IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 24, 2019, accepted June 6, 2019, date of publication June 17, 2019, date of current version June 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923563

Performance Analysis of Multi-Modal Overlay/Underlay Switching Service Levels in Cognitive Radio Networks

ATTAHIRU SULE ALFA[1,3], (Member, IEEE), HAITHAM ABU GHAZALEH[2], (Member, IEEE), AND BODHASWAR T. MAHARAJ[4], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada
[2]Department of Engineering and Computer Science, Tarleton State University, Stephenville, TX 76402, USA
[3]Department of Electrical, Electronic and Computer Engineering (CSIR/UP SARChI ASN Chair), University of Pretoria, Pretoria 0002, South Africa
[4]Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria 0002, South Africa

Corresponding author: Haitham Abu Ghazaleh (abughazaleh@tarleton.edu)

This work was supported in part by Grant N00590/89985 from the National Research Foundation (NRF) of South Africa SARChI ASN Chair, along with a Faculty Development Grant from Tarleton State University.

ABSTRACT Cognitive radio networks have become a popular platform for many systems and applications of the future, and especially for Smart City applications and the Internet of Things. The wireless transceivers used for communicating between the different devices in the cognitive radio networks can operate in either of the two modes, namely overlay and underlay, or a hybrid of these two modes. While operating in the underlay mode, secondary users are likely to experience varying transmission rates due to the fluctuating power levels from the primary users. This has the effect of the channel capacity being dynamic, which forces the secondary user to switch between different transmission rates, or ‘‘service modes’’, during a single networking session. In our previous work, we developed a discrete time queueing model for analyzing the performance of secondary users in such networks with multi-modal and hybrid overlay/underlay switching service levels. In this paper, we extend our previous work to present our novel result for computing the waiting-time distribution of the secondary users. Such results are essential for investigating the sensitivity of the secondary user's performance due to the queueing delays, and especially for real-time applications.

INDEX TERMS Cognitive radio networks, hybrid overlay/underlay modes, discrete-time queues, waiting time distribution.

I. INTRODUCTION

The usage of wireless communication technologies have become widespread and continues to facilitate the interconnectivity between a broad number of users and devices worldwide. In a report that had been published by Cisco [1], it was revealed that there were around 8 billion mobile users connected to the Internet in 2016. This number was further estimated to increase to 11.6 billion devices by 2021, which would exceed the world's projected population of 7.8 billion at that time. Wireless sensor networks [2] are among the plethora of applications that have seen a growing demand for wireless connectivity. Hence, the need for exploring efficient spectrum management techniques that are necessary in meeting such high demands.

The advent of the Internet of Things (IoT) is one of the prominent drivers for the rapid growth in wireless access and connectivity [3]. Networks in the foreseeable future will be expected to efficiently manage a wireless medium that is capable of supporting the variety of network-accessing devices. The network connectivity will further be dominated by objects, or ‘‘things’’, with sensing capabilities that primarily engage in machine-to-machine communication. A vast amount of research continues to investigate the deployment of cognitive radio network of sensors for improving the performance of several and complex applications, such as those in Smart Cities [4]. For instance, a large number of wireless sensor nodes with cognitive radio capabilities are expected to be deployed in cities for supporting a wide range of interconnected services, such as smart traffic monitoring systems and smart parking. Thus, such solutions must be proficient at managing the limited network resources (or bandwidth) for supporting the abundance of devices and data, while maintaining the desired performance demands.

The associate editor coordinating the review of this manuscript and approving it for publication was Theofanis P. Raptis.

In many applications, it is common to have a certain part of the spectrum statically allocated and dedicated to a crowd of users that are administered by a specific wireless network. Wireless networks have traditionally been allocated a fixed part of the spectrum that are dedicated for its own users. An expansion of the network would have typically required additional radio spectrum bands to be sought and purchased. Nowadays, the acquisition of more spectrum for fulfilling the ongoing and increasing demands have been infeasible for many network operators due to the scarcity and steep pricing of the radio spectrum. This had motivated the need to pursue other alternatives and to explore new ways of efficiently utilizing the existing spectrum.

One of the important findings given in a report that was released by The Federal Communications Commission (FCC) [5] is that most of the radio spectrum that had been allocated to licensed users are considerably underutilized. In an effort to address the scarcity of the wireless spectrum, the FCC considered the opportunity of the unlicensed users being allowed access to the wireless channels that are licensed to others, provided that such access is performed with no disruptions and minimal interference to the licensed user's operations. Cognitive Radio (CR) technologies [6] were introduced for the main purpose of exploiting the underutilized licensed spectrum, or ''spectrum holes'' [7], and in the effort of enhancing the overall network performance without the need to acquire additional spectrum. Networks that administer devices with CR technologies would be better equipped at efficiently managing the existing and allocated spectrum bands.

In CR networks, unlicensed users, also known as secondary users (SU), are expected to utilize the channels (when needed) that are licensed to its primary users (PU) in such a manner that does not disrupt the PU's performance. Unlike the overlay mode, an SU that is operating in the underlay mode will be required to strictly manage its power levels for transmission. The SU's power level would need to be sufficiently low to avoid any significant interference to the PU, while also being adequately high enough to provide a satisfactory signal-to-noise ratio (SNR) for its transmission [8]. In such situations, the SU's power levels may fluctuate during its transmission session and relative to the varying channel conditions that are mainly imposed by the PUs. Hence, the SU's transmission rates can be dynamic and the varying channel conditions would further impact the SU's effective transmission rate. Furthermore, the SU in the underlay mode may be in a situation where the best power level that is permissible with the current channel conditions is insufficient and lower than the required minimum. During such instances, the SU's transmissions are suspended until the channel is released by the PU, or the conditions change to permit the transmissions to continue with an adequate SNR.

The results shared in this paper are focused on presenting a method for analyzing the influence of the dynamic switching between the overlay and underlay modes, as well as the switching of the different service rates within the underlay mode, on the overall performance of CR networks. This analysis can aid with reconfiguring the network's resources and mechanisms for the purpose of improving the overall performance in terms of various metrics that are of importance to network designers and operators, such as throughput and latency. Our analyses considers the hybrid overlay/underlay transmission model, also known as sensing-based spectrum sharing [9]. We further focus only on flow-level analyses for evaluating the various traffic statistics and do not consider packet-level analyses in our work.

The various modes are modeled as different ''Service Levels'' for the SU. The highest service level with the maximum transmission rate is achieved while the SU is in the overlay mode, while lower service levels with the varying and lower rates are attained in the underlay mode. Note that the switching between the different service levels is a consequence of the fluctuations in the channel conditions and is not triggered by the network users. For each service level, the service time can follow a particular probability distribution that are not necessarily similar to all the other different service levels. We refer to such behaviors as a multi-modal service event and its analysis is equivalent to examining queues with working vacations, whereby the mode with the highest rate is the server at full service and the remaining modes being equivalent to the working vacation services.

In addition to showing how to model this multi-modal service behavior for SUs, our other main contribution in this paper is the computation of the waiting-time distribution of the SUs that are queued for service. This type of performance measure is essential for assessing the delays incurred by the SUs in the system. Such forms of assessment are especially crucial in real-time applications with temporally-stringent constraints, or applications where the sensory data are valid/relevant for very short periods of time. The information gained from the distribution functions of the waiting times can also be used to aid with maximizing the CR node's energy efficiency, with constraints on the age of the data for transmission [10].

In this paper, we propose an extension to the queueing model presented in our previous work [11] for analyzing the behavior of the SUs in a CR network that is operating in the hybrid overlay/underlay mode. The previous work was focused on the case of the same system but with an infinite buffer capacity. The performance metrics derived in our previous work were also limited to the average number of SUs in the buffer and without the waiting-time analysis. We first present in Section III a detailed description of the system that is being modeled, including a summary of the process that models the switching service levels of the SU being served. In Section IV, we present the details of our discrete-time queueing mode for the more practical case of a system with a finite buffer capacity. In the same section, we further show how to compute the standard performance metrics that are commonly used for analyzing the performance of such systems, such as the average number of SUs in the buffer, along with the waiting-time distribution. We present the same derivations and performance metrics in Section V but for the case of a system with an infinite buffer capacity, which could serve as an approximation for the finite buffer case with very large capacities. Prior to concluding the paper with a few future research directions, some numerical examples are given in Section VI that highlights the application of our proposed model.

II. RELATED WORK

There have been various research work in the past that were aimed at enhancing the power allocation and spectrum sharing mechanisms for SUs in CR networks, such as those given in [12], [13] and [14]. These results rely on having a model that best describes the performance of the CR networks. There have also been many work done at modeling the performance of SUs in CR networks with overlay/underlay spectrum sharing strategies. But all have either imposed relatively simplistic assumptions for the sake of a tractable analysis, or have narrowed their scope of analysis. For instance, the authors in [15] had presented a queueing model than can be used to investigate the impact of service interruptions on the SUs due to the PU's usage behaviors of the licensed channel. However, their analysis was limited to only considering the overlay access mechanism for the SU's mode of operation.

In [16], the authors had presented an M/G/1 queueing model that was proposed for analyzing the performance of a single SU in terms of throughput and latency, within the network coverage area of a single PU. A similar model was also proposed by the authors in [17] that had assumed a system with a finite buffer to yield an M/G/1/K queueing model. Their analysis had also assumed that the wireless channels were subjected to Nakagami-*m* fading and interference. To expand on the single SU analysis, the authors in [18] presented an M/M/1 queueing model for evaluating the performance of CR networks with multiple SUs. Their model assumed heterogeneous arrival and service rates, but further assumed exponentially distributed inter-arrival and service times which are unrealistic. In [19], the authors proposed a spectrum sharing scheme that permits two SUs simultaneously utilizing the same channel and modeled the behavior of the CR network as an M/M/1 queueing system. However, their analysis was limited to investigating the mean queue lengths of backlogged SUs and our proposed analysis considers CR networks where multiple and simultaneous SU transmissions are not permitted.

The majority of the previous work had collectively assumed a single service mode within the underlay access mechanism, while also considering the switching to the overlay mode during an SU's operations. This assumption is insufficient at accurately accounting for the variations in the channel conditions due to the behavior of the PUs. Another assumption that was commonly applied is the memoryless arrival and service behaviors (such as Poisson) that unrealistically models the data traffic behaviors for many networking systems. A more practical assumption was imposed by the authors in [20] and proposed the arrivals of the SU requests as a Markovian Arrival Process (MAP), with the service times following a phase-type distribution. Such assumptions would yield a better accuracy in the results due to the more realistic traffic models. However, their results were constrained to modeling the SU's performance in the overlay mode. In [21], the authors developed the Markov Chain model for the two-mode overlay/underlay switching mechanism. Our results expand on the two service modes to multiple and distinct service levels, while also being used for evaluating various performance measures such as transmission backlog and latency.

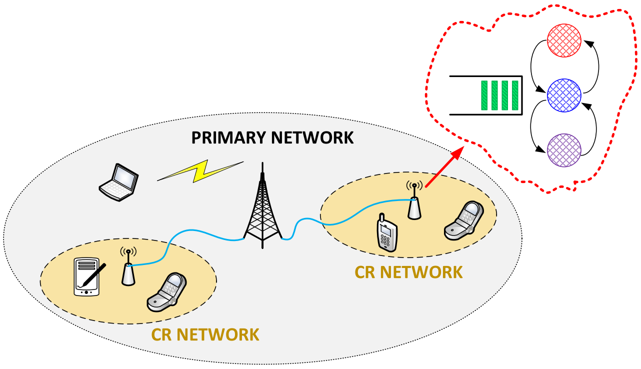

FIGURE 1. A system of cognitive radio networks within a primary network.

In many of the related work, such as [22] and [23], the delay characteristics of the system were inferred from analyzing the queue lengths and their first moments (or mean queue lengths and waiting times). Some of these related work, such as [24], were limited to only evaluating the average waiting times. While these results may provide a good estimate of the system's overall performance, they do not necessarily yield accurate results in terms of the SU's waiting-time behaviors. They further do not disclose the details of the variations in the waiting times that are substantially important for examining the SU's delay tolerances with real-time applications. In [25], the authors had shown how to compute the distribution functions for the waiting times of the CR nodes. However, their analysis was formulated as an M/G/1 system and does not consider the multi-modal service behavior of the SUs.

III. SYSTEM DESCRIPTION

In a typical scenario, there will be a set of CR networks that operate within the coverage area of a primary network, as shown in Fig. 1. The primary network serves the PUs using a dedicated set of frequency channels from the licensed spectrum, whereas the users in the CR network can either operate using the unlicensed spectrum or attempt to utilize a licensed channel from the primary network whenever it is permissible. A user from the CR network is seen as an SU from a primary network's perspective, especially when attempting to utilize a licensed channel that is dedicated to the PUs.

Our analysis consists of modeling the collective behavior of multiple SUs within a CR network that comprises of a single central node (or base station). All SU transmissions are assumed to be managed by the central node and utilizing a single radio channel that is licensed to a PU. This central node can be equivalent to a cluster head in which all the sensor nodes within its network coverage can directly communicate with it. The system can be modeled as a single server queue with a capacity K, where the queueing occurs at the central node for buffering the arrivals of the transmission requests by the SUs, and the single server represents the channel access. We further propose to model the system as a discrete time process where the event changes occur at discrete time epochs.

The arrivals of the transmission requests from the multiple SUs in the network are served in the order of the arrivals, i.e. on a first-come-first-serve (or FIFO) basis. The central node is assumed to manage the priority access of the single licensed channel, with the highest priority given to the PU and followed by the SU that is at the head of the queue. Even though cognitive sensor nodes are potentially capable of operating with the use of multiple frequency channels (both licensed and unlicensed), we restrict our analysis to the nodes that communicate using only one of the frequency channels from the spectrum band that is licensed to the primary network.

In the following subsections, we present a summary of the formulations for the arrival process, the multi-modal service operation, and the switching service process utilized in the proposed queueing model. The full details of these processes can be reviewed in [11].

A. THE ARRIVAL PROCESS

The inter-arrival times of the SUs and their transmission requests are assumed to be modeled by the discrete Markovian Arrival Process (MAP) with the sub-stochastic matrices D_0 and D_1 of dimensions $n \times n$, where $D = D_0 + D_1$ is a stochastic matrix. Let $(A)_{ij}$ be elements of any matrix A, then $(D_k)_{ij},\ k = 0, 1$, captures the probability of transition from a state i to state j, with k arrivals. This type of process can be used to model a general distribution and is suitable for further modeling the correlation between the inter-arrival times. The arrival rate λ of SUs into the system can be calculated as follows,

$$\lambda = \pi D_1 \mathbf{1}, \tag{1}$$

with π being the stationary vector of the Markov Chain represented by the stochastic matrix D that satisfies the equations $\pi D = \pi$ and $\pi \mathbf{1} = 1$. Note that the term $\mathbf{1}$ is defined as a column vector of ones with the appropriate dimensions.

B. THE MULTI-MODAL SERVICE OPERATION

The overlay and underlay access mechanisms in CR networks govern the transmission rates that an SU can utilize for communicating its data. The SUs adjust for these rates by varying its transmission power levels and based on the channel conditions. In the overlay mode, the SU can transmit its data using the maximum power level without the risk of interfering with the PU, thereby achieving the highest transmission rate. However, these transmission rates are reduced while the SU is operating in the underlay mode due to it being forced to lower its power levels for avoiding any service disruptions to the PU.

Many of the models proposed by other researchers assume a homogeneous service behavior while the SU is transmitting in the underlay mode. This neglects the possibility of the SU having to vary its transmission power (and effectively also its transmission rate) to accommodate the changing power levels of the PU during transmission. Such variations are considered in our proposed model and is incorporated into our formulation of the multi-modal switching service process.

Let m be defined as the service mode number such that the set of different modes are $m = 0, 1, 2, \cdots, M$, with $M + 1$ being the total number of service modes. An SU is assumed to initiate its service under any of the first M different modes (i.e. $0 \le m \le M - 1$), with a specific probability. The mode $m = M$ corresponds to the situation where the SU suspends its service due the channel being occupied by the PU and no adequate power level is possible to transmit in the underlay mode. We assume that the SU is transmitting with the highest rate at mode $m = 0$, which corresponds to the SU operating in the overlay mode. The SU is operating in the underlay mode when its service modes are within the range $1 \le m \le M - 1$, where a higher transmission rate is achievable in mode $m = i$ relative to mode $m = i + 1$.

Let H_m be defined as a discrete random variable (with finite supports) that model the service completion time while the SU is operating with the service mode m. We assume the random variable H_m can be modeled by a discrete phase-type (PH) distribution with the representation $(\boldsymbol{\delta}_m, V_m)$ of order ℓ_m, with $v_m = \mathbf{1} - V_m\mathbf{1}$, and for $m = 0, 1, 2, \cdots, M$, such that,

$$Pr\left\{H_m = t\right\} = \boldsymbol{\delta}_m V_m^{t-1} v_m \qquad t = 1, 2, \cdots \tag{2}$$

We assume the service completion time H_m includes both the packet processing time at the node for transmission, along with the transmission time across the wireless medium (including any transmission retrials due to errors and collisions).

During an SU's transmission session, its service mode can switch from one mode to another, and the switching can occur a multiple number of times throughout the session. This implies that the stochastic behavior of the service times can differ within an SUs sessions as its service mode alternates and due to the dynamic channel conditions. We next show how to model the service time behavior for an SU with switching service modes.

C. THE SWITCHING SERVICE PROCESS

The service is performed by a single server with the completion time S being modeled by a probability distribution given by $s_i = Pr\{S = i\}$, for $i = 1, 2, 3, \cdots, \tau$. This service time S corresponds to the total time for an SU to complete its data transmission for a given session. Thus, the service time can vary between 1 unit to τ units of time.

In queueing system models, it is typical to assume that each unit of work that an arriving item brings into the system will receive the same mode of service throughout the lifetime of the session, i.e. a stationary service behavior is assumed for each unit. However, such common assumptions cannot be applied in our analysis due to the changes in the channel conditions that influence the changes to the SU's service mode. An SU that is engaged in a data transmission may undergo the switching of its service mode with different rates.

We assume the switching between the different m service modes to occur at random and define $\theta_{i,j}$ as the probability of the service mode switching from mode i to j, such that $\sum_{j=0}^{M} \theta_{i,j} = 1$. The service mode switching probabilities $\theta_{i,j}$ represent the stochastic variations of the channel conditions, as perceived by the SU under service. The authors in [21] have presented one method for evaluating these probabilities, but is limited to the case of only 3 modes, namely {busy, idle, underlay}. In this work, we assume that each SU in the CR network will experience the same variations in the channel conditions, stochastically. Different SUs may experience dissimilar variations in the channel conditions and thus should be modeled using a different set of probabilities $\theta_{i,j}$. For simplicity, the same set of probabilities $\theta_{i,j}$ were assumed for all SUs in the system, with the assumption being applicable to the case of the SUs being stationed within close proximity of each other.

To capture the changes in the service behavior, we adopt the *remaining time approach* (see [26] for further details) and show how to formulate the switching service process that models the remaining time for completing the SU's workload. Let W be define as the discrete random variable that describes the workload brought by an arriving SU at the head of the queue. This workload is equivalent to the *remaining service time* needed by the SU to complete its transmission. The random variable W is assumed to be modeled by a PH distribution with the representation $(\boldsymbol{\alpha}, T)$ of order n_w with $\mathbf{t} = \mathbf{1} - T\mathbf{1}$. While the SU is in service mode m, the elapsed service time is given by H_m. This service mode can change in the middle of serving any one of the SU's remaining data units with different service behaviors.

Let ϕ_m be defined as the probability that the SU at the head of the queue initiates its service in mode m. Thus, the total service completion time S can be modeled by a PH distribution with representation $(\boldsymbol{\psi}, G)$ of order $n_M = n_w \sum_{m=0}^{M} \ell_m$, such that

$$\begin{aligned} \boldsymbol{\psi} &= \boldsymbol{\alpha} \otimes \boldsymbol{\delta} \\ G &= (T \otimes v \otimes \boldsymbol{\delta}) + I \otimes U_0 \\ \mathbf{g} &= \mathbf{1} - G\mathbf{1} = \mathbf{t} \otimes \mathbf{v}, \end{aligned} \tag{3}$$

where I is an identity matrix of appropriate dimensions and $\otimes$ is the Kronecker product operator. The elements for $\boldsymbol{\psi}$ and G are formulated using the PH distributions $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\delta}_m, V_m)$, along with the following,

$$\begin{aligned} \boldsymbol{\delta} &= [\phi_0 \boldsymbol{\delta}_0,\ \phi_1 \boldsymbol{\delta}_1,\ \phi_2 \boldsymbol{\delta}_2,\ \cdots,\ \phi_M \boldsymbol{\delta}_M] \\ v &= [v_0^T,\ v_1^T, \cdots,\ v_M^T]^T \end{aligned} \tag{4}$$

$$U_0 = \begin{bmatrix} V_0\theta_{0,0} & V_{0,1}\theta_{0,1} & \cdots & V_{0,M}\theta_{0,M} \\ V_{1,0}\theta_{1,0} & V_1\theta_{1,1} & \cdots & V_{1,M}\theta_{1,M} \\ \vdots & \vdots & \cdots & \vdots \\ V_{M,0}\theta_{M,0} & V_{M,1}\theta_{M,1} & \cdots & V_M\theta_{M,M} \end{bmatrix}, \tag{5}$$

with $V_{i,j} = (V_i \mathbf{1}) \otimes \boldsymbol{\delta}_j$.

The distribution of the total service time S that an SU at the head of the queue will require to complete to its transmission can be evaluated as follows.

$$Pr\{S = t\} = \boldsymbol{\psi} G^{t-1} \mathbf{g} \qquad t = 1, 2, \cdots \tag{6}$$

The average service time is given as $\mu_S = \boldsymbol{\psi}(I - G)^{-1}\mathbf{1}$.

During the transitions between the service modes, the transceiver will require a finite amount of time to switch between the different modes. While we do not specifically account for these switching times in the process, we assume them to be factored within the various service modes of the proposed switching service process.

IV. THE MODEL DESCRIPTION WITH $K < \infty$

In this section, we present the model for the practical case of a system with a finite buffer of size $K < \infty$. The system was assumed to have a capacity of holding up to $K - 1$ SUs in a buffer that are awaiting service, with a single SU being in service and at the head of the queue at any given time epoch. Let X_t be defined as the number of SUs awaiting service in the system, including the one that is undergoing service, at times $t = 0, 1, 2, \cdots$, and such that $X_t \leq K$. Furthermore, let Y_t be defined as the arrival phase and Z_t the service phase. The behavior of the SUs in a CR network can be analyzed in discrete time and modeled as a single server queueing system with a finite buffer capacity K. The arrivals of the SUs into the system are assumed to be governed by the MAP with matrices D_0 and D_1. The state of this discrete-time Markov Chain (DTMC) can be represented as (X_t, Y_t, Z_t) with the transition probability matrix $\mathbf{P}$ given as follows.

$$\mathbf{P} = \begin{bmatrix} B & C & & & \\ E & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \\ & & & A_2 & \hat{A}_1 \end{bmatrix}. \tag{7}$$

The block matrices in $\mathbf{P}$ are given as follows,

$$\begin{aligned} &B = D_0, \quad C = D_1 \otimes \boldsymbol{\psi}, \\ &E = D_0 \otimes \mathbf{g}, \quad A_2 = D_0 \otimes \mathbf{g}\boldsymbol{\psi}, \\ &A_0 = D_1 \otimes G, \quad A_1 = D_1 \otimes (\mathbf{g}\boldsymbol{\psi}) + D_0 \otimes G, \end{aligned} \tag{8}$$

with $\hat{A}_1 = A_0 + A_1$.

With discrete-time analysis, it is possible to observe multiple events occurring within a single time epoch, i.e. both an arrival and a departure event in the system and within the same time interval. In our analysis, we assume the case of "late arrivals", i.e. an arrival event is assumed to occur after a service completion event within a discrete time epoch.

A. STEADY-STATE ANALYSIS

Let $\boldsymbol{x}$ be defined as the invariant probability vector of the system described by the transition probability matrix $\mathbf{P}$, such that $\boldsymbol{x} = \left[\, x_{0,1,1}, \;\ldots\; x_{i,j,k}, \;\ldots\; x_{K,n,n_M} \right]$, with the probabilities $x_{i,j,k}$ given as

$$x_{i,j,k} = Pr\{X_t = i, Y_t = j, Z_t = k\}|_{t\to\infty}. \tag{9}$$

The invariant probability matrix $\boldsymbol{x}$ can be futher re-written as $\boldsymbol{x} = [\boldsymbol{x}_0, \;\boldsymbol{x}_1, \;\cdots, \boldsymbol{x}_K]$, where $\boldsymbol{x}_i = Pr\{X_t = i\}|_{t\to\infty}$, such that $\boldsymbol{x}_i = \left[\, x_{i,1,1}, \;\ldots\; x_{i,j,k}, \;\ldots\; x_{i,n,n_M} \right]$ for all $0 \le i \le K$.

The solution for $\boldsymbol{x}$ can be obtained using any of the several algorithms for finite homogeneous Quasi-Birth-Death (QBD) systems [27] and can be calculated as the solution to the following equations.

$$\boldsymbol{x} = \boldsymbol{x}\mathbf{P}, \quad \text{and } \boldsymbol{x}1 = 1. \tag{10}$$

The invariant probability matrix is used to compute the common and relevant performance metrics of the system.

B. PERFORMANCE MEASURES

Using the invariant probability matrix $\boldsymbol{x}$ of the system, the standard and relevant metrics can be derived and used to evaluate the system's performance under varying conditions. Some of the common metrics include the number in the system, the queue length, waiting times in the queue and in the system, among several other measures. The results from these metrics can also provide a quantitative measure of the system's workload, on average, along with how well it is capable of meeting the SUs' service demands. The same metrics can also be applied for determining the optimal system operation parameters under certain constraints, e.g. the capacity K.

1) NUMBER IN THE SYSTEM AND IN THE QUEUE

Consider the distribution of the number in the system (i.e. including the SU in service) irrespective of the phases of arrivals and services. Let $\hat{X}$ be that number, and we define $\hat{x}_i = Pr\{\hat{X} = i\}$, such that

$$\hat{x}_i = \boldsymbol{x}_i\mathbf{1}, \quad \text{for } 0 \le i \le K, \tag{11}$$

with $\mathbf{1}$ being defined as an appropriately-dimensioned column vector of ones.

Similarly, let $\check{X}$ be defined as the number in the queue (i.e. excluding the one in service), with $\check{x}_i = Pr\{\check{X} = i\}$. Thus, we can calculate $\check{x}_i$ as follows,

$$\check{x}_i = \boldsymbol{x}_{i+1}\mathbf{1}, \quad \text{for } 0 \le i < K, \text{ with } \check{x}_0 = \boldsymbol{x}_0\mathbf{1} + \boldsymbol{x}_1\mathbf{1}. \tag{12}$$

Using the steady-state probabilities $\hat{x}_i$ and $\check{x}_i$, the metrics for the average number of SUs in the system $\mathbf{N}_L$ and the queue $\mathbf{N}_Q$ can be evaluated as follows.

$$\mathbf{N}_L = \sum_{i=1}^{K} i\hat{x}_i = \sum_{i=1}^{K} i\boldsymbol{x}_i\mathbf{1}. \tag{13}$$

$$\mathbf{N}_Q = \sum_{i=1}^{K-1} i\check{x}_i = \sum_{i=2}^{K} (i-1)\boldsymbol{x}_i\mathbf{1}. \tag{14}$$

Another metric of interest is the loss probability, p_ℓ, that describes the likelihood of an SU being denied any service by the system due to it having reached its maximum capacity (or buffer is full). This metric can be calculated as follows.

$$p_\ell = \lambda^{-1}\boldsymbol{x}_K(D_1 \otimes G)\mathbf{1}. \tag{15}$$

Note that in calculating p_ℓ, an SU is deemed lost from the system if its arrival occurs during the time epoch when the buffer is full and no other SUs have had their services completed.

Finally, the system's throughput, μ_d can be computed as follows.

$$\mu_d = \sum_{i=1}^{K} \boldsymbol{x}_i(\mathbf{1} \otimes I)\mathbf{g}. \tag{16}$$

2) WAITING-TIME DISTRIBUTION

In this section, we show how to derive the waiting-time distribution of the SU in the queue. We focus mainly on studying the waiting time in the queue from which the waiting time in the system can be easily obtained as a convolution sum of the waiting time in the queue and the service time. In order to analyze for the waiting time, we first need to determine the distribution of the number seen in the system by an arriving SU that is admitted into the system.

Let the vector $\mathbf{z}$ be the one corresponding to such a distribution, i.e. $z_{i,j,k}$ is the probability that an arriving SU finds i SUs in the system, with the phase of arrival given as j and k as the phase of ongoing service. Note that $\boldsymbol{x}$ and $\mathbf{z}$ are of the same dimension. Using the results in [27], we can evaluate the sub-vectors $\mathbf{z}_i$ in $\mathbf{z}$ as follows

$$\mathbf{z}_0 = \rho^{-1}[\boldsymbol{x}_0\, D_1 + \boldsymbol{x}_1(D_1 \otimes \mathbf{g})], \tag{17}$$

and for $1 \le i \le K-1$

$$\mathbf{z}_i = \rho^{-1}[\boldsymbol{x}_i(D_1 \otimes G) + \boldsymbol{x}_{i+1}(D_1 \otimes \mathbf{g}\boldsymbol{\psi})], \tag{18}$$

where

$$\rho = \boldsymbol{x}_0\, D_1\mathbf{1} + \boldsymbol{x}_1(D_1 \otimes \mathbf{g})\mathbf{1} + \sum_{i=1}^{K-1} \boldsymbol{x}_i(D_1 \otimes G)\mathbf{1} + \sum_{i=2}^{K} \boldsymbol{x}_i(D_1 \otimes \mathbf{g}\boldsymbol{\psi})\mathbf{1}. \tag{19}$$

Note that $\rho = \mu_d$, i.e. is equivalent to the system's throughput that can also be computed using equation (16).

Next we define a matrix $B_v^{(k)}$ with elements $(B_v^{(k)})_{i,j}$ as the probability that the service time of k SUs lasts v units of time along with a transition from phase i to phase j, (see [27] for details), such that

$$B_v^{(k)} = GB_{v-1}^{(k)} + (\mathbf{g}\boldsymbol{\psi})B_{v-1}^{(k-1)}, \quad k \ge j \ge 1, \tag{20}$$

where

$$B_k^{(k)} = (\mathbf{g}\boldsymbol{\psi})^k, \quad k \ge 1,\; B_v^{(1)} = G^{v-1}(\mathbf{g}\boldsymbol{\psi}).$$

Finally, let $W^{(q)}$ be the waiting time in the queue with a finite capacity of $K-1$ for an SU. We further define $w_i^{(q)} = Pr\{W^{(q)} = i\}$, such that

$$w_0^{(q)} = \mathbf{z}_0\mathbf{1}, \tag{21}$$

$$w_i^{(q)} = \sum_{k=1}^{i^*} \mathbf{z}_k(\mathbf{1}\otimes I)B_i^{(k)}\mathbf{1}, \quad i \geq 1, \tag{22}$$

where $i^* = \min(i, K-1)$.

In addition to finding the average waiting time in the queue, the solutions $w_i^{(q)}$ can further be used to evaluate the tail behavior of the system delay due to queueing, i.e.

$$Pr\{W^{(q)} > k\} = \sum_{i=k+1}^{\infty} w_i^{(q)}. \tag{23}$$

Let $w_i^{(s)} = Pr\{W^{(s)} = i\}$, be defined as the waiting time in the system which includes both the waiting time in the queue and the service time. The probability $w_i^{(s)}$ can be evaluated as follows,

$$w_i^{(s)} = \sum_{k=0}^{i-1}\left(w_k^{(q)} \times s_{i-k}\right), \quad i \geq 1, \tag{24}$$

where $s_i = \boldsymbol{\psi} G^{i-1}\mathbf{g1}$. Note that each user transmission is assumed to require at minimum 1 unit of time.

V. THE MODEL DESCRIPTION WITH $K = \infty$

For systems with very large buffer sizes, the analysis given in the previous section can yield a DTMC with a very large state space that would subsequently require working with a transition probability matrix with large dimensions. Alternatively, we can simplify the analysis in such situations by assuming an infinite buffer size, i.e. $K = \infty$, provided that the likelihood of the loss of SUs in the system (due to a full buffer) is very small and negligible. The same definitions for (X_t, Y_t, Z_t) given in Section IV applies for the case of the system with $K = \infty$, and with the exception that X_t is unbounded. The transition probability matrix $\tilde{\mathbf{P}}$ for this queueing system with infinite capacity is given as follows, as derived previously in [11].

$$\tilde{\mathbf{P}} = \begin{bmatrix} B & C & & & \\ E & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{25}$$

The definition of the block matrices are the same as those given in equations (8), with the absence of the boundary block matrix $\hat{A}_1$.

A. STEADY-STATE ANALYSIS

Let $\tilde{\boldsymbol{x}}$ be defined as the invariant probability vector of the system with the transition probability matrix $\tilde{\mathbf{P}}$, where $\tilde{\boldsymbol{x}} = \left[\tilde{x}_{0,0,0}, \ \ldots \ \tilde{x}_{i,j,k}, \ \ldots\right]$, and can be calculated as the solution to the following equations.

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}\tilde{\mathbf{P}}, \quad \text{and } \tilde{\boldsymbol{x}}\mathbf{1} = 1. \tag{26}$$

The solution $\tilde{\boldsymbol{x}}$ exists for the case where the system is stable such that following condition is satisfied,

$$\boldsymbol{\pi} A_0 \mathbf{1} < \boldsymbol{\pi} A_2 \mathbf{1}, \tag{27}$$

with $\boldsymbol{\pi} = \boldsymbol{\pi} A$, $\boldsymbol{\pi}\mathbf{1} = 1$, and $A = A_0 + A_1 + A_2$. Note that this condition implies that the rate of departures from the system should always exceed the rate of arrivals, if the system is to remain stable.

The invariant probability matrix can also be re-written as $\tilde{\boldsymbol{x}} = [\tilde{\boldsymbol{x}}_0, \ \tilde{\boldsymbol{x}}_1, \ \tilde{\boldsymbol{x}}_2, \ \cdots]$, where $\tilde{\boldsymbol{x}}_i = Pr\{X_t = i\}|_{t\to\infty}$, such that

$$\tilde{\boldsymbol{x}}_{i+1} = \tilde{\boldsymbol{x}}_i R, \quad i = 1, 2, \cdots, \tag{28}$$

where R is the minimum solution to the matrix quadratic equation

$$R = A_0 + RA_1 + R^2 A_2. \tag{29}$$

Refer to [27] for the various efficient methods available to obtain the solutions for the R matrix and steady-state probabilities $\tilde{\boldsymbol{x}}_i$. These results are required for evaluating the relevant performance metrics given in the subsequent sections.

B. PERFORMANCE MEASURES

Similar to the previous section, we next show how to evaluate some of the standard and common metrics used to analyze the performance of the queueing system, such as the number in the system, the queue length, and waiting times in the queue. The results obtained from these metrics could also serve as an approximation to those obtained from the model of the system with a finite buffer capacity K (see Section IV), where K is very large and the loss probability $p_\ell \approx 0$.

1) NUMBER IN THE SYSTEM AND IN THE QUEUE

The steady-state probabilities $\tilde{\boldsymbol{x}}_i$ can be used to calculate the mean number of SUs in the system $\tilde{\mathbf{N}}_L$ and the queue $\tilde{\mathbf{N}}_Q$, as follows.

$$\tilde{\mathbf{N}}_L = \sum_{i=1}^{\infty} i\tilde{\boldsymbol{x}}_i\mathbf{1} = \tilde{\boldsymbol{x}}_1(I-R)^{-2}\mathbf{1}. \tag{30}$$

$$\tilde{\mathbf{N}}_Q = \sum_{i=2}^{\infty}(i-1)\tilde{\boldsymbol{x}}_i\mathbf{1} = \tilde{\boldsymbol{x}}_1 R(I-R)^{-2}\mathbf{1}. \tag{31}$$

Note that the results for $\tilde{\mathbf{N}}_L$ and $\tilde{\mathbf{N}}_Q$ would always be less than ∞ for a stable system that satisfies the conditions given in equation (27), despite the system being of infinite capacity.

2) WAITING-TIME DISTRIBUTION

For computing the waiting-time distribution of the SUs in the system with the infinite buffer capacity, we apply the same steps followed in the previous section for this derivation. Let the vector $\tilde{\mathbf{z}}_i$ correspond to the probability that an arriving SU finds i SUs ahead of it in the system. Using again the results in [27], we evaluate the vectors $\tilde{\mathbf{z}}_i$ as follows.

$$\tilde{\mathbf{z}}_0 = \lambda^{-1}[\tilde{\boldsymbol{x}}_0 D_1 + \tilde{\boldsymbol{x}}_1(D_1 \otimes \mathbf{g})], \tag{32}$$

and

$$\tilde{\mathbf{z}}_i = \lambda^{-1}[\tilde{\boldsymbol{x}}_i(D_1 \otimes G) + \tilde{\boldsymbol{x}}_{i+1}(D_1 \otimes \mathbf{g}\boldsymbol{\psi})], \quad \text{for } i \geq 1. \tag{33}$$

We can also define the following

$$\tilde{\mathbf{z}}_i = \tilde{\boldsymbol{x}}_1 R^{i-1} F, \quad \text{for } i \geq 1, \tag{34}$$

where

$$F = \lambda^{-1}[(D_1 \otimes G) + R(D_1 \otimes \mathbf{g}\boldsymbol{\psi})], \tag{35}$$

and the result for λ can be computed using equation (1).

Let $\tilde{W}^{(q)}$ be the waiting time in the queue for an SU in the sytem with infinite buffer capacity, and $\tilde{w}_i^{(q)} = Pr\{\tilde{W}^{(q)} = i\}$, such that

$$\tilde{w}_0^{(q)} = \tilde{\mathbf{z}}_0\mathbf{1}, \tag{36}$$

$$\tilde{w}_i^{(q)} = \sum_{k=1}^{i} \tilde{\mathbf{z}}_k(\mathbf{1} \otimes I)B_i^{(k)}\mathbf{1}, \quad \text{for } i \geq 1, \tag{37}$$

where the matrices $B_v^{(k)}$ can be computed using equation (20).

Let $\tilde{w}_i^{(s)} = Pr\{\tilde{W}^{(s)} = i\}$, be defined as the waiting time in the system which includes both the waiting time in the queue and the service time. The probability $\tilde{w}_i^{(s)}$ can be evaluated as follows,

$$\tilde{w}_i^{(s)} = \sum_{k=0}^{i-1} \left(\tilde{w}_k^{(q)} \times s_{i-k}\right), \quad i \geq 1. \tag{38}$$

VI. NUMERICAL EXAMPLES

To demonstrate the application of the queueing models developed in the previous sections, various numerical examples are given in this section to illustrate how the performance of such a CR networking system would behave while varying certain system conditions. We consider both the cases of finite and infinite buffers in the examples to follow. A discrete-event simulation of the same system was also developed in Matlab for the purpose of verifying the analytical results obtained from the proposed model. The analytical results given in this section are shown alongside the numerical results obtained from the simulation of the system with the same parameters. Overall, the simulation results have successfully verified the analytical results.

Throughout this section, the numerical analysis is conducted for the case of the CR system with 4 different modes of service, i.e. $M = 3$. Furthermore, the following values were assigned to the parameters that model the arrival and service behaviors in each of the different modes. These values are fictitious, but they have been chosen to approximately resemble the expected behaviors of such networks and to help demonstrate the differences in performance.

$$D_0 = \begin{bmatrix} 0.3 & (0.45-\gamma) \\ 0.25 & 0.4 \end{bmatrix}, \quad D_1 = \begin{bmatrix} (0.15+\gamma) & 0.1 \\ 0.2 & 0.15 \end{bmatrix},$$

$$\boldsymbol{\alpha} = [0.2 \quad 0.8], \quad T = \begin{bmatrix} 0 & 0.2 \\ 1 & 0 \end{bmatrix},$$

$$\boldsymbol{\delta}_0 = \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \boldsymbol{\delta}_3 = [1\ 0], \quad V_0 = \begin{bmatrix} 0 & 0.15 \\ 0.1 & 0 \end{bmatrix},$$

$$V_1 = \begin{bmatrix} 0 & 0.2 \\ 0.1 & 0 \end{bmatrix}, \quad V_2 = \begin{bmatrix} 0 & 0.25 \\ 0.1 & 0 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$\{\theta_{i,j}\} = \begin{bmatrix} 0.1 & 0.5 & 0.3 & 0.1 \\ 0.6 & 0.1 & 0.1 & 0.2 \\ 0.3 & 0.4 & 0.1 & 0.2 \\ 0.2 & 0.2 & 0.5 & 0.1 \end{bmatrix},$$

$$[\phi_0, \phi_1, \phi_2, \phi_3] = [(0.45-\beta), 0.35, (0.15+\beta), 0.05].$$

Among the 4 different modes of service, an SU operating in mode $m = 0$ will experience the highest service rate due to it also being in the overlay mode of access. In mode $m = 3$, the SU's service is suspended due to the unavailability of a channel and the unsuitable channel conditions. The SU would be operating in the underlay mode of access when it is either in mode $m = 1$ or $m = 2$, with the the average service rate being higher in mode $m = 1$ due to a more favorable channel condition.

The analysis in the following set of examples were conducted by either varying the arrivals of SUs into the system, or varying the service mode initiation probabilities ϕ_m. In the cases where the arrivals are varied, we examine the performance of the CR system by varying the parameter γ in the sub-stochastic matrices D_0 and D_1 of the MAP, and for $0 \leq \gamma \leq 0.4$. This is equivalent to varying the arrival rates into the system within the range of $0.305 \leq \lambda \leq 0.575$. As for the cases of varying ϕ_m, we analyze the system's performance by varying the parameter β within the range of $0 \leq \beta \leq 0.35$. The analysis of the system's behavior with variations in β help to examine the changes in the SU's performance as the likelihood of a transmission being initiated with either the lowest or highest rate is varied. With $\beta = 0$, the average service time $\mu_S = 3.05$ units, whereas $\mu_S = 3.18$ with $\beta = 0.35$. Hence, analyzing the system with $0 \leq \beta \leq 0.35$ is equivalent to examining its performance with varying average service times of $3.05 \leq \mu_S \leq 3.18$.

A. THE FINITE BUFFER CASE OF $K < \infty$

For the finite buffer case, we assume a maximum of 9 SUs can be queued and awaiting their service to be initiated, hence, $K = 10$. In the first set of numerical examples, we examine the performance of the system while only varying $0 \leq \beta \leq 0.35$ for the service mode initiation probabilities, and with $\gamma = 0$. Figs. 2 and 3 show the variations of the average number in the system and the queue with increasing probabilities of an SU initiating its service in mode 2 compared to mode 0. The expected rise in the number of SUs awaiting service in the system is due to the lower transmission rate of the service mode 2 compared to the rate in mode 0. The increase in the number of SUs in the system will also increase the tendency of the buffer reaching its capacity, thus explaining the rise in the loss probabilities shown in Fig. 3. The results in Figs. 4 and 5 show the waiting-time distribution as well as the complementary cumulative distribution (or tail distribution) of the waiting times, respectively, and for the cases when $\beta = 0.05$ and $\beta = 0.35$. The results for the

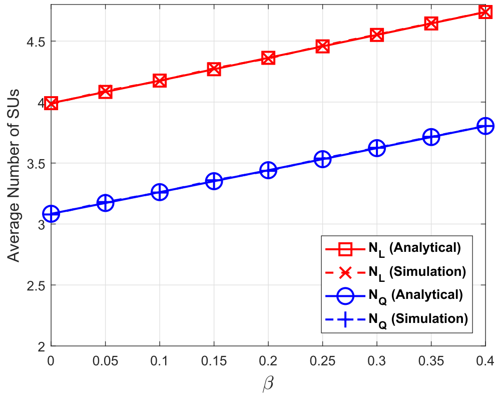

FIGURE 2. Average number of SUs in the system and queue, with varying β and $K = 10$.

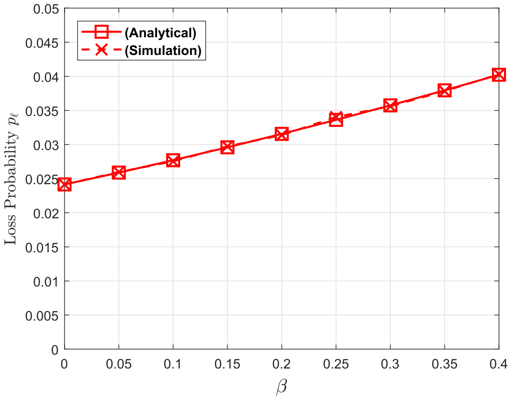

FIGURE 3. The loss probability p_ℓ in the system, with varying β and $K = 10$.

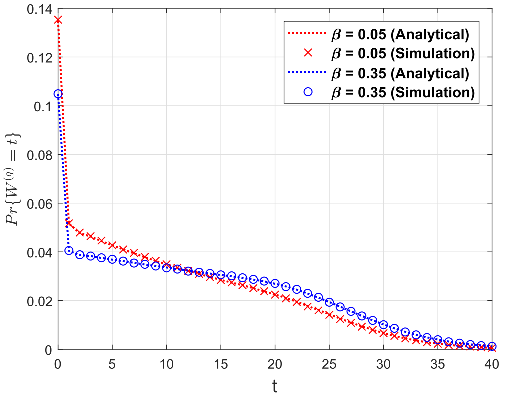

FIGURE 4. Waiting time distribution $Pr\{W^{(q)} = t\}$, with $K = 10$.

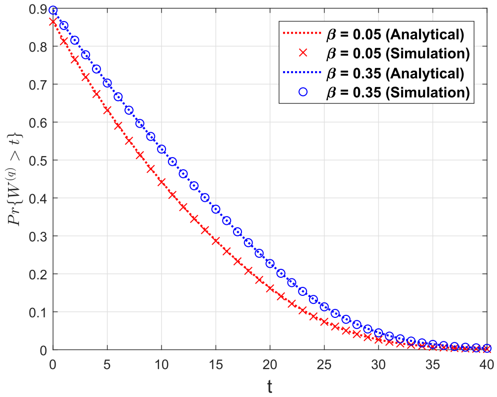

FIGURE 5. Tail behavior of the waiting time distribution $Pr\{W^{(q)} > t\}$ with $K = 10$.

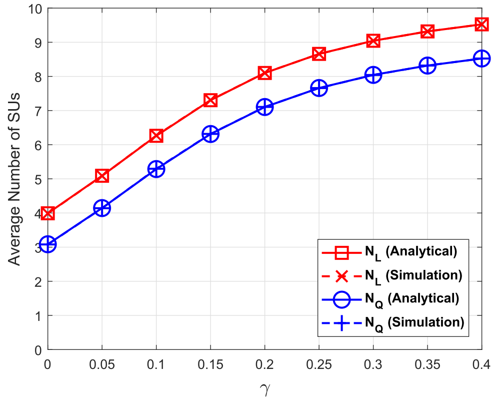

FIGURE 6. Average number of SUs in the system and queue, with varying γ and $K = 10$.

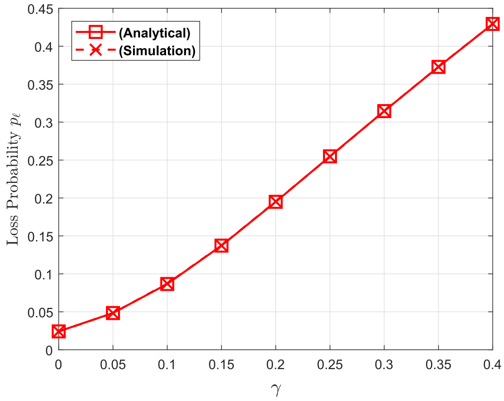

FIGURE 7. The loss probability p_ℓ in the system, with varying γ and $K = 10$.

waiting times exhibit similar trends to those for the number in the system and the queue with increasing β. The waiting-time distribution for $\beta = 0.35$ has a heavier tail implying that SUs are more likely to wait in the queue for longer periods of time prior to receiving service. This is due to the SUs having a higher service time, on average, with $\beta = 0.35$. The results of the average waiting times would also have illustrated the same trends, but the distributions in Figs. 4 and 5 further show the disparity in the variations at different time instances. Such details can be quite crucial for systems with delay sensitive applications. Fig. 4 further shows that SUs in a system with lower β are more likely to have their transmissions delayed by $t = 12$ units of time or less, when compared with SUs operating in a system with high β. This is due to the higher likelihood of SUs initiating their service transmissions in mode 0. Conversely, SUs in systems with higher β are more likely to exhibit a delay of $t > 12$ units of time when compared with SUs operating in systems with lower β.

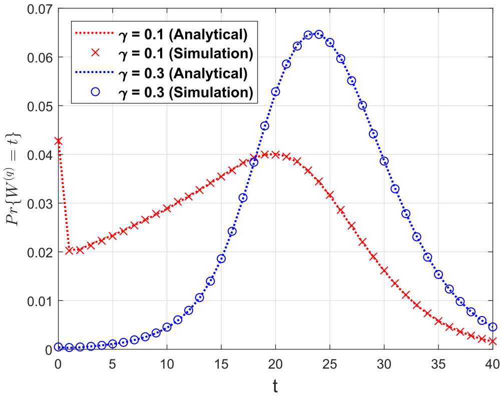

FIGURE 8. Waiting time distribution $Pr\{W^{(q)}=t\}$, with $K=10$ and varying arrival rates.

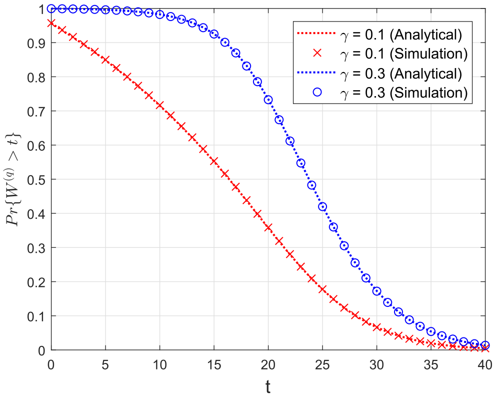

FIGURE 9. Tail behavior of the waiting time distribution $Pr\{W^{(q)}>t\}$ with $K=10$ and varying arrival rates.

In the next set of examples, the same system is analyzed instead with increasing arrival rates. This analysis is accomplished by varying the parameter γ in the sub-stochastic matrices for the MAP, while maintaining a constant service mode initiation probabilities with $\beta=0$ (hence, constant average service rates). Figs. 6 to 9 show the behaviors of the various performance measures with increasing arrival rates of $0\leq\gamma\leq0.4$ (which is equivalent to varying $0.305\leq\lambda\leq0.575$). The results in Fig. 6 illustrate the expected behavior of observing an increase in the average number of SUs in the system with higher arrival rates, which impacts the loss probabilities in the system as shown in Fig. 7. The waiting-time distribution and its tail distribution are shown in Figs. 8 and 9, respectively, for $\gamma=0.1$ and $\gamma=0.3$. These figures show how the SUs are much more likely to wait for longer periods of time in the queue when the arrival rates into the system is high, as given by the heavier tail in the distribution for $\gamma=0.3$. Notice how the difference in the tail distributions is quite significant in Fig. 9, and how these differences change with varying time instances. These results further illustrate that an SU will almost certainly have to wait for at least 10 units of time before its data transmission can be initiated in a system with $\gamma=0.3$. The same is not true for the system with the lower arrival rate of $\gamma=0.1$. This level of detail cannot be observed using the average waiting times alone and without their distributions.

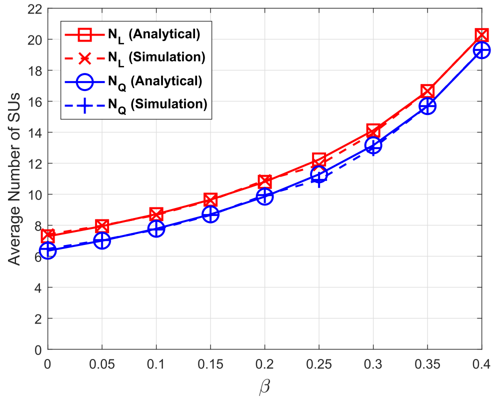

FIGURE 10. Average number of SUs in the system and Queue, with varying β.

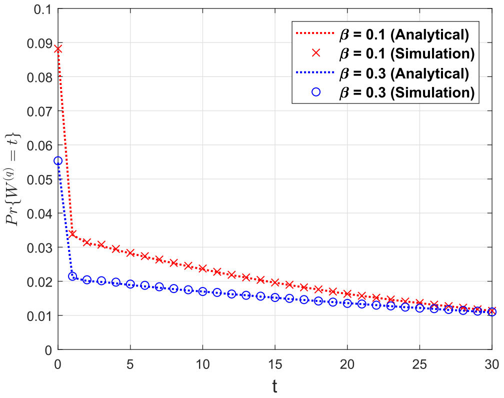

FIGURE 11. Waiting time distribution $Pr\{W^{(q)}=t\}$.

B. THE INFINITE BUFFER CASE OF $K=\infty$

Using the same traffic parameters given at the start of this section, we show the results from analyzing the system with varying β, and for the case where the buffer capacity is infinite. In Fig. 10, the results show the expected rise in the average number of SUs in the system and queue (i.e. $\mathbf{N}_L$ and $\mathbf{N}_Q$, respectively) as the likelihood of the SU's service initiation with the lowest transmission rate is increased. This is due to the longer time needed for the SU to complete its transmission, on average, when initiated in mode $m=2$. The longer service times would further result in an increase in the number of backlogged SUs in the system, as illustrated in the figure with increasing β. Conversely, a lower β tends to reduce the number of backlogged SUs in the system. This is due to the lower average service completion times that resulted from the higher likelihood of the SU's service being initiated in mode $m=0$ with the largest transmission rate.

Figs. 11 and 12 show the waiting time distribution of the SUs in the systems and along with its tail distribution, respectively, for the cases when $\beta=0.1$ and $\beta=0.3$. It is evident from these results that SUs are more likely to have lower waiting times before their service is initiated for the case when $\beta=0.1$, as expected. This is due to the higher likelihood of the SU's service being initiated at mode $m=0$ with the highest transmission rate. The steeper decline in the tail probabilities for the case of $\beta=0.1$ further emphasizes

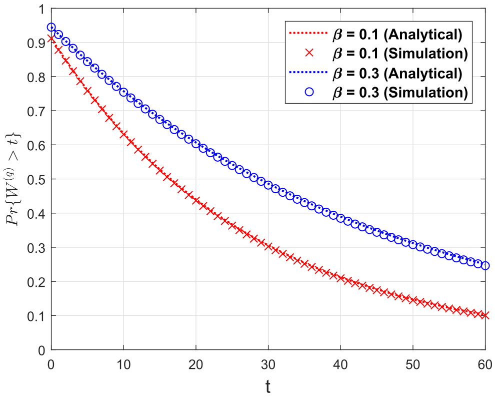

FIGURE 12. Tail behavior of the waiting time distribution $Pr\{W^{(q)} > t\}$.

the improved performance as compared to the case of $\beta = 0.3$.

VII. SUMMARY & FUTURE WORK

The performance of a CR network with multi-modal service switching was modeled as a discrete time single server queueing system and presented in this paper. The model captures the collective behavior of SUs in the presence of a single PU. Unlike the previous work accomplished by others, our proposed model extends beyond the traditional three access modes (i.e. overlay, underlay, and busy) and includes the formulation of a multi-modal switching service behavior that considers the dynamic changes in the channel conditions. In our work, the channel access in the system is assumed to be administered by a centralized node, such as a base station or cluster head. The arrival and service processes were modeled using general distributions that allow for a more accurate analysis. The proposed model considers both the cases of the buffer with a finite and infinite capacity. We further presented a method for computing the waiting-time distribution of the secondary users which is essential for understanding the sensitivity of the secondary user's performance due to the queueing delays, especially for real-time applications.

The dynamic channel conditions for each of the different SUs were assumed to be homogeneous. In other words, each of the SUs were assumed to be under the influence of the same service mode switching probabilities $\theta_{i,j}$. This assumption was necessary for simplifying the formulation of the model and may serve as a valid approximation for a cluster of SUs within close proximity. We intend to extend the model in our future work to consider assigning a distinct $\theta_{i,j}$ for a diverse set, or ''classes'', of SUs. We further intend on developing a method for formulating the service time distributions along with the service mode switching probabilities $\theta_{i,j}$ that considers the dynamics of the channel characteristics in the cognitive radio networking systems. The model presented in this paper considers the SU's behavior from the perspective of a single licensed channel, even though the nodes in the CR networks can have access to a set of these licensed channels. The analysis of the node's behavior with access to multiple licensed channels, along with the channel selection process, will also be considered in our future work.

REFERENCES

[1] (Feb. 2017). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] G. P. Joshi, S. Y. Nam, and S. W. Kim, ''Cognitive radio wireless sensor networks: Applications, challenges and research trends,'' *Sensors*, vol. 13, pp. 11196–11228, Aug. 2013.

[3] S. C. Mukhopadhyay and N. K. Suryadevara, ''Internet of Things: Challenges and opportunities,'' in *Internet Things: Challenges Opportunities* (Smart Sensors, Measurement and Instrumentation), vol. 9. Cham, Switzerland: Springer, 2014, pp. 1–17. doi: 10.1007/978-3-319-04223-7_1.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, ''Internet of Things for smart cities,'' *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[5] *Spectrum Policy Task Force Report*, document ET Docket 02-135, FCC, Nov. 2002.

[6] K. Katzis and H. Ahmadi, ''Challenges implementing Internet of Things (IoT) using cognitive radio capabilities in 5G mobile networks,'' in *Internet of Things (IoT) in 5G Mobile Technologies*. Cham, Switzerland: Springer, 2016, pp. 55–76.

[7] R. Tandra, S. M. Mishra, and A. Sahai, ''What is a spectrum hole and what does it take to recognize one?'' *Proc. IEEE*, vol. 97, no. 5, pp. 824–848, May 2009.

[8] A. Goldsmith, S. A. Jafar, I. Marić, and S. Srinivasa, ''Breaking spectrum gridlock with cognitive radios: An information theoretic perspective,'' *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.

[9] X. Kang, Y.-C. Liang, H. K. Garg, and L. Zhang, ''Sensing-based spectrum sharing in cognitive radio networks,'' *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4649–4654, Oct. 2009.

[10] A. Valehi and A. Razi, ''Maximizing energy efficiency of cognitive wireless sensor networks with constrained age of information,'' *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 643–654, Dec. 2017.

[11] A. S. Alfa, H. A. Ghazaleh, and B. T. Maharaj, ''A discrete time queueing model of cognitive radio networks with multi-modal overlay/underlay switching service levels,'' in *Proc. IEEE Int. Wireless Commun. Mobile Comput. Conf.*, Jun. 2018, pp. 1030–1035.

[12] S. M. K. Badalge, N. Rajatheva, and M. Latva-Aho, ''Overlay/underlay spectrum sharing for multi-operator environment in cognitive radio networks,'' in *Proc. IEEE 73rd Veh. Technol. Conf.*, May 2011, pp. 1–5.

[13] J. Zou, H. Xiong, D. Wang, and C. W. Chen, ''Optimal power allocation for hybrid overlay/underlay spectrum sharing in multiband cognitive radio networks,'' *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1827–1837, May 2013.

[14] M. Usman and I. Koo, ''Access strategy for hybrid underlay-overlay cognitive radios with energy harvesting,'' *IEEE Sensors J.*, vol. 14, no. 9, pp. 3164–3173, Sep. 2014.

[15] H. Li and Z. Han, ''Socially optimal queuing control in cognitive radio networks subject to service interruptions: To queue or not to queue?'' *IEEE Trans. Wireless Commun.*, vol. 10, no. 5, pp. 1656–1666, May 2011.

[16] A. K. Farraj, S. L. Miller, and K. A. Qaraqe, ''Queue performance measures for cognitive radios in spectrum sharing systems,'' in *Proc. IEEE Int. Workshop Recent Adv. Cogn. Commun. Netw.*, Dec. 2011, pp. 997–1001.

[17] T. M. C. Chu, H. Phan, and H.-J. Zepernick, ''On the performance of underlay cognitive radio networks using M/G/1/K queueing model,'' *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 876–879, May 2013.

[18] C. T. Do, N. H. Tran, and C. S. Hong, ''Optimal queueing control in hybrid overlay/underlay spectrum access in cognitive radio networks,'' in *Proc. IEEE 75th Veh. Technol. Conf.*, May 2012, pp. 1–5.

[19] H. M. Tsimba, B. T. Maharaj, and A. S. Alfa, ''Increased spectrum utilisation in a cognitive radio network: An M/M/1-PS queue approach,'' in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[20] S. Tang and B. L. Mark, ''Analysis of opportunistic spectrum sharing with Markovian arrivals and phase-type service,'' *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3142–3150, Jun. 2009.

[21] S. Senthura, A. Anpalagan, and O. Das, ''Throughput analysis of opportunistic access strategies in hybrid underlay—Overlay cognitive radio networks,'' *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2024–2035, Jun. 2012.

[22] S. Wang, J. Zhang, and L. Tong, "Delay analysis for cognitive radio networks with random access: A fluid queue view," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[23] Z. Liang, S. Feng, D. Zhao, and X. S. Shen, "Delay performance analysis for supporting real-time traffic in a cognitive radio sensor network," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 325–335, Jan. 2011.

[24] I. Alabdulmohsin, A. Hyadi, L. Afify, and B. Shihada, "End-to-end delay analysis in wireless sensor networks with service vacation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2014, pp. 2799–2804.

[25] M. Usman, H.-C. Yang, and M.-S. Alouini, "Extended delivery time analysis for cognitive packet transmission with application to secondary queuing analysis," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5300–5312, Oct. 2015.

[26] A. S. Alfa, "Markov chain representations of discrete distributions applied to queueing models," *Comput. Oper. Res.*, vol. 31, pp. 2365–2385, Dec. 2004.

[27] A. S. Alfa, *Applied Discrete-Time Queues*, 2nd ed. New York, NY, USA: Springer-Verlag, 2016. doi: 10.1007/978-1-4939-3420-1.

ATTAHIRU SULE ALFA received the B.Eng. degree from Ahmadu Bello University, Zaria, Nigeria, the M.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, and the Ph.D. degree from the University of New South Wales, Sydney, Australia. He is Professor Emeritus with the Department of Electrical and Computer Engineering, University of Manitoba, and also a UP/CSIR Co-Hosted SARChI Chair Professor with the Department of Electrical, Electronic and Computer Engineering, University of Pretoria. He has authored two books, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System* (Springer in 2010) and *Applied Discrete-Time Queue* (Springer, in 2015, a second edition of the first book). His most recent research focus covers wireless sensor networks, cognitive radio networks, network restoration tools for wireless sensor networks, and the role of 5G on the IoT, with specific interest in the mathematical modeling of those systems. His general research covers, but not limited to, the following areas: 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, and smart cities. Some of his previous works include developing efficient decoding algorithms for LDPC codes, channel modeling, traffic estimation for the Internet, and cross layer analysis. He also works in the application of queuing theory to other areas, such as transportation systems, manufacturing systems, and healthcare systems.

HAITHAM ABU GHAZALEH received the B.Eng. degree in electronics and electrical engineering from the University of Manchester, U.K., in 1999, the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Manitoba, in 2006 and 2010, respectively. He further worked as a Graduate Student Researcher with TRLabs Winnipeg and was involved in the research areas of wireless networking for eHealth and telemedicine applications. From 2009 to 2015, he worked in the Network Planning and Engineering department with MTS Allstream. He is currently an Assistant Professor of electrical engineering with Tarleton State University. His research interests include wireless and network teletraffic modeling and performance analysis, queueing systems, adaptive network resource management, mobility in future generation wireless networks, cognitive radio systems, wireless sensor networks, smart cities, and the Internet of Things. He is also a member of the Association of Professional Engineers and Geoscientists of the Province of Manitoba (APEGM).

BODHASWAR T. (SUNIL) MAHARAJ received the Ph.D. degree in electronic engineering, specializing in wireless communications, from the University of Pretoria, where he is currently a Full Professor and holds the research position of Sentech Chair with Broadband Wireless Multimedia Communications, in 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.