined as the duration from the instant that data arrives at system until the instant of starting transmitting data. Furthermore, the extended data delivery time is defined as the duration from the instant of starting transmitting data until the instant of finishing the whole transmission.

In addition to traffic admission probability α, the overall system time of the secondary users is also affected by channel contention, multiple handoffs, false alarm, and power outage issues. Firstly, the secondary users' channel contention will increase waiting time. Furthermore, a secondary connection may have multiple interruptions from the primary user. Moreover, when false alarm occurs, the secondary users cannot transmit data even though channel is truly idle. Finally, power outage will make data retransmission. These phenomenons will extend the overall system time. In the following subsections, we will investigate the effects of waiting time due to channel contention, multiple handoffs, false alarm probability P_F, and outage probability π_s on $\mathbf{E}[S_s]$.

Effect of Waiting Time Due to Multiple Secondary Users' Contention

When a secondary connection arrives at system, it cannot be transmitted immediately until all the secondary connections in the low-priority queue and the primary connections in the high-priority queue have been served. Hence, when more secondary users access channel, waiting time will increase. Referring to [135], the average waiting time can be expressed as follows:

$$\mathbf{E}[W] = \frac{\frac{1}{2}\lambda_p \mathbf{E}[(Z_p)^2] + \frac{1}{2}\alpha\lambda_s \mathbf{E}[(Z_s)^2]}{(1 - \lambda_p \mathbf{E}[Z_p] - \alpha\lambda_s \mathbf{E}[Z_s])(1 - \lambda_p \mathbf{E}[Z_p])} \ . \tag{8.16}$$

Effects of Multiple Handoffs

The extended data delivery time of each secondary connection consists of the actual service time Z_s in the physical channel and the cumulative delay resulting from multiple handoffs. Let N and D be the number of interruptions for the secondary connection and the handoff delay for each spectrum handoff. Then, we have

$$\mathbf{E}[T] = \mathbf{E}[Z_s] + \mathbf{E}[N]\mathbf{E}[D] \ . \tag{8.17}$$

Referring to [135], it follows that

$$\mathbf{E}[D] = \frac{\mathbf{E}[Z_p]}{1 - \lambda_p \mathbf{E}[Z_p]} \ , \tag{8.18}$$

and

$$\mathbf{E}[N] = \lambda_p \mathbf{E}[Z_s] \ . \tag{8.19}$$

Effect of False Alarm

Now, we consider an observation period with $I+B$ slots, where I and B are the total numbers of idle and busy slots resulting from the secondary connections when false alarms and power outages do not occur. Hence, we have $\frac{B}{I+B} = \lambda_s \mathbf{E}[X_s]$. Recall that ρ_s is the busy probability resulting from the secondary connections. Hence, in the duration of $I+B$, there are $\rho_s(I+B)$ slots are busy, where $\rho_s(I+B)P_F$ slots is the false-alarm busy slots. That is, the total number of busy slots resulting from the secondary connections is increased to $B+\rho_s(I+B)P_F$. From this observation, we have the following relationship:

$$\rho_s = \frac{B+(I+B)\rho_s P_F}{I+B} \quad . \tag{8.20}$$

Solving (8.20), we can have

$$\rho_s = \frac{B}{(I+B)(1-P_F)} = \frac{\lambda_s \mathbf{E}[X_s]}{1-P_F} \quad . \tag{8.21}$$

Finally, because $\rho_s = \lambda_s \mathbf{E}[\widetilde{X}_s]$ in (8.21), the expected actual service time $\mathbf{E}[\widetilde{X}_s]$ can be written as

$$\mathbf{E}[\widetilde{X}_s] = \frac{\mathbf{E}[X_s]}{1-P_F} \quad . \tag{8.22}$$

Effect of Power Outage

Referring to 8.4.1, if $f_s(x)$ is the geometric distribution, i.e.,

$$f_s(x) = (1-\frac{1}{\mathbf{E}[X_s]})^{x-1}(\frac{1}{\mathbf{E}[X_s]}) \quad , \tag{8.23}$$

we can have

$$\mathbf{E}[Z_s] = \frac{\mathbf{E}[X_s]}{(1-P_F)(1-\pi_s)} \quad , \tag{8.24}$$

and

$$\mathbf{E}[(Z_s)^2] = \frac{\mathbf{E}[X_s](2\mathbf{E}[X_s]-1+P_F+\pi_s-P_F\pi_s)}{((1-P_F)(1-\pi_s))^2} \quad , \tag{8.25}$$

where π_s is the outage probability of the secondary connections.

8.5 Numerical Results

In this section, we show the impacts of different system parameters on Θ_p, $\mathbf{E}[S_s]$, and α^*. We consider the following system parameters: $\lambda_p = 0.02$ and $\pi_p = \pi_s = 0.1$. Furthermore, because this report focuses on the latency-sensitive traffic, we can assume that the service time X_p and X_s of the primary and secondary connections follow the geometric distributions (see page 135 in [142]). Note that we only use the geometric distribution as an example here. Indeed, the proposed analytical framework can be applied to any distributions. It only requires the knowledge of the first and the second moments of the data transmission time distributions for the primary and the secondary connections.

Figure 8.1 compares the interference ratio (Θ_p) for various traffic admission probabilities (α). Obviously, Θ_p increases as the arrival rate (λ_s) of the secondary connections increases. Furthermore, because a larger α will lead to much more traffic loads of the secondary connections, the primary connections will be stained with a larger probability. Based on the analytical results, we can design an admission control rule to satisfy the interference constraint on the primary connections. For example, we consider $\lambda_s = 0.019$ and $\Theta_{max} = 1.1175$. In this case, when admission control had not implemented (i.e. $\alpha = 1$), the interference ratio is 1.1325, which is larger than Θ_{max}. In order to satisfy the interference constraint, we must lower the effective traffic loads of the secondary connections by setting $\alpha = 0.3$.

The effects of the traffic admission probability α and the arrival rate λ_s of the secondary connections on the average overall system time ($\mathbf{E}[S_s]$) of the secondary connections in shown in Fig 8.2. We can found that the average overall system time increases as α or λ_s increases. Similarly, we can also develop the admission control rule for the arriving secondary users

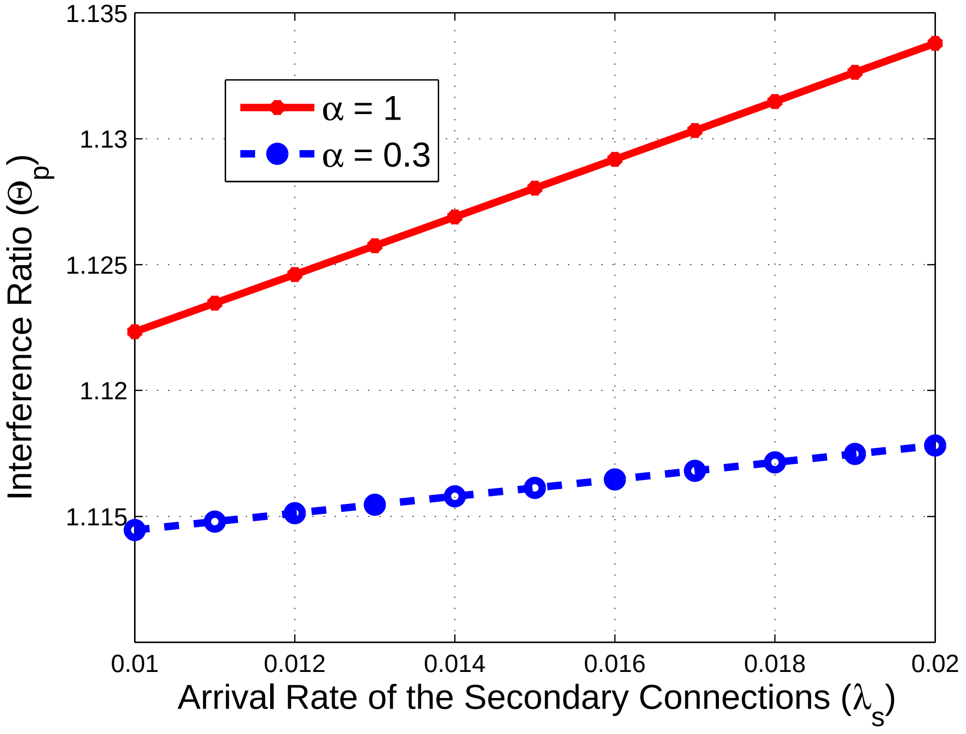


Figure 8.1: Interference ratio (Θ_p) for various arrival rates of the secondary connections, where $P_M = 0.1$.

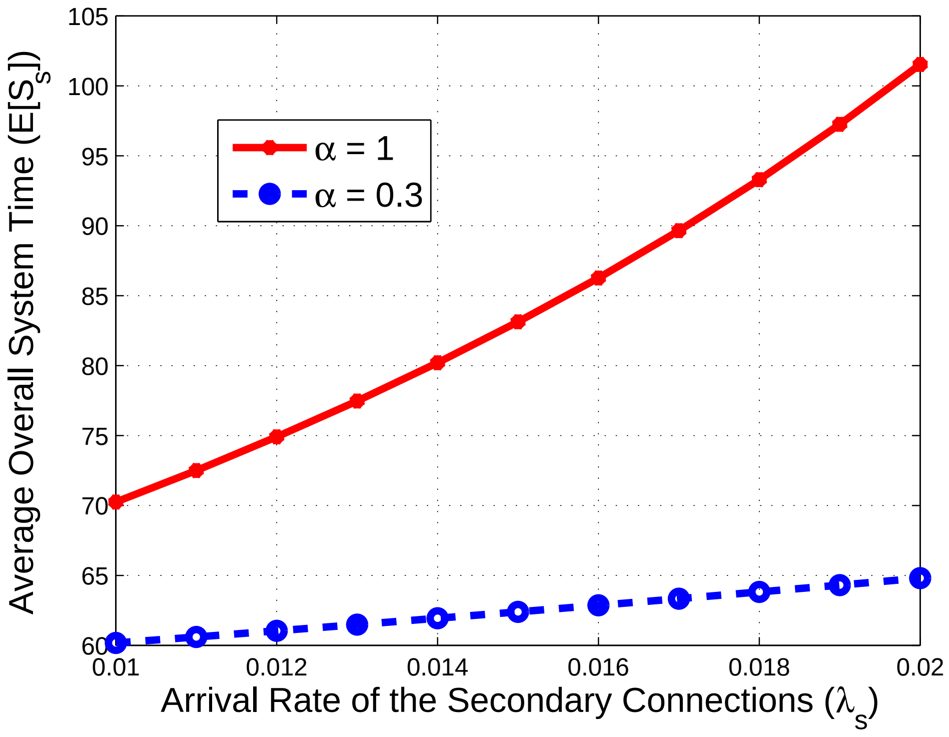


Figure 8.2: Average overall system time ($\mathbf{E}[S_s]$) for various arrival rates of the secondary connections, where $P_F = 0.1$.

subject to their latency requirement. For example, we consider $\lambda_s = 0.016$ and $S_{max} = 63$. If we do not use admission control mechanism, the average overall system time is 86, which is larger than S_{max}. Hence, a part of traffic workloads of the secondary users must be rejected in order to satisfy the delay constraint for the secondary users. In the considered case, we must set $\alpha = 0.3$.

The optimal traffic admission probability for various arrival rates of the secondary connections is shown in Fig. 8.3. This figure shows that α^* decreases as λ_s increases because a larger λ_s implies much more interference and channel contention. Furthermore, a higher P_F will lead to longer ex-

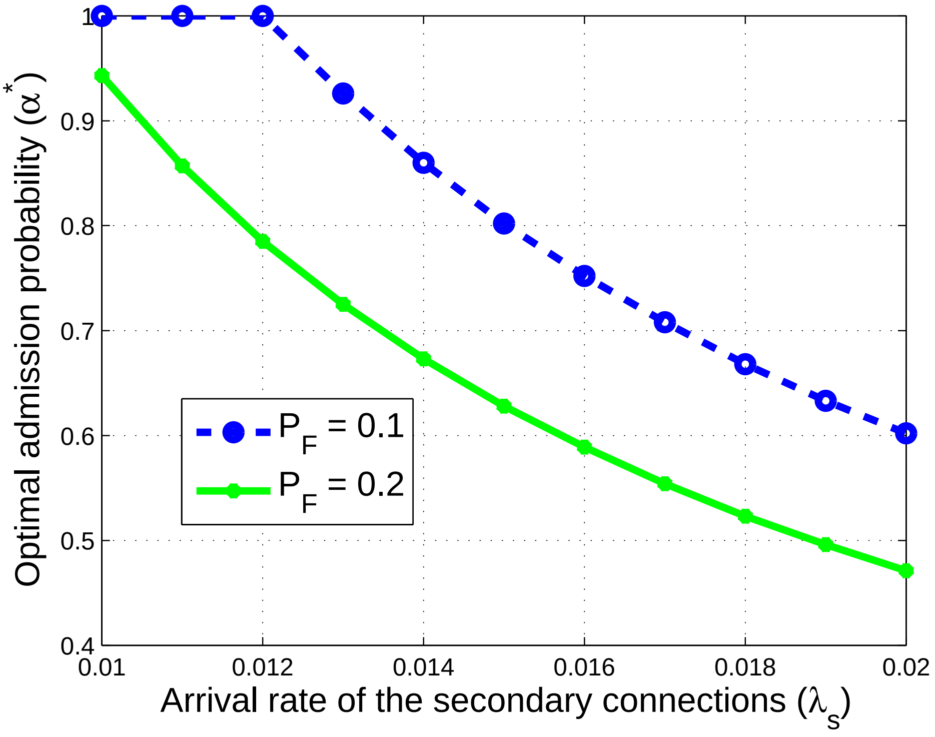


Figure 8.3: Optimal traffic admission probability for various arrival rates of the secondary connections where $\Theta_{max} = 1.13$ and $S_{max} = 75$.

tended data delivery time. Hence, the secondary users must reduce their waiting time in order to maintain the same overall system time requirement. In order to alleviate channel contention between multiple secondary users, the total traffic loads of the secondary users must be lowered. Hence, α^* decreases as P_F increases. For example, when $\lambda_s = 0.014$, we have $\alpha^* = 0.86$ for $P_F = 0.1$ and $\alpha^* = 0.65$ for $P_F = 0.2$, respectively

Chapter 9

Latency Analysis for Spectrum Sharing

Cognitive radio (CR) aims to improve spectrum efficiency by allowing the secondary users to temporally utilize the unused licensed spectrum by the dynamic spectrum sharing (DSS) technique. There are many important challenges needed to be overcome to achieve the objective of DSS. One of key challenges is the fact that the secondary users can only borrow the licensed spectrum from the primary users for a short period of time, and will lose the channel access privilege when the primary user appears. Hence, unlike many available multi-channel MAC protocols for ad hoc networks where throughput is the main performance issue, the MAC protocols for CR networks shall place more emphases on the access latency. Access latency to the borrowed licensed spectrum for the secondary users should be minimized because the primary user may appear to take back its spectrum any time.

In this chapter, we propose an analytical approach to evaluate the latency performance of MAC protocols in CR networks with the dedicated and the embedded control channels. The key ingredient of the analytical approach is the virtual slot concept. The major contributions of this paper are two folds:

1. We find that, in the case of dedicated control channels, an optimal ratio for the control channel bandwidth over the total bandwidth can be found to yield the minimal access delay.

2. From a viewpoint of access delay, the MAC protocols with dedicated control channels and embedded control channels are compared for various data lengths and different number of users.

3. We analyze in which condition dependent of traffic parameters that the dedicated or the embedded control channel scheme should be used.

9.1 Motivation

How to negotiate a spectrum for the communication of between the secondary transmitter and the secondary receiver is a key issue in MAC protocol design for CR network [162]. Intuitively, in order to shorten the negotiation latency when the secondary transmitter and the secondary receiver need to decide the operating channel for their communication, we use control channels to exchange the negotiation information [163]. Basically, the design of control channels can be categorized into two approaches form frequency viewpoint: 1) dedicated control channel scheme, in which control and data frames are transmitted on different spectra; and 2) embedded control channel scheme, where the control and data frames are transmitted on the same spectrum. However, the former scheme encounters the control channel saturation issue and the later scheme encounters channel mismatched issue. These two issues will degrade the delay performance of the secondary users. In this case, one fundamental issue arises: how can the spectrum be dimensioned for control channels in order to minimize the access delay of MAC protocol in CR networks?

9.2 Problem Formulation

In this chapter, we focus on analyzing the CR MAC with dedicated and embedded control channels. For dedicated control channel case, it has control channel saturation and data channel saturation problems. For embedded control channel, it has channel mismatch problem.

In [164], the authors finds that minimizing the per packet delay may not minimize the total packets delay. From Fig. 9.1 and assume there are four packets in the network, one can see that the total packets delay is smaller than four times of the per packet delay. Thus, in this paper, we consider not only the per packet delay but also the total packets delay. The goal of this chapter is to evaluate the per packet and total packets delay by modeling the control and data channel saturation problems in dedicated case and channel mismatch problem in embedded case.

9.2.1 Saturation Problem with Dedicated Control Channel

In [165], the authors showed that splitting the total channel bandwidth leads to the **control channel saturation** issue, thereby degrading the throughput due to the overlong contention duration. Let T_{data} be the data transmission duration, T_c the sum of the contention duration and the control frames transmission duration. When $\frac{\sum_{i=1}^{n} T_{data}}{\sum_{i=1}^{n-1} T_c} < 1$, the control channel saturation issue occurs, as shown in Fig. 9.4.

One way to mitigate the control channel saturation is to increase the

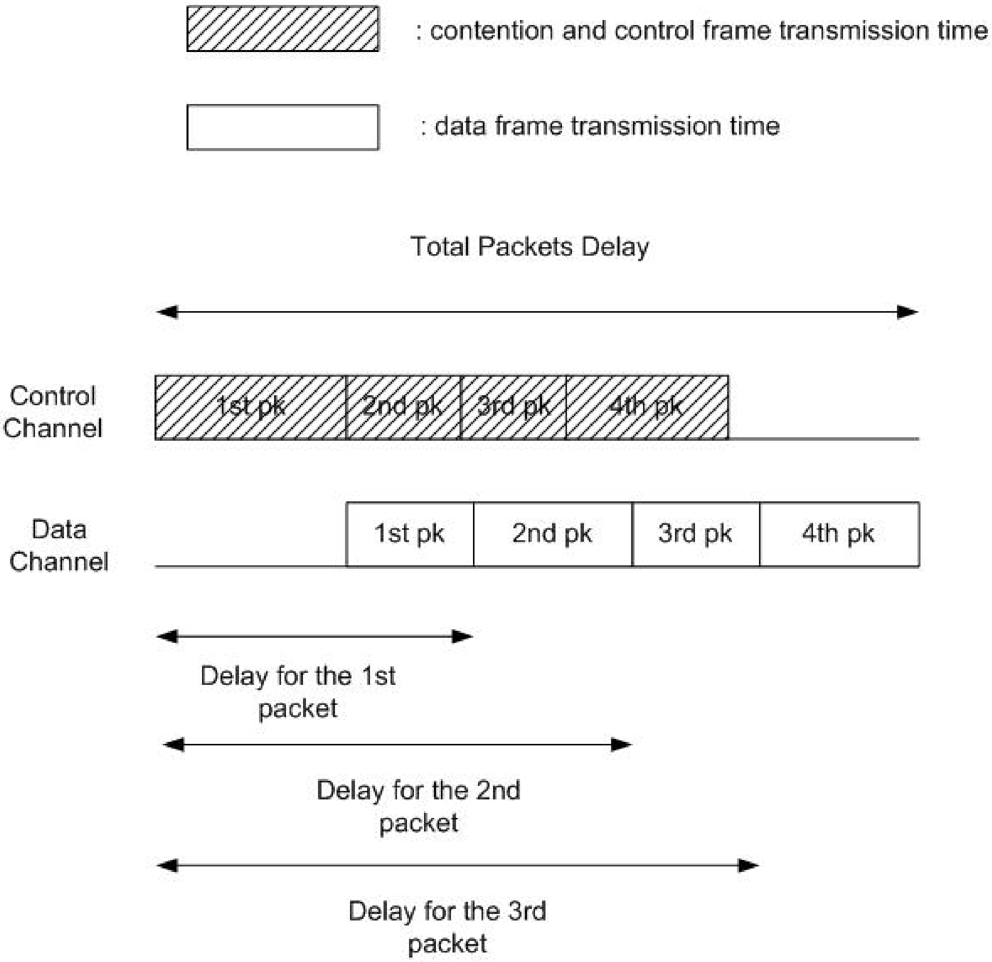


Figure 9.1: Per packet versus total packets delay

control channel bandwidth. If the control channel bandwidth increases too much, i.e., $\frac{\sum_{i=1}^{n} T_{data}}{\sum_{i=1}^{n-1} T_c} > 1$, the data channel saturation issue occurs as shown in Fig. 9.5. In practice, it is hard to adjust the optimal control channel bandwidth for with varying data lengths and contention durations.

9.2.2 Problems with Embedded Control Channel

Mismatched Problem

In Fig. 9.2, we illustrate the channel mismatched problem using the embedded control channel.

1. At T1 - SU 2 doesn't have data to transmit and its default channel CH 3 is busy. Thus, SU 2 switches to the idle channel. However, CH 1 and CH 2 are still busy until T1. If SU 1 has data to transmit at T1, SU 1 switches to CH 1 because it has the longest idle period. However, if SU 2 switches to CH 2, then SU 1 and SU 2 can't exchange data on the same spectrum. We call it the channel mismatched problem.

2. At T2 - SU 1 transmits RTS and waits for the CTS.

3. At T3 - After waiting for a period of RTS timeout and doesn't receive the CTS, transmitter guesses that SU 2 is not in CH 1. Then, SU 1 switches to the second longest idle period CH 2. Receiver SU 2 finds CH 2 is still idle. Thus, SU 2 stays on CH 2. Thus, SU 1 and SU 2 can transmit data on the same spectrum.

4. At T4 - SU 1 receives a CTS from SU 2. Thus, SU 1 knows that SU 2 is on CH 2. Then, SU 1 can transmit data to SU 2 on CH 2.

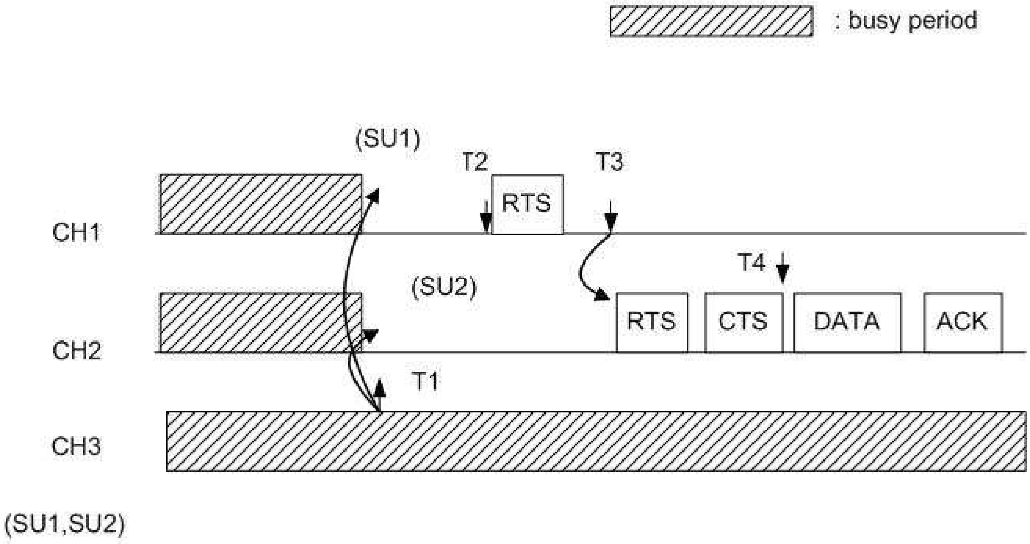


Figure 9.2: Channel mismatched case

Contention Problem

If more than one transmitter switch to the same candidate channel, and all of them send RTS at the same time, collision occurs. The solution to the contention problems is to make each transmitter choose a random backoff size for the transmission order. Collision are reduced due to random start time.

Go Back Timer Problem

Another problem for the CR system with embedded control channel is the go back timer problem: Specifically, how long should SU 2 stay on CH 2 to be a receiver? The solution to this problem is to leave the candidate channel CH 2 under two conditions: (1) when CH 2 becomes busy; (2) when the default channel CH 3 is idle (this timer can be known by observing the value of NAV).

Furthermore, how long SU 1 should stay on other candidate channels to be a transmitter is another issue. To solve this problem, SU 1 should go back to SU 2's default channel when CH 3 becomes idle (this timer can be known from NAV). After that, SU 1 should go back to its default channel when SU 1 transmits to SU 2 successfully.

9.2.3 Definition of Access Delay for CR MAC with Dedicated and Embedded Control Channels

Our goal is to evaluate and minimize the access delay for CR MAC with dedicated and embedded control channels as follows. As to dedicated control channel case, the delay $E[D]$ with considering the control channel and data channel saturation problems:

$$\begin{aligned} E[D] &= P_{cs}E[D|\text{the kth packet with control channel saturation}] \\ &+ P_{ds}E[D|\text{the kth packet with data channel saturation}] \ . \end{aligned} \tag{9.1}$$

where P_{cs} and P_{ds} are the saturation probabilities of the control and data channels, respectively.

As to embedded control channel case, the delay $E[D']$ with considering the channel mismatch problems:

$$E[D'] = E[T]E[N] \ , \tag{9.2}$$

9.2.4 System Model and Assumptions

Now, we introduce the system model that we use to evaluate the access delay. Figure 9.3 shows a system model for CR networks, where there are

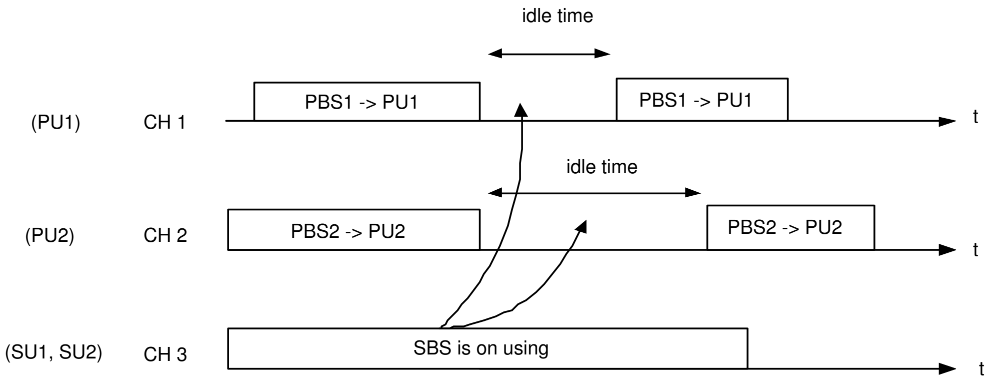


Figure 9.3: SU 1 and SU 2 have two candidate idle channels CH 1, CH 2 as shown by two arrows

primary users 1 and 2 (denoted by PU 1 and PU 2), and two secondary CR users (denoted by SU 1 and SU 2). Let channels (CH) 1 and 2 in the licensed band be assigned to PU 1 and PU 2, respectively, while SU 1 and SU 2 are covered by both the primary base station (PBS) and the secondary base station (SBS). Assume that each secondary user has two radio interfaces, where the one is used for receiving and the other one is used for transmitting. It is assumed that all the secondary users can know other's default channel through the channel discovery procedure and the primary system is TDM based system. Assume that the unlicensed band is already too crowded. Then the CR nodes SU 1 and SU 2 will check whether CH 1 and CH 2 are idle. If so, CH 1 and CH 2 can be lent to SU 1 and SU 2.

9.3 Latency Analysis

In this chapter, we derive an analytical model to calculate the per packet and total packets dynamic access latency for medium access control protocol with dedicated and embedded control channels in the cognitive radio networks.

The rest of this chapter are organized as follows. Section 3.1 analyzes the per packet dynamic access latency for cognitive radio medium access control protocol with dedicated and embedded control channels. In Section 3.2, we analyze the total packets dynamic access latency for cognitive radio medium access control protocol with dedicated and embedded control channels.

9.3.1 Per Packet Latency Analysis

In this section, we analyze the average delay of dynamic spectrum access for the CR MAC protocols with dedicated and embedded control channels as described in Chapter 2. The per packet access delay is defined as the duration starting from the instance that an RTS packet is at the head of the queue and starts to content for the transmission until the instance that the associated acknowledgment (ACK) packet for a data packet is received. The total packets access delay is defined as the total deliver time to send total number packets. It is assumed that each CR node always has packets to transmit. In the case of low traffic load, the dedicated case may be better than the embedded. However, in the case of full load, whether the dedicated control channel is still better than embedded control channel is unknown. Thus, we perform the worst case study to compare the dedicated control channel and the embedded control channels. It is assumed that the channel switching time is zero.

Case I : Dedicated Control Channel

Now we discuss how to calculate the average delay of dynamic spectrum access for the CR MAC protocols with dedicated control channel. Let P_{cs} and P_{ds} be the saturation probabilities of the control and data channels, respectively, both of which can be obtained from Fig. 9.9 for a given ratio of control channel bandwidth over data channel bandwidth. With P_{cs} and P_{ds}, the average access delay of the CR MAC protocol with dedicated control channel $E[D]$ can be expressed as

$$\begin{aligned} E[D] &= P_{cs}E[D|\text{the kth packet with control channel saturation}] \\ &+ P_{ds}E[D|\text{the kth packet with data channel saturation }] \ . \end{aligned} \tag{9.3}$$

Because the probability of the non-saturation event (i.e., $\frac{E[T_{data}]}{E[T_c]} = 1$) is very small, it is assumed that $P_{cs} + P_{ds} \simeq 1$ in (9.3).

(1) Delay With Control Channel Saturation: From Fig. 9.4 and the concept of virtual slots proposed in [166], we express the conditional average access delay given that the control channel saturation occurs as follows:

$$\begin{aligned} &E[D|\text{the kth packet with control channel saturation}] \\ &\qquad = E[T]E[N^{(c)}] + E[T_{data}] \ , \end{aligned} \tag{9.4}$$

where T in this case is the duration of a virtual slot when control channel saturation occurs; $N^{(c)}$ represents the number of virtual slots before the successful transmission of an RTS frame; and T_{data} is the transmission time of data packets. Note that (9.4) means a particular node sends out its data packets at kth order and previous k-1 packets may suffer from control or data channel saturation problems.

Before a specifical node transmits an RTS frame successfully, there exists three possible events in each virtual slot. The first one is an empty slot which has probability $1 - P_{tr}$, the second one is that a collided transmission with probability $(1 - P_s)P_{tr}$, and the last one is successful transmission with probability P_sP_{tr}. Hence, in terms of the durations of a successful transmission t_s, a collided transmission t_f, and an empty slot σ, it is followed that

$$E[T] = (1 - P_{tr})\sigma + (1 - P_s)P_{tr}t_f + P_sP_{tr}t_s \ , \tag{9.5}$$

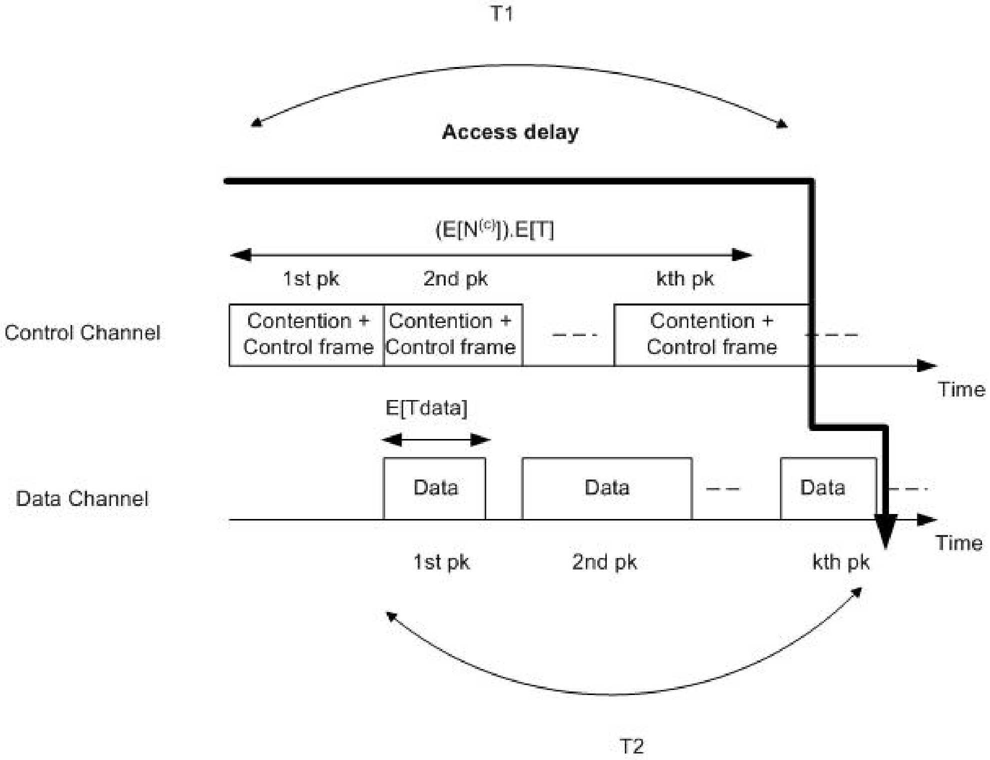


Figure 9.4: The access delay under the control channel saturation (T1≧T2)

where $P_s = \frac{n\tau(1-\tau)^{n-1}}{P_{tr}}$ is the probability of successful frame transmission; $P_{tr} = 1-(1-\tau)^n$ is the probability of at least one frame being transmitted, and n is the number of users. Let p and τ be the probability of collision and that of a station successfully transmitting a packet, respectively. In order to simplify, we consider the infinite retransmission and backoff stages. Then, following the steps in [167], we can get

$$p = 1-(1-\tau)^{n-1} \ , \tag{9.6}$$

and

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1)+pW} \ . \tag{9.7}$$

Then, (9.6) and (9.7) can be solved recursively for given n and W.

According to the IEEE 802.11 MAC protocol,

$$t_s = t_{RTS} + SIFS + t_{CTS} + SIFS \ , \tag{9.8}$$

and

$$t_f = t_{RTS} + EIFS \ , \tag{9.9}$$

respectively, where t_{RTS} and t_{CTS} represent the transmission duration of the RTS and the CTS frames; SIFS and EIFS are the short- and extended-interframe space.

Furthermore, given the minimum backoff window size W and the collision probability p, from [166] it can be shown that

$$\begin{aligned} E[N^{(c)}] &= (1-p)\frac{W+1}{2} + ... \\ &\quad +p^m(1-p)(\frac{W+1}{2} + ... + \frac{2^m W+1}{2}) + ... \\ &= \frac{1}{2}(\frac{W}{1-2p} + \frac{1}{1-p}) \ . \end{aligned} \tag{9.10}$$

Substituting (9.5)-(9.10) into (9.4) for a given value of $E[T_{data}]$, one can obtain the average access delay conditional on the occurrence of control channel saturation.

(2) Delay With Data Channel Saturation: Referring to Fig. 9.5, we discuss how to calculate the access delay with data channel saturation. In the figure, there are three possible events in the control channel: idle, successful transmission, and collision. Assume that a node waits for a duration of $(E[T]E[N^{(c)}]+t_s)$ before successfully transmitting a control frame, where T is the virtual slot duration; $N^{(c)}$ is the number of virtual slots; and t_s is defined in (9.8). Specifically, $(E[T]E[N^{(c)}] + t_s)$ is the duration starting from the beginning of channel contention until the first data frame transmission. Next, this particular user will need further wait for a duration of $(N^{(d)} - 1)E[T_{data}]$ before transmitting its own data frame, where $N^{(d)}$ is the average number of transmitted data frames of other users and this user and $E[T_{data}]$ is the average duration of a data frame. Thus, the average access delay with data channel saturation (denoted by $E[D|\text{data channel saturation occurs}]$) can be expressed as

$$\begin{aligned} &E[D|\text{the kth packet with data channel saturation}] \\ &\quad = \big(E[T]E[N^{(c)}] + t_s\big) + E[T_{data}]E[N^{(d)}] \ . \end{aligned} \tag{9.11}$$

Note that (9.11) means a particular node sends out its data packets at kth order and previous k-1 packets may suffer from control or data channel saturation problems.

In the case when no node successfully transmits a control frame, only two events possibly occur at each virtual slot. The first one is an empty slot with the probability of $\frac{1-P_{tr}}{1-P_sP_{tr}}$; the second one is a collided transmission with the probability of $\frac{(1-P_s)P_{tr}}{1-P_sP_{tr}}$. Hence, the average duration of a virtual slot T in

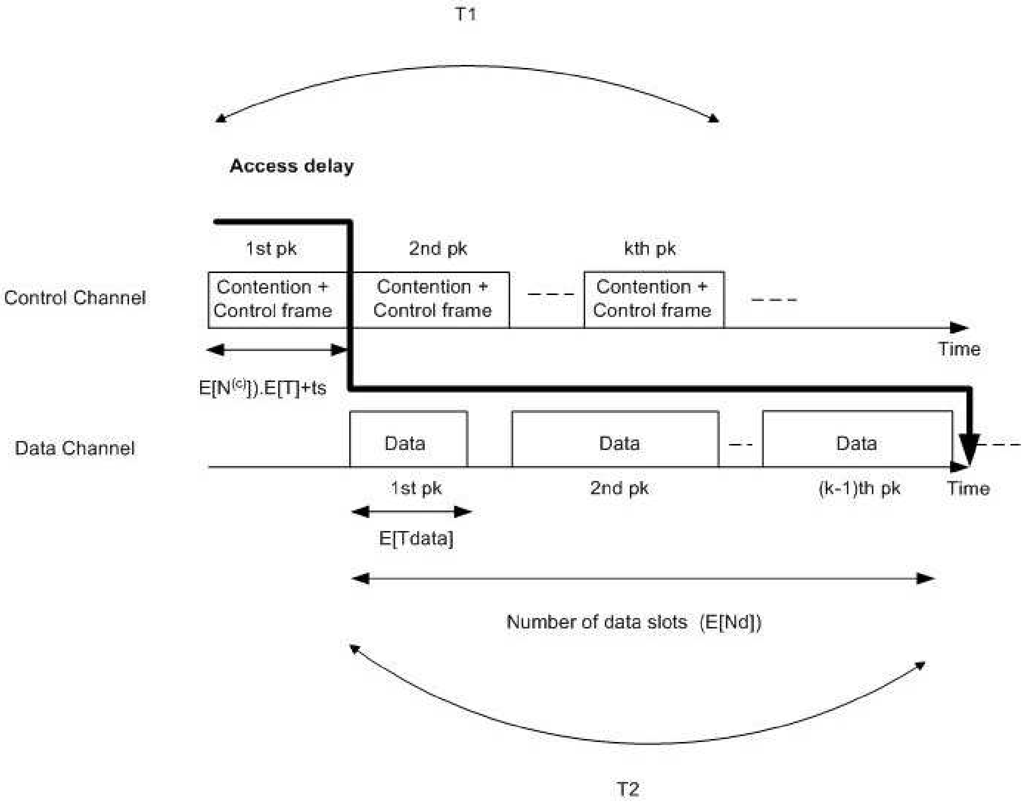


Figure 9.5: The access delay under the data channel saturation (T2≧T1)

this case can be written as

$$E[T] = \frac{1-P_{tr}}{1-P_sP_{tr}}\sigma + \frac{(1-P_s)P_{tr}}{1-P_sP_{tr}}t_f \ . \tag{9.12}$$

Moreover, given the probability without successful transmissions $1 - P_sP_{tr}$, the probability that the number of virtual slots $N^{(c)} = k + 1$ in the control channel until the first data frame transmission is equal to $(1 - P_sP_{tr})^kP_sP_{tr}$. Hence, it is followed that the average number of virtual slots $N^{(c)}$

$$\begin{aligned} E[N^{(c)}] &= \sum_{k=0}^{\infty}(k+1)(1-P_sP_{tr})^kP_sP_{tr} \\ &= \frac{1}{P_sP_{tr}} \ . \end{aligned} \tag{9.13}$$

To calculate $E[N^{(d)}]$, it is assumed that a particular user successfully sends a control frame at the first time with $k - 1$ previous transmitting data frames. The probability $P_s(k)$ that the first data frame of this user as the $k - th$ data frames in the data channel can be expressed as follows

$$P_s(k) = \left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{k-1} \ . \tag{9.14}$$

Then, the average total number $E[N^{(d)}]$ of data frames in the data channel including this user's frame is equal to

$$E[N^{(d)}] = \sum_{k=1}^{\infty} k\left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{k-1} \ . \tag{9.15}$$

Substituting (9.8), (9.12), (9.13) and (9.15) into (9.11) for a given value of $E[T_{data}]$, one can obtain the conditional average access delay given the occurrence of data channel saturation.

Case II : Embedded Control Channel

Now we consider the access delay D' for CR MAC with embedded control channel. When considering a CR MAC with embedded control channels, control channels and data channels use the same frequency spectrum and are not split into two portions in the frequency domain. Applying the virtual slot concept, one can express the average access delay $E[D']$ as

$$E[D'] = E[T]E[N] \ , \tag{9.16}$$

where T is the virtual slot duration, and N is the number of virtual slots before the RTS frame is transmitted successfully.

Consider the average virtual slot duration $E[T]$ first. On the one hand, when establishing a link by dynamic spectrum access, two CR users may choose two different channels in the case of embedded control channel. We call this situation the channel mismatch. In this case, after sending the RTS frame, the transmitter will wait for the CTS until the timeout expires. Let $t_{mismatch}$ be the sum of the time transmitting the RTS frame and the timeout expiration, i.e.,

$$t_{mismatch} = t_{RTS} + DIFS \ . \tag{9.17}$$

On the other hand, if the selected channels are matched, the total transmission time of the control frames and data frames (denoted by t_{match}) is equal to

$$\begin{aligned} t_{match} &= 2DIFS + t_{RTS} + t_{CTS} + 3SIFS \\ &\quad + E[T_{data}] + t_{ACK} \ , \end{aligned} \tag{9.18}$$

where t_{ACK} is the transmission time of the ACK frame; DIFS is the DCF-interframe space, respectively according to the definitions in the IEEE 802.11 WLAN MAC protocol. Denote P_{match} and $P_{mismatch}$ as the probability of the channel match and mismatch, respectively. With m candidate channels, it is followed that

$$P_{match} = \frac{1}{m} \ , \tag{9.19}$$

and

$$P_{mismatch} = \frac{m-1}{m} \ . \tag{9.20}$$

Recall that the probability with an empty slot is $1 - P_{tr}$, and the probability with collided transmission is $(1 - P_s)P_{tr}$. When considering all the four possible events at each virtual slot, the average virtual time slot duration $E[T]$ can be expressed as

$$\begin{aligned} E[T] &= (1 - P_{tr})\sigma + (1 - P_s)P_{tr}t_f \\ &+ P_{match}P_sP_{tr}t_{match} \\ &+ P_{mismatch}P_sP_{tr}t_{misatch} \ , \end{aligned} \tag{9.21}$$

where $T_{mismatch}$, T_{match}, P_{match}, and $P_{mismatch}$ are given by (9.17), (9.18), (9.19), and (9.20), respectively.

Now, we calculate $E[N]$ in (9.27). Consider m candidate channels. With previous $k-1$ mismatched events, the probability of the two CR users select the same channel in the $k-th$ attempt is equal to

$$\frac{m-1}{m} \cdot \frac{m-2}{m-1} \cdots \frac{m-k+1}{m-k+2} \cdot \frac{1}{m-k+1} = \frac{1}{m} \ . \tag{9.22}$$

Before a CR node transmits RTS, it needs to wait for a contention duration. The number of virtual slots before the successful transmission of an RTS frame can be denoted by $N^{(c)}$ as shown in (9.10). With previous $k-1$ mismatched events, this node will undergo $kE[N^{(c)}]$ virtual slots on average. The probability of the occurrence of the first matched event in k events is $1/m$ according to (9.22). Thus, $E[N]$ can be expressed as

$$E[N] = \sum_{k=1}^{m} \frac{1}{m} \cdot kE[N^{(c)}] = \frac{m+1}{2}E[N^{(c)}] \ . \tag{9.23}$$

Substituting (9.21) and (9.23) into (9.27) for a given value of $E[T_{data}]$, one can obtain the average access delay for embedded control channel .

9.3.2 Total Packets Latency Analysis

Case I : Dedicated Control Channel

Now we discuss how to calculate the average whole packets delay of dynamic spectrum access for the CR MAC protocols with dedicated control channel. The similar analytical principle as the per packets delay. With P_{cs} and P_{ds}, the average whole packets access delay of the CR MAC protocol with dedicated control channel $E[D_{overall}]$ can be expressed as

$$\begin{aligned} E[D_{overall}] &= P_{cs}E[D_{overall}|\text{the last packet with control channel saturation}] \\ &+ P_{ds}E[D_{overall}|\text{the last packet with data channel saturation}] \ . \end{aligned} \tag{9.24}$$

<u>**(1) Total Packets Delay With Control Channel Saturation:**</u> Because we now calculate the whole packets delay under the control channel saturation. Thus, we calculate the delay until the last packet has been sent. From Fig. 9.6, we express the conditional whole packets access delay given that the control channel saturation occurs as follows:

$$\begin{aligned} &E[D_{overall}|\text{the last packet with control channel saturation}] \\ &\qquad = c(E[T]E[N^{(c)}] + t_s) + E[T_{data}] \ , \end{aligned} \tag{9.25}$$

where c is the whole number of packets in the network and $(E[T]E[N^{(c)}]+t_s)$ means the average delay to send a packet and can be obtained from (9.12), (9.13), (9.8) and (9.9).

Substituting (9.8), (9.9), (9.12) and (9.13) into (9.26) for a given value of $E[T_{data}]$ and c, one can obtain the conditional whole packets delay given the event that control channel saturation occurs.

<u>**(2) Total Packets Delay With Data Channel Saturation:**</u> Now, we discuss how to calculate the whole packets delay with data channel satura-

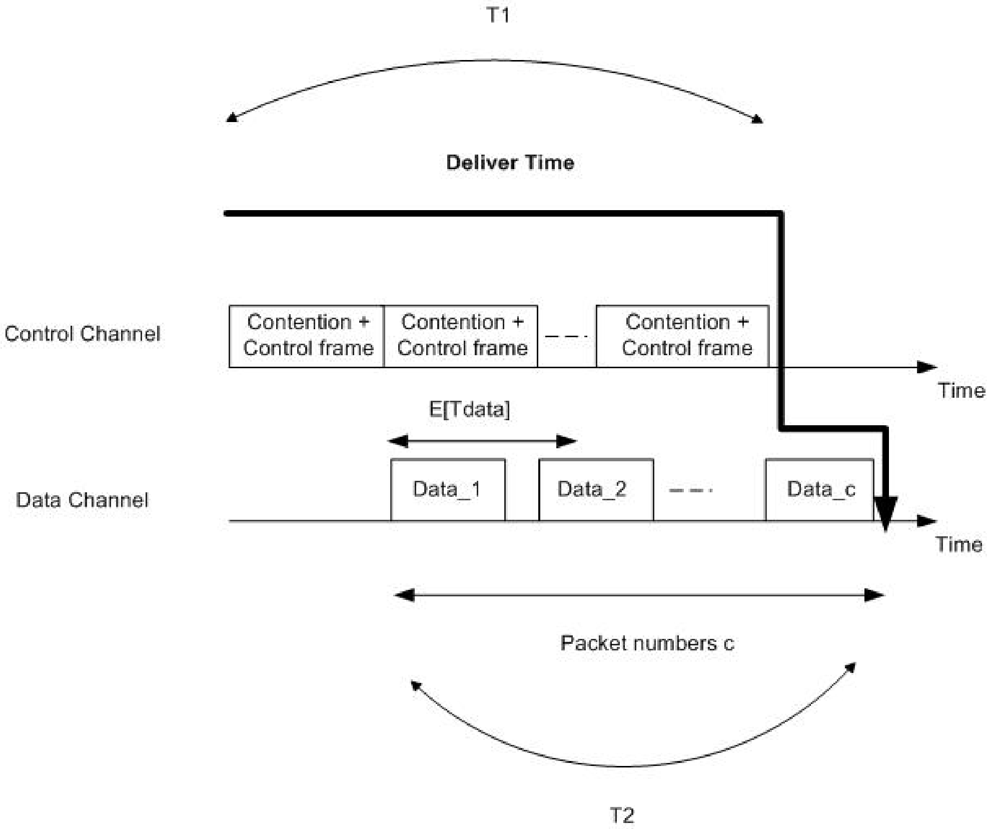


Figure 9.6: The whole packets access delay under the control channel saturation (T1≧T2)

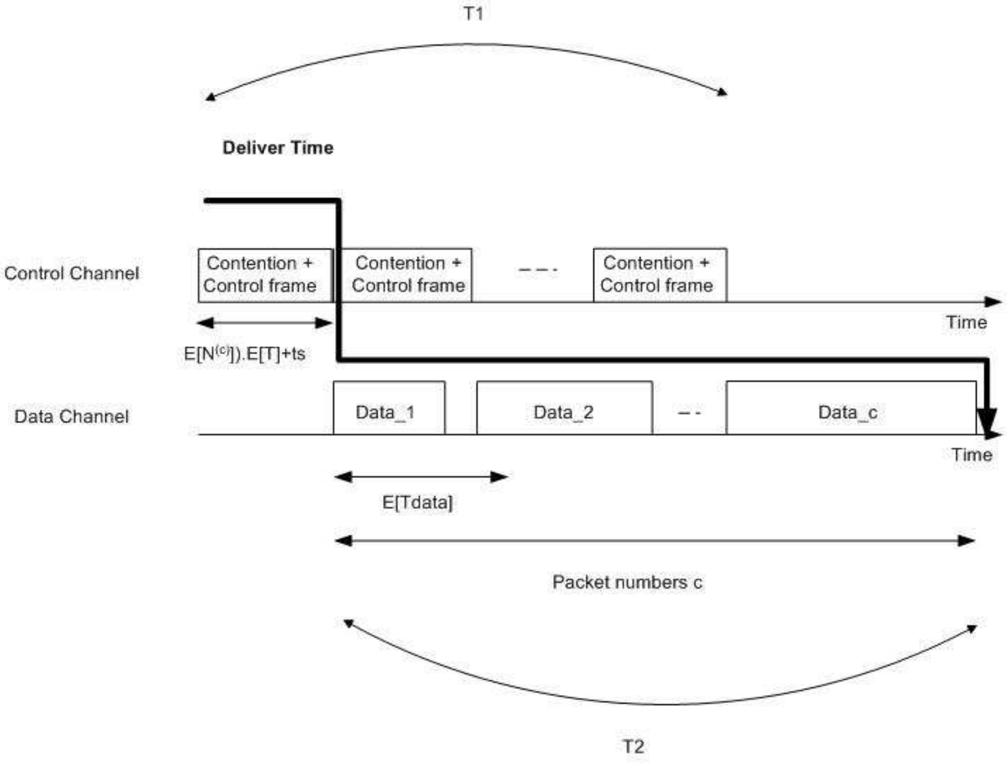


Figure 9.7: The whole packets access delay under the data channel saturation (T2≧T1)

tion. We consider the whole packets delay of whole c packets as shown in Fig. 9.7. Firstly, the system wait a duration of the first user's control frame transmission time equals to $(E[T]E[N^{(c)}]t_s)$. Because the data channel saturation occurs, we will calculate the latency resulted from c data frames rather than that resulted from control frames. Because there are c data frames need to be transmitted, we can express the conditional whole packets access delay given that the data channel saturation occurs as follows:

$$\begin{aligned} &E[D_{overall}|\text{the last packet with data channel saturation}] \\ &\qquad = cE[T_{data}] + (E[T]E[N^{(c)}]t_s) \ , \end{aligned} \tag{9.26}$$

where c is the whole number of packets in the network and $(E[T]E[N^{(c)}]+t_s)$ can be obtained from (9.12), (9.13), (9.8) and (9.9).

Substituting (9.8), (9.9), (9.12) and (9.13) into (9.26) for a given value of $E[T_{data}]$ and c, one can obtain the conditional whole packets delay given the event that data channel saturation occurs.

Case II : Embedded Control Channel

Now we consider the whole packets access delay $D'_{overall}$ for CR MAC with embedded control channel. When considering a CR MAC with embedded control channels, control channels and data channels use the same frequency spectrum and are not split into two portions in the frequency domain. Thus, from the average viewpoint, the whole packets can be seen that distribute uniformly on the whole number of channels (we denote this number as m). Then, we calculate the average time duration to send a complete control frames and a data frame that equal to $E[T']E[N^{(c)}]t'_s$. From Fig. 9.8, one can express the average whole packets access delay for CR MAC with embedded

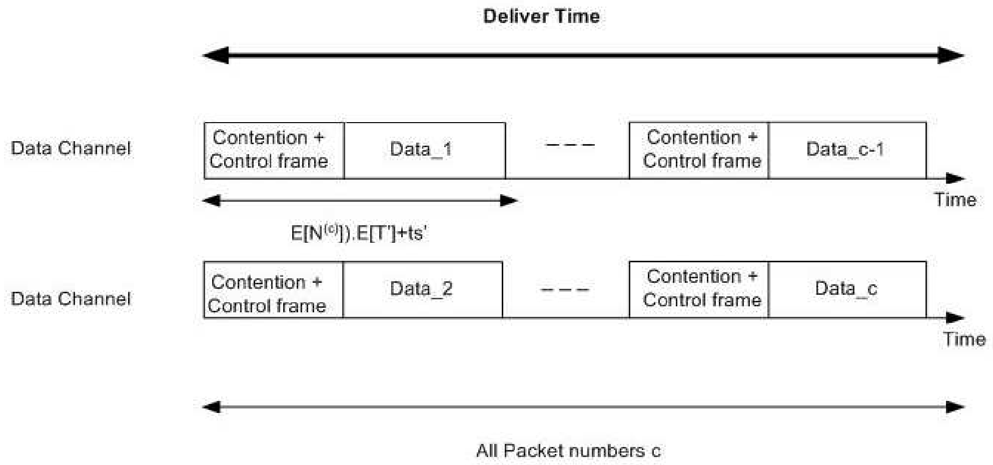


Figure 9.8: The whole packets access delay for CR MAC with embedded control channel

control channels $E[D'_{overall}]$ as

$$E[D'_{overall}] = \frac{c}{m}(E[T']E[N^{(c)}]t'_s)m = c(E[T']E[N^{(c)}]t'_s) \ , \tag{9.27}$$

where T' is the virtual slot duration, t'_s is the total transmission time of the control frames and a data frame and $E[N^{(c)}]$ can be obtained from (9.13).

Moreover, base on the similar analysis principle of (9.12), we can express the average duration of a virtual slot T' on the data channel as follows:

$$E[T'] = \frac{1-P_{tr}}{1-P_sP_{tr}}\sigma + \frac{(1-P_s)P_{tr}}{1-P_sP_{tr}}t'_f \ . \tag{9.28}$$

where t'_f is the total duration when collision happens.

According to the IEEE 802.11 MAC protocol, we have

$$t'_s = 2DIFS + t_{RTS} + 3SIFS + t_{CTS} + E[T_{data}] + t_{ACK} \ , \tag{9.29}$$

and

$$t'_f = DIFS + t_{RTS} + 2SIFS + t_{CTS} \ . \tag{9.30}$$

Table 9.1: System parameters for the control and the data channel

RTS	160 bits
CTS	112 bits
ACK	112 bits
Slot time, σ	20 μsec
SIFS	10 μsec
DIFS	50 μsec
EIFS	60 μsec

Substituting (9.29), (9.30), (9.28) and (9.13) into (9.27) for a given value of whole packet numbers c, one can obtain the conditional whole packets delay for CR MAC with embedded control channel.

9.4 Numerical Results

In this section, we first show the relation between the delay of dedicated control channel and the control channel bandwidth ratio. Then we compare the per packet and total packets access delay performance of the dedicated control channel and embedded control channel for various user numbers and data sizes. As to the analytical result of the total packets delay, we use 2000 packets to simulate the total packets delay. The related system parameters are listed in Table 9.1. The basic transmission rate is used in each control channel. The data rate is used in each channel for embedded case. In our simulation, it is assumed that data length and the contention period are exponential distribution [168].

9.4.1 Control Channel Saturation Probability

To calculate the average access delay of the CR MAC with dedicated control channels, we need to obtain P_{cs} and P_{ds}. Based on the assumptions that the data length and contention period are exponentially distributed, we simulate the control channel saturation probability in Fig. 9.9. Given average data length and contention duration, some samples are randomly generated. Then we count the samples satisfying the criterion $\frac{\sum_{i=1}^{n} T_{data}}{\sum_{i=1}^{n-1} T_c} < 1$ to obtain P_{cs}. Note that $P_{ds} \simeq 1 - P_{cs}$. Figure 9.9 plots the control channel saturation probability versus the control channel bandwidth ratio. In this figure, one can see that the control channel saturation probability decreases as the the control channel bandwidth ratio increases. The value of control channel saturation probability can be used to calculate the access delay for various control channel bandwidths ratio, which will be shown in the next figure.

9.4.2 Per Packet Access Latency

Effect of Control Channel Bandwidth Ratio on the Access Delay for Dedicated Control Channels

Figure 9.10 shows the effect of control channel bandwidth ratio on the access delay of the CR MAC with dedicated control channels. In the figure, one can see the existence of an optimal control channel bandwidth ratio with the minimum access delay. In the cases with 15 and 20 users, the access delay is minimum when the ratio of the control channel bandwidth over the total bandwidth is about 0.2. Referring to the data channel saturation as show in Fig. 9.5, one can see that a higher control channel bandwidth ratio results in longer access delay because of less bandwidth left for transmitting data frames. Noteworthily, the optimal control bandwidth ratios are different for

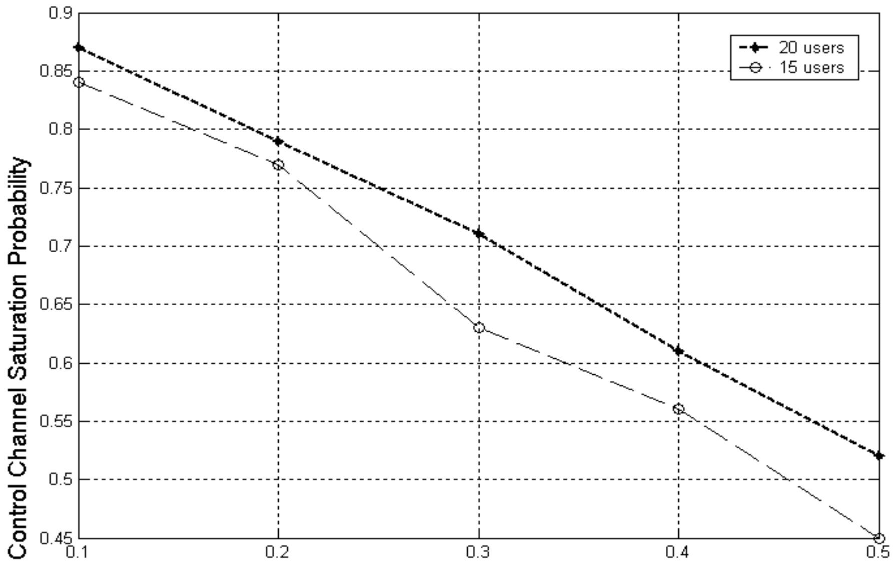


Figure 9.9: The effect of control channel bandwidth ratio on the control channel saturation probability

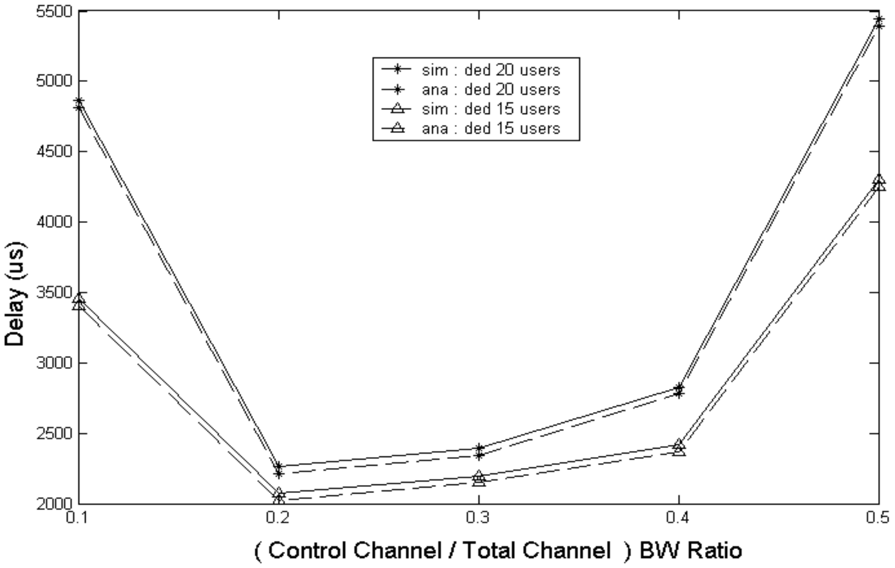


Figure 9.10: The effect of control channel bandwidth ratio on the access delay of the CR MAC for dedicated control channels

various data sizes and contention periods. From this viewpoint, it is difficult to design a predefined optimal bandwidth ratio for a CR MAC with dedicated control channels.

Effect of Number of Users on Access Latency

Figure 9.11 shows the effect of number of users on the CR MAC with dedicated and embedded control channels for the control bandwidth ratio of 0.1 and 0.2. When the ratio is 0.1, the delay for the CR MAC with embedded control channels is always lower than that for the CR MAC with dedicated control channel. For the number of users is 20, the access delay with embedded control channels is 2.5 msec, but the access delay with dedicated control channels is about 4.6 msec. When the optimal ratio 0.2 is considered, the delay for CR MAC with embedded control channel is lower than that for CR

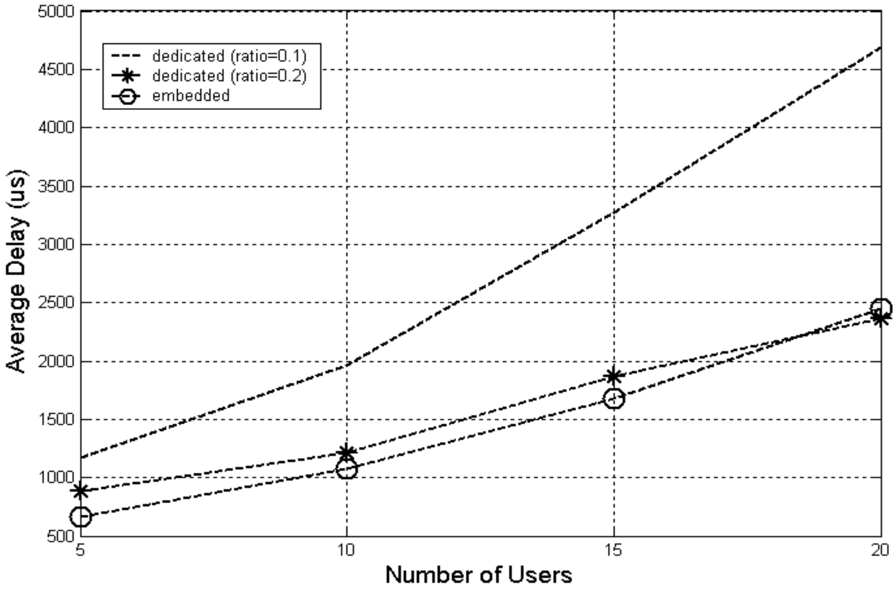


Figure 9.11: The effect of the different user numbers on the access delay with dedicated and embedded control channels

MAC with dedicated control channel only for the case of 17 users. Nevertheless, the optimal control bandwidth ratios for the dedicated control channels varies for different data sizes and contention periods. Thus, it may be hard to trace the optimal bandwidth ratio dynamically.

Effect of Number of Users and Data Lengths on Confidence Interval Comparison of Access Latency

Because the confidence interval can jointly consider the mean and standard deviation, we use confidence interval to compare these two control channel methods.

First, Figure 9.12 shows the confidence interval (90 percent confidence) of the per packet access latency for CR MAC with dedicated and embedded control channels under small packet length (300 bytes). It shows that when the number of users is not very large (17,18 users), the per packet access delay for the CR MAC with embedded control channels is usually shorter than that the CR MAC with dedicated control channels. This is because that the data transmission rate of embedded case is higher than that of dedicated case. Thus, when the user numbers is small, the embedded case can get smaller delay with higher data rate. But, when the user numbers is large, embedded may suffer from higher contention delay than dedicated case.

Then, Figure 9.13 shows the confidence interval (90 percent confidence) of the per packet access latency for CR MAC with dedicated and embedded control channels under large packet length (2000 bytes). It shows that the per packet access delay of the CR MAC with dedicated control channel is shorter than the CR MAC with embedded control when user numbers is large (over 24 users) and the average data length is large (2000 bytes). The reason is the same as the previous small data length case.

Access Latency Comparison for Three Data Channels Case

The previous results of per packet delay is considered under two data channels environments. Because the typical IEEE 802.11 MAC protocol has three data channels, we now discuss the access delay by extending to three primary data channels.

First, Figure 9.14 shows the per packet access latency for CR MAC with dedicated and embedded control channels under small packet length (300 bytes). It shows that the per packet access delay for embedded case is usually shorter than dedicated case when the number of users is not very large (15 users).

Then, Figure 9.15 shows the per packet access latency for CR MAC with

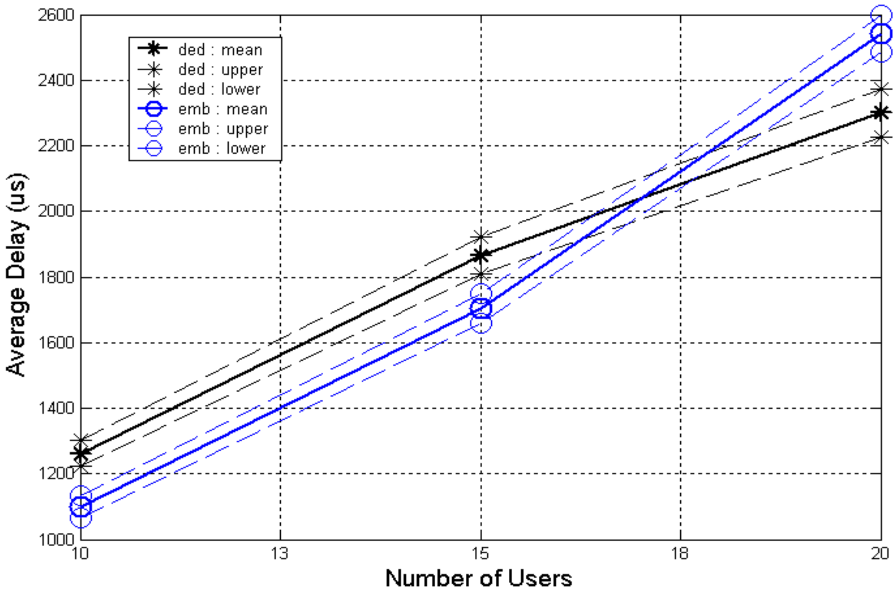


Figure 9.12: Confidence interval of access latency for CR MAC with dedicated and embedded control channels under small data length

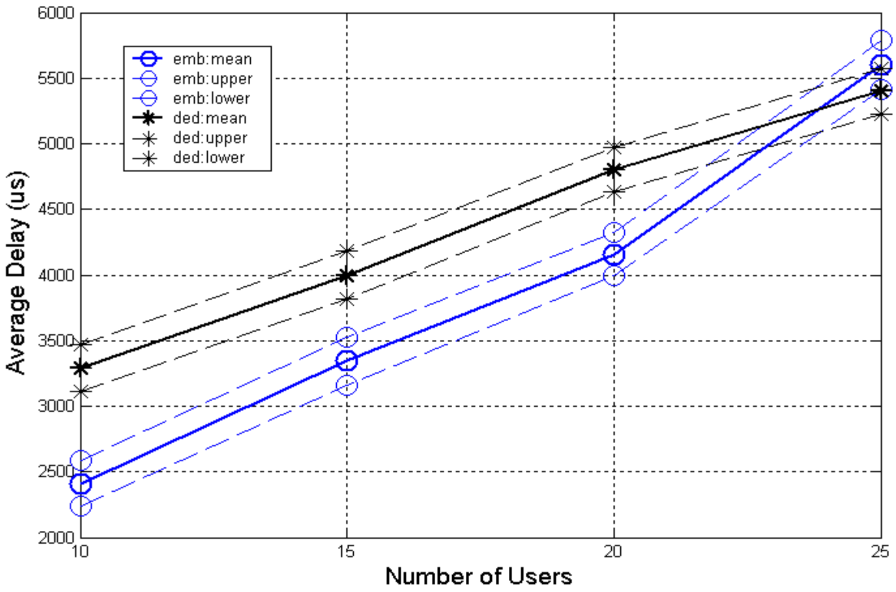


Figure 9.13: Confidence interval of access latency for CR MAC with dedicated and embedded control channels under large data length

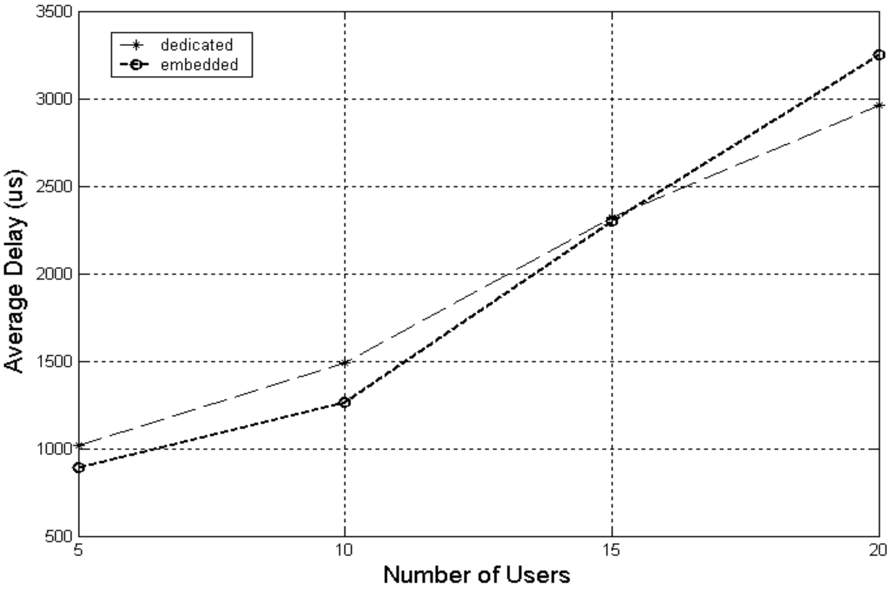


Figure 9.14: Access latency for CR MAC with dedicated and embedded control channels under small data length

dedicated and embedded control channels under large packet length (2000 bytes). It shows that the per packet access delay of the CR MAC with dedicated control channel is shorter than the CR MAC with embedded control when user numbers is large (23 users).

9.4.3 Total Packets Access Latency

Effect of Number of Users on Access Latency

Figure 9.16 shows the effect of different number of users on the total packets delay of the CR MAC with dedicated and embedded control channels. This figure shows that the total packets latency of dedicated case is higher than that of embedded case when the user numbers is not large(under 18 users).

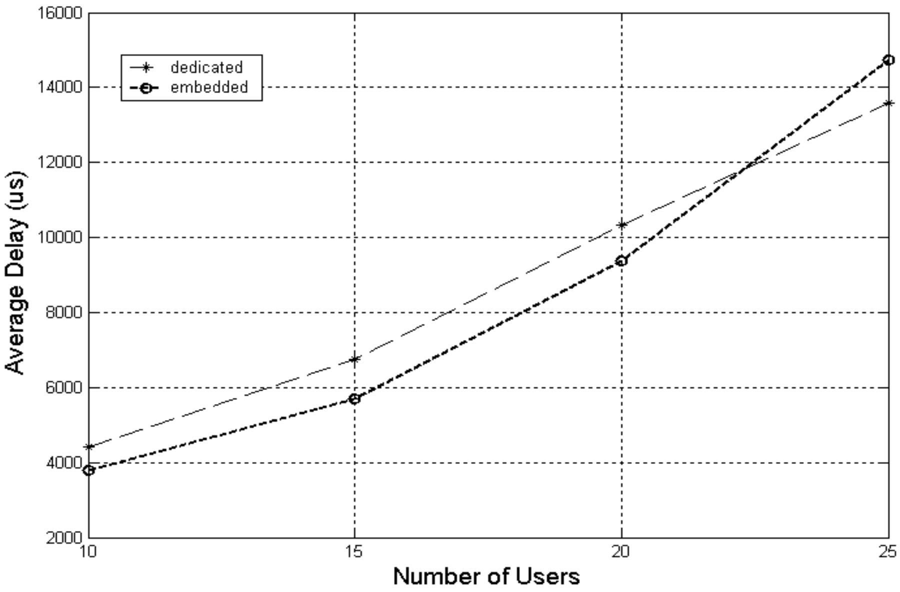


Figure 9.15: Access latency for CR MAC with dedicated and embedded control channels under large data length

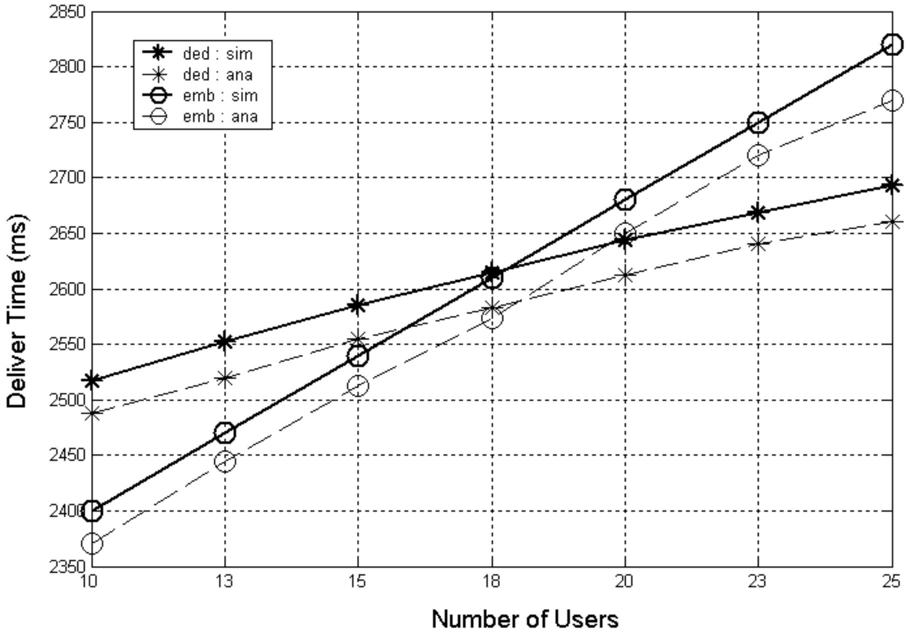


Figure 9.16: The effect of number of users on the overall delay of CR MAC with dedicated and embedded cases

Effect of Number of Users and Data Lengths on Confidence Interval Comparison of Access Latency

First, Fig. 9.17 shows the confidence interval (90 percent confidence) of the total packets access latency for CR MAC with dedicated and embedded control channels under small data lengths (300 bytes). It shows that the total packets access delay of the CR MAC with dedicated control channel is longer than the CR MAC with embedded control when the average data length is small (300 bytes). Because the embedded case has higher data rate and nodes transmit the small data length, the data transmission time is no longer a bottle neck. Thus, the dedicated will have higher access delay because the dedicated case has lower data rate.

Then, Figure 9.18 shows the confidence interval (90 percent confidence) of the total packets access latency for CR MAC with dedicated and embedded control channels under large data length (2000 bytes). It shows that when the number of users is not large (under 24 users), the total packets access delay for the CR MAC with embedded control channels is shorter than the CR MAC with dedicated control channels. The reason is that when the data length is large, the data channel of dedicated case is usually fully used. Thus, when the user numbers is large, the contention delay will become a bottle neck to embedded case and decrease the system throughput even though the embedded case has higher data rate.

Access Latency Comparison for Three Data Channels Case

The previous results of total packets delay is considered under two data channels environments. Because the typical IEEE 802.11 MAC protocol has three data channels, we now discuss the access delay by extending to three primary data channels.

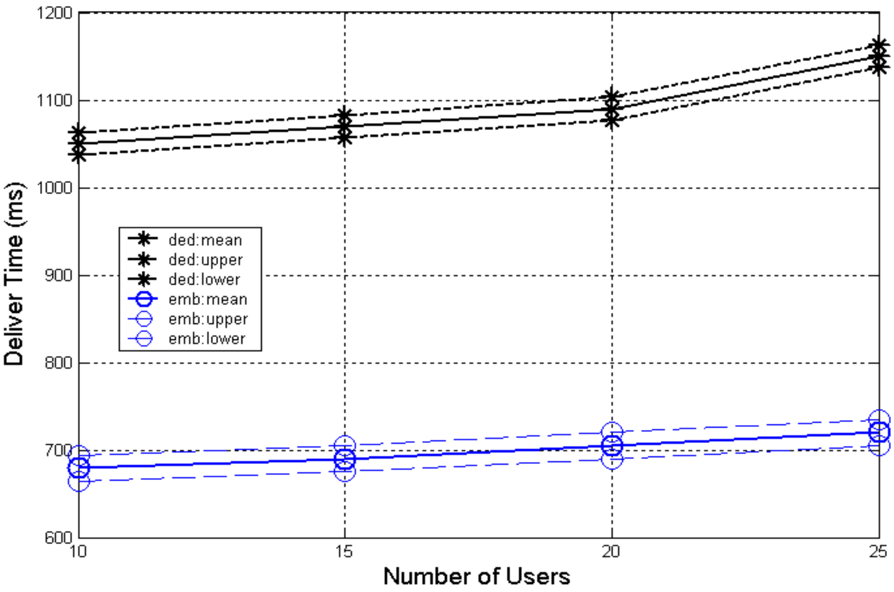


Figure 9.17: Confidence interval of access latency for CR MAC with dedicated and embedded control channels under small data length

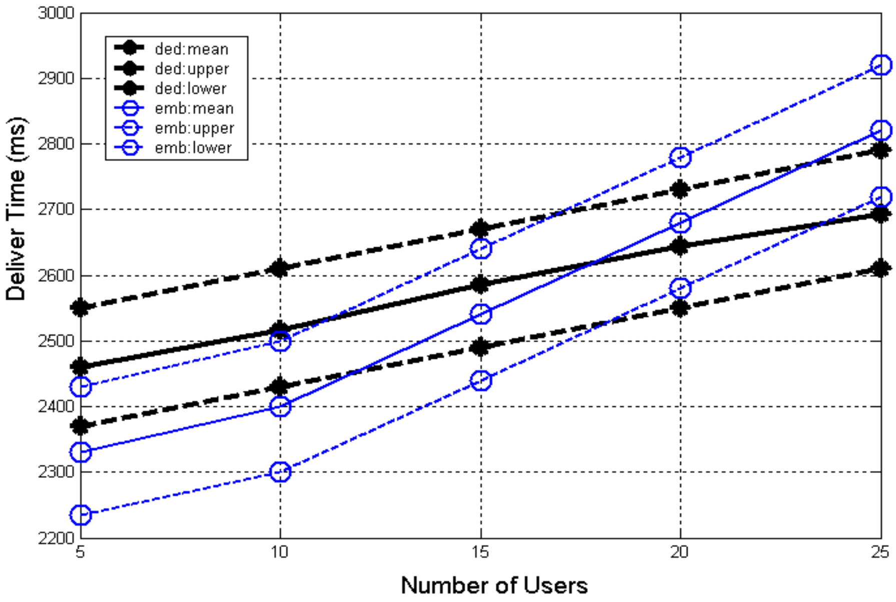


Figure 9.18: Confidence interval of access latency for CR MAC with dedicated and embedded control channels under large data length

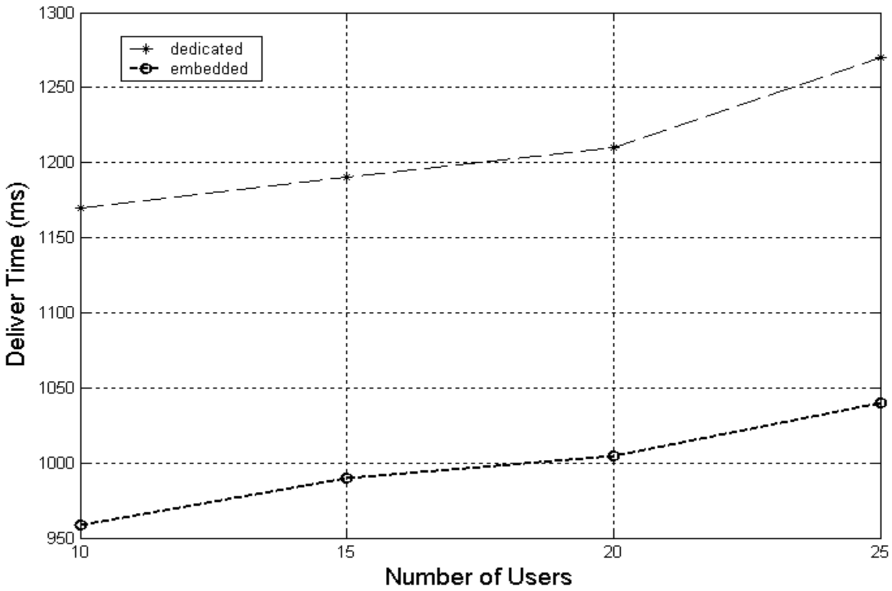


Figure 9.19: Total packets access latency for CR MAC with dedicated and embedded control channels under small data length

First, Figure 9.19 shows the total packets access latency for CR MAC with dedicated and embedded control channels under small packet length (300 bytes). It shows that the total packets access delay of the dedicated case is longer than the embedded case when the average data length is small.

Then, Figure 9.20 shows the total packets access latency for CR MAC with dedicated and embedded control channels under large packet length (2000 bytes). It shows that the total packet access delay of the CR MAC with dedicated control channel is shorter than the CR MAC with embedded control when user numbers is large (over 17 users).

Hence, from the standpoint of access latency, we find that the legacy IEEE 802.11 CSMA/CA MAC protocol is worth extending to CR MAC by adding the channel search capability under some system characteristics.

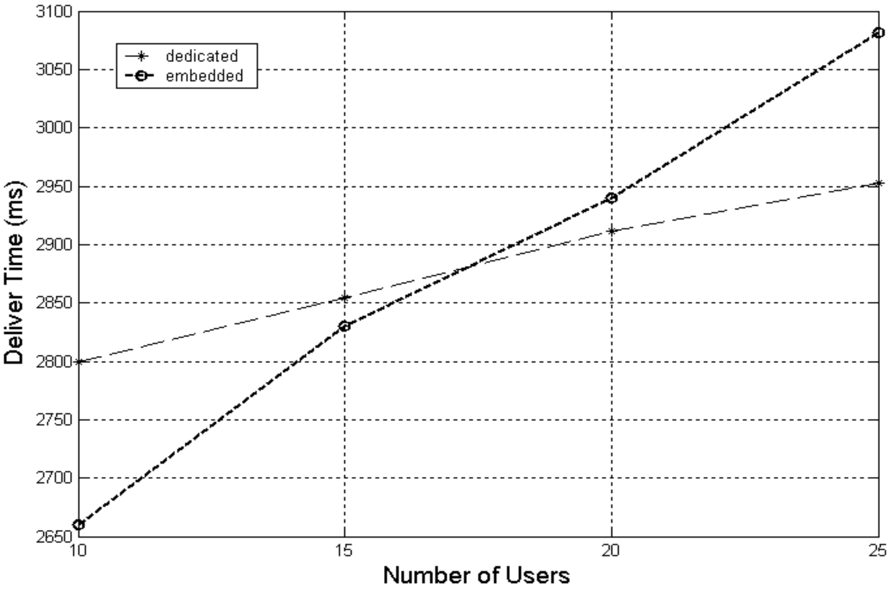


Figure 9.20: Total packets access latency for CR MAC with dedicated and embedded control channels under large data length

We can conclude that DSA protocol in CR networks can dedicate a certain portion of spectrum for sending control frames if user numbers is large (i.e., 25 users or higher than 1.13 times of the typical 15-user case); otherwise, the embedded control channel shall be considered. Furthermore, from total packets delay viewpoint, when the data length is small (i.e., 300 bytes or lower than 0.29 times of the typical 1024-byte case), the embedded control channel method is suggested to be used even though the user numbers is large.

Chapter 10

Location-Aware Concurrent Transmission for Spectrum Sharing

Our main contribution of this chapter is to provide the idea of utilizing location awareness to facilitate frequency sharing in a concurrent transmission manner. Specific achievements are summarized in the following.

- We show that a CR device having location information of other nodes can concurrently transmit a peer-to-peer data in the presence of an infrastructure-based connection in some region. We also dimension the concurrent transmission (or the scanning-free) region for CR users. Note that a concurrent transmission region of a CR system is equivalent to a scanning-free region. Nevertheless, the wide-band spectrum sensing procedure is still needed but is initiated only when the CR user is outside the concurrent transmission region. Therefore, the energy consumption of CR systems with location awareness capability can be reduced significantly.

- Based on the CSMA/CA MAC protocol, a physical/MAC cross-layer analytical model is developed to compute the coexistence probability of a peer-to-peer connection and an infrastructure-based connection. Based on this analytical model, we find that concurrent transmission of the secondary CR users and the primary users in the legacy system can significantly enhance the total throughput over the pure legacy system.

10.1 Motivation

In the literature, the coexistence issue of the hybrid infrastructure-based and overlaying ad hoc networks has been addressed but in different scenarios. In [169–172], the idea of combining ad hoc link and infrastructure-based link was proposed mainly to extend the coverage area of the infrastructure-based network. That is, the coverage area of ad hoc networks is not overlapped with that of the infrastructure-based network. In the present hybrid ad hoc/infrastructure-based network, as shown in Fig. 10.1, the peer-to-peer CR users are located within the coverage area of the existing legacy wireless network. In [173], to further improve the throughput of a wireless local area network (WLAN), it was suggested that an access point (AP) could dynamically switch between the infrastructure mode and the ad hoc mode. In our considered scenario, the decision of establishing ad hoc connections is made by the CR users in a distributed manner.

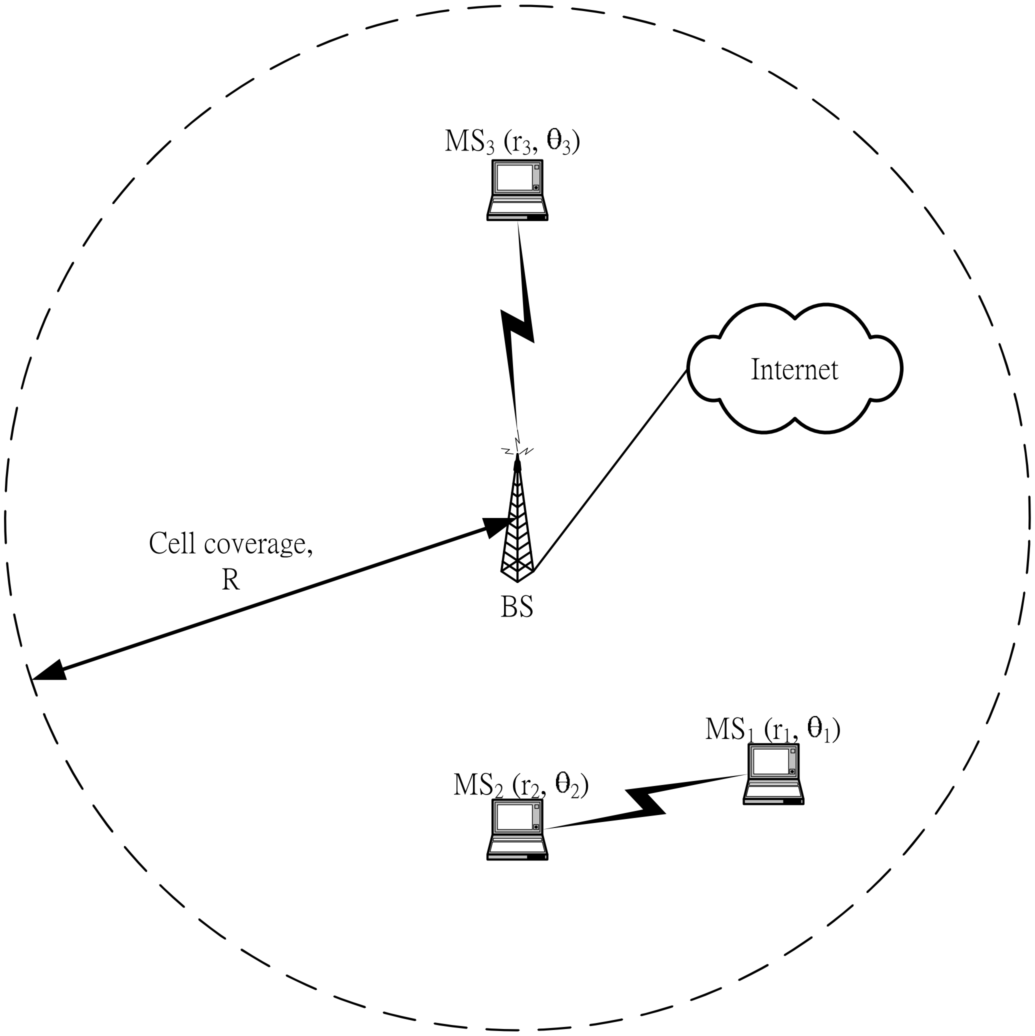

MS$_3$ (r_3, θ_3)
Internet
Cell coverage,
R
BS
MS$_1$ (r_1, θ_1)
MS$_2$ (r_2, θ_2)

Figure 10.1: An illustrative example for the coexistence of two CR devices establishing an overlaying cognitive ad hoc link and a primary user connecting to the infrastructure-based network, where all the devices (MS$_1$, MS$_2$, and MS$_3$) use the same spectrum simultaneously.

10.2 System Model

In this section, we define the generic system model discussed in this report. Figure 10.1 illustrates a hybrid ad hoc/infrastructure-based network consisting of two CR devices (MS_1 and MS_2) and a primary user MS_3. Assume that the secondary CR users MS_1 and MS_2 try to make a peer-to-peer connection, and the primary user MS_3 has been connected to the base station (BS) or access point (AP) of the legacy infrastructure-based system. In the figure, MS_1, MS_2 and MS_3 are located at (r_1, θ_1), (r_2, θ_2) and (r_3, θ_3), respectively; the coverage area of the base station is πR^2. All the primary and secondary users stay fixed or hardly move.

We assume the CR devices can perform the positioning technique to acquire their relative or absolute position by using GPS or detecting the signal strength from the BSs of legacy systems [174–179]. The location information is broadcasted by using the geographical routing protocols [180–182]. Although both the positioning and geographical routing may waste time and consume the energy, they have no need to be processed for every data transmission. They are only performed when a new node joins or the node changes its position. Furthermore, with the help of upper layer, the location information is already stored in the device. Therefore, compared to the spectrum sensing at every transmission, we believe the additional energy consumption and memory space due to the positioning and location update is relatively small. The overhead and optimal reserved resources for acquiring the location information are beyond the scope of this paper and have been studied in some research works [183, 184].

Based on the CSMA/CA MAC protocol, multiple users contend the channel, and only one mobile station within the coverage of the base station can establish an infrastructure-based communication link at any instant. To set up an extra peer-to-peer ad hoc connection in the same frequency band of the primary user, two secondary CR users have to ensure that the current infrastructure-based link quality cannot be degraded. Here, we consider that both primary and secondary users have identical transmit power. It is reasonable to assume only one secondary user can be established a link after the contention at one instance due to the similar interference range. Denote SIR_i and SIR_a as the received signal-to-interference ratios (SIRs) of the infrastructure-based and ad hoc links, respectively. Then we can define the coexistence (or concurrent transmission) probability (P_{CT}) of the infrastructure-based link and CR-based ad hoc link in an overlapped area as follows:

$$P_{CT} = P\{(SIR_i > z_i) \cap (SIR_a > z_a)\}, \tag{10.1}$$

where z_i and z_a are the required SIR thresholds for the infrastructure-based and ad hoc links, respectively. To obtain the concurrent transmission region, it is crucial to calculate the coexistence probability of both the infrastructure and ad hoc links. If the link quality of the primary user cannot be guaranteed, CR devices have to sense and change to other frequency bands.

We consider the following propagation model [185]:

$$P_r = \frac{P_t h_{bs}^2 h_{ms}^2 G_{bs} G_{ms} 10^{\frac{\xi}{10}}}{r^{\alpha}} \ , \tag{10.2}$$

where P_r and P_t are the received and transmitted power levels at a mobile station, respectively; h_{bs} and h_{ms} represent the antenna heights of the base station and the mobile station, respectively; G_{bs} and G_{ms} stand for the antenna gains of the base station and the mobile station, respectively; r is the distance between the transmitter and receiver; α is the path loss exponent; $10^{\frac{\xi}{10}}$ is the log-normally distributed shadowing component.

10.3 Signal-to-Interference Ratio Analysis

10.3.1 Uplink SIR Analysis

In the *uplink case* when the primary user MS$_3$ transmits data to the base station, denote $SIR_i^{(u)}$ as the uplink SIR of MS$_3$ and let P_{30} and P_{10} be the received power from MS$_3$ and MS$_1$ at the base station, respectively. Then from (10.2), we have

$$SIR_i^{(u)} = (\frac{r_1}{r_3})^\alpha = \frac{P_{30}}{P_{10}} \ , \tag{10.3}$$

where r_1 and r_3 are the distances between MS$_1$ and MS$_3$ to the base station, respectively. Similarly, the SIR of a peer-to-peer ad hoc link from MS$_1$ to MS$_2$ can be written as

$$SIR_a = \frac{P_{12}}{P_{32}} = (\frac{d_{23}}{d_{12}})^\alpha \ , \tag{10.4}$$

where P_{12} is the received power at MS$_2$ from MS$_1$ and P_{32} is the interference power from MS$_3$; d_{12} and d_{23} are the distances from MS$_1$ and MS$_3$ to MS$_2$, respectively. Substituting (10.3) and (10.4) into (10.1), the concurrent transmission probability $P_{CT}^{(u)}$ in the *uplink case* can be written as

$$P_{CT}^{(u)} = P\{(r_3 z_i^{1/\alpha} < r_1 < R) \cap (d_{12} < \frac{d_{23}}{{z_a}^{1/\alpha}})\} \triangleq \frac{R_{CT}^{(u)}}{\pi R^2} \ . \tag{10.5}$$

Note that $R_{CT}^{(u)}$ denotes the concurrent transmission region where MS$_1$ can connect to MS$_2$ without interfering the uplink signal of MS$_3$ to the base station. As shown in Fig. 10.2, the condition $(r_3 z_i^{1/\alpha} < r_1 < R)$ leads to a donut-shaped area consisting of two circles centered at the base station with the radii of $r_3 z_i^{1/\alpha}$ and R, respectively. Meanwhile, the condition $(d_{12} < d_{23}/z_a^{1/\alpha})$ yields the circular area centering at MS$_2$ with a radius of $d_{23}/z_a^{1/\alpha}$. From the figure, the region $R_{CT}^{(u)}$ can be computed as

$$R_{CT}^{(u)} = \pi(\frac{d_{23}}{z_a^{1/\alpha}})^2 - A_1 - A_2 \ , \tag{10.6}$$

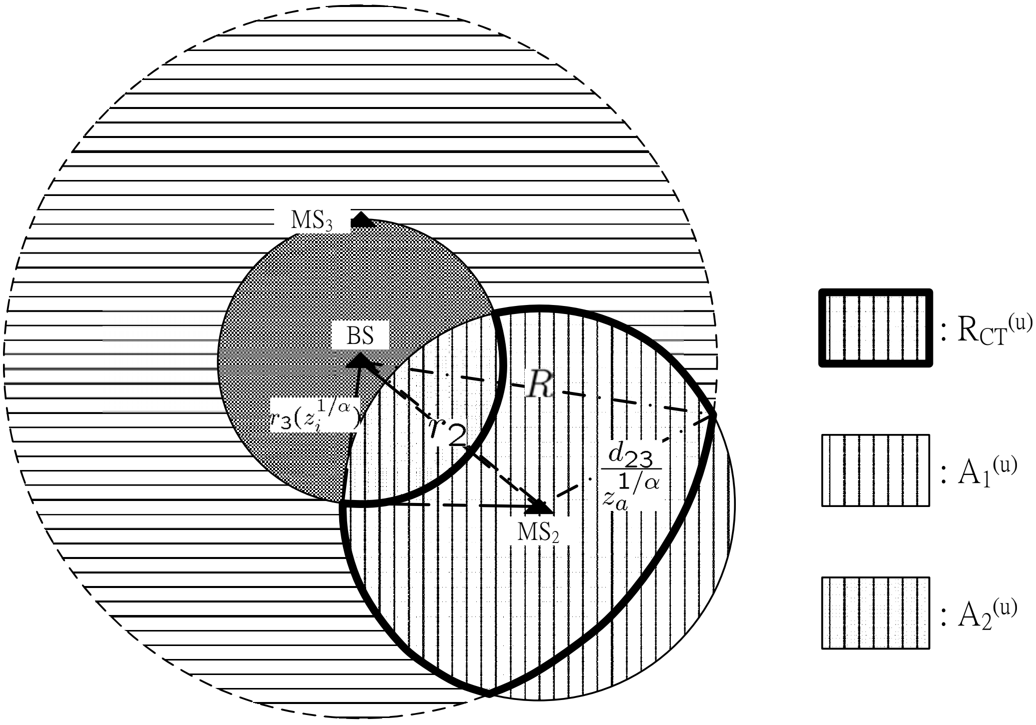


Figure 10.2: Physical representation of the coexistence probability for the concurrent transmission of overlaying CR-based ad hoc link and infrastructure uplink transmission.

where

$$A_1 = (\frac{d_{23}}{z_a^{1/\alpha}})^2(\pi - \theta') - R^2\theta + 2\Delta \tag{10.7}$$

and

$$A_2 = (\frac{d_{23}}{z_a^{1/\alpha}})^2\phi - (r_3 z_i^{1/\alpha})^2\phi' - 2\Delta' \ . \tag{10.8}$$

The definitions of parameters θ, θ', ϕ, ϕ', Δ, and Δ' and the detailed derivation of (10.6), (10.7) and (10.8) are discussed in Appendix E.

10.3.2 Downlink SIR Analysis:

Now we consider the *downlink case* when the base station sends data to the primary user MS$_3$. Denote $SIR_i^{(d)}$ as the infrastructure link's SIR in the downlink direction. Then from (10.2), we have

$$SIR_i^{(d)} = \frac{P_{03}}{P_{13}} = (\frac{h_{bs}}{h_{ms}})^2(\frac{d_{13}}{r_3})^\alpha \ , \tag{10.9}$$

where P_{03} and P_{13} are the MS$_3$'s received powers from the base station and MS$_1$, respectively; d_{13} stands for the distance from MS$_1$ to MS$_3$; h_{bs}, h_{ms} and r_3 are given in (10.2) and (10.3). Similarly, the ad hoc link's SIR from MS$_1$ to MS$_2$ can be expressed as

$$SIR_a = \frac{P_{12}}{P_{02}} = (\frac{h_{ms}}{h_{bs}})^2(\frac{r_2}{d_{12}})^\alpha \ , \tag{10.10}$$

where P_{12} and P_{02} are the received powers at MS$_2$ from MS$_1$ and the base station, respectively; r_2 represents the distance between MS$_2$ and the base station; d_{12}, h_{bs} and h_{ms} are defined in (10.4) and (10.2).

Substituting (10.9) and (10.10) into (10.1), we can obtain the concurrent transmission probability $P_{CT}^{(d)}$ of a CR-based peer-to-peer ad hoc link and an infrastructure-based downlink transmission as follows:

$$P_{CT}^{(d)} = P\{(d_{13} > r_3 z_i'^{1/\alpha}) \cap (d_{12} < r_2 z_a'^{1/\alpha}) \cap (r_1 < R)\} \triangleq \frac{R_{CT}^{(d)}}{\pi R^2} \ , \tag{10.11}$$

where $z_i' = z_i h_{ms}^2 / h_{bs}^2$ and $z_a' = 1/z_a \cdot h_{ms}^2/h_{bs}^2$. From (10.11), the concurrent transmission region $R_{CT}^{(d)}$ in the *downlink case* is shown in Fig. 10.3. The criterion $(d_{13} > r_3 z_i'^{1/\alpha})$ results in the region outside the circle centered at MS$_3$ with a radius of $r_3 z_i'^{1/\alpha}$, while the criterion $(d_{12} < r_2 z_a'^{1/\alpha})$ yields the region inside the circle centered at MS$_2$ with radius $r_2 z_a'^{1/\alpha}$. At last, $r_1 < R$ because MS$_1$ is assumed to be uniformly distributed within a cell of radius R.

The coexistence probability of the CR-based ad hoc link and the infrastructure-based downlink can be obtained by calculating the area of $R_{CT}^{(d)}$. The distances from the AP to the intersections of the two circles with radii of $r_3 z_i'^{1/\alpha}$

and $r_2 z_a'^{1/\alpha}$ are denoted by r^+ and r^- as shown in Fig. 10.3. In Appendix F, we have shown that

$$\begin{aligned} r^+ &= \frac{1}{d_{23}^2}\{r_2 r_3[r_2 r_3(z_a'^{\frac{2}{\alpha}} + z_i'^{\frac{2}{\alpha}}) + \cos(\theta_2 - \theta_3)(d_{23}^2 - r_2^2 z_a'^{\frac{2}{\alpha}} - r_3^2 z_i'^{\frac{2}{\alpha}})] + \\ &\quad \sin(\theta_2 - \theta_3)\delta\} \end{aligned} \tag{10.12}$$

and

$$\begin{aligned} r^- &= \frac{1}{d_{23}^2}\{r_2 r_3[r_2 r_3(z_a'^{\frac{2}{\alpha}} + z_i'^{\frac{2}{\alpha}}) + \cos(\theta_2 - \theta_3)(d_{23}^2 - r_2^2 z_a'^{\frac{2}{\alpha}} - r_3^2 z_i'^{\frac{2}{\alpha}})] - \\ &\quad \sin(\theta_2 - \theta_3)\delta\} \ , \end{aligned} \tag{10.13}$$

where

$$\delta = \sqrt{2 r_3^2 z_i'^{\frac{2}{\alpha}}(d_{23}^2 + r_2^2 z_a'^{\frac{2}{\alpha}}) - (d_{23}^2 - r_2^2 z_a'^{\frac{2}{\alpha}})^2 - r_3^4 z_i'^{\frac{4}{\alpha}}} \ . \tag{10.14}$$

With the values of r^+ and r^-, $R_{CT}^{(d)}$ can be calculated in the following two cases:

1. $max(r^+, r^-) \le R$: In this case, referring to Fig. 10.3(a), the area of $R_{CT}^{(d)}$ can be expressed as

$$R_{CT}^{(d)} = \pi(d_{23} z_a'^{1/\alpha})^2 - A_1 - A_2 \ , \tag{10.15}$$

where

$$A_1 = (r_2 z_a'^{1/\alpha})^2(\pi - \theta') - R^2\theta + 2\Delta \ ; \tag{10.16}$$

$$A_2 = (r_2 z_a'^{1/\alpha})^2\phi - (r_3 z_i'^{1/\alpha})^2\phi' - 2\Delta' \ . \tag{10.17}$$

2. $max(r^+, r^-) > R$: As shown in Fig. 10.3(b), the area of $R_{CT}^{(d)}$ can be expressed as

$$R_{CT}^{(d)} = \pi(d_{23} z_a'^{1/\alpha})^2 - A_1 - A_2 + A_3 \ , \tag{10.18}$$

where

$$A_1 = (r_2 z_a'^{1/\alpha})^2(\pi - \theta') - R^2\theta + 2\Delta \ ; \tag{10.19}$$

$$A_2 = (r_2 z_a'^{1/\alpha})^2 \phi - (r_3 z_i'^{1/\alpha})^2 \phi' - 2\Delta' \; ; \tag{10.20}$$

$$\begin{aligned} A_3 \;=\; & \Delta'' + [(r_3 z_i'^{1/\alpha})^2 \psi_2 - \frac{1}{2}(r_3 z_i'^{1/\alpha})^2 \sin\psi_2] + \\ & [(r_2 z_a'^{1/\alpha})^2 \psi_3 - \frac{1}{2}(r_2 z_a'^{1/\alpha})^2 \sin\psi_3] - \\ & [R^2 \psi_1 - \frac{1}{2} R^2 \sin\psi_1] \; . \end{aligned} \tag{10.21}$$

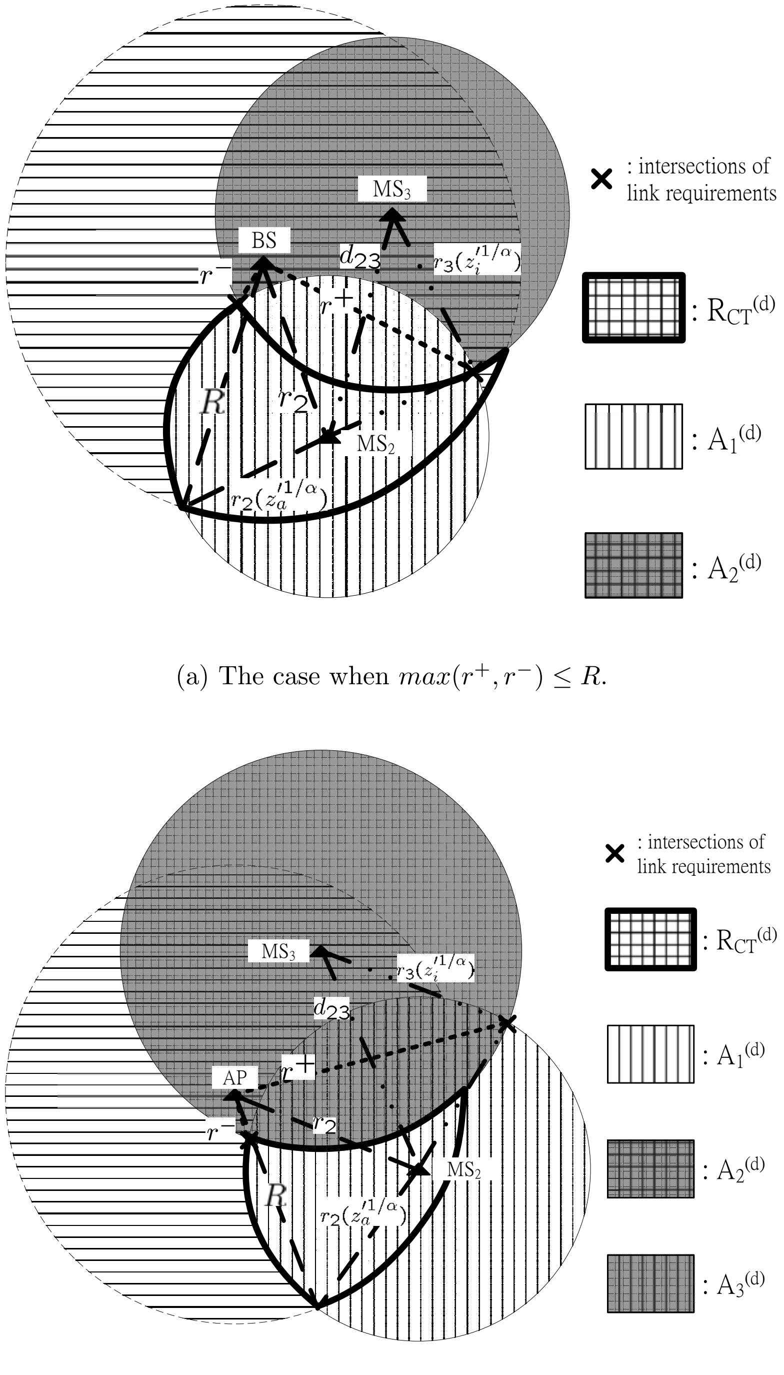
The detailed derivations of (10.15) and (10.18) and the definitions of the parameters θ, θ', ϕ, ϕ', ψ_1, ψ_2, ψ_3, Δ, Δ', and Δ'' are given in Appendices G and H, respectively.

10.3.3 Multiple Ad Hoc Connections Coexisting with One Infrastructure Link

After evaluating the concurrent transmission probability of the infrastructure link and the one overlaying CR-based ad hoc link, one may be interested in knowing how many secondary users can concurrently establish ad hoc links together with the primary user. This question is non-trivial since it needs to consider the interference from a set of ad hoc links to the infrastructure link, and vice versa. Besides, different from the pure infrastructure network, both the locations of the transmitter and receiver in an ad hoc link are random.

Instead of calculating the maximum number of ad hoc links, we suggest constructive procedures enabling CR devices to establish ad hoc links in the presence of an infrastructure transmission. The detailed procedures are summarized as follows:

1. Consider a network in which all the primary and secondary users are fixed, and the CR device can learn the locations of its receiver and neighbors by the routing mechanism [15]. Here, we assume that l ad



(a) The case when $max(r^+, r^-) \leq R$.

(b) The case when $max(r^+, r^-) > R$.

Figure 10.3: The area of concurrent transmission region R_{CT} in downlink cases.

hoc links have been established and coexisted with the infrastructure link at the same time. Before establishing a new ad hoc connection, the CR device has to overhear the channel and memorizes the locations of all the existing transmitters.

2. With the location information, the new CR device starts evaluating the concurrent transmission region R_{CT}. The device should consider the interference from the infrastructure link as well as other existing ad hoc links, and vice versa. Denote the indices $\{p, m, n, k\}$ as the primary user, the transmitter and receiver of the new ad hoc link, and the transmitter of other existing ad hoc link, respectively. Using similar procedures in deriving (10.5), the three conditions in the infrastructure uplink case can be written by

$$r_m \geq \left(\frac{1}{\frac{1}{z_i}(\frac{1}{r_p})^\alpha - \sum_k (\frac{1}{r_k})^\alpha}\right)^{\frac{1}{\alpha}} \; ; \tag{10.22}$$

$$d_{mn} \leq \left(\frac{1}{z_a((\frac{1}{d_{pn}})^\alpha + \sum_k (\frac{1}{d_{kn}})^\alpha)}\right)^{\frac{1}{\alpha}} \; ; \tag{10.23}$$

$$r_m \leq R \; , \tag{10.24}$$

where r_i and d_{ij} are the distances between the base station and CR device j to i, respectively. Similarly, from (10.11), the three criteria in the downlink case are

$$d_{mp} \geq \left(\frac{1}{\frac{1}{z_i}(\frac{1}{r_p})^\alpha - \sum_k (\frac{1}{d_{kp}})^\alpha}\right)^{\frac{1}{\alpha}} \; ; \tag{10.25}$$

$$d_{mj} \leq \left(\frac{1}{z_a((\frac{1}{r_n})^\alpha + \sum_k (\frac{1}{d_{kn}})^\alpha)}\right)^{\frac{1}{\alpha}} \; ; \tag{10.26}$$

$$r_m \leq R \; . \tag{10.27}$$

3. Since the concurrent transmission regions $R_{CT}^{(u)}$ and $R_{CT}^{(d)}$ are known, the CR device can determine whether it can concurrently transmit data

together with the infrastructure link and other ad hoc connections by the primary user and other secondary CR users.

10.4 Shadowing Effects

In the previous section, we only consider the impact of path loss on the concurrent transmission probability of CR-based network overlaying the infrastructure-based system. However, even though the CR device is located inside the concurrent transmission region R_{CT}, a peer-to-peer ad hoc connection may not be able to coexist together with the primary user's infrastructure link due to shadowing. Thus, it is important to investigate the reliability of concurrent transmissions of the hybrid infrastructure and CR-base ad hoc network when shadowing is taken into account.

Shadowing can be modeled by a log-normally distributed random variable [186]. Represent $10^{\frac{\xi_{ij}}{10}}$ as the shadowing component in the propagation path from users i to j, where ξ_{ij} is a Gaussian random variable with zero mean and standard deviation of σ_ξ. Thus, the uplink and downlink SIRs in both the infrastructure-based connection and CR-based ad hoc link are modified as:

- *uplink case:*

$$SIR_i^{(u)}(\xi_{30}, \xi_{10}) = \frac{10^{\frac{\xi_{30}}{10}}/r_3^\alpha}{10^{\frac{\xi_{10}}{10}}/r_1^\alpha} \;; \tag{10.28}$$

$$SIR_a^{(u)}(\xi_{12}, \xi_{32}) = \frac{10^{\frac{\xi_{12}}{10}}/d_{12}^\alpha}{10^{\frac{\xi_{32}}{10}}/d_{23}^\alpha} \;; \tag{10.29}$$

- *downlink case:*

$$SIR_i^{(d)}(\xi_{03}, \xi_{13}) = \frac{10^{\frac{\xi_{03}}{10}}/r_3^\alpha}{10^{\frac{\xi_{13}}{10}}/d_{13}^\alpha} \;; \tag{10.30}$$

$$SIR_a^{(d)}(\xi_{12},\xi_{02}) = \frac{10^{\frac{\xi_{12}}{10}}/d_{12}^{\alpha}}{10^{\frac{\xi_{02}}{10}}/r_2^{\alpha}} \ . \tag{10.31}$$

Note that the index 0 represents the base station and ξ_{30} of (10.29) in the uplink case is equivalent to ξ_{03} of (10.31) in the downlink case. Let $\boldsymbol{\xi} = (\xi_{10},\xi_{30},\xi_{12},\xi_{32})$ and $\boldsymbol{\xi'} = (\xi_{13},\xi_{03},\xi_{12},\xi_{02})$ and assume that these shadowing components are identical and independently distributed. Taking shadowing into account, the concurrent transmission probability P_{CT} can be represented by

- *uplink case:*

$$P_{CT}^{(u)}(\boldsymbol{\xi}) = P\{(r_3(z_i 10^{\frac{\xi_{10}-\xi_{30}}{10}})^{1/\alpha} < r_1 < R) \cap (d_{12} < d_{23}(z_a 10^{\frac{\xi_{12}-\xi_{32}}{10}})^{1/\alpha})\} \ ; \tag{10.32}$$

- *downlink case:*

$$\begin{aligned} P_{CT}^{(d)}(\boldsymbol{\xi'}) \ &= \ P\{(d_{13} > r_3(z_i 10^{\frac{\xi_{13}-\xi_{03}}{10}})^{1/\alpha}) \cap (d_{12} < r_2(z_a 10^{\frac{\xi_{12}-\xi_{02}}{10}})^{1/\alpha}) \\ &\quad \cap (r_1 < R)\} \ . \end{aligned} \tag{10.33}$$

We define the reliability of uplink concurrent transmission $F_{CT}^{(u)}(\xi)$ as the probability that, in the region R_{CT}, a CR device can successfully establish an ad hoc link in the presence of the primary user's uplink transmission subject to the shadowing effect. That is,

$$F_{CT}^{(u)}(\boldsymbol{\xi}) = P\{(SIR_i^{(u}(\xi_{30},\xi_{10}) > z_i) \cap (SIR_a^{(u)}(\xi_{12},\xi_{32}) > z) | MS_1 \in R_{CT}^{(u)}\} \tag{10.34}$$

Note that $F_{CT}^{(u)}(\boldsymbol{\xi}) = 1$ when shadowing is not considered. Substituting (10.28) and (10.29) into (10.34), we can obtain

$$\begin{aligned} F_{CT}^{(u)}(\boldsymbol{\xi}) \ &= \ P\{(\xi_{30}-\xi_{10} > 10\log_{10}(z_i(\frac{r_3}{r_1})^{\alpha})) \cap (\xi_{12}-\xi_{32} > 10\log_{10}(z_a(\frac{d_{12}}{d_{23}})^{\alpha})) \\ &\quad |MS_1 \in R_{CT}^{(u)}\} \ . \end{aligned} \tag{10.35}$$

Assume that ξ_{ij} have the same standard deviation for all i and j and let $\xi_i^{(u)} = \xi_{30} - \xi_{10}$, $\xi_a^{(u)} = \xi_{12} - \xi_{32}$. Then, $\xi_i^{(u)}$ and $\xi_a^{(u)}$ becomes a Gaussian random variable with $\mathsf{N}(0, 2\sigma_\xi)$. Hence, it is followed that

$$\begin{aligned} F_{CT}^{(u)}(\boldsymbol{\xi}) &= P\{\xi_i^{(u)} \geq 10\log_{10}(z_i(\frac{r_3}{r_1})^\alpha) | MS_1 \in R_{CT}^{(u)}\} \cdot \\ &\quad P\{\xi_a^{(u)} \geq 10\log_{10}(z_a(\frac{d_{12}}{d_{23}})^\alpha) | MS_1 \in R_{CT}^{(u)}\} \\ &= Q(\frac{10\log_{10}(z_i(\frac{r_3}{r_1})^\alpha)}{2\sqrt{2}\sigma}) \cdot Q(\frac{10\log_{10}(z_a(\frac{d_{12}}{d_{23}})^\alpha)}{2\sqrt{2}\sigma}) \ , \end{aligned} \tag{10.36}$$

where $Q(x) = \frac{1}{\pi}\int_x^\infty \exp^{-x^2} dx$.

Following similar procedures in the uplink case, we can also obtain the reliability of downlink concurrent transmission:

$$\begin{aligned} F_{CT}^{(d)}(\boldsymbol{\xi'}) &= P\{(SIR_i^{(d)}(\xi_{03}, \xi_{13}) > z_i) \cap (SIR_a^{(d)}(\xi_{12}, \xi_{02}) > z) | MS_1 \in R_{CT}^{(d)}\} \\ &= Q(\frac{10\log_{10}(z_i(\frac{r_3}{d_{13}})^\alpha)}{2\sqrt{2}\sigma}) \cdot Q(\frac{10\log_{10}(z_a(\frac{d_{12}}{r_2})^\alpha)}{2\sqrt{2}\sigma}) \ . \end{aligned} \tag{10.37}$$

10.5 MAC Layer Throughput Analysis

In this section, the MAC layer throughput performance of the considered hybrid infrastructure and overlaying CR-based ad hoc network is evaluated from a PHY/MAC cross-layer perspective. The main task here is to incorporate the interference from both the infrastructure and ad hoc links into the throughput evaluation model in the MAC layer.

In this paper, the CSMA/CA MAC protocol with the binary exponential backoff algorithm is considered because it is widely deployed in many license-exempt frequency bands. However, the CSMA/CA MAC protocol may not be used to establish the CR-based ad hoc link since the clear channel assessment (CCA) by measuring the received signal strength (RSS) may forbid the transmissions in the presence of infrastructure link. To remove this constraint, we use the location and channel station information to replace the RSS measurement for CCA in the traditional CSMA/CA MAC protocol. Therefore, the CR device can establish the ad hoc connection once the new connection does not injure the existing primary infrastructure link.

Next, we first summarize the throughput calculation in the traditional CSMA/CA MAC protocol [187, 188]. Assume N stations always transmit data packets in the network, and let W and $2^b W$ be the minimum and maximum backoff window sizes, respectively. Given the stationary transmission probability τ that a station transmits packet in a given slot and the successful transmission probability $p_s(N)$, the throughput $S(N)$ of the CSMA/CA MAC protocol can be expressed as

$$S(N) = \frac{p_{tr} p_s(N) \mathrm{E}[P]}{(1 - p_{tr})\sigma + p_{tr}(1 - p_s(N))T_c + p_{tr} p_s(N) T_c} \ , \tag{10.38}$$

where $p_{tr} = 1 - (1 - \tau)^N$; E[P], T_s, T_c, and σ are the average payload size, the average successful transmission duration, the average collision duration, and the slot duration. The stationary transmission probability τ is a function of the packet loss probability p_L, that is

$$\tau(p_L) = \frac{2}{1 + W + p_L W \sum\limits_{i=0}^{b-1} (2p_L)^i} \ . \tag{10.39}$$

Note that both the packet loss probability p_L and the successful transmission probability $p_s(N)$ are influenced by the radio channel effect and the multiuser capture effect in the physical layer [188].

Then, we evaluate the total throughput performance of the concurrent transmission in the hybrid infrastructure and overlaying CR-based ad hoc network. Here, we assume N_{CR} CR devices and N non-CR devices using the same frequency band in the coverage of a base station. Since the CR device can establish an ad hoc connection without interfering the existing infrastructure link, the total throughput of such a hybrid network S_{CT} is independently contributed by the two links. The throughput of the infrastructure link and the CR-based ad hoc connection are denoted by S_i and S_a, respectively. The total throughput S_{CT} then can be expressed as

$$S_{CT} = S_i(N) + S_a(N_{CR}P_{CT}) \ . \tag{10.40}$$

Since N non-CR devices contend for data transmission to the base station, the throughput of infrastructure-based link S_i is the same as (10.38). However, because only $N_{CR}P_{CT}$ CR devices have the opportunity to establish the connection, the throughput of a CR-based ad hoc link S_a is similar to S_i, but the number of contending stations changes to $N_{CR}P_{CT}$.

10.6 Numerical Results

In this section, we first investigate the concurrent transmission probability of the infrastructure and overlaying CR-based ad hoc network. Then we apply the proposed cross-layer analytical model to evaluate the total throughput performance in this hybrid network. Figure 10.1 illustrates the considered network topology, where MS_1, MS_2 and MS_3 are the CR-based ad hoc transmitter, receiver and infrastructure primary user, respectively. The stations MS_2 and MS_3 are, respectively, located at $(r_2, -\frac{\pi}{2})$ and $(r_3, \frac{\pi}{2})$, where r_2 and r_3 are the distances between the base station to MS_2 and MS_3; whereas MS_1 is uniformly distributed in the cell with radius $R = 100$ meters. In addition, we also perform the simulation to verify the proposed analytical model for the concurrent transmission probability P_{CT}. In the simulation, 10^4 points, which are uniformly distributed in the region πR^2, represent the possible locations of the ad hoc transmitter MS_1. The probability P_{CT} is calculated by counting the number of points where MS_1 can successfully establish an

Table 10.1: System Parameters for Concurrent Transmission in an Overlaying Ad Hoc Cognitive Radio Network

MAC/PHY header	224/192 bits
ACK/RTS/CTS	304/352/304 bits
DATA payload	16000 bits
Bit rate	1 Mbps
Slot time	20μs
SIFS/DIFS	10/50μs
Min contention window	32
Maximum backoff stage	5
Transmission power, P_t	20 dBm
Noise power, N_0	-90 dBm

ad hoc link to MS$_2$ in the presence of the infrastructure link (base station to MS$_3$). As shown in the following figures, the results in the analytical model agrees well with that in the simulation. The other system parameters are listed in Table I.

10.6.1 Uplink Concurrent Transmission Probability

Figure 10.4 shows the impact of the primary user's location on the uplink concurrent transmission probability $P_{CT}^{(u)}$, where the transmission power $P_t =$ 20 dBm and noise power $N_0 = -90$ dBm, respectively; the required link SIR threshold is 0 dB or 3 dB. First, one can see that the analytical results match the simulation results well. Second and more importantly, there exists an optimal concurrent transmission probability $P_{CT}^{(u)}$ against the distance r_3 from the primary user MS$_3$ to the base station. Note that for $z_i = 0$ dB, the maximal $P_{CT}^{(u)} = 0.45$ at $r_3 = 40$ meters; and for $z_i = 3$ dB the maximal

$P_{CT}^{(u)} = 0.22$ at $r_3 = 26$ meters. This phenomenon can be explained as follows. On the one hand, when MS$_3$ approaches to the base station, it is also closer to the CR-based ad hoc receiver, thereby causing higher interference and decreasing the concurrent transmission probability. On the other hand, when MS$_3$ moves away from the base station, its uplink SIR decreases due to the weaker signal strength and thus yields a lower $P_{CT}^{(u)}$. Hence, an optimal primary user's location can be found in the sense of maximizing the uplink concurrent transmission probability $P_{CT}^{(u)}$.

Figure 10.5 shows the impact of MS$_2$'s locations on the uplink concurrent transmission probability $P_{CT}^{(u)}$. As shown in the figure, as the CR-based ad hoc user moves away from the base station, the concurrent transmission probability monotonically increases from 10% to 50% because the interference from the infrastructure-based link to the ad hoc connection decreases.

10.6.2 Downlink Concurrent Transmission Probability

Figure 10.6 shows the downlink concurrent transmission probability $P_{CT}^{(d)}$ versus the distance r_3 of the primary user MS$_3$ to the base station when user MS$_2$ is located at $(50, -\frac{\pi}{2})$. For the SIR requirement $z_i = z_a = 0$ dB, $P_{CT}^{(d)} = 25\%$ is a constant in the range of $r_3 \leq 100$ meters. This is because the interference transmitted from the base station to the ad hoc users is independent of the locations of the primary user, MS$_3$. However, a more stringent SIR requirement $z_i = z_a = 3$ dB yields a lower and decreasing downlink concurrent transmission probability when r_3 increases.

Figure 10.7 shows the impact of CR user MS$_2$'s locations on the downlink concurrent transmission probability. Similar to Fig. 10.5, $P_{CT}^{(d)}$ also monotonically increases when CR user MS$_2$ moves away from the base station. However, comparing Figs. 10.5 and 10.7, the uplink's concurrent transmis-

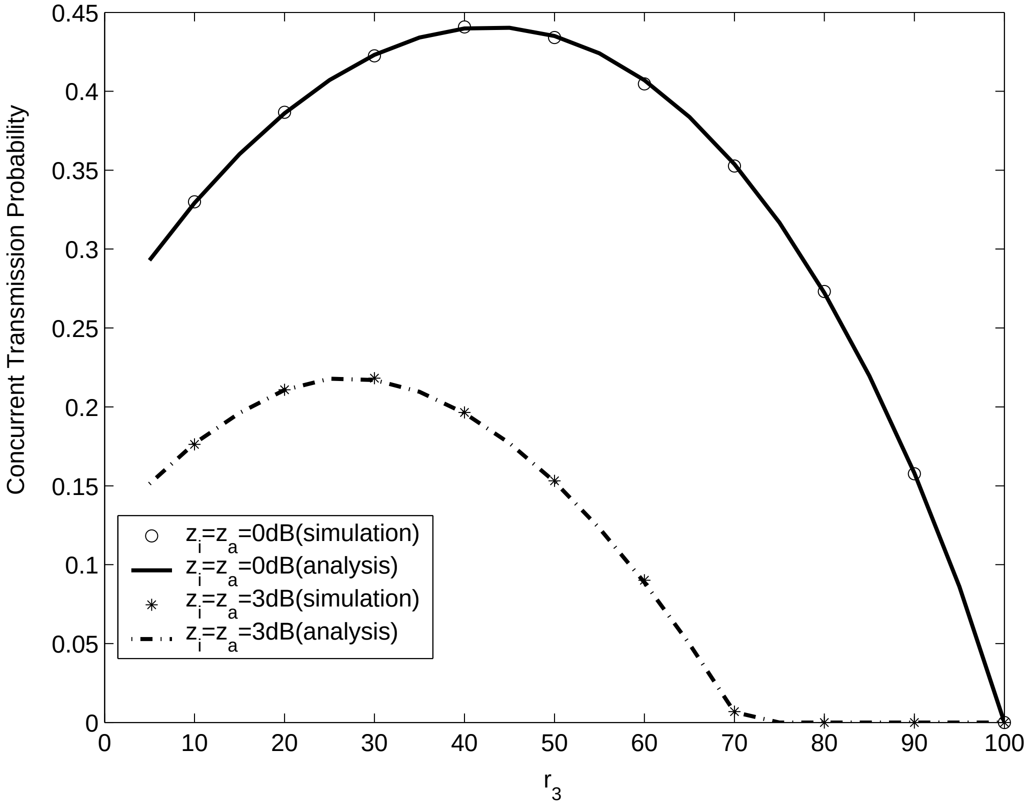


Figure 10.4: The concurrent transmission probability $P_{CT}^{(u)}$ versus the infrastructure uplink user's locations as the ad hoc receiver MS_2 is located at $(50, -\frac{\pi}{2})$, where r_3 is the distance between the base station and the primary user MS_3.

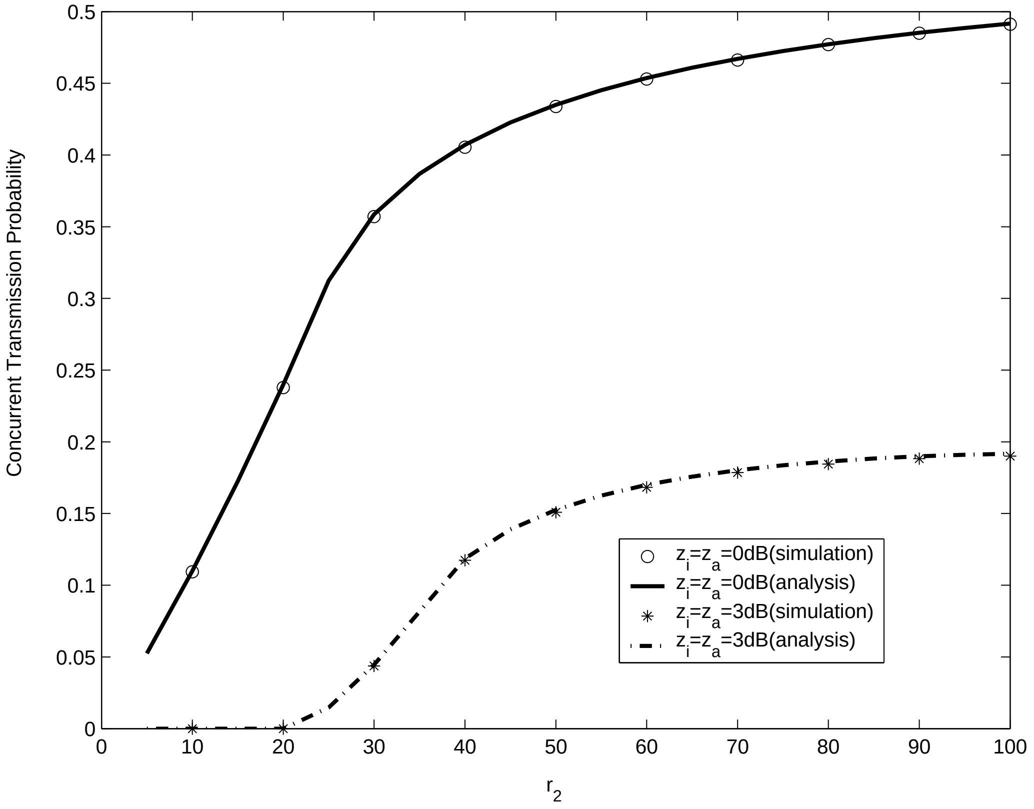


Figure 10.5: The concurrent transmission probability $P_{CT}^{(d)}$ versus the CR-based ad hoc receiver's location as the infrastructure uplink user MS_3 is located at $(50, \frac{\pi}{2})$, where r_2 is the distance between the base station and ad hoc link receiver MS_2.

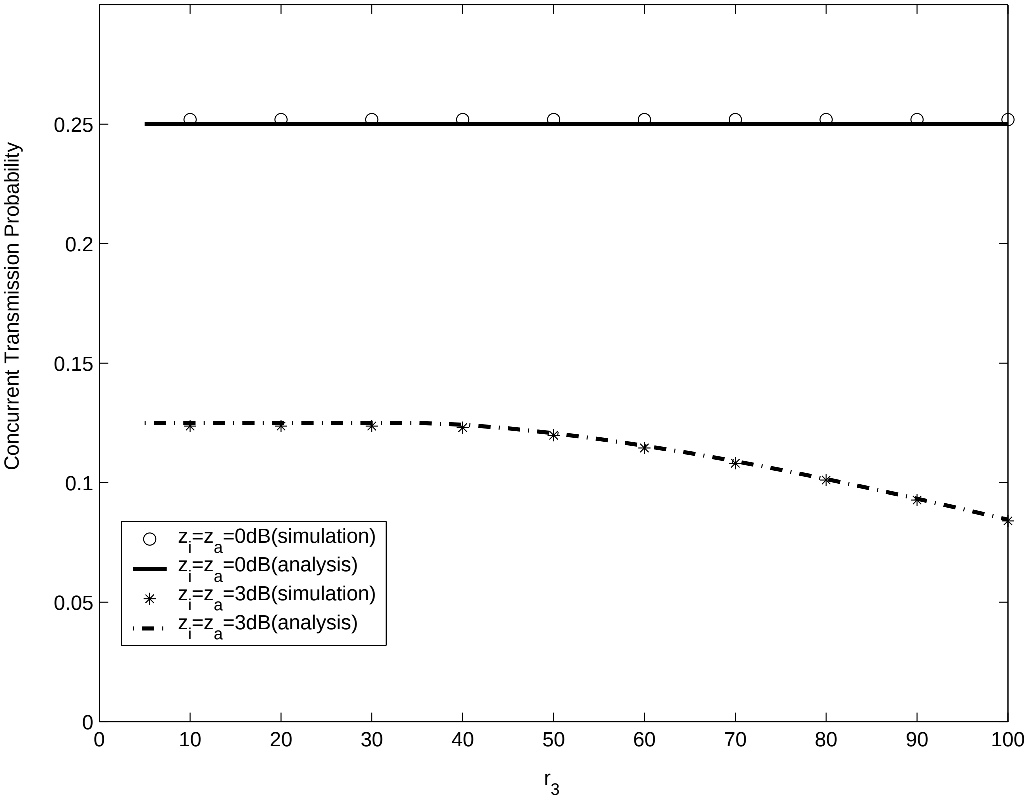


Figure 10.6: Impact of primary user MS_3's location on the downlink concurrent transmission probability $P_{CT}^{(d)}$ as the ad hoc receiver MS_2 is located at $(50, -\frac{\pi}{2})$.

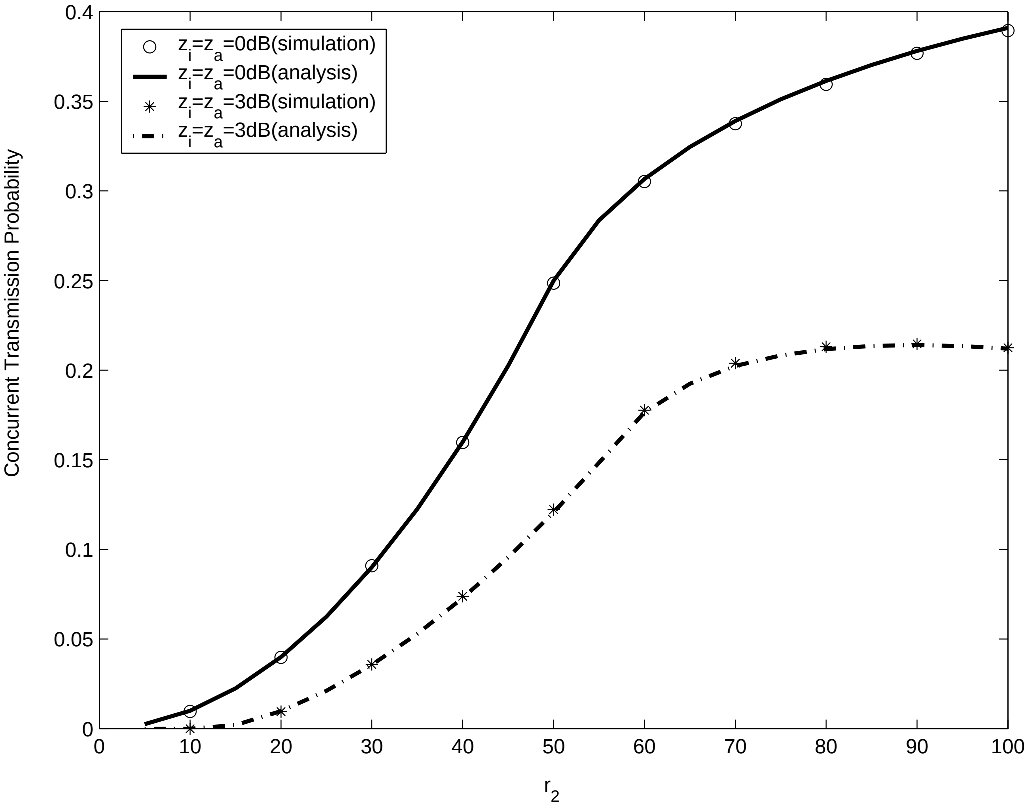


Figure 10.7: Impact of CR-based ad hoc receiver MS_2's location on the concurrent transmission probability $P_{CT}^{(d)}$ as the infrastructure downlink user MS_3 is located at $(50, \frac{\pi}{2})$.

sion probability is higher than that of the downlink's. For $z_i = z_a = 0$ dB and $r_2 = 100$ meters, $P_{CT}^{(u)} = 49\%$ and $P_{CT}^{(d)} = 39\%$, respectively. This is because in the considered scenario the interference to the ad hoc user from the infrastructure-based uplink transmission is weaker than that from the downlink transmission.

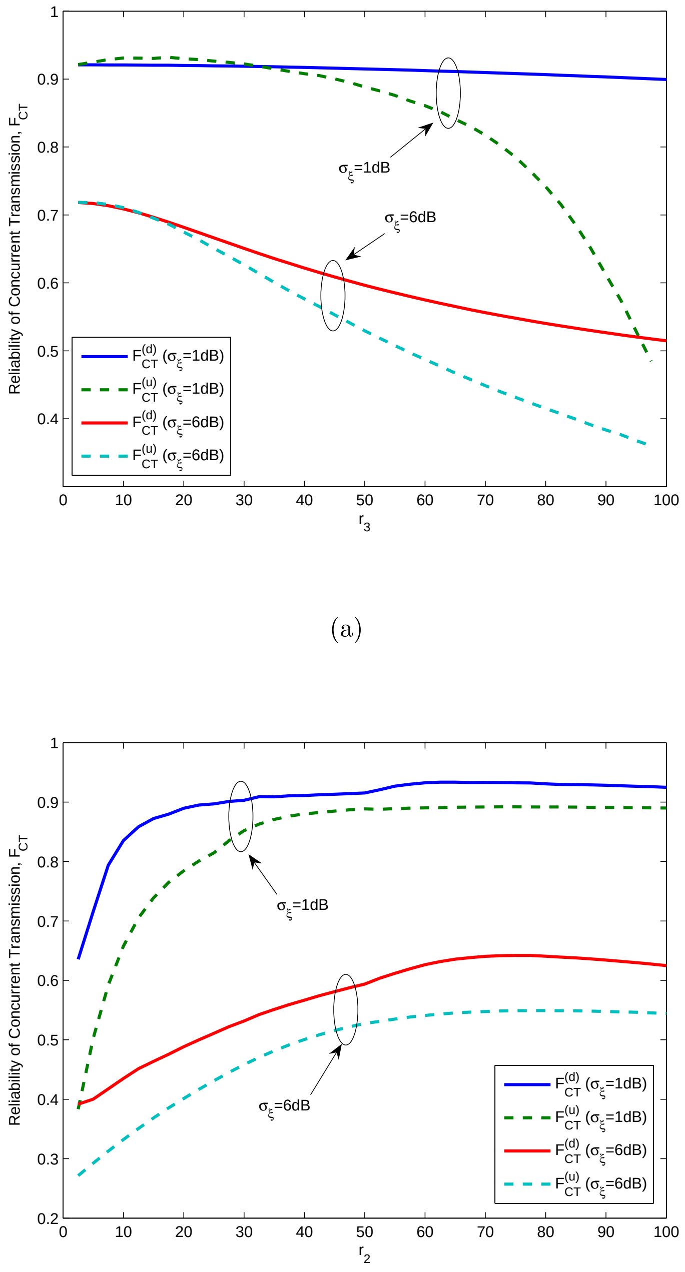
10.6.3 Effects of Shadowing on the Concurrent Transmission

Figures 10.8(a) and (b) illustrate the reliability of the concurrent transmissions with various shadowing standard deviations versus r_3 and r_2, respectively. In general, comparing $\sigma_\xi = 6$ dB and $\sigma_\xi = 1$ dB, one can find that the larger shadowing variance leads to a lower reliability for both uplink and downlink concurrent transmissions. For example, in Fig. 10.8(a), when the primary user's distance to the base station r_3 in the range of $0 \sim 100$ meters, $F_{CT}^{(d)}$ is larger than 0.9 for $\sigma_\xi = 1$ dB, whereas it decreases to $0.6 \sim 0.7$ for $\sigma_\xi = 6$ dB. However, when the primary user moves to the cell edge, the reliability of uplink and downlink concurrent transmissions decreases due to shadowing and weaker received signal strength. As shown in Fig. 10.8(a), for $\sigma_\xi = 6$ dB, $F_{CT}^{(d)}$ and $F_{CT}^{(u)}$ decrease from 0.7 to 0.5 and 0.4, respectively. Since the uplink signal strength is weaker than the downlink signal, the reliability of uplink concurrent transmission is usually more sensitive to shadowing effects than downlink concurrent transmission, especially when the primary user is at the cell edge. In Fig. 10.8(b), it is shown that, subject to the influence of shadowing, the reliability of uplink and downlink concurrent transmissions increases when the receiver MS_2 of the ad hoc link approaches to the cell edge. For $\sigma_\xi = 1$ dB, $F_{CT}^{(u)}$ and $F_{CT}^{(d)}$ increase from 0.4 and 0.63 to 0.89 and 0.92 as r_2 increases to 100 meters; for $\sigma_\xi = 6$ dB $F_{CT}^{(u)}$ and $F_{CT}^{(d)}$ also increase from 0.29 and 0.4 to 0.54 and 0.62. Clearly, the interference from the primary user to the ad hoc user becomes weaker when ad hoc users moves away from the base station. As a result, the reliability of concurrent transmission increases and the shadowing effect on the reliability remains constant as $r_2 > 30$ meters for $\sigma_\xi = 1$ dB and $r_2 > 60$ meters for $\sigma_\xi = 6$ dB.

10.6.4 Total Throughput of Cognitive Ad Hoc Networks Overlaying Infrastructure-based System With Concurrent Transmission

Figure 10.9 demonstrates the total throughput of the CR-based ad hoc link and the infrastructure-based uplink transmissions for various numbers of ad hoc users and different locations of primary users. The total throughput is normalized to the infrastructure-based uplink capacity. As shown in the figure, in the worst case at $r_3 = 50$ meters the total throughput with the concurrent transmission is still 145% higher than the pure infrastructure-based uplink, and the total throughput reaches a maximum of 173% at $r_3 = 10$ meters.

Figure 10.10 shows the total throughput performance of the concurrent transmission of infrastructure-based downlink and ad hoc link. In this case, the concurrent transmission probability is constant for various locations of primary users as shown in Fig. 10.6. Thus the throughput is mainly affected by the number of ad hoc users. For $N_{CR} = 50$, the total throughput is 157% when $10 < r_3 < 100$ meters. However, when $r_3 = 50$ meters, the total throughput improves from 148% to 173% as N_{CR} is changed from 100 to 10.



(a)

(b)

Figure 10.8: Impacts of shadowing on the reliability of downlink $F_{CT}^{(d)}$ (solid line) and uplink $F_{CT}^{(u)}$ (dotted line) concurrent transmission against the locations of (a) the primary user MS_3 and (b) the ad hoc user MS_2 in the cases of $\sigma_\xi = 1$ and 6 dB, respectively.

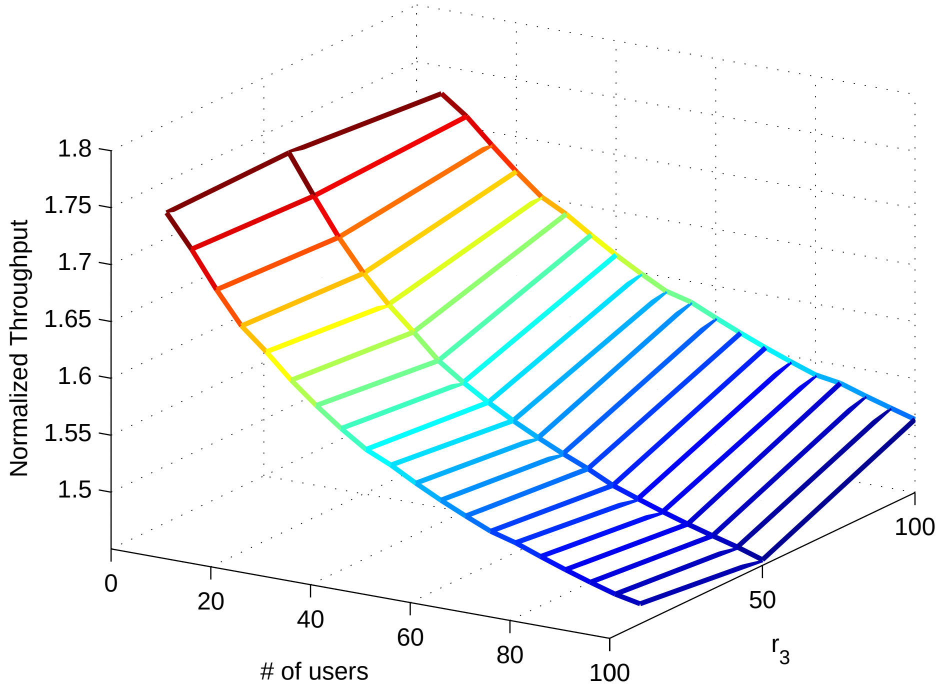


Figure 10.9: Total throughput performance of the uplink concurrent transmission.

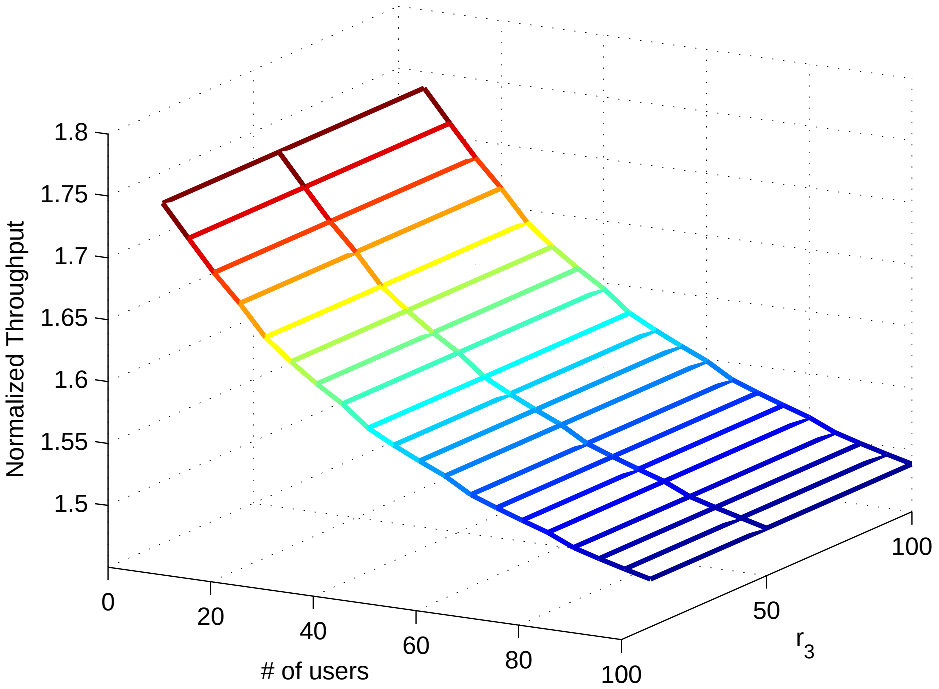


Figure 10.10: Total throughput performance of the downlink concurrent transmission.

Chapter 11

Neighbor-Aware Cognitive MAC Protocol for Spectrum Sharing

In this chapter, we focus on the cognitive MAC protocol design, which is different to the conventional scheme with two objectives: the avoidance of primary user's transmissions and short access delay. In additional to the objectives of high spectrum utilization and QoS provisioning in traditional networks, the cognitive MAC protocol for CR devices has to determine whether its transmission will interfere the primary user at the current and future time period. Moreover, the CR device also demands to access the channel only in a short period of time because primary users have the highest priority to access the channel. Due to the short available transmission time, the fairness from the aspect of access delay among users is more important than the amount of delivered bit for the cognitive MAC protocol, which is also different from the requirement in the legacy MAC design.

11.1 Motivation

According to [17–19], the main functionality of a cognitive MAC protocol, as shown in Fig. 1.4, can be summarized as follows:

- ***observe*** stage - to sense the surrounding environment and record the spectrum usage of the existing legacy systems;

- ***plan*** stage - to evaluate if a temporary ad hoc link can be established without interfering current users;

- ***decide*** stage - to determine the transmit power, frequency, the time and the duration of the frame transmission;

- ***act*** stage - to perform transmission with specified resources at the scheduled time.

To achieve the aforementioned objectives, we design an enhanced CSMA/CA MAC protocol for the spectrum access of secondary users. The CSMA/CA MAC protocol has the preliminary function of spectrum avoidance to the primary users. To start with, we examine the CSMA/CA MAC protocol by referring the four stages of the cognition cycle in Fig. 1.4. First, from the viewpoint of the ***observe*** stage, the cognitive MAC protocol is required to record the spectrum usage of primary users and to collect the traffic characteristics, such as the delay-sensitive or non-real-time data traffic. For the CSMA/CA MAC protocol, most recent research results, instead of identifying the interference, focus on either sensing the carrier transmission in the surrounding environment or avoiding collisions [107–109]. Thus, the functions of recording the spectrum usage and traffic characteristics are not fully considered in the current CSMA/CA MAC protocols.

Second, in the ***plan*** stage of the cognition cycle, the cognitive MAC protocol shall determine whether the requested frame transmission from the secondary user will interfere the primary user's connection. Because the cognitive MAC protocol only permits the secondary user to utilize the spectrum of the legacy system during the spare time of the primary user's transmissions, the access delay in the cognitive MAC protocol for secondary users shall be small. The standard deviation of the access delay in a cognitive MAC should be reduced to make all the secondary users have the equal opportunities of accessing the channel. However, the fairness problem in terms of equal access delay is not emphasized in many modified CSMA/CA MAC protocols [114–117]. Furthermore, a cognitive MAC protocol shall differentiate the priority for various traffic types with QoS provisioning. Although the authors in [109–113] suggested adjusting the transmission probability with different contention window (CW) sizes and different lengths of black bursts to differentiate the traffic types, the issue of avoiding interference to the legacy system was not fully considered yet.

Third, in the ***decide*** stage, the cognitive MAC protocol schedules frame transmissions for secondary users to satisfy the QoS requirement, especially for delay-sensitive traffic. In previous works, some researchers suggested to reserve time slots prior to delay-sensitive frame transmissions [118–123]. However, such reservation methods require a polling process or additional handshaking procedure to coordinate frame transmissions. These methods consume battery energy and waste the valuable bandwidth in sending management frames. Thus, how to design a *distributed* mechanism to reserve the transmissions for high priority frames becomes an issue.

At last, in the ***act*** stage of the cognition cycle, the cognitive MAC protocol synchronizes stations and execute the transmission at the specified time.

To synchronize the clock of each station, the methods designating a centralized controller to broadcast "*beacon*" signals or utilizing the global clock provided by Global Positioning System (GPS) were suggested in [16,122,123]. However, both methods require additional devices.

Here, we propose such a generic cognitive MAC protocol in overlaying ad hoc networks with emphases on achieving the aforementioned objectives: high spectrum utilization, QoS satisfactory and short access delay. Specifically, in the ***observe*** stage, we propose a mechanism of establishing the neighbor list to help stations to recognize the spectrum opportunities. In the ***plan*** stage, an improved contention resolution mechanism, consisting of the gating mechanism, linear backoff algorithm and stall avoidance scheme, is suggested to enhance the performance of throughput, access delay and fairness from the aspect of short access delay for CR devices. In the ***decide*** stage, a novel invited reservation procedure is developed to ensure a secondary user with QoS provisioning. At last, in the ***act*** stage, a distributed frame synchronization mechanism is proposed to coordinate frame transmissions among secondary users without a centralized controller.

11.2 Neighbor List Establishment in *Observe* Stage

To have the knowledge of the spectrum usage by the existing legacy system, we suggest a neighbor list establishment mechanism to record frame transmissions from primary and secondary users in the ***observe*** stage of the cognition cycle. Here, we assume the cognitive MAC protocol can cooperate with other spectrum sensing, identification, and allocation mechanisms to obtain the spectrum usage of primary users. Then, we partition the observed frames into three categories and respectively store the necessitated information into: Primary user information table (PIT), Reservation Information Table (RIT) and Contention Information Table (CIT). The functions of each table are described as follows.

- Primary user Information Table (PIT) stores the spectrum usage of primary users, including:
  - the address of the PU;
  - the repetition period of the PU's transmission;
  - the frame length of the PU's transmission.

  The PIT records the transmission time of the PU to avoid interfering the existing legacy system. Recall that the PU is assumed to periodically transmit packets using the TDMA MAC protocol. The secondary user can avoid interfering the primary user transmission by acquiring the period and frame length. On the other hand, the neighbor list establishment can incorporate with advance traffic models to estimate the information of PU's transmissions [189, 190]. With the knowledge of PU's transmissions, the secondary user can determine whether its transmission will cause the interference.

- Reservation Information Table (RIT) saves the reservation information of delay-sensitive traffic flows for secondary users including:
  - the source address of the delay-sensitive traffic flow;
  - the sequence number of the delay-sensitive traffic flow.
  - the next packet length in the delay-sensitive traffic flow;

The RIT collects the reservation information in the MAC header of the delay-sensitive data and its corresponding ACK frames for secondary users. The header in the proposed MAC protocol is similar to that in IEEE 802.11 WLAN [16], except for the duration field in the MAC header for delay-sensitive data and its ACK frames. In our proposed cognitive MAC protocol, the duration field represents the length of the next delay-sensitive frame in the flow instead of the length of the current frame as in the IEEE 802.11 WLAN. This duration field will be set to zero if the delay-sensitive traffic flow has no remaining packets. By overhearing the MAC header of frame transmissions in all the reserved flows, CR devices can update its RIT and remove the canceled flow from the list. In addition, the RIT can also use the received order of the observed information for the sequence of reserved flow transmissions. The transmission sequence and frame length incorporating with the newly proposed distributed frame synchronization mechanism help secondary users to recognize the time whether it can transmit packet without interfering to the transmission of PUs and reserved frames, which will be detailed in Section 11.5.

- Contention Information Table (CIT) records the properties of non-real-time traffic including:
  - the source address of the non-real-time data traffic flow.
  - the transmission time of the observed frame;
  - the number of non-real-time data traffic flows;

  The CIT provides the information with the number of non-real-time traffic flows, which will be used to reduce the collisions and improve the channel throughput in the ***act*** stage of the cognition cycle.

To correctly establish PIT, RIT and CIT, a CR user is designed to observe the status of frame transmissions around its neighborhood for an observe period T_{obv} before any frame transmission. The duration of T_{obv} must be longer than the period of the legacy system to ensure the secondary user has the knowledge of the periodic frame transmissions for primary users. Furthermore, T_{obv} shall also be longer than the maximum repetition period between two successive delay-sensitive frames to prevent unnecessary real-time traffic flow establishment. The optimal value for T_{obv} can be obtained through heuristic search but beyond the scope of this paper.

Another interesting point is that the continuous table update is inefficient in terms of energy consumption. For this issue, the proposed MAC protocol can incorporate with some well-known power management, e.g. power saving mode in IEEE 802.11 [16]. The node without packet transmissions can enter the sleeping mode, in which the station turns off all the unnecessary functions. The node will sleep over a fixed period of time and wake up to check whether other nodes have packets to itself. An extra transmission window, like ATIM window in IEEE 802.11, is preserved after all reserved frame transmissions to indicate the packet transmission in the later future. The station having packets to the sleeping node sends the indicative message in this window; otherwise, the sleeping node returns to the sleep mode until the next ATIM window or the time it has packet to send.

11.3 Contention Resolution in *Plan* Stage

In the ***plan*** stage of the cognition cycle, the cognitive MAC protocol has two major functions. One is to prevent CR users from interfering the legacy system, and the other is to make them efficiently and effectively access the unused spectrum in a short available transmission time. To this end, we suggest three improved approaches as follows:

1. gating mechanism - to forbid the transmissions that may interfere to primary users or collide with other CR users;
2. linear backoff algorithm - to expedite the link establishment of delay-sensitive traffic flows;
3. stall avoidance scheme - to speed up the transmission of stalled non-real-time data packets.

The three above mechanisms help to achieve the objectives of high throughput, low access delay, and fairness for secondary users.

11.3.1 Gating Mechanism

The gating mechanism is used to avoid interfering the primary user of the legacy system and to reduce the collision among CR users. The basic idea is cooperating the spectrum usage information obtained from the spectrum sensing, identification and allocation techniques to prevent from interfering the primary users. Recall that the PIT stores the information of primary user transmissions. The gating mechanism postpones the secondary user transmitting packets when the primary user appears on the channel. In addition, we also suggest the modified p-persistent CSMA algorithm to improve the efficiency of spectrum usage for CR users, where the optimal value of p can be computed according to the number of *nrt-nodes* in CIT [187].

The detailed procedure of the proposed gating mechanism is described as follows:

1. When a frame of a CR user is requested for transmission, the gating mechanism first checks whether a legacy user occupies the channel from the information in PIT.
    - If so, the transmission of this CR user is deferred.
    - Otherwise, the optimal transmission probability p is calculated based on the neighborhood information in CIT.
2. Apply the p-persistent algorithm to determine whether the frame can be transmitted:
    - If the frame is granted for transmission, the CR user immediately sends the frame.
    - Otherwise, the frame will be deferred and again contend for the channel access.

According to the proposed procedure, one may argue that it still cause the interference with the legacy system using the CSMA/CA MAC protocol by suppressing the bandwidth. However, most existing systems using the CSMA/CA MAC are operated on unlicensed frequency bands. Both legacy and CR devices have the equal right to access these frequency bands, and thus we believe that the bandwidth suppression is not an issue for secondary users.

11.3.2 Linear Backoff Algorithm

To expedite the channel access in supporting delay-sensitive application, we suggest that the link establishment of delay-sensitive traffic flows shall follow the linear backoff algorithm instead of increasing the CW size exponentially as in the legacy CSMA/CA MAC protocol. That is, if the request for sending the first frame of a delay-sensitive traffic flow is collided, the CW size (CW_{rt}) for that particular frame increases according to the following principle:

$$CW_{rt} = \min(CW_{max}, CW_{min} \times (N_{req} - 1)), \tag{11.1}$$

where N_{req} is the number of attempts for sending the frame; CW_{max} and CW_{min} are the maximum and minimum CW sizes in the contention resolution mechanism, respectively.

Figure 11.1 shows the CW sizes for the linear and binary exponential backoff algorithms. As shown in the figure, the CW size in the linear backoff algorithm increases less slowly than that in the binary exponential backoff algorithm. Therefore, the channel access of the first frame in a delay-sensitive traffic flow can be faster than that of the non-real-time data flows. As long as the delay-sensitive traffic flow is successfully established, the remaining frames are sent in the reserved time slot according to the proposed invited reservation procedure (which will be discussed in Section 11.4). Based on our design, because only the first frame contends for accessing the channel, the number of attempts of establishing a delay-sensitive traffic flow is much fewer than that of non-real-time traffic flows. Thus, the proposed MAC protocol can avoid the collisions issue of the linear backoff algorithm, while reducing the access delay in the link establishment of delay-sensitive traffic flows.

11.3.3 Stall Avoidance Scheme

In order to improve the fairness for the CR users, we develop a stall avoidance scheme aiming to reduce the transmission delay of the *nrt-nodes* with excessive buffered frames. The specific goal of the suggested approach is to minimize the variance of the access delay among all the *nrt-nodes.* Due to the short available transmission time of the spectrum in an overlaying cognitive

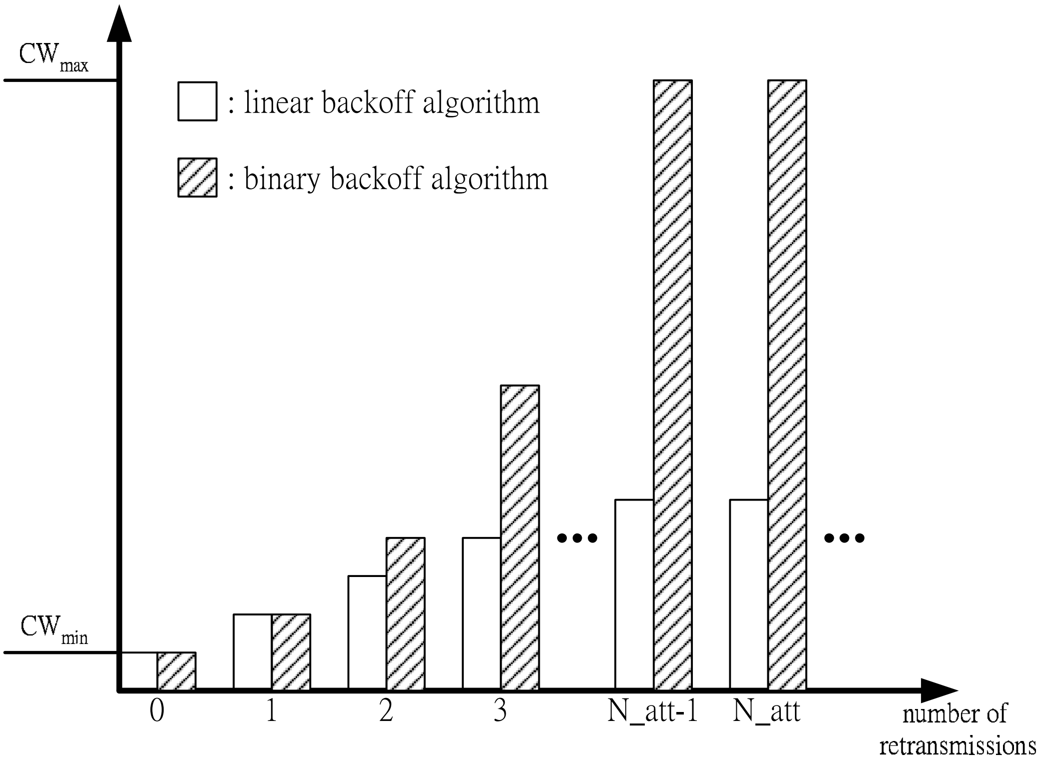


Figure 11.1: Comparison of CW size between linear and binary exponential backoff algorithms.

ad hoc network, the small variance of the access delay makes CR users have equal opportunities to access the channel. Here, the access delay includes the waiting time in the queue and the channel access time. Therefore, obviously, reducing the variance of access delay implies to speed up the back-logged frame transmission.

The suggested stall avoidance scheme with respect to *nrt-nodes* is described as follows. Select a pre-determined threshold $\mathrm{Q_{th}}$ for the maximum allowable buffered data frames and the guaranteed CW size for the stalled *nrt-nodes* CW_{stall}, where

$$CW_{stall} < CW_{min}. \tag{11.2}$$

If the number of buffered frames in an *nrt-node* is more than $\mathrm{Q_{th}}$, the CW size of the subsequent frames in the queue is reduced to CW_{stall}. Because a smaller CW size leads to a higher transmission probability, the lagging frames in a stalled *nrt-node* with CW_{stall} can be transmitted earlier than others, thereby improving the fairness performance among *nrt-nodes.* Both $\mathrm{Q_{th}}$ and CW_{stall} are system parameters, which optimal values can be obtained through heuristic search but beyond the scope of this paper.

One may argue that reducing the CW size worsens the network congestion in a crowded system and thus causes the instability for a network. However, this situation may seldom happen because secondary users in a cognitive network have plenty of channels, and the number of secondary users choose and access on the same channel is small compared to the legacy system. Furthermore, our simulation results shown in the later section illustrate that the system up to 140 stations can still remain stable. Therefore, we believe the system instability is not a severe problem for the proposed MAC protocol.

11.4 Invited Reservation Procedure in *Decide* Stage

Next, another key challenge in designing the cognitive MAC protocol lies in the way of periodically transmitting delay-sensitive traffic flows because any connection in a cognitive ad hoc network cannot interfere the legacy system. To solve this problem, we propose an invited reservation procedure in the ***decide*** stage of the cognition cycle.

11.4.1 Invited Reservation Procedure

The invited reservation procedure is designed for supporting the delay-sensitive application. Based on this procedure, the receiver sends the real-time clear-to-send (rt-CTS) control frame to reserve time slots for the transmitter sending subsequent frames of a reserved delay-sensitive traffic flow. Like the clear-to-send (CTS) control frame, the duration field of the MAC header in the rt-CTS frame defines the length of current frame transmission and thus can be used to forbid the transmissions from the stations in receiver's neighborhood. Because the delay-sensitive frame transmissions are controlled by the receiver of reserved flows, the collisions due to the hidden node problem can be somehow alleviated. For example, Fig. 11.2 illustrates a scenario where the transmitter STA 1 establishes a delay-sensitive traffic flow to the receiver STA 2 in the presence of a hidden node STA 3. In the figure, the receiver STA 2 sends rt-CTS inviting STA 1 to transmit reserved frames. Upon receiving the rt-CTS control frame for STA 1, the hidden node STA 3 recognizes the incoming reserved transmission and halts sending packets. Therefore, the invited reservation procedure can reduce the dropping rate of delay-sensitive traffic flows, especially in an environment with hidden nodes.

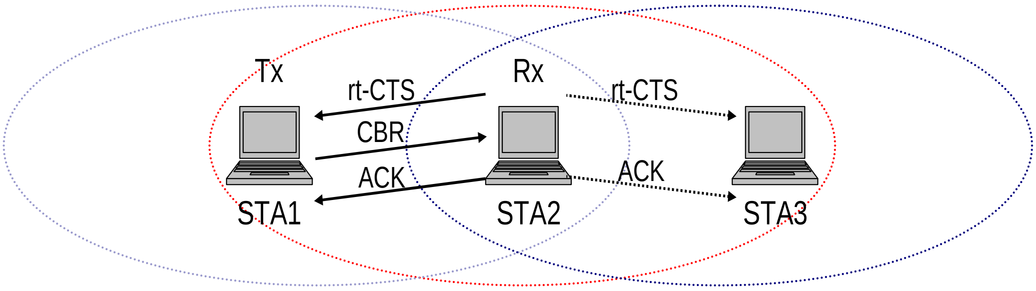


Figure 11.2: An illustration of the invited reservation procedure.

Furthermore, recall that the MAC duration field of delay-sensitive data and its ACK frames represent the frame length of the next packet in the reserved flow. This value will be set to zero if the reserved flow is cancelled, and thus the receiver knows whether the sender has packets once the flow is established. On the other hand, since the receiver can learn the spectrum usage time of primary users in ***observe*** stage of the cognition cycle, the receiver always has the knowledge to adjust the invitation without interfering primary users.

In our design, the transmission based on the reception of rt-CTS may induce an issue that the transmission from the hidden node of the receiver may cause collisions at the sender. In fact, assuming that the hidden node also follows the proposed MAC protocol, it can transmit packets only after it receives all the data or ACK frames from all the reserved flows in RIT. In other words, the hidden node has to wait for the end of all delay-sensitive frame transmissions and then transmits packets accordingly. Therefore, we believe the considered situation may not happen in our proposed MAC protocol.

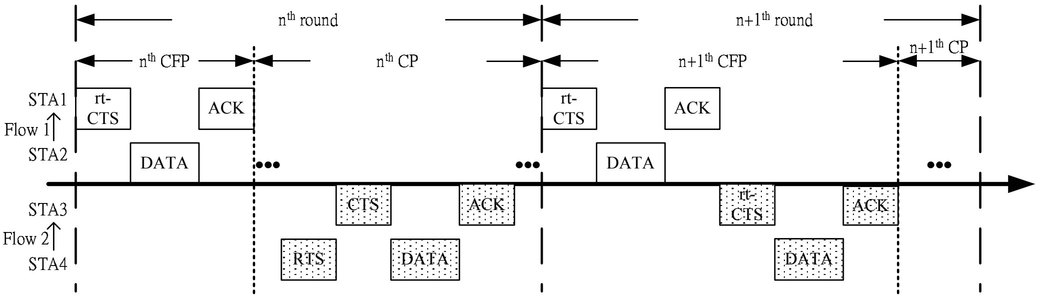


Figure 11.3: The timing diagram for the invited reservation procedure.

11.4.2 Link Establishment with Invited Reservation Procedure

Next, the problem is how to establish a delay-sensitive traffic flow using the invited reservation procedure. In our design, the first packet of a delay-sensitive traffic flow is used for the link establishment through the random access on the channel by request-to-send/clear-to-send (RTS/CTS) handshaking. The first packet transmission also reserves the length and sequence for the next packet transmission. Once the link is successfully established, the receiver periodically sends the rt-CTS control frame to reserve time for the reserved delay-sensitive flow. Figure 11.3 illustrates an example for the link establishment. In Flow 2 (STA 4 → STA 3), STA 4 follows the RTS/CTS handshaking procedure to send the first packet of a delay-sensitive traffic flow during the n$^{\mathrm{th}}$ contention period (CP). As long as the flow is established, STA 3 periodically sends rt-CTS with reserved information to its sender STA 4. Accordingly, without contention, STA 4 transmits the rest packets of the reserved flow in succeeding contention free periods (CFPs).

However, the random access for the link establishment of a delay-sensitive traffic flow induces the collision with other non-real-time data frames and leads packet drop due to the increased access delay. To alleviate the impacts of contentions, we design the linear backoff algorithm for the first packet transmission to shorten the access delay. The linear backoff algorithm decreases the CW size of the collided frame to expedite the link establishment of a delay-sensitive traffic flow. The simulation results in the later section demonstrate that the dropping rate of delay-sensitive frames in the proposed MAC protocol is almost negligible compared to that in the IEEE 802.11 DCF mode. Therefore, we believe the impact of contentions to the link establishment for delay-sensitive traffic is insignificant because most of packets are reserved and sent during an acceptable period.

On the other hand, for a reservation based MAC protocol, one important issue is the starvation problem for non-real-time data traffic. As shown in Fig. 11.3, the fixed total transmission time in each round is partitioned into two periods: the CFP and CP. To prevent the starvation problem, the time duration of the two periods shall be appropriately allocated so that the delay constraints for the delay-sensitive traffic flow can be satisfied, while its impact on the non-real-time data transmission can be limited to an acceptable level. However, precisely controlling the duration of CFP and CP in a distributed way is sophisticated for CR device. Instead, in this paper, the stall avoidance scheme is designed to avoid the bandwidth suppression by expediting the stalled frame transmission. Recall that the stall avoidance scheme will decrease the CW size to CW_{stall} if an *nrt-node* has excessive buffered frames, and $CW_{stall} < CW_{min}$. The stalled non-real-time data frame with the small CW size CW_{stall} can have a higher probability to win the channel contention and prohibits from *rt-nodes* establishing a new delay-sensitive traffic flow. Therefore, in this way, the access delay of non-real-time data frames can be still controlled within a reasonable range without sacrificing the delay constraint for delay-sensitive traffic.

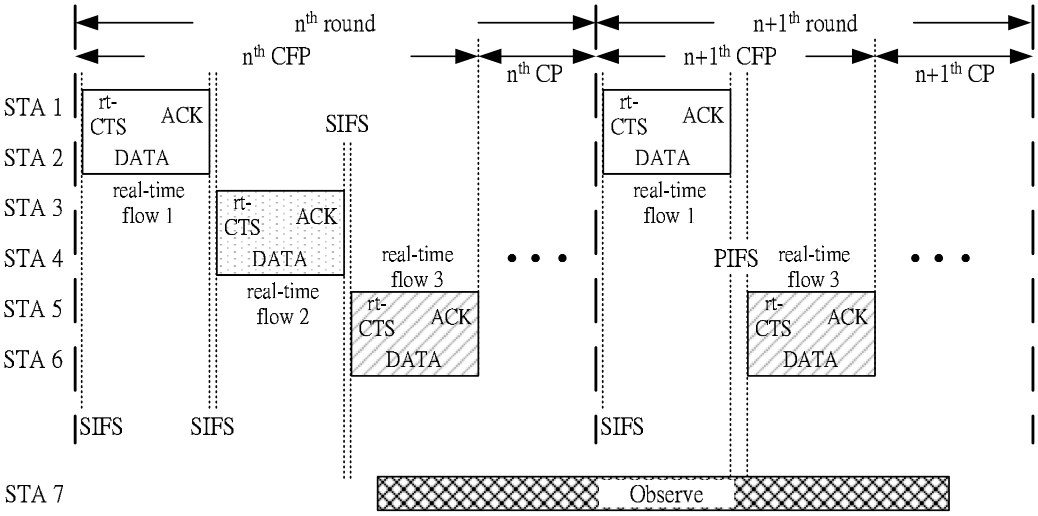


Figure 11.4: The timing diagram for the new proposed distributed frame synchronization mechanism.

11.5 Distributed Frame Synchronization Mechanism in *Act* Stage

Another important issue in the ***act*** stage of the cognition cycle is to develop a distributed approach to ensure the frame synchronization among CR users. The objective of frame synchronization is to inform stations the starting time of the CFP and CP in each round. In the legacy IEEE 802.11 WLAN, the access point or the designated central controller broadcasts the "*beacon*" signal as the start of each round. However, broadcasting "*beacon*" signals not only increases energy consumption, but wastes the valuable transmission time for secondary users. To this end, we propose a new distributed frame synchronization mechanism as follows.

The basic idea of the proposed distributed synchronization algorithm is to let the secondary user transmission follow the sequence of reserved flows in RIT. Since a CR user establishes its neighbor list in the ***observe*** stage, the information in RIT can be applied to identify the first and last stations transmitting the delay-sensitive frames. Thus, when the channel is available, the receiver of the first reserved flow in RIT broadcasts rt-CTS frame to start a new CFP. During the CFP, the *rt-nodes* transmit frames based on the sequence in RIT, whereas the *nrt-nodes* wait until receiving the ACK frame from the receiver of the last flow in RIT. If the primary user become active during CFP, the receiver of the delay-sensitive traffic flow will halt sending the rt-CTS and resume it when the spectrum turns idle. Therefore, through the information stored in PIT and RIT, all the CR users can access the channel in the designated period without influencing the primary users and the transmissions in the reserved time.

Figure 11.4 illustrates an example for the proposed distributed synchronization mechanism. Assume that STAs 1∼6 establish delay-sensitive traffic flows. At the start of the n^{th} CFP, when STA 1 senses the channel is available, it immediately sends rt-CTS to start a new CFP and waits for receiving the frame from STA 2. After sending the ACK frame to STA 2, STA 1 waits for a fixed duration and repeats the above procedure until flow 1 is terminated. In the meanwhile, STAs 3 and 5 overhear the channel and recognize a new CFP. When the previous reserved transmissions are finished, STA 3 send rt-CTS to STA 4 to start flow 2. At last, when the transmissions of the last flow in RIT are ended, i.e., flow 3 from STA 6 to STA5, the CP in the n^{th} round starts. All the *nrt-nodes* are allowed to contend for the channel during this duration until the next CFP starts. Note that the proposed distributed frame synchronization is only needed when delay-sensitive traffic flows exit. The legacy CSMA/CA MAC protocol with the suggested gating mechanism is enough when only non-real-time traffic exists.

Three interesting scenarios are discussed as follows. First, when a CR user just joins the CR network or turns on the power, it may not recognize the time when the CFP and CP start. This particular user may access the channel in CFP and collide the on-going transmission. To avoid this kind of collisions, the previously suggested neighbor list establishment mechanism has been carefully designed to resolve this problem. Recall that every CR user overhears the spectrum for at least a $\mathrm{T_{obv}}$ period to correctly establish PIT, RIT and CIT before transmission. Since the duration of $\mathrm{T_{obv}}$ is longer than the maximum repetition period between two successive frames of a reserved flow, a new user can be aware of the first and last transmissions in CFP based on the information in its RIT. Thus, it can easily recognize the starting time of the two periods in each round. Take STA 7 in Fig. 11.4 as an example. Since it observes the channel for a $\mathrm{T_{obv}}$ period and establishes PIT and RIT, STA 7 can recognize that CFP and CP respectively start after STAs 1 and 5 send the rt-CTS and ACK frames. Therefore, STA 7 will access the channel during CP without colliding with the transmission of primary users or reserved flows.

Another interesting issue is when some CR users can not maintain frame synchronization due to the failure reception of rt-CTS, data and ACK frames in the lossy wireless channel. Under this situation, the *nrt-node* may send the data frame in CFP and cause collisions to other reserved flows. To deal with this problem, the invited reservation procedure is designed with following principle. If not hearing the rt-CTS frame longer than a duration of PCF inter-frame spacing (PIFS), the receiver of the next reserved flow in RIT immediately sends the rt-CTS frame to reserve time slots. Note that in the IEEE 802.11 WLAN, PIFS is longer than SIFS, but shorter than data inter-frame spacing (DIFS). In this way, the rt-CTS frame always has the highest transmission priority, and thus it can protect the transmissions in CFP from the interference of unsynchronized stations. As shown in Fig. 11.4, when STA 5 identifies a new CFP, it waits for rt-CTS or data frame from STA 3 or 4. Once the channel is idle longer than PIFS as shown in the n+1$^{\mathrm{th}}$ CFP, STA 5 directly sends rt-CTS to prevent the non-real-time frame transmissions from unsynchronized stations. Therefore, the frame synchronization can be maintained even though some frames are lost due to signal outage.

At last, one may be curious that whether the new distributed frame synchronization mechanism is backward compatible and interoperable with the legacy DCF and PCF modes in IEEE 802.11 WLAN. On the one hand, the transmission in DCF mode has no influence to that in our invited reservation procedure. Because the maximum spacing between any two delay-sensitive frames in our MAC protocol (PIFS) is shorter than the minimum duration of two data frames in DCF mode (DIFS), i.e., PIFS $<$ DIFS. The frame transmission in DCF mode cannot disturb the reserved frame transmission in CFP. On the other hand, the neighbor list establishment helps secondary users to recognize the transmission in the PCF mode and treat them like primary user transmission. Therefore, with the gating mechanism and information in PIT, all the secondary user transmissions in our MAC protocol are postponed until the end of PCF mode to avoid interfering the transmission in PCF mode.

11.6 Throughput Analysis

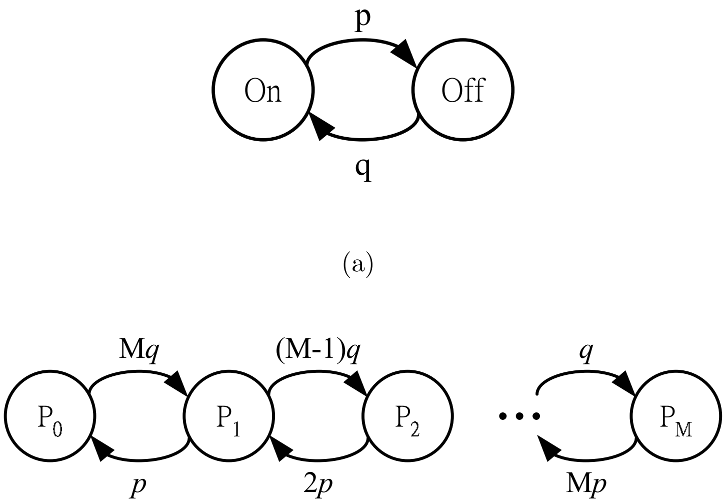
In this section, we analyze the throughput of our proposed cognitive MAC protocol with mixed delay-sensitive and non-real-time data traffic flows. To ease the analysis, we make the following assumptions: (1) the spectrum usage information is correctly obtained by the spectrum sensing, identification, and allocation mechanisms; (2) the channel is ideal without transmission errors; (3) a fixed number of *nrt-nodes* always have packets to send; (4) the delay-sensitive traffic of a fixed number of *rt-nodes* is characterized by the "on/off" model with the exponentially distributed inter-arrival and departure time [191]. Recall that CR users can send packet only in the spare time of primary user transmissions. Thus, we only consider the throughput performance during the time available for CR users.

11.6.1 Mixed non-real-time and delay-sensitive traffic flows

Next, we consider a mixed traffic model with delay-sensitive and non-real-time data traffic flows. Denote M and K the number of *rt-nodes* and *nrt-nodes*, respectively, and let $n_{rt}(t)$ be the number of *rt-nodes* requesting for frame transmissions at the time instant t. For simplicity, it is assumed that only one delay-sensitive traffic flow requests to establish in each round.

The delay-sensitive traffic is assumed to be modeled by an interrupted Poisson process (IPP), as shown in Fig. 11.5(a). In the figure, the "On" state represents a talk spurt, whereas the "Off" state is for a silent spurt [191]. The durations for both states are exponentially distributed with a mean value of $1/q$ and $1/p$, respectively. In addition, an M-stage Markov-modulated Poisson process (MMPP) shown in Fig. 11.5(b) is applied to model multiple delay-sensitive traffic flows. Each state in the figure stands for the number of *rt-nodes* requesting for frame transmissions, and thus the state probability (P_i) can be expressed as

$$P_i = P\{n_{rt}(t) = i\} = \left(\begin{array}{c} M \\ i \end{array} \right) \rho^i P_0 \ , \qquad (11.3)$$



(a)

(b)

Figure 11.5: (a) Interrupted Poisson process model for delay-sensitive traffic. (b) Markov modulated Poisson process (MMPP) model for one type delay-sensitive traffic with M users.

where $\rho = q/p$.

Denote $T(M,K)$ the throughput of M *rt-nodes* and K *nrt-nodes* in an overlaying cognitive ad hoc network, which can be given by

$$\begin{aligned} T(M,K) &= E[\text{throughput of } M \textit{ rt-nodes} \text{ and } K \textit{ nrt-nodes}] \\ &= \sum_{i=0}^{M-1} P_i \cdot \left(\frac{iL_{rt}}{L} T_{rt} + \frac{L - iL_{rt}}{L} T_{nrt}(K) \right) \\ &= T_{nrt}(K) + (T_{rt} - T_{nrt}(K)) \sum_{i=0}^{M-1} P_i \frac{iL_{rt}}{L} \\ &= T_{nrt}(K) + (T_{rt} - T_{nrt}(K)) \frac{L_{rt}}{L} \frac{M\rho}{1+\rho} \ , \end{aligned} \tag{11.4}$$

where L is the entire duration of two secondary transmission periods, i.e., CFP and CP; L_{rt} is the total duration for sending delay-sensitive data frames as well as the rt-CTS and ACK control frames; $T_{nrt}(K)$ represents the re-

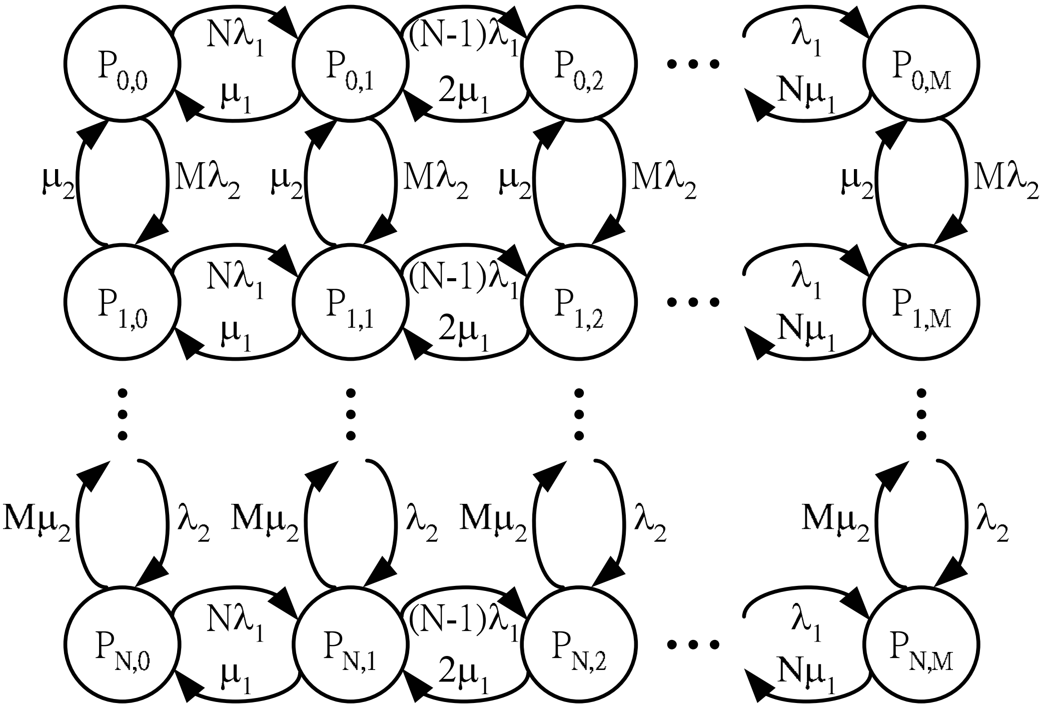


Figure 11.6: Two-dimensional Markov chain for the analysis of two delay-sensitive traffic types.

ceived frames from K *nrt-nodes* in the CP; T_{rt} contains the delay-sensitive data frames. Note that L is assumed to be fixed because it excludes the duration of primary user transmissions.

The above analysis can be extended to the mixed traffic model containing multiple types of delay-sensitive traffic by using a multi-dimension Markov chain. Take two delay-sensitive traffic types as an example and let N and M be the numbers of CR users sending these traffic types, respectively. Then, the state probability $P_{i,j}$ of the two-dimensional Markov chain model in Fig. 11.6 can be expressed as

$$P_{i,j} = \begin{pmatrix} N \\ i \end{pmatrix} \begin{pmatrix} M \\ j \end{pmatrix} \rho_1 \rho_2 P_{0,0} \ , \qquad (11.5)$$

where $\rho_1 = q_1/p_1$ and $\rho_2 = q_2/p_2$ are similar to the definitions in (11.3). Thus, the total throughput $T(M, N, K)$ can be written as

$$T(M,N,K) = T_{nrt}(K) + (T_{rt1} - T_{nrt}(K))\,\frac{L_1}{L}\frac{M\rho_1}{1+\rho_1} + (T_{rt2} - T_{nrt}(K))\,\frac{L_2}{L}\frac{N\rho_2}{1+\rho_2} \tag{11.6}$$

where all the parameters are defined in (11.4).

11.7 Simulation

In this section, we demonstrate the performance of the proposed cognitive MAC protocol through the NS-2 simulator [192]. We also use the IEEE 802.11 DCF mode with the RTS/CTS handshaking for the performance comparison. The RTS/CTS DCF mode is naturally a good candidate for secondary user in a cognitive network because the carrier sensing avoids interfering primary user transmissions at the moment of any packet transmission. Furthermore, most of current cognitive MAC protocol still use the CSMA/CA MAC protocol to resolve the collisions among secondary user [190, 193, 194]. Thus, the performance of the CSMA/CA MAC protocol with RTS/CTS handshaking can be a baseline in a cognitive network.

11.7.1 Simulation Environment

To begin with, we explain the considered simulation environment. In this paper, we consider two network topologies, in which nodes are 1) allocated in one cluster and 2) distributed to three clusters in string-type topology. In the first network topology, all the primary and secondary users are located in the same cluster with the area of 10 m × 10 m. On the other hand, in the second network topology, primary users are only located in Region II, but secondary users are distributed at all the three regions, as shown in Fig. 11.7.

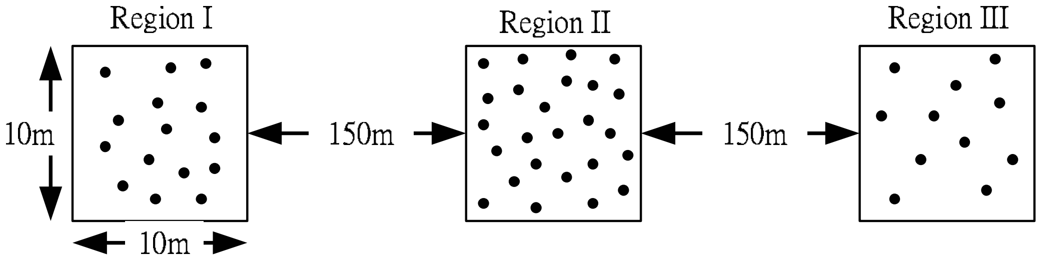


Figure 11.7: The considered network topology for the simulation, in which nodes are distributed to three clusters in a string-type topology.

Furthermore, the nodes in region I are hidden from that in region III, and vice versa. The secondary users in regions I and III can send packets only to those in region II. In our simulation, the primary user is assumed to adopt the TDMA MAC protocol, whereas secondary users use the proposed MAC protocol or the CSMA/CA MAC with RTS/CTS handshaking for packet transmissions. In addition, it is assumed that the transmission is considered success if only one user accesses on the channel, while the collision takes place when more than one user transmit packets at the same time. The other simulation parameters are listed in Table 11.1.

Furthermore, the following traffic models are considered in the simulation.

- **Voice traffic** is characterized by an interrupted Poisson process. In the "On" state, an 164-byte packet is generated every 20.48 msec, which is equal to 64 Kbps. By contrast, the simulator stops generating any packet in the "Off" state. The duration of the "On" and "Off" states follows the exponential distribution with the average durations of 1 and 1.3 seconds, respectively.

- **Telnet data traffic** is modeled by a Poisson process with the packet length of 950 and 60 bytes. The packet inter-arrival rates are determined by the offered load.

Table 11.1: System Parameter for the Simulation of Cognitive MAC Protocol

Area of one cluster	10 m×10 m
Distance of each cluster	160 m
Data Rate	2 Mbps
Slot time	20 μsec
SIFS	10 μsec
PIFS	30 μsec
DIFS	50 μsec
Minimum CW size	31
Maximum CW size	1023
CW for stall avoidance (CW_{stall})	15
Maximum frame transmission times	7
Number of primary users	10
Number of rt nodes	10
Number of nrt nodes	80
Period of delay-sensitive flow	20 msec
Period of primary user transmission	20 msec
Packet size of primary user and delay-sensitive flows	164 bytes

- **FTP data traffic** is assumed that the simulator continuously generates packets with three frame sizes: 950, 500 and 60 bytes if a node successfully send the previous one.

In our simulation, only half of users have packets to send, and the others are the corresponding receivers. Without any notice in the following, the network always consists of 40 Telnet data flows in the presence of 10 voice traffic flows, which are respectively generated and received by 80 *nrt-nodes*, 10 *rt-nodes* and 10 primary users.

11.7.2 Performance Measurements

The performance of the proposed cognitive MAC protocol is evaluated in terms of the normalized throughput, mean access delay, fairness, and dropping rate. For clearance, we define these performance metrics as follows.

- **Normalized throughput** is defined as the ratio of the number of successfully received bits to the amount of total transmitted bits. Note that the successfully received bits account for both the successfully received non-real-time data frames and the delay-sensitive frames as well as the primary user frame transmissions. The received frames from delay-sensitive traffic and primary users will not be counted if the 20-msec delay constraint is not satisfied.

- **Mean access delay** accounts for the average duration a non-real-time data frame requires when it is generated by the transmitter until it is successfully received by the receiver. By this definition, the access delay includes the waiting time in the queue and the latency of the channel access. The frame, which number of transmission times is larger than the maximum retry limit, will not be taken into account for the access delay calculation.

- **Fairness** is evaluated by the maximum standard deviation of the frame access delay among all the non-real-time traffic flows. Note that the available transmission time for the secondary user in a cognitive network is short. The reduced delay variance among traffic flows represents that all secondary users have equal opportunities to send packets on the channel during the short transmission time. Thus, the maximum standard deviation of access delay among user is more appropriate for measuring the fairness performance in a cognitive network.

- **Dropping rate** is defined as the ratio of the number of dropped frames to the total number of transmitted frames. The frame is considered to be dropped if the access delay of the delay-sensitive frame or the primary user transmission is beyond 20 msec.

11.7.3 Numerical Results

First, we examine the dropping rate of primary user transmissions in the network topology that all user are located in one cluster, as shown in Fig. 11.8. Because of the gating mechanism and the information stored in PIT, the proposed cognitive MAC protocol dynamically stops secondary users sending frames at the time when primary users access the channel. Surprisingly, the dropping rate of primary users in the legacy CSMA/CA MAC protocol is less than 3%. This phenomenon results from the carrier sensing that the node performs before sending packets. This carrier sensing satisfies the basic requirement of spectrum sensing in a cognitive network. However, the carrier sensing only detects the channel at the time instance when it executes instead

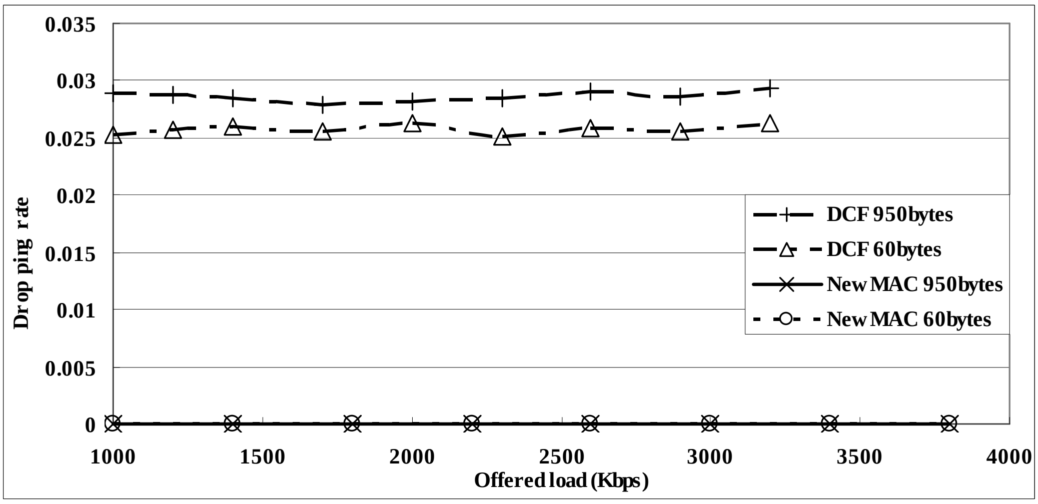


Figure 11.8: Dropping rate of the primary user's transmissions in the network topologies that all secondary users are located in one cluster.

of the whole duration for the frame transmission. The collision still happens when the primary user appears at the time during the secondary user transmission. In our proposed cognitive MAC protocol, the gating mechanism with the information in PIT ensures the entire duration of the secondary user transmission has no influence to primary users. Although the current simulation only considers the periodic traffic, the proposed method still can cooperate with other advance traffic models to estimate the time and length of primary user transmissions. Therefore, we believe the proposed scheme still provides valuable contribution for the cognitive MAC protocol design.

Figure 11.9 compares the normalized throughput of the proposed cognitive MAC protocol and the legacy CSMA/CA MAC protocol in the one-cluster network topology. For both the frame sizes of 60 and 950 bytes, the throughput in the proposed MAC protocol are 100% better than those in the CSMA/CA MAC protocol. The improvements mainly result from the invited reservation procedure, which can control the delay for the delay-

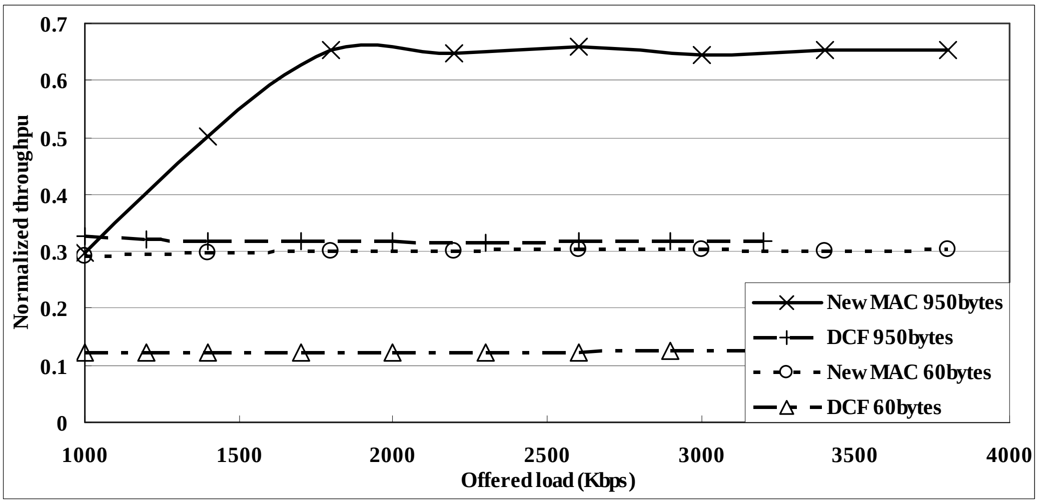


Figure 11.9: Throughput comparison of the proposed cognitive MAC protocol with the traditional CSMA/CA MAC protocol.

sensitive frames. Thus, most of the delay-sensitive frames can be counted for throughput computation. On the contrary, all the delay-sensitive frames in the legacy CSMA/CA MAC protocol still contend for the channel with other non-real-time frames. The contention leads the access delay of delay-sensitive frames is beyond the delay constraint due to retransmissions, and causes the neglect of those frame in throughput calculation. Furthermore, the gating mechanism in the proposed cognitive MAC protocol reduces the collisions in the CP. Hence, the throughput of the proposed MAC is better than that of the CSMA/CA MAC protocol in supporting mixed-type traffic flows.

Figures 11.10 and 11.11 compare the mean access delay and the fairness performances between the two MAC protocols. For the proposed cognitive MAC protocol, the mean access delay and its maximum standard deviation in sending the Telnet data frames is less than 0.2 sec and 1 sec, respectively. However, for the legacy CSMA/CA MAC protocol, the considered perfor-

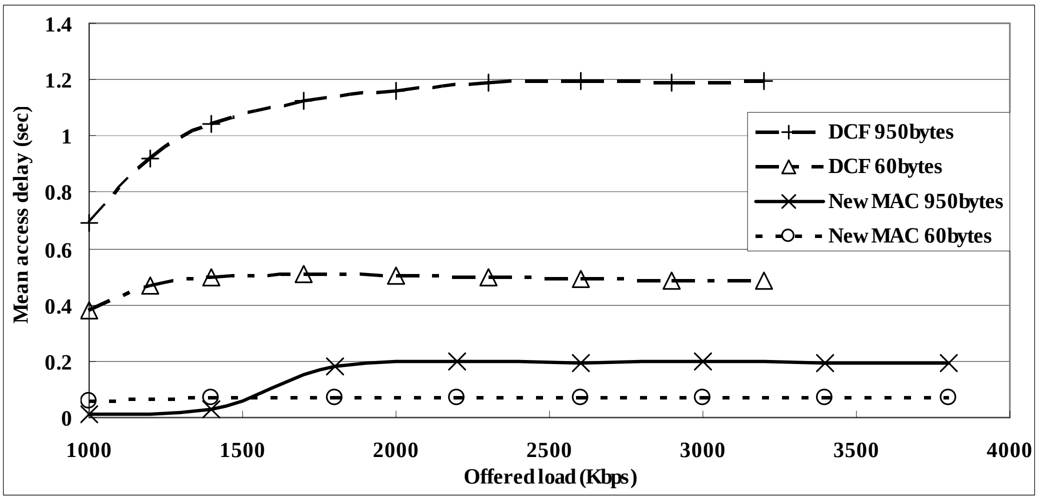


Figure 11.10: Comparison of mean access delay between the proposed cognitive MAC protocol and the CSMA/CA MAC protocol.

mance matrices are increased to 1.2 sec and 3.6 sec, respectively. The long access delay and its maximum standard deviation in the CSMA/CA MAC protocol is due to the long waiting time in the queue. However, when the non-real-time data frames are back-logged, the proposed stall avoidance scheme can effectively reduce the CW size of the stalled frames to expedite the transmissions. Therefore, the proposed cognitive MAC protocol can reduce the waiting time and improve the access delay and its standard deviation.

An interesting phenomenon shown in Figs.11.10 and 11.11 is that both the delay and its standard deviation become stable at high traffic load. Recall that the computation of access delay only counts the frame which the number of transmissions is less than the retry limit, i.e., seven in our simulation. Therefore, the limited number of frame transmission attempts for the calculation of access delay confines the values of its average and standard deviation even at high traffic load.

Figure 11.12 compares the dropping rate of delay-sensitive traffic for sec-

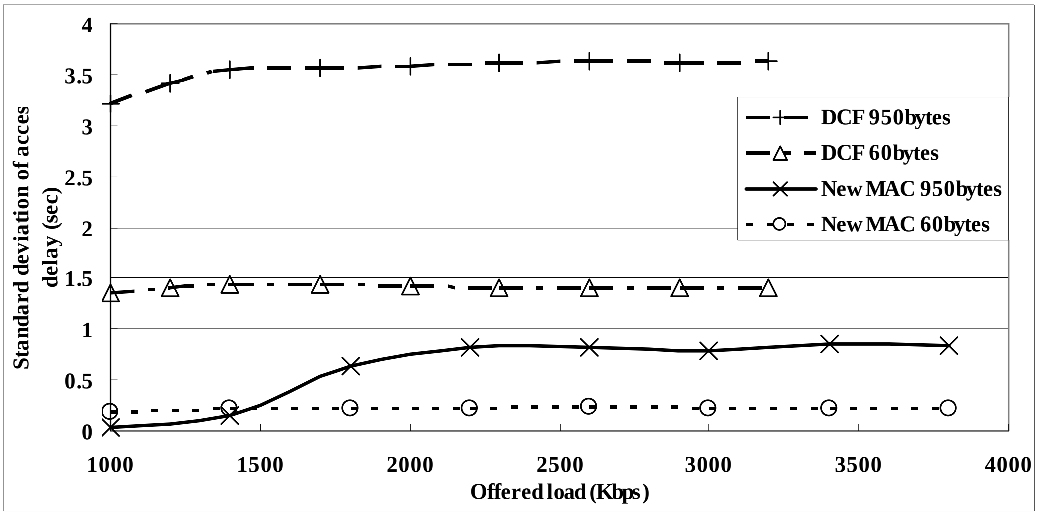


Figure 11.11: Comparison of fairness among the proposed cognitive MAC protocol and the CSMA/CA MAC protocol.

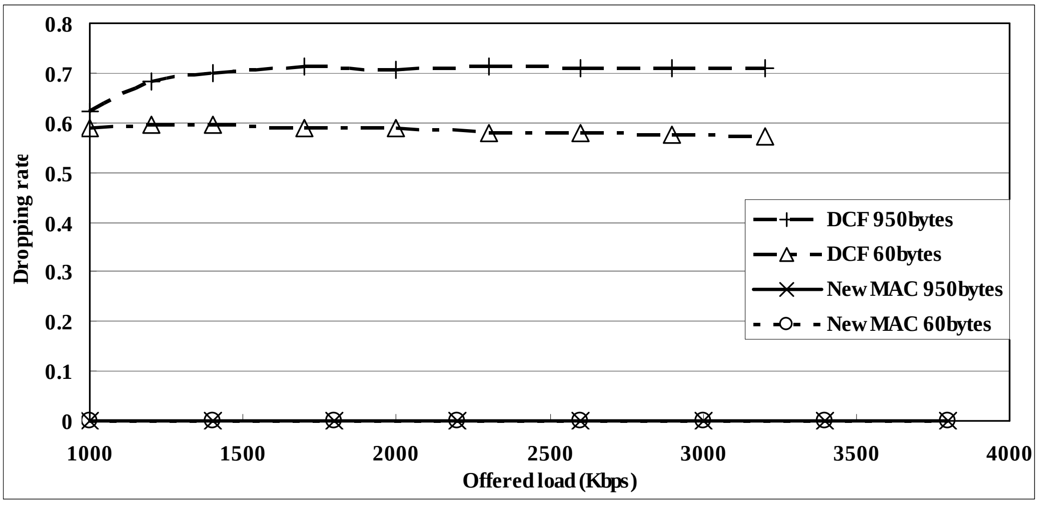


Figure 11.12: Dropping rate of delay-sensitive traffic for secondary users.

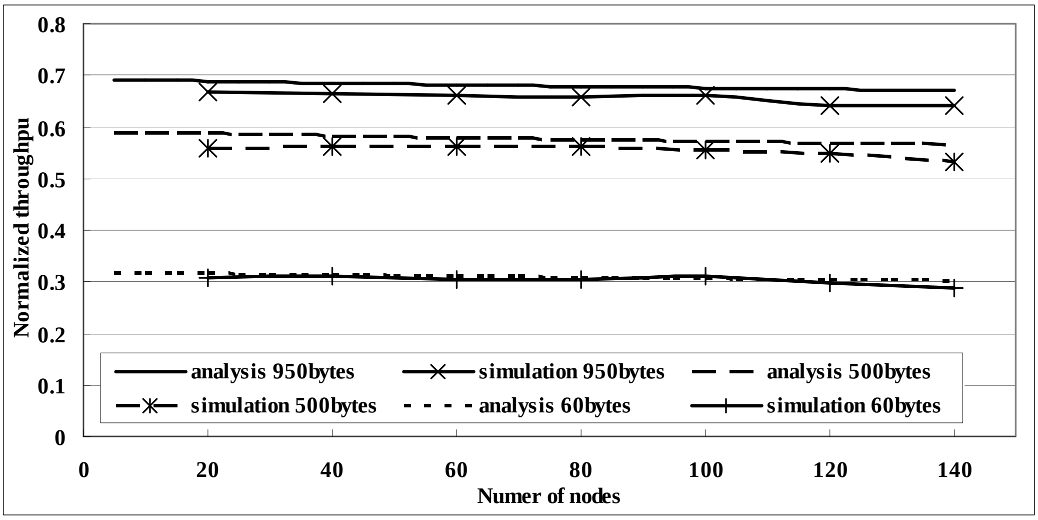


Figure 11.13: Throughput of the proposed cognitive MAC protocol in the mixed voice and FTP data traffic.

ondary users in two considered MAC protocols. As shown in the figure, the dropping rate in the proposed MAC protocol is lower than 0.1%, while in the legacy CSMA/CA MAC protocol the dropping rate can be higher than 50%. Because delay-sensitive frames in the legacy CSMA/CA MAC protocol contend for the channel with data frames using the similar priority, the retransmission due to the collisions causes the access delay beyond the delay constraint. By contrast, with the help of the invited reservation procedure, the proposed MAC protocol can guarantee the delay-sensitive frames to be received within the predefined constraint of 20 msec. Although, in the proposed MAC protocol, the first frame of delay-sensitive traffic flow still contends with other non-real-time data frames, the linear backoff algorithm helps to quickly establish the traffic flow and reserve the succeeding frame transmissions. Therefore, the dropping rate of the proposed MAC protocol is almost negligible compared to the CSMA/CA MAC protocol.

To validate our results, Fig. 11.13 shows the normalized throughput

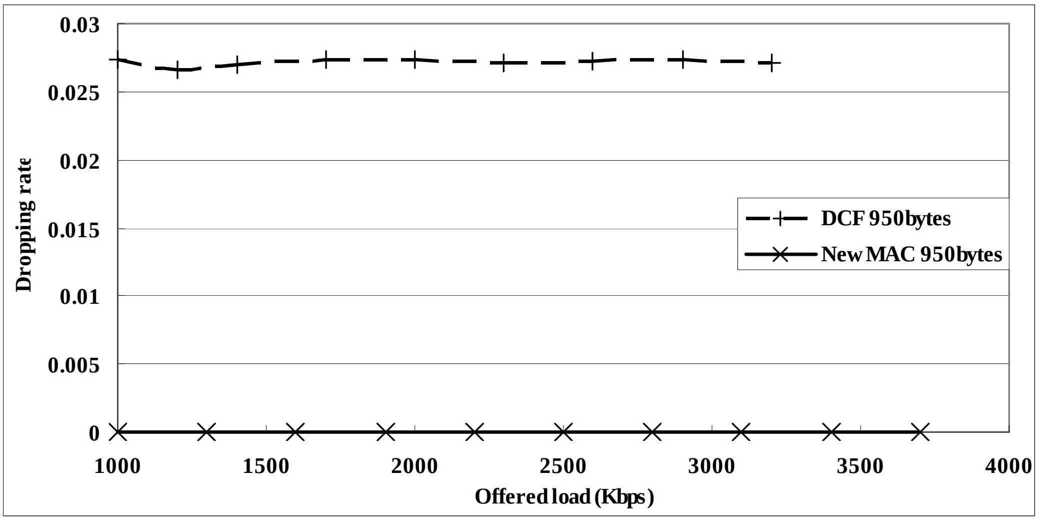


Figure 11.14: Dropping rate of primary user transmissions in the network topology that secondary users are distributed in three separated clusters.

with the mixed delay-sensitive and non-real-time data traffic in the proposed cognitive MAC protocol by simulations and analysis. The considered scenario includes 10 *rt-nodes* and various number of *nrt-nodes* establishing voice and FTP data traffic flows, respectively. As shown in the figure, the result from the analytical model by (11.4) are close that from simulations, especially for the case with a small packet size. Even for the large packet size, the discrepancy between analysis and simulation is still less than 3%. As shown in the figure, the throughput still remains stable even in the system with more than 100 *nrt-nodes*. This observation illustrates that the use of small CW size in the linear backoff algorithm and stall avoidance scheme do not deteriorate the network stability in a crowded environment.

At last, the network topology that all secondary users are distributed in three string-type clusters, as shown in Fig. 11.7, is considered. Figure 11.14 shows the dropping rate of primary user transmission. Similar to the observation shown in Fig. 11.8, the dropping rate in the proposed cogni-

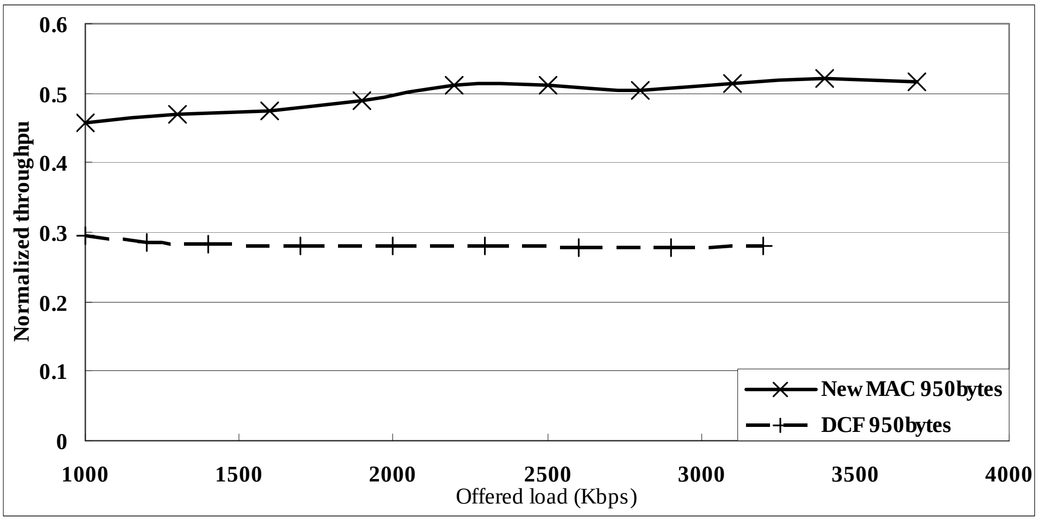


Figure 11.15: Comparison of throughput performance between the proposed cognitive MAC protocol and the CSMA/CA MAC protocol in the network topology that secondary users are distributed in three separated clusters.

tive MAC protocol is negligible compared to the legacy CSMA/CA MAC protocol. In addition, as shown in Figs. 11.15 and 11.16, the proposed cognitive MAC protocol for secondary user transmissions still outperforms the legacy CSMA/CA MAC protocol in terms of throughput and dropping rate. Like the RTS/CTS handshaking in the legacy CSMA/CA MAC protocol, broadcasting rt-CTS from the receiver of reserved flows in RIT reduces the collision in the reserved slot. Thus, the hidden node issue only cause a minor impact to the throughput and dropping rate in the proposed cognitive MAC protocol.

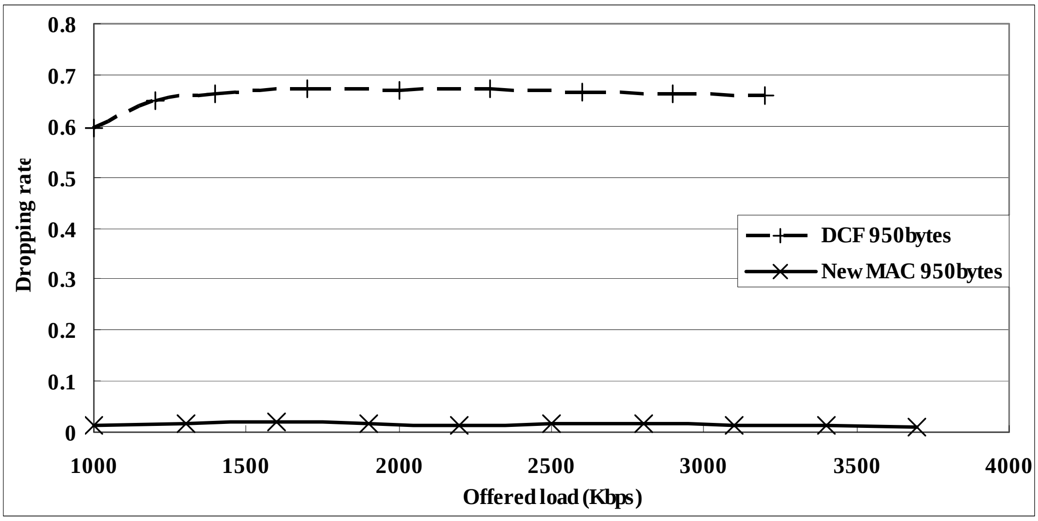


Figure 11.16: Comparison of dropping rate for secondary users between the proposed cognitive MAC protocol and the CSMA/CA MAC protocol in the network topology that secondary users are distributed in three separated clusters.

Chapter 12

Conclusions

In this report, we have developed analytical framework based on the preemptive resumption priority (PRP) M/G/1 queueing theory to characterize the general channel usage behaviors with multiple handoffs from a macroscopic viewpoint. Based on this model, we can evaluate the effects of multiple handoffs for the QoS performance of the secondary users, and then provide important insights into the designs of spectrum decision, spectrum mobility, and spectrum sharing algorithms. This report includes the following research topics:

1. Modeling techniques for cognitive radio networks;
2. Load-balancing spectrum decision;
3. Proactive spectrum handoff;
4. Optimal proactive spectrum handoff;
5. Reactive spectrum handoff;
6. Interference-avoiding spectrum sharing.

The contributions from this research are listed as follows.

1. Introduce a queueing-theoretical framework to characterize the effects of the general channel usage behaviors with multiple handoffs.
2. Design system parameters for load-balancing multiuser spectrum decision schemes to evenly distribute the traffic loads of secondary users to multiple channels.
3. Propose a traffic-adaptive spectrum handoff scheme, which changes the target channel sequence of spectrum handoffs based on traffic conditions.
4. Determine the optimal target channel sequence for the proactive spectrum handoff.
5. Provide a framework to determine whether the spectrum sensing technology can effectively shorten the extended data delivery time under various sensing time and traffic parameters.
6. Develop a cross-layer admission control rule for the secondary users.
7. Analyze the latency performance of various MAC protocols.
8. Propose a location-aware concurrent transmission MAC protocol.
9. Propose a neighbor-aware cognitive MAC protocol.

In the following, we summarize the results from the above contributions.

12.1 Modeling Techniques for Cognitive Radio Networks

In this part, the preemptive resume priority (PRP) M/G/1 queueing network model has been proposed to evaluate the QoS performances for the connection-based spectrum management techniques in the non-hopping and the hopping modes. This analytical framework provided a systematic viewpoint to integrate the designs of the spectrum management techniques and can help evaluate their QoS performances for various traffic arrival rates and service time distributions. There still exists many open problems for spectrum management techniques. On top of the proposed model, these open issues can be solved from the systematic viewpoint and then provide better traffic-adaptive solutions.

12.2 Load-Balancing Spectrum Decision

In this part, an analytical framework has been proposed to design the system parameters for the sensing-based and the probability-based spectrum decision schemes. The proposed model integrated with the PRP M/G/1 queueing systems can evaluate the effects of multiple interruptions from the primary connections and sensing errors (false alarm and missed detection) of the secondary connections on the overall system time for the two considered spectrum decision schemes. Based on this analytical model, the optimal number of candidate channels for the sensing-based spectrum selection method and the optimal channel selection probability for the probability-based spectrum selection method can be obtained analytically for various sensing time and traffic parameters. We found that the probability-based scheme can reduce the overall system time compared to the sensing-based scheme when the traffic loads of the secondary users is light, whereas the sensing-based scheme performs better in the condition of heavy traffic loads. This observation provide an important insight into design a traffic-adaptive spectrum decision scheme in the presence of sensing errors.

12.3 Proactive Spectrum Handoff

In this part, we have used the PRP M/G/1 queueing network model to characterize the spectrum usage behaviors with multiple handoffs. We studied the latency performance of the secondary connections by considering the effects of (1) generally distributed service time; (2) various operating channels; and (3) queueing behaviors of multiple secondary connections. The proposed model can accurately estimate the extended data delivery time of different proactively designed target channel sequences. On top of this model, we showed the extended data delivery time of the secondary connections based on the *always-staying* and the *always-changing* sequences in the IEEE 802.22 standard. If the secondary users can adaptively adopt the better target channel sequence according to traffic conditions, the extended data delivery time can be improved significantly compared to the existing target channel selection methods, especially for the heavy traffic loads of the primary users.

12.4 Optimal Proactive Spectrum Handoff

In this part, we have investigated the **Cumulative Handoff Delay Minimization Problem**. We formulated an optimization problem of determining a target channel sequence for multiple handoffs with the objective of minimizing the cumulative handoff delay for the newly arriving secondary user's connection. In order to solve this problem, we developed a state diagram to characterize the evolution of the target channel sequence. Based on this model, an optimal solution can be found by the proposed dynamic programming algorithm with time complexity of $O(LM^2)$. Furthermore, we suggested a suboptimal greedy strategy to select the target channels for spectrum handoffs with time complexity of $O(M)$. We proved that only six permutations of the target channel sequences are needed to be compared when the suggested greedy strategy is adopted. Numerical results show that the performance of this greedy strategy can approach the optimal solution.

12.5 Reactive Spectrum Handoff

In this part, we have investigated the effects of reactive spectrum handoff on the channel utilization and the extended data delivery time of the secondary users' connections by considering the three key design features for spectrum handoffs, consisting of (1) heterogeneous arrival rates of the primary users; (2) various arrival rates of the secondary users; (3) handoff processing time. Firstly, we propose a PRP M/G/1 queueing network model to characterize the spectrum usage behaviors between the primary and the secondary connections with multiple handoffs . Next, we develop a state diagram to characterize the effect of multiple handoff delay on the extended data delivery time of the secondary users' connections. Based on the proposed unifying model, an insightful study to quantify the effect of the three design features on the channel utilizations and the extended data delivery time under various traffic arrival rates and service time distributions can be provided. More importantly, these analytical results can facilitate the designs of admission control rule for the secondary users and can provide a framework to determine whether the spectrum sensing technology can effectively shorten the data delivery time under various sensing time.

12.6 Interference-Avoiding Spectrum Sharing

In this part, we proposed an adaptively arrival rate control mechanism by adjusting traffic admission probability of the secondary connections in order to maintain the interference constraint of the primary users and the latency requirement of the secondary users. Although a larger traffic admission probability for the secondary connections can increase channel utilization, it leads to more interference on the primary connections as well as more contention between the secondary connections. In order to find the best traffic admission probability, we formulate this issue as a cross-layer optimization problem by considering the effects of sensing errors and power outage in the physical layer, traffic admission probability in the MAC layer, and the traffic statistics as well as the QoS constraints in the application layer. The analytical results show the optimal traffic admission probabilities under various cross-layer parameters and provide important insight into the design tradeoffs between the system-level performance measure (channel utilization) and the user-level performance measures (interference ratio and transmission latency).

12.7 Latency Analysis for Spectrum Sharing

There are two major contributions in this thesis. First, we propose an analytical model to evaluate the per packet dynamic channel access latency for the CR MAC with dedicated and embedded control channels. The per packet access latency for DSA with dedicated and embedded control channels have not been seen in the literature. By using the analytical model, the optimal bandwidth ratio for the dedicated control channel case can be obtained and can know how system parameters affect the latency of the dedicated and embedded control channels. Second, we propose an analytical model to evaluate the total packets dynamic channel access latency for CR MAC with dedicated and embedded control channels. By jointly considering the per packet and total packets cases, a complete design guideline is proposed. In principle, one can know that the legacy IEEE 802.11 CSMA/CA MAC protocol is worth extending for the CR MAC by adding a channel search capability in some suggested conditions.

12.8 Location-Aware Concurrent Transmission for Spectrum Sharing

In this part, we identified a critical region R_{CT} in which the overlaying cognitive ad hoc users and the primary user can concurrently transmit data without causing interference to each other. If the location information of other nodes is available, such a concurrent transmission region can be easily identified. There are two major advantages of identifying the concurrent transmission opportunity. First, the overall throughput of the concurrently transmitted data obtained by combining both the overlaying cognitive ad hoc networks and the legacy infrastructure-based system is much higher than that of the pure infrastructure-based system. Our numerical results show that, in the uplink case, the concurrent transmission region subject to 1 dB and 6 dB shadowing standard deviation can be up to 45% out of the entire cell area with about 90% and 60% reliability, respectively. Second, if such a concurrent transmission opportunity can be identified first, it is clear that the need of the time- and energy-consuming wide-band spectrum scanning process required by most existing cognitive radio systems can be reduced dramatically.

12.9 Neighbor-Aware Cognitive MAC Protocol for Spectrum Sharing

In this part, we have proposed a cognitive MAC protocol to establish an overlaying cognitive ad hoc network with QoS provisioning in the presence of the legacy wireless systems. The proposed mechanisms can supplement the insufficiency of the legacy CSMA/CA MAC protocol to fulfill the goals of the cognitive wireless networks. With respect to the four stages in the cognition cycle, we suggest the following techniques:

- **Neighbor list establishment** in the ***observe*** stage: to help CR users having the knowledge of the spectrum usage by the primary and other CR users;

- **Improved contention resolution algorithm** in the ***plan*** stage: to prevent CR users from interfering the existing legacy system and to allow CR users to efficiently and fairly access the channel in the short spare time of the spectrum usage by primary users.

- **Invited reservation procedure** in the ***decide*** stage: to schedule the transmissions of delay-sensitive traffic with satisfactory QoS requirements for secondary user without interfering the legacy system and to dynamically allocate the bandwidth for various traffic types to avoid the starvation issue for low priority traffic.

- **Distributed frame synchronization** in the ***act*** stage: to distributively coordinate the frame transmissions among CR users.

Through the simulations by NS-2, we demonstrate that even in the environment with hidden nodes, the throughput performance of the proposed MAC protocol is at least 50% better than that of the legacy CSMA/CA MAC protocol. The mean access delay and its maximum standard deviation of the proposed MAC protocol are 5 times less than the CSMA/CA MAC protocol. At last, instead of more than 50 % for the legacy CSMA/CA MAC protocol, the dropping rate of delay-sensitive traffic for the proposed MAC protocol is almost negligible.

12.10 Suggestions for Future Research

The proposed PRP M/G/1 queueing network model provides a systematic method to help the design of spectrum management technologies. It can capture the general behaviors for the connection-based channel usage, including the effects of channel selection, spectrum sensing time, multiple interruptions, channel switching between different channels, and generally distributed service time simultaneously. The contributions of this report are summarized in Table 12.1, where the signs " ∘ " and " × " indicate that the issue "has" and "has not" been discussed, respectively. Specifically, we have investigated the spectrum decision, mobility, and sharing issues in the non-hopping mode as well as the spectrum mobility issue in the hopping mode. The spectrum decision and sharing issues in the hopping model are still needed to be solved.

Some interesting research issues that can be extended from the proposed model include the following:

1. For the spectrum sensing issue, an interesting issue is to consider the se-

Table 12.1: Summary of This Report

	Spectrum Management Techniques with Multiple Handoffs	
	Non-hopping Model	Hopping Mode
Spectrum Decision (Chapter 4)	○	×
Spectrum Mobility (Chapters 5-7)	○	○
Spectrum Sharing (Chapter 8)	○	×

quential sensing. For example, in the sensing-based spectrum decision scheme in Chapter 4, we assume that the secondary user performs wideband sensing to find the idle channel when a new connection request arrives at CR network. In fact, the secondary user can perform sequential sensing to find the idle channel. In this case, the sensing procedures can be terminated once one idle channel is found, and thus the sensing time can be shorter than $n\tau$. Furthermore, in Chapters 4 and 8, we assume that the class-A missed detection and false alarm probabilities are constant. In fact, as transmission time increases, missed detection and false alarm probabilities can be reduced because the sensing results at the previous sensing phases can be employed to improve the accuracy of the sensing results at the following sensing phases. Hence, it is also worthwhile to investigate the effects of variable missed detection and false alarm probabilities in Chapters 4 and 8. Besides, the effects of the class-B missed detection is another important research issue.

2. From the viewpoint of *spectrum decision* issue, it is worthwhile determining the optimal distribution probability vector for the probability-based spectrum decision method when the secondary connections may have different opinions on the observed traffic statistics $\lambda_p^{(k)}$, λ_s, $f_p^{(k)}(x)$,

and $f_s(x)$. Second, it would be interesting to see how to analyze the latency performance for the spectrum decision methods in the hopping mode.

3. The proposed model assumes that the *spectrum mobility* functionality can help the interrupted secondary user resume its unfinished data transmission on the suitable channel. This resumption policy can be characterized by the preemptive resumption priority queueing network. However, in other scenarios, the interrupted secondary user may need to retransmit the whole connection rather than resuming the unfinished transmission. In this situation, a CR network should be modeled by the preemptive repeat priority queueing network. It is also worthwhile to investigate the latency performances resulted from different transmission policies.

4. For the *spectrum sharing* issue, an interesting issue is to involve the distributed channel contention behaviors into the proposed model. In the proposed model, we assume that the FCFS scheduling policy is adopted. For a distributed medium access control (MAC) protocol such as the carrier sense multiple access (CSMA) protocol in Chapter 11, the channel contention time and retransmission in the MAC layer should be taken into account when calculating the latency performance of the secondary users.

Cognitive radio (CR) is an emerging technique to promote spectrum efficiency. Through the CR technique, we believe that the dream of freely connecting people anywhere anytime is no longer an impossible mission but will be come true in the near future. After all, the development of technology is to satisfy the need that people want.

Bibliography

[1] R. W. Brodersen, A. Wolisz, D. Cabric, S. M. Mishra, and D. Willkomm, “CORVUS: A Cognitive Radio Approach for Usage of Virtual Unlicensed Spectrum,” *Berkeley Wireless Research Center (BWRC) White paper*, 2004.

[2] J. Mitola and G. Q. Maguire, “Cognitive Radio: Making Software Radios More Personal,” *IEEE Personal Communications*, vol. 6, pp. 13–18, Aug. 1999.

[3] S. Haykin, “Cognitive Radio: Brain-empowered Wireless Communications,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[4] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie, “Cognitive Networks,” *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Nov. 2005.

[5] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, “NeXt Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey,” *Computer Networks Journal (Elsevier)*, vol. 50, pp. 2127–2159, Sep. 2006.

[6] P.-Y. Huang and K.-C. Chen, "A Cognitive CSMA-based Multichannel MAC Protocol for Cognitive Radio Networks," *APSIPA Annual Summit and Conference (ASC)*, Oct. 2009.

[7] D. Shiung and K.-C. Chen, "On the Optimal Power Allocation of a Cognitive Radio Networks," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2009.

[8] S.-Y. Lien, C.-C. Tseng, and K.-C. Chen, "Carrier Sensing based Multiple Access Protocols for Cognitive Radio Networks," *IEEE International Conference on Communications (ICC)*, May 2008.

[9] O. Jo and D.-H. Cho, "Seamless Spectrum Handover Considering Differential Path-loss in Cognitive Radio Systems," *IEEE Communications Letters*, vol. 13, no. 3, pp. 190–192, Mar. 2009.

[10] M. D. Silvius, R. Rangnekar, A. B. MacKenzie, and C. W. Bostian, "The Smart Radio Channel Change Protocol: A Primary User Avoidance Technique for Dynamic Spectrum Sharing Cognitive Radios to Facilitate Co-existence in Wireless Communication Networks," *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, Jun. 2009.

[11] A. Sgora and D. D. Vergados, "Handoff Prioritization and Decision Schemes in Wireless Cellular Networks: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 4, pp. 57–77, 2009.

[12] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "CRAHNs: Cognitive Radio Ad Hoc Networks," *Ad Hoc Networks*, vol. 7, no. 5, pp. 810–836, 2009.

[13] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "A Survey on Spectrum Management in Cognitive Radio Networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, Apr. 2008.

[14] Q. Zhao, L. Tong, and A. Swami, "A cross-layer approach to cogintive MAC for spectrum agility," *Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 200–204, Nov. 2005.

[15] S. Krishnamurthy, M. Thoppian, S. Venkatesan, and R. Prakash, "Control channel based MAC-layer configuration, routing and situation awareness for radio networks," *Proceedings of IEEE Military Communications*, pp. 1–6, Oct. 2005.

[16] *IEEE Std. 802.11, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specification*, Nov. 1999.

[17] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, pp. 13–18, Aug. 1999.

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

[19] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie, "Cognitive networks," *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 352–360, Nov. 2005.

[20] Q. Zhao and A. Swami, "A Decision-theoretic Framework for Opportunistic Spectrum Access," *IEEE Wireless Communications Magazine*, vol. 14, no. 4, pp. 14–20, Aug. 2007.

[21] X. Zhu, L. Shen, and T.-S. P. Yum, "Analysis of Cognitive Radio Spectrum Access with Optimal Channel Reservation," *IEEE Communications Letters*, vol. 11, no. 4, pp. 304–306, Apr. 2007.

[22] C. Zhang, X. Wang, and J. Li, "Cooperative Cognitive Radio with Priority Queueing Analysis," *IEEE International Conference on Communications (ICC)*, Jun. 2009.

[23] H. Tran, T. Q. Duong, and H.-J. Zepernick, "Average Waiting Time of Packets with Different Priorities in Cognitive Radio Networks," *IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, May 2010.

[24] Y. Zhu, Q. Zhang, Z. Niu, and J. Zhu, "On Optimal QoS-aware Physical Carrier Sensing for IEEE 802.11 Based WLANs: Theoretical Analysis and Protocol Design," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1369–1378, April 2008.

[25] A. Banaei and C. N. Georghiades, "Throughput Analysis of a Randomized Sensing Scheme in Cell-based Ad-hoc Cognitive Networks," *IEEE International Conference on Communications (ICC)*, Jun. 2009.

[26] J. Gambini, O. Simeone, Y. Bar-Ness, U. Spagnolini, and T. Yu, "Packet-wise Vertical Handover for Unlicensed Multi-standard Spectrum Access with Cognitive Radios," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5172–5176, Dec. 2008.

[27] J. Gambini, O. Simeone, U. Spagnolini, Y. Bar-Ness, and Y. Kim, "Cognitive Radio with Secondary Packet-By-Packet Vertical Handover," *IEEE International Conference on Communications (ICC)*, May 2008.

[28] G. Noh, J. Lee, and D. Hong, “Stochastic Multichannel Sensing for Cognitive Radio Systems: Optimal Channel Selection for Sensing with Interference Constraints,” *IEEE Vehicular Technology Conference Fall*, Sep. 2009.

[29] C.-M. Lee, J.-S. Lin, Y.-P. Hsu, and K.-T. Feng, “Design and Analysis of Optimal Channel-hopping Sequence for Cognitive Radio Networks,” *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2010.

[30] A. T. Chronopoulos, M. R. Musku, S. Penmatsa, and D. C. Popescu, “Spectrum Load Balancing for Medium Access in Cognitive Radio Systems,” *IEEE Communications Letters*, vol. 12, no. 5, pp. 353–355, May 2008.

[31] I. Malanchini, M. Cesana, and N. Gatti, “On Spectrum Selection Games in Cognitive Radio Networks,” *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2009.

[32] H.-P. Shiang and M. van der Schaar, “Queuing-based Dynamic Channel Selection for Heterogeneous Multimedia Applications Over Cognitive Radio Networks,” *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 896–909, Aug. 2008.

[33] Y. Song, Y. Fang, and Y. Zhang, “Stochastic Channel Selection in Cognitive Radio Networks,” *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2007.

[34] H.-J. Liu, S.-F. Li, Z.-X. Wang, W.-J. Hong, and M. Yi, “Strategy of Dynamic Spectrum Access Based-on Spectrum Pool,” *IEEE Interna-*

tional Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Sep. 2008.

[35] A. W. Min and K. G. Shin, "Exploiting Multi-channel Diversity in Spectrum-agile Networks," *IEEE International Conference on Computer Communications (INFOCOM)*, Apr. 2008.

[36] B. Hamdaoui, "Adaptive Spectrum Assessment for Opportunistic Access in Cognitive Radio Networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 922–930, Feb. 2009.

[37] A. Sabharwal, A. Khoshnevis, and E. Knightly, "Opportunistic Spectral Usage: Bounds and a Multi-band CSMA/CA Protocol," *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 533–545, 2007.

[38] J. Jia, Q. Zhang, and X. S. Shen, "HC-MAC: A Hardware-constrained Cognitive MAC for Efficient Spectrum Management," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 106–117, Jan 2008.

[39] M. Felegyhazi and J.-P. Hubaux, "Game Theory in Wireless Networks: A Tutorial," *EPFL Technical Report: LCA-REPORT-2006-002*, 2006.

[40] 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*, vol. 25, no. 3, pp. 589–600, April 2007.

[41] Q. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic Spectrum Access via Periodic Channel Sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 785–796, Feb. 2008.

[42] O. Mehanna, A. Sultan, and H. E. Gamal, "Blind Cognitive MAC Protocols," *IEEE International Conference on Communications (ICC)*, Jun. 2009.

[43] S. Senthuran, A. Anpalagan, and O. Das, "A Predictive Opportunistic Access Scheme for Cognitive Radios," *IEEE Vehicular Technology Conference Fall*, Sep. 2009.

[44] L. Yang, L. Cao, and H. Zheng, "Proactive Channel Access in Dynamic Spectrum Networks," *Physical Communication*, vol. 1, no. 2, pp. 103–111, 2008.

[45] R.-T. Ma, Y.-P. Hsu, and K.-T. Feng, "A POMDP-based Spectrum Handoff Protocol for Partially Observable Cognitive Radio Networks," *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2009.

[46] M. Hoyhtya, S. Pollin, and A. Mammela, "Performance Improvement with Predictive Channel Selection for Cognitive Radios," *IEEE International Workshop on Cognitive Radio and Advanced Spectrum Management (CogART)*, Feb. 2008.

[47] S.-U. Yoon and E. Ekici, "Voluntary Spectrum Handoff: A Novel Approach to Spectrum Management in CRNs," *IEEE International Conference on Communications (ICC)*, May 2010.

[48] Y. Song and J. Xie, "Common Hopping Based Proactive Spectrum Handoff in Cognitive Radio Ad Hoc Networks," *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2010.

[49] L.-C. Wang and A. Chen, "On the Performance of Spectrum Handoff for Link Maintenance in Cognitive Radio," *IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, May 2008.

[50] B. Wang, Z. Ji, K. J. Ray Liu, and T. C. Clancy, "Primary-prioritized Markov Approach for Dynamic Spectrum Allocation," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 1854–1865, Apr. 2009.

[51] I. Suliman and J. Lehtomaki, "Queueing Analysis of Opportunistic Access in Cognitive Radios," *IEEE International Workshop on Cognitive Radio and Advanced Spectrum Management (CogART)*, May 2009.

[52] H. Li, "Queuing Analysis of Dynamic Spectrum Access Subject to Interruptions from Primary Users," *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, Jun. 2010.

[53] P. Zhu, J. Li, and X. Wang, "A New Channel Parameter for Cognitive Radio," *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, Aug. 2007.

[54] S. Wang and H. Zheng, "A Resource Management Design for Cognitive Radio Ad Hoc Networks," *IEEE Military Communications Conference (MILCOM)*, Oct. 2009.

[55] F. Borgonovo, M. Cesana, and L. Fratta, "Throughput and Delay Bounds for Cognitive Transmissions," *Advances in Ad Hoc Networking*, vol. 265, pp. 179–190, Aug. 2008.

[56] C.-T. Chou, Sai Shankar N, H. Kim, and K. G. Shin, "What and How Much to Gain from Spectral Agility?" *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 576–588, Apr. 2007.

[57] J. Heo, J. Shin, J. Nam, Y. Lee, J. G. Park, and H.-S. Cho, "Mathematical Analysis of Secondary User Traffic in Cognitive Radio System," *IEEE Vehicular Technology Conference Fall*, Sep. 2008.

[58] Y. Zhang, "Spectrum Handoff in Cognitive Radio Networks: Opportunistic and Negotiated Situations," *IEEE International Conference on Communications (ICC)*, Jun. 2009.

[59] F. Capar, I. Martoyo, T. Weiss, and F. Jondral, "Comparison of Bandwidth Utilization for Controlled and Uncontrolled Channel Assignment in a Spectrum Pooling System," *IEEE Vehicular Technology Conference Spring*, May 2002.

[60] B. Ishibashi, N. Bouabdallah, and R. Boutaba, "QoS Performance Analysis of Cognitive Radio-based Virtual Wireless Networks," *IEEE International Conference on Computer Communications (INFOCOM)*, Apr. 2008.

[61] D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Optimal Admission Control in Cognitive Radio Networks," *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, Jun. 2009.

[62] W. Ahmed, J. Gao, and M. Faulkner, "Performance Evaluation of a Cognitive Radio Network with Exponential and Truncated Usage Models," *IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, Oct. 2009.

[63] M. Huang, R. Yu, and Y. Zhang, "Call Admission Control with Soft-QoS Based Spectrum Handoff in Cognitive Radio Networks," *Interna-*

tional Conference on Wireless Communications and Mobile Computing (IWCMC), Jun. 2009.

[64] I. Suliman, J. Lehtomaki, T. Braysy, and K. Umebayashi, “Analysis of Cognitive Radio Networks with Imperfect Sensing,” *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2009.

[65] S. Tang and B. L. Mark, “Performance Analysis of a Wireless Network with Opportunistic Spectrum Sharing,” *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2007.

[66] ——, “Modeling and Analysis of Opportunistic Spectrum Sharing with Unreliable Spectrum Sensing,” *IEEE Transactions on Wireless Communications*, vol. 8, pp. 1934–1943, Apr. 2009.

[67] ——, “Modeling an Opportunistic Spectrum Sharing System with a Correlated Arrival Process,” *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2008.

[68] ——, “An Analytical Performance Model of Opportunistic Spectrum Access in a Military Environment,” *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2008.

[69] E. W. M. Wong and C. H. Foh, “Analysis of Cognitive Radio Spectrum Access with Finite User Population,” *IEEE Communications Letters*, vol. 13, no. 5, pp. 294–296, May 2009.

[70] M. M. Rashid, M. J. Hossain, E. Hossain, and V. K. Bhargava, “Opportunistic Spectrum Access in Cognitive Radio Networks: A Queueing Analytic Model and Admission Controller Design,” *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2007.

[71] ——, "Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5259–5269, Oct. 2009.

[72] Y. Zhang, "Dynamic Spectrum Access in Cognitive Radio Wireless Networks," *IEEE International Conference on Communications (ICC)*, May 2008.

[73] Sai Shankar N, "Squeezing the Most Out of Cognitive Radio: A Joint MAC/PHY Perspective," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, Apr. 2007.

[74] Sai Shankar N, C.-T. Chou, K. Challapali, and S. Mangold, "Spectrum Agile Radio: Capacity and QoS Implications of Dynamic Spectrum Assignment," *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2005.

[75] I. Suliman and J. Lehtomaki, "Optimizing Detection Parameters for Time-slotted Cognitive Radios," *IEEE Vehicular Technology Conference Spring*, Apr. 2009.

[76] S.-L. Wu, C.-Y. Lin, Y.-C. Tseng, and J.-P. Sheu, "A New Multi-Channel MAC Protocol with On-Demand Channel Assignment for Multi-Hop Ad Hoc Networks," *In Proceedings of International Symposium on Parallel Architectures, Algorithms and Networks*, pp. 232–237, 2000.

[77] J. Chen and S. T. Sheu, "Distributed Multichannel MAC Protocol for IEEE 802.11 Ad Hoc Wireless LANs," *Computer Communications*, pp. 1000–1013, Dec 2004.

[78] W. C. Hung, K. Law, and A. Leon-Garcia, "A Dynamic Multi-Channel MAC for Ad Hoc LAN," *21st Biennial Symposium on Communications*, Apr 2002.

[79] N. Jain, S. R. Das, and A. Nasipuri, "A Multichannel CSMA MAC Protocol with Receiver-Based Channel Selection for Multihop Wireless Networks," *Computer Communications and Networks*, pp. 15–17, Oct 2001.

[80] C. Y. Chang, H. C. Sun, and C. C. Hsieh, "MCDA: A Efficient Multichannel MAC Protocol for 802.11 Wireless LAN with Directional Antenna," *19th International Conference on Advanced Information Networking and Applications*, pp. 64–67, Mar 2005.

[81] H. Koubaa, "Fairness-Enhanced Multiple Control Channels MAC for Ad Hoc Networks," *IEEE Vehicular Technology Conference*, pp. 1504–1508, Jun 2005.

[82] J. So and N. Vaidya, "Multi-Channel MAC for Ad Hoc Networks: Handling Multi-Channel Hidden Terminals Using A Single Transceiver," *In Proceedings of MobiHoc '04*, pp. 222–233, May 2004.

[83] P. Tan and M. C. Chan, "AMCM: Adaptive Multi-Channel MAC Protocol for IEEE 802.11 Wireless Networks," *Third International Conference on Broadband Communications, Networks, and Systems*, Oct 2006.

[84] J. H. Chen and Y. D. Chen, "AMNP: Ad Hoc Multichannel Negotiation Protocol for Multihop Mobile Wireless Networks," *IEEE International Conference on Communications*, pp. 3607–3612, Jun 2004.

[85] J. Chen, S. T. Sheu, and C. A. Yang, “A New Multi-Channel Access Protocol for IEEE 802.11 Ad Hoc Wireless LANs,” *IEEE Personal, Indoor and Mobile Radio Communications*, pp. 7–10, Sept 2003.

[86] S.-C. Lo and C.-W. Tseng, “A Novel Multi-Channel MAC Protocol for Wireless Ad Hoc Networks,” *IEEE Vehicular Technology Conference*, pp. 46–50, 2007.

[87] P. Bahl, R. Chandra, and J. Dunagan, “SSCH: Slotted Seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad Hoc Wireless Networks,” *The ACM Annual International Conference on Mobile Computing and Networking*, pp. 216–230, Sept 2004.

[88] J. N. M and I. T. L., “Spread Spectrum Medium Access Control with Collision Avoidance Protocol for Multichannel Network,” *Proceedings of the IEEE INFOCOM* , pp. 776–783, 1999.

[89] Z. Tang and J. J. Garcia-Luna-Aceves, “Hop-Reservation Multiple Access (HRMA) for Ad Hoc Networks,” *Proceedings of the IEEE INFOCOM* , pp. 194–201, Mar 1999.

[90] L. Ma, X. Han, and C. C. Shen, “Dynamic Open Spectrum Sharing MAC Protocol for Wireless Ad Hoc Networks,” *In proceedings of IEEE DySPAN 2005* , pp. 203–213, Nov 2005.

[91] P. Pawelczak, R. V. Prasad, L. Xia, and I. G. M. M. Niemegeers, “Cognitive Radio Emergency Networks - Requirments and Design,” *In proceedings of IEEE DySPAN 2005* , pp. 601–606, Nov 2005.

[92] S. Krishnamurthy, M. Thoppian, S. Venkatesan, and R. Prakash, “Control Channel Based MAC-Layer Configuration, Routing and

Situation Awareness for Cognitive Radio Networks," *Proceedings of IEEE Military Communications*, pp. 455–460, Oct 2005.

[93] R. Maheshwari, H. Gupta, and S. R. Das, "Multichannel MAC Protocols for Wireless Networks," *Sensor and Ad Hoc Communications and Networks*, pp. 393–401, 2006.

[94] 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*, pp. 589–600, Apr 2007.

[95] A. Tzamaloukas and J. J. Garcia-Luna-Aceves, "Channel-Hopping Multiple Access," *IEEE International Conference on Communications*, pp. 415–419, 2000.

[96] B. Alawieh, Y. Zhang, C. Assi, and H. Mouftah, "Improving Spatial Reuse in Multihop Wireless Networks," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 3, pp. 71–91, Third Quarter 2009.

[97] C. R. Lin and C.-Y. Liu, "Enhancing the Performance of IEEE 802.11 Wireless LAN by Using a Distributed Cycle Stealing Mechanism," *4th International Workshop on Mobile and Wireless Communications Network*, pp. 564–568, 2002.

[98] F. Wang, M. Krunz, and S. Cui, "Spectrum Sharing in Cognitive Radio Networks," *IEEE International Conference on Computer Communications (INFOCOM)*, Apr. 2008.

[99] Y. Song and J. Xie, "Optimal Power Control for Concurrent Transmissions of Location-aware Mobile Cognitive Radio Ad Hoc Networks," *IEEE Global Communications Conference (GLOBECOM)*, Nov. 2009.

[100] L.-C. Wang and A. Chen, “Effect of Location Awareness on Concurrent Transmissions for Cognitive Ad Hoc Networks Overlaying Infrastructure-based Systems,” *IEEE Transactions on Mobile Computing*, vol. 8, no. 5, pp. 577–589, May 2009.

[101] J. Zhu, X. Guo, L. L. Yang, W. S. Conner, S. Roy, and M. M. Hazra, “Adapting Physical Carrier Sensing to Maximize Spatial Reuse in 802.11 Mesh Networks,” *Wiley Wireless Communications and Mobile Computing*, vol. 4, no. 8, pp. 933–946, Dec. 2004.

[102] N. A. M. Maung, T. Noguchi, and M. Kawai, “Maximizing Aggregate Throughput of Wireless Ad Hoc Networks Using Enhanced Physical Carrier Sensing,” *International Conference on Distributed Computing Systems Workshops*, June 2008.

[103] G. Holland, N. Vaidya, and P. Bahl, “A Rate-adaptive MAC Protocol for Multi-hop Wireless Networks,” *ACM International Conference on Mobile Computing and Networking*, pp. 236–251, July 2001.

[104] T.-S. Kim, H. Lim, and J. C. Hou, “Understanding and Improving the Spatial Reuse in Multihop Wireless Networks,” *IEEE Transactions on Mobile Computing*, vol. 7, no. 10, pp. 1200–1212, Oct. 2008.

[105] M. Park, S.-H. Choi, and S. M. Nettles, “Cross-layer MAC Design for Wireless Networks Using MIMO,” *IEEE Global Communications Conference (GLOBECOM)*, vol. 5, 2005.

[106] H. Dai, K. W. Ng, and M. Y. Wu, “An Overview of MAC Protocols with Directional Antennas in Wireless Ad hoc Networks,” *IEEE International Multi-Conference on Computing in the Global Information Technology*, pp. 84–90, July 2006.

[107] C. Zhu and M. S. Corson, “A five-phase reservation protocol for mobile ad hoc networks,” *INFOCOM, IEEE International Conference on Computer Communications*, vol. 1, pp. 322–331, Mar. 1998.

[108] J. J. Garcia-Luna-Aceves and A. Tzamaloukas, “Receiver-initiated collision avoidance in wireless networks,” *ACM Wireless Networks*, vol. 8, no. 2-3, pp. 249–263, Mar.-May 2002.

[109] J. L. Sobrinho and A. S. Krishnakumar, “Quality-of-service in ad hoc carrier sense multiple access networks,” *IEEE Journal on Selected Areas in Communication*, vol. 17, no. 8, pp. 1353–1368, Aug. 1999.

[110] *IEEE Std. 802.11e/D3.2, “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Medium Access Control(MAC) Enhancement for Quality of Service (QoS)”*, Aug. 2002.

[111] A. Veres, A. T. Campbell, M. Barry, and L. Sun, “Supporting service differentiation in wireless packet networks using distributed control,” *IEEE Journal on Selected Areas in Communication*, vol. 19, no. 10, pp. 2081–2093, Oct. 2001.

[112] F. Calì, M. Conti, and E. Gregori, “IEEE 802.11 protocol: design and performance evaluation of an adaptive backoff mechanism,” *IEEE Journal on Selected Areas in Communication*, vol. 18, no. 19, pp. 1774–1786, Sep. 2000.

[113] ——, “Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit,” *IEEE/ACM Transcations on Networking*, vol. 8, no. 6, pp. 785–799, Dec. 2000.

[114] A. Lindgren, A. Almquist, and O. Schelen, “Evaluation of quality of service scheme for IEEE 802.11 wireless LANs,” *in Proceedings of*

the 26th Annual IEEE Conference on Local Computer Networks (LCN 2001), pp. 384–391, Nov. 2001.

[115] T. Nandagopal, T. E. Kim, X. Gao, and V. Bharghavan, "Achieving MAC layer fairness in wireless packet networks," *The ACM Annual International Conference on Mobile Computing and Networking*, pp. 87–98, Aug. 2000.

[116] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed fair scheduling in a wireless LAN," *The ACM Annual International Conference on Mobile Computing and Networking*, pp. 167–178, Aug. 2000.

[117] Y. Kwon, Y. Fang, and H. Latchman, "Fast collision resolution (FCR) MAC algorithm for wireless local area network," *IEEE Global Telecommunications Conference*, vol. 3, pp. 2250–2254, Nov. 2002.

[118] H. Zhu, M. Li, I. Chlamtac, and B. Prabhakaran, "A survey of quality of service in IEEE 802.11 networks," *IEEE Wireless Communications*, vol. 11, pp. 6–14, Aug. 2004.

[119] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Transactions on Communications*, vol. 37, no. 8, pp. 885–890, Aug. 1989.

[120] C. R. Lin and M. Gerla, "Asynchronous multimedia multihop wireless networks," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 1, pp. 118–125, Apr. 1997.

[121] S. Elnoubi and A. M. Alsayh, "A packet reservation multiple access (PRMA)-based algorithm for multimedia wireless system," *IEEE*

Transactions on Vehicular Technology, vol. 53, no. 1, pp. 215–222, Jan. 2004.

[122] S. T. Sheu and T. F. Sheu, "DBASE: a distributed bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LAN," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 3, pp. 1558–1567, Apr. 2001.

[123] S. Jiang, J. Rao, D. He, X. Ling, and C. C. Ko, "A simple distributed PRMA for MANETs," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 293–304, Mar. 2002.

[124] A. Chen, L. C. Wang, Y. T. Su, Y. X. Zheng, B. Yang, D. S. Wei, and K. Naik, "NICE-a dencentralized medium access control using neighborhood information classification and estimation for multimedia application in ad hoc 802.11 wireless LANs," *IEEE International Conference on Communications*, vol. 1, pp. 208–211, May 2003.

[125] A. Chen, L. C. Wang, C. W. Wang, and D. S. L. Wei, "NICER - a distributed wireless mac protocol for mobile ad hoc networks," *IEEE Vehicular Technology Conference*, vol. 3, pp. 1448–1452, Oct. 2003.

[126] W. Hu, D. Willkomm, G. Vlantis, M. Gerla, and A. Wolisz, "Dynamic Frequency Hopping Communities for Efficient IEEE 802.22 Operation," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 80–87, May 2007.

[127] H.-J. Liu, Z.-X. Wang, S.-F. Li, and M. Yi, "Study on the Performance of Spectrum Mobility in Cognitive Wireless Network," *IEEE Singapore International Conference on Communication Systems (ICCS)*, Jun. 2008.

[128] H. Su and X. Zhang, "Channel-hopping Based Single Transceiver MAC for Cognitive Radio Networks," *IEEE Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2008.

[129] Y.-C. Liang, Y. Zeng, E. C. Peh, and A. T. Hoang, "Sensing-throughput Tradeoff for Cognitive Radio Networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, Apr. 2008.

[130] P. Wang, L. Xiao, S. Zhou, and J. Wang, "Optimization of Detection Time for Channel Efficiency in Cognitive Radio Systems," *IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2007.

[131] W.-Y. Lee and I. F. Akyildiz, "Optimal Spectrum Sensing Framework for Cognitive Radio Networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3845–3857, Oct. 2008.

[132] C. R. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. J. Shellhammer, and W. Caldwell, "IEEE 802.22: The First Cognitive Radio Wireless Regional Area Network Standard," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 130–138, Jan. 2009.

[133] IEEE 802.11, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, supplement to IEEE 802.11 Standard, Sept. 1999.

[134] L. Kleinrock, *Queueing Systems - Volume 2: Computer Applications.* John Wiley & Sons Inc., 1975.

[135] L.-C. Wang, C.-W. Wang, and F. Adachi, "Load-balancing Spectrum Decision for Cognitive Radio Networks," *accepted by IEEE Journal on Selected Areas in Communications (JSAC)*, 2010.

[136] X. Liu and Z. Ding, "ESCAPE: A Channel Evacuation Protocol for Spectrum-agile Networks," *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Apr. 2007.

[137] C.-W. Wang, L.-C. Wang, and F. Adachi, "Modeling and Analysis of Multi-user Spectrum Selection Schemes in Cognitive Radio Networks," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2009.

[138] X. Li and S. A. Zekavat, "Traffic Pattern Prediction and Performance Investigation for Cognitive Radio Systems," *IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2008.

[139] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal Selection of Channel Sensing Order in Cognitive Radio," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 297–307, Jan. 2009.

[140] C.-W. Wang, L.-C. Wang, and F. Adachi, "Performance Gains for Spectrum Utilization in Cognitive Radio Networks with Spectrum Handoff," *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Sep. 2009.

[141] C.-W. Wang and L.-C. Wang, "Modeling and Analysis for Proactive-decision Spectrum Handoff in Cognitive Radio Networks," *IEEE International Conference on Communications (ICC)*, Jun. 2009.

[142] Chee-Hock Ng and Boon-Hee Soong, *Queueing Modelling Fundamentals with Applications in Communication Networks, 2nd.* John Wiley & Sons Inc., 2008.

[143] Q. Zhao and B. M. Sadler, “A Survey of Dynamic Spectrum Access: Signal Processing, Networking, and Regulatory Policy,” *IEEE Signal Processing Magazine*, pp. 79–89, May 2007.

[144] N. K. Jaiswal, “Preemptive Resume Priority Queue,” *Operations Research*, vol. 9, no. 5, pp. 732–742, Sep./Oct. 1961.

[145] *Draft Standard for Wireless Regional Area Networks Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE 802.22 Working Group.

[146] S. Srinivasa and S. A. Jafar, “The Throughput Potential of Cognitive Radio: A Theoretical Perspective,” *IEEE Communications Magazine*, pp. 73–79, May 2007.

[147] Q. Shi, D. Taubenheim, S. Kyperountas, P. Gorday, and N. Correal, “Link Maintenance Protocol for Cognitive Radio System with OFDM PHY,” *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Apr. 2007.

[148] C.-W. Wang, L.-C. Wang, and F. Adachi, “Modeling and Analysis for Rective-decision Spectrum Handoff in Cognitive Radio Networks,” *IEEE GLOBECOM*, Dec. 2010.

[149] D. Willkomm, J. Gross, and A. Wolisz, “Reliable Link Maintenance in Cognitive Radio Systems,” *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, Nov. 2005.

[150] J. Tian and G. Bi, “A New Link Maintenance and Compensation Model for Cognitive UWB Radio Systems,” *International Conference on ITS Telecommunications Proceedings*, Jun. 2006.

[151] L.-C. Wang and C.-W. Wang, "Spectrum Handoff for Cognitive Radio Networks: Reactive-sensing or Proactive-sensing?" *IEEE International Performance Computing and Communications Conference (IPCCC)*, Dec. 2008.

[152] L.-C. Wang, Y.-C. Lu, C.-W. Wang, and D. S.-L. Wei, "Latency Analysis for Dynamic Spectrum Access in Cognitive Radio: Dedicated or Embedded Control Channel?" *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2007.

[153] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, Mar./Apr. 1982.

[154] S. K. Bose, *An Introduction to Queueing Systems.* Kluwer Academic/Plenum Publishers, New York, 2002.

[155] ETSI, "Universal Mobile Telecommunications System (UMTS); Selection Procedures for The Choice of Radio Transmission Technologies of The UMTS," *Technical Report UMTS 30.03 version 3.2.0*, April 1998.

[156] C. R. Stevenson, C. Cordeiro, E. Sofer, and G. Chouinard, "Functional Requirements for The 802.22 WRAN Standard," *IEEE 802.22-05/0007r46*, Sep. 2005.

[157] C.-W. Wang, L.-C. Wang, and F. Adachi, "Optimal Admission Control in Cognitive Radio Networks with Sensing Errors," *IEICE Tech. Rep.*, vol. 109, no. 440, pp. 491–496, Mar. 2010.

[158] L.-C. Wang and C.-W. Wang, "Spectrum Management Techniques with QoS Provisioning in Cognitive Radio Networks," *International Symposium on Wireless and Pervasive Computing (ISWPC)*, May 2010.

[159] L.-C. Wang, C.-W. Wang, and C.-J. Chang, "Modeling and Analysis for Spectrum Handoffs in Cognitive Radio Networks," *Submitted to IEEE Transactions on Mobile Computing*, 2010.

[160] M. S. Jang and D. J. Lee, "Optimum Sensing Time Considering False Alarm in Cognitive Radio Networks," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2009.

[161] R. Rao and A. Ephremides, "On the Stability of Interacting Queues in a Multi-access System," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 918–930, Sep. 1988.

[162] C. Han, J. Wang, and S. Li, "A Spectrum Exchange Mechanism in Cognitive Radio Contexts," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2006.

[163] C. Cormio and K. R. Chowdhury, "A Survey on MAC Protocols for Cognitive Radio Networks," *Ad Hoc Networks*, vol. 7, no. 7, pp. 1315–1329, Sep. 2009.

[164] G. I. P. P. Nico and A. S. Pomportsis, "Distributed Protocols for Ad Hoc Wireless LANs: A Learning-Automata-Based Approach," *Ad Hoc Networks*, pp. 419–431, 2004.

[165] J. Deng, Y. S. Han, and Z. J. Haas, "Analyzing Split Channel Medium Access Control Schemes with ALOHA Reservation," *in Ad-Hoc, Mobile, and Wireless Networks-ADHOC-NOW '03* , pp. 128–139, 2003.

[166] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas, "IEEE 802.11 Packet Delay - A Finite Retry Limit Analysis," *Proceedings of IEEE Global Telecommunications Conference* , pp. 950–954, Dec 2003.

[167] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, pp. 535–547, March 2000.

[168] D. D. Kouvatsos, N. Vlachos, C. Mouchos, and A. Tsokanos, "IP Fragmentation and Framing over Optical Networks : Towards Performance Optimization," *HET-NETs '05*, 2005.

[169] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communication*, vol. 19, no. 10, pp. 2105–2115, Oct. 2001.

[170] E. Yanmaz, O. K. Tonguz, S. Mishra, H. Wu, and C. Qiao, "Efficient dynamic load balancing algorithms using icar systems: a generalized framework," *IEEE Vehicular Technology Conference*, vol. 1, pp. 586–590, Sep. 2002.

[171] E. H.-K. Wu, Y.-Z. Huang, and J.-H. Chiang, "Dynamic adaptive routing for heterogeneous wireless network," *IEEE Global Telecommunications Conference*, vol. 6, pp. 3608–3612, Nov. 2001.

[172] Y. D. Lin and Y. C. Hsu, "Multihop cellular: a new architecture for wireless communications," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 3, pp. 26–30, Mar. 2000.

[173] J. Chen, S.-H. G. Chan, J. He, and S. C. Liew, "Mixed-mode WLAN: the integration of ad-hoc mode with wireless LAN infrastructure," *IEEE Global Telecommunications Conference*, vol. 1, pp. 231–235, Dec. 2003.

[174] D. Niculescu and B. Nath, "Ad hoc positioning system (APS)," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 3, pp. 1734–1743, Mar. 2001.

[175] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *IEEE Computer*, vol. 34, no. 8, pp. 57–66, Aug. 2001.

[176] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 3, pp. 1734–1743, Mar. 2003.

[177] S. J. Ingram, D. Harmer, and M. Quinlan, "Ultrawideband indoor positioning systems and their use in emergencies," *Position Location and Navigation Symposium*, pp. 706–715, Apr. 2004.

[178] D. Madigan, E. Einahrawy, R. P. Martin, W.-H. Ju, P. Krishnan, and A. S. Krishnakumar, "Bayesian indoor positioning systems," *INFOCOM, IEEE International Conference on Computer Communications*, vol. 2, pp. 1217–1227, Mar. 2005.

[179] G. Sun, J. Chen, W. Guo, and K.-J. R. Liu, "Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 12–23, Jul. 2005.

[180] Y.-C. Tseng, S.-L. Wu, W.-H. Liao, and C.-M. Chao, "Location awareness in ad hoc wireless mobile networks," *IEEE Computer*, vol. 34, no. 6, pp. 46–52, Jun. 2001.

[181] M. Mauve, A. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks," *IEEE Network*, vol. 15, no. 6, pp. 30–39, Nov./Dec. 2001.

[182] X. Hong, K. Xu, and M. Gerla, "Scalable routing protocols for mobile ad hoc networks," *IEEE Network*, vol. 16, no. 4, pp. 11–21, Jul./Aug. 2002.

[183] T. Park and K. G. Shin, "Optimal tradeoffs for location-based routing in large-scale ad hoc networks," *IEEE/ACM Transcations on Networking*, vol. 13, no. 2, pp. 398–410, Apr. 2005.

[184] H. Celebi and H. Arslan, "Adaptive positioning systems for cognitive radios," *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 78–84, Apr. 2007.

[185] T. S. Rappaport, *Wireless Communications: Principle and Practice, 2nd edition.* New Jersey: Prentice Hall, Inc., 2002.

[186] G. L. Stüber, *Principle of Mobile Communication, 2nd edition.* Boston/Dordrecht/London: Kluwer Academic Pulishers, 2001.

[187] G. Bianchi, "Performance analysis of IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communication*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[188] L. C. Wang, S. Y. Huang, and A. Chen, "A cross-layer throughput performance investigation for csma/ca-based wireless local area network with directional antennas and capture effect," *IEEE Wireless Communications and Networking Conference*, vol. 3, pp. 1879–1884, Mar. 2004.

[189] M. Oner and F. Jondral, "Extracting the channel allocation information in a spectrum pooling system exploiting cyclostationarity," *IEEE International Symposium on Personal, Indoors, and Mobile Radio Communications*, vol. 1, pp. 551–555, Sep. 2004.

[190] Q. Zhao, L. Tong, and A. Swami, "Decentralized cognitive MAC for dynamic spectrum access," *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 224–232, Nov. 2005.

[191] C. H. Ng, *Queuing Modeling Fundamentals.* New York: John Wiley & Sons, Ltd., 1996.

[192] The Network Simulator- NS2. [Online]. Available: http://www.isi.edu/nsnam/ns

[193] N. Choi, M. Patel, and S. Venkatesan, "A full duplex multi-channel MAC protocol for multi-hop cognitive radio networks," *Cognitive Radio Oriented Wireless Networks and Communications, International Conference on*, pp. 1–5, Jun. 2006.

[194] D. Zheng and J. Zhang, "Protocol design and throughput analysis of frequency-agile multi-channel medium access control," *IEEE Transcations on Wireless Communications*, vol. 5, no. 10, pp. 2887–2895, Oct. 2006.

Appendix

A Distribution Probability Vector for the Sensing-based Channel Selection Scheme

The probability that a secondary user can select channel k for its operating channel is determined inherently based on the traffic patterns for the sensing-based spectrum decision scheme. According to the sensing outcomes, this probability consists of three components. First, we consider the case that false alarm dose not occur at the idle channel k. When the channels in $\Im \subseteq \Omega - \{k\}$ are also actually idle and false alarms do not occur at the channels in $\Re \subseteq \Im$, channel k will be selected with probability $\frac{1}{1+|\Re|}$. Secondly, we consider the case when a false alarm occurs at the idle channel k. If false alarms also occur at all the remaining idle channels, the secondary user will randomly select one channel from all candidate channels to be its operating channel. In this case, channel k is selected with probability $1/|\Omega|$. Thirdly, we consider the case when channel k is actually busy. With the similar argument in the previous case, the secondary user will randomly select one channel if false alarms occur at all the idle channels. In this case, channel k will be selected with probability $1/|\Omega|$. On the other hand, channel k cannot be selected when $k \notin \Omega$. From these observations, we can have

$$p_{sb}^{(k)} = \begin{cases} (1-\rho^{(k)})(1-P_F)\times \\ \displaystyle\sum_{\Im \subseteq \Omega - \{k\}} \left[\prod_{i \in \Im} (1-\rho^{(i)}) \prod_{j \in \Omega - \{k\} - \Im} \rho^{(j)} \sum_{\Re \subseteq \Im} \frac{1}{1+|\Re|} (1-P_F)^{|\Re|} (P_F)^{|\Im|-|\Re|} \right] + \\ \left[(1-\rho^{(k)})P_F + \rho^{(k)} \right] \times \\ \displaystyle\sum_{\Im \subseteq \Omega - \{k\}} \left[\prod_{i \in \Im} (1-\rho^{(i)}) \prod_{j \in \Omega - \{k\} - \Im} \rho^{(j)} (P_F)^{|\Im|} \right] \times \frac{1}{|\Omega|} \ , & k \in \Omega \\ \\ 0 \ , & k \notin \Omega \end{cases}$$

B Derivation of $\omega_i^{(k)}$

First, we consider the type-i secondary connections whose default channels are channel η. Denote $\omega_{i,\eta}^{(k)}$ as the arrival rate of these secondary connections at channel k, $p_{i,\eta}^{(k)}$ as the probability that these secondary connections are interrupted again at channel k, and $\Phi_{i,\eta}^{(k)}$ as the effective service time of these secondary connections at channel k. Hence, we have

$$\omega_i^{(k)} = \sum_{\eta=1}^{M} \omega_{i,\eta}^{(k)} \ , \tag{B1}$$

$$p_i^{(k)} = \sum_{\eta=1}^{M} \frac{\omega_{i,\eta}^{(k)}}{\omega_i^{(k)}} p_{i,\eta}^{(k)} \ , \tag{B2}$$

$$\mathbf{E}[\Phi_i^{(k)}] = \sum_{\eta=1}^{M} \frac{\omega_{i,\eta}^{(k)}}{\omega_i^{(k)}} \mathbf{E}[\Phi_{i,\eta}^{(k)}] \ , \tag{B3}$$

and

$$\mathbf{E}[(\Phi_i^{(k)})^2] = \sum_{\eta=1}^{M} \frac{\omega_{i,\eta}^{(k)}}{\omega_i^{(k)}} \mathbf{E}[(\Phi_{i,\eta}^{(k)})^2] \ . \tag{B4}$$

Next, we derive $\omega_{i,\eta}^{(k)}$ as follows. For the type-i secondary connections whose default channels are channel η, their operating channels are channel $s_{i,\eta}$

after the i^{th} interruption. In addition, for the type-0 secondary connections on their default channel $\eta = s_{0,\eta}$, we have $\omega_{0,\eta}^{(\eta)} = \lambda_s^{(\eta)}$. Moreover, the type-$(i-1)$ secondary connections at channel $s_{i-1,\eta}$ will turn into the new arrivals of the type-i secondary connections at channel $s_{i,\eta}$ when they are interrupted again. Hence, we have

$$\begin{aligned}
\omega_{i,\eta}^{(k)} &= \begin{cases} 0 & , \quad k \neq s_{i,\eta} \\ \lambda_s^{(\eta)} & , \quad k = s_{i,\eta}, \text{ and } i = 0 \\ \omega_{i-1,\eta}^{(s_{i-1,\eta})} p_{i-1,\eta}^{(s_{i-1,\eta})} & , \quad k = s_{i,\eta}, \text{ and } i \geq 1 \end{cases} \\
&= \begin{cases} 0 & , \quad k \neq s_{i,\eta} \\ \lambda_s^{(\eta)} \prod_{j=0}^{i-1} p_{j,\eta}^{(s_{j,\eta})} & , \quad k = s_{i,\eta} \end{cases} \\
&= \begin{cases} 0 & , \quad k \neq s_{i,\eta} \\ \lambda_s^{(k)} \prod_{j=0}^{i-1} \lambda_p^{(s_{j,\eta})} \mathbf{E}[\Phi_{j,\eta}^{(s_{j,\eta})}] & , \quad k = s_{i,\eta} \end{cases} .
\end{aligned} \tag{B5}$$

Note that $p_{j,\eta}^{(s_{j,\eta})} = \lambda_p^{(s_{j,\eta})} \mathbf{E}[\Phi_{j,\eta}^{(s_{j,\eta})}]$ according to (5.18). Because $\lambda_s^{(k)}$ and $\lambda_p^{(k)}$ are given in advanced as well as $\mathbf{E}[\Phi_{j,\eta}^{(s_{j,\eta})}]$ has been derived in Appendix C, we can obtain the closed-form expression for $\omega_{i,\eta}^{(k)}$. Finally, substituting (B5) into (B1), $\omega_i^{(k)}$ in (5.17) can be evaluated.

C Derivations of $\mathbf{E}[\Phi_i^{(k)}]$ and $\mathbf{E}[(\Phi_i^{(k)})^2]$

Here we only show how to evaluate $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$ and $\mathbf{E}[(\Phi_{i,\eta}^{(k)})^2]$ because we can obtain $\mathbf{E}[\Phi_i^{(k)}]$ and $\mathbf{E}[(\Phi_i^{(k)})^2]$ in (5.7), (5.12), and (5.3) when $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$ and $\mathbf{E}[(\Phi_{i,\eta}^{(k)})^2]$ are known according to (B3) and (B4). Denoted $f_{i,\eta}^{(k)}(\phi)$ and $F_{i,\eta}^{(k)}(\phi)$ as the probability density function (pdf) and the cumulative density function (cdf) of $\Phi_{i,\eta}^{(k)}$ where $i \geq 0$, respectively. Moreover, let $\Gamma_p^{(k)}$ be the inter-arrival time of the primary connections at channel k. Then, for the type-0 secondary connections whose default channels are channel $\eta = s_{0,\eta}$, their effective duration at channel k can be expressed as follows:

$$\Phi_{0,\eta}^{(k)} = \begin{cases} 0 & , \quad k \neq s_{0,\eta} \\ \min(\Gamma_p^{(\eta)}, X_s^{(\eta)}) & , \quad k = s_{0,\eta} \end{cases} , \tag{C1}$$

where $\min(a,b)$ is the minimum function. It returns the minimal value among a and b. Hence, the cdf of $\Phi_{0,\eta}^{(k)}$ can be expressed as follows:

$$F_{0,\eta}^{(k)}(\phi) = \begin{cases} 0 & , \quad k \neq s_{0,\eta} \\ 1 - [(1 - A_p^{(\eta)}(\phi))(1 - F_s^{(\eta)}(\phi))] & , \quad k = s_{0,\eta} \end{cases} . \tag{C2}$$

where $A_p^{(k)}(\gamma)$ and $F_s^{(k)}(\phi)$ are the cdfs of $\Gamma_p^{(k)}$ and $X_s^{(k)}$, respectively. They are system parameters and are known in advance.

Next, let $\widetilde{\Phi}_{i,\eta}^{(k)}$ be the remaining transmission time when a type-i secondary connection whose default channel is channel η is interrupted at channel k. Furthermore, the pdf and the cdf of $\widetilde{\Phi}_{i,\eta}^{(k)}$ are denoted by $\widetilde{f}_{i,\eta}^{(k)}(\phi)$ and $\widetilde{F}_{i,\eta}^{(k)}(\phi)$, respectively. According to the definition of $\widetilde{\Phi}_{i,\eta}^{(k)}$, we have

$$\widetilde{\Phi}_{0,\eta}^{(k)} = \begin{cases} 0 & , \quad k \neq s_{0,\eta} \\ X_s^{(\eta)} - \Phi_{0,\eta}^{(\eta)} & , \quad k = s_{0,\eta} \end{cases} . \tag{C3}$$

Because the pdf of $X_s^{(\eta)}$ has been known and the pdf of $\Phi_{0,\eta}^{(\eta)}$ can be derived by differentiating (C2), we can derive $\widetilde{F}_{0,\eta}^{(\eta)}(\phi)$ as follows:

$$\begin{aligned} & \widetilde{F}_{0,\eta}^{(\eta)}(\phi) \\ = & \mathbf{Pr}(\widetilde{\Phi}_{0,\eta}^{(\eta)} \leq \phi) \\ = & \mathbf{Pr}(X_s^{(\eta)} - \Phi_{0,\eta}^{(\eta)} \leq \phi) \\ = & \mathbf{Pr}(X_s^{(\eta)} - \min(\Gamma_p^{(\eta)}, X_s^{(\eta)}) \leq \phi) \\ = & \iint_{x - \min(\gamma,x) \leq \phi} a_p^{(\eta)}(\gamma) f_s^{(\eta)}(x) \eta\gamma\eta x \ , \end{aligned} \tag{C4}$$

where $a_p^{(k)}(\gamma)$ and $f_s^{(k)}(x)$ are the pdfs of $\Gamma_p^{(k)}$ and $X_s^{(k)}$, respectively. Then, we can obtain $\widetilde{f}_{0,\eta}^{(k)}(\phi)$ by differentiating $\widetilde{F}_{0,\eta}^{(k)}(\phi)$. Furthermore, according to the total probability principle, we have

$$\widetilde{f}_{0,\eta}^{(k)}(\phi) \;=\; \begin{cases} 0 & , \quad k \neq s_{0,\eta} \\ \mathbf{Pr}(N<1)\widetilde{f}_{0,\eta}^{(\eta)}(\phi|N<1)+\mathbf{Pr}(N\geq 1)\widetilde{f}_{0,\eta}^{(\eta)}(\phi|N\geq 1) & , \quad k = s_{0,\eta} \end{cases} . \tag{C5}$$

Because the remaining transmission time is zero when a secondary connection does not encounter any interruption during its transmission period, we have $\widetilde{f}_{0,\eta}^{(\eta)}(\phi|N<1)=\delta(\phi)$ where $\delta(\phi)$ is the delta function. Then, we can revise (C5) as follows:

$$\widetilde{f}_{0,\eta}^{(\eta)}(\phi|N\geq 1)=\frac{\widetilde{f}_{0,\eta}^{(\eta)}(\phi)}{\mathbf{Pr}(N\geq 1)}=\frac{\widetilde{f}_{0,\eta}^{(\eta)}(\phi)}{p_{0,\eta}^{(\eta)}}=\frac{\widetilde{f}_{0,\eta}^{(\eta)}(\phi)}{\lambda_p^{(\eta)}\mathbf{E}[\Phi_{0,\eta}^{(\eta)}]} \;. \tag{C6}$$

When a type-0 secondary connection is interrupted on its default channel $\eta = s_{0,\eta}$, its remaining transmission time will turn into the transmission time of the type-1 secondary connection at channel $s_{1,\eta}$. That is, the events $\{\Phi_{1,\eta}^{(s_{1,\eta})} < \phi\}$ and $\{\min(\Delta_p^{(s_{1,\eta})}, \widetilde{\Phi}_{0,\eta}^{(s_{0,\eta})}) < \phi|N \geq 1\}$ are equivalent. Then, following the similar argument as in (C2) and (C6), we can have

$$F_{1,\eta}^{(k)}(\phi) \;=\; \begin{cases} 0 & , \quad k \neq s_{1,\eta} \\ 1-[(1-A_p^{(s_{1,\eta})}(\phi))(\widetilde{F}_{0,\eta}^{(s_{0,\eta})}(\phi|N\geq 1))] & , \quad k = s_{1,\eta} \end{cases} . \tag{C7}$$

and

$$\widetilde{f}_{1,\eta}^{(s_{1,\eta})}(\phi|N\geq 2)=\frac{\widetilde{f}_{1,\eta}^{(s_{1,\eta})}(\phi)}{\mathbf{Pr}(N\geq 2)}=\frac{\widetilde{f}_{1,\eta}^{(s_{1,\eta})}(\phi)}{\prod_{j=0}^{1}p_{j,\eta}^{(s_{j,\eta})}}=\frac{\widetilde{f}_{1,\eta}^{(s_{1,\eta})}(\phi)}{\prod_{j=0}^{1}\lambda_p^{(s_{j,\eta})}\mathbf{E}[\Phi_{j,\eta}^{(s_{j,\eta})}]} \;, \tag{C8}$$

where $\widetilde{F}_{0,\eta}^{(s_{0,\eta})}(\phi|N \geq 1)$ can be derived by integrating $\widetilde{f}_{0,\eta}^{(s_{0,\eta})}(\phi|N \geq 1)$ in (C6).

Repeating the similar discussions, the general forms of $F_{i,\eta}^{(k)}(\phi)$ and $\widetilde{f}_{i,\eta}^{(s_{i,\eta})}(\phi|N \geq$

$i + 1)$ for any i can be expressed as follows:

$$F_{i,\eta}^{(k)}(\phi) \;=\; \begin{cases} 0 & , \quad k \neq s_{i,\eta} \\ 1 - [(1 - A_p^{(s_{i,\eta})}(\phi))(\widetilde{F}_{i-1,\eta}^{(s_{i-1,\eta})}(\phi|N \geq i))] & , \quad k = s_{i,\eta} \end{cases} , \tag{C9}$$

and

$$\widetilde{f}_{i,\eta}^{(s_{i,\eta})}(\phi|N \geq i{+}1) = \frac{\widetilde{f}_{i,\eta}^{(s_{i,\eta})}(\phi)}{\mathbf{Pr}(N \geq i+1)} = \frac{\widetilde{f}_{i,\eta}^{(s_{i,\eta})}(\phi)}{\prod_{j=0}^{i} p_{j,\eta}^{(s_{j,\eta})}} = \frac{\widetilde{f}_{i,\eta}^{(s_{i,\eta})}(\phi)}{\prod_{j=0}^{i} \lambda_p^{(s_{j,\eta})}\mathbf{E}[\Phi_{j,\eta}^{(s_{j,\eta})}]} , \tag{C10}$$

where $\widetilde{F}_{i-1,\eta}^{(s_{i-1,\eta})}(\phi|N \geq i)$ is the integration of $\widetilde{f}_{i-1,\eta}^{(s_{i-1,\eta})}(\phi|N \geq i)$. Because the arrivals of the primary connections follow the Poisson process, we have $A_p^{(k)}(\gamma) = 1 - e^{-\lambda_p^{(k)}\gamma}$. Based on the relationships of (C9) and (C10), we can derive the functions $f_{i,\eta}^{(k)}(\phi)$ from $F_{i,\eta}^{(k)}(\phi)$, and thus $\mathbf{E}[\Phi_{i,\eta}^{(k)}]$ and $\mathbf{E}[(\Phi_{i,\eta}^{(k)})^2]$ can be also evaluated.

D Derivation of $\mathbf{Pr}(\widetilde{X}_p = x)$

Now, we show how to derive $\mathbf{Pr}(\widetilde{X}_p = x)$ from $f_p(x)$. Let $f_p(x) = \mathbf{Pr}(X_p = x)$. We can have

$$\mathbf{Pr}(\widetilde{X}_p = \widetilde{x}) = \sum_{x=1}^{\widetilde{X}_p} \mathbf{Pr}(\widetilde{X}_p = \widetilde{x}|X_p = x)\mathbf{Pr}(X_p = x) \ . \tag{D1}$$

Referring to [135], we find that $\mathbf{Pr}(\widetilde{X}_p = \widetilde{x}|X_p = x)$ follows the negative binomial distribution with parameter $P_M\rho_s$. That is,

$$\mathbf{Pr}(\widetilde{X}_p = \widetilde{x}|X_p = x) = \binom{\widetilde{x}-1}{\widetilde{x}-x}(1 - P_M\rho_s)^x(P_M\rho_s)^{\widetilde{x}-x} \ . \tag{D2}$$

E Derivation of (10.6)

The region $R_{CT}^{(u)}$ shown in (10.6) is composed of three sections with the areas of $\pi(\frac{d_{23}}{z_a^{1/\alpha}})^2$, A_1, and A_2. Figure E1 shows all the parameters used in deriving the area of A_1 and A_2 in (10.7) and (10.8), respectively. In (10.7), the section A_1 is composed of two fan-shaped areas with the measures of $(\frac{d_{23}}{z_a^{1/\alpha}})^2(\pi - \theta')$ and $R^2\theta$, and two identical triangles with the lengths of R, r_2, and $\frac{d_{23}}{z_a^{1/\alpha}}$, where

$$\theta = cos^{-1}\frac{R^2 + r_2^2 - \frac{d_{23}}{z_a^{1/\alpha}}}{2Rr_2} , \tag{E1}$$

$$\theta' = cos^{-1}\frac{r_2^2 + (\frac{d_{23}}{z_a^{1/\alpha}})^2 - R^2}{2r_2(\frac{d_{23}}{z_a^{1/\alpha}})} , \tag{E2}$$

$$\Delta = \sqrt{s(s-R)(s-r_2)(s-\frac{d_{23}}{z_a^{1/\alpha}})} , \tag{E3}$$

and

$$s = \frac{R + r_2 + \frac{d_{23}}{z_a^{1/\alpha}}}{2} . \tag{E4}$$

Similarly, in (10.8), the section A_2 is also made up by two fan-shaped areas with the measures of $(\frac{d_{23}}{z_a^{1/\alpha}})^2\phi$ and $(r_3 z_i^{1/\alpha})^2\phi'$, and the triangle with the lengths of r_2, $r_3 z_i^{1/\alpha}$, and $\frac{d_{23}}{z_a^{1/\alpha}}$, where

$$\phi = cos^{-1}\frac{r_2^2 + (\frac{d_{23}}{z_a^{1/\alpha}})^2 - (r_3 z_i^{1/\alpha})^2}{2r_2(\frac{d_{23}}{z_a^{1/\alpha}})} , \tag{E5}$$

$$\phi' = cos^{-1}\frac{r_2^2 + (r_3 z_i^{1/\alpha})^2 - (\frac{d_{23}}{z_a^{1/\alpha}})^2}{2r_2(r_3 z_i^{1/\alpha})} , \tag{E6}$$

$$\Delta' = \sqrt{s'(s'-r_2)(s'-r_3 z_i^{1/\alpha})(s'-\frac{d_{23}}{z_a^{1/\alpha}})} , \tag{E7}$$

and

$$s' = \frac{r_2 + r_3 z_i^{1/\alpha} + \frac{d_{23}}{z_a^{1/\alpha}}}{2} . \tag{E8}$$

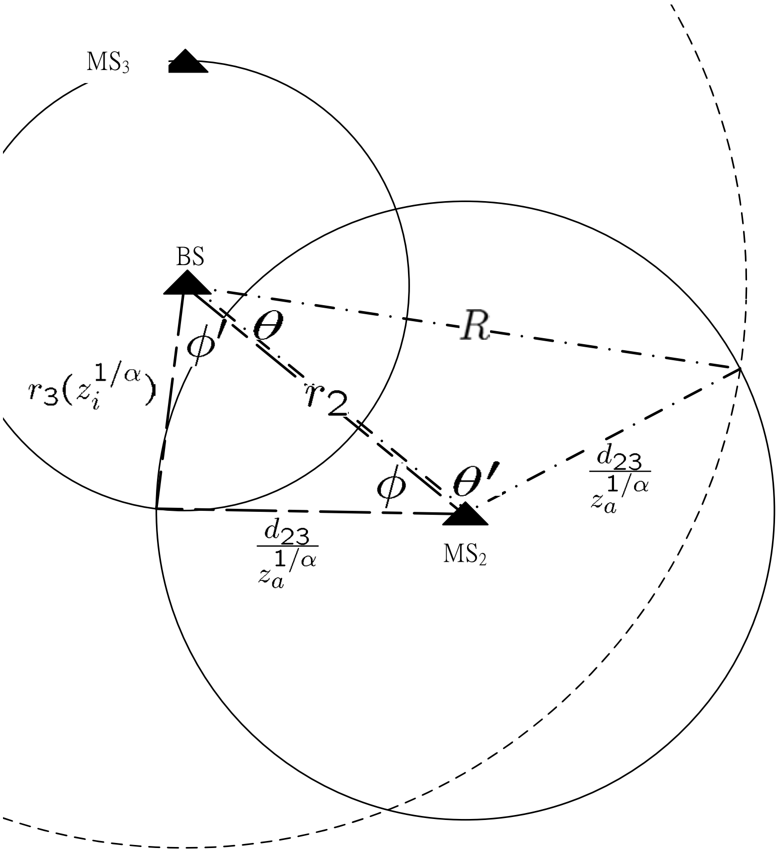


Figure E1: Definition of parameters in the calculation of the area of region $R_{CT}^{(u)}$ in the infrastructure uplink case.

F Derivation of (10.12) and (10.13)

The distances between the base station to the intersections by the two circles with the radii of $r_3 z_i'^{1/\alpha}$ and $r_2 z_a'^{1/\alpha}$ and centered at (r_3, θ_3) and (r_2, θ_2) are denoted by r^+ and r^-, respectively. Given the locations of intersections (x, y), they can be obtained by jointly solving the equations as follows:

$$\begin{cases} (x - r_2 cos\theta_2)^2 + (y - r_2 sin\theta_2)^2 = (r_2 z_a'^{1/\alpha})^2 \ ; \\ (x - r_3 cos\theta_3)^2 + (y - r_3 sin\theta_3)^2 = (r_3 z_a'^{1/\alpha})^2 \ . \end{cases} \tag{F1}$$

The distances between the base station and the intersections r^+ and r^-, respectively, shown in (10.12) and (10.13) can be obtained by solving the locations (x, y) from (F1).

G Derivation of (10.15)

In the infrastructure downlink case, when $max(r^+, r^-) \leq R$, the region $R_{CT}^{(d)}$ is composed of three sections with the areas of $\pi(d_{23} z_a'^{1/\alpha})^2$, A_1, and A_2. Figure G1 details the parameters used to calculate the areas of these sections in (10.16) and (10.17). In (10.16), the region A_1 is made of two fan-shaped areas with the measures of $(r_2 z_a'^{1/\alpha})^2(\pi - \theta')$ and $R^2\theta$, and two identical triangles with the lengths of R, r_2, and $r_2 z_a'^{1/\alpha}$, where

$$\theta = cos^{-1} \frac{R^2 + r_2^2 - (r_2 z_a'^{1/\alpha})}{2Rr_2} \ , \tag{G1}$$

$$\theta' = cos^{-1} \frac{r_2^2 + (r_2 z_a'^{1/\alpha})^2 - R^2}{2r_2(r_2 z_a'^{1/\alpha})} \ , \tag{G2}$$

$$\Delta = \sqrt{s(s - R)(s - r_2)(s - r_2 z_a'^{1/\alpha})} \ , \tag{G3}$$

and

$$s = \frac{R + r_2 + r_2 z_a'^{1/\alpha}}{2} \ . \tag{G4}$$

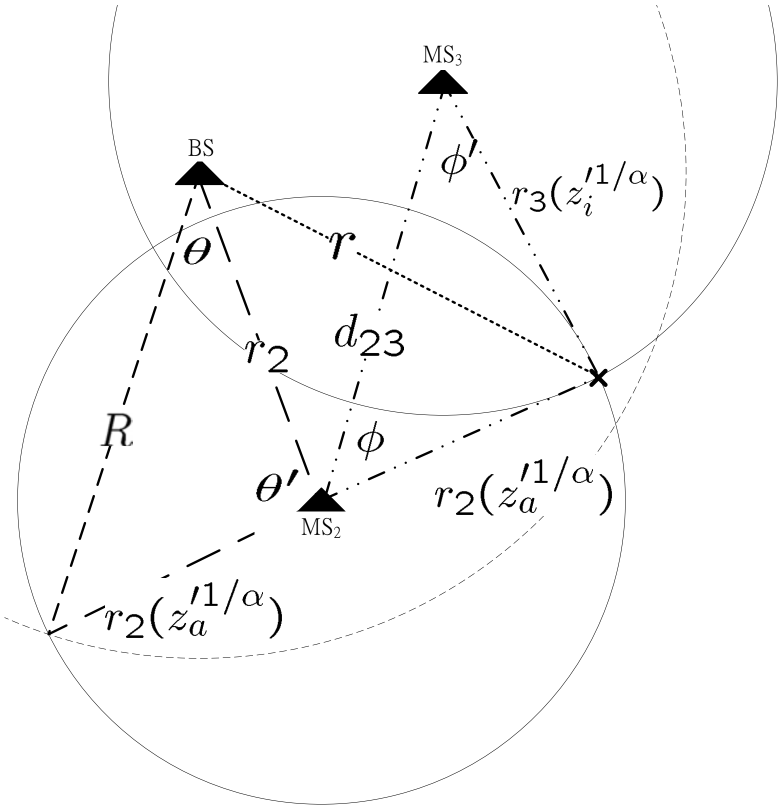


Figure G1: Definition of parameters in the calculation of the area of region $R_{CT}^{(d)}$ in the infrastructure downlink case when $max(r^+, r^-) \leq R$.

Similarly, the section A_2 in (10.17) is also composed of two fan-shaped areas with the measures of $(r_2 z_a'^{1/\alpha})^2\phi$ and $(r_3 z_i'^{1/\alpha})^2\phi'$, and the triangle with the lengths of d_{23}, $r_2 z_a'^{1/\alpha}$, and $r_3 z_i'^{1/\alpha}$, where

$$\begin{aligned} \phi &= cos^{-1}\frac{d_{23}^2 + (r_2 z_a'^{1/\alpha})^2 - (r_3 z_i'^{1/\alpha})^2}{2d_{23}(r_2 z_a'^{1/\alpha})} , && \text{(G5)} \\ \phi' &= cos^{-1}\frac{d_{23}^2 + (r_3 z_i'^{1/\alpha})^2 - (r_2 z_a'^{1/\alpha})^2}{2d_{23}(r_3 z_i'^{1/\alpha})} , && \text{(G6)} \\ \Delta' &= \sqrt{s'(s' - d_{23})(s' - r_3 z_i'^{1/\alpha})(s' - r_2 z_a'^{1/\alpha})} , && \text{(G7)} \end{aligned}$$

and

$$s' = \frac{d_{23} + r_2 z_a'^{1/\alpha} + r_3 z_i'^{1/\alpha}}{2} . \qquad \text{(G8)}$$

H Derivation of (10.18)

When $max(r^+, r^-) > R$, the region $R_{CT}^{(d)}$ can be separated into four sections with the areas of $\pi(d_{23} z_a'^{1/\alpha})^2$, A_1, A_2, and A_3. Figure H1 shows all the parameters in deriving (10.19), (10.20), and (10.21). In (10.19), the section A_1 can be divided into two fan-shaped areas with the measures of $(r_2 z_a'^{1/\alpha})^2(\pi - \theta')$ and $R^2\theta$, and the triangle with the lengths of R, r_2, and $r_2 z_a'^{1/\alpha}$, where

$$\begin{aligned} \theta &= cos^{-1}\frac{R^2 + r_2^2 - (r_2 z_a'^{1/\alpha})}{2Rr_2} , && \text{(H1)} \\ \theta' &= cos^{-1}\frac{r_2^2 + (r_2 z_a'^{1/\alpha})^2 - R^2}{2r_2(r_2 z_a'^{1/\alpha})} , && \text{(H2)} \\ \Delta &= \sqrt{s(s-R)(s-r_2)(s - r_2 z_a'^{\frac{1}{\alpha}})} , && \text{(H3)} \end{aligned}$$

and

$$s = \frac{R + r_2 + r_2 z_a'^{\frac{1}{\alpha}}}{2} . \qquad \text{(H4)}$$

In addition, the section A_2 is made of two fan-shaped areas with the measures

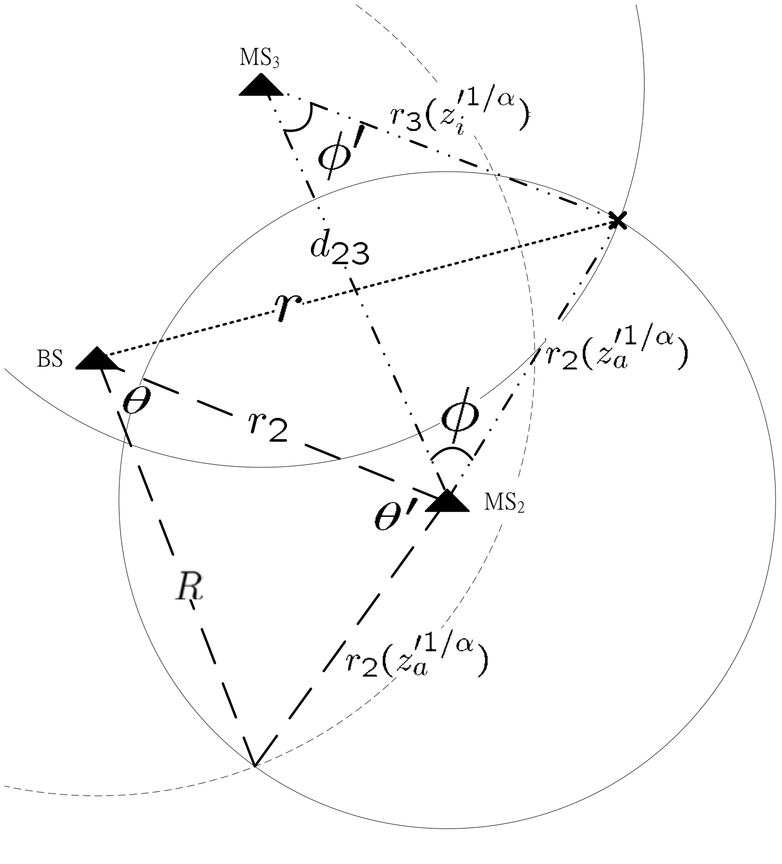


Figure H1: Definition of parameters in the calculation of the areas of sections A_1 and A_2 in (10.18) in the infrastructure downlink case when $max(r^+, r^-) > R$.

of $(r_2 z_a^{\prime 1/\alpha})^2\phi$ and $(r_3 z_i^{\prime 1/\alpha})^2\phi'$, and the triangle with the lengths of d_{23}, $r_3 z_i^{\prime 1/\alpha}$, and $r_2 z_a^{\prime 1/\alpha}$, where

$$\begin{aligned}
\phi &= cos^{-1}\frac{d_{23}^2 + (r_2 z_a^{\prime 1/\alpha})^2 - (r_3 z_i^{\prime 1/\alpha})^2}{2d_{23}(r_2 z_a^{\prime 1/\alpha})} \ , && \text{(H5)}\\
\phi' &= cos^{-1}\frac{d_{23}^2 + (r_3 z_i^{\prime 1/\alpha})^2 - (r_2 z_a^{\prime 1/\alpha})^2}{2d_{23}(r_3 z_i^{\prime 1/\alpha})} \ , && \text{(H6)}\\
\Delta' &= \sqrt{s'(s' - d_{23})(s' - r_3 z_i^{\prime\frac{1}{\alpha}})(s' - r_2 z_a^{\prime\frac{1}{\alpha}})} \ , && \text{(H7)}
\end{aligned}$$

and

$$s' = \frac{d_{23} + r_2 z_a^{\prime\frac{1}{\alpha}} + r_3 z_i^{\prime\frac{1}{\alpha}}}{2} \ . \qquad \text{(H8)}$$

At last, in (10.21), the section A_3 is constructed by three arc-shaped areas with the measures of $(r_3 z_i^{\prime 1/\alpha})^2\psi_2 - \frac{1}{2}(r_3 z_i^{\prime 1/\alpha})^2 sin\psi_2$, $(r_2 z_a^{\prime 1/\alpha})^2\psi_3 - \frac{1}{2}(r_2 z_a^{\prime 1/\alpha})^2 sin\psi_3$, and $R^2\psi_1 - \frac{1}{2}R^2 sin\psi_1$, and the triangle with the lengths of a, b, and c, where

$$\begin{aligned}
\psi_1 &= cos^{-1}\frac{R^2 + r_2^2 - r_2^2 z_a^{\prime\frac{2}{\alpha}}}{2Rr_2} + cos^{-1}\frac{R^2 + r_3^2 - r_3^2 z_i^{\prime\frac{2}{\alpha}}}{2Rr_3} - \\
&\quad cos^{-1}\frac{r_2^2 + r_3^2 - d_{23}^2}{2r_2 r_3} \ , && \text{(H9)}
\end{aligned}$$

$$\begin{aligned}
\psi_2 &= cos^{-1}\frac{d_{23}^2 + r_3^2 z_i^{\prime\frac{2}{\alpha}} - r_2^2 z_a^{\prime\frac{2}{\alpha}}}{2d_{23}(r_3 z_i^{\prime\frac{1}{\alpha}})} + cos^{-1}\frac{d_{23}^2 + r_3^2 - r_2^2}{2d_{23}r_3} - \\
&\quad cos^{-1}\frac{r_3^2 + r_3^2 z_i^{\prime\frac{2}{\alpha}} - R^2}{2r_3(r_3 z_i^{\prime\frac{1}{\alpha}})} \ , && \text{(H10)}
\end{aligned}$$

$$\begin{aligned}
\psi_3 &= cos^{-1}\frac{d_{23}^2 + r_2^2 z_a^{\prime\frac{2}{\alpha}} - r_3^2 z_i^{\prime\frac{2}{\alpha}}}{2d_{23}(r_2 z_a^{\prime\frac{1}{\alpha}})} + cos^{-1}\frac{d_{23}^2 + r_2^2 - r_3^2}{2d_{23}r_2} - \\
&\quad cos^{-1}\frac{r_2^2 + r_2^2 z_a^{\prime\frac{2}{\alpha}} - R^2}{2r_2(r_2 z_a^{\prime\frac{1}{\alpha}})} \ , && \text{(H11)}
\end{aligned}$$

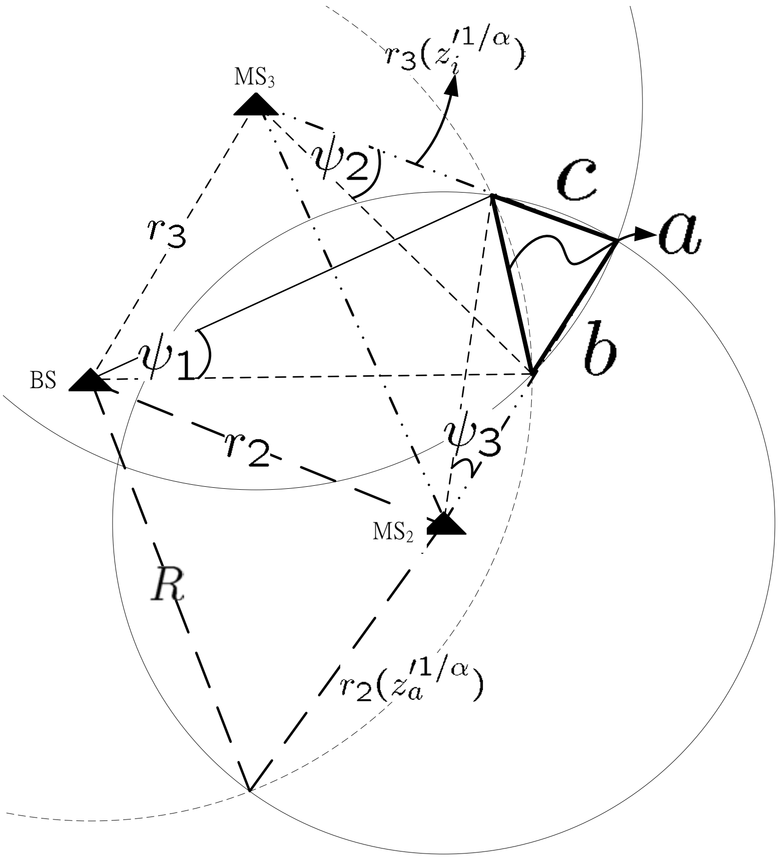


Figure H2: Definitions of parameters in the calculation of the area of section A_3 in (10.18) in the infrastructure downlink case when $max(r^+, r^-) > R$.

$$
\begin{aligned}
\Delta'' &= \sqrt{s''(s''-a)(s''-b)(s''-c)}\ , && \text{(H12)}\\
s'' &= \frac{a+b+c}{2}\ , && \text{(H13)}\\
a &= 2Rsin\frac{\psi_1}{2}\ , && \text{(H14)}\\
b &= 2(r_3 zi'^{\frac{1}{\alpha}})sin\frac{\psi_2}{2}\ , && \text{(H15)}
\end{aligned}
$$

and

$$
c = 2(r_2 z_a'^{\frac{1}{\alpha}})sin\frac{\psi_3}{2}\ . \qquad \text{(H16)}
$$

Figure H2 shows all the parameters in deriving (10.21).