IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received July 22, 2019, accepted August 10, 2019, date of publication August 13, 2019, date of current version August 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935126

A Virtual Control Layer Resource Allocation Framework for Heterogeneous Cognitive Radio Network

SHI WANG[1,2], (Member, IEEE), SUNIL MAHARAJ[2], (Senior Member, IEEE), AND ATTAHIRU S. ALFA[2,3], (Member, IEEE)

[1]School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China
[2]Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria 0028, South Africa
[3]Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada

Corresponding author: Sunil Maharaj (sunil.maharaj@up.ac.za)

This work was supported by the Sentech Chair in BWMC and DTI THRIP Program.

ABSTRACT To make best use of the various radio resources in the heterogeneous network enabled by the software defined network or cognitive radio (CR) technology, a resource allocation framework with flexibility and quick reconfigurability in the network control layer is an important issue. In this paper, the authors propose a channel allocation framework with configurable objectives and high computing efficiency in the complicated context of a multi-user multi-channel CR network cell. First, a channel allocation protocol named the distribution probability matrix (DPM) is applied to model the channel allocation scenarios quantitatively. Then, a queueing analytical framework using DPM is built to model the CR system and comprehensive performance evaluations of every individual secondary user are obtained separately. An overall performance evaluation of the CR system is carried out using the concept of weighted throughput, which is introduced to represent the importance of the users and the feature to distinguish different types of users. Then, a parameter named overtime probability (OP) is introduced to describe the measure of delay approximately with high computing efficiency. Thereby, an optimization of the system is formulated to maximize the overall weighted throughput under delay constraints represented by OP and a hill climbing algorithm is developed to find the solution in terms of DPM. The numerical results reveal how to allocate the resources to achieve the optimization objective under various system settings and prove the computing efficiency of the framework.

INDEX TERMS Cognitive radio, control layer, delay, hill climbing algorithm, optimization, resource allocation, weighted throughput.

I. INTRODUCTION

The demand for wireless data transmission is growing rapidly nowadays, with a massively increasing number of wireless devices and various applications. Besides exploring more radio spectrum bands to satisfy this demand, future wireless communication technologies also aim at converging existing technologies with emerging ones to create stronger and faster networks. Among the technologies, software defined network (SDN) [1] and cognitive radio (CR) technology [2], [3] provide the functionality to manage radio resources and devices from different wireless technologies.

Orchestrating and managing radio spectrum resources from different domains are important issues to enable the applications of SDN and CR technology. The related protocols or frameworks for radio resource allocation are applied in a virtual control layer between the application layer and infrastructure layer in SDN [4] or a central node in cognitive radio networks (CRNs) or clusters. Compared to the traditional telecommunication system, the crucial functionalities that the resource allocation(RA) frameworks must provide are the flexibility to fit the different data transmission demands of users and fast reconfigurability to guarantee quality of service (QoS) under various scenarios.

Two preliminary objectives are essential to be considered in the RA framework, but have not been studied well yet:

The associate editor coordinating the review of this article and approving it for publication was Hayder Al-Hraishawi.

The first is how to make the RA objectives configurable to adapt to the different types of services in order to optimize the service provider's benefit. The second one is how to reduce the computational cost, introduced by the complexity of both the network structures and RA objectives, in order to improve the QoS potential across different systems. The access of the radio resource is limited and dynamic [5] in a heterogeneous network (HetNet) enabled by the SDN or CR. Queueing theory is an effective tool to study the dynamic behavior of a communication system and the applications of queueing theory are well developed [6]. In this research, a queueing analytical framework is proposed for the objectives mentioned above.

The first step in modeling the RA objectives is to distinguish different types of users, who may have various characteristics [7], such as profit or importance to the service provider. According to the publications that investigate the method to represent these features of the users, weighted throughput is considered appropriate. The weights admit various interpretations, including levels of importance, ''utility'', and price [8]. Then, using the concept of weighted throughput, the way to schedule different jobs with different weights when resources are insufficient is studied in [9] and [10]. However, the effective application of the weighted throughput in the HetNet has not been well studied sufficiently and requires further research.

The next step is to model RA protocol using a queueing analytical framework. In related publications, channel allocation protocols such as opportunistic scheduling [11], [12] and random allocation [13] are commonly applied to evaluate the performance of the system. However, the protocols are not configurable because they either have a fixed objective built in (opportunistic scheduling) or do not consider any objectives at all (random allocation). To make the channel allocation protocol adaptable to various scenarios, the authors applied a configurable and flexible channel allocation protocol named the distribution probability matrix (DPM) [14]. The flexibility of DPM enables the proposed framework to describe all possible channel allocation results and consequently to analyze the performance of any given channel allocation scenario. In addition, when dealing with multiple users, such as secondary users (SUs) in CRN, they are usually described by an identical model in the related literature. This makes it complicated to analyze multiple users separately, especially when the users are of different types. To improve this, the authors developed a queueing analytical framework using the DPM that is capable of modeling the HetNet system with every individual user independently configurable. Moreover, the performance of every individual user can be obtained separately using this framework.

Performance evaluations during the RA process that are used to optimize the RA objectives consume a large part of the computational resources and time. Among the performance evaluations, the calculation of the measure of delay is complicated if the delay is one of the concerns in the RA objectives. A lot of research on delay analysis in the HetNet focuses on the process of transmission of the packets ahead of the tagged one [15]. This method is accurate; however it requires consideration of every step of the transmission of the packets ahead, which makes the calculation time-consuming. Since the radio resources are dynamic in the HetNet, performance evaluations must be carried out first to direct the allocation of the radio resources. Thus, to make the best use of the limited and dynamic resources available to the system, not only the accuracy of the performance evaluation is of concern, but also its efficiency. Considering this, the authors developed a method to estimate the measure of delay with high efficiency.

In summary, the main contributions of this research that adopt the approaches above are listed below:

- the design and implement of a novel channel allocation protocol named DPM that can analyze the performance of any given channel allocation scheme with multiple users.
- advanced and realistic performance evaluation that is able to configure and analyze every individual user independently within the system, such as weighted throughput and fast delay evaluation, from which the authors formed a realistic RA optimization problem.
- a heuristic algorithm to find the solution of the above-mentioned RA optimization problem.

The detailed process are as follows. First, the authors model a multi-user multi-channel cognitive radio network, which can also describe a SDN HetNet by ignoring the activity of the primary user (PU). Then, a queueing analytical framework that is able to evaluate each individual user in the system simultaneously is established to reveal the relationship between various performance measures and channel allocation protocols. With the help of the framework, an objective considering both the overall performance in terms of weighted throughput and the delay constraints that are used to guarantee the transmission of SUs with low priority is formulated as an optimization problem to find the optimal channel allocation solutions. Analysis of how the system settings and environmental parameters affect the optimal allocation results is carried out afterwards.

The rest of the paper is organized as follows. The system model and assumptions are introduced in Section II. Then, the calculation of performance evaluations and the way in which the optimization problem is built are shown in Section III. Numerical results and discussions are shown in Section IV. Section V concludes the paper.

II. SYSTEM MODEL AND ASSUMPTIONS

A. BASIC MODEL SETTINGS

In this research, an overlay [16] CRN with multiple SUs carrying different type of services inside a CRN cell or cluster is considered. Multiple PU channels are available to this SU system while there is no PU transmitting through them. Perfect control information exchange between the PU and SU system is assumed. The SU system comprises a base station (or a center node) and multiple SUs of different types working within the coverage of one or more PU networks. The number

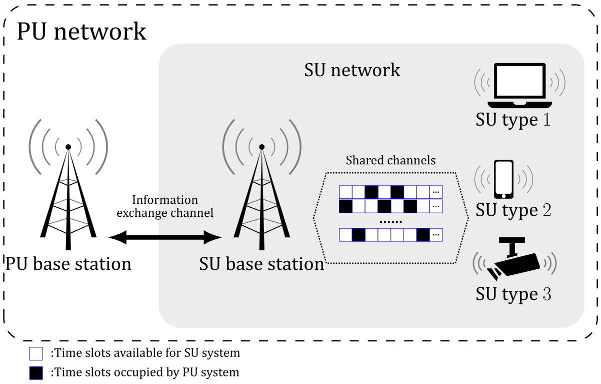


FIGURE 1. System model.

of SUs is denoted by M. The number of PU channels that are available to the SU system is denoted by J. Both the PU and SU systems work under the same time slot structure that is mutually synchronized. In this research, the focus is on uplink data transmission. Each SU has a finite buffer to store the packets waiting to be transmitted. The buffer size of the i-th SU is denoted by K_i. An example of the system model is shown in Figure 1. This study can also be applied to investigate a SDN-based HetNet by ignoring the existence of the PUs.

The channel allocation protocol is the rule that the SU system uses to determine which SU can transmit on the channel where the PU is sensed to be absent. The common objective of the channel allocation protocol in the CRN is to satisfy QoS requirements such as maximizing the throughput, minimizing the packet transmission failure rate and minimizing the delay.

B. PRIMARY USER ACTIVITY MODEL

In an overlay CRN, the SU system can use a channel when there is a PU transmitting on it. In this research, the PU activities on a channel are modeled as a time-homogeneous first-order Markov process with two states. This process is independent of the behavior of the SU system. Let $O^j(t) = \{0, 1\}$ represent the PU activity state of the j-th channel at the t-th time slot. When there is a PU transmitting, the state is called "busy" and $O^j(t) = 0$, otherwise the state is called "free" and $O^j(t) = 1$. The transition matrix of the process can be written as:

$$\boldsymbol{P}^j_{PU} = \begin{pmatrix} p^j_{b\to b} & 1 - p^j_{b\to b} \\ 1 - p^j_{f\to f} & p^j_{f\to f} \end{pmatrix}, \tag{1}$$

where $p^j_{b\to b} = \Pr\{O^j(t) = 0 | O^j(t-1) = 0\}$ denotes the probability that there is PU transmitting on the j-th channel (the "busy" state), given that there was a PU transmitting on the channel (the "busy" state) in the previous time slot, and $p^j_{f\to f} = \Pr\{O^j(t) = 1 | O^j(t-1) = 1\}$ denotes the probability that the j-th channel state changes from "free" to "free".

C. SECONDARY SYSTEM TRANSMISSION MODEL

The SU system transmits the data within time slots that are synchronized with the PU system. An SU time slot consists of three parts: the spectrum sensing part, the channel allocation part and the data transmission part.

The SU base station and the mobile stations keep sensing the existence of the PU transmission and the base station collects all the information at the beginning of a time slot to estimate whether the PU is absent on every PU channel. In our model perfect spectrum sensing is assumed. Besides the PU's presence, the transmission conditions (e.g. the signal-to-noise ratio (SNR) at transceivers) of each SU mobile station are also collected by the SU base station and used to guide the channel allocation.

According to the estimation of available PU channels in the spectrum sensing part, the SU system allocates the available channels to the SUs to transmit data in the current time slot. The channel allocation protocol applied in this paper is named the DPM [14]. This protocol works as follows: First, an J-by-M distribution probability matrix is set up as indicated below:

$$\boldsymbol{D} = \begin{pmatrix} d^1_1 & \cdots & d^1_M \\ \vdots & \ddots & \vdots \\ d^J_1 & \cdots & d^J_M \end{pmatrix}, \tag{2}$$

where d^j_i represents the probability that the SU system allocates the j-th channel to the i-th SU when the channel is not occupied by the PU. It is assumed that one PU channel can only be assigned to one SU, thus in $\boldsymbol{D}$, $\sum_{i=1}^{M} d^j_i = 1$ holds for any $j \in [1, J]$.

If a PU channel is allocated to one SU, the SU transmits its data during the remainder of the time slot. To make better use of the channel resources, an adaptive modulation and coding (AMC) scheme is applied by the SU system [17]. The modulation schemes are determined by the channel condition and the channel condition is modeled as a Markov process. A transition matrix of channel condition states denoted by $\boldsymbol{P}^j_{\mathrm{CS}}$ is used to represent this model. Details of the data transmission are shown in Appendix A. As the error recovery method, a stop-n-wait automatic request and repeat (ARQ) is applied. The average packet error rate is Per.

After data transmission, the packets arriving during the current time slot will enter the SU buffer. If the total number of packets exceeds the buffer size, the overflow packets will be rejected. This model is illustrated in Figure 2.

D. QUEUEING MODEL

1) ARRIVAL PROCESS

The number of arrival packets at each SU during one time slot is assumed to follow a Batch Bernoulli process [18]. A probability vector $\boldsymbol{\alpha}_i$ describes the process of the i-th SU as follows:

$$\boldsymbol{\alpha}_i = \big(\alpha_i(0), \alpha_i(1), \cdots, \alpha_i(v_i)\big), \tag{3}$$

where $\alpha_i(j)$ is the probability that j packets arrive during one time slot at the i-th SU and v_i is the maximum number of packets that can arrive at the i-th SU during one time slot.

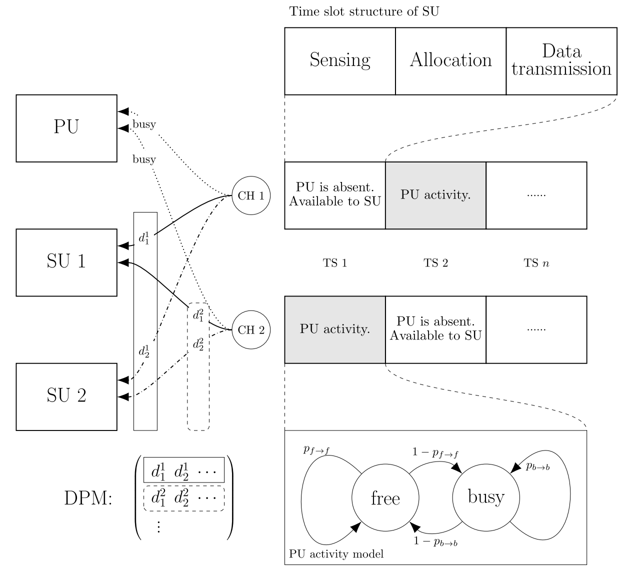


FIGURE 2. **Illustration of DPM, time slot structure of SU system and the PU activity model.**

2) QUEUEING MODEL

With the analysis above, the queueing analytical model of the SU system can be built. To make the proposed model capable of evaluating the performance of the SU system comprehensively, the setup of the state space of the system must include as many details as possible about the entire SU system, such as the states of the number of packets in the buffer of each SU, the states of PU activities on each channel and the state of channel conditions. The states of the number of packets in the buffer of each SU are denoted by:

$$\boldsymbol{b} \in \mathcal{B} = \{(b_1, b_2, \cdots, b_M) | b_i \in \{0, 1, \cdots, K_i\}, i \in [1, M]\}, \tag{4}$$

where $b_i \in \{0, 1, \cdots, K_i\}$ is the number of packets in the buffer of the i-th SU. The PU occupancy state of all J channels at the t-th time slot is represented by:

$$\boldsymbol{o} \in \mathcal{O} = \left\{ \left(O^1, O^2, \ldots, O^J \right) | O^j \in \{0, 1\} \right\}, \tag{5}$$

where O^j represents the PU activity state of the j-th channel at the t-th time slot. The channel condition state of J channels to M SUs is denoted by:

$$\boldsymbol{c} \in \mathcal{C} = \{(c_1, c_2, \cdots, c_M) \mid c_i \in [1, N_{\mathrm{SNR}}(i)]\}, \tag{6}$$

where c_i is the channel condition state of the PU channels, and $N_{\mathrm{SNR}}(i)$ is the number of channel condition states of the i-th SU. Details are discussed in Appendix A.

Finally, the state space of the system can be denoted using $\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}$ in Equations (4) to (6) as follows:

$$\Phi \triangleq \{(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) | \boldsymbol{b} \in \mathcal{B}, \boldsymbol{o} \in \mathcal{O}, \boldsymbol{c} \in \mathcal{C}\}. \tag{7}$$

Then, the transition matrix of the system $\boldsymbol{P}(\boldsymbol{D})$ can be derived, given any DPM $\boldsymbol{D}$. The derivation of $\boldsymbol{P}(\boldsymbol{D})$ is shown in Appendix B.

III. PERFORMANCE EVALUATION AND OPTIMIZATION METHOD

A. PERFORMANCE MEASURES DERIVATION

The QR algorithm [19] can be applied to the transition matrix $\boldsymbol{P}(\boldsymbol{D})$ to get the steady probability vector: $\boldsymbol{\pi}'$, which is a 1 by $\prod_{i=1}^{M}(K_i + 1) \times 2^J \cdot \prod_{i=1}^{M} N_{\mathrm{SNR}}(i)$ vector. Then, $\boldsymbol{\pi}'$ is reorganized in the form of $\pi(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$ by mapping a vector $\{\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}\}$ to any system state.

1) QUEUE LENGTH DISTRIBUTION

Let $P(q_i = l)$ denote the probability that l packets are waiting to be transmitted in the i-th SU's buffer; it can be calculated as follows:

$$P(q_i = l) = \sum_{\boldsymbol{b} \in B_i(l)} \boldsymbol{\pi}(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}), \tag{8}$$

where $B_i(l) = \{(b_1, \cdots, b_M) | b_i = l, b_j \in [0, K_j], j \neq i\}$, which represents the set of states of SU buffers when there are l packets in the i-th SU's buffer.

2) AVERAGE THROUGHPUT

In this research, the expected number of packets that can be transmitted during one time slot is used to represent the average throughput of an SU. First, the distribution of the number of packets the i-th SU can transmit during one time slot under each system state denoted by $\boldsymbol{t}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$ can be obtained as:

$$\boldsymbol{t}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) = \sum_{A \in \boldsymbol{A}} d_i^j O^j \boldsymbol{\mu}_i(A, \boldsymbol{o}, \boldsymbol{c}), \tag{9}$$

where $\boldsymbol{\mu}_i(A, \boldsymbol{o}, \boldsymbol{c})$ is the distribution of number of packets that can be transmitted by the i-th SU when the allocation result is $A \in \boldsymbol{A} = \{a_1, \cdots, a_J \mid a_j \in \{0, 1, \cdots, M\}\}$. $\boldsymbol{A}$ is the set of all possible allocation results: $a_j = 0$ represents that the j-th channel is occupied by PU, and $a_j = i$ represents that the j-th channel is allocated to the i-th SU. The PU occupancy state is $\boldsymbol{o}$ and the channel condition state is $\boldsymbol{c}$ from eq. (26). d_i^j is the element of the DPM $\boldsymbol{D}$ and O^j is the occupancy state of the j-th channel obtained from $\boldsymbol{o}$. Then, the average throughput under state $(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$ can be calculated as:

$$\mathrm{Thr}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) = \sum_{\beta} \min(\beta, b_i) \cdot t_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}, \beta), \tag{10}$$

where $t_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}, \beta)$ is the $(\beta + 1)$-th element of $\boldsymbol{t}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$, which is the probability that β transmissions have been made by the i-th SU under $(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$. Thus the average throughput of the i-th SU can be derived as:

$$\overline{\mathrm{Thr}}_i = \sum_{(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) \in \Phi} \mathrm{Thr}(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) \cdot \boldsymbol{\pi}(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}). \tag{11}$$

B. WEIGHTED THROUGHPUT

With the average throughput of every SU obtained by the proposed framework, it is easy to carry out the overall analysis based on the average throughput. Here, the concept of weighted throughput mentioned in Section I is applied:

A parameter of weight is assigned to each SU that represents the level of importance. The weights of all SUs is represented by a set of positive parameters: $\boldsymbol{w} = \{(w_1, w_2, \cdots, w_M) \mid w_i \in \mathbb{R}^+\}$, where w_i is the weight of the throughput of the i-th SU. The total weighted throughput of the SU system given the DPM $\boldsymbol{D}$ can be obtained as follows:

$$\mathrm{W}(\boldsymbol{D}) = \sum_{i=1}^{M} w_i \cdot \overline{\mathrm{Thr}_i}. \tag{12}$$

With the help of the weighted throughput $\mathrm{W}(\boldsymbol{D})$, evaluation and optimization of the performance of a CR system that serves SUs with different priorities can be carried out.

C. MEASURE OF DELAY

1) DISCUSSION ON CALCULATION OF DELAY

The common approach to delay calculation is to make use of the arrival rates, service rates and the buffers' settings to evaluate the length of time it takes a packet to be transmitted from its arrival. In the queueing analysis of the CR system, the delay is usually calculated as the time interval from when one tagged packet arrives to when it has been transmitted, which is actually the transmission time of the packets ahead of the tagged one. Therefore, all transmissions of the packets in front of the tagged packet need to be considered step by step, which makes the analysis complicated, especially when there are multiple transmission rates (e.g. when the AMC is used).

Accuracy is one of the advantages of the calculation of delay mentioned above. However, the computing resource consumption of the calculation of delay is high. Therefore, if computing efficiency is required in a CR system, it may be inappropriate to apply this calculation to evaluate the system. For instance, a CR system uses the measure of delay as a reference to allocate the PU channels at the beginning of a time slot, which is probable if the performance of delay is a matter of concern. In this CR system, the longer it takes to calculate the measure of delay, the less time remains for the data transmission. To make the calculation of delay simple, a different approach to the measure of delay is proposed based on the following idea: When a packet was transmitted, there should be a certain number of packets remaining in the buffer. Thus, the time interval during which all the remaining packets arrive at the buffer can be used to represent the measure of delay of the transmitted packet. The principle behind this approach is that since the information on the service rates are already contained in the number of packets left in the buffer (which is represented by the distribution of the queue length), it is possible to use the distribution of the queue length as the replacement of the service rates to estimate the measure of delay.

2) CALCULATION OF THE MEASURE OF DELAY

According to the discussion above, the measure of delay is defined as the waiting time of a packet from the time slot when it arrives at the buffer to the time slot when it leaves the buffer. In the time slot a packet is being transmitted, it is assumed that there are x packets in the buffer. Thus, the estimation of the time interval during which x packets have arrived can be used to estimate the waiting time of the packet being transmitted. The actual waiting time t' consists of two parts: The first part is the time t during which x packets have arrived; the second part is the time slots when continuous packet rejection occurs when the buffer is full. It is obvious that when the buffer size of the SU in the model goes to infinity, the rejection probability of packets converges to zero. Thus, to trade accuracy for efficiency, only the first part can be used to estimate the measure of delay. The error caused by the omission of the second part can be eliminated by increasing the buffer size in our model.

Then, the probability that the packet arrival time t is greater than a preset length of time is used to estimate the measure of delay in the framework. This probability is named overtime probability (OP) and the preset length of time is named the delay reference. The OP is obtained as follows: First, the delay reference is represented by an integer d_{ref}, which is the number of time slots. For the i-th SU, when there are x packets in the buffer, the probability that arrival time t equals j time slots is:

$$P_i(t=j) = \sum_{y=1}^{v} \alpha_i(y) \cdot \mathrm{Conv}(\boldsymbol{\alpha}_i, j-1, x-y), \tag{13}$$

where $\mathrm{Conv}(\boldsymbol{\alpha}, i, j)$ is the j-th element of vector $\underbrace{\boldsymbol{\alpha} * \boldsymbol{\alpha} * \cdots * \boldsymbol{\alpha}}_{i\text{‘}\boldsymbol{\alpha}\text{’}s}$, $*$ is the discrete convolution of vectors and $\boldsymbol{\alpha}_i$ is the vector that represents the arrival process of the i-th SU in eq. (3).

Finally, the OP can be obtained, given the delay reference d_{ref} time slots:

$$\begin{aligned} P_i(d_{\mathrm{ref}}) = P(q_i = 0)\alpha(0)^{d_{\mathrm{ref}}} + \sum_{x=1}^{K_i} P(q_i = x) \\ \cdot \Big(1 - \sum_{i=1}^{d_{\mathrm{ref}}} \sum_{y=1}^{v} \alpha_i(y) \cdot \mathrm{Conv}(\boldsymbol{\alpha}_i, i-1, x-y)\Big), \end{aligned} \tag{14}$$

where $P(q_i = l)$ is the probability that there are l packets in the i-th SU's buffer as in eq. (8).

D. OPTIMIZATION PROBLEM FORMULATION AND SOLUTION

Now, using the proposed framework, the weighted throughput of the SU system and delay measures of each SU given any DPM $\boldsymbol{D}$ can be obtained. The method to obtain the optimal channel allocation using the criterion including requirements of the weighted throughput and the delay can be investigated thereafter. In this research, an optimization objective is set up with the requirements of the weighted throughput and the delay as follows: Find a DPM $\boldsymbol{D}$ that can maximize the overall weighted throughput of the whole SU system, while the measure of delay of every SU, which is represented by OP in eq. (14), is no greater than a threshold P_{delay}. Here, the weighted throughput represents the utility or benefits that the CR system can get from serving the SUs, while the delay constraints represent the guarantee of the minimum services that one SU can get from the CR system. The optimization problem can be written in the following form:

$$\begin{aligned} &\underset{\boldsymbol{D}}{\text{maximize}}\ W(\boldsymbol{D}) \\ &\text{subject to } P_i(d_{\text{ref}}) < P_{\text{delay}}, \quad \text{for } \forall i \in [1, M]. \end{aligned} \tag{15}$$

With the same concern of computing efficiency as mentioned in the delay calculation part, the SU system needs an efficient method to obtain an approximate solution of DPM under certain precisions rather than a closed form solution of the DPM. To find an approximate DPM solution of our optimization problem efficiently, a modified hill climbing algorithm [20] is applied.

The detail of the algorithm is shown in Algorithm 1. The main processes of this algorithm is: First, set a series of decreasing steps(probability): S. Then, pick the first step as the initial step and pick an arbitrary DPM as the initial start DPM. Then, construct a set of neighborhood DPM L. The distance between every DPM in L and the start DPM is the length of the step(as the Manhattan Distance in each DPM dimension). Calculate the weighted throughput and delay reference of every DPM in L and find the DPM that satisfies $P_i < P_{\text{delay}}$ with the maximum weighted throughput. Reconstruct L using the DPM we get and repeat this until we cannot find a better weighted throughput. Then, use the next step in the step series and set the the optimal DPM we get above as the initial start DPM. Repeat the whole process until all the steps in S are used. The maximum weighted throughput and the DPM we get are the optimum solutions.

One concern about the validity of hill climbing algorithm is that it may be stuck at the local optimum if the target function is not convex. Luckily, although the convexity of the weighted throughput is difficult to prove or to falsify, the numerical results (for example, the data in Figure 3 and Figure 7) show that in most cases, the weighted throughput is a convex function of DPM.

Algorithm 1 Discrete Space Hill Climbing Algorithm

Input: Weighted throughput function: W
  Delay evaluation function of SU i: P_i
  Delay reference: d_{ref}
  Set of step lengths: $S = \{\text{Step}_1, \cdots, \text{Step}_n\}$
  Initial DPM: $\boldsymbol{D}_{\text{start}} = \begin{pmatrix} d_1^1 & \cdots & d_M^1 \\ \vdots & \ddots & \vdots \\ d_1^J & \cdots & d_M^J \end{pmatrix}$
Output: optimal weighted throughput: W_{opt}
  optimal DPM: $\boldsymbol{D}_{\text{opt}}$

```
D ← D_start;
i ← 1
Step ← Step_i;
repeat
    W_current ← W(D);
    Let L be a set of DPM;
    for j in [1, J − 1] do
        for m in [1, M] do
            if d_m^j + Step ∈ [0, 1] then
                d_m^j ← d_m^j + Step
                add D to L
            end if
            if d_m^j − Step ∈ [0, 1] then
                d_m^j ← d_m^j − Step
                add D to L
            end if
        end for
    end for
    findOpt ← False;
    for all D in L do
        T(D) ← maximum in P_i of all SU given D
        if W(D) > W_current AND T(D) < d_ref then
            D_next ← D;
            W_current ← W(D);
            findOpt ← True;
        end if
    end for
    if findOpt then
        D ← D_next;
    else
        if Step = Step_n then
            endCondition ← True;
        else
            i ← i + 1;
            Step ← Step_i;
        end if
    end if
until endCondition = True;
W_opt ← W_current;
D_opt ← D;
```

IV. NUMERICAL RESULTS AND ANALYSIS

In the following part, the framework is implemented to reveal its potential to configure the radio resources to achieve certain RA objectives. Evaluation of the performance of the SU system and qualified DPM from eq. (15) under various system configurations are carried out. To make the results simple, intuitive and meanwhile without loss of generality, a model with two users and two PU channels is built. The default settings of the parameters are as follows:

- Channel condition: $P_{\text{target}} = 0.01, f_m = 10\,\text{Hz}$
- PU activity model: $p^1_{b\to b} = p^2_{b\to b} = p^1_{f\to f} = p^2_{f\to f} = 0.5$
- Arrival process: $\alpha_1 = \alpha_2 = \{0.5, 0.5\}$
- SU buffer size: $K_1 = 6, K_2 = 6$
- ARQ setting: $\text{Per} = 0.01$.

A. THROUGHPUT AND WEIGHTED THROUGHPUT

The average throughputs of both SU can be obtained simultaneously and the weighted throughput of the SU system under

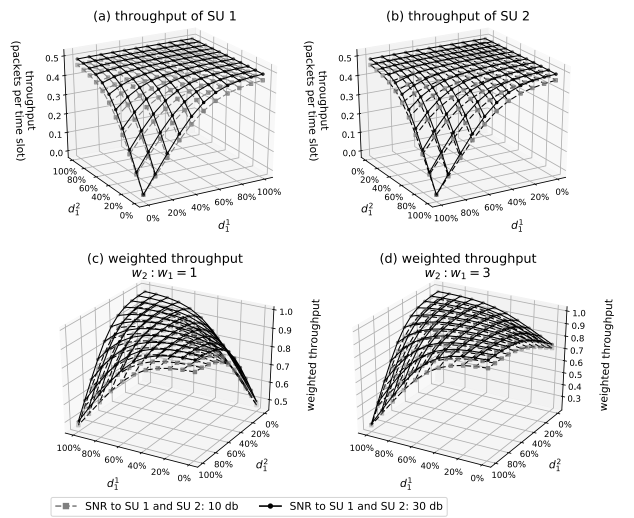


FIGURE 3. **The throughput of each SU and weighted throughput function of the SU system versus the DPM. d_i^j is the probability that the j-th channel is allocated to the i-th SU.**

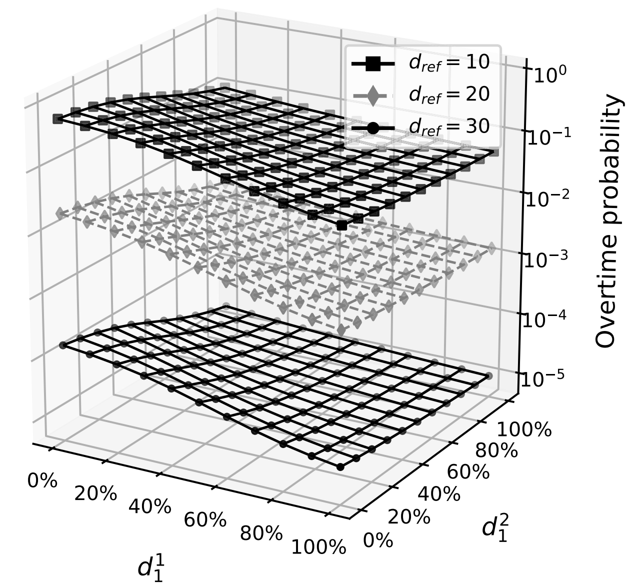


FIGURE 4. **Overtime probability versus DPM under different d_{ref}. d_i^j is the probability that the j-th channel is allocated to the i-th SU. The vertical axis is the overtime probability.**

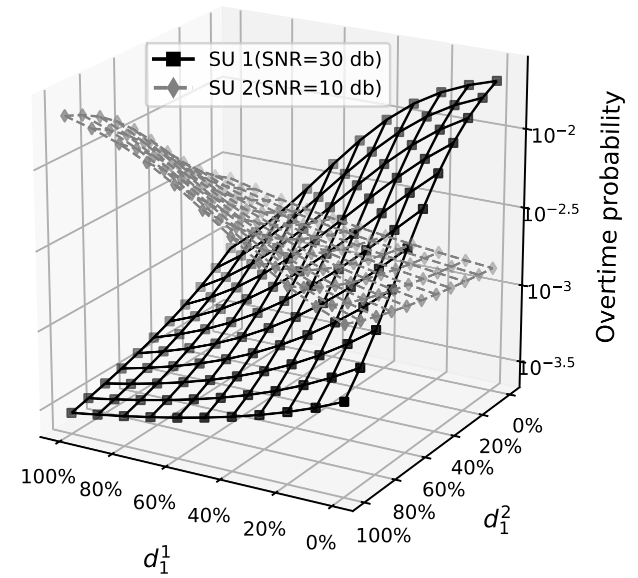


FIGURE 5. **Overtime probability of SUs with different SNRs versus DPM. $d_{\mathrm{ref}} = 20$. The SNR of SU 1 is 30dB, the SNR of SU 2 is 10dB. The vertical axis is the overtime probability.**

any weights can be also obtained directly with one single run of calculation using the proposed framework. The framework makes it convenient to compare the throughputs between different SUs and analyze the overall performance of the SUs. The comparison of throughputs of both SUs and the analysis of the weighted throughput with different weight settings under different channel conditions are shown in Figure 3. In Figure 3(a)(b), the influence of DPM on the throughput of both SUs is shown. It is evident that improvement of throughput caused by increased distribution probability is greater when the DPM is lower. However, when the DPM to an SU is high (greater than 50%), the further increase of the DPM does not improve the throughput significantly. The effect of this trend on the weighted throughput is shown in Figure 3(c) and (d). Figure 3(c) shows the weighted throughput function versus the DPM when the throughputs of both SU have the same weights ($w_1 = w_2$), while Figure 3(d) shows the weighted throughput function versus the DPM when the weight of SU 2 is greater than that of SU 1 ($w_2 : w_1 = 3$). All the trends of the weighted throughput function are identical when the SNR conditions (in terms of SNR) of SUs vary.

B. THE OVERTIME PROBABILITY

In this part, the OP that serves as an estimation of the measure of delay in this research is analyzed. Figure 4 illustrates how the delay reference affects the relationship between the probability OP and DPM. The OP of SUs with different SNRs versus the DPM is shown in Figure 5 and it is obvious that an SU with a higher SNR has a lower OP. The advantages of the proposed framework are as follows: First, the OP is determined only by the queue length distribution and the delay reference d_{ref}. Since the queue length distribution is independent of the reference, as long as the queue length distribution under certain system settings and environmental parameters is obtained, the OP with any delay reference can be determined by simple calculations. Second, since our calculation of the measure of delay is an efficient approximation, the computing time is significantly reduced compared to other delay calculation methods. A comparison of run time statistics between our method (Method I) and an example (Method II) of other methods in the related literature, such as [15], is shown in Figure 6. The details of Method II are shown in Appendix C. After repetitive calculations under identical settings, significant improvement on run time using the OP can be observed.

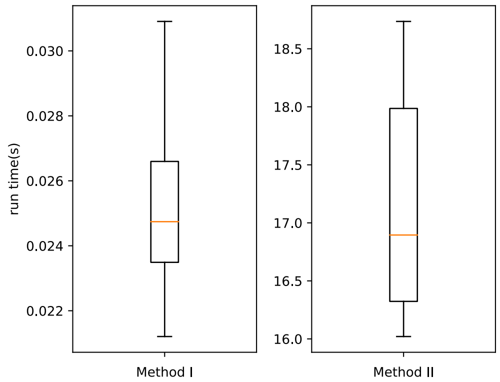


FIGURE 6. **Comparison of delay computing time statistics.**

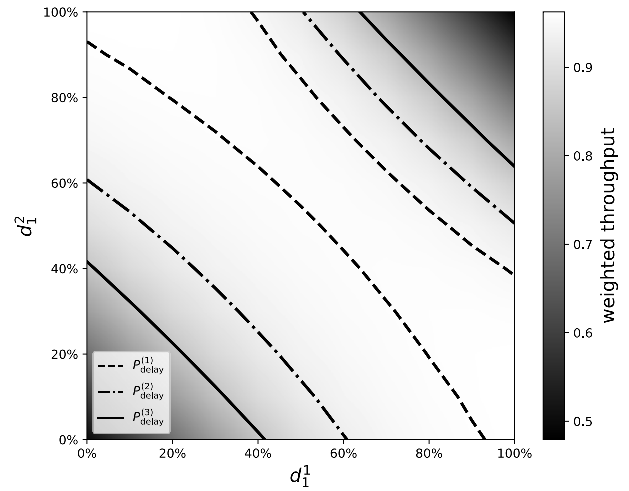


FIGURE 7. **Illustration of the optimization problem. The settings are as follows: SNR of SU1 and SU2: 10dB, $w_2 : w_1 = 10$, $d_{ref} = 20$, $P_{delay}^{(1)} = 2 \times 10^{-3}$, $P_{delay}^{(2)} = 4 \times 10^{-3}$, $P_{delay}^{(3)} = 8 \times 10^{-3}$.**

TABLE 1. **Optimization process with no delay constraints.**

iterations	d_1^1	d_1^2	$W(\boldsymbol{D})$	OP 1	OP 2
start	0.5	0.5	1.127183	5.136×10^{-3}	2.85×10^{-08}
1	0.4	0.4	1.138202	6.441×10^{-3}	1.61×10^{-08}
2	0.3	0.5	1.139135	6.421×10^{-3}	1.60×10^{-08}
3	0.2	0.6	1.141875	6.362×10^{-3}	1.55×10^{-08}
4	0.1	0.7	1.146248	6.265×10^{-3}	1.46×10^{-08}
5	0	0.8	1.151963	6.133×10^{-3}	1.36×10^{-08}
6	0	0.9	1.156396	5.398×10^{-3}	1.70×10^{-08}
7	0	0.91	1.156426	5.337×10^{-3}	1.75×10^{-08}
8	0	0.909	1.156427	5.343×10^{-3}	1.74×10^{-08}
			Parameter Settings		
SNR of SU 1	5dB	$\boldsymbol{\alpha}_1$	$\{0.5, 0.5\}$		
SNR of SU 2	30dB	$\boldsymbol{\alpha}_2$	$\{0.2, 0.8\}$		
d_{ref}	20	P_{delay}	1		

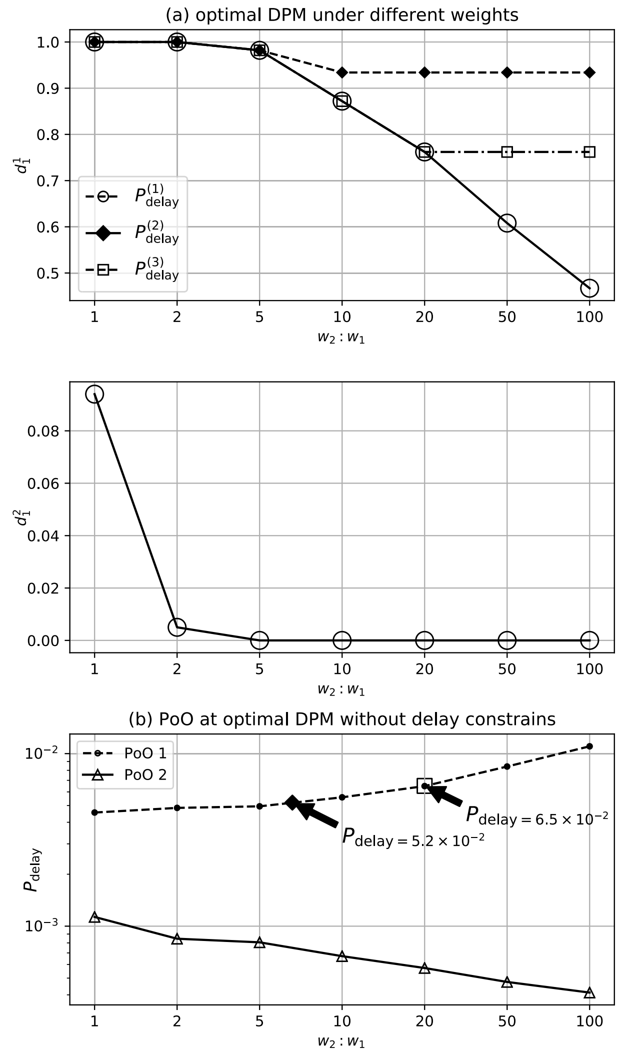


FIGURE 8. **Optimal DPM versus weights under different delay constraints. $P_{delay}^{(1)} = 1$, $P_{delay}^{(2)} = 6.5 \times 10^{-3}$, $P_{delay}^{(3)} = 5.2 \times 10^{-3}$.**

C. OPTIMIZATION AND DISCUSSION

After the weighted throughput and the OP have been obtained with the proposed framework, the proposed hill climbing algorithm can be applied to find the optimal DPM that can achieve the best weighted throughput under delay constraints. An example of the optimization problem is illustrated in Figure 7. The objective of the algorithm is to find the best weighted throughput when the OPs of both SUs are feasible. The feasible area in Figure 7 is the area between the contour lines of delay constraint P_{delay}.

The processes of the algorithm under different settings are shown in Table 1 and Table 2. It is obvious that if there are no delay constraints the algorithm can converge to the solution rapidly and the rate of convergence becomes worse if there are delay constraints.

The optimal DPM under different weights and delay constraints is shown in Figure 8. From Figure 8(a), it is evident that when there is no delay constraint ($P_{delay} = 1$), the channels need to be allocated more frequently to SU 2 to achieve better overall weighted throughput as the weight of SU 2 increases. However, the decrease in the DPM of one SU (d_1^1 and d_1^2) makes its OP increase. Thus when there are delay constraints, the increase of optimal DPM caused by the increase of weight reaches a critical point where the OP is equal to the delay constraint P_{delay}. The weight setting under

TABLE 2. Optimization process with delay constraints.

iterations	d_1^1	d_1^2	$W(\boldsymbol{D})$	OP 1	OP 2
start	0.5	0.5	1.127183	5.136×10^{-3}	2.85×10^{-8}
1	0.4	0.6	1.128324	5.124×10^{-3}	2.83×10^{-8}
2	0.3	0.7	1.131703	5.088×10^{-3}	2.75×10^{-8}
3	0.2	0.8	1.137177	5.031×10^{-3}	2.62×10^{-8}
4	0.1	0.9	1.144493	4.955×10^{-3}	2.43×10^{-8}
5	0	1	1.153264	4.864×10^{-3}	2.20×10^{-8}
6	0	0.99	1.153931	4.910×10^{-3}	2.14×10^{-8}
7	0	0.98	1.154517	4.958×10^{-3}	2.09×10^{-8}
8	0	0.97	1.155022	5.007×10^{-3}	2.03×10^{-8}
9	0	0.96	1.155448	5.058×10^{-3}	1.98×10^{-8}
10	0	0.95	1.155796	5.111×10^{-3}	1.93×10^{-8}
11	0	0.94	1.156067	5.165×10^{-3}	1.88×10^{-8}
12	0	0.939	1.156090	5.170×10^{-3}	1.88×10^{-8}
13	0	0.938	1.156112	5.176×10^{-3}	1.87×10^{-8}
14	0	0.937	1.156134	5.181×10^{-3}	1.87×10^{-8}
15	0	0.936	1.156154	5.187×10^{-3}	1.86×10^{-8}
16	0	0.935	1.156174	5.192×10^{-3}	1.86×10^{-8}
17	0	0.934	1.156193	5.198×10^{-3}	1.85×10^{-8}
Parameter Settings					
SNR of SU 1	5dB	$\boldsymbol{\alpha}_1$	$\{0.5, 0.5\}$		
SNR of SU 2	30dB	$\boldsymbol{\alpha}_2$	$\{0.2, 0.8\}$		
d_{ref}	20	P_{delay}	5.2×10^{-3}		

which the delay constraints meet is shown in Figure 8(b). The changes of the optimal DPM with delay constraints are also shown in Figure 8(a).

V. CONCLUSION

In this paper, the authors applied a method named DPM that works at the network control layer to describe the channel allocation protocol quantitatively in a multi-user multi-channel CRN. Using the DPM as one of the configurable parameters, a queueing analytical framework is established to obtain various performance measures. Using the performance measures obtained, first, an overall performance evaluation of the whole SU system is introduced with the concept of weighted throughput. The weight of each SU, which represents the level of importance to the CR system, can be configured flexibly to evaluate the benefit gained by providing services to various types of users. Then, an approximation calculation of the measure of delay, named OP, is introduced. The calculation of OP proves to be more efficient than the other measure of delay calculation methods. The OP is used to describe the delay constraints. Afterwards, an optimization objective of the SU system is set up as follows: Given all the environmental parameters and system settings as input, to determine how to set the DPM to maximize the overall weighted throughput of the whole SU system while each SU is under certain delay constraints represented by the OP. Applying a modified hill climbing algorithm, the numerical solution of the optimization problem is obtained. With the series of numerical results, the effects of various factors, such as channel conditions, delay constraints and weights on the optimal DPM are investigated. The proposed analytical framework provides comprehensive performance evaluations and key factors such as system settings and environmental parameters can easily be configured. Moreover, the performance evaluations and optimization solution obtained by the framework can serve as a comprehensive and accurate reference to design the inter-cell or inter-cluster RA allocation algorithms for the entire network. The proposed framework can serve as a convenient tool at the virtual network control layer in a CRN or HetNet, which is capable of managing and making efficient use of radio resources.

APPENDIX A
SECONDARY USER DATA TRANSMISSION MODEL

In this paper, the transmission mode on the physical layer is convolutionally coded M_n-ary rectangular or square quadrature amplitude modulation (QAM) modes, which are adopted from the HIPERLAN/2 or IEEE 802.11a standards [21].

The channel condition that determines what modulation scheme should be used at the secondary users (SU's) transceiver is divided into N^j_{SNR} states according to the signal-to-noise ratio (SNR). Under each state, the modulation schemes are set to guarantee that the average packet error rates equal a preset packet error rate P_{target}. Thus, the boundary of these states in terms of SNR can be written in [22] as:

$$\Gamma_n = \frac{1}{g_n} \ln\left(\frac{a_n}{P_{\text{target}}}\right), \tag{16}$$

where $\Gamma_n \in \{\Gamma_0, \Gamma_1, \ldots, \Gamma_{N^j_{\text{SNR}}}\}$ are the SNR boundaries of each state. Here, $\Gamma_0 = 0$ and $\Gamma_{N^j_{\text{SNR}}+1} = \infty$ and the $N^j_{\text{SNR}}+1$ SNR boundaries of the N^j_{SNR} states can be obtained. When the SNR at an SU's receiver is low ($\gamma \in (\Gamma_0, \Gamma_1)$), the SU cannot make any transmission and when $\gamma \in (\Gamma_n, \Gamma_{n+1})$, $n \in \{1, 2, \ldots, N^j_{\text{SNR}} - 1\}$, the SU is transmitting using mode n.

The model through which the channel condition state evolves from time slot to time slot can be analyzed by finite-state Markov model [25] and can be applied in the form of finite-state Markov chain in the proposed model. Without loss of generality, the channel model is assumed to be Rayleigh: the SNR is exponentially distributed with the probability density function:

$$p(\gamma^j_i) = \frac{1}{\bar{\gamma}^j_i} \exp\left(-\frac{\gamma^j_i}{\bar{\gamma}^j_i}\right), \quad \gamma^j_i \geq 0 \tag{17}$$

where γ^j_i is the SNR at the i-th SU on the j-th channel and $\bar{\gamma}^j_i$ is the average SNR. The steady state probability of the i-th SU on the j-th channel in state k can consequently be obtained by:

$$\pi^j_{\text{CS}i}(k) = \int_{\Gamma_k}^{\Gamma_{k+1}} p(\gamma^j_i)\, \mathrm{d}x. \tag{18}$$

Let $\boldsymbol{P}^j_{\text{CS}i} = \begin{pmatrix} p_{0,0} & \cdots & p_{0,N^j_{\text{SNR}}-1} \\ \vdots & \ddots & \vdots \\ p_{N^j_{\text{SNR}}-1,0} & \cdots & p_{N^j_{\text{SNR}}-1,N^j_{\text{SNR}}-1} \end{pmatrix}$ denote the state transition probability matrix of the i-th SU on the j-th channel, where $p_{a,b}$ is the transition probability from state a to state b. According to [23] and [24], the transition probabilities can be obtained by:

$$p_{k,k+1} \approx \frac{f_m T_p}{\pi^j_{\mathrm{CS}i}(k)} \sqrt{\frac{2\pi\Gamma_{k+1}}{\bar{\gamma}^j_i}} \exp\left(-\frac{\Gamma_{k+1}}{\bar{\gamma}^j_i}\right), \quad (19)$$

$$p_{k,k-1} \approx \frac{f_m T_p}{\pi^j_{\mathrm{CS}i}(k)} \sqrt{\frac{2\pi\Gamma_{k}}{\bar{\gamma}^j_i}} \exp\left(-\frac{\Gamma_{k}}{\bar{\gamma}^j_i}\right), \quad (20)$$

where f_m is the maximum Doppler frequency of the mobile station. The transition is assumed to occur only between adjacent states ($p_{j,k} = 0$, for any $|j-k| > 1$, and $p_{k,k} = 1-p_{k,k+1}-p_{k,k-1}$). With eqs. (16) to (20), the transition probability matrix $\boldsymbol{P}^j_{\mathrm{CS}i}$ can be obtained. Let $\boldsymbol{V}_{\mathrm{rate}} = \{V_{\mathrm{rate}}(k)|k \in \{0, \cdots, N^j_{\mathrm{SNR}}\}\}$ represent the number of packets that can be transmitted in one time slot when the channel condition state is k. The settings of the parameters above are listed in Table 3 and $\boldsymbol{V}_{\mathrm{rate}} = \{0, 1, 2\}$.

TABLE 3. Transmission modes with convolutionally coded modulation.

	Mode 1	Mode 2
Modulation	BPSK	QPSK
Coding rate R_c	1/2	1/2
Rate (bits/symbol)	0.5	1
a_n	274.7229	90.2514
g_n	7.9932	3.4998
Γ_n(dB)	-1.5331	1.0942

APPENDIX B
THE DERIVATION OF THE TRANSITION MATRIX

The state space of the SU system is denoted by:

$$\Phi \triangleq \{\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}\}. \quad (21)$$

The transition matrix between all the channel states $\{\boldsymbol{o}, \boldsymbol{c}\}$ is denoted by matrix $\boldsymbol{B}$ and $\boldsymbol{B}$ is obtained as:

$$\boldsymbol{B} = (\boldsymbol{P}^1_{\mathrm{PU}} \otimes \cdots \otimes \boldsymbol{P}^J_{\mathrm{PU}}) \otimes \boldsymbol{P}_{\mathrm{CS},1} \otimes \cdots \otimes \boldsymbol{P}_{\mathrm{CS},M}, \quad (22)$$

where $\boldsymbol{P}_{\mathrm{CS},i}$ is the channel condition state transition probability matrix of the i-th SU in Appendix A.

First, the distribution of the number of packets that can be transmitted needs to be derived. The distribution of transmitted packets of each SU through each channel can be obtained, given the channel condition state $\boldsymbol{c}$ according to the AMC settings in Appendix A. The number of transmission attempts that the i-th SU can make through the j-th channel is denoted as C^j_i.

$$C^j_i = \begin{cases} V_{\mathrm{rate}}(c_i) & \text{for } O^j \neq 0 \\ 0 & \text{for } O^j = 0. \end{cases} \quad (23)$$

From the ARQ settings, the probability that x packets can be transmitted can be calculated as given below:

$$\mu(x) = \begin{cases} \binom{C^j_i}{x}(1-\mathrm{Per})^x & \text{for } 0 \leq x \leq C^j_i \\ \mathrm{Per}^{C^j_i} & \text{for } x = 0, \end{cases} \quad (24)$$

where $\binom{C^j_i}{x}$ is C^j_i choose x. Here, $\boldsymbol{A} = \{a_1, \cdots, a_J \mid a_j \in \{0, 1, \cdots, M\}\}$, which is used to represent the channel allocation results. $a_j = 0$ represents that the j-th channel is occupied by PU, and $a_j = i$ represents that the j-th channel is allocated to the i-th SU. Then, the distribution of the number of packets that can be transmitted by the i-th SU through the j-th channel can be calculated:

$$\boldsymbol{\mu}^j_i = \begin{cases} \{\mu^j_i(0), \ldots, \mu^j_i(C^j_i)\} & \text{for } a_j = i \\ \{0\} & \text{otherwise.} \end{cases} \quad (25)$$

The distribution of the number of packets that can be transmitted by the i-th SU given the channel state $\{\boldsymbol{o}, \boldsymbol{c}\}$ can be obtained as follows:

$$\boldsymbol{\mu}_i(\boldsymbol{A}, \boldsymbol{o}, \boldsymbol{c}) = \boldsymbol{\mu}^1_i * \boldsymbol{\mu}^2_i * \cdots * \boldsymbol{\mu}^J_i, \quad (26)$$

where $*$ is convolution. The transition matrix describes that the number of packets in the i-th SU's buffer changes from b_i to b'_i under $\boldsymbol{A}$:

$$\boldsymbol{P}(b_i, b'_i, \boldsymbol{A}) = \begin{bmatrix} \boldsymbol{\alpha}_i\boldsymbol{\Gamma}(0,0)\boldsymbol{\mu}_i(\boldsymbol{A}) & \cdots & \boldsymbol{\alpha}_i\boldsymbol{\Gamma}(0,K)\boldsymbol{\mu}_i(\boldsymbol{A}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\alpha}_i\boldsymbol{\Gamma}(K,0)\boldsymbol{\mu}_i(\boldsymbol{A}) & \cdots & \boldsymbol{\alpha}_i\boldsymbol{\Gamma}(K,K)\boldsymbol{\mu}_i(\boldsymbol{A}) \end{bmatrix}, \quad (27)$$

where $\boldsymbol{\Gamma}(i, i')$ is an auxiliary matrix; the elements of the matrix are derived as follows:

$$\Gamma_{v,c}(i, i') = \begin{cases} 1 & \text{when } i' = \min(v + i - \min(i, c), K) \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

Let set $\boldsymbol{b} = \{b_1, b_2, \cdots, b_M\}$ denote the number of packets in each SU's buffer at the current time slot, and $\boldsymbol{b}' = \{b'_1, b'_2, \cdots, b'_M\}$ denote the number of packets at the next time slot. Then, a coefficient matrix can be obtained, given $\boldsymbol{b}$, $\boldsymbol{b}'$ and $\boldsymbol{A}$:

$$\boldsymbol{T}(\boldsymbol{b}, \boldsymbol{b}', \boldsymbol{A}) = \boldsymbol{P}(b_1, b'_1, \boldsymbol{A}) \otimes \boldsymbol{P}(b_2, b'_2, \boldsymbol{A}) \otimes \cdots \otimes \boldsymbol{P}(b_M, b'_M, \boldsymbol{A}). \quad (29)$$

Given the DPM $\boldsymbol{D}$ and the PU occupancy state $\boldsymbol{o}$, the coefficient matrix can be obtained:

$$\boldsymbol{T}_{\boldsymbol{b},\boldsymbol{b}'}(\boldsymbol{D}, \boldsymbol{o}) = \sum_{\mathrm{A}\in\mathcal{A}} \left(\prod_{j=1}^{J} d^j_{a_j} \boldsymbol{T}(\boldsymbol{b}, \boldsymbol{b}', \boldsymbol{A})\right), \quad (30)$$

where d^j_i is from the DPM $\boldsymbol{D}$ and $\mathcal{A}$ is the set of all possible allocation results. All the coefficients according to the index of channel condition states are enumerated, and extended to a matrix with the same dimensions as $\boldsymbol{B}$:

$$\boldsymbol{T}_{\boldsymbol{b},\boldsymbol{b}'} = \begin{bmatrix} \boldsymbol{T}_{\boldsymbol{b},\boldsymbol{b}'}(\boldsymbol{D}, \boldsymbol{o}(1)) \\ \vdots \\ \boldsymbol{T}_{\boldsymbol{b},\boldsymbol{b}'}(\boldsymbol{D}, \boldsymbol{o}(\phi)) \\ \vdots \\ \boldsymbol{T}_{\boldsymbol{b},\boldsymbol{b}'}(\boldsymbol{D}, \boldsymbol{o}(N_{\mathrm{CS}})) \end{bmatrix} \otimes \boldsymbol{1}_{1\times(\prod_{i=1}^{M}(K_i+1)\cdot N_{\mathrm{CS}})}, \quad (31)$$

where $\boldsymbol{o}(\phi)$ is the occupancy state that can be obtained given the system states ϕ, N_{CS} is the total number of system states.

To enumerate all coefficient matrices, β_{index} is mapped to any $\{b_1, b_2, \cdots, b_M\}$ using the following equation:

$$\text{index} = \sum_{i=1}^{M} b_i \times \text{base}_{i-1}, \tag{32}$$

where $\text{base}_i = \prod_{j=i}^{M}(K_i + 1)$.

The transition matrix is divided into $\prod_{i=1}^{M}(K_i + 1) \times \prod_{i=1}^{M}(K_i + 1)$ blocks and obtained as follows:

$$\boldsymbol{P}(\boldsymbol{D}) = \begin{bmatrix} \boldsymbol{T}_{\beta_0,\beta_0} \circ \boldsymbol{B} & \cdots & \boldsymbol{T}_{\beta_0,\beta_B} \circ \boldsymbol{B} \\ \vdots & \ddots & \vdots \\ \boldsymbol{T}_{\beta_B,\beta_0} \circ \boldsymbol{B} & \cdots & \boldsymbol{T}_{\beta_B,\beta_B} \circ \boldsymbol{B} \end{bmatrix}, \tag{33}$$

where $\circ$ is the entrywise product.

APPENDIX C
DELAY CALCULATION METHOD II

In this part, a method to calculate the measure of delay using the idea applied in related literature, such as [15], is introduced. The measure of delay is represented by the expected time to transmit all packets that are in the buffer when the packet we consider arrives. Since the adaptive modulation and coding (AMC) is applied and the transmission rate varies, the state space is extended with another state f that represents the number of packets that have already been transmitted by a selected secondary user (SU) to trace the transmission process of the packets ahead of the tagged one. The extended system state is $\Phi' \triangleq \{\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}, f\}$. The delay of a tagged packet can be interpreted as the first passage time from the system state when this packet arrives to the system state that all packets ahead of it are transmitted (number of packets in the i-th SU buffer $b_i = 0$ or f equals the number of packets in the buffer).

To calculate the first passage time above, the transition matrix based on $\boldsymbol{P}(\boldsymbol{D})$ is modified (to make it concise, the $\boldsymbol{D}$ in the equations is omitted), given that there are F packets in the selected SU's buffer. First, an extending band matrix is built, given $\{\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}\}$ and $\boldsymbol{t}_i$ obtained as in eq. (9):

$$\boldsymbol{P}_{\text{ex}}(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}, f) = \begin{bmatrix} P_{0,0} & \cdots & P_{0,f} \\ \vdots & \ddots & \vdots \\ P_{f,0} & \cdots & P_{f,f} \end{bmatrix}, \tag{34}$$

where

$$P_{j,j'} = \begin{cases} t_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}, j' - j) & \text{for } j' - j \in [0, \max(\boldsymbol{t}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}))] \\ 0 & \text{otherwise,} \end{cases}$$

where $\max(\boldsymbol{t}_i(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}))$ is the maximum number of packets that can be transmitted when the system state is $(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$. Then, the $\boldsymbol{P}$ is extended row by row as illustrated below:

$$\boldsymbol{P}'_{\text{delay}}(f) = \begin{bmatrix} \boldsymbol{P}(\phi(0)) \otimes \boldsymbol{P}_{\text{ex}}(\phi(0), f) \\ \vdots \\ \boldsymbol{P}(\phi(R_m)) \otimes \boldsymbol{P}_{\text{ex}}(\phi(R_m), f), \end{bmatrix} \tag{35}$$

where $R_m = \prod_{i=1}^{M}(K_i + 1) \cdot N_{\text{CS}}$ is the size of $\boldsymbol{P}$, and $P(\phi(r))$ is the r-th row of matrix $\boldsymbol{P}$. $\phi(i)$ represents the system state of the i-th row in $\boldsymbol{P}$. The absorbing states are reached when all the packets in the buffer are transmitted ($b_i = 0$ or $f = F$). Then, $\boldsymbol{P}_{\text{delay}}$ is obtained by removing the absorbing states from $\boldsymbol{P}'_{\text{delay}}$. The average first passage time to the absorbing states under each state is denoted as $\boldsymbol{S}$, which is a $(K_s \cdot \prod_{i \neq s}(K_i + 1) \cdot N_{\text{CS}})$-by-1 column vector ($s$ is the label of the selected SU). Then, $\boldsymbol{S}$ can be solved from the following equation system:

$$(\boldsymbol{I} - \boldsymbol{P}_{\text{delay}}) \cdot \boldsymbol{S} = \boldsymbol{1} \tag{36}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{1}$ is a column vector whose elements are all 1 and the size is the same as $\boldsymbol{S}$. Let $\boldsymbol{S}(F)$ be the solution of $\boldsymbol{S}$ given F. Thus the average first passage time from the state of F packets in the buffer to the empty buffer state is:

$$d(F) = \sum_{\boldsymbol{b} \in B_s(F)} S(F, \boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}) \cdot \boldsymbol{\pi}(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c}), \tag{37}$$

where $B_s(F)$ is the set of all states of SU buffers when the selected SU's buffer has F packets in it and $S(F, \boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$ are the elements in $\boldsymbol{S}$ when the system state is $(\boldsymbol{b}, \boldsymbol{o}, \boldsymbol{c})$. Finally the average delay of the i-th SU can be obtained as:

$$\overline{\text{Dly}} = \sum_{t=1}^{K_i} d(t). \tag{38}$$

REFERENCES

[1] F. Granelli, A. A. Gebremariam, M. Usman, F. Cugini, V. Stamati, M. Alitska, and P. Chatzimisios, "Software defined and virtualized wireless access in future wireless networks: Scenarios and standards," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 26–34, Jun. 2015.

[2] J. Mitola and G. Q. Maguire, Jr., "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Apr. 1999.

[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[4] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.

[5] B. S. Awoyemi, B. T. J. Maharaj, and A. S. Alfa, "Solving resource allocation problems in cognitive radio networks: A survey," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 176, 2016.

[6] A. S. Alfa, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System*. New York, NY, USA: Springer, 2010.

[7] W. Xu, R. Qiu, and J. Cheng, "Fair optimal resource allocation in cognitive radio networks with co-channel interference mitigation," *IEEE Access*, vol. 6, pp. 37418–37429, 2018.

[8] V. Rodriguez and D. J. Goodman, "Power and data rate assignment for maximal weighted throughput in 3G CDMA," in *Proc. IEEE Wireless Commun. Netw. (WCNC)*, vol. 1, Mar. 2003, pp. 525–531.

[9] P. Baptiste, M. Chrobak, C. Dürr, W. Jawor, and N. Vakhania, "Preemptive scheduling of equal-length jobs to maximize weighted throughput," *Oper. Res. Lett.*, vol. 32, no. 3, pp. 258–264, 2004.

[10] F. Y. L. Chin, M. Chrobak, S. P. Y. Fung, W. Jawor, J. Sgall, and T. Tichý, "Online competitive algorithms for maximizing weighted throughput of unit jobs," *J. Discrete Algorithms*, vol. 4, no. 2, pp. 255–276, 2006.

[11] M. Rashid, M. Hossain, E. Hossain, and V. Bhargava, "Opportunistic spectrum scheduling for multiuser cognitive radio: A queueing analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5259–5269, Oct. 2009.

[12] Y. B. Zikria, F. Ishmanov, M. K. Afzal, S. W. Kim, S. Y. Nam, and H. Yu, "Opportunistic channel selection MAC protocol for cognitive radio ad hoc sensor networks in the Internet of Things," *Sustain. Comput., Inform. Syst.*, vol. 18, pp. 112–120, Jun. 2018.

[13] J. Wang, A. Huang, L. Cai, and W. Wang, "On the queue dynamics of multiuser multichannel cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1314–1328, Mar. 2013.

[14] S. Wang, B. T. Maharaj, and A. S. Alfa, "Queueing analysis of performance measures under a new configurable channel allocation in cognitive radio," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9571–9582, Oct. 2018.

[15] L. B. Le, E. Hossain, and T. S. Alfa, "Delay statistics and throughput performance for multi-rate wireless networks under multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3234–3243, Nov. 2006.

[16] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006.

[17] A. J. Goldsmith and S.-G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, May 1998.

[18] T. G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. New York, NY, USA: Springer-Verlag, 2012.

[19] J. G. F. Francis, "The QR transformation a unitary analogue to the LR transformation—Part 1," *Comput. J.*, vol. 4, no. 3, pp. 265–271, 1961.

[20] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

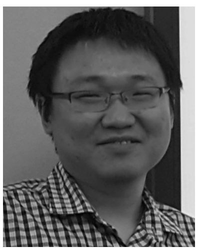
[21] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 172–180, May 2002.

[22] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.

[23] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[24] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[25] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels—A survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.



SHI WANG received the B.Eng. and M.Eng. degrees from the Beijing University of Telecommunication, Beijing, in 2006 and 2009, respectively, and the Ph.D. degree in engineering from the University of Pretoria, Pretoria, South Africa, in 2018. He was a Lecturer with the School of Electronic and Information Engineering, Liaoning Technical University, Huludao, China. His research interests include cognitive radio networks, radio resource allocation in wireless networks, and queuing theory.

SUNIL MAHARAJ received the Ph.D. degree in engineering the area of wireless communications from the University of Pretoria, where he is currently a Full Professor and holds the research position of Sentech Chair in Broadband Wireless Multimedia Communications with the Department of Electrical, Electronic, and Computer Engineering. His research interests include OFDM-MIMO systems, massive MIMO, cognitive radio resource allocation, and 5G cognitive radio sensor networks.

ATTAHIRU S. ALFA is currently a Professor Emeritus with the Department of Electrical and Computer Engineering, University of Manitoba, and a UP/CSIR co-hosted SARChI Chair Professor with the Department of Electrical, Electronic, and Computer Engineering, University of Pretoria. He also works in the application of queuing theory to other areas, such as transportation systems, manufacturing systems, and healthcare systems. He has authored two books, *Queuing Theory for Telecommunications: Discrete Time Modeling of a Single Node System* (Springer, 2010) and *Applied Discrete-Time Queue* (Springer, 2015). He has also coauthored a recent book *Optimization Methods for User Admissions and Radio Resource Allocations for Multicasting Over High Altitude Platform* (River Publishers as a series in communications, 2019). His research interests include queuing theory and applications, optimization, performance analysis and resource allocation in telecommunication systems, modeling of communication networks, analysis of cognitive radio networks, modeling and analysis of wireless sensor networks, developing efficient decoding algorithms for LDPC codes, channel modeling, traffic estimation for the Internet, and cross layer analysis.