UNLOCKING 5G SPECTRUM POTENTIAL FOR INTELLIGENT IOT:
OPPORTUNITIES, CHALLENGES, AND SOLUTIONS

Dynamic Spectrum Management through Resource Virtualization with M2M Communications

Abdallah Moubayed, Karim Hammad, Abdallah Sham, and Hanan Lutfiya

Abstract

Wireless spectrum licensing has increased due to the continuous evolution and use of cellular technology. The increase in the number of mobile-connected devices and global data traffic demand has led to a significant increase in demand for spectrum access with studies showing that there is unexploited capacity in the spectrum. This is especially critical for 5G networks where the service requirements are extremely stringent. To that end, this article proposes dynamic spectrum management through the combination of two innovative architectures, wireless resource virtualization (WRV) and machine-to-machine (M2M) communications. WRV allows for better utilization of the spectrum, while reducing both the capital and operational expenditures. On the other hand, M2M communications can help boost capacity and improve quality of service by leveraging spectrum access across multiple radio technologies. In this article, a brief discussion of multi-radio access technology heterogeneous networks is given. Then the problem of dynamic spectrum management through resource virtualization with M2M communications is described. To the best of our knowledge, such a combined framework has not been previously proposed. Two different algorithms are proposed to evaluate the performance of the considered architecture, namely a decomposition-based algorithm and a greedy-based algorithm. Simulation results show that such architecture can boost the overall capacity of the system by achieving higher data rates. Moreover, it is shown that the number of possibly supported M2M pairs is increased while using the same spectrum.

Introduction

The licensing of wireless spectrum has grown due to the proliferation and evolution of cellular technology [11]. This has caused the demand for spectrum access to grow dramatically due to the increasing number of new consumers for services such as satellite digital audio broadcasting [11]. It is projected that the number of mobile-connected devices will reach 11.6 billion by 2021 with the monthly global mobile data traffic demand reaching 49 exabytes [2]. However, due to the spectrum being mostly allocated and divided, providing more spectrum to expand existing services or offer new ones has become more challenging [11]. However, U.S. studies have shown that the issue is not the lack of spectrum, but rather spectrum access [11]. This means that the spectrum’s capacity is not being exploited to its full extent [11]. Moreover, service providers (SPs) are searching for creative ways to meet the growth in data services’ demand rate while simultaneously improving the average gain per user [11]. This is especially critical for fifth generation (5G) networks where the service requirements are extremely stringent [3].

Wireless resource virtualization (WRV) is one promising solution that has been proposed to meet the increasing rate demand in 5G networks [1, 4]. WRV refers to the sharing of the physical infrastructure and slicing of the wireless resources among coexisting networks in a dynamic manner to better utilize the available resources [1, 4]. WRV can be extremely advantageous to participating SPs. First of all, since SPs share resources, resource utilization improves. This in turn reduces the number of resources that remain idle. Second, the capital and operational expenditures (CAPEX and OPEX) can be minimized by up to 80 and 27 percent, respectively, when adopting WRV [1, 4]. Additionally, higher peak rates can be achieved with WRV due to aggregated and shared radio resources. Finally, due to the increase of user numbers within the cells, multi-SP multiplexing gain is achieved. It is worth noting that WRV's benefits can be further amplified when adaptive resource allocation algorithms are adopted, particularly in systems that use orthogonal frequency-division multiple access (OFDMA), such as in LTE-A downlink [1, 4].

Another proposed solution for 5G networks is the adoption of a multiple-radio-access-technology (multi-RAT) heterogeneous networks (Het-Nets) architecture. This architecture aims to boost capacity and improve the quality of service (QoS) by leveraging spectrum access across multiple radio technologies [5]. It seeks to aggregate the various radio technologies into one common converged network that is seamless to the end user. The aim is to develop resource allocation techniques that can utilize the available resources in the different RATs in an efficient manner. Multi-RATs also open the door to improved performance gains through efficient diversity dimen-
Multi-RAT HetNets have been heavily proposed in the literature as a promising architecture for 5G systems. The idea is to aggregate the different radio access technologies together into one converged network in order to offer more capacity and enhance the QoS for users.

The idea originated from the fact that the spectrum in the licensed band is heavily assigned. Hence, the need to make better use of the unlicensed band arose in order to increase the capacity of the system and offload some of the load placed on the cellular network. Early works suggested the integration of Wi-Fi into the Third Generation Partnership Project (3GPP) networks to alleviate the congestion in the cellular networks with simulation results showing $2-3 \times$ gains in terms of system capacity and users’ QoS. The proposed architecture was then further developed to include many other access technologies to further improve the gain by exploiting the diversity available in multiple dimensions (i.e., spectral, temporal, and frequency).

**Technologies**

There are different technologies that have been proposed in the literature to be used in multi-RAT HetNets. The below list shows some of these technologies:

- Long Term Evolution/Long Term Evolution-Advanced (LTE/LTE-A)
- Wi-Fi
- Legacy networks (GSM, UMTS, etc.)
- D2D communications
- Internet of Things (IoT)-based M2M communications (Bluetooth, Zigbee, etc.)

**Challenges**

Although HetNets have several advantages and are considered a promising architecture to meet the requirements of 5G networks, adopting them produces a unique and diverse set of challenges. These challenges are grouped into three main categories: physical-related (pertain to the physical layer of the network), user-related (pertain to the user experience), and core-network-related (pertain to the functions and load in the core network). These are summarized in Fig. 1.
Spectrum efficiency: With the scarcity and expensiveness of the spectrum, especially the licensed band, it is important to efficiently use the available spectrum to maximize the performance gain of using multiple technologies [8]. Padaki et al. showed that the maximum spectrum efficiency achievable per user with optimal transmission power ranges between 15–25 b/s/Hz [9]. This means that if a user is assigned one resource block (RB) in an LTE system, their rate would only be around 4.5 Mb/s. However, with the increasing growth of data, the spectrum efficiency has to increase further to ensure that users have a satisfying QoS.

Joint radio operation/interference management: This challenge pertains to the resource allocation, which in turn affects the spectrum efficiency and resource utilization. The idea is that resource allocation should be done jointly between the multiple RATs. This ensures that interference is minimized and that the spectrum is used efficiently [10]. It was shown in [11] that for the scenario involving one LTE macrocell and 30 WiFi cells, the average user throughput reaches around 320 kb/s in the presence of a joint radio resource allocation scheme. Moreover, the energy efficiency reaches 180 kb/s/J for the joint scheme. This highlights the significance of adopting a joint radio resource allocation.

Quality of experience: The perceived quality of experience (QoE) of users is vital. Since users will be using different RATs and protocols, this should be done in a seamless manner that does not affect the users’ QoE [8].

RAT selection: How to efficiently assign which RAT the user will employ is also vital to ensure ubiquitous connectivity. This is dependent on several metrics like the perceived channel conditions within the different RATs as well as the load within the RAT. Also, the users’ QoS requirements and resource availability will affect this selection process. Therefore, each user should be assigned to a specific RAT to maximize the overall performance of the network [8]. Wu et al.’s work showed almost a 10 percent improvement in average aggregate throughput when taking users’ preference into consideration compared to the maximum signal-to-interference-plus-noise ratio (SINR)-based scheme. Furthermore, the average normalized power consumption of low-power users was reduced to around 15 percent using the preference-aware scheme. Thus, the aggregate throughput that can be achieved in a real-life scale network can reach hundreds of gigabits per second. Therefore, it is crucial to develop efficient RAT selection algorithms to further improve the network throughput and performance while also making full use of the available spectrum.

Inter-RAT handover management: Since multiple radio access technologies (RATs) are employed, it is important to have an efficient inter-RAT handover mechanism since users will be constantly moving from one RAT to another [8]. Extending Rao et al.’s work [12] to a multi-RAT scenario with thousands of cells and millions of users, the amount of messages/day exchanged will reach the hundreds of millions (average message size = 7.5 bytes). This in turn will put more load on the core network as it has to deal with the excess signaling overhead. This will also have an impact on the possible spectrum gains since more resources would be needed to exchange these messages.

Mobility management: Determining where the user is and to which BS (also RAT) it belongs is a challenging issue in multi-tier HetNets. Since the different technologies handle the mobility in a different manner and using different entities, it is vital to determine a way to efficiently converge these entities/mechanisms so that the user mobility is managed appropriately [10]. This has added importance given the amount of messages exchanged between the core network and the users for mobility purposes, which might also have an impact on the spectrum utilization.

Inter-RAT resource coordination, utilization, and scheduling: With the different RATs used, each of which uses a different spectrum band and different access protocols, the coordination, utilization, and scheduling of the available resources is a challenging aspect. Therefore, it is crucial to be able to efficiently utilize the resources available to maximize the performance gain of the multi-RAT architecture [8]. By combining the work in [1] and [13], it can be shown that the overall complexity to schedule and coordinate the available resource can be prohibitively high. For example, the search space when \( J = 15 \) MCSs, \( L = 25 \) RBs (for bandwidth of 5 MHz), \( |C| = 5 \) users/cell, and \( N = 100 \) cells would be around \( 6.38 \times 10^{40} \) feasible solutions. Therefore, it is essential that low-complexity algorithms are developed that can efficiently manage the radio resources by maximizing the resource utilization.

Network Management: Since different RATs and protocols are employed, it is more challenging to manage the whole network. Therefore, it is important to have efficient algorithms/architectures that can manage the network to achieve the desired requirements for all users [8].

Figure 2. Illustrative example.

**Figure 2. Illustrative example.**

**System Model and Problem Description**

As mentioned earlier, the proposed architecture is a multi-RAT HetNet-based architecture that leverages the benefits of resource virtualization and M2M communication (e.g., Bluetooth or Zigbee) in order to have more efficient management and utilization of the available spectrum. In what follows, the gen-
eral model considered in this work is presented. Moreover, the channel model used to represent the channel conditions is given. Additionally, a mathematical description of the problem is provided containing the objective and the constraints considered.

**General Model**

A single cell in the downlink phase of an LTE-A system is assumed in this article. A set of SPs are supported by an infrastructure provider (InPr) that manages the cell. Each SP supports two sets of users, namely a set of cellular users (CU) and a set of machine-to-machine (M2M) pairs. These pairs can communicate locally using any protocol such as Bluetooth or Zigbee. Moreover, they can communicate with the BS as per LTE-A Release-12 [14, 15]. The SPs’ frequency bands are assumed to be contiguous with each band containing multiple RBs. This constraint ensures that users are able to access the available sub-channels. A minimum number of RBs is allocated to each SP based on the pre-determined access ratio agreement between them and the InPr. The proper allocation of RBs to the users of the SPs ensures isolation between them. It is assumed that the BS and all users transmit at constant power, which is set to their maximum power. Moreover, it is assumed that perfect channel state information (CSI) of all users is available at the BS [1, 4].

Figure 2 shows an illustrative example with two SPs, each supporting two CUs and two M2M pairs (SP1 supports CU1, CU2, M1, and M2, while SP2 supports CU3, CU4, M3, and M4). Each SP originally has five RBs to allocate. It is shown that in the non-WRV case, users CU1 and CU2 only have access to RBs 1 through 5, while CU3 and CU4 can only receive RBs 6 to 10. The same applies to M2M pairs M1 to M4, respectively. In contrast, the set of RBs is aggregated into one bigger pool in the case of WRV, giving the users access to a larger set of possible RBs to be allocated, which can result in higher throughput. As can be seen, CU1 would get RBs 1, 3, 6, and 7, while CU2 and CU4 having the same allocation. This can result in higher throughput due to the higher number of RBs allocated (as is the case, e.g., with CU1) or better channel conditions (CU3 in this case). The same applies to the M2M pairs, which will have a limited number of RBs to share in the non-WRV case. However, with WRV, they can share more RBs with possibly better channel conditions, which can result in higher throughput for them as well.

**Channel Model**

This work assumes that the model describing the channel between the BS, also denoted as eNodeB, and the user consists of two components. The first is a deterministic distance-dependent macroscopic path loss component. This component follows the model proposed in [14]. The second component represents the random shadow fading path loss. This component is modeled as a Gaussian random variable as shown in [1, 4]. In contrast, the channel condition between any two users is assumed to be distance-dependent and follows the same model given in [14]. The model considers two different cases, The first is a line-of-sight (LoS) case. In this case, the distance separating the two machines is less than 300 m. The second case is a non-line-of-sight (NLoS) case where the distance separating the two machines is larger than 300 m.

**Problem Description**

Similar to the formulation in [11], this work assumes that one M2M pair can share the RBs allocated to one cellular user. This is so that the aggregate throughput of the network is...
improved. Hence, an M2M pair can only share RBs allocated to a CU of the same SP rather than being assigned dedicated RBs. This is to avoid any possible cross-pricing issues with other SPs. The formulated optimization problem is designed so that the performance of the existing cellular network is not degraded. Each scheduling round consists of allocating the available RBs. Additionally, it is assumed that an infinitely backlogged model is adopted as users continuously have data to receive. Two distinct user sets are served by the eNodeB. The first set represents the set of CUs of the different SPs. The second set represents the set of M2M pairs belonging to the different SPs. Moreover, $C_m$ denotes the set of CUs, and $MTC_m$ denotes the set of M2M pairs belonging to SP $m$.

A CU experiences interference only when an M2M pair transmits on the same RBs allocated to the CU. This interference is dependent on the M2M pair’s transmission power as well as the channel gain between the CU and the pair. The downlink SINR observed by a particular CU at a specific RB is the same as that given in [1]. On the other hand, the SINR of an M2M pair is the same as that given for a D2D pair in [1]. A binary variable is used to indicate whether an M2M pair shares the RB(s) allocated to a CU or not. Therefore, this variable is set to 1 if the M2M pair shares the RB(s) and is 0 otherwise.

The CUs’ and M2M pairs’ rate is defined as per Shannon’s capacity model based on their respective SINRs. The aim is to maximize the sum rate of both the CUs and M2M pairs. This optimization problem can be modeled as an integer nonlinear program (INLP) in a similar manner to that in [1]. The problem contains two distinct decision variable sets. The first set is a set of integers denoting the number of RBs allocated to a particular CU. The second set is a set of binary integers denoting whether an M2M pair shares the RB(s) allocated to a particular CU. The formulation contains seven constraints. The first two constraints ensure that both the CUs and M2M pairs’ threshold SINR is adhered to. The third constraint mandates that an M2M pair shares at most one CU’s RB(s). Additionally, the fourth constraint ensures that a maximum of one M2M pair shares any CU’s RB(s). This is to avoid any interference caused by the M2M pairs to each other. The fifth constraint guarantees that the pre-agreed access ratio illustrated by the number of RBs assigned to each SP is respected. Lastly, the final two constraints ensure that both the CUs and M2M pairs achieve a minimum rate. It is worth noting that the assumption that an admission control process is implemented a priori guarantees that solving such a problem is feasible [1]. It is also worth mentioning that the described model only takes into account one type of data traffic for each set of users. To add different data types, other constraints need to be considered within the model.

**DECOMPOSITION-BASED ALGORITHM**

As shown in [1], regular nonlinear or integer optimization techniques cannot be used to solve such a problem. Therefore, the decomposition-based algorithm takes advantage of the fact that resource allocation for CUs should be done first to solve this problem. This is because of the assumption that the performance of the cellular network should not be degraded. Moreover, it is assumed that M2M pairs are not allocated dedicated RBs, but rather can only share allocated RBs. Therefore, the original optimization problem is decomposed into two simple sub-problems.

The first sub-problem represents the resource allocation problem for CUs under the assumption that no interference exists from M2M pairs. Therefore, this can be modeled as a binary integer programming problem that can be solved to optimality. The second sub-problem represents the M2M users’ resource sharing problem. The potential interference caused to the CUs is taken into consideration in this sub-problem and is defined to be one of the optimization problem’s constraints. This sub-problem can also be modeled as a binary integer programming problem that can be solved to optimality.

Despite the adoption of a decomposition-based algorithm to solve the initially described problem, it is still computationally prohibitive to optimally solve these sub-problems [1]. Therefore, it is crucial that heuristic algorithms with lower complexity are developed to solve these problems.

**GREEDY-BASED HEURISTIC ALGORITHM**

The proposed heuristic is a greedy and exhaustive search-based algorithm. It is divided into two parts. The first part solves the CUs’ resource allocation problem, while the second part solves the M2M pairs’ resource sharing problem. Figure 3 provides a flowchart of the algorithm.

**HEURISTIC I: CUS’ RESOURCE ALLOCATION PROBLEM**

The same heuristic algorithm proposed in [1] is adopted in this work to solve the resource allocation problem for CUs. The heuristic is decom-
The proposed heuristic is a greedy and exhaustive search-based algorithm. It is divided into two parts. The first part solves the CUs’ resource allocation problem, while the second part solves the M2M pairs’ resource sharing problem.

This part uses an exhaustive-search-based procedure to solve the M2M pairs’ resource sharing problem. CUs are first sorted in decreasing order of their rate requirement. This is because one RB is assigned to each user sequentially. The second part is a greedy manner based on the SLA agreement to ensure that each SP gets the minimum number of RBs agreed on with the InPr.

As shown in [1], this algorithm’s order of complexity is the product of the total number of RBs to be allocated and the aggregate number of CUs in the cell. That is because one RB is assigned in each iteration by searching among all the possible users.

**HEURISTIC II: M2M PAIRS’ RESOURCE SHARING PROBLEM**

This part uses an exhaustive-search-based procedure to solve the M2M pairs’ resource sharing problem. CUs are first sorted in decreasing order of number of allocated RBs so as to maximize the possible number of RBs that each M2M pair can share. The interference caused by each M2M pair to the CU with the highest number of allocated RBs is calculated. The CU’s resulting SINR is compared to its threshold SINR. The M2M pair chosen is the one that satisfies the SINR constraint and causes the least interference to the CU. The algorithm moves to the next CU if none of the M2M pairs satisfies the initial CU’s SINR constraint.

This algorithm’s complexity is dependent on the complexity of its two sub-algorithms: the adopted sorting algorithm and the resource sharing decision algorithm. It can be shown that the complexity of optimal sorting algorithms such as Radix sorting is on the order of the number of CUs using lower MCSs that have lower spectral efficiencies. The second observation is that M2M communication can reduce the effect of the worsening channel conditions as the decay is not as fast as expected. A third observation is that employing WRV can help improve the sumrate by up to 10 percent. This is because with the bigger sample size of resources available for each SP, users have a better chance of being allocated a better channel. This emphasizes the positive impact that WRV can have on the achievable sumrate. Last but not least, it is noticeable that the greedy-based algorithm performs comparably to the decomposition-based solution, thus highlighting its efficiency.

Table 2 shows the average rate per cellular user, average rate per M2M pair, and average combined rate of the first scenario. For the second scenario, the maximum number of M2M pairs that can be supported is shown. This is done for the three SPs with and without WRV. Note that the results shown in the table are based on those obtained using the decomposition-based algorithm. Results show that employing WRV can help SPs achieve better performance in terms of average rate per user (average 9 percent per CU and 10.97 percent per M2M pair) as well as the average number of M2M pairs that can be supported (up to 15.79 percent more users). This implies that employing WRV improves the spectral efficiency and resource utilization as it allows a higher rate to be obtained using the decomposition-based algorithm.
be transmitted and more M2M pairs to be supported. Hence, it can be deduced that the use of WRV has improved the spectrum utilization by increasing the users’ transmission rate and the number of concurrent connections using the same spectrum.

**Conclusion and Future Research Directions**

In this article, the problem of wireless resource virtualization with M2M communication underlaying the LTE-A network is described. It is shown that the problem can be formulated as an INLP problem. Due to the high complexity of such problems, a decomposition-based algorithm that divides the main problem into two smaller linear binary integer programming subproblems is presented. Using the decomposition-based algorithm, each of the two subproblems is solved to optimality. Moreover, a greedy-based algorithm that solves each of the two subproblems is described and developed. Results show that wireless resource virtualization increases the system spectrum utilization. Also, M2M communication helps reduce the impact of worsening channel conditions. Furthermore, the proposed greedy-based algorithm performs comparably to the decomposition-based algorithm while being less computationally complex.

This work can be extended in multiple directions. One possible extension is considering a heterogeneous traffic model that includes different data traffic types with different QoS requirements for different users. This would consider a more general case where users rarely have only one type of data traffic. This can be applied to both cellular users as well as M2M pairs.

**References**


**Table 2. Performance gain results.**

<table>
<thead>
<tr>
<th></th>
<th>First scenario</th>
<th>Second scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg. rate/ CU (kb/s)</td>
<td>Avg. rate/ M2M pair (kb/s)</td>
</tr>
<tr>
<td>No WRV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP1</td>
<td>6711.8</td>
<td>17922.0</td>
</tr>
<tr>
<td>SP2</td>
<td>3385.9</td>
<td>8961.2</td>
</tr>
<tr>
<td>SP3</td>
<td>846.5</td>
<td>1994.0</td>
</tr>
<tr>
<td>With WRV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP1</td>
<td>7056.7</td>
<td>20618.0</td>
</tr>
<tr>
<td>SP2</td>
<td>3446.3</td>
<td>9458.0</td>
</tr>
<tr>
<td>SP3</td>
<td>1019.1</td>
<td>2240.3</td>
</tr>
</tbody>
</table>

**Improvement (%)**

|                  |                |                 |                  |
| SP2              | 1.80           | 5.54            | 3.55             |
| SP3              | 20.39          | 12.35           | 5.68             | 15.79                    |

**Biographies**

AbdAllah MoubAyed (amoubaye@uwo.ca) received his B.E. and M.Sc. degrees in electrical engineering from the Lebanese American University, Lebanon, and King Abdullah University of Science and Technology, Saudi Arabia, in 2012 and 2014, respectively. He is currently pursuing a Ph.D. in electrical and computer engineering at Western University, London, Ontario, Canada. His research interests include wireless communication, resource allocation, performance and optimization modeling, machine learning and data analytics, computer network security, and e-learning.

Karim HammAd (khammad@uwo.ca; khammad@aaau.edu) received his B.Sc. and M.Sc. degrees in electronics and communications engineering from the Arab Academy for Science, Technology and Maritime Transport, Egypt, in 2005 and 2009, respectively, and his Ph.D. degree in electrical and computer engineering from Western University in 2016. He is currently a postdoctoral fellow at York University, Ontario, Canada. His research interests include wireless network cross-layer design, physical layer security, and digital circuit design.

AbdAllah ShAmi (abdallah.shami@uwo.ca) received his B.Sc. degree in electrical and computer engineering from Lebanon University in 1997 and his Ph.D. degree in electrical engineering from the City University of New York in September 2002. He joined Western University in July 2004, where he is currently a professor in the Department of Electrical and Computer Engineering. His current research interests are in the areas of network optimization, cloud computing, and wireless networks.

HanAn lutfiyya (hlutfiyya@uwo.ca) received her B.Sc. degree from Yarmouk University, Jordan, her M.Sc. degree from the University of Iowa, Ames, and her Ph.D. degree in 1992 from the University of Missouri, Rolla, all in computer science. She has been with the Computer Science Department at Western University since 1992, where she is a professor and chair of the Computer Science Department. Her research interests include distributed systems management and software engineering.