# Opportunistic Channel Access Using Reinforcement Learning in Tiered CBRS Networks

Matthew Tonnemacher*‡, Chance Tarver†, Vikram Chandrasekhar*, Hao Chen*, Pengda Huang*, Boon Loong Ng*, Jianzhong (Charlie) Zhang*, Joseph R. Cavallaro†, and Joseph Camp‡

*Standards and Mobility Innovation Lab, Samsung Research America, Richardson, TX 75082
†Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005
‡Department of Electrical Engineering, Southern Methodist University, Dallas, TX 75205

*Abstract*—The upcoming deployments of devices on the new 3.5 GHz, Citizens Broadband Radio Service (CBRS) is expected to enable innovation by lowering the barrier to entry into LTE and other technologies. With a three-tiered spectrum-sharing solution, the CBRS band promises to allow coexistence of federal incumbent users, priority licensees, and general users. While there have been many works studying cellular traffic offloading to unlicensed bands (e.g., Licensed Assisted Access) or minimizing interference in Cognitive Radio Networks, there has been comparatively little work on maximizing spatial reuse of spectrum in a shared spectrum CBRS network. Hence, the essence of this work is to leverage listen-before-talk (LBT) schemes over a CBRS network for increasing the spatial reuse at secondary (general) users while minimizing the interference footprint on higher-tier (incumbent) users. In this work, we propose LBT schemes that allow opportunistic access to licensed, CBRS spectrum and test our LBT schemes on a custom testbed with multiple software-defined radios and a real-time signal analyzer. We find that by allowing LBT spectrum sharing in a two carrier, two eNB scenario, we see upwards of 50% user perceived throughput (UPT) gains for *both* eNBs. Furthermore, we examine the use of Q-learning to adapt the energy-detection threshold (EDT), combating problematic topologies such as hidden and exposed nodes. When adapting the EDT of opportunistically transmitting nodes, we see up to 350% gains in average secondary node UPT in certain difficult topologies with merely a 4% reduction in primary node UPT.

## I. Introduction

Additional spectrum availability and increased efficiency in the use of existing resources are needed to accommodate the rapidly increasing density and subsequent data demands of wireless devices around the world. This need was recognized by the President of the United States in 2010 when he called for an additional 500 MHz of wireless spectrum to be made available within ten years [1]. It was later a finding of the President's Council of Advisors on Science and Technology (PCAST) that sharing the spectrum would be essential to meet the wireless challenges that are currently seen [2]. This recommendation was embraced by the Federal Communications Commission (FCC) in a notice of proposed rulemaking where they suggest a three-tiered, database-managed, spectrum-sharing scheme for the 3550 – 3700 MHz band, which would become the Citizens Broadband Radio Service (CBRS) [3]. Even before
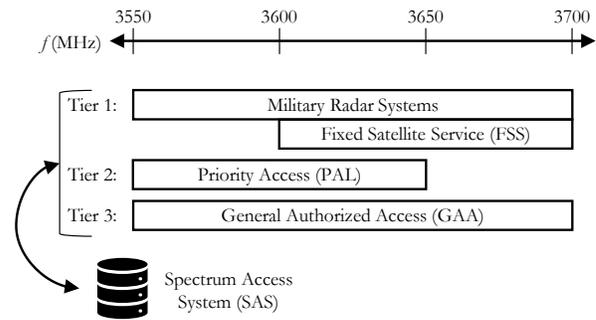
Fig. 1. Three-tiered spectrum sharing with central database management in the CBRS band.

the presidential memorandum, these themes could be seen in other spectrum policies such as in TV White Space where unlicensed devices can utilize unused portions of ultra-high-frequency (UHF) bands that were licensed for TV stations [4]. The CBRS band is quickly developing and will see deployment in commercial devices soon with the launch of the Qualcomm X20 LTE modem, which supports operations on that band [5].

In the standard CBRS architecture, there is a three-tiered system managed by a dynamic database called the Spectrum Access System (SAS). This is illustrated in Figure 1. The top tier consists of incumbent users, the second tier consists of Priority Access Licenses (PALs), and the bottom tier is for General Authorized Access (GAA). It is expected that many users will operate in a pseudo-unlicensed fashion on the GAA tier, and in competitive markets, some carriers may choose to purchase a PAL license to ensure a minimum QoS.

At the same time that the federal government is opening up new bands for shared use, there is increasing congestion on the unlicensed bands. For example, 802.11 Wi-Fi and Bluetooth devices densely occupy the unlicensed 2.4 GHz and 5 GHz industrial, scientific and medical (ISM) radio bands. Cell providers are increasingly interested in using free, unlicensed spectrum to supplement their licensed networks, exacerbating this crowding. The idea of using unlicensed bands to supplement licensed networks has been pushed to multiple standards such as 3GPP's licensed assisted access (LAA) [6], LTE-U [7], and MulteFire [8].

It is clear that spectrum sharing will become more prevalent in future medium access policies. The cornerstone of these policies is the requirement for a way to manage the sharing. This is traditionally done through contention-based protocols such as a listen-before-talk (LBT) scheme like Carrier-Sense Multiple Access (CSMA) for 802.11 devices. In these schemes, the channel must be measured to be idle for a certain amount of time before it is accessed. In this paper, we present and compare two LTE-based LBT schemes for use in the CBRS infrastructure. Specifically, we examine the trade-off between GAA user gain and PAL user interference.

In the current release of the CBRS standard, there is no formal mechanism for GAA users to access PAL allocated channels. Rather, it is framed such that sharing could occur without explanation for how it should be done. When it comes to opportunistic random access, LBT schemes are a proven method that can be used to allow GAA users to access PAL spectrum opportunistically. Ideally, under such a scheme, the secondary GAA nodes (SNs) gain additional spectrum resources, increasing throughput, while the primary PAL nodes (PNs) are unaffected. While it is impossible to achieve such perfect coordination due to the inability for secondary nodes to predict future PN packet arrival, the network operation would then be similar to Wi-Fi.
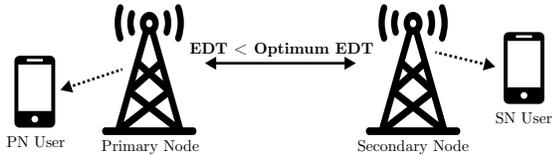


(a) Exposed node case with a PN inside sensing range of SN, unnecessarily preventing SN transmission.
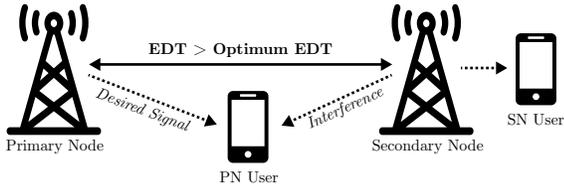


(b) Hidden terminal case where a PN user is impacted by SN interference, despite PN being outside of SN sensing range.

Fig. 2. Example primary node (PN)–secondary node (SN) topologies where an adaptive EDT could benefit the network.

If LBT is adopted, a statically-defined policy for all devices will likely be vulnerable to poor performance in common topologies that can arise such as hidden and exposed terminals, shown in Figure 2 [9]. These situations are especially heinous in tiered access topologies, where it is imperative that the PNs and incumbents receive minimal impact from opportunistic SNs. One of the most popular random access schemes, Wi-Fi, has historically had problems with hidden/exposed terminals. However, the Wi-Fi medium access control protocol, CSMA/CA with RTS/CTS, effectively mitigated these issues by using virtual carrier sensing, which has both transmitter/receiver broadcast a handshake to alert other nodes of a pending transmission.

In the CBRS tiered architecture, however, such a scheme is not possible, as the primary node does not necessarily engage in LBT behavior. Thus, if LBT is to be used for opportunistic access, hidden and exposed terminals must be dealt with in another way. To mitigate this issue, we propose a novel reinforcement Q-learning technique to adapt an energy-detection threshold (EDT) for secondary nodes in a shared spectrum environment. We show that by using machine learning, we can increase both SN and PN gains over LBT schemes that use a static EDT.

In this paper, we enable additional sharing and improved throughput by proposing listen-before-talk (LBT) schemes that allow opportunistic access to licensed LTE spectrum and test our LBT schemes on a custom testbed with multiple software-defined radios and a real-time signal analyzer. In doing so, we have the following five contributions: (*i.*) We design and evaluate two LBT schemes to be used in CBRS networks for PN-SN spectrum sharing. (*ii.*) We show that while one scheme ends up having higher performance, both schemes greatly improve SN UPT with a minor decrease in PN UPT. (*iii.*) We find that the decreased PN UPT is a function of PN traffic load and problematic network topologies between PN/SN users. (*iv.*) To reduce the negative consequences of spectrum sharing on the PN, we formulate a novel Q-learning algorithm that adjusts SN opportunistic access via learning an optimal EDT for carrier sensing. (*v.*) By using average and differential PN buffer occupancy as the environmental observations, we find the detriment to PN UPT from spectrum sharing can be greatly reduced.

## II. CBRS OVERVIEW

The CBRS band is a three-tiered, shared-spectrum platform managed dynamically by the Spectrum Access System (SAS) as shown in Figure 1. The incumbent users in the band include Department of Defense (DoD) radar systems as well as Fixed Satellite Service (FSS). To protect this tier, there will be an Environmental Sensing Capability (ESC). ESC nodes will monitor the band and notify the SAS in the case that an incumbent becomes present. The SAS will then move lower-tier users to unoccupied channels.

In the next tier, users may purchase spectrum in 10 MHz chunks via an auction to become PALs. Unlike other traditionally licensed bands, purchasing a channel does not tie the licensee to a specific frequency. Instead, the SAS can dynamically assign the PAL to any channel in the band. Additionally, the license is only valid for three years and is for a census tract (a geographic area that is roughly sized according to a fixed populous) as opposed to longer terms and larger areas seen in other bands. In a given census tract, only 7 PAL licenses will be issued, guaranteeing that a minimum of 80 MHz of the band is always available for GAA use.

The bottom tier is designed to be similar to an unlicensed band to enable a low-cost, flexible solution for a large group of potential applications. The SAS may assign GAA users to any channel in the full 150 MHz band as long as they do not interfere with a user of higher priority.

All devices are managed by the SAS, which is the true innovation of this band. It is motivated by the TV White Space database system, but it is designed to be more dynamic. The primary functions of the SAS include determining available frequencies and assigning them to devices, determining maximum power levels for devices, enforcing exclusion zones, protecting PALs from GAA users, and facilitating coordination between GAA users. While dynamic, it is an explicit goal of the SAS not to micromanage the spectrum, leaving such fine-grained tuning to the operators.

One of the main advantages of the CBRS band is the low barrier to entry leading to many possible use cases. The GAA tier can be accessed in an almost unlicensed-like fashion at no cost by registering with the SAS. This creates many opportunities for neutral host or private LTE networks. For example, an arena could deploy a neutral-host infrastructure on 3.5 GHz using the GAA tier. Enterprise campuses could deploy private LTE networks that cover their entire campus. In addition, existing cell providers could augment their networks by utilizing this band as a GAA user. In select, competitive markets, users may choose to purchase a priority license to reduce interference and guarantee a minimum quality of service. It is also possible to use the CBRS band for a fixed-wireless, infrastructure-type deployment.

Two major organizations are devoted to developing standards on the CBRS band. The first is the Wireless Innovation Forum (WInnForum). The WInnForum is developing standards for certifying devices and developing a SAS protocol and is not tied to any specific wireless technology. The other major organization is the CBRS Alliance. The CBRS Alliance is developing standards for operating LTE on the CBRS band, which they recently have named OnGo [10].

The CBRS Alliance imposes additional structure to be able to facilitate coordination amongst GAA users. In [11], a coexistence group (CxG) is created which is managed by a coexistence manager (CxM). Multiple GAAs can be a part of a CxG, and the SAS will allocate a pool of spectrum to the CxG instead of to individual CBSDs. The CxM can then enable more fine-grained sharing among the members of its group.
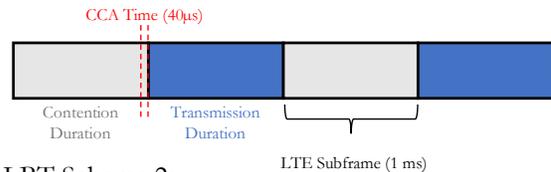
## III. LISTEN BEFORE TALK

Although the CBRS band allows for spectrum sharing between multiple tiers, there is currently no mechanism to facilitate this sharing. It could be possible for GAA users to take advantage of spectrum holes in the PAL spectrum. To do this, the GAA users need to be able to sense the spectrum. Wi-Fi, one of the more popular random access schemes available, has used LBT for sharing the spectrum between multiple users to great success. In fact, a version of LBT is essential in any shared-spectrum environment and is legally required in many countries for operation on an unlicensed band.

LBT has even been adopted in some LTE standards, for example 3GPP's LAA specification. The LBT scheme used in LAA is as follows. Whenever a device needs to transmit, it needs an initial clear-channel assessment (CAA). It must sense that the channel is idle for at least 34 $\mu$s. If so, it can transmit

for the length of one transmit opportunity (TxOP). If there is additional traffic to send, an exponential random backoff mechanism is used, similar to Wi-Fi [12].

However, in the context of CBRS Alliance LTE devices, this mechanism may not be ideal. Given the existence of CxMs in CBRS, it is possible to tailor the LBT scheme specifically for LTE devices. Although there could be many ways for performing LBT, we develop and compare the performance of two schemes that seem to be the natural choices: sensing at the end (Scheme 1) or the beginning (Scheme 2) of a subframe. These are shown in Figure 3, and each is evaluated below. Similar schemes have also been considered for LTE/Wi-fi coexistence in [13]. For the developed LBT schemes, we consider only the downlink, assuming that devices are operating similarly to LAA in Release 13 of LTE where the CBRS carrier is considered as a supplemental downlink carrier [12].
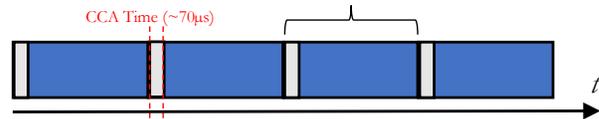


Fig. 3. Comparison of two proposed LBT schemes for CBRS.

### A. LBT Scheme 1: Subframe Puncturing

In LBT Scheme 1, an entire subframe functions as a contention window. However, only the last 40 $\mu$s in a subframe is used for the CCA. If the channel is determined to be idle, the next subframe is used for transmission. Using this scheme has an advantage of not altering the structure of a subframe. However, it results in, at most, a $50\%$ transmission duty cycle. Moreover, the scheme could measure the channel to be idle during the contention window only for the primary node to start transmitting on the next subframe, leading to a collision. As this is a tiered access system, collisions with the PN need to be avoided at all costs.

### B. LBT Scheme 2: Symbol Puncturing

In LBT Scheme 2, the potential transmitter always senses in the first symbol of every subframe. If successful, we transmit in the remaining 13 symbols of the subframe. This has the advantage of sensing at the time that a PN would come in, reducing the likelihood of a collision, given synchronization of subframe boundaries between PN and SN. This scheme sacrifices a single symbol out of each subframe. However, the first symbol in LTE is typically used for the control channel, so this scheme may require altering the subframe structure, though this omission of the control channel may be inconsequential in cases where cross-carrier scheduling is used.

## C. LBT Scheme Comparison

There are apparent differences between the LBT schemes by construction. Given that the best-case duty cycle for Scheme 1 and Scheme 2 are 50% and 92%, respectively, Scheme 2 is preferable. However, there is a tradeoff between performance and implementation complexity between the two schemes. With Scheme 1, we do not require any modification to a subframe structure.

To evaluate the performance of these LBT schemes in a CBRS-like framework, we first simulate their operation in various scenarios. In Figure 4, we show each scenario considered to help understand the effect of the LBT scheme on both the PN and SN. For a baseline, we consider in Figure 4a, Scenario 1, where two operators are operating on their carriers without any sharing. They operate in an "On/Off" mode, meaning they are only transmitting if there are connected UEs with traffic. In Figure 4b, Scenario 2, we test the result when both operators engage in mutual sharing onto each other's carrier. In Figure 4c, Scenario 3, we consider the case that a single operator is on its carrier and performs LBT on another carrier that is entirely unoccupied. This case represents an upper bound on gains for an SN. In Figure 4d, Scenario 4, we share on a single component carrier to see the realistic gains for Op. 2 when sharing and the effect it has on the PN, Op 1.

For the sake of comparing the performance of each LBT scheme, we considered the scenarios with no sharing (Scenario 1) and sharing with a PN (Scenario 4) from Figures 4a and 4d as before and after cases. We report the change in the user-perceived throughput (UPT) for operator 1 (Op. 1) and operator 2 (Op. 2). In this figure, UPT is given by

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1}{P_{total}} \left[ \sum_{j=1}^{P_{served}} \frac{M \cdot r_{ij}}{t_{ij}} + \frac{b_i}{t_{serving,i}} \right] \quad (1)$$

where $N$ is the number of UEs served by the eNB, and $i$ indexes the UEs. $P_{total}$ is the total number of packets, elaborated by $P_{total} = P_{served} + P_{serving}$, where $P_{served}$ and $P_{serving}$ are the number of packets served and being served, respectively. $M$ is the number of bits per packet, $r_{ij}$ is the ratio of successfully transmitted bits over all bits in the packet to UE $i$ for packet $j$, and $t_{ij}$ is the time taken to send the same packet. $b_i$ is the number of bits sent to UE $i$ as a partial packet still in flight, and $t_{serving,i}$ is the time spent by the packet.

We simulated our spectrum-sharing scheme using MATLAB by reusing the 3GPP LAA evaluation assumptions for an indoor scenario [14]. Figure 5 shows this topology, and the rest of the simulation settings are as follows:

- Two operators with four small cells each in a single floor building (Figure 5)
- 18 dBm TX power
- 10 randomly distributed UEs per operator
- -72 dBm EDT
- 20 MHz system bandwidth
- 10 drops simulated
- 20,000 subframes per simulation

Figure 6 shows the simulation results. For each test, we show the mean, median, 5th, and 95th percentile to illustrate



(a) Scenario 1 – No Sharing    (b) Scenario 2 – Mutual Sharing

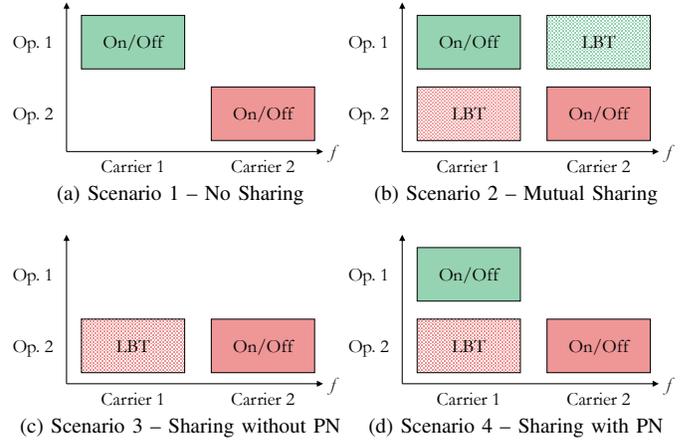(c) Scenario 3 – Sharing without PN    (d) Scenario 4 – Sharing with PN

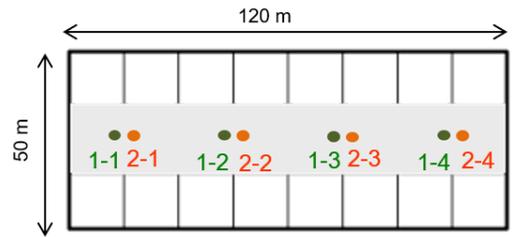Fig. 4.  Simulation scenarios considered.



Fig. 5.  3GPP indoor scenario for LAA coexistence evaluations with two operators and four nodes per operator [14].

the variance across all simulations. In Figure 6a, we see a maximum increase in UPT for the SN of 40%. However, for the same case, there can be a 10% reduction in UPT for the PN. In Figure 6b, we see a nearly 80% increase in performance for the SN with a similar drop in performance for the PN. Overall, LBT Scheme 2 performed significantly better for both the PN and SN. So, in the next subsection, we select LBT Scheme 2 for use in additional simulations to determine the possible spectrum-sharing gain.

### D. Simulations with Static EDT

Figure 7 shows the results for simulating LBT Scheme 2 across different spectrum sharing scenarios for two different traffic arrival rates to see the effect on UPT. In Figure 7a, the average traffic arrival rate was 0.5 MB/s for PNs and SNs. The first cluster of results is the Scenario 1 baseline from Figure 4a, where both operators are on their carriers without sharing the spectrum. The second cluster of results shows Scenario 2 from Figure 4b, where each operator mutually shares its primary spectrum with the other operator. Here, we can see that each operator experiences an increase in the mean UPT by about 25%.

To see a (coarse) upper bound on maximum achievable spectrum sharing gain, the third cluster shows Scenario 3 from Figure 4c. Here, we can see that an operator can achieve a maximum of 133% gain when adopting a spectrum sharing scheme.
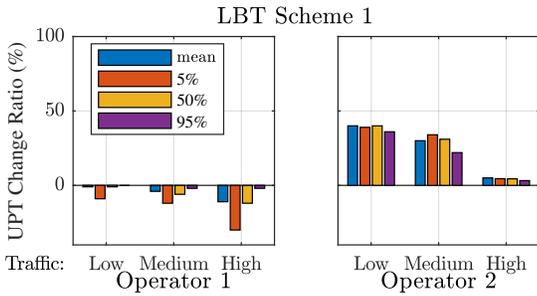
(a) UPT change when 1 Op. shares using subframe puncturing LBT (Scheme 1).
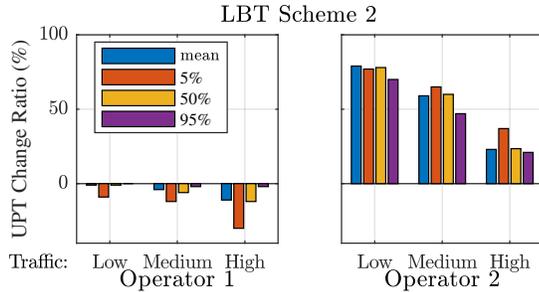


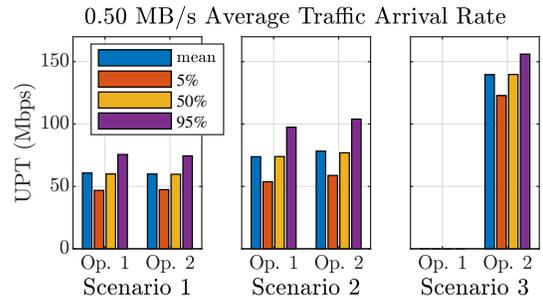(b) UPT change when 1 Op. shares using symbol puncturing LBT (Scheme 2).

Fig. 6. Performance of each LBT scheme for different traffic loads.



(a) UPT for spectrum-sharing scenarios when both operators use a relatively slow average traffic arrival rate of 0.50 MB/s.



(b) UPT for spectrum-sharing scenarios when both operators use a relatively fast average traffic arrival rate of 1.05 MB/s.

Fig. 7. Performance of LBT for two different average traffic arrival rates.

The simulation is repeated for the case of a higher-traffic arrival rate of 1.05 MB/s in Figure 7b. Similar to the results for the slower traffic rate, when each operator engages in a mutual sharing as in Scenario 2, each operator sees an improvement in UPT. For the higher traffic case, the gains are doubled with approximately 50% increase in UPT for both operators.
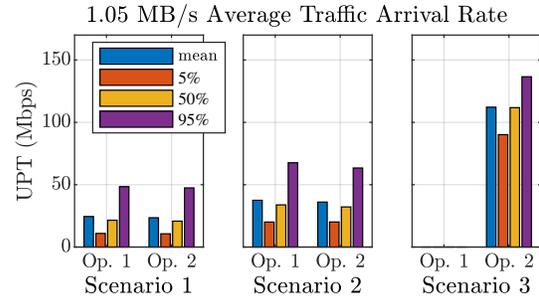
We then use LBT Scheme 2 and consider the performance for various EDTs. In Figure 8a, we show the results when we consider different EDTs for Scenario 2 from Figure 4b. Here, we see that there is an "optimal" EDT around -52 dBm. These results highlight that for different scenarios, there may be different "ideal" EDTs.

In Figure 8b, we plot the UPT vs. EDT for Scenario 4 from Figure 4d. In this figure we can see the effect of a higher EDT. Here, a higher EDT at Op. 2 implies more frequency of channel access at the expense of increased downlink interference at Op. 1. As the EDT becomes larger, UPT of Op. 1 decreases, and UPT of Op. 2 increases.

It is worth noting that the UPT decreases for the PN are significantly smaller in the low-traffic load case and more significant in the high-traffic load case for both schemes. Ideally, in situations where the PN has a high-traffic load, the SN would behave more passively when on the PN's carrier. In the next section, we will explore the use of machine learning in adjusting SN EDT to improve LBT performance in hidden and traffic-heavy node scenarios. Using our algorithm, we show that scenario-specific, poor-LBT performance can be significantly reduced.
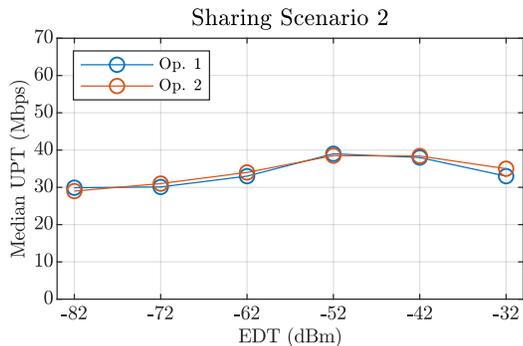
### E. Shared-Spectrum Testbed

To further evaluate the LBT schemes outlined so far, we developed a shared-spectrum testbed shown in Figure 9. The testbed consists of the following:

- 4 USRP SDRs with the possibility of including more for larger tests with many nodes and UEs.
- Ideal ethernet backhaul via python/UDP for statistics
- Real-time spectrum analyzer

The nodes of the testbed can be arranged to emulate various topologies such as hidden/exposed nodes with additional possibilities of including mobility. This allows us to see how well the LBT schemes behave under real channel conditions where there may be many reflections, etc. that change quickly.

We modified the National Instruments LTE Application Framework to implement our LBT schemes. Figure 10 shows an example result from the real-time signal analyzer spectrogram. Here, we use LBT Scheme 1 where the SN uses a subframe for contention and a subframe for transmission. We show Scenario 2, where each operator can engage in spectrum sharing. In this case, Op. 2 is under heavy load while Op. 1 is not. Op. 2 augments its services by aggregating onto Op. 1's spectrum while Op. 1 has no traffic. For this demo, we restrict each operator only to occupy half of each 20-MHz carrier so that we can easily distinguish between the operators on the spectrogram. In this figure, red is a higher power, and green is a lower power. We can see that Op. 2 is using Carrier 1 with a 50% duty cycle. We can also see that when Op. 1 begins using its spectrum,

(a) Mutual sharing on 2 carriers, Scenario 2.
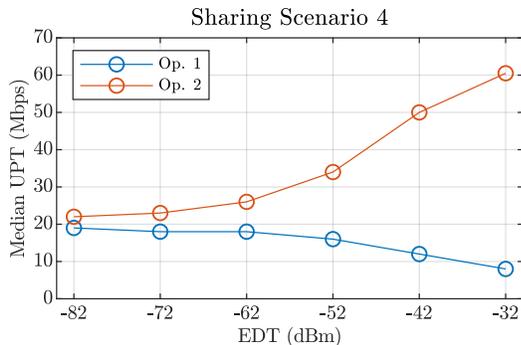


(b) Sharing on 1 carrier, Scenario 4.

Fig. 8. Median UPT vs EDT for tests with a 20 MHz system bandwidth and a traffic arrival rate of 1.05 MB/s.
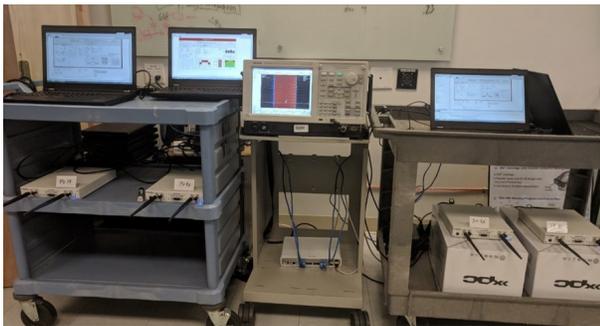


Fig. 9. Photograph of the shared-spectrum testbed. Four USRPs connected to host PCs running LabVIEW Communications with a real-time signal analyzer.

Op. 2 detects this and waits for the carrier to become available again before transmitting.

In this testbed, the SN synchronizes with the timing of the PN of a channel. Using the existing LTE synchronization signals, the SN detects the subframe boundaries, measures the energy in the channel at the appropriate time, and then if the PN is decided to be absent based on the EDT, transmits to its users during the available TXOP. This result highlights the feasibility for the SN to sync to a PN for performing LBT in real-time.

## IV. REINFORCEMENT LEARNING

While LBT schemes have been successfully implemented in 802.11 standards to great effect, certain network situations can result in poor performance. While other applications of LBT
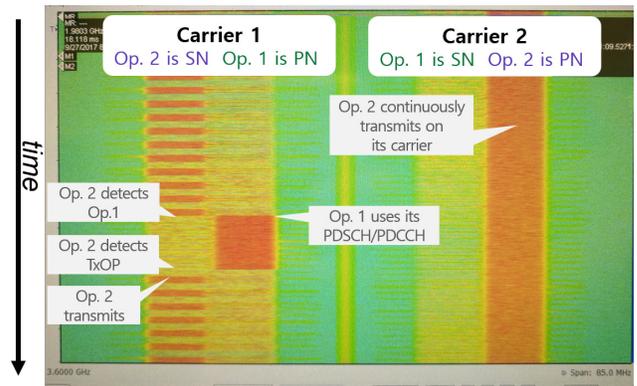


Fig. 10. Real-time signal analyzer spectrogram for LBT Scheme 1, Scenario 2 from the testbed.

may have ways of reducing network problems in topologies that include hidden and exposed terminals such as the collision avoidance in CSMA, similar schemes are not applicable when applying LBT to LTE technology. In this section, we build a machine learning framework that dynamically adjusts the EDT of the SN to enable more efficient spectrum sharing in the PN's carrier.

### A. Reinforcement Learning Primer

Before we detail our algorithm, we first introduce the general reinforcement-learning strategy. Roughly speaking, reinforcement learning addresses the general problem of learning from interaction to achieve a goal. The learner and decision maker is called the agent. Everything outside the agent that the agent interacts with is called the environment.

Agents interact with the environment via actions. For each action, $\alpha$, that the agent executes, it influences the state of the environment and receives an evaluative feedback, or reward, $r$. This reward is used to learn/adapt its subsequent actions, should it encounter the same state in a subsequent time slot. We define periodic time intervals, $t = 0, T, 2T, ...,$ in which each agent represents its observation, $o$, of the surrounding environment at time $t$ as a state $s \in S$, where $S$ designates a finite set of environmental states. In summary, at each step $t$:

The agent:
- Executes action $\alpha_t$
- Receives observation $\alpha_t$ of $s_t$
- Receives reward $r_t$

The environment:
- Receives action $\alpha_t$
- Emits observation $o_{t+1}$ of $s_{t+1}$
- Emits scalar reward $r_{t+1}$

At each time step, the agent implements a mapping from states to probabilities of selecting each possible action. This mapping is called the agent's policy $\pi_t$, where $\pi_t(s, \alpha)$ is the probability that $\alpha_t = \alpha$ if $s_t = s$. Reinforcement learning methods specify how the agent changes its policy as a result of its experience. The agent's goal is to maximize the total amount of reward it receives over the long term.

One popular reinforcement-learning algorithm is Q-learning [15]. This model-free learning strategy can be used to learn an optimal decision policy for any Markov decision process. We adopt Q-learning with the objective to minimize interference at an incumbent or PN due to spectrum sharing with an opportunistic SN. In this scenario, each SN acts as an agent adapting its action in response to the reward obtained for its previous action.

### B. Target Improvement Areas

Our objective is to use reinforcement learning to assist the SN in harvesting unused bandwidth from the PN in an optimal fashion. Specifically, we leverage Q-learning to dynamically adjust the SN's EDT to maximize network UPT while subsequently minimizing the impact on PN UPT. We identify two scenarios in which EDT adjustment can mitigate poor LBT performance:

*a) Hidden Terminals:* In a hidden terminal topology illustrated in Figure 2b, a UE served by the primary node potentially sees a significant interference if a secondary node transmits at the same time. If the queue size at the PN increases, a possible reason is because of interference from a (hidden) SN. In this scenario, the network would benefit if the SN had a more conservative EDT.

*b) Load Adaptation:* The SN node opportunistically adapts to fluctuations in offered traffic at PNs. If the PN traffic load is low, it may be able to use lower modulation and coding schemes (MCS) while maintaining a similar quality of service. By using more robust coding, higher interference can be tolerated without an increase in packet loss. Thus, the SN EDT can be reduced, allowing for more aggressive SN behavior, depending on the distance of the PN. Alternatively, if the PN traffic load is high, the SN EDT should be increased to prevent interference, even if the PN is further away, allowing higher PN MCS schemes to be used.

With these scenarios in mind, we now present our Q-learning algorithm along with scenario-specific results.

### C. Q-Learning Algorithm Description

In designing the reinforcement-learning algorithm, our objective is to determine a policy (sequence of state/action pairs) by which the agent (SN eNB) adapts its EDT based on observations taken during the latest epoch to maximize long-term rewards. In our setup, we assume that the PN shares transmit buffer occupancy/queue length information with the overarching CBRS network architecture, making this information available to the SN. In turn, the SN uses this information as the environmental observation for the reinforcement learning. This assumption is based on the CxG structure that is present in the CBRS Alliance. The buffer occupancy and queue length information could be shared through this mechanism if added to the CBRS standard. Since the CBRS standard is still new with no commercial deployments, this could be feasible.

The basic Q-learning implementation is as follows. Let epoch $m$, with duration $T$, refer to time interval $mT \leq t < (m+1)T$. The epoch duration, $T$, needs to be long enough (e.g., 10s - 100s of sub-frames) to avoid adapting to short-lived flows. At time $t = mT$, the agent chooses an action which maximizes its Q-table. At time $t = (m+1)T$, the agent receives the observation of the environment state from its last action, receives associated reward, updates the Q-table, and then chooses an action $\alpha_{m+1}$ for epoch $m + 1$. Figure 11 depicts this iterative process.
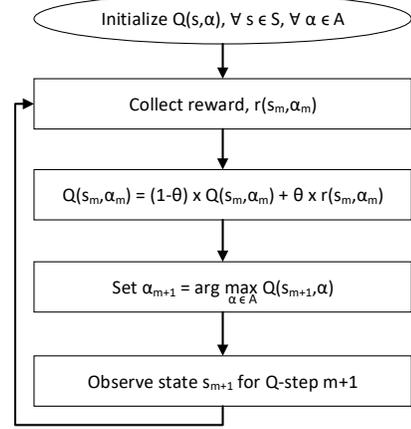


Fig. 11.   Overview of Q-learning algorithm.

Given state space $S$, the environment lies in one of two states $s \in \{1, 2\}$ defined in Table I. Here, $L_m$ is the instantaneous PN transmit queue size at the end of epoch $m$, and $\gamma_1$ is a threshold used to differentiate high and low traffic loads. $\gamma_1$ selection can be used to adjust the relative weight between PN and SN users. We define these states such that there are binary light/heavy traffic conditions to reduce variables in our performance evaluation. However, this definition can easily be extended to multiple states if it is necessary to define more nuanced packet load conditions by defining multiple thresholds.

TABLE I
STATES FOR THE Q-LEARNING

| State | Average Primary Node Queue Size | Comment |
|---|---|---|
| 1 | $0 \leq L_m < \gamma_1$ | Primary node traffic load is light |
| 2 | $L_m \geq \gamma_1$ | Primary node traffic load is heavy |

The agent is rewarded or punished according to the intuitive guidelines listed in Table II. In general, a positive reward is given if the PN's state improves or if a higher EDT threshold is chosen at the SN without a negative impact on the PN. Consequently, a negative reward is given in cases where the PN transitions to a worse state or if the SN chooses a low EDT value without any benefits.

More specifically, we define numerical rewards according to the state transition and the average buffer occupancy over the previous epoch, $B_m$. Table III shows the detailed reward scheme used by our Q-learning algorithm. Here, $\gamma_2 \in (0, 1)$ is the buffer occupancy threshold, $\gamma_3$ is the action threshold (EDT setting threshold in dBm), and $\gamma_4$ is the scalar reward value. Each is a tunable parameter that can help control the Q-learning to better tailor it for various goals and constraints. $Z_m$ can be considered as a soft reward when outcomes are between actionable thresholds.

TABLE II
REWARD INTUITION

| Reward | Conditions |
|---|---|
| Positive | • The state remains in 1, and the last action was to choose a high EDT value (e.g. -62 dBm).<br>• The state changes from 2 to 1 following the last epoch. |
| Negative | • The state remains in 1, and the last action was to choose low EDT value (e.g. -77 dBm).<br>• The state changes from 1 to 2 following the last epoch, and the buffer occupancy is large. |

TABLE III
REWARDS FOR THE Q-LEARNING

| $(s_m, s_{m+1})$ | Reward $r(s_m, \alpha_m)$ |
|---|---|
| $(1,1)$ | • $\gamma_4$, if $B_m \leq \gamma_2$ and $\alpha_m \geq \gamma_3$<br>• $-\gamma_4$, if $B_m \leq \gamma_2$ and $\alpha_m < \gamma_3$<br>• $Z_m$, otherwise |
| $(1,2)$ | • $\gamma_4$, if $B_m \leq \gamma_2$ and $\alpha_m \geq \gamma_3$<br>• $-\gamma_4$, if $B_m \leq \gamma_2$ and $\alpha_m < \gamma_3$<br>• $-\gamma_4$, if $B_m > \gamma_2$ |
| $(2,1)$ | • $0$, if $B_m \leq \gamma_2$<br>• $\gamma_4$, otherwise |
| $(2,2)$ | • $0$, if $B_m \leq \gamma_2$<br>• $Z_m$, otherwise |

As mentioned earlier, the selection of $\gamma_1$ plays an important role in how the learning algorithm behaves between the thresholded values, with higher values allowing the SN to be more responsive to changes in the PN queue length. These threshold values need to be tuned experimentally for different deployments, as there is no absolute rule for how they should be set. In general, each threshold contributes in one way or another to how quickly the algorithm adapts to changes. Depending on the specific situation, more or less rapid responses could be advantageous.

The agent updates the Q-table after each action according to Equation 2. Here, $\theta \in (0,1)$ is a discount factor that is used to control the importance of the reward, $r(s_m, \alpha_m)$, in terms of updating the Q-function. Larger $\theta$ values will prioritize longer-term reward, while a lower $\theta$ applies more weight to the next iteration reward.

$$Q(s_m, \alpha_m) = \theta Q(s_m, \alpha_m) + (1 - \theta) r(s_m, \alpha_m) \quad (2)$$

The next action is chosen at each epoch according to the probability distribution in Equation 3.

$$P(\alpha_{m+1}) = \begin{cases} 1 - \epsilon, & \arg\max_{\alpha \in A} Q(s_{m+1}, \alpha) \\ \epsilon, & \mathrm{rand}(\alpha \in A) \end{cases} \quad (3)$$

Here, $\epsilon \in [0,1]$ is an exploration parameter, allowing for occasional random actions to be taken. In general, allowing for exploration prevents the learning algorithm from getting locked into suboptimal operation by filling in more of the Q-table than would occur otherwise. Furthermore, the exploration

probability can be reduced over time as more iterations of the algorithm have occurred.

### D. Simulations with Adaptive EDT

To evaluate the performance of the adaptive EDT, we perform system simulation in MATLAB. We examine several distinct scenarios to examine how the performance of LBT compares with and without the Q-learning based adaptive EDT.

*1) Hidden Node – Mitigating Interference:* For the first simulation, we consider the topology shown in Figure 12. All of the PN UEs are placed equidistant from the PN and SN so that if the SN is transmitting, the SINR that they would receive would be approximately 0 dB. This emulates a hidden-node case, where the distance between the PN and SN is far greater than the distance between the SN and the PN UEs. In Figure 13, we compare PN UPT with a fixed, -62 dBm EDT to the adaptive EDT using Q-learning. In this figure, the upper bound on PN transmission is the situation in which there is no secondary user; thus, the PN can transmit interference free. We can see that when using a fixed EDT, the PN UPT drops drastically as expected in a hidden node scenario. However, by allowing the EDT to increase in response to the detection of increasing buffer occupancy at the PN, the penalty received by the PN is greatly reduced.
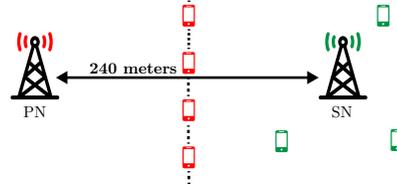


Fig. 12. Hidden node test topology where the PN UEs are equidistant from the PN and SN.
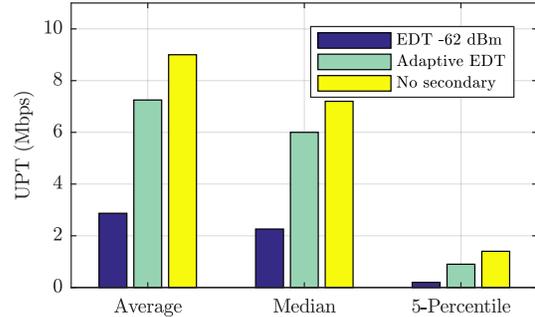


Fig. 13. PN performance when its UEs are hidden terminals to the SN. An adaptive EDT in the SN allows the SN to reduce its interference to the PN.

*2) Adapting to PN Load:* In the next set of simulations, we have four nodes with two operators in a shared-carrier topology, as shown in Figure 14. In this scenario, the PN load is effectively doubled, as each SN needs to defer to two different PNs. The UE distribution for each node is randomized in proximity around each node. We present the simulation results in Figure 15. We can see that there is not a significant change in the PN UPT for any scheme, as they all approach the upper bound. For the case with no secondary node and the case with a -72 dBm fixed EDT, this is to be expected,

as the SNs will be able to sense the PN and defer for nearly every PN transmission. However, when using an adaptive EDT, the SN UPT is increased by a factor of four. This is because the reinforcement learning can adaptively shrink when buffer occupancy remains low at the PN, by taking advantage of momentarily light traffic loads and/or transmissions to UEs located further away from the SN.
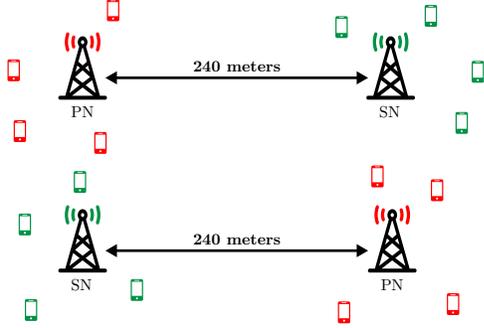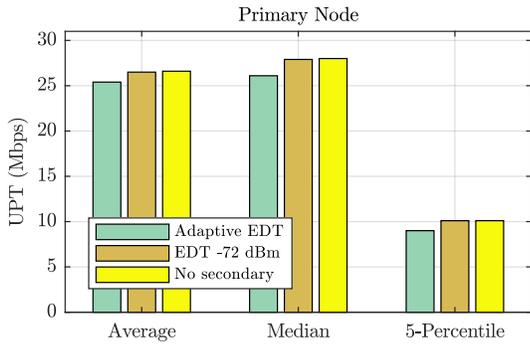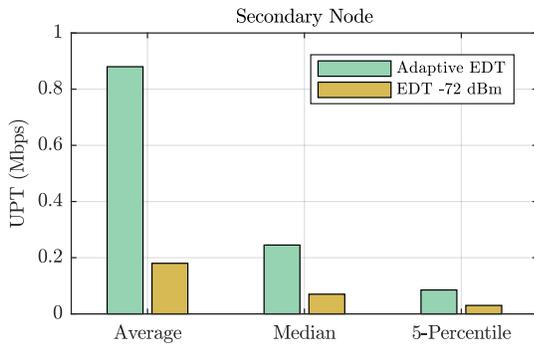


Fig. 14. Four node topology.



(a) PN UPT for the case where the SN adapts to PN load.



(b) SN UPT for the case where the SN adapts to PN load.

Fig. 15. Performance of PN and SN with adaptive versus fixed EDTs.

*3) Effects of Q-Learning Parameters:* In the next set of simulations, we have two nodes with two operators, a subset of the scenario in Figure 14. We perform the simulation with two different $\gamma_1$ settings for the Q-learning. Figure 16 shows these simulation results. In Figure 16a, we show the UPT of the PN. In the case of a fixed EDT of -62 dBm, our highest considered EDT, we see the lowest performance for the PN as the SN will not as readily defer to the PN. Inversely, for a

fixed EDT of -72 dBm, our lowest considered EDT, the SN will defer heavily to the PN, resulting in a performance similar to the case where there is no secondary node. When using an adaptive EDT, by changing the value of $\gamma_1$, we can balance the performance of the SN and PN. In Figure 16b, we can see the complementary performance of the SN.
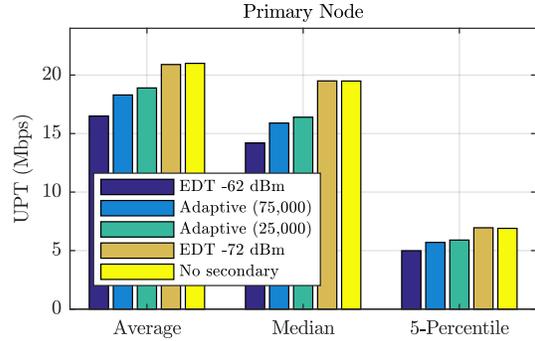


(a) PN UPT with differing $\gamma_1$ values.
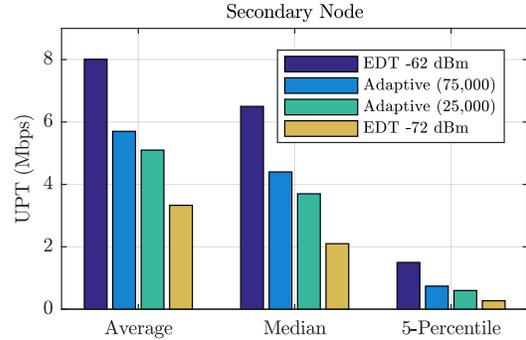


(b) SN UPT with differing $\gamma_1$ values.

Fig. 16. Spectrum sharing with different $L_m$ threshold values.

*4) Multi-Node Scenario:* In our final set of simulations, we use the LAA indoor scenario as outlined previously (shown in Figure 5). Figure 17 shows the benefit of Q-learning for the case where Op. 1 has at first a low, 0.125 MB/s average traffic arrival rate and then a high, 1.05 MB/s one while the SN traffic is kept high at 1.05 MB/s. For the low-traffic case, there is only about a 5% gain in UPT when using Q-learning because the SNs were already exploiting the many spectrum holes created by the limited traffic activity at PNs. However, when Op. 1 has a higher traffic load, the SN significantly benefits from Q-learning where the adaptive EDT leads to a median UPT improvement of over 30%.

## V. RELATED WORK

Adopting machine learning to wireless systems has been considered in previous works with promising results. In [16], the authors use Q-learning for dynamically choosing the channel for cells as opposed to static assignments. In [17], a decentralized Q-learning scheme is used for reducing the interference seen by 802.22, PN users. Although Q-learning has been considered for power allocations and channel assignments, it has not been used, to the best of the authors' knowledge, for adapting a dynamic EDT for SNs in a shared spectrum environment.
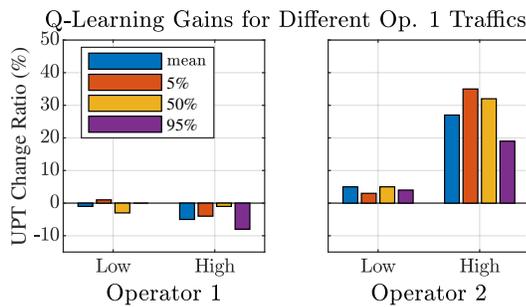
Fig. 17. Change in UPT for Op. 1 and Op. 2 when Op. 1 has a 0.125 MB/s (Low) and a 1.05 MB/s (High) average traffic arrival rates while Op. 2 always has a 1.05 MB/s traffic arrival rate.

Overall, CBRS is still an emerging standard, but we can learn from similar experiences on other bands. When considering the coexistence of GAA users, there are many similarities to unlicensed bands which has been studied extensively. Notably, there is substantial work that has been done for the coexistence of LTE and Wi-Fi nodes in unlicensed bands for LAA [12], [13], [18], [19] with [20] using reinforcement learning to alter the duty cycle of the LTE nodes. However, for CBRS, we only consider the case of LTE nodes coexisting with another LTE node that has a higher priority, which is a new area. There has not yet been extensive studies for CBRS-specific performance improvements; so far, there are only initial proof of concept demonstrations reported. In [21], a field trial of CBRS devices (CBSDs) working with the SAS was shown where they suggest improvements to the SAS protocol based on their results, and recently Verizon has deployed a CBRS network in Florida [22].

## VI. Conclusions

In this paper, we examined the challenges of using LBT for PAL/GAA spectrum sharing in CBRS networks by evaluating two different LBT schemes and showing that they can be used to greatly improve SN UPT with a minor decrease in PN UPT. To reduce the negative consequences of spectrum sharing on the PN, we presented a novel, Q-learning algorithm that adjusts SN opportunistic access via learning an optimal EDT for carrier sensing. We showed that by using average and differential PN buffer occupancy as the environmental observations, the SN can improve their throughput by up to 350% with only marginal losses to the PN UPT (4%). In future work, we can extend the intelligence globally from the local learning framework presented in this work, to jointly optimize within and across different shared-spectrum deployments, and examine how this work can scale to situations with multiple SNs.

## References

[1] B. Obama, "Presidential Memorandum: Unleashing the Wireless Broadband Revolution," June 2010. [Online]. Available: https://obamawhitehouse.archives.gov/the-press-office/presidential-memorandum-unleashing-wireless-broadband-revolution

[2] President's Council of Advisors on Science and Technology, "Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth," Executive Office of the President, July 2012. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_final_july_20_2012.pdf

[3] FCC 15-47. Federal Communications Commission, Apr. 2015. [Online]. Available: https://apps.fcc.gov/edocs_public/attachmatch/FCC-15-47A1.pdf

[4] FCC 08-260. Federal Communications Commission, Nov. 2008. [Online]. Available: https://apps.fcc.gov/edocs_public/attachmatch/FCC-08-260A1.pdf

[5] "Snapdragon x20 modem with category 18 gigabit class lte," Mar 2018. [Online]. Available: https://www.qualcomm.com/products/snapdragon/modems/x20

[6] D. Flore, "Laa standardization: coexistence is the key," 3GPP, July 2016. [Online]. Available: http://www.3gpp.org/news-events/3gpp-news/1789-laa_update

[7] M. Chmaytelli, "LTE-U Forum: Ensuring LTE and Wi-Fi fairly coexist in unlicensed spectrum," Qualcomm, Mar. 2016. [Online]. Available: https://www.qualcomm.com/news/onq/2015/03/03/lte-u-forum-ensuring-lte-and-wi-fi-fairly-coexist-unlicensed-spectrum

[8] "MulteFire Release 1.0 Technical Paper," MulteFire Alliance, July 2017. [Online]. Available: https://www.multefire.org/wp-content/uploads/MulteFire-Release-1.0-whitepaper_FINAL.pdf

[9] L. Wang, K. Wu, and M. Hamdi, "Combating Hidden and Exposed Terminal Problems in Wireless Networks," IEEE Trans. Wireless Commun., vol. 11, no. 11, pp. 4204–4213, Nov. 2012.

[10] K. Mun, "OnGo: New Shared Spectrum Enables Flexible Indoor and Outdoor Mobile Solutions and New Business Models," May 2018. [Online]. Available: https://www.cbrsalliance.org/wp-content/uploads/2018/04/Mobile-Experts-OnGo.pdf

[11] CBRS Coexistence Technical Specification, CBRS Alliance, 2018. [Online]. Available: https://www.cbrsalliance.org/wp-content/uploads/2018/04/CBRS-Coexistence-Technical-Specification.pdf

[12] H. J. Kwon et al., "Licensed-Assisted Access to Unlicensed Spectrum in LTE Release 13," IEEE Commun. Mag., vol. 55, no. 2, pp. 201–207, Feb. 2017.

[13] B. Jia and M. Tao, "A Channel Sensing Based Design for LTE in Unlicensed Bands," in 2015 IEEE International Conference on Communication Workshop (ICCW), June 2015, pp. 2332–2337.

[14] 3GPP, "Study on Licensed-Assisted Access to Unlicensed Spectrum," 3GPP, Technical Report (TR) 36.889, June 2015, version 13.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2579

[15] C. J. C. H. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, 1989.

[16] J. Nie and S. Haykin, "A Q-learning-based Dynamic Channel Assignment Technique for Mobile Communication Systems," IEEE Trans. Veh. Technol., vol. 48, no. 5, pp. 1676–1687, Sept. 1999.

[17] A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," IEEE Trans. Veh. Technol., vol. 59, no. 4, pp. 1823–1834, May 2010.

[18] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink Performance Analysis of LTE and WiFi Coexistence in Unlicensed Bands with a Simple Listen-Before-Talk Scheme," in 2015 IEEE 81st Veh. Technol. Conf., May 2015, pp. 1–5.

[19] N. Rupasinghe and I. Güvenç, "Licensed-assisted access for WiFi-LTE coexistence in the unlicensed spectrum," in 2014 IEEE GC Wkshps, Dec. 2014, pp. 894–899.

[20] ——, "Reinforcement learning for licensed-assisted access of LTE in the unlicensed spectrum," in Proc. of 2015 IEEE Wireless Commun. and Netw. Conf., Mar. 2015, pp. 1279–1284.

[21] M. Palola et al., "Field trial of the 3.5 GHz citizens broadband radio service governed by a spectrum access system (SAS)," in Proc. of DySPAN 2017, Mar. 2017, pp. 1–9.

[22] "You don't need high grade Navy radar systems to spot which companies just achieved another industry milestone for customers," May 2018. [Online]. Available: http://www.verizon.com/about/news/you-dont-need-high-grade-navy-radar-systems-spot-which-companies-just-achieved-another