Spectrum Allocation and QoS Provisioning Framework for Cognitive Radio With Heterogeneous Service Classes

Rahman Doost-Mohammady, *Member, IEEE*, M. Yousof Naderi, *Member, IEEE*, and Kaushik Roy Chowdhury, *Member, IEEE*

***Abstract*—Cognitive radio (CR) networks will enable dynamic spectrum re-use and thereby accelerate the adoption of high bandwidth services in available licensed frequencies with better channel characteristics. However, the possibility of the licensed user reclaiming the channel raises additional concerns on how best to reserve resources for secondary users (SUs) that are likely to have different qualities of service (QoSs) depending on their application requirements. This paper addresses the problem of spectrum resource management for co-located SUs with both streaming and intermittent data by efficiently identifying the number of backup channels that will ensure seamless end to end service. The contributions of this paper are threefold: First, a comprehensive analytical framework based on queueing theory is devised to calculate the theoretical delay in accessing the spectrum depending on the required QoS, with guidelines on how to optimize the set of back-up channels for possible future use; second, a method of spectrum allocation for SUs with these different QoS demands is formulated, especially as they co-exist and affect the performance of each other; third, a case study of applying these techniques in a novel application area of wireless medical telemetry is presented. Results reveal that the simulated spectral efficiency of the channel allocation using our approach matches closely with our theoretical predictions, within a 5% bound.**

***Index Terms*—Dynamic spectrum access, channel allocation, QoS, queuing, Markov chain.**

I. Introduction

THE spectacular growth in wireless applications has raised concerns on whether the current state of spectrum availability will scale proportionately. The unlicensed (ISM) bands are being used by millions of wireless devices for streaming video and essential data communication. Over the past few years, consumers have transitioned from the 900 MHz to the 2.4 GHz band, and an increasing number of new product developments today target the 5 GHz ISM band. The effort to identify additional wireless spectrum with markedly reduced congestion has led to the radically different concept of cognitive radio, wherein individual radios identify portions of the spectrum, and opportunistically transmit when the licensed or primary users (PUs) are not currently active [1], [2]. When multiple different secondary users (SUs) identify the same set of available channels, the task of allocating them adequate portions of the spectrum is a non-trivial task. Not only must the individual quality of service (QoS) demands be met, but also the new assignment of spectrum to one set of users may in-turn impact adversely the performance of a different set of users that have already been allotted the same portion of the spectrum. The key questions that this paper aims to address are: i) under which conditions must SUs be allowed to share spectrum, and when must they be assigned completely exclusive spectrum? ii) what must be the size of the spectrum *chunks* that can be allotted per application? iii) how can SUs meet their QoS needs through identifying an optimal amount of backup spectrum (note that this spectrum is only marked for future use, and not reserved right away), in case they are interrupted by the PU's return? Finally, To highlight the practical aspect of this research, we present a case study of a practical problem that affects the medical community in the Boston area. We demonstrate how our approach can benefit the wireless medical telemetry service (WMTS), through measurements and using stored traces of spectrum usage from extensive spectrum surveys conducted at hospital sites.

The motivation of our work stems from a need to have a rigorous mathematical framework that considers the spectrum usage activity of the PUs, the latency and bandwidth requirement of the SUs, and returns an efficient spectrum allocation scheme. At a high level, we devise separate analytical formulations for *streaming* and *non-streaming* categories of applications as in other works such as [3], [4]. Within each of these categories, depending on the packet arrival rate at a given node (γ), the required bit rate (R), the link-layer successful packet transfer time (ℓ), and the packet length (L) chosen for the application, further grouping is possible. Before initiating transmission, the SU informs the above four parameters to the controlling BS, which undertakes a centralized resource allocation and informs each SU which PU channel(s) must be used to satisfy its required QoS. To meet the rate requirement R, we utilize channel aggregation and find the best set of contiguous PU channels, such that the cumulative bandwidth suffices for that SU. To satisfy the delay requirement ℓ, a critical concern for both categories of applications, we use *a priori* statistical knowledge of PU activity in terms of inter-arrival and active (or "on" time). In our approach, each such *group*

Manuscript received June 24, 2013; revised December 11, 2013; accepted April 10, 2014. Date of publication April 24, 2014; date of current version July 8, 2014. The associate editor coordinating the review of this paper and approving it for publication was M. Bennis.

The authors are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: doost@ece.neu.edu; naderi@ece.neu.edu; krc@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2014.2319307

of SUs (constructed on the basis of similarity in the above metrics) is assigned a precisely calculated number of *backup* channels. The selection of the number of backup channels is an important factor in our design—too few may lead to long-term service disruptions, while too many make inefficient use of the spectrum. Moreover, our model formulation has a key difference from classical backup channel estimation for cellular networks: here, the backup channels are only marked for possible use on case the current channel of the SU is taken away. Thus, these channels can be used by the PU at any time on a higher priority-basis, or they can be the fall back option for other SUs of the group who had to give back their initially assigned spectrum to a returning PU in its default channel.

The specific contributions of our work are as follows:

- We cast the problem of active and backup channel allocation for the streaming category as a 2-dimensional Markov process having the properties of quasi-birth-and-death (QBD) that ensure an appropriate bounded channel access delay for *streaming CR nodes*. To the best of our knowledge this is the first application of QBD for spectrum allocation for time bounded real-time applications.
- A different Markov process based on an adapted version of the original 802.11 distributed coordinated function (DCF) model [5] is devised for *non-streaming CR nodes*. Unlike the classical case where each node has similar channel access rights, in our work, the PUs have priority of using the channel. This changes the probability distributions of the state transitions, and impacts the critical functions of backoff, countdown, among others.
- Our method focuses on minimizing the overall spectrum usage, and serving as many SUs as possible with their QoS requirements met. It factors in a rich set of possibilities including SUs with different QoS requirements, PUs with varying activity patterns, and flexible channel boundaries.
- We demonstrate a new application area for CR, that requires efficient utilization of the WMTS bands in a medical environment composed of heterogeneous devices with different bandwidth and QoS requirements. This band covers the ranges 608–614 MHz (digital television or DTV channel 37), 1395–1400 MHz (lower-L band), and 1427–1432 MHz (upper-L band), and is susceptible to either interference or high priority traffic from DTV, utility providers and government installations. The medical telemetry applications must share these bands on a low-priority basis, and thus, play the role of SUs.

The rest of this paper is organized as follows: We describe the related work in Section II. Section III explains in detail the analysis framework, and the channel allocation scheme. Section IV describes a case study for medical telemetry using our approach. Section V provides a comprehensive simulation study, and finally, Section VI concludes the paper.

II. Related Work

While spectrum sensing has received a lot of attention over the past several years, the problem of QoS provisioning for the SUs merits more research, as ultimately, applications will drive future adoption of CR technology. For continuous traffic generating SUs, and with exact knowledge of the channel gains for the entire licensed spectrum and the PU activity in them, a Markovian framework is presented in [6] that derives the queueing delay performance of the SU packets. This approach assigns SUs to channels on which they experience the best channel gain, after which a PDF for packet delay for each SU is derived. In [7], [8], call drop and call blocking probability in a secondary network is studied via a Markov chain analysis and based on exponential inter-arrival time of PUs. Along similar lines, [9]–[11] formulate different call admission strategies for ensuring that QoS, expressed in terms of call dropping and blocking rate, is achieved. Through a Markovian analysis, they minimize the dropping rate while attempting to meet a user-defined call blocking rate constraint. For non-continuous traffic, the transmission delay and packet drop performance under unslotted CSMA/CA is analyzed in [12]. However, all these works assume that there is identical statistical behavior of the PUs on all channels, which does not reflect practical observations.

TABLE I
List of Channel Allocation With QoS Provisioning Works in the Literature

Reference	Allocation Objective	Target QoS metric	PU heterogeneity	SU heterogeneity
[9]–[11]	Call Droping Prob.	Call Blocking Prob	No	No
[13]	Fairness in Airtime Share	Call Blocking Prob.	No	Yes
[4], [14]	Spectrum Utilization	Call Blocking Prob.	No	Yes
Our work	Spectrum Utilization	Blocking delay	Yes	Yes

In this paper we consider a general case with heterogeneous PU activity in the licensed spectrum. Our model is further complicated by considering different types of traffic—streaming and non-streaming, each of which may have further fine-grained requirements of latency, bandwidth, among other QoS features. Our analytical work involves devising two different Markov-chain based frameworks for these two traffic types. In the streaming traffic, the average wait time for a streaming node to access a free licensed channel is derived and matched with the QoS-specified delay that serves as the permissible upper bound. For the non-streaming traffic, the average wait time of a single packet transmission is derived under the assumption of CSMA/CA between multiple contending SUs. In summary, the heterogeneous spectrum-usage and channel definitions, and the inclusion of a spectrum allocation algorithm that ensures that the user-specified QoS needs are met, differentiates our work from the existing state-of-the-art. Table I lists other works in the literature that propose joint QoS provisioning and channel allocation with entries that marks their respective target QoS metrics, the assumptions made on the heterogeneity of PU channels and also the heterogeneity of SU demands.

III. System Model and Problem Formulation

A. Network Model

In this section we describe the various network entities, details of the QoS assumptions, and the overview of our approach.

- **Central BS**: The BS accepts new data transmission requests made by the SUs. Each node submits a QoS vector in the form (γ, R, ℓ, L) to the BS. In this vector, γ indicates the packet arrival rate of the nodes, assuming a Poisson

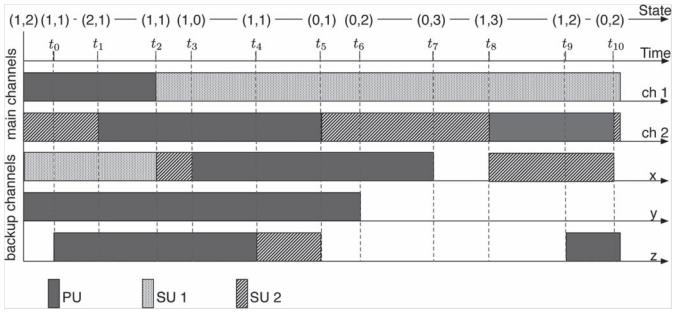


Fig. 1. An example of channel assignment for 2 streaming nodes with 3 backup channels. Random PU arrivals and departures trigger SU channel movements and the corresponding Markov process state transition.

arrival distribution. For the streaming case, γ is set to ∞. Also R is the required rate, ℓ is the packet delay constraint and L is the length of each packet the node is transmitting. The BS groups these requests based on their QoS requirements. For e.g., the non-streaming nodes are separated from their streaming counterparts. If nodes vary in any one of these parameters, i.e., different arrival rates, latencies, or packet lengths, they form their own individual subset where *all* these requirements are the same for the nodes of that particular subset. We assume that PU arrivals and departures are accurately detected at the BS and communicated to the SUs. The BS has the knowledge of statistical PU arrival rates.

- **Streaming SUs**: Each SU for a given streaming group is assigned its own channel since this channel will be continuously utilized for streaming data. Hence, such a node will not have to contend with any other SU nodes for channel access. Despite this, the operation of the SU could be disrupted when the PU takes over the channel. Therefore, to guarantee the continuous operation of a group of streaming SUs on their channels, a set of backup channels is identified for the group. These backup channels will be used by SUs when their own default channel is occupied by the PU. Our approach identifies the number of backup channels that needs to be assigned per group in such a way that the average packet queueing delay for nodes of that group is below that of the threshold specified in the QoS vector. Note that these backup channels are also contained within the licensed spectrum, and consequently, may also be claimed by PUs. Hence, marking them of future use does not guarantee their availability. We begin the analytical formulation under the limiting assumption of fixed PU arrival rates λ and μ for the default and the backup channels, respectively, in Section III-B. We relax this assumption in Section III-D, for the case of heterogeneous PU arrival and departure rates.
- **Non-streaming SUs**: For non-streaming groups, multiple SUs, provided they have the same QoS requirements, are assigned to a single channel and allowed to contend for the spectrum, provided long-term delay and packet transmission rate threshold are met. For this, the number of nodes assigned to a given channel, and using classical CSMA/CA at the link layer, must be carefully decided, as we show in Section III-C.

B. Delay Analysis for Streaming Type Allocation

For a network composed of M streaming nodes, we need to identify the lowest number of backup channels N, such that the affected streaming node is ensured continuous use of the spectrum with average delay below or equal to ℓ. In determining N, an underestimation results in an increase in the blocking probability for the streaming nodes, while an overestimation results in inefficient use of the spectrum. If a node's main channel is occupied, and all N channels are busy, then it must await in a queue for either one of the backup channels or its own original channel to become available. Our approach involves modeling this system as a queueing problem, where the M nodes are *customers* that randomly arrive at a queue serviced by N backup channels as *servers*. Considering the delay constraint ℓ of the QoS vector, the mean queueing time must be kept below this threshold.

Since the number of available backup channels (here, servers in the queueing problem) is varying owing to the random arrival and departure of PUs on these channels, we model the problem with a two-dimensional continuous-time Markov process with sate space $S = \{(m(t), n(t)) : 0 \leq m(t) \leq M, 0 \leq n(t) \leq N\}$, where $m(t)$ is the number of nodes out of their original channel due to PU presence and either seeking a backup channel or operating on one (i.e., nodes that had to vacate their earlier default channel), and $n(t)$ is the number of backup channels not occupied by PUs (they may, however, be used by SU nodes) at an arbitrary time t. Fig. 1, shows an example, with a network of $M = 2$ SUs assigned with 2 default channels—channel 1 and 2 shown in the top half of the figure. There are 3 backup channels-channels x, y, and z shown in the bottom half. The notations in the parenthesis denote the current state of the system. For e.g., in slot $[t_0, t_1]$, the channel 1 is used by the PU (thus displacing SU 1), while SU 2 still has access to channel 2. Since one SU is out of its default channel, $m(0) = 1$. Looking at the backup channels, we find that SU 1 is using channel x, while channels y and z are occupied by other PUs. Hence, the number of backup channels not used by PUs is $n(0) = 1$. Thus, the state at time t_0 is defined as (1,1).

Upon arrival of a PU on its default channel the affected SU will switch to any available backup channel (see instants t_1, t_8 in Fig. 1). On the other hand, if the default channel becomes available (even if the SU is operating on a perfectly fine backup channel) it immediately resumes using it. For e.g., in instant t_5, when the PU vacates its default channel 2, the SU 2 leaves the backup channel z and returns back to channel 2. If the SU is dislodged from its own default channel and no backup channel is available, it will wait in the queue until one of them is available. For e.g., at time t_1, SU 2 vacates its default channel 2, and all other channels are either occupied by the PUs (default channels 1 and 2, and backup channels y and z) or by other SUs (backup channel x). Thus, it must now enter into a wait state, behind any already existing SUs in the wait queue.

Assuming a fixed Poisson PU arrival rate λ and departure rate λ_{on} for all M default channels, and μ and μ_{on} as the corresponding rates for all N backup channels, our two-dimensional Markov process is shown in Fig. 2.

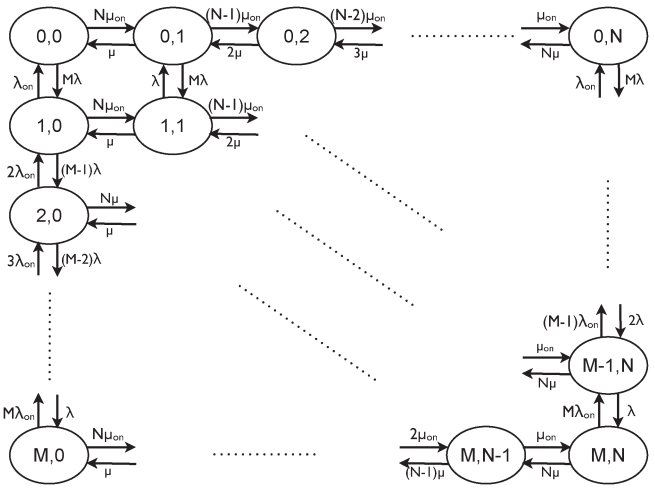


Fig. 2. Finite quasi-birth-death process representing M main channels with PU arrival rate λ and PU "on" rate λ_{on}, and N backup channels with PU arrival rate μ and PU "on" rate with μ_{on}.

Starting from the state (0,0), any one (out of N) backup channels can become available for use if the PU exits, which occurs with the rate $N\mu_{on}$. Likewise, the arrival of a single PU (with the rate μ) will result in the transition to state (0,0) from state (0,1). Similarly, the chain can be extended to the terminal state $(0, N)$ in the horizontal plane, and the state $(M, 0)$ in the vertical plane. In general:

- the transition rate to state $(i + 1, j)$, i.e., when a streaming node requests a backup channel, is $(M - i)\lambda$.
- the transition rate to state $(i - 1, j)$, i.e., when a node currently served by a backup channel, or in the queue waiting for one, reclaims its own default channel due to PU leaving, is $i\lambda_{on}$.
- the transition rate to state $(i, j - 1)$, i.e., a backup channel becoming free for use, is $j\mu$.
- the transition rate to state $(i, j + 1)$ or a backup channel becoming occupied by the PU, is $(N - j)\mu_{on}$.

The Markov process of Fig. 2 has all the properties of a quasi-birth-and-death (QBD) process [15] where it has M levels and N phases at each level. It is straightforward to show that the transition matrix Q of the QBD process in Fig. 2 has the following form:

$$Q = \begin{pmatrix} A_1^{(0)} & A_0^{(0)} & \mathbf{0} & \cdots & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & \mathbf{0} & \cdots & \\ & \ddots & \ddots & \ddots & & \\ \mathbf{0} & \cdots & A_2^{(k)} & A_1^{(k)} & A_0^{(k)} & \cdots \\ & & \ddots & \ddots & \ddots & \\ \mathbf{0} & \cdots & & & A_2^{(M)} & A_1^{(M)} \end{pmatrix}$$

where each of the matrices $A_0^{(k)}$, $A_1^{(k)}$ and $A_2^{(k)}$ are given as

$$\begin{aligned} A_0^{(k)} &= (M-k)\lambda I_{(N+1)}, \\ A_1^{(k)} &= T + \tilde{A}_1^{(k)}, \\ \tilde{A}_1^{(k)} &= -((M-k)\lambda + k\lambda_{on})\, I_{(N+1)}, \\ A_2^{(k)} &= k\lambda_{on} I_{(N+1)}. \end{aligned}$$

T is the transition matrix specific to each level of the two dimensional Markov process:

$$T = \begin{pmatrix} -N\mu_{on} & N\mu_{on} & & & \\ \ddots & \ddots & & & \\ & k\mu & -k\mu-(N-k)\mu_{on} & (N-k)\mu_{on} & \\ & & \ddots & \ddots & \ddots \\ & & & N\mu & -N\mu \end{pmatrix}$$

and $I_{(N+1)}$ is the identity matrix of size $(\mathrm{N}+1)$-by-$(\mathrm{N}+1)$. Matrix Q represents an inhomogeneous (level-dependent) finite QBD process since arrivals to and departures from level k depend on k (in this case it is a function of k). Since the properties of this type of Markov process are relatively unexplored because of its generality [16], no closed form expression for the steady state probability vector exists. In other words, to obtain the steady state probability vector Π, the set of equations given by $\Pi Q = 0$ and normalization condition $\sum_{i,j} \pi_{i,j} = 1$ needs to be solved. However, several numerical algorithms are proposed to accelerate the computation of stationary distribution of the process [17] that can be used for faster convergence. The irreducibility of matrix Q is trivially deduced, and therefore, a steady-state probability vector Π exists. The set of equations given by $\Pi Q = 0$ and normalization condition $\sum_{i,j} \pi_{i,j} = 1$ for our QBD process can be expanded as the following:

$$\begin{aligned} &\Pi_0 T + \Pi_0 \tilde{A}_1^{(0)} + \Pi_1 A_2^{(1)} = 0, \\ &\Pi_0 A_0^{(0)} + \Pi_1 T + \Pi_1 \tilde{A}_1^{(1)} + \Pi_2 A_2^{(2)} = 0, \\ &\Pi_{k-1} A_0^{(k-1)} + \Pi_k T + \Pi_k \tilde{A}_k^{(k)} + \Pi_{k+1} A_2^{(k+1)} = 0, \\ &k = 1 \ldots M-1, \\ &\Pi_{M-1} A_0^{(M-1)} + \Pi_M T + \Pi_M \tilde{A}_1^{(M)} = 0, \\ &\sum_{i=0}^{M} \sum_{j=0}^{N} \pi_{i,j} = 1, \\ &\mathbf{\Pi_i} = (\pi_{i,0}, \pi_{i,1}, \ldots, \pi_{i,N}). \end{aligned} \tag{1}$$

By solving this set of equations, the average length of the queue can be simply calculated as the following:

$$L = \sum_{i>j} (i-j)\pi_{i,j}. \tag{2}$$

Equation (2) shows that there will be nodes waiting in the queue (hence contributing to the queue length) when the number of nodes forced out of their original channels is more than the number of available backup channels.

Subsequently, by using Little's Law, the average waiting time in the queue is obtained:

$$W = \frac{L}{\lambda_e} = \frac{L}{\sum_{i,j}(M-i)\lambda\pi_{i,j} + \sum_{i\geq j} j\mu_{on}\pi_{i,j}}. \tag{3}$$

Above λ_e is the effective queue arrival rate, i.e., the rate at which nodes join the queue to get served by a backup channel. The rate expression has two parts as shown in the denominator in (3). The first part is the queue arrival rate from nodes requesting a backup channel right after PU arrival at their main channel (t_1 instance in Fig. 1). The second part concerns a node that is already operating on a backup channel. but has to leave it and wait its turn in the queue because the PU arrives

at this specific backup channel, and no other backup channel is immediately available (for e.g., the t_3 instance in Fig. 1, where SU 2 is operating on backup channel x, and this too gets claimed by a returning PU).

The above formulation is specific to the case when fixed PU arrival and departure rates are assumed for each of the set of default and backup channels. For a general case of unequal PU arrival and departure rates for all of the main and backup channels, any mathematical analysis become intractable as for arbitrary values of M and N a representation of any Markov process can not be easily conceived. This is due to exponential increase in the number of states as M and N increase. In fact, it is proved in [18] that the general problem of job-server queueing with any number of servers more than two, when servers go out of service (as in our case) with heterogeneous rates is mathematically intractable. In Section III-D, by using the same formulation, we solve the general problem of heterogeneous arrival and departure rate by means of an approximation algorithm.

C. Delay Analysis for Non-Streaming Type Allocations

We let non-streaming nodes share a channel, as they do not transmit continuously. Though this improves spectrum utilization efficiency (compared to issuing each node a dedicated but seldom accessed channel), the need to ensure that packing excessive number of these SUs within a channel does not lower the performance below their QoS threshold. Formally, given a maximum possible transmission rate R for a given PU channel, with the PU arrival rate of λ, the maximum number of nodes K that can be serviced in that channel needs to be calculated. We assume that all these nodes have the common QoS vector (γ, R, ℓ, L).

- *IEEE 802.11 DCF preliminaries*: In the 802.11 DCF model, a CSMA/CA mechanism is employed. Each node with a packet to transmit contend with other nodes for channel access during a contention window. The contention period starts when the nodes senses the channel Idle for duration of Distributed Inter-Frame Spacing (DIFS) as defined in IEEE 802.11 standard. Then, a back off timer is initially set with a randomly chosen time in the range $[0, W]$ and starts counting down. At the end of each slot time of the backoff timer, the node senses the channel and if it detects a transmission on the channel, it freezes the timer until the transmission is over. When the timer hits zero, the node attempts a packet transmission which may be successful (with an ACK is received before $ACKTimeout$ time) or result in a collision with other transmissions on the channel (No ACK). If the transmission is succcessful, the receiver of the packet tries sending an ACK message to the transmitter node after a duration of Short Inter-Frame Spacing (SIFS), which is the time needed for a node's radio to switch from receiving mode to transmission mode. In case of a collision, the node doubles the range of backoff window, i.e. $[0, W_1]$ where $W_1 = 2W$ and repeats the contention process. This process will go on in case of further collisions each time doubling the range up to mth stage so that at stage i, $W_i = 2^i W$. After the mth stage, the contention window range will not increase further. The contention process continue until the packet is successfully transmitted.
- *Our revisions to the 802.11 DCF model*: We revise the Markov processes described in [5], [19] to capture the behavior of such a network by incorporating the presence of the PU in the network. Therefore, collisions in the network are caused either with PU or other SUs attempting transmission at the same time. The fundamental assumption in [5] is that the probability of a packet encountering a collision p, after a transmission attempt is fixed over time and independent of the transmitting node. Even with the presence of PU in our revised model, this assumption holds true, for the SUs, since PU transmission equally affects the activity of all nodes, e.g. freezing their back-off counter. This does not violate the assumption that the probability of transmission attempt τ and probability of collision p for each node is independent of that of others. However note that the dependence of any SU transmission attempt on the PU transmission, of course, remains. Based on derivations given in [19], τ, the probability that an arbitrary node starts transmission at a randomly chosen slot is:

$$\tau = \frac{2(1-2p)q}{q[(W+1)(1-2p)+Wp(1-(2p)^m)]+2(1-q)(1-p)(1-2p)} \tag{4}$$

where q is the probability of having packets to transmit. In the following, we provide our extensions to the model in [19] with revised formulation of the variables used in (4) taking the PU activity into account in network of K SUs. The probability of collision for any node either with $(K-1)$ other nodes or the PU is given as follows:

$$p = 1 - (1-\tau)^{K-1}(1-P_{on}). \tag{5}$$

P_{on} is the probability of the arrival of PU with Poisson arrival rate λ during a time necessary for a node transmission to be successful (T_s) and is simply given as:

$$P_{on} = 1 - e^{-\lambda T_s}. \tag{6}$$

Probability q, which is the probability that there is at least one packet to be transmitted at each slot, can be approximated with the following relation as a function of γ, specified by the QoS vector [19]:

$$q = 1 - e^{-\gamma E_s}. \tag{7}$$

E_s is the average slot time spent by the channel in any state including an idle-channel slot, successful transmission (T_s), SU collision (T_l), or PU-SU collision (T_I), which results in interference to PU. The idle state includes both fixed slot time with length σ during which the nodes decrement their backoff counter and also the times of PU appearance on the channel which freezes the back off timer for all SUs. Now, E_s is given as:

$$E_s = (1-e^{-\lambda\sigma})\left(\frac{1}{\lambda_{on}} + \sigma\right) + e^{-\lambda\sigma}\{(1-P_t).\sigma + P_t P_s.T_s + P_t P_I.T_I + P_t(1-P_s-P_I).T_l\}. \tag{8}$$

In the above, the first term indicates a fixed backoff slot during which a PU appears on the channels. This state occurs with probability $1-e^{-\lambda\sigma}$ and lasts for $((1/\lambda_{on})+\sigma)$ in average. In case of no PU activity. In case of PU appearance, either an idle (no activity on the channel) or SU transmission attempt can occurs where the latter takes only σ and the latter will take T_s, T_I, and T_l in case of successful transmission, PU-SU collision and SU collision respectively. Each of these occurs with the following probabilities: P_t is the probability that at least one node attempts a transmission at an arbitrary slot, P_s is the conditional probability that a given packet transmission on the channel is successful, and P_I is the probability that any transmission attempt of SUs collide with the PU. Clearly, the complement of sum of P_s and P_I will indicate the probability of SU collision. These probabilities are given as

$$P_t = 1-(1-\tau)^K, \tag{9}$$

$$P_s = \frac{K\tau(1-\tau)^{K-1}(1-P_{on})}{P_t}, \tag{10}$$

$$P_I = P_{on}, \tag{11}$$

$$T_s = \frac{L}{R} + SIFS + \frac{L_{ack}}{R} + DIFS, \tag{12}$$

$$T_I = \frac{L}{2R} + \frac{1}{\lambda_{on}}, \tag{13}$$

$$T_l = L + AckTimeout. \tag{14}$$

Here, L is the packet length, R is channel rate and $AckTimeout$ is the permissible timeout duration for the acknowledgement to arrive. For fixed K (number of contending SUs on the channel) and known W (initial backoff length) and m (number of backoff stages), values of τ, p, and q can be obtained numerically by solving the nonlinear system of equations comprising (4)–(14). Finally, the mean packet transmission delay is derived as below [20]

$$\Delta = \sum_{i=0}^{\infty} \frac{2^{\min(i,m)}W-1}{2} p^i E_s + \sum_{i=1}^{\infty} i p^i (1-p) T_l + T_s.$$

This can be simplified to

$$\Delta = E_s\left[\frac{W}{2}\frac{(2p)^m-1}{2p-1} + \frac{W}{2}\frac{(p)^m}{1-p} - \frac{2}{2-p} - 1\right] + \frac{p}{1-p}T_l + T_s. \tag{15}$$

Using the above derivations, we find the solution to the problem of how many nodes with a given QoS vector may be assigned to a single channel. We formulate a simple optimization problem that maximizes K for a given channel (and repeated over multiple channels) such that $\Delta \le \ell$, using the expression from (15).

To constrain the amount of interference to PU, we assume the probability of SU-PU collision can not exceed a threshold P_{th}, or we need $P_I \le P_{th}$. Using (11) and (6), this inequality can be simplified as

$$T_s \le \frac{1}{\lambda}\log\frac{1}{1-P_{th}}. \tag{16}$$

Using (12) and assuming fixed L_{ack} for all nodes from any QoS class, we obtain a constraint on the packet length L of SUs operating on the channel:

$$L \le R\left(\frac{1}{\lambda}\log\frac{1}{1-P_{th}} - DIFS - \frac{L_{ack}}{R} - SIFS\right). \tag{17}$$

TABLE II
LIST OF NOTATIONS USED IN FREQUENCY ALLOCATION ALGORITHM

Symbol	Definition
s	Number of PU channels required for a request in the range $s_{min}, ..., s_{max}$
$\mathcal{H}_s$	PU arrival rate histogram of all channels of length s
Ψ	Set of current channel allocation requests by nodes
Ψ^c_{str}	Set of all streaming requests for channels of length c
Ψ^c_{nstr}	Set of all non-streaming requests for channels of length c
Ψ^c	Set of all requests of for channels of length c
$\ell(\Psi)$	Set of delay components of the QoS vector for all requests in set Ψ

The above inequality will set a constraint on the type of QoS classes that be allocated to certain PU channels. We will use as a metric in our channel allocation algorithm in the next section.

D. Frequency Allocation Algorithm

In the previous sections, we obtained the formulations for deriving the analytical QoS, given a set of licensed channels that the SUs use. In this section, we describe a greedy heuristic algorithm that is used to allocate channels, assuming heterogeneous PU behavior in them, to the SUs such that their QoS thresholds are met. The general class of such resource allocation problems are NP-hard [21], and hence, we seek a low-complexity heuristic approach.

At first, we present Algorithm 1 that allocates channels to streaming nodes using the analysis of Section III-B. Then, we propose Algorithm 2 for non-streaming nodes based on the analysis given in Section III-C. Finally, Algorithm 3 uses Algorithms 1 and 2 to allocate available channels to input nodes of various types and QoS requirements. Table II lists the notations used in the presented algorithms.

Algorithm 1 Allocating spectrum to streaming nodes with required delay ℓ and s PU channels

```
1: function allocate_streaming (m, s, ℓ, H_s)
2:     while m > 0 do
3:         set M = min(|b_1|, m), λ = λ^{b_1}, λ_on = λ_on^{b_1}.
4:         while M > 0 do
5:             for each b_i in H_s in ascending order do
6:                 N_i = |b_i| (|b_i| − M if i = 1).
7:                 μ = μ^{b_i}, μ_on = μ_on^{b_i}.
8:                 if W(M, N_i, λ, λ_on, μ, μ_on) < ℓ then
9:                     Search for smallest subset of b_i with size
                       N_opt that W(M, N_opt, λ, λ_on, μ, μ_on) < ℓ.
10:                    Allocate M ch. in b_1 and N_opt ch. in b_i
                       as backup to M out of m nodes.
11:                    m = m − M.
12:                    update H_s.
13:                    Break
14:                else
15:                    Increment i and Continue search.
16:                end if
17:            end for
18:            if no set of ch. were found so that D < ℓ then
19:                decrement M.
20:            end if
21:        end while
22:    end while
23: end function
```

Algorithm 2 Allocating spectrum to non-streaming nodes with required delay ℓ and s PU channels

1: **function** ALLOCATE_NONSTREAMING $(n, s, \ell, L, \mathcal{H}_s)$
2: $i = 1$.
3: **while** $n > 0$ **do**
4: $\lambda = \lambda^{c_1}$, $\lambda_{on} = \lambda_{on}^{c_1}$.
5: **if** $\mathcal{N}(\lambda, \lambda_{on}, \ell, L) > 0$ **then**
6: $l = \min(\mathcal{N}(\lambda, \lambda_{on}, \ell, L), n)$
7: assign l nodes of n to c_i.
8: $n = n - l$.
9: update $\mathcal{H}^s$.
10: **end if**
11: Increment i.
12: **end while**
13: **end function**

1) Algorithm 1 for Streaming Nodes: To allocate channels to m streaming nodes in the same QoS class, we utilize the analytical model given in Section III-B to find m main channels of length s and a minimal number of backup channels so that their delay requirement ℓ is met. However, that model demands channels with identical PU arrival rate λ and PU departure rate λ_{on} for the main and also for any potential set of backup channels. Therefore to deal with spectrum of channels with heterogeneous arrival and departure rates λ and λ_{on}, we construct a 2-dimensional *histogram* out of all λ and λ_{on} values of the existing channels. The PUs that are placed within a histogram bin of width w_{b1} and w_{b2} (for the given range of λ and λ_{on} values) are treated to have alike PU arrival and departure rates. In this regard, the arrival and departure rates of the channels in the same histogram bins are approximated to the upper bound arrival rate and lower bound departure rate of than bin (worst case for all the channels of the same bin).

Formally, let the minimum and maximum PU arrival rates in the set of all existing channels be $\lambda^{\min}$ and $\lambda^{\max}$ respectively. Also let $\lambda_{on}^{\min}$ and $\lambda_{on}^{\max}$ be the respective minimum and maximum PU departure rates. Fixing the bin widths at w_{b1} and w_{b2}, we get $B_1 = \lceil(\lambda^{\max} - \lambda^{\min})/w_{b1}\rceil$ and $B_1 = \lceil(\lambda_{on}^{\max} - \lambda_{on}^{\min})/w_{b2}\rceil$ each being the number of x-axis and y-axis bins of the histogram respectively.

Our proposed algorithm for streaming nodes works on the 2D-Histogram $\mathcal{H}_c$ as its input. It determines the systemic order in which groups of channels are chosen from $\mathcal{H}_c$ as main channels and backup channels and then assigned to groups of streaming nodes. Given m streaming nodes, our algorithm performs with a greedy allocation strategy, starting with the bins indicating least PU activity (best channels for SU operation) that have the smallest arrival rate λ and the largest departure rate λ_{on}. Thus, we start from the upper left corner of the histogram and sweep all the bins in the zig-zag order, shown in Fig. 3(a).

As some bins are empty and do not contain any channels with matching λ and λ_{on} ranges, the algorithm finds the first non-empty bin in the order of bins determined in Fig. 3(a). At the beginning of each iteration, the algorithm starts with the first non-empty bin, and without loss of generality, we refer to such a bin as b_1 and the number of channels in it as $|b_1|$. At start,

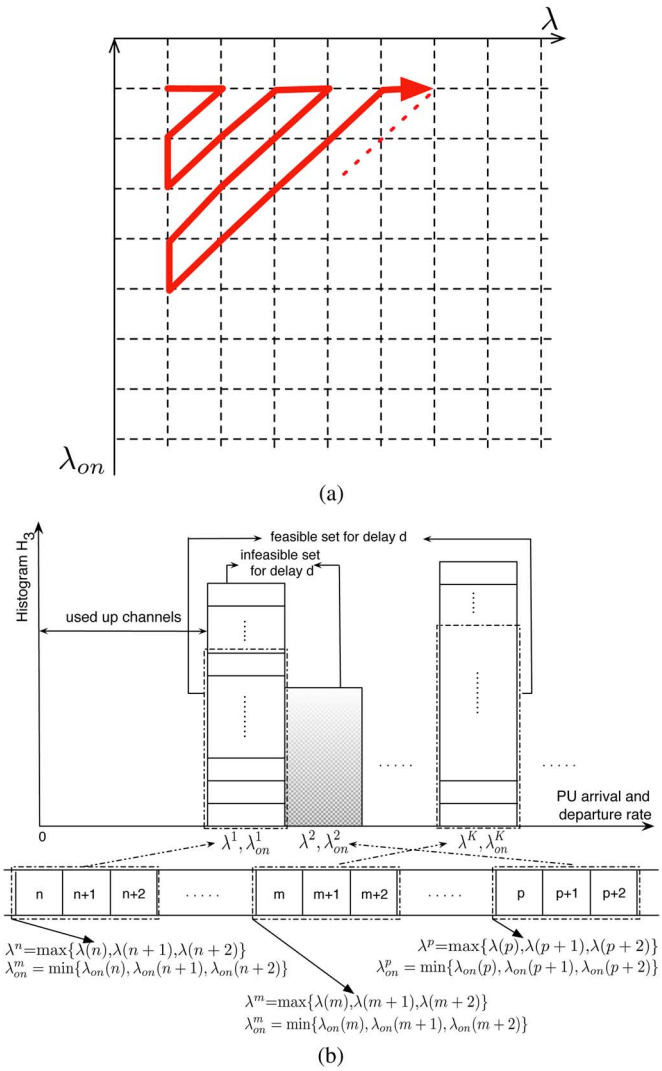


Fig. 3. (a) Traversing order of the 2D-Histogram during allocation. (b) A histogram H_3 shown with PU arrival and departure shown in one dimension. PU channels are aggregated with $s = 3$ to form the histogram.

Algorithm 1 (line 1) considers the allocation of M channels as main channels, namely the minimum of m, which is also the number of given nodes and number of channels in bin b_1, or $M = \min(m, |b_1|)$. It then searches for the appropriate set of backup channels on the histogram that satisfies the delay requirement ℓ for M default channels, based on the analytical formulation of Section III-B.

To find the best choice of a set of backup channels for the selected M main channels, we iterate over the remaining bins of the histogram starting from b_1 onwards (lines 5–17). At the ith iteration, we set $N_i = |b_i|$. We also refer to (3) as $D = W(M, N_i, \lambda, \lambda_{on}, \mu, \mu_{on})$, which returns the mean delay for M default channels with arrival and departure rate (λ, λ_{on}), and N_i backup channels with representative arrival and departure rates (μ, μ_{on}) of bin b_i. If $D > \ell$, we continue the iteration over next bin b_{i+1}. If $D \leq \ell$. Then, bin b_i contains a potential set of backup channels for the M main channels in b_1. In this case, the smallest subset of channels in b_i that satisfies the delay requirements for M nodes needs to be found. For this purpose, a binary search is performed to find an $N_{opt} \leq N_i$. When a set of feasible and optimized main and backup channels with mean

delay $D \leq \ell$ is found, the histogram is updated to exclude these channels as they are allocated. The same process is repeated for any remaining set of $m - M$ given nodes. If at the end of iterations, no bin with potential set of backup channels is found, we decrement M and repeat the iterations until a set of backup channel is found for M main channels in b_1.

As the complexity of algorithm 1 is dependent of distribution and range of the histogram, the QoS requirement of SUs and specific results of the analytical model, it is challenging to compute a general average complexity for it. However, the worst case complexity of algorithm 1 can be obtained as follows: In the worst case, all the histogram bins are searched for a feasible solution and this feasible solution is always found in the last bin. Let the total number of channels be P and the total number of bins be $H = B_1 B_2$, as B_1 and B_2 defined earlier. Also the complexity of solving set of n linear equations using Gaussian elimination is $O(n^3)$ [22], therefore to solve (1) for i main and j backup channels, we have a complexity of $(i+1)^3(j+1)^3$. Considering a uniform histogram (again the worst case in search for channels), the number of channels in each bin would be:

$$P_1 = P_2 = \cdots = P_H = \frac{P}{H}. \tag{18}$$

Then, the complexity in the worst case can be easily shown to be $O(P^7)$:

$$\sum_{i=1}^{P_1} \sum_{j=P_1,\ldots,P_H} (i+1)^3 (j+1)^3 = \frac{1}{4}(H-1)\left(\frac{P}{H}\right)^4 \left(\left(\frac{P}{H}\right)^3 + 6\left(\frac{P}{H}\right)^2 + 13\left(\frac{P}{H}\right)^2 + 12\right).$$

2) Algorithm 2 for Non-Streaming Nodes: Given n non-streaming nodes, similar to algorithm 1, a 2D-histogram of nodes is used and channels are picked by traversing the bins of the histogram as shown in Fig. 3(a). Based on the results of Section III-C, one channel can be allocated to several nodes. Therefore, at each bin of the histogram b_i, each channel c is examined with (17) to check whether it can operate on than channel given the packet length L in its QoS vector and then the number of nodes it can accommodate so that their delay ℓ is satisfied is determined. Here, we refer to such quantity as $n = \mathcal{N}(\lambda, \lambda_{on}, \ell, L)$, which is obtained by doing a binary search over range $[0, n]$ and using (4)–(15) to find the maximum number of nodes n_{opt}, a channel c can be allocated to. We iterate over bins and the channels within bins as long as any nodes are left and for each channel, we allocate as many nodes out of n as possible (see lines 4–14 of Algorithm 2). Then, the allocated channel is removed from the histogram and the remaining number of SUs is also updated. This procedure continues until all n SUs are accommodated.

The worst case complexity of the algorithm, since it iterates over all channels, is $O(PE \log n)$, P being the number of channels. Also $\log n$ is due to the binary search that is performed to find the maximum number of nodes that can allocated to each channel. This is multiplied by E, which is the complexity of solving the non-linear equation obtained from (4)–(15) for number of nodes during the binary search.

3) Algorithm 3 for General Nodes: Algorithm 3 uses the previous algorithms to allocate nodes from various QoS classes, namely various rate and delay requirements. For nodes that need higher bandwidth than a single PU channel, multiple contiguous PU channels can be aggregated to ensure sufficient available bandwidth is sufficient for the required rate. Let Ψ_s denote a set of nodes that need s aggregate channels to satisfy its data rate. Also Ψ_s^{nstr} is a subset of Ψ_s with only non-streaming type nodes. Equivalently, Ψ_s^{str} is for streaming type nodes. We also refer to $\ell(\Psi_s)$ as the set of delay values in the QoS vector of nodes in set Ψ_s.

Algorithm 3 Frequency allocation algorithm

1: **for** $s = s_{\max}$ to $s_{\min}$ **do**
2: Let Ψ_{str}^s and Ψ_{nstr}^s be sets of streaming and non-streaming nodes with s required bins.
3: $D_s = \ell(\Psi_{str}^s)$, $D_s = \{\ell_1, \ldots, \ell_p\}$, $\ell_1 < \cdots < \ell_p$.
4: $D_n = \ell(\Psi_{nstr}^s)$, $D_n = \{\ell'_1, \ldots, \ell'_q\}$, $\ell'_1 < \cdots < \ell'_q$.
5: Form $\mathcal{H}_s$ of the available channels of length s.
6: **for** $j = 1$ to q **do**
7: allocate_nonstreaming $(|\Psi_{nstr}^s(\ell'_j)|, s, \ell'_j, L, \mathcal{H}_s)$.
8: **end for**
9: **for** $i = 1$ to p **do**
10: allocate_streaming $(|\Psi_{str}^s(\ell_i)|, s, \ell_i, \mathcal{H}_s)$.
11: **end for**
12: **end for**

At first, requests are sorted in descending order based on the required number of PU channels s. This sorting will be independent of the streaming or non-streaming nature of the requests and their delay requirements. The channel allocation procedure begins with the maximum $s = s_{\max}$ and is repeated for descending values, down to minimum $s = s_{\min}$. Starting the allocation from $s_{\max}$ is aimed towards minimizing the number of unused fragments at the end of allocation procedure for efficient use of the spectrum [23], [24].

At each iteration over values of s, the set of delay values for each of the Ψ_s^{str} and Ψ_s^{nstr} are formed and sorted in ascending order. In the next steps, non-streaming and streaming requests of Ψ^s are processed respectively by algorithms 1 and 2 with ascending order of their delay requirement. In other words, the nodes with lower delay requirement are allocated first due to their stricter QoS. Also at each iteration, we choose to allocate non-streaming nodes prior to streaming ones due the additional constraint for the packet length L of non-streaming QoS given in (17).

To construct $\mathcal{H}_s$, the entire available spectrum must first be divided in channels of length s, and the availability of each channel, with respect to its allocation status, must be evaluated. An intuitive example of this concept is shown at the lower part of the Fig. 3(b) where $s = 3$. For each aggregate channel of s PU channels, the overall λ would be the maximum of all λ values of individual PU channels. Moreover, the overall λ_{on} for the channel will be the minimum of all λ_{on} values, as it indicates the rate at which the whole aggregate channel is vacated by the PU. The overall λ and λ_{on} for all aggregate channels will be used in forming the histogram $\mathcal{H}_s$.

IV. Wireless Medical Telemetry: A Case Study Using Real-World QoS Constraints

In this section, we study a real world scenario where the channel allocation framework of Section III is used to efficiently solve the problem of dynamic spectrum allocation in the WMTS bands. Although the FCC has allocated the WMTS bands for medical use, there are several issues that impair free access. First, there are no effective regulations protecting medical telemetry in channel 37 from the harmful interference caused by the power leakage from DTV transmissions in the adjacent channels 36 and 38. In fact, there are many documented cases of interruptions in hospital communication due to this DTV interference [25]. This adjacent channel interference effectively narrows the use of this channel (that represents almost 40% of all WMTS bandwidth). Given the critical nature of hospital communication, this breach must be immediately detected and corrective actions taken [26]. A second cause for concern is the non-uniform access rights in the L bands. Portions of these bands are shared by utility metering telemetry and government radar installations, which have priority or primary access right. Thus, the medical telemetry devices must be aware if these primary users (PUs) are present, and choose different portions of the spectrum, if indeed this is so.

In a medical environment composed of heterogeneous devices with different bandwidth, QoS, and access priority requirements, the problem of frequency allocation in these bands where interference from different sources is common is a challenging task. In the following, we provide the mapping of our analytical framework to this practical problem, and provide comprehensive simulation results in the subsequent section for this specific scenario.

A. Mapping of our QoS Framework for WMTS Bands

The algorithm of Section III-D can allocate small portions of the spectrum dynamically within the WMTS band to devices based on the type and duration of transmission, thereby increasing the potential for frequency re-use and the resulting channel capacity. The algorithm is particularly useful because in WMTS, the bandwidth for each device is relatively small (in the order of several KHz) and therefore the number of devices using the WMTS band could be relatively high (thousands). This necessitates a very efficient algorithm that can quickly and efficiently allocate channels to all these devices. Also, medical telemetry involves transmitting scalar data at set duty cycles, one-shot alarms, streaming information, among others, each with different bandwidth, latency requirements that must be jointly considered [27] which also fits the general streaming and non-streaming categories discussed in earlier sections. In deploying new nodes, the existing legacy medical telemetry transmissions that are not equipped with dynamic spectrum access, as well as PUs in the designated portions of the WMTS spectrum (i.e., the utility transmissions) must be protected, thereby necessitating a dynamic spectrum access-based solution.

Our analytical framework requires a statistical knowledge of the PU occupancy within the WMTS bands. In the next section we explain the methods we used to characterize the WMTS band through real experiments and extract the PU arrival and departure statistics in that band. Some preliminary measurements are described in our earlier work in [28].

B. Characterizing the WMTS Bands

To obtain a probabilistic model of channel occupancy on the WMTS channel 37 and the L bands, we performed a measurement study at several hospitals in Boston's Longwood area. We measured the spectrum usage on this channel using the USRP2 platform. The received power on every band was measured with a fine grained resolution, taking 1024-point FFT, i.e., obtaining a 6100 Hz resolution for each FFT bin (The resolution is appropriately chosen so that it fits the 6.25 KHz which is the commonly used medical telemetry bandwidth). Using the noise floor determination technique in [28], we extracted the active medical telemetry signals for each bin. Since these signals are temporally intermittent, we performed a statistical analysis at each bin on the inter-arrival and ON times of these signals. We then fit an exponential distribution function on these time samples. Therefore, the PU activity in each bin is captured with two λ and λ_{on} values, representing the arrival and departure rate of their respective exponential distributions. Fig. 4(a) and (b) show the PU inter-arrival and ON time statistics of three sample bins at all three bands within WMTS, namely DTV channel 37, lower-L band and upper-L band respectively. These statistics in Fig. 4(a) are fit with exponential distribution of mean 10.11, 18.75, and 10.82 with 95% confidence interval of [9.95,10.31], [18.45,19.35], and [10.57,11.11] respectively. Also in Fig. 4(b) exponential fit on measured PU ON times is undertaken with mean 2.29, 2.39 and 2.08 each with 95% confidence interval of [2.19,2.40], [2.39,2.81], and [1.93,2.24] respectively.

The exponentially-distributed PU activity assumption made in this section will be used for efficient channel allocation for the SUs in the network. In channel 37, these measurements represent all legacy medical telemetry activity, where devices are not equipped with dynamic spectrum access methods. In the L band, the observed channel activity jointly captures both the existing legacy medical telemetry and utility metering applications.

V. Performance Evaluation

In this section, we undertake a thorough simulation through ns-2 (packet level simulation for CSMA/CA based non-streaming nodes) as well as in MATLAB (for streaming nodes with continuous channel usage) to demonstrate the performance benefit in the WMTS band in terms of spectrum efficiency, as well as verify the theoretical findings on spectrum allocation from Section III for both streaming and non-streaming nodes. We also show the near-optimal spectral utilization efficiency of our greedy approximation approach in Section III-D. In these studies, we vary a metric called as the *load factor*, i.e., the number of nodes that have the same streaming requirements of latency and bandwidth. The PU activity statistics are acquired from real measurements described in Section IV from hospitals in the Boston area. Based on the activity model of the channels,

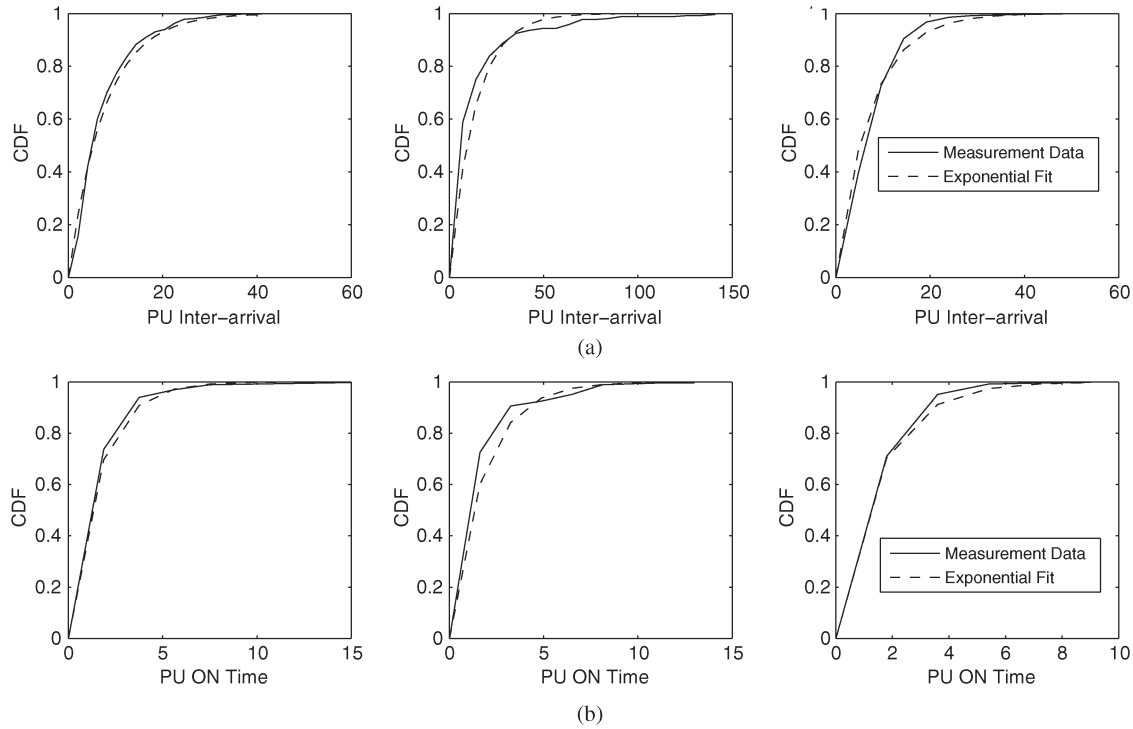


Fig. 4. (a) Exponential CDF fitting for PU inter-arrival time of three sample bins centered at 608.028 MHz, 1395.691 MHz, and, 1428.897 MHz. (b) Exponential CDF fitting for PU ON time of three sample bins centered at 608.028 MHz, 1395.691 MHz, and 1428.897 MHz.

TABLE III
APPLICATION SPECIFICATIONS IN A HOSPITAL CASE-STUDY [28]

Application Type	Size Kb/Packet	Avg Rate kb/s	Events per hr	Latency ms
Telemetry	2.6	12.8	Stream	500
Telemetry Diagnostic	5.1	25.6	Stream	500
Telemetry Alarm	1.0	0.1	10/h	500
Clinician Notifier	2.6	0.1	20/h	500
BCMA	0.4	0.1	30/h	500
Infusion Pump Status	1.0	1	Stream	500
Infusion Pump Alarm	1.0	0.1	1/h	500

we try to accommodate additional SU nodes in the empty portions of the band. We use the actual measured activity pattern on the WMTS channels, as a reference for the activity of PUs in our simulations. Also for the SU nodes, we consider 7 types of telemetry applications with specifications given in Table III as from the previous paper [28].

Similar to [28], a realistic wireless planning of a typical hospital with total area of 18580 m^2 is considered where the number of operating application nodes in each row in Table III is estimated by the values in vector (60,21,22,20,19,81,18). We use these values as a reference for our simulation study and choose random locations for each application node in a square area of 140 $\times$ 140 m. Also packet arrival events are created with Poisson distribution for each non-streaming application with the given rates of the above table, and the channel allocation algorithm is run for them in MATLAB. We then perform a packet level simulation in ns-2 to verify the validity of our allocation, based on the comparison of the delay from analytical and simulation findings. For the streaming case, since there is no channel contention (each node being allotted a dedicated channel), we verify the performance and analytical derivations through MATLAB.

TABLE IV
GROUPS OF MAIN AND BACKUP CHANNELS WITH VARIOUS PU ACTIVITY

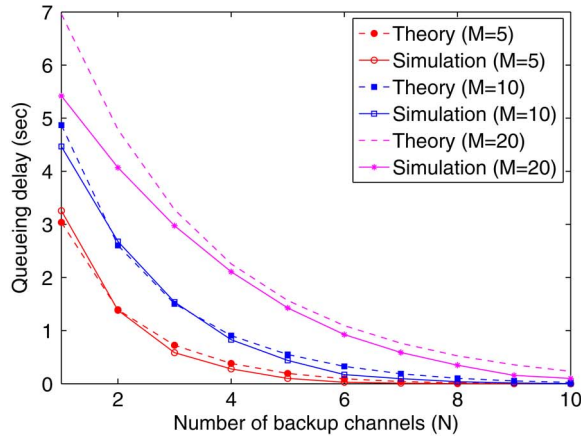
Group	PU activity	λ	λ_{on}	μ	μ_{on}
I	High	0.05	0.1	0.09	0.06
II	Medium	0.024	0.1	0.046	0.1
III	Low	0.005	1.0	0.011	1.0

A. Streaming Nodes

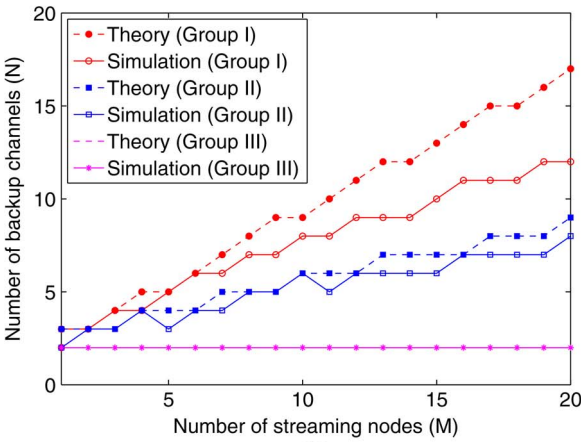
Three different sets of statistics (λ, λ_{on}, μ, μ_{on}) for the streaming and backup channels obtained from measurements in Section IV and used in the following discussion are shown in Table IV.

These three groups are examples of *high*, *medium*, and *low* usage channels by the PUs respectively. Intuitively, lower values of λ and μ indicate sparse arrivals of the PU. For λ_{on} and μ_{on}, lower values specify longer active duration for a given arrival event. Fig. 5(a) compares the theoretical and simulated queueing delay incurred for three sets (5, 10, 20) of streaming nodes, for the *medium* group. We observe that the simulation results very closely matches the analysis in Section III-B on all ranges of N. Fig. 5(b) shows the trend for the required number of backup channels to keep the queueing delay below 500 ms as the number of streaming channels are varied.

For the *high* group, where the number of backup channel increases almost linearly with the number of streaming channels, the simulation results indicate a slightly smaller number of backup channels than what the theoretical model predicts. This is largely due to the limited simulation time (3600 s) in the case of larger set of channels. However in the case of *medium* and specifically *low* usage groups, where the number of required backup channels is lower, the difference between theoretical



(a)



(b)

Fig. 5. (a) Incurred queueing delay both in theory and simulation for 5, 10, and 20 streaming channels while varying number of backup channels for the *medium* group, i.e., $\lambda = 0.024$, $\lambda_{on} = 0.1$, $\mu = 0.1$, $\mu_{on} = 0.1$. (b) Number of backup channels N vs. number of main channels M to keep the average queueing delay below 500 s shown for three PU activity groups of Table IV both in theory and simulation.

prediction and what simulation indicates is small. Overall, we find that the theoretical prediction used for channel allocation always keeps the average delay within the required bound.

To verify the spectral efficiency of the frequency allocation algorithm in Section III-D, we present a theoretical estimate of the channel allocation in a heterogeneous PU usage regime for the WMTS band in this paper. Assume that for a frequency range $\mathcal{F}$, the probability of any frequency f being available for secondary use is known and represented by function $P_{off}(f)$ on the domain $\mathcal{F}$. Then at any arbitrarily small frequency range df, the effective bandwidth is $P_{off}(f)\ df$. To achieve a minimum effective bandwidth b, we should have:

$$\int_{f_1}^{f_2} P_{off}(f)\ df \geq b. \tag{19}$$

To allocate bandwidth b in the most efficient manner, f_1 and f_2 in the above equation must be found in $\mathcal{F}$ in such a way that $f_2 - f_1$ is minimized. We use the same measurement statistics for WMTS presented before to get a discrete $P_{off}(f)$ over WMTS band. At each bin we have $P_{off} = 1 - (\lambda/\lambda_{on})$

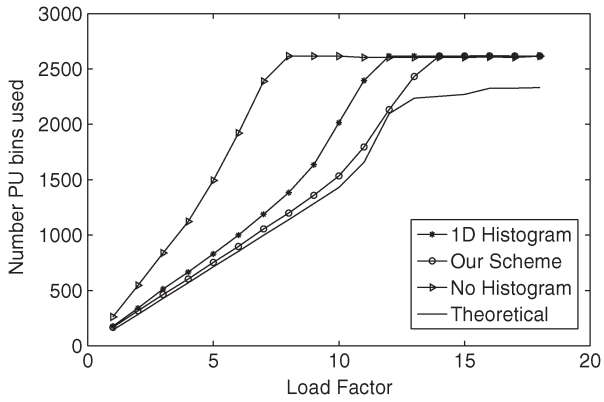


Fig. 6. Comparison of amount of allotted spectrum versus streaming request load resulted by algorithm 3, simplified of algorithm 3 based on 1-D histogramming method, simple method with no histogramming, and theoretical estimate.

as in [29]. We started with the vector (60,21,81) as the number of nodes for the streaming applications in Table III. We measured the number of PU bins allocated to all requests using Algorithm 3, and also the theoretical estimate for the number of PU channels used by the same set of requests using (19). We solved (19) numerically by choosing df as low as 300 Hz. We repeated the simulations by then scaling the number of nodes for each application given in vector (60,21,81), by a *load factor* (i.e., an integral multiplier). Fig. 6 shows the amount of allocated spectrum, in term of number of PU channels, of our algorithm compared to the theoretical estimate. Apart from the theoretical estimate and for the sake of comparison, we also compare the spectral efficiency of our allocation algorithm in Section III-D with its simplified version which uses only a 1D-histogram of PU arrival rate instead 2-D histogram of PU arrival and departure rates and also another simple algorithm which does not use any histogramming at all. The former is similar to the algorithm 1, but only uses a one-dimensional histogram of the PU arrival rate. The latter method merely uses the analysis of Section III-B to find a proper set of backup channels for each individual streaming node without any grouping or histograming. Fig. 6 shows how closely the spectral efficiency of matches of the theoretical estimate. The simplified histogramming algorithm just mentioned performs slightly below algorithm 3 in terms of efficiency. Also the simple algorithm with individual allocation of streaming nodes is least efficient of all.

B. Non-Streaming Nodes

In order to validate the model given in Section III-C, we performed simulations in the ns-2 network simulator with the environmental parameters set up to closely match the assumptions used in our model. We set the parameters of Table V as inputs to both to the ns-2 simulator for the packet-level simulation, and also for our MATLAB implementation that was used for the mathematical analysis of the model.

In the ns-2 simulation, the nodes contend over a single channel that is occasionally occupied by a PU with a known arrival rate. We observe that the average encountered delay of the nodes in our simulations closely follows that of the

TABLE V
SIMULATION PARAMETERS

Simulation Parameter	Default Value
MAC header	28 bytes
PHY header	16 bytes
ACK	14 bytes
Payload size	325 bytes
Slot time	20 μs
DIFS	50 μs
SIFS	10 μs
ACK timeout	500 μs
CW_{min}	32
CW_{max}	1024
Retry limit	7

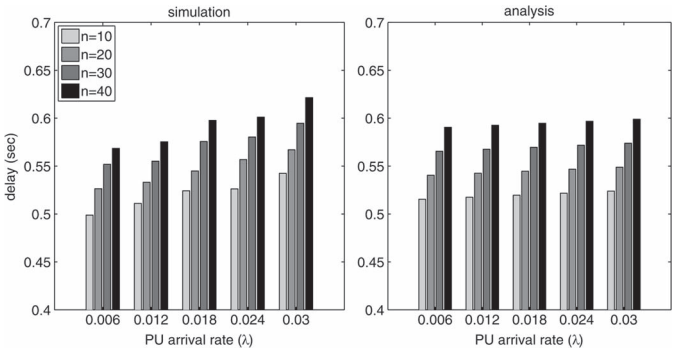


Fig. 7. Average delay vs. PU arrival rate λ with varying number of nodes with mean packet inter-arrival time of 120 sec contending over a single channel.

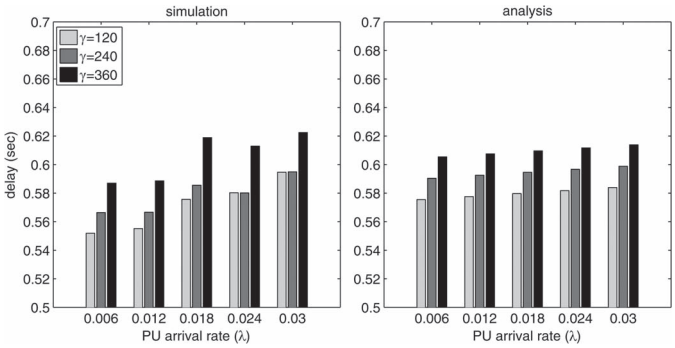


Fig. 8. Average delay vs. PU arrival rate λ with varying packet arrival rate γ for 30 nodes contending over a single channel.

mathematical analysis. Fig. 7 shows the average delay versus PU arrival rate λ for the channel, while we vary the number of contending nodes from 10 to 40, each having packet arrival rate (1/120). In another trial, we varied the Poisson packet arrival rate γ for 30 contending nodes, and plotted the average delay versus the PU arrival rate as shown in Fig. 8. Both plots verify that simulation results closely follow the results of MATLAB analysis.

VI. CONCLUSION

We have formulated a channel allocation scheme for networks of heterogeneous QoS classes, where the spectrum PU occupancy statistics is also varying in different channels, and identified performance bounds for this approach. We proposed two Markovian models that calculates the average delay of the nodes for general streaming and non-streaming QoS classes that are in close agreement with simulation studies. We used the aforementioned models to devise a greedy algorithm with polynomial time complexity that assigns SUs with channels, while ensuring their QoS requirements are met. We show that the spectral efficiency of the given allocation scheme converges to the theoretical bounds for a practical case study, using real-world measurement traces for wireless medical telemetry bands.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their very constructive comments.

REFERENCES

[1] J. Mitola, "Cognitive radio: An integrated agent architecture for software defined radio," Ph.D. dissertation, Royal Inst. Technol. (KTH), Stockholm, Sweden, 2000.

[2] I. F. Akyildiz, W. Y. Lee, and K. R. Chowdhury, "CRAHNs: Cognitive radio *ad hoc* networks," *Ad Hoc Netw. (Elsevier) J.*, vol. 7, no. 5, pp. 810–836, Jul. 2009.

[3] B. Wang, D. Zhao, and J. Cai, "Joint connection admission control and packet scheduling in a cognitive radio network with spectrum underlay," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3852–3863, Nov. 2011.

[4] A. Alshamrani, X. Shen, and L. Xie, "QoS provisioning for heterogeneous services in cooperative cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 819–830, Apr. 2011.

[5] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[6] M. Rashid, J. Hossain, E. Hossain, and V. Bhargava, "Opportunistic spectrum access in cognitive radio networks: A queueing analytic model and admission controller design," in *Proc. IEEE GLOBECOM*, 2007, pp. 4647–4652.

[7] P. Tang, Y. Chew, L. Ong, and M. Haldar, "Performance of secondary radios in spectrum sharing with prioritized primary access," in *Proc. IEEE MILCOM*, 2006, pp. 1–7.

[8] Y. Kondareddy, N. Andrews, and P. Agrawal, "On the capacity of secondary users in a cognitive radio network," in *Proc. IEEE SARNOFF Symp.*, 2009, pp. 1–5.

[9] X. Zhu, L. Shen, and T. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," *IEEE Commun. Lett.*, vol. 11, no. 4, pp. 304–306, Apr. 2007.

[10] R. Yu, Y. Zhang, M. Huang, and S. Xie, "Cross-layer optimized call admission control in cognitive radio networks," *Mobile Netw. Appl. (Springer)*, vol. 15, no. 5, pp. 610–626, Oct. 2010.

[11] C. An, H. Ji, and P. Si, "Dynamic spectrum access with QoS provisioning in cognitive radio networks," in *Proc. IEEE GLOBECOM*, 2010, pp. 1–5.

[12] D. Zhu, J. Park, and B. Choi, "Performance analysis of an unslotted CSMA in the multi-channel cognitive radio networks," in *Proc. 5th Intl. Conf. QTNA*, 2010, pp. 156–161.

[13] Y. Xing, R. Chandramuli, S. Mangold, and S. S. N, "Dynamic spectrum access in open spectrum wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 626–637, Mar. 2006.

[14] L. Lou and S. Roy, "Analysis of dynamic spectrum access with heterogeneous networks: Benefits of channel packing scheme," in *Proc. IEEE GLOBECOM*, 2009, pp. 1–7.

[15] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore, MD, USA: John Hopkins Univ. Press, 1981.

[16] J. Kharoufeh, "Level-dependent quasi-birth-and-death processes," in *Wiley Encyclopedia of Operations Research and Management Science*. New York, NY, USA: Wiley, 2011.

[17] D. Gaver, P. Jacobs, and G. Latouche, "Finite birth-and-death models in randomly changing environments," *Adv. Appl. Probab.*, vol. 16, pp. 715–731, 1984.

[18] D. Yue, W. Yue, and R. Tian, "Analysis of two-server queues with a variant vacation policy," in *Proc. ISORA*, 2010, pp. 483–491.

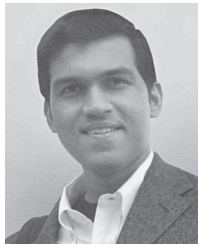
[19] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "Unsaturated throughput analysis of IEEE 802.11 in presence of non ideal transmission channel and capture effects," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1276–1286, Apr. 2008.
[20] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions," *IEEE/ACM Trans. Netw.*, vol. 15, no. 1, pp. 159–172, Feb. 2007.
[21] Q. Xin and J. Xiang, "Joint QoS-aware admission control, channel assignment, power allocation for cognitive radio cellular netoworks," in *Proc. MASS*, 2009, pp. 294–303.
[22] A. Stuart and J. Voss, *Matrix Analysis and Algorithms*. Coventry, U.K.: University of Warwick, 2009, ser. Lecture Notes.
[23] Y. Yuan, P. Bahl, R. Chandra, T. Moscibroda, and Y. Wu, "Allocating dynamic time-spectrum blocks in cognitive radio networks," in *Proc. ACM MobiHoc*, 2007, pp. 130–139.
[24] E. Coffman, P. Robert, F. Simatos, S. Tarumi, and G. Zussman, "Channel fragmentation in dynamic spectrum access systems—A theoretical study," in *Proc. SIGMETRICS*, 2010, pp. 333–344.
[25] R. Hampton, "Lessons learned from interference to wireless medical telemetry service systems," *Biomed. Instrum. Technol.*, vol. Suppl., pp. 37–39, 2006.
[26] Medical Connectivity. [Online]. Available: http://medicalconnectivity.com/2008/04/27/an-assessment-of-wireless-med ical-telemetry-system-wmts
[27] S. D. Baker and D. H. Hoglund, "Medical-grade mission-critical wireless networks," *IEEE Eng. Med. Biol. Mag.*, vol. 27, no. 2, pp. 86–95, Mar./Apr. 2008.
[28] R. Doost-Mohammady and K. R. Chowdhury, "Enhancing wireless medical telemetry through dynamic spectrum access," in *Proc. ICC*, 2012, pp. 1603–1608.
[29] W. Lee and I. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3845–3857, Oct. 2008.

Rahman Doost-Mohammady (M'14) received the B.Sc. degree in computer engineering from Sharif University of Technology, Tehran, Iran, in 2007 and the M.Sc. degree in embedded systems from Delft University of Technology, Delft, the Netherlands, where he started his research on cognitive radio networks.

Since 2010, he has been a Research Assistant with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, where he is also currently toward the Ph.D. degree on the implementation issues and applications of cognitive radio networks. He has also won the Best Paper Award in the Cognitive Radio and Networks Symposium at the IEEE International Conference on Communications in 2012.

M. Yousof Naderi (M'09) received the B.Sc. degree in computer engineering from Shahid Beheshti University (formerly National University of Iran), Tehran, Iran, in 2008 and the M.Sc. degree (with honors) in communication and computer networks from Sharif University of Technology, Tehran, in 2010. He is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.

His current research interests include the design and experimentation of novel communication protocols, algorithms, and analytical models specialized for wireless energy harvesting networks, cognitive radio networks, multimedia sensor networks, and cyber-physical systems.



Kaushik Roy Chowdhury (M'09) received the B.E. degree in electronics engineering with distinction from VJTI, Mumbai University, Mumbai, India, in 2003 the M.S. degree in computer science from the University of Cincinnati, Cincinnati, OH, USA, in 2006, and the Ph.D. degree from Georgia Institute of Technology, Atlanta, GA, USA, in 2009. His M.S. thesis was given the Outstanding Thesis Award jointly by the Department of Electrical and Computer Engineering and by the Department of Computer Science, University of Cincinnati.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. He currently serves on the editorial board of the *Elsevier Ad Hoc Networks* and *Elsevier Computer Communications* journals. His expertise and research interests include wireless cognitive radio ad hoc networks, energy harvesting, and intra-body communications.

Dr. Chowdhury is the recipient of multiple Best Paper Awards at the IEEE International Conference on Communications.