Opportunistic Spectrum Access: Challenges, Architecture, and Protocols

C. Santivanez, R. Ramanathan, C. Partridge, R. Krishnan, M. Condell, S. Polit

Internetwork Research Department,
BBN Technologies, Cambridge, MA, USA

{csantiva,ramanath,craig,krash,mcondell,spolit}@bbn.com

Abstract

We consider the concept of *opportunistic spectrum access (OSA)* – whereby radios identify unused portions of licensed spectrum, and utilize that spectrum without adverse impact on the primary licensees. OSA allows both dramatically higher spectrum utilization and near-zero deployment time, with an obvious and significant impact on both civilian and military communications. We discuss two broad classes of challenges to OSA: *spectrum agility*, which involves wideband sensing, opportunity identification, coordination and use; and *policy agility*, which enables regulatory policies to be applied dynamically using machine understandable policies. Focusing on spectrum agility, we present an architecture based on an OSA adaptation layer. We describe protocols for OSA, including a hole information protocol, idle channel selection and use, and a novel access protocol for the coordination channel. We present a simulation study, discuss insights, and show that our OSA system can provide an order-of-magnitude performance improvement in throughput over a legacy system.

1 Introduction

Imagine traffic laws in which each lane in the highway is dedicated to particular makes of car – BMWs and Saabs use lane 1, Toyotas and Fords lane 2, and so on. A Toyota cannot use lane 1 even if that lane is empty! Such a scheme is in many ways similar to the regulatory regime governing spectrum allocation today. Allocation is based on services – e.g. land mobile, public safety, and broadcast television – and regulations forbid a device from using an empty portion of the spectrum (even though the device may be capable of doing so) unless the particular service has been allocated that spectrum.

Not surprisingly, this regulatory regime results in large portions of unused spectrum [1]. This underutilization is true both spatially and temporally. That is, there are a number of instances of spectrum that are used only in certain geographical areas, and a number of instances of spectrum that are used only for short periods of time. Detailed

measurements have verified that spectrum occupancy is very low – a typical utilization based on data from [3] is illustrated in Figure 1. The utilization of available spectrum is thus highly inefficient, leading to *apparent* (or artificial) spectrum scarcity. Another consequence of current regulatory policy is *deployment difficulty* – each new kind of communicating device has to be individually certified.

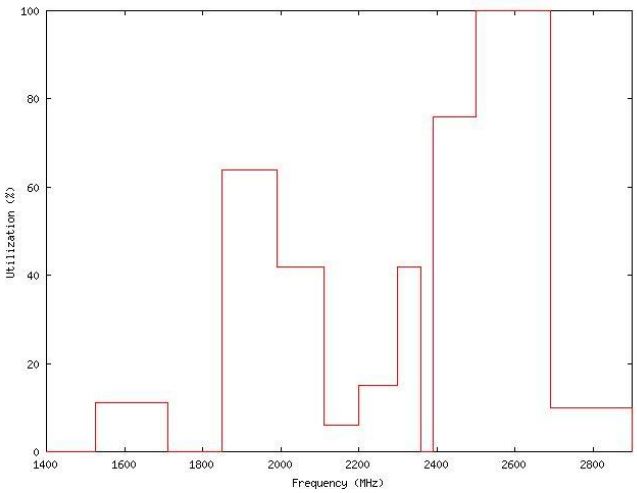

Figure 1: Peak hour spectrum occupancy measurements around Dupont circle (Washington DC), 1.4 – 2.9 GHz [3].

Meanwhile, the Wireless Internet continues to grow at a rapid pace, both in terms of number of users and their airtime. The unlicensed bands are getting congested in one place/time while swathes of spectrum lie unused elsewhere. It is clear that in order to support the growth of the Wireless Internet, we must find a way to access this unused potential.

In recent years, the concept of *opportunistic spectrum access* has emerged as a way to dramatically improve spectrum utilization. The basic idea is this: a device first senses the spectrum it wishes to use and characterizes the presence, if any, of primary[1] users. Based on that information, and regulatory policies, the device identifies communication opportunities ("holes") in frequency, time, or even code – and transmits using those opportunities in a manner that limits the interference perceived by primary users. Opportunistic spectrum access allows dramatically higher spectrum utilization. It also enables near-zero deployment time through radios that can opportunistically retarget their services to a new portion of the spectrum as needed — with obvious and significant impact on both civilian and military communications. Opportunistic spectrum access is also

[1] A *primary* user (sometimes referred to as *incumbent*) is one who holds the rights (license) to the spectrum in question. A *secondary* user is one who is authorized to use licensed spectrum opportunistically without causing unacceptable interference to primary users.

referred to as *dynamic spectrum access*, and is often included as part of the larger concept of *cognitive radios.*

While conceptually simple, the realization of opportunistic spectrum access is challenging. Several problems must be solved: sensing over a wide frequency band; identifying the presence of primary users and determining the nature of opportunities; coordinating the use of these opportunities with other nodes; and most importantly, the definition and application of interference-limiting policies, and adherence to these policies while utilizing the opportunities.

Fortunately, recent technological advances in a number of areas can be brought to bear on this problem. First, the emergence of Software Defined Radios has enabled the RF-level programmability and spectrum agility essential to opportunistic spectrum access. Second, wideband sensing technologies have come a long way due to faster digital signal processors and tunable filters. Third, the use of waveforms that can be adapted to fit a specified spectral profile – e.g., waveforms that can occupy non-contiguous frequencies – is beginning to be better understood. Additionally, there is widespread interest in this flexibility from regulatory bodies such as the FCC (see, for instance, [1][2]). Furthermore, the impetus to develop *secondary markets* for spectrum purchased at auction (very expensively) but not in use is adding to the urgency for a change in the regulatory regime, and is making it worthwhile for organizations to invest in this technology.

Devices with opportunistic spectrum access capability will be expected to operate over a wide range of frequencies and within different geopolitical regions. We believe that opportunistic access requires the devices to be both *spectrum-agile* and *policy-agile*. A spectrum-agile device can operate over a wide range of frequencies. A policy-agile device understands the constraints under which it operates: which frequencies are available, and the rules for opportunistically using those frequencies. As these rules change according to location (and decisions of policy makers and primary users) a radio, especially a mobile radio, must be able to shift from one set of policies to another easily.

In this paper, we discuss the challenges we need to surmount in solving the spectrum agility and policy agility problems. We then present a simple architecture for opportunistic spectrum access with a view to harvesting the "low hanging fruit". Within the context of this architecture, we describe novel protocols for achieving spectrum agility, and show using simulations that significant performance gains are possible even with simple approaches.

We have organized the remainder of the paper as follows. First we survey related research, concepts, regulations and standards activities. Next, in section 3, we discuss the hard problems in policy-responsive opportunistic spectrum access. Section 4 is devoted to our solution approaches for spectrum agility, including simulation-based insights. Section 5 concludes this paper with some remarks on future research directions.

2 Related Work

Opportunistic spectrum access is part of a continuum of adaptation, agility and co-existence that spans several levels of sophistication, from simple "listen-before-talk" to "reasoning about environment". In the following paragraphs we survey several efforts in this continuum.

A simple way of co-existing with primaries is *dynamic frequency selection (DFS)* – a method first specified by the ITU and later by the FCC, and being developed by the IEEE 802.11h subcommittee. DFS is a harmonized set of rules for Wireless LANs to share the spectrum with primary users (mostly military radar). DFS detects other devices using the same radio channel and switches to a new "clean" channel if required. The protocol has mechanisms for the access point to instruct the terminals to switch to the new channel.

The emergence of a number of different radio technologies – e.g. 802.11.x, 802.15.x, Bluetooth, Hiperlan etc. – that share the unlicensed spectrum has given rise to the problem of destructive interference between these systems. To address the problem, *spectrum etiquette protocols* have been designed [4][5][6] so that these technologies can co-exist in the same band. Spectrum etiquette is a set of rules to be followed by all users of the spectrum so that fair and conflict-free access to the radio resource is enabled. For infrastructure-oriented networks, [7] proposes a coordinated, spatially aggregated spectrum access via a regional spectrum broker.

The next level of sophistication comes in generalizing such access to a much wider band with a multitude of diverse services, coordinating use of opportunities in cooperative and non-cooperative modes, and utilizing non-contiguous frequency holes. A radio platform called the *adaptive spectrum radio (ASR)* that demonstrates the principles for dynamically accessing the spectrum is described in [8]. The ASR adapts its frequency and modulation to exploit spectrum gaps both in frequency and time. The ASR uses an adaptive form of Orthogonal Frequency Division Multiplexing (OFDM) that exploits spectrum gaps through the use of non-contiguous carriers. A key part of such access is spectrum sensing, which is of much recent interest [9][10].

At the far end of our adaptation and co-existence continuum lies the concept of *cognitive radio*. First developed by Mitola [11], cognitive radio refers to a device that has knowledge of its capabilities, internal state and the radio environment. Further, the knowledge is represented in a form that allows for automated reasoning to satisfy the needs of the user. It allows expressive negotiations among peers about the use of radio spectrum across fluents of space, time, and user context [12]. A language for representing radio-domain knowledge, called RKRL, is given in [11]. A cognitive radio is self-aware and "knows that it knows." In

its extreme, the concept accommodates adaptation through learning[2].

Finally, this section will not be complete without mentioning the crucial "enablers" that make adaptive/opportunistic spectrum access or cognitive radios possible. Chief among these is Software Defined Radios (SDRs). Although the concept is a decade old, it is only recently that they have become "prime time" in the commercial arena [13], open source [14] and military [15]. SDRs provide the waveform agility required for spectrum agility and policy agility. Another enabler is simply the interest and encouragement from the FCC. In November 2002, the FCC released its Spectrum Policy Task Force (SPTF) report [1]. A key part of the SPTF report was the introduction of the concept of *interference temperature*. The metric is a step toward the construction of receivers that can tolerate a pre-determined amount of interference, enabling an opportunistic transmitter to determine the bound on its own transmission's spectral density and adjust parameters accordingly.

3 Challenges: Spectrum and Policy Agility

Opportunistic Spectrum Access can be divided into two broad classes of problems, and corresponding functional blocks.

1. *How do we access the opportunities in an interference-limiting manner?* How does a device sense and identify opportunities, coordinate their use, and actually use them? Current transceiver technology uses a single frequency channel, or a fixed set of channels in a single band. Moving to a *spectrum-agile* physical layer and developing a MAC layer to support agility is a significant challenge.[3]
2. *How do we control such access to be in accordance with regulatory policies in a certifiable way?* How can the regulatory policies be expressed, how can they be interpreted by the device? Given a set of possible transmission strategies, how can a device determine which ones are permitted (excluded) by policy? How can a certifying agency be sure that a device will indeed adhere to the policy? How do we support multiple (concurrent) policy authorities? In general, we need to define a framework for *policy-agile* control of the spectrum access behaviors.

We note that the problem of determining the appropriate regulatory policy or mandating the level of interference tolerance is *not* addressed. That is the job of the regulatory agencies and spectrum incumbents. Rather we simply are interested in finding ways to express the constraints, which regulators and spectrum incumbents impose on devices, and in finding ways to dynamically exploit the available spectrum, consistent with the constraints.

The remainder of this section elaborates on these two problems, and the challenges therein.

3.1 Spectrum Agility

We begin by developing the generic problem in a little more detail. For ease of discussion, we use an idealized model as follows. Say we are given a geographical region G, and F KHz of spectrum divided into n equal-sized channels of c KHz each ($F = nc$). A set of primaries, each licensed to use some subset of the c channels, operate in G. The primaries may be mobile, and thus, even if primaries use a fixed subset, the occupancy in a given channel is time varying. A set of *secondaries* – nodes utilizing opportunistic spectrum access – occupy the same region G.

The goal is for the secondaries to communicate using some subset of the c channels. The constraint is that the secondary nodes should not interfere with the primaries. In this subsection we shall use a simple version of this constraint, namely that the power level of a secondary at a primary should not exceed its (known) interference threshold. The next subsection will consider the accommodation of multiple and far more complex policy constraints. An optimization criterion could be that throughput (say, assuming infinite load) should be maximized.

Is this problem solvable? A moment's reflection raises more questions: Can we communicate with the primaries? Or is their usage pattern somehow accessible? Do the primaries send as well as receive – if they only receive, how do we know they are present? Can the secondaries only communicate using the channels not used by primaries or do they have access to a (pre-assigned) common "control" channel? And what is the nature of the network – for both primaries and secondaries?

Indeed, there are several versions of the problem depending upon our answers to these questions. Given the complexity of the problem, it is necessary to break it up into sub-problems for easier consideration. The problem can be decomposed into three sub-problems, and corresponding solution functional blocks: *opportunity awareness*, by which the nodes determine the number and nature of the opportunities; *opportunity allocation*, which is essentially the medium access control problem in this new context; and *opportunity use*, which is the problem of efficiently communicating how the (possibly noncontiguous) channels are to be used. The interplay between these functional blocks is depicted in Figure 2. We consider each of these functional blocks in more detail in the next three subsections.

[2] We note that the term "cognitive radio" has come to have several meanings, and therefore in this paper we shall refrain from using the term. Instead, we shall use specific terms as appropriate to describe the idea in question.

[3] We observe that a radio can be agile in multiple ways – in observable frequencies, accessible frequencies, transmit parameters, etc. Devices may be more agile in some dimensions than in others.

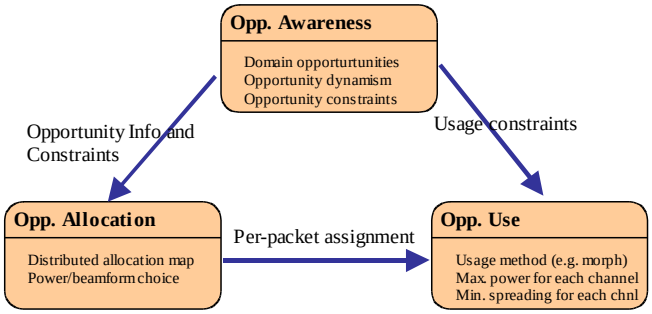

Figure 2: Spectrum agility functional decomposition

3.1.1 Opportunity Awareness

The awareness problem consists of three distinct sub problems, each offering a rich array of research issues. The decomposition and interplay is depicted in Figure 3. We now delve deeper into what makes each of these problems hard, and the challenges we need to overcome in solving them.

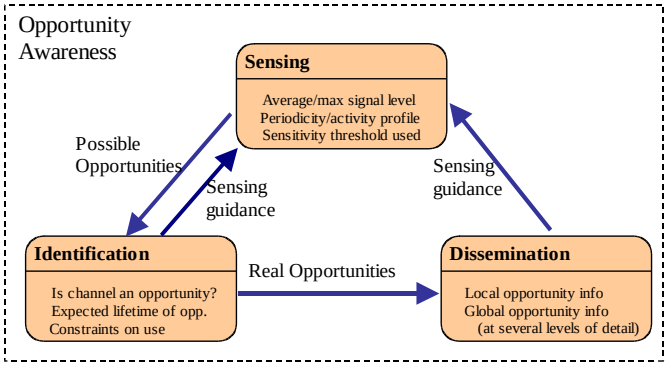

Figure 3: Opportunity Awareness functional decomposition

Wideband Sensing

Sensing a narrow channel is routinely done (e.g. Wireless LAN cards [16]) and is easy. The hard problem comes when the bandwidth F to be sensed is large, as is necessary for opportunistic spectrum access. Suppose we need to sense on 10000 channels, each of 10 KHz (that is, F = 100 MHz[4]) Continuous analog sensing is not feasible, since it would require as many as 10000 analog filters. Thus, digital signal processing techniques have to be employed. For example, the entire band needs to be sampled and a Fast Fourier Transformation (FFT) needs to be used to obtain the component of energy at every slot. This is a processing-intensive operation. In particular, there is a tradeoff between sensing interval and the accuracy of sensing and the processing cost. Further, a large sensing interval also results in more idle periods being sensed and therefore a discrepancy between peak and average power levels.

An alternative to sensing, with its attendant challenges, is to have a way to tell the secondaries how the primaries are using the channels. For instance, the primaries could register their allotment and optionally the planned usage, and have the secondaries access this database. While sensing is the more commonly proposed approach, there is at least one proposal to the FCC along the lines of the non-sensing database oriented approach, with some compelling arguments in its favor [17].

Identification

Another part of the awareness is *identification* of opportunities that can be used in an interference-free manner. Note that sensing merely tells you the characteristics of the channel. If a channel is sensed free, it may or may not be prudent to use it — for instance, it may be an emergency band that is rarely used, but when it is used, you really don't want to be interfering. Similarly, even if a channel is occupied, it may be acceptable to transmit up to some power level known not to cause interference for the primaries.

The problem of opportunity identification is two-fold. First, interference occurs at a receiver and you can only sense transmitters. Thus, if a node is not a transmitter (e.g. TV receiver) there is no way to detect it except indirectly by detecting the primary signal. Second, it is not clear what to do after detecting a primary signal. For instance, if the signal is strong, do we infer that we are close to the receiver and hence should not transmit? Or do we infer that the receiver is also likely to get a strong signal and therefore even if we proceed to transmit, the SNR at the primary receiver will be adequate? And vice-versa for a weak signal.

A conservative approach is to simply avoid a channel that has any kind of signal, weak or strong. However, that will needlessly refrain from using many legitimate opportunities and is not a satisfactory solution.

Another issue is the "hidden node" problem. In one version of this problem, an obstruction between a primary transmitter and a secondary sensor prevents sensing. A silent primary receiver is in good reception range of both the primary transmitter and the secondary node. In this case, even the detection of the primary transmitter does not work as a solution to a silent receiver.

The FCC, in partial recognition of the difficulty of these issues, concluded that receivers must be made more tolerant of interference than they are now. They proposed the concept of interference temperature, as mentioned in section 2. Even this, however, does not address the full extent of the problem because the interference at a node is the sum total of all transmissions, and all transmitting secondary nodes need to cooperate to ensure that the sum of their interferences is less than the interference temperature ceiling. And if there are multiple incompatible secondary networks operating in the same region, this is an extremely difficult problem.

[4] This is the target for the XG program.

These difficulties may be somewhat alleviated if we can assume certain capabilities/features. For instance, the ability to be much more sensitive than a primary, or the ability to sense a "pilot", which, due to its increased processing gain goes further (a good example is the DTV pilot) help somewhat. If the primaries are "chatty" (that is, they transmit something every so often), interference can be limited. We shall use this assumption in designing our protocol (see section 4).

Dissemination

In order to be able to allocate opportunities optimally, a node may need to be aware of both opportunities identified by itself, *and* also those identified at other nodes. Depending upon the operation of the opportunity allocation (channel access) algorithm, the radius of knowledge will vary. At a minimum a transmitter needs to know the opportunity profile at the intended receiver(s). Due to problems such as hidden terminal, a band that looks clear at the transmitter may not be available at the receiver. This observation leads to the problem of *dissemination* of opportunities, that is, nodes need to send their opportunity information to other nodes in the network.

Conveying opportunity information is potentially extremely bandwidth intensive. For instance, suppose we sent a bitmap (where a 1 is a "hole" and 0 is a "wall") of the status of 10000 channels. If every node advertised its own map, and the map were to be sent globally, and the map changed rapidly (a few times a second), we would easily need megabits of bandwidth simply for control traffic.

Thus, we need to consider *approximations* of the opportunity information, There are three dimensions along which approximations are possible – not sending upon every change, not sending to all the nodes, and making the information more coarse grained (e.g. representing an eight-channel profile of 10010000 with a two-by-four channel profile of 10). The challenge is to architect a dissemination mechanism that provides the right level of approximation. There is also a regime of possibilities that combine the dimensions. For instance, one could send coarse grained information to distant nodes more often, and finer grained information to nearby nodes less often.

Another problem regarding dissemination is: *which channel do we use for dissemination*? Since dissemination is done to facilitate opportunity allocation, we have a chicken-and-egg situation where we cannot send control information because we aren't allocated channels and we cannot allocate channels because we can't exchange control information. One way to resolve this dilemma is to use a dedicated control channel for opportunistic spectrum accessing nodes. That is, the regulatory agencies would allot a channel specifically for this purpose. However, this approach is impractical in many situations — for instance if we needed to set up a tactical network in hostile territory.

3.1.2 Opportunity Allocation and Use

Opportunity Allocation is the process of deciding which secondary node will use which opportunity and for how long. This is similar to the medium access control in wireless networks except that we are dealing with discontiguous channels of varying size that need to be allocated commensurate with the needs of communicating nodes.

Nonetheless, the similarity leads us to consider approaches that have worked for medium access control in (fixed channel) wireless networks. These approaches can be broadly classified into *contention-based* (e.g. CSMA/CA) and *contention-free* (e.g. TDMA, FDMA) approaches. One simple idea is to extend the Request-to-Send (RTS) and Clear-to-Send (CTS) handshake in CSMA/CA to negotiate a set of opportunities for the ensuing DATA transmission. Similarly, in the contention-free domain, one could use a combined TDMA and FDMA procedure for assigning nodes to time slots and channels.

Each of these approaches presents hard problems especially for opportunistic spectrum access in ad hoc networks. For instance, with CSMA/CA, on what channel should a node listen for an RTS? How do we ensure that neighboring nodes in another network do not also end up negotiating the same channel? With this contention-free approach, the assignment process, which has to be distributed, must take into account the dynamism of both the topology and the opportunity information, and it must react in real-time to assign opportunities – even harder than the already hard problem of dynamic TDMA.

Finally, we consider Opportunity Use. There are two problems here: first, we need waveforms that can use discontiguous portions of the spectrum; second, we need to ensure that the transmission does not affect existing users in terms of the power spectral density – which measures the power delivered by a signal in a 1 Hz band – or interference temperature [1] they perceive by controlling transmitter parameters appropriately.

It is a hard challenge to dynamically adapt the waveform to fit a power spectral density (PSD) profile (calculated based on sensed information and policy constraints). For example, given a PSD vector for a range of frequencies, how to dynamically construct a waveform that satisfies particular data rate and power constraints for the device?

This dynamic adaptation is especially challenging since we must take into account the nonlinear characteristics of the signal chain and the power amplifier of the radio to avoid unacceptable emissions or interference resulting from distortion. Specific policy requirements such as low emission thresholds at adjacent channels and sharp notch-out requirements at particular frequencies may further constrain the waveform.

In general, the waveform may be adapted across a wide range of parameters including frequency band of operation, data rate, time, power, bandwidth, modulation level, and coding. The beamform may need to be adapted as well if

directional antennas are used and directional constraints are placed on the emission.

Interference limits, or more generally (as discussed in the next section) policy constraints, only dictate *what* are the measurable constraints on the emission – for example, "field strength of any emission by the device should not exceed 1500 uV/m at a distance of 3 meters" or "99% of the power must be contained in the bandwidth (1.2 MHz) of F1." It is up to the device to figure out which parameter set, amongst a large number of such sets, ought to be used to achieve the result. Given the complexity of policies and device capabilities, this is one of the most challenging problems in this area.

3.2 Policy Agility

Current devices support only a small number of modes of operation and a limited range of intended operating environments, and therefore, all the relevant policy sets that apply can be hard-coded into the radio. However, devices capable of opportunistic spectrum access will be expected to operate over a wide range of frequencies and within different geopolitical regions.

To operate in multiple regions, in a hard-coded policy environment, one would have to hard code in all the possible policies, and combinations of policies, that the radio might encounter. That's clearly a combinatorial process and hard to do right.

Furthermore, the notion of opportunistic spectrum access being a nascent one, it is not yet clear what the right policies are. Policies will likely undergo many revisions as we gain more experience with the deployment and interference characteristics of such devices. Hard coding the policies would make it very difficult to change the policies once they are in place.

Another reason is accreditation. If a device were not policy agile then each policy change would require re-design, re-implementation and re-accreditation for each configuration of a system. That is, accreditation is an $m \times n$ problem[5] – for each of m policies, each of n device configurations/types, a separate accreditation needs to be done.

These problems can be addressed by making devices policy agile. This means two things: first, there must be a way to encode policies so that they are *machine readable and understandable*; second, there must be a way for the device to *reason* about which behaviors are applicable, given a set of policies. We note here that the results of the reasoning should be verifiable; in other words, the reasoning should be capable of also producing an automated proof of correctness of the results.

Our vision for the use of machine-understandable policies for policy agility is shown in Figure 4. On the left, policies are expressed or encoded in a standard format and then loaded into the device. The device then constrains its operation in accordance with these policies. In order to change the policies, we simply need to load a new version. For instance, operating in a different country would merely require downloading from a different website or using a new smart card.

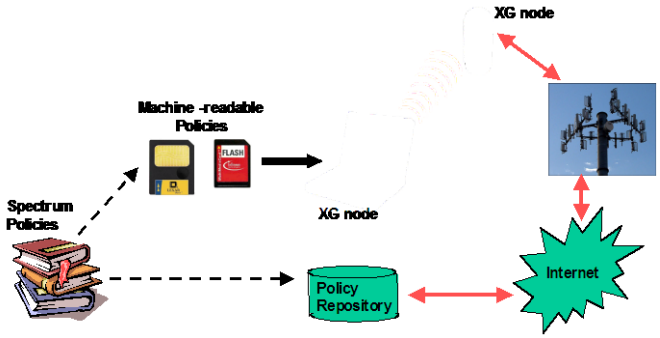

Figure 4: With machine understandable policies, changing policies dynamically is very easy -- simply use a new smart card or download from a web site.

The need for machine-understandable policies leads us to the problem of developing a *language* for expressing policies. This is a hard problem. There is a vast diversity in the primitive objects that make up regulatory policies – frequency, power spectral density, mathematical formulae, geography, database accesses, etc. Finally, it is not sufficient to be able to simply *express* the information – it must be done in a manner that conveys the structural relationships amongst the objects so that the device can *reason* about policies so that not every single fact has to be encoded.

The problem of developing a language actually consists of two sub-problems that ought to be decoupled: i) identifying the "primitive" objects for representing policies, and ii) identifying the inter-relationships between them – that is, an *ontology*. The ontology is intimately tied to the reference domain, namely regulatory policies, whereas the representation format only needs to have the power to represent all possible inter-relationships.

A more comprehensive discussion of the policy agility problem can be found in [20] and a policy language framework in [21].

4 Architecture and Protocols for Opportunistic Spectrum Access

It should be clear from the discussion in the previous section that developing a solution to opportunistic spectrum access in all its generality is hard and awaits more years of research. In this section, we describe what might be considered a small first step – an architecture and a set of very simple protocols for supporting spectrum agility in a wireless network using some simplifying assumptions. These protocols were developed at BBN as part of the DARPA neXt Generation (XG) program, and have been implemented within an OPNET simulation model. Our main objective is not necessarily to design the most efficient or

[5] Avoiding the creation of $m \times n$ problems has long been a hallmark of good communications design (see [19]).

robust solution, but to harvest the "low hanging fruit" using simple protocols, and study the benefits of spectrum agility. We only address spectrum agility, not policy agility, due to space constraints.

The target scenario is essentially as mentioned in section 3.1 – a set of primaries emit radio signals in a band, which is segmented into channels (we call them *frequency slots*), and a set of secondaries seek to communicate using "holes" in the spectrum assigned to the primaries. As is apparent from the discussion in the previous section, the problem has many variants and is hard in general. Therefore, in keeping with the goal of simplicity, we shall make the following simplifying assumptions.

1. The spectrum is divided into some number of fixed-size *frequency slots*. Primaries are assigned a subset of these slots (perhaps varying with time). Secondaries are allowed to transmit in (a set of) unused slots.
2. We assume that the primaries are chatty, that is, they are not silent for more than a small, apriori known amount of time.
3. Secondaries are assigned a small, dedicated *coordination channel* that is known apriori to all secondaries free of any other emitters.
4. Secondaries have two transceivers – the *coordination transceiver* is always tuned to the coordination channel; and the *data transceiver* is frequency-agile and may dynamically adjusts its transmission frequency.
5. Secondaries have some mechanism that can tell a secondary node from a primary one.
6. Spectrum occupancy, that is, the hole information, changes slowly enough, say not more than once in few tens of seconds (equivalently, we can ignore changes faster than that).

4.1 System Architecture

We divide the secondary node into three layers: *physical*, *OSA adaptation*, and *media access control*, as shown in Figure 5. In other words, OSA functionality is inserted between the MAC and the physical layer and can work with both conventional (OSA-unaware) as well as OSA-aware MACs. The OSA adaptation layer constitutes a particular and simple instantiation of the functions identified in Figure 5. We briefly describe each component below.

The *sensing interface* performs the opportunity awareness function. It receives from the physical layer sensor an array of receive power values for each channel and maintains a time-varying *Hole Information Array (HIA).* The HIA contains, for each channel in the band, an entry 0 or 1 indicating, respectively, whether the channel is free (a "hole") or is occupied (a "wall"). All secondaries use the same power P_s, and a channel is considered an opportunity if and only if the secondary can transmit at power P_s without interfering with any primary.

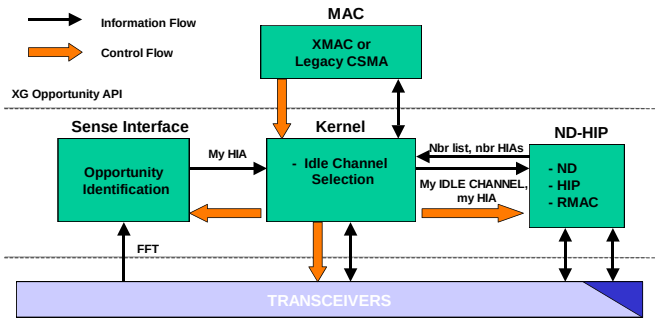

Figure 5: Our OSA system architecture consists of an OSA adaptation layer below the MAC layer that provides transparent opportunistic spectrum access.

The *sensing interface* performs the opportunity awareness function. It receives from the physical layer sensor an array of receive power values for each channel and maintains a time-varying *Hole Information Array (HIA).* The HIA contains, for each channel in the band, an entry 0 or 1 indicating, respectively, whether the channel is free (a "hole") or is occupied (a "wall"). All secondaries use the same power P_s, and a channel is considered an opportunity if and only if the secondary can transmit at power P_s without interfering with any primary.

The *Neighbor Discovery and Hole Information Protocol (ND-HIP)* module performs the opportunity dissemination function. Each node periodically broadcasts its HIA as part of a control message called the *Hole Information Protocol (HIP)* packet. One purpose of the HIP packet is to serve as "beacons" or "Hello" messages for the purpose of discovering neighbors, as in many ad hoc routing protocols (e.g. OLSR [18]). Specifically, a node considers another node as its neighbor if the moving window average of the number of received HIP packets exceeds a threshold. The HIP packets are sent using the coordination transceiver.

The HIP packet contains some other fields: the *idle channel* that the data transceiver must be tuned to when in receive mode; and a list of *rendezvous times.* This is in support of a MAC that we call the *Rendezvous MAC (R-MAC),* which is a lightweight MAC to efficiently utilize the coordination channel. We now briefly describe the idle channel selection and rendezvous MAC.

The idle channel selection is done by the OSA "kernel" in our architecture given in Figure 5, which also centralizes the adaptation layer's functions and presents a well-defined interface to the MAC layer. The idle channel for a node M is chosen such that all 1-hop neighbors of M can transmit to M on it. That in turn implies that if *c* is an idle channel, then M and all of its 1-hop neighbors should have a hole in *c*. Thus, each node collects the HIA's from its neighbors, and finds overlapping holes, and assigns the idle channel for itself. In particular, the idle channel is the maximal set of frequency slots such that each slot is a hole for this node as well as each of its neighbors. This process is illustrated in Figure 6, where a 2-slot idle channel is found. As shown there, the idle channel need not be a single frequency slot. It can be a contiguous or non-contiguous set of slots.

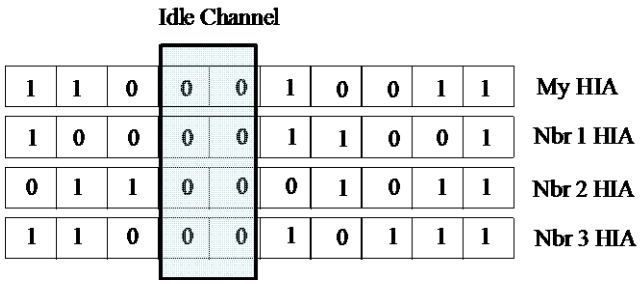

Figure 6: Selection of the idle channel based on own and neighboring HIA

Once the idle channel is selected, it is included in each node's HIP packet so that all its neighbors know.

The motivation for the Rendezvous MAC (R-MAC) is that we expect any dedicated channel to be of low capacity, low enough that without intelligent arbitration, HIP packets may easily saturate it. The core idea behind R-MAC is that if a node knows, when transmitting one packet, of the time and duration of its next packet transmissions, it may append this information to the current packet transmission. In our system, this is true due to the largely periodic nature of HIP transmissions. Neighbors avoid transmitting on the "reserved time". Before transmitting, a node senses the coordination channel, and if it is free, the node transmits the packet. If the channel is busy, R-MAC enters a backoff state and reschedules the packet for a random time into the future. No RTS/CTS or ACK is employed since there are multiple destinations. R-MAC does not *eliminate* collisions, but it reduces it considerably, resulting in adequate utilization.

We note that R-MAC is not the MAC protocol for actual data packets, only for the HIP packets within the coordination channel. The MAC layer protocol for data packets may be either a legacy (conventional) MAC or an OSA-compatible MAC, as shown in Figure 5. When a legacy MAC is used, it is unaware of the OSA adaptation layer and the fact that its transmissions are on dynamically changing bandwidth. Since the idle channel grows or shrinks in accordance with the available frequency slots, this approach still provides improvement over using legacy MAC on a non-OSA system – the legacy MAC sees increased channel capacity.

A much more detailed description of our system and each of the protocols can be found in [22].

4.2 Simulation Results

We have implemented a high-fidelity simulation model, using OPNET, of our OSA system with the protocols described in the previous section. To show how the system exploits unused spectrum, experiments were conducted on a very simple network consisting of 4 nodes in a line, separated by 250m each. Since we wanted to accentuate the negative impact of the coordination channel we increased the size of the coordination channel to 1.5MHz.

Figure 7 shows how the algorithms take advantage of the unused portion of the spectrum. For these plots, we have divided the time into 40ms intervals and consider a frequency band used if a transmission occurred over any part of the 40 ms interval.

The upper left plot shows the way a *legacy* (i.e. current, fixed spectrum allocation) system used the frequency. The primary nodes are assigned 95 MHz of spectrum, so the legacy secondary radios can use only the remaining 5 MHz. The frequency band has been divided in frequency slots of 100 KHz each, and the number of frequency slots occupied by each – and not the actual frequencies – is shown. The lower area represents the primary nodes utilization. For example, in the interval <0, 3> two primary nodes are active occupying 380 slots (i.e. 38 MHz). Similarly, in the interval <16, 21> no primary node is active. Averaging over the interval <1, 21> (i.e. after initialization) and over their assigned frequency bands, the primary nodes' present an utilization of 40%. Looking at the legacy secondary node's bandwidth utilization, we can see the legacy nodes always use the same amount of bandwidth (the unassigned 5 MHz) and are constantly transmitting, with the exception of the interval <0, 1> (initialization) when they are only active for short bursts sending neighbor discovery beacons.

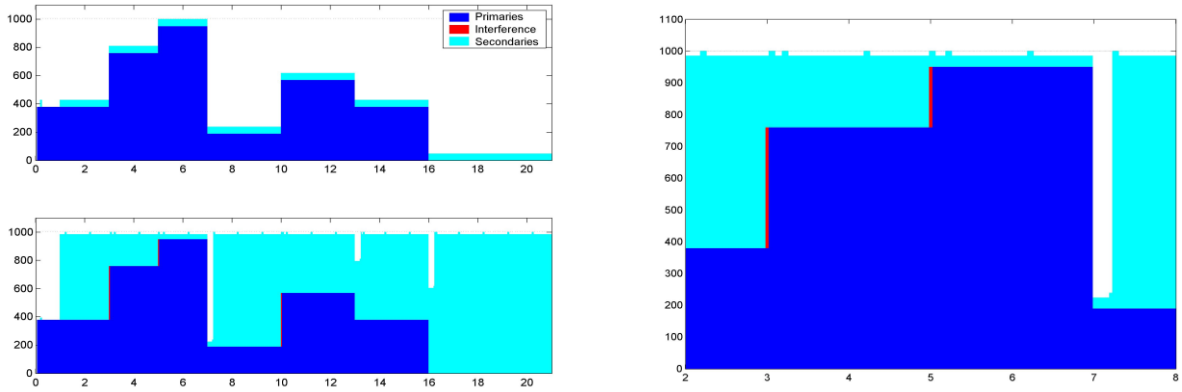

Figure 7: Spectrum utilization by legacy (left upper) and OSA (left lower). Right side is a magnification of the OSA utilization. Bandwidth is in units of number of 100 KHz frequency slots used. This figure is best viewed in color.

The lower left plot illustrates the way OSA radios exploit the available spectrum. After the one-second-initialization period is completed, we see that OSA fills the spectrum gaps almost completely, except for very small periods of time. Dark lines represent the periods where OSA radios interfere with primary nodes. This interference is due to the latency on sensing the channel. It may take up to two sensing intervals (8 ms each) for the radio to detect that a primary node has become active, resulting on interference periods of up to 16 ms per each time a primary node becomes active. However, since our time step is 40 ms, we paint the whole 40 ms interval as "interfered".

The right plot is a magnification of the lower left plot (OSA spectrum utilization) for the interval <2, 8>. This allows a more detailed examination of the OSA system's behavior. First, we notice that our OSA system does not occupy the entire 1000 frequency slots all the time. Typically there will be a small gap, corresponding to the coordination channel's 15 slots. Only when HIP packets are sent (periodically every second and also event-driven upon primary nodes' activation at times 3s and 5s) will OSA occupy the entire spectrum. Thus, the coordination channel and OSA control packets reduces the DATA throughput achieved. Second, we see that upon a primary node's activation (for example at time 3 sec.), OSA transmissions will interfere with primary nodes until their local sensors detect the primary node signal, and before any control packet is sent they will stop transmitting on the primary node's frequency band. This implies that they will not be able to communicate with any neighbor whose IDLE channel contains parts of that band (all the 4 nodes in our example). The OSA nodes will then recompute their IDLE channel and include it together with their new HIA information in a HIP packet that is broadcasted after some random time (to avoid collisions from all opportunistic nodes sensing the primary node's signal at the same time).

The throughput comparison between our system with OSA and a legacy system without OSA as a function of the bandwidth utilization of primary nodes is given in Figure 8. Also given is the throughput achievable by an ideal (best possible) system that is restricted to use QPSK modulation (otherwise the comparison will not be fair). Since we set a large coordination channel of 1.5 MHz, it is expected that when the primary nodes' utilization approaches 100%, the legacy system, able to transmit over 5 MHz, outperforms an OSA system which can only use 3.5 MHz. We note that for a primary node utilization of 40% (typical of current spectrum occupancy measurements [3]), using OSA gives an *order-of-magnitude* performance improvement over legacy systems and is within 35% of the ideal.

While our simple solution for spectrum agility does enable opportunistic spectrum access, it is clearly leaves something to be desired. The need for an a priori dedicated coordination may not be realistic in many environments, and the question is if we can design a protocol that does not require it. It is also clear that CSMA has limitations with spectrum agile protocols and we need to investigate TDMA based algorithms, which is quite challenging. And if TDMA is used, there would be a need to disseminate hole information further than just 1 hop and this brings into question scalability issues.

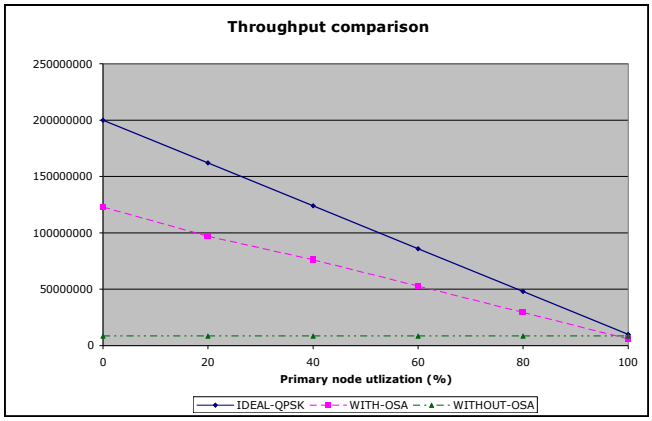

Figure 8: Throughput comparison as a function of primary node utilization with and without OSA

5 Concluding Remarks

The "spectrum scarcity" that is made so much of nowadays is *artificial* – a consequence of an archaic allocation policy that results in severe under-utilization of spectrum. Opportunistic spectrum access has the potential to dramatically increase spectrum utilization by allowing secondary (non-licensed) users to opportunistically re-use primary (licensed) spectrum in an interference-limiting manner. However, realizing this potential requires not only making devices spectrum agile but also policy agile.

We discussed a number of hard problems that arise in the spectrum- and policy-agility of devices, and presented an architecture for opportunistic spectrum access. We formulated a simplified version of the general problem of sensing, disseminating and accessing, and described protocols that harvest the "low hanging fruit" of unused spectrum.

Much work remains to be done. Some problems worthy of further research include:

- Formulating the opportunistic spectrum access problem more rigorously and determining the theoretical capacity bounds in this new operational regime.
- Relaxing many of the assumptions we made: Opportunistic spectrum access when primaries are non-chatty or silent, addressing the hidden node problem, constructing a coordination channel "on-the-fly".
- Dealing with rapidly changing hole information in a large network. This likely requires multi-level compression techniques for the Hole Information Array and managing the dissemination so that it is scalable, yet allows an adequate snapshot of spectrum occupancy.
- Building a prototype for demonstrating end-to-end policy responsive opportunistic spectrum access where

the user modifies the policies and the device changes its behavior to conform to new policy.

This paper represents a first step toward policy responsive opportunistic spectrum access and shows that the *major* problems are tractable and that a system that is spectrum- and policy-agile is feasible. It is clear that opportunistic spectrum access will be key part of the evolution of the Wireless Internet.

Acknowledgments

The authors wish to thank Preston Marshall, DARPA, for some of the ideas in this paper and for support for our XG project. We also thank Daniel Ugarte, Northeastern University for help with the simulations.

References

[1] Federal Communications Commission, "Spectrum Policy Task Force: Report," ET-Docket 02-135, November 2002, http://www.fcc.gov/sptf

[2] Federal Communications Commission, "Facilitating Opportunities for Flexible, Efficient, and Reliable Spectrum Use Employing Cognitive Radio Technologies," *Notice of Proposed Rule Making (NPRM),* Released Dec. 30, 2003.

[3] The New America Foundation and Shared Spectrum Company, "Dupont Circle Spectrum Utilization during Peak hours, http://newamerica.net/Download_Docs/pdfs/Doc_File_183_1.pdf

[4] X. Jing, D. Raychaudhari, "A Spectrum Etiquette Protocol for Efficient Coordination of Radio Devices in Unlicensed Bands," *Proc. PIMRC,* Beijing, 2003.

[5] S. Mangold, K. Challapalli, "Coexistence of Wireless Networks in Unlicensed Frequency Bands," Wireless World Research Forum #9, Zurich Switzerland July 2003.

[6] N. Golmie, R.E. Van Dyck, A. Soltanian, A. Tonnerre, O. Rebala, "Interference Evaluation of Bluetooth and IEEE 802.11b Systems," *Wireless Networks,* 9(3), pp. 201-211. 2003.

[7] M.M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, J. Evans, "DIMSUMnet: New Directions in Wireless Networking using Dynamic spectrum Access," *Proc. IEEE WoWMom, June 2005.*

[8] W. Horne, P. Weed, D. Schaefer, "Adaptive Spectrum Radio: A Feasibility Platform on the Path to Dynamic Spectrum Access," *Fifth Annual International Symposium on Advanced Radio Technologies,* Mar. 2003.

[9] S. Nandagopalan, C. Cordeiro, K. Challapalli, "Spectrum Agile Radios: Utilization and Sensing Architectures," *Proc. IEEE Dynamic Spectrum Access Networks (DYSPAN),* Nov. 2005

[10] G. Ganesan, G. Li, "Cooperative Spectrum Sensing in Cognitive Radio Networks", *Proc. IEEE Dynamic Spectrum Access Networks (DYSPAN),* Nov. 2005

[11] J. Mitola III, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," *PhD Dissertation, Royal Institute of Technology (KTH) Sweden,* May 2000.

[12] J. Mitola III, G.Q. Maguire, "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Communications,* August 1999.

[13] http://www.vanu.com

[14] http://www.gnu.org/software/gnuradio/

[15] http://jtrs.army.mil

[16] M.S. Gast, ``802.11 Wireless Networks: The Definitive Guide'', O'Reilly and Associates, April 2002.

[17] B. Krenik, "Clearing interference for cognitive radio," EETimes Online, http://www.eet.com/in_focus/mixed_signals/showArticle.jhtml?articleID=29100649

[18] T. Clausen et al (ed), "Optimized Link State Routing Protocol" IETF MANET RFC 3626, Oct 2003.

[19] M.A. Padlipsky, "A Perspective on the ARPANET Reference Model," *Proc. IEEE INFOCOM '83*. 1983. San Diego, CA.

[20] BBN Technologies, "The XG Vision, version 2.0", Request for Comment (RFC), accessible at http://www.ir.bbn.com/projects/xmac/vision.html

[21] BBN Technologies, "XG Policy Language Framework version 1.0," Request for Comment (RFC) accessible at http://www.ir.bbn.com/projects/xmac/pollang.html

[22] C. Santivanez et al, "XOSA: Exploiting Opportunistic Spectrum Access for Wireless Ad Hoc Networks," BBN Technical Report No, 8410, Dec. 2004.