This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2009 proceedings.

A Resource Allocation Algorithm for Real-Time Streaming in Cognitive Networks

Diego Piazza†, Pamela Cosman‡, Laurence B. Milstein‡ and Guido Tartara†
†Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy
‡ECE department, University of California, San Diego, USA

***Abstract*—Cognitive radios have been proposed as a means to implement efficient reuse of the licensed spectrum. Commonly, wireless networks are characterized by a fixed spectrum assignment policy. The limited available spectrum and the inefficiency in the spectrum usage necessitate a new communication paradigm to exploit the existing wireless spectrum opportunistically. We consider a simple single-cell scenario with two data up-links, one licensed to use the spectral resource (primary) and the other unlicensed (secondary or cognitive). It is assumed that the cognitive user accesses the channel only when the channel is sensed idle. An ON-OFF channel model is used for the primary link, where traffic statistical characteristics are taken into account. We study a practical resource allocation algorithm that assigns the uplink to the secondary users according to a channel-and-queues aware scheduler when primary link OFF periods are sensed. We fit the resource allocation algorithm to the widely investigated orthogonal frequency division multiple access (OFDMA) scheme and we exploit multiuser diversity by applying a smart power allocation within independent OFDMA subchannels. A video encoder rate control is introduced in order to limit the video frame loss due to overflow that trades the video frame loss probability with the overall encoding quality. Lastly, the performance of the cognitive network model is investigated under the proposed resource allocation algorithm.**

***Index Terms*—Cognitive networks, Dynamic resource allocation, scheduling, OFDMA.**

I. Introduction

Based on evidence that a fixed (licensed) spectrum allocation can result in a highly inefficient resource utilization [1], cognitive radio prescribes the coexistence of licensed (primary) and unlicensed (secondary or cognitive) radio nodes on the same bandwidth. While the first group is allowed to access the spectrum at any time, the second seeks opportunities for transmission by exploiting the idle periods of primary nodes [2]. The main requirement is that the activity of secondary nodes should be transparent to the primary, so as not to interfere with the licensed use of the spectrum. Cognitive radio, a term first coined by Mitola [3], is a highly flexible alternative to the classic mode of operation. By sensing and adapting to its environment, a cognitive radio is able to avoid interference and fill voids in the wireless spectrum, thus increasing spectral efficiency. Although the gains to be made by the combination of cognitive radios and primary spectrum licensing seem intuitive, the fundamental theoretical limits of the gains to be made by this coupling have only recently been explored [4]. In [5] and [6], the capacity limits of cognitive networks are studied at the physical layer. Resource allocation algorithms based on Markov chain channel models, where the primary link might be either busy or idle, have been proposed [7], while in contributions such as [8] game theory is used to dynamically obtain spectrum sharing rules.

In essence, cognitive radio relies on the access to knowledge of the primary users' activity by the secondary users. However, obtaining sufficiently accurate information on the radio environment (e.g., on the primary activity) at the cognitive nodes is one of the key problems in the implementation of cognitive networks [9].

Resource allocation in multiuser single-antenna wideband OFDMA systems is a largely researched problem (e.g, [10]). The key advantage of such systems with respect to single-carrier systems is the possibility of considering frequency as an additional resource to be allocated. Although the maximum normalized average throughput is not increased with respect to that in a single-carrier system, a more efficient use of the resources is possible, especially if the bandwidth is considerably larger than the coherence bandwidth of the channel and the channel is varying slowly with respect to the scheduling updates. This is mainly because randomness in the system is increased by the wideband resources, and this feature can be exploited by a smart scheduler [11] that takes into account the variability of the channel. Further, a "cross-layer" approach to the resource allocation problem has been addressed [12] with the so-called channel-and-queue-aware scheduling. That is, both users' channel quality and queue status are accounted for when allocating resources in order to meet some real-time requirements.

In this paper, we introduce a resource allocation algorithm for cognitive networks. In particular, we specialize it to the case of an OFDMA uplink, where secondary users have to meet some real-time requirements. We revise the well-known ON-OFF channel model, taking into account both the traffic model and the physical channel characteristics. In particular, we use a Pareto distribution [13] for modeling the burstiness of stochastically heavy-tailed primary traffic in order to test the ability of the resource allocation algorithm to fulfill the secondary users' real-time constraints. A time-correlated fading channel model is used at the physical level for secondary users. A channel-and-queues aware modified proportional fair (PF) scheduling algorithm is employed in each of the OFDMA subchannels in order to assign the uplink resources to the best users according to a certain utility function. Users with heavily loaded queues (i.e., number of bits queued above a certain

This work was supported in part by the Office of Naval Research under grant number N000140810081

978-1-4244-2948-6/09/$25.00 ©2009 IEEE

threshold) are prioritized and, at the same time, users with nearly empty queues are penalized by applying a correction factor to the standard PF algorithm presented in [11]. Power allocation is performed for each subchannel for throughput optimization under a total transmitted power constraint. Real-time constraints apply to the system due to the video streaming traffic of the secondary users.

The paper is organized by covering in Section II a model of the cognitive network, particularly regarding the ON-OFF channel model, while the channel-and-queues aware scheduling and power allocation algorithm is presented in Section III. Numerical results are discussed in Section IV.

II. System Model

We consider the uplink of an OFDMA system, where the overall frequency band B is subdivided into M orthogonal subchannels of bandwidth $\Delta B = B/M$. Data frames are organized in time slots of duration T_S seconds that we consider as the system's discrete time unit. Furthermore, we assume the the M uplink subchannels are fading independently.

No interference is allowed from secondary users to primary users, and hence that secondary users exploit each uplink subchannel only when the corresponding primary subchannel is idle. Thus, the primary link usage can be conveniently modeled as a set of ON-OFF processes, one for each orthogonal subchannel, where we label as OFF periods the time slots where the primary users are not transmitting on that subchannel. Furthermore, we assume that instantaneous and error-free information about the ON-OFF process is available at the secondary users' scheduler.

A. Primary Traffic Model

An ON-OFF model is commonly used to capture the burstiness and dependency structure of traffic streams. It is well known that Internet traffic appears similar in different time scales, which is called self-similarity [13]. Statistical analysis shows that the ON-OFF periods of self-similar traffic exhibit long-tailed characteristics, which are described by a Pareto distribution as follows:

$$P(X > s) = \left(\frac{s_{min}}{s}\right)^{\alpha}, \; s > s_{min} > 0; \alpha > 0, \qquad (1)$$

where s_{min} is the smallest possible value of the discrete random variable X, which represents the number of time slots, each of duration T_S seconds, in which the primary channel is ON. The OFF period length is modeled by an exponentially distributed variable with mean s_{off} [12]. The primary channel is considered to be either ON or OFF during a time slot, and the average probabilities of finding the channel busy or idle are P_{ON} and $P_{OFF} = 1 - P_{ON}$, respectively. Channel availability for secondary users can be conveniently modeled as a set of M independent ON-OFF processes running simultaneously (one for each subchannel).

B. Secondary Users' System Model

Secondary users transmit on the same OFDMA subchannels as primary users. Each of the K secondary users is entitled to transmit on the uplink only during OFF periods in order not to cause interference to licensed users. A resource allocation algorithm schedules the secondary users and the transmit power for each of the M subchannels (Fig.1).

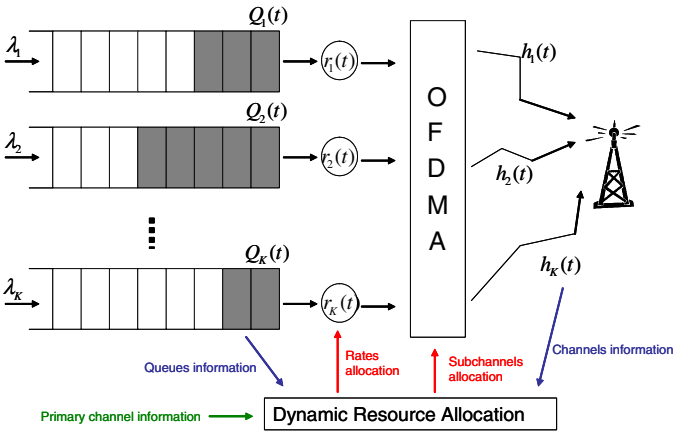

Fig. 1. Model of the secondary cognitive system.

Real-time video traffic is buffered in finite-length queues and $Q_k(s)$ is the time-varying level of the $k-th$ user's queue, which is filled by the video encoder at an average rate of λ_k. We indicate by L_Q the maximum length of a queue. Let $\mathbf{h}_k(s) = [h_{k,1}(s)...h_{k,M}(s)]$ be the set of time-varying fading channel gains between the $k-th$ user and the base station (BS) on the M subchannels at time slot s. Thus, the signal-to-noise ratio at the BS in the $m-th$ subchannel at time slot s is

$$SNR_m(s) = \frac{P_{\widehat{k},m}(s) h^2_{\widehat{k},m}(s)}{P_N} \qquad (2)$$

where $P_{\widehat{k},m}(s)$ is the power allocated to the user $\widehat{k}$ scheduled in the $m-th$ subchannel and P_N is the Gaussian noise power. A total transmitted power constraint has to be met in any time slot (e.g., for regulatory purposes), so that $\sum_{m,\widehat{k}} P_{\widehat{k},m}(s) = P_T$.

C. Queues and Channel State Information

Information about secondary users' queues and uplink channel states (i.e., the set of OFDMA subchannels) is required at the BS in order to perform scheduling and power allocation (Fig.1). We assume that within each subchannel's OFF period, the BS has exact knowledge of $\mathbf{Q}(s) = [Q_1(s), ..., Q_K(s)]$ and $\mathbf{H}(s) = [\mathbf{h}_1^T(s), ..., \mathbf{h}_K^T(s)]$. Furthermore, secondary users and the BS are assumed to be able to sense instantaneously any ON-OFF subchannel transitions. However, no knowledge is assumed on the duration of OFF-ON periods, and the resource allocation is made on a slot-by-slot basis.

D. Video Encoding

The encoder converts the video frames into a bit stream that is written into the encoder output buffer. The data rate of the bit stream can vary due to the type of frame that is to be encoded, and also due to the video content (e.g., a scene cut) [14]. The encoder is free to allocate the rate within the frames so as to optimize some (quality) criterion, while making sure that, on average, it is able to adhere to the target constant bit rate (CBR), λ_k. We choose to model the encoded bit stream arrival process $\lambda_k(s)$ with a truncated Gaussian distribution with $E[\lambda_k(s)] = \lambda_k$, $Var(\lambda_k(s)) = \sigma_k^2$ and $\lambda_k(s) > 0$.

III. Dynamic Resource Allocation

The resource allocation algorithm is performed only in subchannels that at a given time slot are sensed idle by secondary users, that is, during primary subchannels' OFF periods. The goal of the algorithm is to schedule the user in each available subchannel in a manner that optimizes a certain utility function, and to allocate the total available power across the subchannels in order to maximize the overall throughput.

A. Scheduling Utility Function

The proportional fair (PF) scheduling algorithm was proposed [11] as a tradeoff between diversity gain, fairness and packet delay. Let $r_{k,m}(s)$ be the transmission data rates achievable by user k on the $m-th$ subchannel at time slot s on the basis of the channel state $h_{k,m}(s)$ and the system's adaptive modulation and coding (AMC) scheme, and let $T_1(s), \ldots, T_K(s)$ be the average throughputs in a sliding window of s_c slots. The proportional fair algorithm chooses the user to transmit on subchannel m that has the largest ratio

$$\mu_{k,m}(s) = \frac{r_{k,m}(s)}{T_k(s)}. \tag{3}$$

If $\widehat{k}_m$ is the user selected by the scheduler at time slot s in subchannel m, the update of $T_k(s)$ for the next time slot is given by

$$T_k(s+1) = \begin{cases} (1-\frac{1}{s_c})T_k(s) + \frac{1}{s_c} r_{k,m}(s) & k = \widehat{k}_m \\ (1-\frac{1}{s_c})T_k(s) & k \neq \widehat{k}_m \end{cases} \tag{4}$$

The ratio $\frac{r_{k,m}(s)}{T_k(s)}$ is proportional to the rate supported by the user's uplink, and thus to the channel gain, and inversely proportional to the average throughput. The parameter s_c is related to the latency (or delay). Thus, the PF algorithm schedules a user when its instantaneous channel quality is high relative to its average channel condition over the time-scale s_c.

We propose a modified version of the PF scheduling algorithm (MPF) that also takes into account the queues' state $\mathbf{Q}(s)$. Specifically, the utility function in (3) becomes

$$\eta_{k,m}(s) = \mu_{k,m}(s)\pi_k(s), \tag{5}$$

where we introduce a correction factor $\pi_k(s)$ defined as

$$\pi_k(s) = \begin{cases} e^{\frac{Q_k(s)-\alpha_k}{\epsilon_k}} & Q_k(s) < \alpha_k \\ e^{\frac{Q_k(s)-\beta_k}{\epsilon_k}} & Q_k(s) > \beta_k \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

In (6), α_k and β_k are scheduler-defined thresholds that determine which level of emptiness or fullness a user's queue has to reach to begin to penalize or prioritize the user. The parameter ϵ_k can be regarded as a shaping factor that determines how fast the penalization or the prioritization is carried out depending on the queue's level.

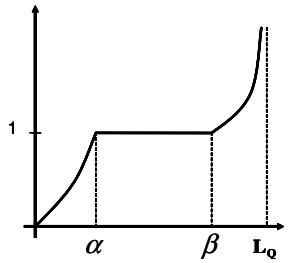

Fig. 2. Scheduling utility function as a function of the level in the user's queue

Accordingly, when the queue level is between the thresholds α_k and β_k, a standard PF algorithm is performed (Fig.2). The scheduling algorithm is able to react to a potential overflow of a given user's buffer by prioritizing that user (through the utility function $\pi_k(s)$), but it also takes into account the uplink channel condition (through $\mu_{k,m}(s)$). Alternatively, a user whose queue is near empty would be penalized until the queue fills up to the threshold α_k. The prioritization process relies on fair behavior of the users' encoders. In fact, a cheating user could arbitrarily use a larger encoding data rate in order to maintain a full queue and higher priority than fair users.

B. Resource Allocation Algorithm

Let the m-th primary subchannel to be in the OFF state at time slot s, i.e., no primary users transmit on the m-th subchannel at time slot s, and let the scheduler have exact knowledge of $\mathbf{Q}(s)$ and $h_{k,m}(s)$ for $k = 1, ..., K$, i.e., the queue state vector and the set of users' uplink m-th subchannel gains, respectively. Suppose that the set $h_{1,m}(s), ..., h_{K,m}(s)$ can be mapped into a corresponding data rate set $r_{1,m}(s), \ldots, r_{K,m}(s)$ for an equal power distribution over $N(t)$ subchannels (i.e., $P_{k,m}(s) = P_T/N(s)$), where $N(s) \in [0, M]$ is the number of primary subchannels in the OFF state at time slot s. Thus, the k-th user's signal-to-noise ratio in the m-th subchannel $SNR_m(s) = \frac{P_T h_{k,m}(s)^2}{N(s)P_N}$ can be mapped to the corresponding data rate $r_{k,m}(s)$ according to a predefined bit loading table (e.g., as in Table I [15]).

TABLE I
Downlink parameters (IEEE 802.16)

Modulation	*Coding rate*	*SNR range* [dB]	*Spectral efficiency* [bit/carrier]
QPSK	1/2	9.4-11.2	1
QPSK	3/4	11.2-16.4	3/2
16-QAM	1/2	16.4-18.2	2
16-QAM	3/4	18.2-22.7	3
64-QAM	2/3	22.7-24.4	4
64-QAM	3/4	$>$ 24.4	9/2

For each user in each subchannel that is in the OFF state, we can compute the utility function $\eta_{k,m}(s)$ according to (5) and (6), and we can schedule the user that maximizes the utility function:

$$\widehat{k}_m(s) = \text{argmax}_k\{\eta_{k,m}(s)\}. \quad (7)$$

Thus, we have a secondary user scheduled in each of the $N(s)$ subchannels where primary users are not transmitting.

We are now able to find the optimal power allocation over the $N(s)$ scheduled users. In particular, we choose to allocate the transmitted power that maximizes the overall throughput:

$$\mathbf{P}_{opt}(s) = \text{argmax}_{\mathbf{P}}\{\sum_m r_{\widehat{k}_m,m}(s)\}, \quad (8)$$

where $\mathbf{P}_{opt}(s)$ is the optimal power allocation vector of $N(s)$ elements, and $\mathbf{P}$ is the matrix corresponding to all allowed power allocations. Although it is certainly a suboptimal approach, searching a limited number of possible power allocations limits the delay in the resource allocation algorithm. Before scheduling is performed for the next time slot, the average throughput update is carried out according to (4).

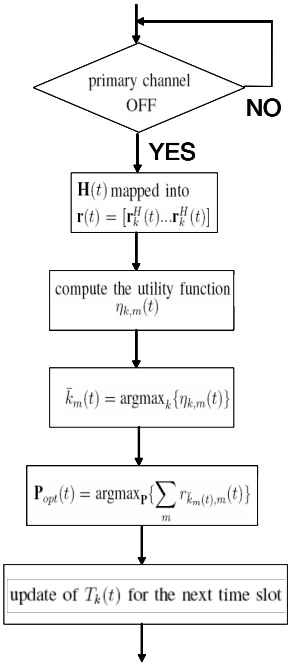

Fig. 3. Resource allocation algorithm workflow

C. MAC-Application Layer Interaction

The resource allocation algorithm is performed only if at least one subchannel is in the OFF state ($N(s) > 0$), thus, no outflow from the secondary users' queues is possible if all the subchannels are used by primaries. Each secondary user's video encoder has knowledge of the corresponding queue, so that we can force the encoder to lower the average video stream bit rate λ_k in order to reduce the probability of overflow when secondary users are not allowed to transmit (i.e., $N(s) = 0$). Also, more complicated data rate control algorithms have been implemented in video encoders [14] that could be suitable to provide an overflow prevention mechanism.

IV. Numerical Results

Simulations have been carried out for an OFDMA system with $B = 5$MHz, $M = 4$ subchannels, and each time slot is $T_S = 5$msec long. For each subchannel a primary channel ON-OFF process was simulated with parameters s_{min} and s_{off} chosen to obtain the desired OFF probability, P_{OFF}. A time-correlated flat-fading process with mean signal-to-noise ratio (SNR) of 18dB is independently run for each of $K = 6$ secondary users for each subchannel, and the bit loading corresponding to the users' SNR is taken from Table I [15]. The maximum queue length is equal for all the users and it is set to $L_Q = 2$ Mb.

In Fig. 4, the performance of the resource allocation algorithm is evaluated. We compare the video frame loss probability of the standard PF (dashed curve) and the MPF (solid curve) scheduling algorithms, both with and without power optimization (i.e., each user scheduled transmits at P_T/M), for $P_{OFF} = 0.3$ and $P_{OFF} = 0.5$. We assume the secondary channel to be error-free once the bit loading is made, thus, video frame loss is estimated by accounting for the related overflow occurrences. Note that there is a performance gain

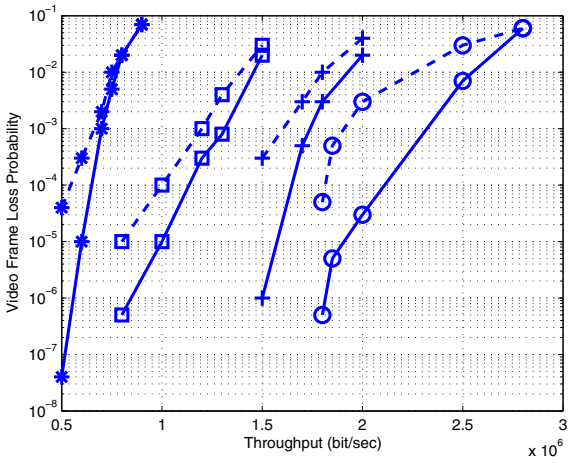

Fig. 4. Standard PF (dashed) and MPF (solid) performances for $P_{OFF} = 0.3$ and $P_{OFF} = 0.5$ without power optimization ($'*'$ and $'\square'$ curves, respectively), and with power optimization ($'+'$ and $'o'$ curves, respectively).

in all cases considered comparing PF and MPF scheduling. As expected, a larger throughput is achievable for increasing P_{OFF}, as the channel is available more often, and for a similar data rate, we expect a lower probability of overflow. Although a very simple power optimization scheme has been used, in which only $\pm$3dB power steps are allowed, the throughput of the system is nearly doubled for a given video frame loss probability.

In Fig. 5, we take into account the possibility of controlling the video stream rate when secondary users are not allowed to

transmit (i.e., $N(s) = 0$). In particular, we apply a correction factor ρ to λ_k with $\rho_1 = 1$, $\rho_2 = 0.8$, $\rho_3 = 0.5$ and $\rho_4 = 0.2$. The performance gain achieved by MPF over PF still holds,

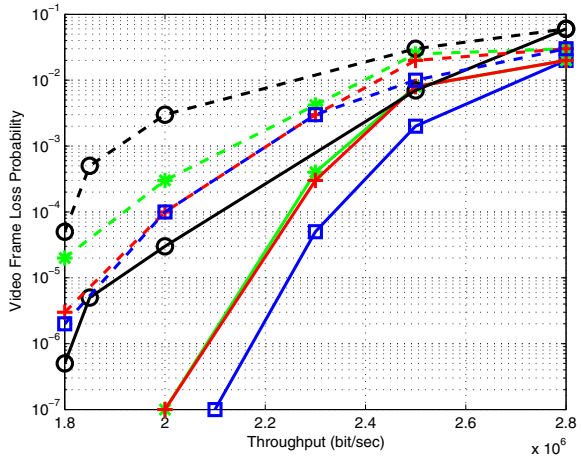

Fig. 5. Standard PF (dashed) and MPF (solid) performances for $P_{OFF} = 0.5$ with power optimization and video encoder rate control with $\rho_1 = 1$ ($'o'$ curves),$\rho_2 = 0.8$ ($'*'$),$\rho_3 = 0.5$ ($'+'$) and $\rho_4 = 0.2$ ($'\square'$).

and the video encoder data rate control provides a means to fairly reduce the overflows. Overall, the probability of video frame loss is significantly decreased, at the expense of a lower quality encoding when all the subchannels are in the ON periods, caused by the reduction of the average video encoder data rate $\rho\lambda_k$.

V. Conclusions

In this paper, we have introduced a resource allocation algorithm for secondary users in an OFDMA cognitive network with real-time constraints. A rather simple ON-OFF process is used to model the primary user activity on the uplink. In particular, a Pareto distribution is used for taking into account the burstiness of the ON periods. The secondary users' real-time constraint is modeled through the adoption of finite-length queues that are filled by the video encoder continuously, and emptied by the transmitter in a bursty fashion during the OFF periods. The scheduling algorithm is a modified version of the well-known PF scheduler, where a queue-aware utility function has been assumed to work along with the usual channel-driven PF rule. Simulation results show that our scheduling algorithm achieves better performance with real-time traffic because of its ability to prioritize near-to-overflow users. Furthermore, we introduce a throughput maximization scheme when allocating power to the transmitting users. Even rather simple optimization schemes achieve a huge throughput gain without affecting the overall algorithm complexity. Finally, a MAC-application layer interaction has been proposed according to which the video encoder is forced to reduce the video streaming data rate when no secondary users are allowed to transmit on the uplink subchannels. A tradeoff between quality degradation due to video frame loss and that due to the video encoding data rate control emerges. More sophisticated rate control algorithms could be proposed that also take into account the encoding process (I, B and P frames) in order to reduce the quality degradation.

References

[1] Federal Communications Commission Spectrum Policy Task Force, FCC Report of the Spectrum Efficiency Working Group, November 2002. http://www.fcc.gov/sptf/files/SEWGFinalReport_1.pdf

[2] S. Haykin, Cognitive radio: brain-empowered wireless communications, IEEE Journal on Selected Areas Commun., vol. 23, no. 2, pp. 201-220, Feb. 2005.

[3] J. Mitola III, Cognitive Radio: an Integrated Agent Architecture for Software Defined Radio. Ph.D Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2000.

[4] N. Devroye, P. Mitran and V. Tarokh, Achievable rates in cognitive radio, *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1813-1827, May 2006

[5] A. Jovicic and P. Viswanath, Cognitive radio: an information-theoretic perspective, submitted http://lanl.arxiv.org/PScache/cs/pdf/0604/0604107.pdf

[6] S.A. Jafar and S. Srinivasa, "Capacity limits of cognitive radio with distributed and dynamic spectral activity," Selected Areas in Communications, IEEE Journal on , vol.25, no.3, pp.529-537, April 2007

[7] Y. Xing, R. Chandramouli, S. Mangold and S. Shankar N, Dynamic spectrum access in open spectrum wireless networks, IEEE Journ. Selected Areas Commun., pp. 626-637, vol. 24, no. 3, March 2006.

[8] R. Etkin, A. Parekh and D. Tse, "Spectrum sharing for unlicensed bands," New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on , vol., no., pp.251-258, 8-11 Nov. 2005

[9] G. Ganesan and Y. Li, "Cooperative spectrum sensing in cognitive radio networks," New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on , vol., no., pp.137-143, 8-11 Nov. 2005

[10] M. Realp, R. Knopp and A.I. Perez-Neira, "Resource allocation in wideband wireless systems," Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on , vol.2, no., pp. 852-856 Vol. 2, 11-14 Sept. 2005

[11] P. Viswanath, D. N. Tse, and R. Laroia, "Opportunistic beamforming using dumb antenna," *IEEE Trans Inf. Theory*, vol. 48, no. 6, pp. 1277-1294, June 2002.

[12] G. Song Li, L.J. Cimini, Jr. and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE , vol.3, no., pp. 1939-1944 Vol.3, 21-25 March 2004

[13] W. Willinger and M. S. T. Vern Paxson, Self-similarity and heavy tails: Structural modeling of network traffic, A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Robert J. Adler, Raisa E. Feldman and Murad S. Taqqu, editors. Birkhuser, Boston, 1998.

[14] A. Leontaris and P. C. Cosman, "Compression Efficiency and Delay Tradeoffs for Hierarchical B-Pictures and Pulsed-Quality Frames," *IEEE Trans. on image processing*, VOL. 16, NO. 7, JULY 2007.

[15] Standard Specifications 802.16-2004, available at http://ieee802.org/16