Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 7385627, 12 pages
https://doi.org/10.1155/2017/7385627

Research Article

Resource Allocation in Heterogeneous Buffered Cognitive Radio Networks

B. S. Awoyemi,[1] B. T. Maharaj,[1] and A. S. Alfa[1,2]

[1] *Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria 0002, South Africa*
[2] *Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 5V6*

Correspondence should be addressed to B. S. Awoyemi; awoyemibabatunde@gmail.com

Received 9 May 2017; Revised 9 June 2017; Accepted 20 June 2017; Published 25 July 2017

Academic Editor: Gonzalo Vazquez-Vilar

Resources available for operation in cognitive radio networks (CRN) are generally limited, making it imperative for efficient resource allocation (RA) models to be designed for them. However, in most RA designs, a significant limiting factor to the RA's productivity has hitherto been mostly ignored, the fact that different users or user categories do have different delay tolerance profiles. To address this, in this paper, an appropriate RA model for heterogeneous CRN with delay considerations is developed and analysed. In the model, the demands of users are first categorised and then, based on the distances of users from the controlling secondary user base station and with the assumption that the users are mobile, the user demands are placed in different queues having different service capacities and the resulting network is analysed using queueing theory. Furthermore, to achieve optimality in the RA process, an important concept is introduced whereby some demands from one queue are moved to another queue where they have a better chance of enhanced service, thereby giving rise to the possibility of an improvement in the overall performance of the network. The performance results obtained from the analysis, particularly the blocking probability and network throughput, show that the queueing model incorporated into the RA process can help in achieving optimality for the heterogeneous CRN with buffered data.

1. Introduction

In recent times, there has been a deservedly growing interest in cognitive radio networks (CRN) as a possible driver for next generation (xG) wireless communications. This special interest in CRN hinges on its promise of much higher resourcefulness, particularly considering the spectrum usage. In its design, CRN enables an allotted spectrum space to be used by two different networks, a primary and a secondary one, under certain preconditions agreed upon by both networks [1]. With this kind of arrangement, much better utilisation of the rather scarce spectrum becomes inevitable. This promise has triggered considerable research effort to develop and describe the CRN, as well as to address possible challenges to its introduction and implementation.

In CRN, primary users (PUs) generally take priority in the usage of the resources, especially spectrum, because they are the original or licensed owners of it. The secondary users (SUs) in the network must devise ways to achieve and maintain acceptable quality of service (QoS) despite the stringent conditions under which they have to operate [2]. In heterogeneous CRN especially, the demands of the SUs are usually different from one SU to another or from one category of users to another, and the CRN should be capable of meeting the different demands efficiently and timeously [3]. To make this possible, resource allocation (RA) models that capture the essential peculiarities and dynamics of the heterogeneous users in CRN and that can optimally assign the available resources fairly and favourably are required [4].

In developing RA models for CRN, an important and realistic criterion for categorising demands of the heterogeneous users is their level of delay tolerance that is permissible for an acceptable QoS. Depending on the kind of service being provided, different users may have differing delay tolerance characteristics. Usually, because resources for SUs' transmission are limited or sometimes even temporarily unavailable (e.g., because of PUs' transmission), SUs, depending on the kind of service intended to be provided, may keep their data for transmission in a buffer (queue) and wait (usually for an acceptable time duration) for the requisite resources to be available for their transmission to be completed. Those delay instances and/or durations for SUs, and the properties of the queue developed as a result, are critical for the CRN's QoS provisioning and applying queueing models in analysing the queue characteristics can help in improving their performance significantly. That is the main interest of this paper. The paper therefore develops and analyses an appropriate queueing model for heterogeneous CRN with users having different delay priorities or delay profiles. The analyses of the model show that, by investigating and exploiting the delay characteristics of users in CRN using queueing theory, a significant improvement in the optimal allocation of resources for the heterogeneous buffered CRN can be realised.

2. Related Literature

Developing appropriate RA models for CRN has been a recent research focus and significant progress has already been made in this regard. In earlier works of the authors, a comprehensive survey of various approaches that have been investigated and employed in achieving optimal (or near-optimal) solutions for RA in CRN has been carried out, with their pros and cons highlighted and discussed [5]. Optimal RA solutions for heterogeneous CRN were investigated in [6, 7] and the works were further strengthened in [8] to include QoS provisioning. The concept of cooperative diversity was introduced in [9] to address the problem of interference mitigation, thereby achieving even better results.

Similarly, other authors have also been (and are currently being) involved in investigating RA solutions for CRN. The authors in [10–12] analysed RA problems in heterogeneous multiple-input multiple output CRN. In their works, the dual optimisation problems developed focused on optimising the transmit power and transmission time allocation of the SUs. To achieve their objectives, firstly, power allocation to each SU was optimised (on assumption of a constant transmission time), and secondly, optimal scheduling of the SUs was realised. In [13, 14], the authors studied RA in CRN, first with cooperative relays and then with imperfect spectrum sensing. The developed problems were nonconvex and nondeterministic polynomial-time-hard (NP-hard). By separating the problems into two, subchannel allocation and power allocation, a convex programming conversion was achieved and suboptimal solutions realised. All of these works and many more similar works all show that RA in CRN is an active research area and some interesting results have been achieved already. However, in the above-mentioned works and in most other related works on RA in CRN, no consideration has been given to the delay requirements of the SUs.

The above-mentioned works have all addressed RA problems in CRN but without any consideration of queueing capabilities in achieving better (and/or optimal) solutions. Works on RA in the CRN that have incorporated some kind of time delay or data queueing in their RA problem formulations are indeed very few. The available ones, as obtained by the authors in the process of this research work, are briefly reviewed. In [15], the authors, in developing their RA model, made provision for users that require real-time communication and gave them priority over non-real-time users. The real-time users were given priority in that they were admitted first into the network and given sufficient resources to transmit their data. Thereafter, an optimal number of non-real-time users were made to share the remaining resources of the network. Thus, the delay profiles of the users were used as a means of controlling the number of users admitted to the network. Authors in [16] classified the SUs into delay-sensitive and delay-tolerant SUs, with the delay-sensitive SUs having the higher priority in the allocation of resources, as their delay requirements must always be met first. In the analysis, the delay requirements of the delay-sensitive SUs were simply transformed into a constant rate requirement using very elementary ideas of queueing theory.

A study on band allocation in CRN when both PUs and SUs have data queues was carried out in [17]. In order to satisfy the required QoS for the SUs, the authors proposed that each SU be probabilistically assigned to a PU band and showed that such arrangement gives a better performance than either random or fixed allocations for the SUs. In [18], the authors developed a model that separates the users by the amount of data backlog they have in their queues waiting to be transmitted. The allocation of resources was carried out based on the size of the backlog; users with small backlogs were given just sufficient resources to transmit their data, while users with large backlogs shared the remaining resources among themselves in a manner that was fair and efficient. In the above-mentioned works, the significance of time delays and the effects on the overall performance of the RA problems had not been studied. Moreover, improving overall network capacity by developing appropriate queueing models to address the RA problems was never carried out, making this work uniquely different.

In this work therefore, the more realistic possibility of users having buffered data (queues) is incorporated into the RA problems of CRN and its effects are investigated. Particularly, the RA model developed in this paper addresses the very likely circumstance where heterogeneous SUs have data in their buffer waiting to be transmitted. With the SUs having different delay tolerance profiles, the model seeks to find a unique method of transmitting the SUs' data in a manner in which QoS is guaranteed for each SU, while the overall productivity of the network is enhanced. By categorising the SUs and developing efficient queueing and RA mechanisms to best satisfy their requirements, a much greater capacity can be realised for the network. The contributions in this work are summarised as follows:

(i) Developing and analysing a queueing model that captures delay characteristics of various categories of users in buffered heterogeneous CRN.

(ii) Investigating optimality in the RA process for the heterogeneous CRN by studying the impact of varying the values of user demands into different queues and determining their effect on the overall performance of

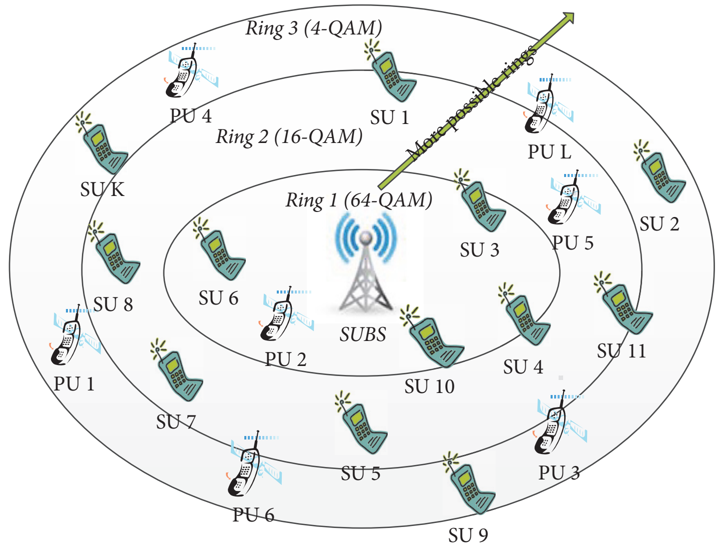

Figure 1: Representation of the underlay heterogeneous CRN. Users are assumed to be placed in virtual rings of different distance ranges from the SUBS.

the network. The variation in arrivals is achieved by moving a factor θ of demands from users in a farther queue to a nearer one. θ itself is such that it can be changed continuously within a certain range, and an optimal value for θ that maximises the productivity of the network is realised.

3. System Model

The system model is shown in Figure 1. The model is a development on the work presented in [19]. The referenced work had considered an overlay network in which the SUs query a database to ascertain the PUs' transmission and had designed the SUs' connections according to the periodicity of the PUs' traffic patterns. However, in this model, a centralised, underlay heterogeneous CRN is developed. The SUs are made to transmit on the entire PUs' bandwidth within an acceptable interference power limit. This makes it possible to concentrate on analysing the SU network of the CRN separately and intently, so as to achieve the utmost for the SUs, and this without causing significant harm to the PUs.

The SUs are classified based on their delay profiles as either delay-sensitive (DS) or delay-tolerant (DT) users, the classes differentiated by the duration of delay time for acceptable service. The SUs are assumed to be both mobile and capable of performing adaptive modulation and coding (AMC) to dynamically change their modulation and coding schemes. Moreover, the SUs, depending on their distance to the secondary user base station (SUBS), are placed in virtual rings. The nearest ring to the SUBS operates with the highest AMC scheme, while the farthest ring operates with the lowest AMC scheme. Data transmission requests of users within a ring are placed in a queue of that ring and service (transmission of data) is carried out using the available subchannels. The queues therefore act as a buffer, should there be a possible delay in immediate transmission due to insufficiency in resource availability. There are N subchannels, which automatically correspond to the number of parallel servers in each ring (or queue). Queues are finite with a maximum length Y. As a result of the mobility of SUs, arrival rates into queues can be adjusted so that the maximum productivity of the network can be realised. To accomplish this, a fraction of the demands of users, particularly the DT demands, which have a high delay profile, is moved from a farther ring (queue) to a closer ring (queue), so that the demands can possibly be transmitted at a higher rate. The intent is that, by such an arrangement, a likely reduction in both the time and energy consumed in transmitting the data can be achieved, making it possible for significant improvement in the capacity of the SUs of the CRN.

3.1. Queueing Model. The queueing model is set up to determine the capacity of the system. The queueing analysis shows the significance of the fraction of demands that is moved between queues. To make the model manageable and easier to analyse, only two concentric rings are considered, meaning that there are two parallel queues, each served by multiple servers (subchannels). The analysis can however be extended to three or more rings. In the initial analysis, the closest ring to the SUBS is assigned to transmit at a modulation scheme of 64-QAM (6 bits per symbol) and the farthest ring is assigned to transmit with a modulation scheme of 4-QAM (2 bits per symbol). Arrivals into each of the queues follow a Poisson distribution with arrival rates λ_1 for queue 1 and λ_2 for queue 2. Service is exponential with rates μ_1 for queue 1 and μ_2 for queue 2. μ_1 and μ_2 correspond to the data rate of the AMC scheme operated in each ring, meaning 6 bits per symbol and 2 bits per symbol for queue 1 and queue 2, respectively. Service per unit time in queue 1 is therefore significantly faster than in queue 2, since it operates at a higher service rate. Some of the arrivals into each queue are DT demands and users can move from one ring to another. The essence is to determine whether the CRN's productivity can be improved by moving the DT demands of the farther ring to the ring nearer to the SUBS. In the model therefore, a fraction θ, which represents the DT demands of queue 2, is moved to queue 1 where the demand is capable of being transmitted at a higher data rate. The challenge is to be able to find the value θ, the fraction of the demands from the farther queue to be moved to the nearer queue that will not be counterproductive but will rather optimise the total productivity of the network.

4. Analysis of Model

For the analysis carried out in this work, the model is restricted to two rings (queues) for simplicity. To analyse the queueing model developed, the following parameters are defined:

(i) Total number of subchannels (which is also equivalent to the number of multiple servers in each queue) $= N$

(ii) Arrival rate into queue 1 $= \lambda_1$

(iii) Arrival rate into queue 2 $= \lambda_2$

(iv) Fraction of queue 2 (DT demands) moved to queue 1 $= \theta$

(v) Total arrival into queue 1 $= \lambda_1 + \theta\lambda_2$

(vi) Total arrival into queue 2 $= \lambda_2 - \theta\lambda_2$

(vii) Total arrival into the network $\lambda = \lambda_1 + \lambda_2$, with the traffic intensities given as $\rho_1 = (\lambda_1 + \theta\lambda_2)/N\mu_1$; $\rho_2 = (\lambda_2 - \theta\lambda_2)/N\mu_2$; $\rho = \rho_1 + \rho_2$

(viii) Service rate of queue 1 $= \begin{cases} n\mu_1, & 0 \le n \le N; \\ N\mu_1, & N \le n \le Y \end{cases}$

(ix) Service rate of queue 2 $= \begin{cases} n\mu_2, & 0 \le n \le N; \\ N\mu_2, & N \le n \le Y \end{cases}$

The queueing model is a continuous-time Markov chain (CTMC) queue with a finite buffer. The model is shown in Figure 2. From Figure 2, it can be observed that the total arrival into queue 1 is the addition of the original arrival λ_1 and the fraction of arrival to queue 2 that is redirected to queue 1. Similarly, the total arrival to queue 2 becomes what is left of the original arrival after the fraction $\theta\lambda_2$ has been taken away. If there are no arrivals into queue 2 or no part of queue 2 is moved to queue 1, then $\theta\lambda_2 = 0$ and arrival to queue 1 is simply limited to λ_1.

Let $X = \{(k,l),\ 0 \le k \le Y,\ 0 \le l \le Y\}$ be the state space of the combined queues, where $k(l)$ represents the number of data packets in the system from queue 1(2). Hence, $(k,l) \in X$. The state space diagram is shown in Figure 3. $x_{k,l}(t)$ (for simplicity, this is subsequently written as $x_{k,l}$) is the probability that at time t there are (k,l) data packets in the system, implying that there are k packets in queue 1 and l packets in queue 2 available for transmission (including the packets in service). The system can now be studied at steady state using the stationary distribution conditions. If $\mathbf{x} = [x_{0,0}\ x_{0,1}\ \cdots\ x_{0,Y}\ x_{1,0}\ x_{1,1}\ \cdots\ x_{1,Y}\ \cdots\ x_{Y,0}\ x_{Y,1}\ \cdots\ x_{Y,Y}]$ is the steady state vector, the conditions $\mathbf{0} = \mathbf{xQ}$ and $1 = \mathbf{xe}$ for CTMCs hold, where $\mathbf{Q}$ is the generator matrix and $\mathbf{e}$ is the identity matrix.

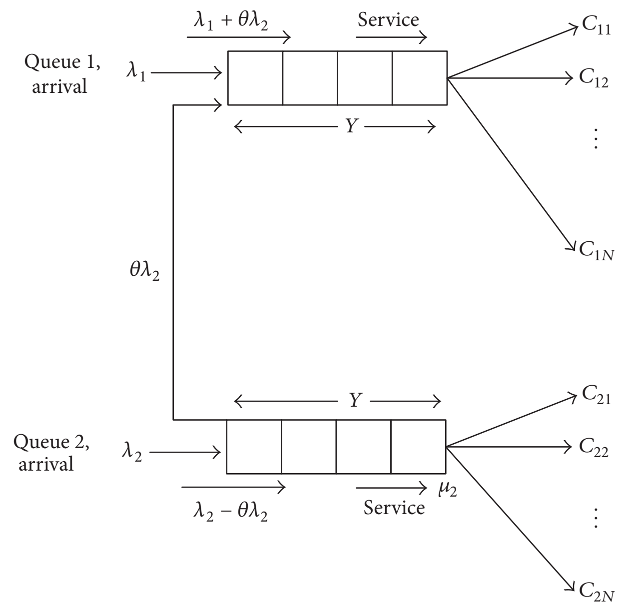

Figure 2: Queueing model for the heterogeneous CRN.

From the state space diagram as shown in Figure 3 and using the steady state conditions, the equilibrium balance equations are obtained as

$$0 = -\lambda x_{0,0} + \mu_2 x_{0,1} + \mu_1 x_{1,0},$$

$$0 = (\lambda_2 - \theta\lambda_2) x_{0,0} - (\lambda + \mu_2) x_{0,1} + 2\mu_2 x_{0,2} + \mu_1 x_{1,1},$$

$$0 = (\lambda_2 - \theta\lambda_2) x_{0,1} - (\lambda + 2\mu_2) x_{0,2} + 3\mu_2 x_{0,3} + \mu_1 x_{1,2},$$

$$0 = (\lambda_2 - \theta\lambda_2) x_{0,N-1} - (\lambda + N\mu_2) x_{0,N} + N\mu_2 x_{0,N+1} + \mu_1 x_{1,N},$$

$$0 = (\lambda_2 - \theta\lambda_2) x_{0,Y-1} - (\lambda_1 + \theta\lambda_2 + N\mu_2) x_{0,Y} + \mu_1 x_{1,Y},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{0,0} - (\lambda + \mu_1) x_{1,0} + \mu_2 x_{1,1} + 2\mu_1 x_{2,0},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{0,1} + (\lambda_2 - \theta\lambda_2) x_{1,0} - (\lambda + \mu_1 + \mu_2) x_{1,1} + 2\mu_2 x_{1,2} + 2\mu_1 x_{2,1},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{0,2} + (\lambda_2 - \theta\lambda_2) x_{1,1} - (\lambda + \mu_1 + 2\mu_2) x_{1,2} + 3\mu_2 x_{1,3} + 2\mu_1 x_{2,2},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{0,N} + (\lambda_2 - \theta\lambda_2) x_{1,N-1} - (\lambda + \mu_1 + N\mu_2) x_{1,N} + N\mu_2 x_{1,N+1} + 2\mu_1 x_{2,N},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{0,Y} + (\lambda_2 - \theta\lambda_2) x_{1,Y-1} - (\lambda_1 + \theta\lambda_2 + \mu_1 + N\mu_2) x_{1,Y} + 2\mu_1 x_{2,Y},$$

$$0 = (\lambda_1 + \theta\lambda_2) x_{N-1,0} - (\lambda + N\mu_1) x_{N,0} + \mu_2 x_{N,1} + N\mu_1 x_{N+1,0},$$

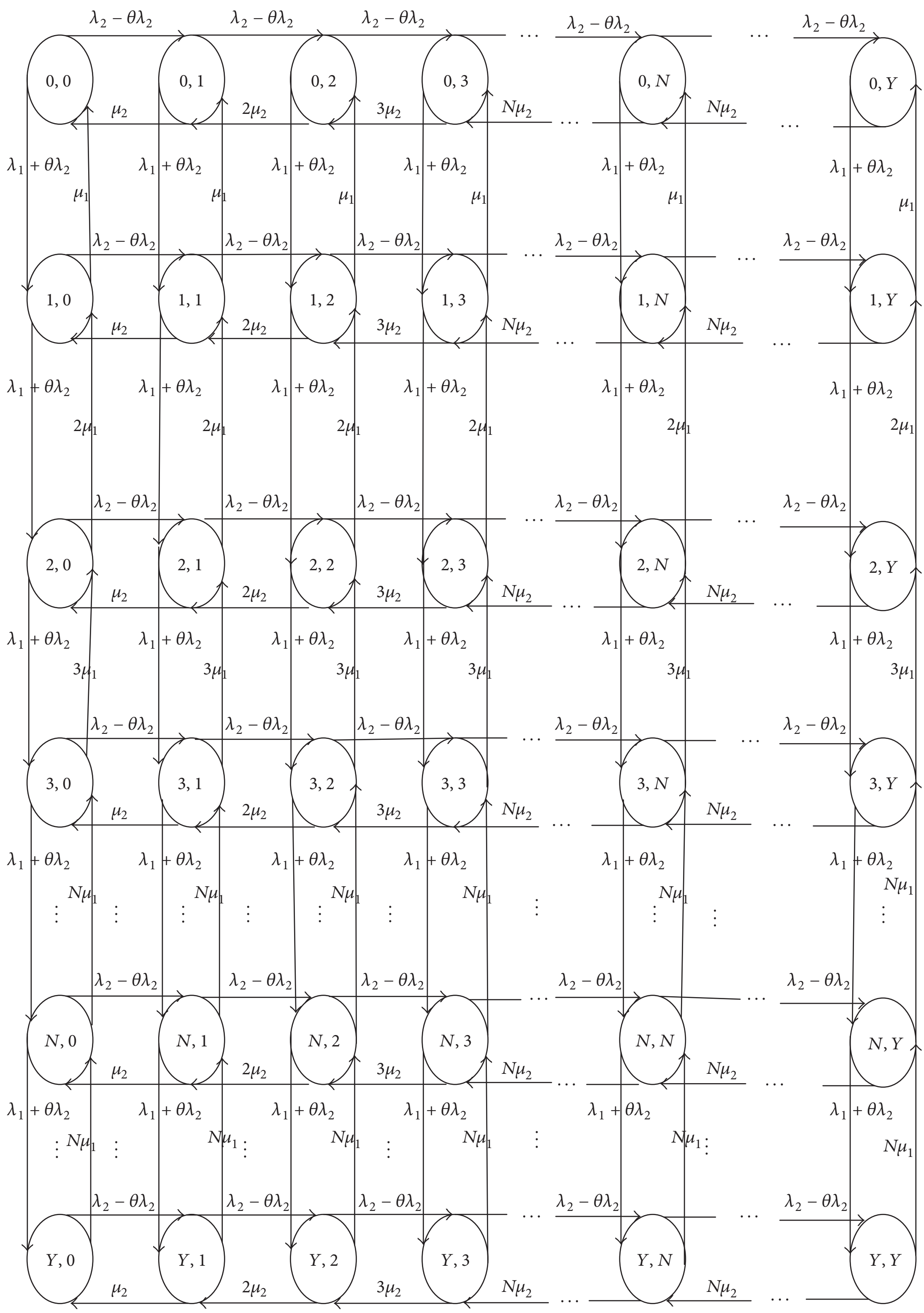

Figure 3: State space for the queueing model of RA in heterogeneous CRN.

$$
\begin{aligned}
0 &= (\lambda_1 + \theta\lambda_2)\, x_{N-1,1} + (\lambda_2 - \theta\lambda_2)\, x_{N,0} \\
&\quad - (\lambda + N\mu_1 + \mu_2)\, x_{N,1} + 2\mu_2 x_{N,2} + N\mu_1 x_{N+1,1}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{N-1,2} + (\lambda_2 - \theta\lambda_2)\, x_{N,1} \\
&\quad - (\lambda + N\mu_1 + 2\mu_2)\, x_{N,2} + 3\mu_2 x_{N,3} \\
&\quad + N\mu_1 x_{N+1,2}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{N-1,N} + (\lambda_2 - \theta\lambda_2)\, x_{N,N-1} \\
&\quad - (\lambda + N\mu_1 + N\mu_2)\, x_{N,N} + N\mu_2 x_{N,N+1} \\
&\quad + N\mu_1 x_{N+1,N}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{N-1,Y} + (\lambda_2 - \theta\lambda_2)\, x_{N,Y-1} \\
&\quad - (\lambda_1 + \theta\lambda_2 + N\mu_1 + N\mu_2)\, x_{N,Y} + N\mu_1 x_{N+1,Y}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{Y-1,0} - (\lambda_2 - \theta\lambda_2 + N\mu_1)\, x_{Y,0} \\
&\quad + \mu_2 x_{Y,1}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{Y-1,1} + (\lambda_2 - \theta\lambda_2)\, x_{Y,0} \\
&\quad - (\lambda_2 - \theta\lambda_2 + N\mu_1 + \mu_2)\, x_{Y,1} + 2\mu_2 x_{Y,2}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{Y-1,2} + (\lambda_2 - \theta\lambda_2)\, x_{Y,1} \\
&\quad - (\lambda_2 - \theta\lambda_2 + N\mu_1 + 2\mu_2)\, x_{Y,2} + 3\mu_2 x_{Y,3}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{Y-1,N} + (\lambda_2 - \theta\lambda_2)\, x_{Y,N-1} \\
&\quad - (\lambda_2 - \theta\lambda_2 + N\mu_1 + N\mu_2)\, x_{Y,N} + N\mu_2 x_{Y,N+1}, \\
0 &= (\lambda_1 + \theta\lambda_2)\, x_{Y-1,Y} + (\lambda_2 - \theta\lambda_2)\, x_{Y,Y-1} \\
&\quad - (N\mu_1 + N\mu_2)\, x_{Y,Y}.
\end{aligned} \tag{1}
$$

The set of linear equations (1), being finite, can be solved simultaneously in a number of ways. Three different approaches for solving these types of equations and obtaining solutions are discussed.

4.1. Standard Approach. The set of linear equations (1) can be solved using any of the well-established standard numerical analysis methods for solving simultaneous linear equations. These methods are broadly classified into two which are the direct approach (such as by using backward substitution or Gaussian elimination) and the iterative approach (such as Jacobi or Gauss-Seidel methods). Using any of these approaches, state probabilities are obtained, which can then be used to obtain the various performance measures of interest. These approaches have been used extensively in obtaining state probabilities, most especially for networks with manageable queue sizes. Detailed analysis on these methods has been well presented in [20] and it is therefore not necessary to repeat them in this work. References [21, 22] are classic examples of the use of these approaches in analysing queueing models and obtaining state probabilities for CRN. The major challenge with this approach is the high complexity and computational demands of the analysis, especially for large networks. Due to the complex nature of CRN, it is arguable whether the approach is best suited for its RA optimisation determination, especially when heterogeneous considerations are made, as developed in this work.

4.2. Equivalent Two-Class System. Although the developed queueing problem of RA in CRN may be solved using the standard approach, other (presumably better) methods for obtaining solutions can as well be investigated. The second approach being explored here is referred to as an equivalent two-class, first-come first-served system. By a careful consideration of the developed model, it is observed that the overall system can be closely comparable to a two-class, nonpriority queueing model in which customers are served using the first-come first-served discipline with different arrival rates and service rates [23]. According to [23], such a system can be regarded as an M/G/1 queue, where G is a mixture of two exponential distributions (a hyperexponential distribution) and packet arrivals are grouped into a single arrival stream. The analysis of this system has already been carried out in works like [23] and gives the following expressions for the queue lengths $L_q^{(1)}$, $L_q^{(2)}$, and L_q representing the queue length of queue 1, queue 2, and the overall queue, respectively:

$$
\begin{aligned}
L_q^{(1)} &= \frac{(\lambda_1 + \theta\lambda_2)(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho}, \\
L_q^{(2)} &= \frac{(\lambda_2 - \theta\lambda_2)(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho}, \\
L_q &= \frac{\lambda(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho}.
\end{aligned} \tag{2}
$$

4.3. State Reduction. The final solution approach to be investigated is the use of state reduction. This approach is the most significant of the three approaches considered in this work. State reduction is a technique for obtaining steady state probabilities for finite (and even infinite) state Markov chains. The approach is developed by Grassmann, Taksar, and Heyman and is generally referred to as the GTH algorithm [24]. This approach is very important in that it is employed to study the effect of and to find optimal values for the parameter θ for different possible CRN scenarios. By varying θ, the arrival rates into each queue can be adjusted and the effects on the overall performance of the network investigated. After obtaining the equilibrium probabilities of the Markov chain through state reduction, the optimum value of the parameter θ is then obtained through Newton's method. It is important to note that the definition of θ is not necessarily limited to the fraction of the DT demands being moved from one queue to the other alone, as is being considered in this work. It could be defined to be any other factor, for instance, a fraction of the higher priority demands. The important question to be addressed is how to determine the value of θ that will maximise the overall productivity of the network. The derivations and analysis of how the state reduction approach is used in obtaining state probabilities and the optimal value of θ are provided in Appendices A and B.

4.4. Performance Metrics. Performance measures employed to demonstrate the effect of θ on the network, as considered in this paper, are the blocking probability and system throughput. Both of these measures are obtainable using the steady state probabilities. Other performance measures, such as the average number of packets in the queue or in the system and the average waiting time of a packet in the queue or before the packet's transmission is completed, are easily obtainable from the two measures evaluated but for the sake of brevity, they are not considered in this paper. The blocking probability is the probability that a packet that arrives in the network is blocked or dropped because of meeting a full system (i.e., all servers are fully engaged and the waiting queue is full) and is therefore not served. In steady state, the blocking probability P_B is defined as

Blocking probability, P_B = Pr{a packet arrives to meet a full system},

$$P_B = \Pr\{\text{system is full}\} = x_{Y,Y}. \tag{3}$$

The system throughput is defined as the number of total arrivals that are eventually served, that is, the number of total arrivals that are not dropped or blocked. The throughput is in effect the effective arrival rate, represented as λ_e. The throughput is given as follows:

Throughput = effective arrival rate, λ_e. That is:

Throughput = total arrival rate, λ × Pr{system is NOT full}, and hence,

$$\text{Throughput} = \lambda\left(1 - P_B\right) = \lambda\left(1 - x_{Y,Y}\right). \tag{4}$$

4.5. Scalability of the Model. It is important to discuss the scalability of the model. If the rings (and invariably, the queues) are extended beyond two, the network becomes larger and indeed, it becomes more difficult to analyse such networks. However, analysing queueing models with more than two queues, though difficult, is not impossible to achieve. Two propositions on how this can be achieved are provided. First, if the queues are within a limited number, say three or four, then it is possible to analyse the network using the exact same ideas presented in this section, that is, by using stationary distribution conditions to obtain equilibrium balance equations which are then solved using any of the approaches already discussed to obtain state probabilities. It has to be stated, however, that the number of equations to be solved grows exponentially with the number of queues being considered, meaning that the computational complexity can be significantly increased when the networks are large. The second proposition, which may be more appropriate for larger networks with higher number of rings (and queues), is to use the matrix-analytic approach developed by Neuts and Miller [25, 26]. This approach has been proven to be useful in solving both continuous-time and discrete-time queueing models with a considerably high level of accuracy and reliability. The computational complexity is also manageable, even for large networks.

5. Results and Discussion

In this section, numerical results of the developed model are presented. The results are obtained using the analysis from the state reduction method, which, as earlier argued for the developed model, is the best of the methods presented for obtaining state probabilities and studying the effect of the fraction of packets moved from queue 2 to queue 1, θ. For the purpose of validation, however, Gaussian elimination (an example of direct method) was also used to obtain state probabilities and the values obtained correspond exactly to those obtained using state reduction. Generally, values of state probabilities are usually typically very close, irrespective of the methods employed in obtaining them. The major difference is always in the level of complexity in the computations when different methods are employed.

For ease of analysis, the network model, comprising two separate queues, is limited to $N = 2$ servers (subchannels) in each queue. A queue length of $Y = 2$ is equally employed in each queue. The model, parameters used (number of servers, queue length, etc.), and the results obtained are comparable to, and validated by, the work in [27]. The model is however easily scalable, and the results obtained are fair representations of larger networks. For the first consideration (as presented in the first set of plots), SU demands in queue 1 are served at a data rate of 6 bits per symbol (64-QAM modulation) for each unit of time, while demands in queue 2 are served at a rate of 2 bits per symbol (4-QAM modulation) for each unit of time. In the second consideration, the service rates at both queues are reduced to study the effect of such reduction on the overall performance of the network. SU demands in queue 1 are therefore served at 2 bits per symbol (4-QAM modulation) for each time unit, while demands in queue 2 are served at a rate of 1 bit per symbol (BPSK modulation) for each time unit. In all analyses, the arrival rate to each of the queues is gradually increased from 1 to 10 bits per symbol per unit of time. The value of θ is varied between 0 and 100% of λ_2. Performance results of blocking probability and throughput of the heterogeneous CRN are presented and discussed.

Figures 4 and 5 present the blocking probability and throughput performance measures, respectively, when the service rate is at 6 bits per symbol per unit time for queue 1 and 2 bits per symbol per unit time for queue 2. From the results, it can be observed that, at low arrival rates, the blocking probability, which is the probability of finding the system full, is very low, implying that the system can effectively service almost all arrivals. The effective arrival rate, or throughput, is thus very high, in fact, close to the total number of arrivals in the system. Again, it can be observed that, by gradually increasing arrivals, the effective arrival steadily increases, although at an increasing blocking probability. It therefore implies that, by increasing the arrival rate, the system throughput can indeed be increased, albeit at a decreasing rate. Eventually, a highest possible value of the throughput is obtained because of the obvious limitation in server capacity. Also, the results show that, by increasing θ, the blocking probability decreases while the throughput

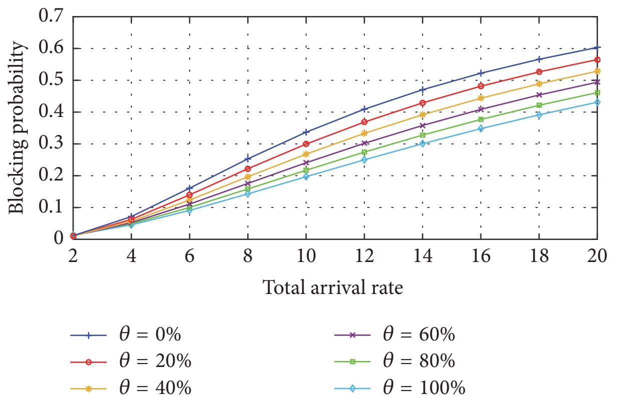

Figure 4: Blocking probability against total arrival rate for different θ values. Service rates are 6 bits/symbol/unit time in queue 1 and 2 bits/symbol/unit time in queue 2.

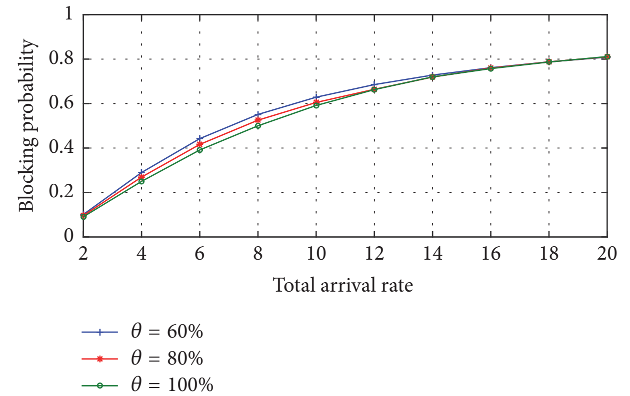

Figure 6: Blocking probability against total arrival rate for different θ values. Service rates are 2 bits/symbol/unit time in queue 1 and 1 bit/symbol/unit time in queue 2.

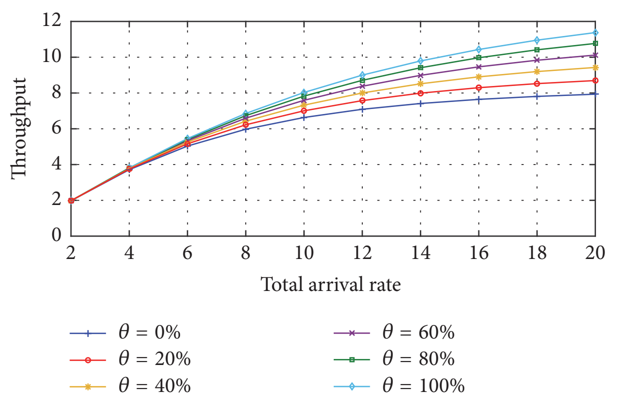

Figure 5: Throughput against total arrival rate for different θ values. Service rates are 6 bits/symbol/unit time in queue 1 and 2 bits/symbol/unit time in queue 2.

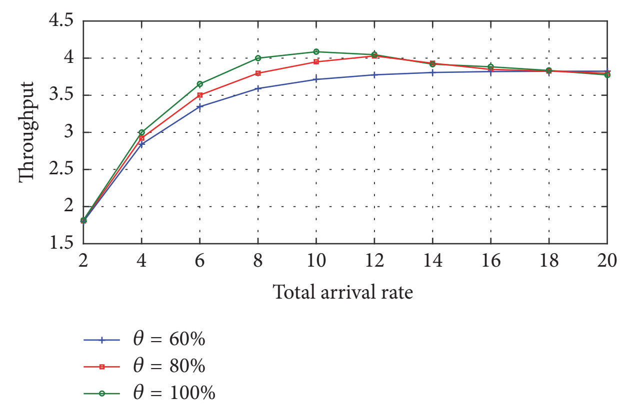

Figure 7: Throughput against total arrival rate for different θ values. Service rates are 2 bits/symbol/unit time in queue 1 and 1 bit/symbol/unit time in queue 2.

increases, signifying an improvement in the overall performance of the network.

In Figures 6 and 7, the results of both blocking probability and throughput are likewise presented; only in this instance, the service rates of the servers are 2 bits per symbol per unit time in queue 1 and 1 bit per symbol per unit time in queue 2, respectively. The results show similar trends to those presented in Figures 4 and 5, except that, at some point, the improvement in performance due to an increase in θ is completely eliminated and rather, a gradual reduction in performance is observed. This is because, as more and more of the demands of queue 2 are moved to queue 1 in the hope of being served more quickly, a tipping point (which invariably corresponds to the optimum θ value) is reached after which an additional increase in the value of queue 2 demands moved to queue 1 no longer improves the overall performance. Rather, the blocking probability increases more significantly, owing to the continuous increase in queue 1, resulting in a decrease in the network throughput as well as a poorer overall network.

It is very significant to observe from a comparison of the results presented in Figures 4 and 5 and Figures 6 and 7 that if all the demands of queue 2 are DT and queue 1 has a very high service rate (high enough to conveniently accommodate demands from both queues 1 and 2, as observable in Figures 4 and 5), it would be practically unnecessary to have any service at all in queue 2. This is because, in Figures 4 and 5, even at $\theta = 100\%$, that is, after moving the entirety of the demands in queue 2 to queue 1 (with the assumption that all the demands are DT), the throughput performance had still not declined. In comparison with Figures 6 and 7, where the service rate in queue 1 is relatively lower, after some optimum value of θ, an increase in its value results in a decline in the overall performance of the system because the capacity of queue 1 has been overstretched and therefore both the blocking probability and the throughput performances begin to depreciate. It is therefore important for any given problem formulation always to find the optimum value of θ and to allow only that fraction of queue 2 demands to be moved to queue 1 in order to achieve an overall best (optimum) performance for the network.

Figures 8 and 9 describe the performance of the blocking probability and the throughput as a function of θ for different values of arrival. From the plots, it is possible to determine

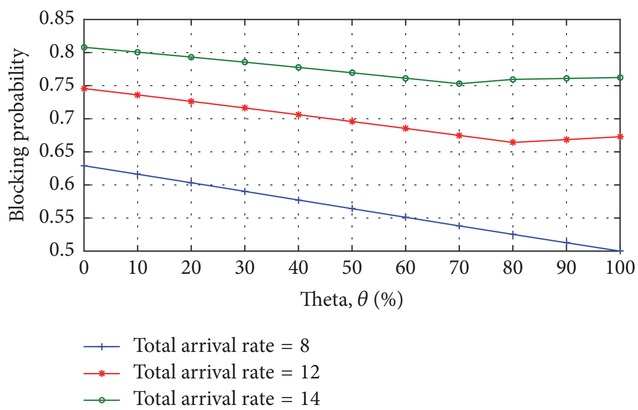

Figure 8: Blocking probability against increasing θ values for different total arrival rate. Service rates are 2 bits/symbol/unit time in queue 1 and 1 bit/symbol/unit time in queue 2.

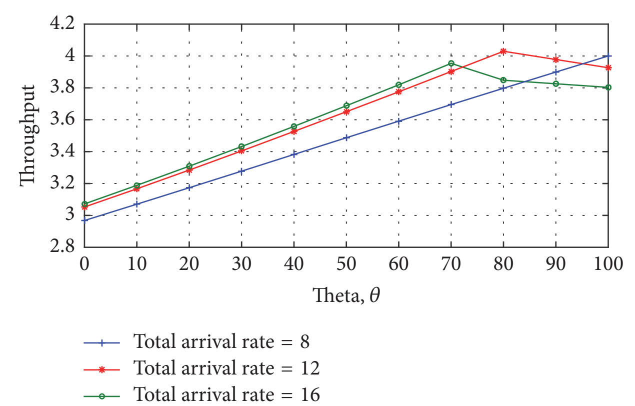

Figure 9: Throughput against increasing θ values for different total arrival rate. Service rates are 2 bits/symbol/unit time in queue 1 and 1 bit/symbol/unit time in queue 2.

the value of the total rewards and to observe its effects on the overall network performance. According to Figure 9, the total rewards for any given value of θ are simply the difference between the throughput value at the given θ and the throughput value at $\theta = 0\%$, for any value of the arrival rates. Similarly, the optimal value of θ for each of the arrival rates considered can easily be observed. At the total arrival rate of 8 bits per symbol per unit time, an optimal value for θ was never realised because, even at $\theta = 100\%$, improvement in performance was still being observed. However, for total arrival rates of 12 and 16 bits per symbol per unit time, an optimal value of θ is realised at the peaks of the plots, after which the performance of the network begins to depreciate. Finally, it should be noted that a change in either the arrival or service rates of the queues will shift the optimum value of θ, either to the right or to left of the plots. Therefore, a universal optimum value of θ cannot necessarily be realised; only specific values for specifically developed problem formulations and network parameters (such as arrival rates, service rates) are feasible.

6. Conclusions

In this paper, a queueing model has been developed and studied for RA optimisation in heterogeneous CRN with different users having buffered data in queues waiting for transmission. In the model, by leveraging the different delay tolerance profiles of user demands and the mobility of users, a queueing system is developed that can achieve a greater overall capacity in the RA of the network while still satisfying the varying demands of each of the users or user categories. In the RA model developed, the user demands are classified as either DS or DT, while a fraction θ of the DT demands in the queue farther from the SUBS is moved to the queue nearer to the SUBS for possible faster transmission. This is achievable because the ring (or equivalently, the queue) nearer to the SUBS transmits at a higher rate, which implies that its service rate (and thus its capacity) is higher. The ring can therefore accommodate a larger number of SUs (and their demands) than the ring farther from the SUBS. By changing θ over a given range, its effect on the overall performance of the network is analysed and optimal values for θ, depending on prevalent arrival or/and service conditions, are realised. The results of the various analyses presented show that, with the developed model, an improvement in blocking probability and an optimal throughput can be effectively achieved for the heterogeneous CRN with buffered data.

Appendix

A. Obtaining the Steady State Probabilities and Achieving Optimal θ Values through State Reduction

For easier depiction, let i represent the state (k, l). Also, for each $i \in X$, let r_i be the additional data rate (reward) achieved through the movement of θ data packets from queue 1 to queue 2. From any state $i \in X$, the rate of going to another state $j \in X$ is q_{ij}, where q_{ij}'s form the elements of the generator matrix Q. Implicitly, both r_i and q_{ij} depend on the parameter θ. Hence, to achieve an optimal total data rate (maximum throughput), an optimal value of θ has to be obtained such that the total additional data rates (total rewards) are maximised when the system is in equilibrium (steady state). This therefore defines the problem as that of determining the optimal value of θ when the network is in steady state. Let T be the total expected rewards (total additional data rate due to θ); the objective of the optimisation problem is given as

$$\max T(\theta) = \sum_{i \in X} x_i r_i, \tag{A.1}$$

and the constraints are the limitations in both arrival and service rates of the network, as well as the time delays and the minimum data rate for an acceptable QoS.

Since θ is a variable entity, differential calculus can be employed in finding its optimal value. Let $T'(\theta)$ and $T''(\theta)$ be the respective first and second derivative of $T(\theta)$ with respect to θ (for ease of reference, $T(\theta)$, $T'(\theta)$, and $T''(\theta)$

are subsequently represented as T, T', and T'', resp.). T is maximised by solving $T' = 0$ for θ. To solve $T' = 0$ using Newton's method, T'' has to be obtained. Also, since the problem is a maximisation problem, T'' has to be negative. To obtain T' and T'', intermediate results from state reduction are used [27]. State reduction is first employed to obtain the steady state vector $\mathbf{x}$, the values of which are then used to obtain T and its derivatives.

Using the state reduction method to obtain $\mathbf{x}$ requires a relaxation of the steady state equation $\mathbf{xe} = \sum_{i\in X} x_i = 1$, so that the sum of x_i is no longer 1. A different solution is obtained, say, g_i and let its vector be $\mathbf{g}$. To differentiate g_i from x_i, g_i is referred to as the chance that there are i data packets in the network, while x_i is the probability that there are i data packets in the network at any given time. It can immediately be observed that the sum of g_i is not 1. Rather, a certain state, say state b, is given a chance $g_b = 1$ and the chances for all other states are calculated accordingly. Then, to find the probability x_i of being in state i, the chance g_i is divided by the sum of all chances $G = \sum_{i\in X} g_i$. Hence,

$$x_i = \frac{g_i}{\sum_{i\in X} g_i}, \qquad x_i = \frac{g_i}{G}, \tag{A.2}$$

or $g_i = x_i G$. Differentiating twice with respect to θ, this becomes

$$g_i' = x_i' G + x_i G', \qquad g_i'' = x_i'' G + 2x_i' G' + x_i G'', \tag{A.3}$$

and hence,

$$x_i' = \frac{g_i' - x_i G'}{G}, \qquad x_i'' = \frac{g_i'' - 2x_i' G' - x_i G''}{G}. \tag{A.4}$$

The values g_i, g_i', g_i'' are obtained using the state reduction (or GTH) algorithm. The algorithm is provided in Appendix B. Once these values have been obtained, substituting into the set of equations above gives the corresponding x_i, x_i', x_i''. Also, after obtaining g_i, g_i', g_i'', obtaining the optimal value for θ is achieved by applying Newton's method. This is carried out as follows.

Assume that there are $M + 1$ states numbered from 0 to M and M is finite. Recall $T = \sum_{i=0}^{M} x_i r_i$. But $x_i = g_i/G$. Substituting for x_i gives

$$T = \frac{\sum_{i=0}^{M} g_i r_i}{G}, \qquad TG = \sum_{i=0}^{M} g_i r_i. \tag{A.5}$$

Taking first and second derivatives with respect to θ becomes

$$T'G + TG' = \sum_{i=0}^{M} \left(g_i' r_i + g_i r_i'\right), \qquad T''G + 2T'G' + TG'' = \sum_{i=0}^{M} \left(g_i'' r_i + 2g_i' r_i' + g_i r_i''\right). \tag{A.6}$$

The above submissions yield the following equations for T, T', and T'':

$$T = \frac{\sum_{i=0}^{M} g_i r_i}{G}, \qquad T' = \frac{\sum_{i=0}^{M} \left(g_i' r_i + r_i' g_i\right) - TG'}{G}, \qquad T'' = \frac{\sum_{i=0}^{M} \left(g_i'' r_i + 2g_i' r_i' + r_i'' g_i\right) - 2T'G' - TG''}{G}. \tag{A.7}$$

To obtain the optimal value of θ, an approximate value of θ is first chosen, say θ_m, and a new value θ_{m+1}, presumably a better approximation, is obtained according to Newton's method as follows:

$$\theta_{m+1} = \theta_m - \frac{T'}{T''}. \tag{A.8}$$

Substituting for T' and T'' gives

$$\theta_{m+1} = \theta_m - \frac{\sum_{i=0}^{M} \left(g_i' r_i + r_i' g_i\right) - TG'}{\sum_{i=0}^{M} \left(g_i'' r_i + 2g_i' r_i' + r_i'' g_i\right) - 2T'G' - TG''}. \tag{A.9}$$

In the above formulation, maximising θ is subject to $T'' < 0$. If not, then the sign of T' has to be considered. If $T' > 0$, the rewards can be increased by increasing θ. If $T' < 0$, the rewards can be increased by decreasing θ.

B. The State Reduction and/or GTH Algorithm (See [27])

The values for g_i, g_i', and g_i'', required to obtain the steady state probabilities (and their derivatives), x_i, x_i', and x_i'', can be obtained through state reduction as explained below.

Recall that $g_i = x_i G$.

Let $\mathbf{g} = \mathbf{x}G$, and hence,

$$\mathbf{g}Q = 0, \tag{B.1}$$

where Q is the generator matrix. Taking first and second derivatives with respect to θ, we have

$$\mathbf{g}'Q = -\mathbf{g}Q', \qquad \mathbf{g}''Q = -2\mathbf{g}'Q' - \mathbf{g}Q''. \tag{B.2}$$

The set of equations given above can be solved using state reduction to obtain $\mathbf{g}$, $\mathbf{g}'$, and $\mathbf{g}''$. The state reduction algorithm is as follows.

The states are numbered from 0 to M, giving a total of $M+1$ states. M is finite. For each state, we find a steady state equation using $\mathbf{g}G = 0$.

For state j, the steady state equation is given as

$$\sum_{i=0}^{M} g_i q_{ij} = 0, \tag{B.3}$$

where q_{ij} are the elements of the generator matrix Q. By using Gaussian elimination, equations $m+1, m+2, \ldots, M$ can be used to eliminate $g_{m+1}, g_{m+2}, \ldots, g_M$ from equations $0, 1, 2, \ldots, m$ so that a new set of equations is obtained. The new set of equations is represented by

$$\sum_{i=0}^{m} g_i q_{ij}^{m} = 0. \tag{B.4}$$

Grassmann, Taksar, and Heyman (GTH) did show that q_{ij}^{m} can indeed be interpreted as the transition rates of a continuous-time Markov chain, meaning that it is unnecessary to calculate the diagonal elements q_{ii}^{m}. Specifically, if s_m is defined as $-q_{mm}^{m}$, then,

$$s_m = -q_{mm}^{m} = \sum_{i=0}^{m} q_{mj}^{m}. \tag{B.5}$$

We can use normal elimination methods to find all q_{ij}^{m} recursively, starting with $q_{ij}^{M} = q_{ij}$ and then calculating $q_{ij}^{M-1}, q_{ij}^{M-2}, \ldots, q_{ij}^{1}$. This gives, for $i, j < m$,

$$q_{ij}^{m-1} = q_{ij}^{m} - \frac{q_{im}^{m} q_{mj}^{m}}{q_{mm}^{m}}. \tag{B.6}$$

Substituting for s_m becomes

$$q_{ij}^{m-1} = q_{ij}^{m} + \frac{q_{im}^{m} q_{mj}^{m}}{s_m}. \tag{B.7}$$

Better still, a more recent approach to state reduction, modified by Grassmann [27], which gives a more convenient and quicker recursion, is obtained by making

$$\begin{aligned} q_{ij}^{m} &= q_{ij}^{m+1} + \frac{q_{i\,m+1}^{m+1} q_{m+1\,j}^{m+1}}{s_{m+1}}, \\ q_{ij}^{m} &= \left(q_{ij}^{m+2} + \frac{q_{i\,m+2}^{m+2} q_{m+2\,j}^{m+2}}{s_{m+2}} \right) + \frac{q_{i\,m+1}^{m+1} q_{m+1\,j}^{m+1}}{s_{m+1}}. \end{aligned} \tag{B.8}$$

Hence,

$$q_{ij}^{m} = q_{ij} + \sum_{p=m+1}^{M} \frac{q_{ip}^{p} q_{pj}^{p}}{s_p}. \tag{B.9}$$

For $m > i$, this gives

$$q_{im}^{m} = q_{im} + \sum_{p=m+1}^{M} \frac{q_{ip}^{p} q_{pn}^{p}}{s_p}. \tag{B.10}$$

For $m > j$, this gives

$$q_{mj}^{m} = q_{mj} + \sum_{p=m+1}^{M} \frac{q_{np}^{p} q_{pj}^{p}}{s_p}. \tag{B.11}$$

b_{ij} is now defined as

$$\begin{aligned} b_{ij} &= \frac{q_{ij}^{j}}{s_j}, \quad i < j, \\ b_{ij} &= q_{ij}^{i}, \quad i > j. \end{aligned} \tag{B.12}$$

The b_{ij} can be calculated by row, starting with row M then continuing with row $M-1$ and so on. Once b_{ij} are calculated, normal back-substitution can be used to obtain g_j thus.

Set $g_0 = 1$ and evaluate $g + i = 1, 2, \ldots, M$ as follows:

$$g_j = \sum_{i=0}^{j-1} g_i b_{ij} = b_{0j} + \sum_{i=1}^{j-1} g_i b_{ij}, \quad j > 0. \tag{B.13}$$

The equations for g_i' and g_i'' do have the same structure as the equations for g_i. This can be verified by using the fact that $g_0 = 1$ and writing $\sum_{i=0}^{M} g_i q_{ij} = 0$ as follows:

$$q_{0j} + \sum_{i=1}^{M} g_i q_{ij} = 0, \quad j \geq 0. \tag{B.14}$$

Since $g_0 = 1$, g_0' must be 0. Also, the jth equation obtained from $\mathbf{g}'Q = -\mathbf{g}Q'$ has the jth element of $\mathbf{g}Q'$ as its constant term. If this constant term is denoted as q_{0j}^{*}, then,

$$q_{0j}^{*} + \sum_{i=1}^{M} g_i' q_{ij} = 0, \quad j \geq 0. \tag{B.15}$$

The last two equations are similar, except that all q_{0j} are replaced by q_{0j}^{*}. Consequently, if b_{ij}^{*} are the coefficients obtained by eliminating g_i' from $\mathbf{g}'Q = -\mathbf{g}Q'$, $b_{ij}^{*} = b_{ij}$ except for $i = 0$. From the algorithm given above,

$$b_{0j}^{*} = \frac{\left(q_{0j}^{*} + \sum_{p=j+1}^{M} b_{ip} b_{pj} \right)}{s_j}, \quad j = M, M-1, \ldots, 1. \tag{B.16}$$

Again using the fact that $g_0' = 0$,

$$g_j' = b_{0j}^{*} + \sum_{i=1}^{j-1} g_i' b_{ij}. \tag{B.17}$$

Similarly, if $q_{0j}^{**} = (2\mathbf{g}'Q' + \mathbf{g}Q'')_j$, b_{0j}^{**} can be obtained by a similar representation to b_{0j}^{*} and using back-substitution to find g_j'' from the equation

$$g_j'' = b_{0j}^{**} + \sum_{i=1}^{j-1} g_i'' b_{ij}. \tag{B.18}$$

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Y. Zeng, Y. C. Liang, A. Hoang, and R. Zhang, "A review on spectrum sensing for cognitive radio: Challenges and solutions," *EURASIP Journal on Advances in Signal Processing*, vol. 1, Article ID 381465, 2010.

[2] L. Wang, W. Xu, Z. He, and J. Lin, "Algorithms for optimal resource allocation in heterogeneous cognitive radio networks," in *Proceedings of the 2009 2nd Conference on Power Electronics and Intelligent Transportation System, PEITS 2009*, pp. 396–400, December 2009.

[3] M. Kaplan and F. Buzluca, "A dynamic spectrum decision scheme for heterogeneous cognitive radio networks," in *Proceedings of the 2009 24th International Symposium on Computer and Information Sciences, ISCIS 2009*, pp. 697–702, September 2009.

[4] L. Zheng and C. W. Tan, "Cognitive radio network duality and algorithms for utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 3, pp. 500–513, 2013.

[5] B. S. Awoyemi, B. T. J. Maharaj, and A. S. Alfa, "Solving resource allocation problems in cognitive radio networks: a survey," *Eurasip Journal on Wireless Communications and Networking*, vol. 2016, no. 1, article no. 176, 2016.

[6] B. S. Awoyemi, B. T. Maharaj, and A. S. Alfa, "Resource allocation for heterogeneous cognitive radio networks," in *Proceedings of the 2015 IEEE Wireless Communications and Networking Conference, WCNC 2015*, pp. 1759–1763, March 2015.

[7] B. Awoyemi, B. Maharaj, and A. Alfa, "Optimal resource allocation solutions for heterogeneous cognitive radio networks," *Digital Communications and Networks*, vol. 3, no. 2, pp. 129–139, 2017.

[8] B. S. Awoyemi, B. T. Maharaj, and A. S. Alfa, "QoS provisioning in heterogeneous cognitive radio networks through dynamic resource allocation," in *Proceedings of the 12th IEEE AFRICON International Conference, AFRICON 2015*, September 2015.

[9] D. Xu, Z. Feng, and P. Zhang, "Resource allocation for heterogeneous services in multiuser cognitive radio networks," *International Journal of Communication Systems*, pp. 1–14, 2016.

[10] M. G. Adian, H. Aghaeinia, and Y. Norouzi, "Optimal resource allocation for opportunistic spectrum access in heterogeneous MIMO cognitive radio networks," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 1, pp. 74–83, 2016.

[11] M. G. Adian and H. Aghaeinia, "Optimal resource allocation in heterogeneous MIMO cognitive radio networks," *Wireless Personal Communication*, vol. 76, no. 1, pp. 23–39, 2014.

[12] M. G. Adian and H. Aghaeinia, "Optimal resource allocation for opportunistic spectrum access in multiple-input multiple-output–orthogonal frequency division multiplexing based cooperative cognitive radio networks," *IET Signal Processing*, vol. 7, no. 7, pp. 549–557, 2013.

[13] S. Wang, M. Ge, and C. Wang, "Efficient resource allocation for cognitive radio networks with cooperative relays," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2432–2441, 2013.

[14] S. Wang, Z.-H. Zhou, M. Ge, and C. Wang, "Resource allocation for heterogeneous cognitive radio networks with imperfect spectrum sensing," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 3, pp. 464–475, 2013.

[15] A. Alshamrani, X. S. Shen, and L. L. Xie, "QoS provisioning for heterogeneous services in cooperative cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 819–830, 2011.

[16] C. Shi, Y. Wang, and P. Zhang, "Joint spectrum sensing and resource allocation for multi-band cognitive radio systems with heterogeneous services," in *Proceedings of the 2012 IEEE Global Communications Conference, GLOBECOM 2012*, pp. 1180–1185, December 2012.

[17] A. El Shafie, A. Sultan, and T. Khattab, "Band allocation for cognitive radios with buffered primary and secondary users," in *Proceedings of the 2014 IEEE Wireless Communications and Networking Conference, WCNC 2014*, pp. 1508–1513, April 2014.

[18] P. Mitran, L. B. Le, and C. Rosenberg, "Queue-aware resource allocation for downlink OFDMA cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 10, pp. 3100–3111, 2010.

[19] B. Zhao and V. Friderikos, "A queuing-based delay-tolerant scheme for energy efficiency over cognitive radio networks," in *Proceedings of the 2012 IEEE Globecom Workshops, GC Wkshps 2012*, pp. 326–330, December 2012.

[20] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*, Princeton University Press, Princeton, NJ, USA, 2009.

[21] A. S. Zahmati, X. Fernando, and A. Grami, "Steady-state Markov chain analysis for heterogeneous cognitive radio networks," in *Proceedings of the 33rd IEEE Sarnoff Symposium*, pp. 1–5, IEEE, April 2010.

[22] A. S. Zahmati, X. Fernando, and A. Grami, "A continuous-time Markov chain model and analysis for cognitive radio networks," *International Journal of Communication Networks and Distributed Systems*, vol. 8, no. 3/4, pp. 195–212, 2012.

[23] D. Gross, J. F. Shortle, and J. M. Thompson, *Fundamentals of Queueing Theory*, John Wiley & Sons, Hoboken, NJ, USA, 4th edition, 2008.

[24] I. Sonin, "The state reduction and related algorithms and their applications to the study of Markov chains, graph theory, and the optimal stopping problem," *Advances in Mathematics*, vol. 145, no. 2, pp. 159–188, 1999.

[25] M. Neuts, *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*, Algorithmic Approach, Dover Publications, 1981.

[26] D. R. Miller, "Computation of steady-state probabilities for $M/M/1$ priority queues," *Operations Research. The Journal of the Operations Research Society of America*, vol. 29, no. 5, pp. 945–958, 1981.

[27] W. K. Grassmann, "Optimizing steady state Markov chains by state reduction," *European Journal of Operational Research*, vol. 89, no. 2, pp. 277–284, 1996.