QoS-Aware Energy-Efficient Downlink Predictive Scheduler for OFDMA-Based Cellular Devices

Karim Hammad, *Student Member, IEEE*, Serguei L. Primak, *Member, IEEE*, Mohamad Kalil, *Member, IEEE*, and Abdallah Shami, *Senior Member, IEEE*

***Abstract*—We propose a predictive energy-efficient scheduling scheme that optimizes the user equipment (UE)'s bits/joule metric subject to quality-of-service (QoS) constraints in downlink orthogonal frequency-division multiple access (OFDMA) systems. This is achieved by minimizing the number of wake-up transmission time intervals (TTIs), where the UE receiver circuit is ON, in a longer time horizon than studied before. The proposed predictive scheduler is supported by a ray-tracing (RT) engine that increases the scheduler's knowledge by long-term information about the users' propagation characteristics. A convex multiobjective binary-integer-programming formulation for the problem is presented to optimize both the UE's *energy efficiency* (EE) and QoS. The multiobjective formulation is then used for benchmarking of a low-complexity and computationally efficient heuristic scheduler. The results show that the proposed schedulers have significantly improved the UE's EE and the overall capacity of the system, compared with a recently published EE scheduling scheme while maintaining target QoS.**

***Index Terms*—Cloud radio access networks (C-RANs), energy–efficient communications, orthogonal frequency-division multiple access (OFDMA), ray tracing (RT), resource allocation.**

I. Introduction

A RAPID growth of cellular system designs and standards over the past ten years has significantly enlarged the wireless market volume. Today's statistics show that over 1 billion users worldwide are connected to social networking media such as Facebook and YouTube [1] and that approximately 40% of them are mobile users [2]. However, analysts predict that these numbers will continue to grow dramatically over the coming years [1] due to two major factors. The first is the unabated advancements in the mobile devices industry, particularly with smartphones. Their high computing capabilities allow them to replace many other important devices such as GPS receivers, cameras, and laptop computers. The other factor is the increasing popularity of multimedia services such as Voice over Internet Protocol (VoIP), video streaming, social networking, interactive gaming, web browsing, etc.

From a technical point of view, the emergence of the aforementioned services over the currently deployed fourth-generation (4G) networks [i.e., long-term evolution (LTE)] has introduced various challenges for the system design from both the network and user equipment (UE) sides. From the network side, most operators seek to maximize capacity (i.e., spectral efficiency) and reduce the operation cost, including the energy efficiency (EE). These goals present challenges, particularly in situations of potentially increasing numbers of users and heavy-load traffic connections, while having to maintain stringent QoS requirements. Whereas from the UE side, the intensive and complex circuitry of a 4G device is quite rigorous on the current smartphone battery technology. This results either in a fast depletion of the battery energy, or it may limit the implementation of a fully functional 4G device. Therefore, a main stream of research has recently been established and devoted for enabling green communications (i.e., energy-efficient or energy-aware communication systems) [3]. The future generation of mobile communications, known as fifth generation (5G) [4], will address the EE as a fundamental aspect of the system.

A. Related Work

An energy-efficient design for wireless systems should encompass both the network and the UE sides. Although the majority of the system's energy consumption resides in the network side [5], most recent studies were focusing on optimizing the UE energy consumption either in the uplink [6]–[8] or in the downlink [9]–[13]. This is due to the need to increase the UE's battery lifetime per charge. Consequently, in this paper, we focus on minimizing the UE's energy consumption in the downlink, a subject less studied in the literature.

In [9], it is showed that optimizing the UE power consumption inherently requires optimization of the base-station (BS) downlink transmit power. Hence, the optimization formulation was designed to improve the EE for both of the BS and UE. The idea was based on buffering BS downlink traffic for some transmission time intervals (TTIs) and then transmitting these data in the minimum possible number of time slots constrained by a fixed bit rate constraint. However, the implemented heuristic did not consistently fulfill the data rate constraint. Unlike in the earlier approach, in [10], only the EE from the BS side is considered. The objective in [10] was to design an optimal energy-efficient resource-allocation scheme with delay provisioning for delay-sensitive traffic in downlink orthogonal frequency-division multiple access (OFDMA)-based wireless

Manuscript received October 2, 2015; revised January 29, 2016; accepted March 25, 2016. Date of publication April 20, 2016; date of current version February 10, 2017. The review of this paper was coordinated by Dr. F. Gunnarsson.

The authors are with the Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON N6A 3K7, Canada (e-mail: khammad@uwo.ca; slprimak@uwo.ca; mkalil3@uwo.ca; ashami2@uwo.ca).

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

Digital Object Identifier 10.1109/TVT.2016.2555293

access networks. The model of the scheduling problem used the effective capacity concept to provide the statistical delay provisioning. Thus, the problem was modeled as maximizing the effective capacity-based EE under statistical delay constraints. Utilizing the effective capacity method, such as the model in [10], in [11], an adaptive resource-allocation scheme is proposed for downlink heterogeneous mobile wireless networks. The scheme dynamically assigns power levels and time slots, and derives the admission control conditions for different real-time mobile users to satisfy various statistical delay-bound QoS requirements. The channel state information (CSI) is taken into consideration, which was estimated at the receiver and sent back to the transmitter, for adaptive modulation and adaptive power control. In a different context, in [12], the problem of improving the EE in the downlink of an OFDM-based cognitive radio (CR) network is considered. The objective was to design an energy-efficient resource-allocation scheme that maximizes the overall EE of the CR system while considering proportional fairness and rate requirements among the secondary users. This is in addition to keeping the interference to the primary users below their tolerable thresholds.

Unlike the studies mentioned earlier, in [13], a green resource allocation (GRA) scheme is proposed as an alternative approach to the well-known Third-Generation Partnership Project LTE discontinuous reception (DRX) power management scheme [14]. To minimize the UE energy consumption in the downlink, the scheduling of the BS downlink transmissions to the UE us optimized to a fewer time slots while turning off the receiver circuit in the unused slots. The scheduling was formulated as a nonlinear-integer programming problem. In contrast to this paper, which focuses on the EE problem, our previous work [15] in the area of predictive scheduling has focused on maximizing the network's average throughput (i.e., spectral efficiency), subject to fairness constraints in time-division-multiple-access-based systems.

B. *Scope and Contribution*

In this paper, we consider the LTE frequency-division duplexing (FDD) mode system with a frame structure type 1, where two time slots make one subframe (i.e., of duration 1 ms) [16]. Combining ten subframes (i.e., used for scheduling) makes one frame with a length of 10 ms. We noted that the work in [13], such as many studies in both downlink and uplink [17], depends on scheduling time granularity of one subframe or at most one frame. In this paper, we further expand the solution space of the scheduling problem for optimizing UE's energy consumption in the downlink while maintaining users' QoS requirements. The key strategy, as used in [13], is minimizing the number of wake-up TTIs for the UE's receiver circuit but in a longer timescale, spanning multiple future frames (i.e., 10-ms LTE radio frames) of the UE's channel.

The problem's time expansion is supported by preestimating the users' propagation channel over multiple future frames. This is done using an advanced ray-tracing (RT)-based central downlink scheduler system implemented at the BS site. The direct result of increasing the knowledge about the user's CSI is that the scheduler's capability of increasing the UE's EE and meeting the QoS requirements becomes higher compared with previously implemented schedulers. In other words, the UE opportunistically consumes less energy for the same amount of data received based on the future statistics of the propagation channel. This idea is shown in Fig. 1. It shows that the predictive scheduler with two frames of time granularity would be able to rearrange downlink transmissions to UEs into fewer TTIs compared with that of the traditional single-frame scheduler. This results in better EE and possibly offloading spectral resources that help in admitting more UEs. However, the underlying increase in the solution space of the optimization problem results in a substantial growth of the computational complexity. We then address this complexity by designing a less-complex heuristic algorithm that approximates the optimal scheduler performance. More details about the optimization problem and its relaxation is provided in Sections IV and V.

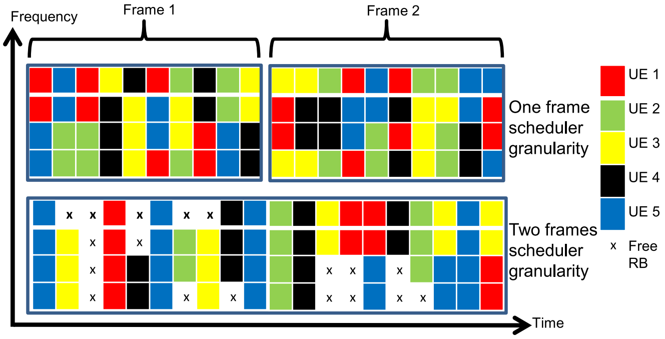
Frequency
Frame 1
Frame 2
One frame scheduler granularity
Two frames scheduler granularity
Time
UE 1
UE 2
UE 3
UE 4
UE 5
x Free RB

Fig. 1. Energy-efficient predictive scheduling.

The contributions of this paper are summarized as follows.

- We propose an optimal framework that minimizes the energy consumption of the UE receiver circuit while satisfying a constant rate (i.e., effective bandwidth) constraint. The framework utilizes the ray-tracing channel prediction model, and it considers both the modulation and coding scheme (MCS) and UE circuit operation time.
- To assure feasible solutions, we propose a second formulation for the optimization problem through relaxing the rate constraint using the penalty method to cope with the channel capacity limitations.
- After investigating the dominant factors that affect the UE's power consumption budget in the downlink, we further modify the optimization problem by allowing the scheduler to focus solely on optimizing the number wake-up TTIs for the UE.
- To address the complexity of the optimization problem, we deduce a heuristic algorithm to solve the scheduling problem in the final formulation with a comparable performance but significantly lower complexity.

The remainder of this paper is organized as follows. Section II presents the system model and design objectives. The motivation for utilizing the RT channel prediction model with the proposed RT-based scheduler system is discussed in Section III. The optimal formulation and the iterative algorithm of the proposed scheduler are described in Sections IV and V,

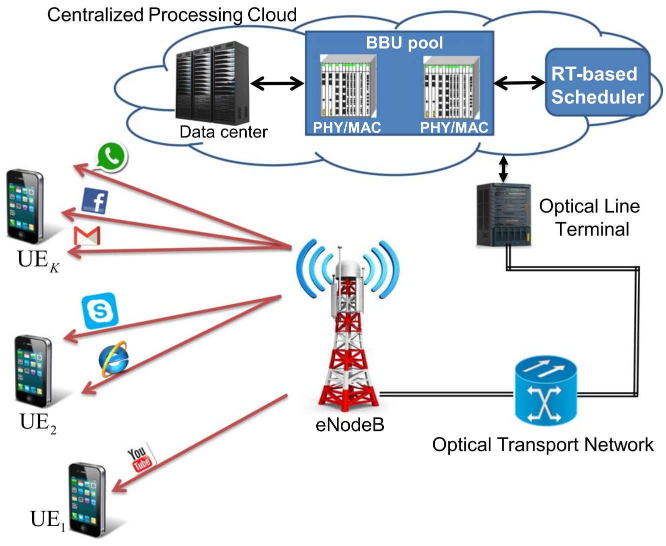

Fig. 2. C-RAN-based model.

respectively. Simulation results are provided in Section VI. Finally, Section VII concludes this paper.

II. System Description

A. System Model

We consider a single cell of a mobile cloud computing (MCC) LTE downlink multiuser system. It is based on the evolving concept of cloud radio access networks (C-RANs) [18], [19]. Online computational resources can be used for the computationally demanding RT prediction of the evolved Node B (eNB)–UE channel, as explained later within the MCC framework. The arrangement allows the transfer of the prediction cost from the eNB to the C-RAN. This is shown with the aid of Fig. 2. A C-RAN architecture is based on centralizing multiple baseband (BB) processing units (i.e., traditionally located at every BS site) and forming a pool of shared wireless resources within a centralized processing cloud [18]. In addition to the BB unit (BBU) pool, the cloud also integrates a data center and an RT-based downlink scheduling system. The data center is responsible for establishing users' traffic connections based on standard QoS requirements. The RT-based scheduling system mainly integrates an RT engine (i.e., for predicting the downlink CSI) and a central scheduler. The system's detailed structure and operation will be discussed in Section III. The eNB tower is connected to the central processing cloud via an optical transport network.

We assume a single eNB located at the center of cell. The overall cell bandwidth is divided equally into N 180-kHz resource blocks (RBs) consisting of 12 adjacent subcarriers. The FDD LTE frame type 1 duration is 10 ms and is composed of ten 1-ms subframes (i.e., each subframe represents a TTI) [16]. When the normal cyclic prefix is used, each subframe consists of 14 OFDMA symbols, each with duration of 66.67 μs.

The eNB transmits H traffic connections (i.e., bearers) to each one of K UEs. As our model addresses the UE's EE in the downlink, we use $P_t^{(k)}$ to denote the total downlink power consumed by the receiver circuit of UE k. More details about the calculation of the components of $P_t^{(k)}$ are provided in Section II-B. Each user connection is associated with different QoS requirements, depending on the traffic type [e.g., VoIP, video streaming, and File Transfer Protocol (FTP)]. However, without loss of generality, we assume that the QoS requirements for all traffic connections (of each UE) are preprocessed by the data center. Then, the data center translates them into a single connection request (or reservation) with a target average transmission rate $\bar{R}_D$. That rate is calculated based on the effective bandwidth theory (i.e., the dual concept of the effective capacity [10]) and will simultaneously meet all of the user's individual connections requirements. From the UE side, multiple traffic connections with different QoS parameters are further prioritized (i.e., intrascheduling) according to their QoS class identifier (QCI) priority (see [16, Tab. 13.1]). It is known that intrascheduling user's connections are preceded by UE's interscheduling. This two step process is vital particularly when the system's capacity prevents the scheduler from allocating enough resources to accommodate the user's target rate (i.e., $\bar{R}_D$).

To satisfy the users QoS requirements, we assume that the central processing cloud requests data connections between eNB and users with a total time duration of T and a target average bit rate for each UE of $\bar{R}_D^{(k)}$. Starting from this assumption, the eNB ultimately aims to schedule the transmission of the data for each user's requested connection in a way that satisfies two major goals. The first is to maintain the average connection's downlink rate for each UE by adequate allocation of resources, i.e., the rate should be preserved throughout the requested connection duration T at the data center's designated value $\bar{R}_D^{(k)}$. The second goal is to minimize the energy consumed by the UE's hardware to receive and decode the eNB's downlink traffic. The key strategy behind reducing the UE receiver's energy consumption is minimizing the number of TTIs where the UE receiver circuit is scheduled to be in active mode. More details about this strategy are provided in Sections II-B and IV. Ideally, in an OFDMA-based system, for the eNB to satisfy the user's requested connection rate requirement, the following constraint must hold all the time:

$$\frac{1}{T}\sum_{m=1}^{M}\sum_{n=1}^{N} B_k(m,n) \geq \bar{R}_D^k \quad \forall\, k \tag{1}$$

where M is the total number of TTIs within a requested connection of total duration T such that $T = MT_{\text{TTI}}$ (i.e., T_{TTI} is equal to 1 ms), $B_k(m,n)$ is the number of received bits by user k during TTI m over RB n, and $\bar{R}_D^{(k)}$ is the data center's requested effective bandwidth for user k within the connection time T. It should be noted that the requested average connection rate $\bar{R}_D^{(k)}$ is selected to accommodate the QoS requirements of multiple traffic buffers for UE k (i.e., $\bar{R}_D^{(k)} = \sum_{h=1}^{H} \bar{R}_D^{(k)}(h)$, where h is the UE's connection index).

The importance of the constraint (1) lies in carefully selecting a suitable time horizon (i.e., further explained later in Section IV when defining τ) during which the scheduler successively allocates resources throughout the requested connection time T. This can be illustrated by looking at the following subconstraint:

$$\frac{1}{\tau}\sum_{m=m_o}^{m_o+G-1}\sum_{n=1}^{N} B_k(m,n) \geq R_D^{(k)} \quad \forall k, m_o \tag{2}$$

where G is the number of TTIs considered in the time horizon τ (i.e., equal to $G \times 1$ ms) during which the RT engine predicts the channel, $m_o \in \{1, 1+G, 1+2G, \ldots, 1+M-G\}$ is the initial TTI index in the observed horizon within the connection time T, $R_D^{(k)}$ is the quasi-instantaneous target rate of user k within the horizon τ.

The key idea behind the subconstraint (2) is that changing the value of τ provides the network with a two fold control on the UE's EE and network's QoS requirements (including the packet delay). For instance, consider a user receiving a delay-sensitive VoIP connection with a total duration of 50 s (i.e., $T = 50$ s) at a standard instantaneous rate of 13.6 kb/s (i.e., $R_D^{(k)} = 13.6$ kb/s). The 13.6 kb/s corresponds to generating a voice packet of 244 bits every 20 ms plus an extra 28 bits for the compressed Internet Protocol/User Datagram Protocol/Real-Time Transfer Protocol header (as discussed in [20]). Satisfying the 13.6 kb/s for $T = 50$ s with the aid of (2) having τ set to any value less or equal to 100 ms (i.e., ten LTE frames) will ensure a packet delay bounded by 100 ms (i.e., VoIP packet delay budget) throughout the 50-s call time. In this case, both of the connection delay and rate requirements are met. In addition to meeting the QoS requirements, utilizing accurate future predictions of the user's CSI by an RT-based mechanism (i.e., explained in Section III) in the 100-ms horizon better optimizes the UE's EE compared with traditional shorter term scheduling.

In the same context, it is also worth noting that the scheduler is always protected by an admission control system which helps the scheduler avoid admitting users' connections over reaching the network's capacity. Thus, after adapting the value of τ accordingly, the eNB's scheduler can safely utilize (2) to support both admitted guaranteed bit rate (GBR) and non-GBR connections.

To simplify the analysis in this paper, we only investigate improving the UE's EE constrained by the GBR requirement (i.e., effective bandwidth) on an instantaneous basis. This framework is supported by selecting the scheduler's granularity (i.e., τ) less than or equal to 100 ms. Thus, the delay analysis is not considered in this paper.

B. UE Circuit Power Consumption

The UE transceiver circuit could be seen as a composition of BB and radio-frequency (RF) stages. A simplified block diagram for those stages is shown in Fig. 3. The components of those stages are the major source of energy consumption inside any cellular device. To investigate the EE of our scheduling scheme, the LTE UE power consumption model developed in [21] is utilized in our analysis to measure the UE energy consumption while in the receive mode (i.e., downlink). The model originally accounts for the power consumption of both of the transmit and receive processing paths. However, in this paper, we only consider the power model of the receiver section. This model was defined as follows [21]:

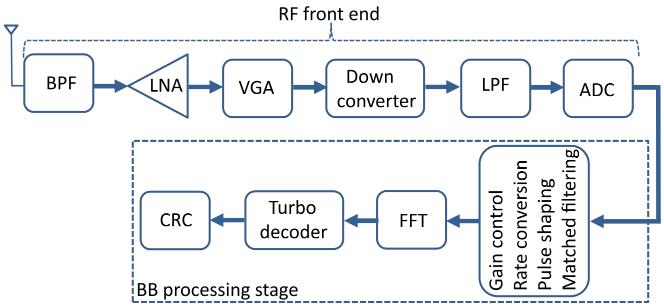

Fig. 3. Simplified block diagram for LTE UE downlink physical-layer processing chain.

$$P_t^{(k)} = \underbrace{m_{\text{idle}} \underbrace{P_{\text{idle}}}_{\text{constant}}}_{\text{UE OFF}} + \underbrace{\overline{m_{\text{idle}}}\left(\underbrace{P_{\text{on}} + P_{\text{rx}}}_{\text{constant}} + \underbrace{P_{\text{BB}} + P_{\text{RF}}}_{\text{variable}}\right)}_{\text{UE ON}} \text{ W} \tag{3}$$

where $P_t^{(k)}$ is the total power consumption of the UE's k receiver circuit, m_{idle} is a logical variable that determines whether the UE is OFF (i.e., idle state) or ON (i.e., wake-up state), P_{idle} is the power consumed when the UE is in the idle state and is equal to 0.5 W, P_{on} is the power consumed when the UE is awake from the idle state and is equal to 1.53 W, P_{rx} is the base power consumed by the receiver circuit while in the operation state and is equal to 0.42 W, P_{BB} is the power consumed by the BB stage of the receiver circuit, and P_{RF} is the same for the RF stage.

Details about the power components in (3) can be found in [21]. For simplicity, and for the rest of this paper, we set P_c to denote the constant power term in (3) that is either equal to P_{idle} (when the UE is OFF) or $P_{\text{on}} + P_{\text{rx}}$ (when the UE is ON), whereas P_k denotes the variable power consumed by UE k when UE is ON and is equal to $P_{\text{BB}} + P_{\text{RF}}$. In [21], both P_{BB} and P_{RF} are modeled by fitting a first-order polynomial to experimental circuit power measurements employing the least-mean-square-error criteria as follows:

$$P_{\text{BB}} = 1.923 + (2.89 \times 10^{-3} \times B_r) \text{ W} \tag{4}$$

$$P_{\text{RF}} = 1.889 - (1.11 \times 10^{-3} \times P_r) \text{ W} \tag{5}$$

where B_r is the downlink bit rate in megabits per second, and P_r is the received signal power in dBm.

C. Channel Model

The multipath fading downlink channel between the eNB and each UE is modeled using the deterministic RT approach [22] with the aid of the RT engine residing in the central processing cloud of Fig. 2. The question of how often (or for how long) the channel is modeled during the user's connection is going to be answered in the following. The received signal power at the UE side is calculated by squaring the vector summation of all of the complex polarized electric field components arriving at the UE antenna. Each polarized field vector, which differs in magnitude and phase, corresponds to a separate received radio ray scattered from objects in the surrounding environment such as buildings, trees, and ground. The total received signal power in the far zone of the transmitting antenna is, therefore, as described in [23, Ch. 2], as follows:

$$P_r^{(k)} = \frac{\lambda^2 \beta}{8\pi\eta_o} \left| \sum_{i=1}^{I} \left[E_{\theta,i}^k \sqrt{\left|G_\theta^k(\theta_i,\phi_i)\right|} e^{j\psi_\theta} + E_{\phi,i}^k \sqrt{\left|G_\phi^k(\theta_i,\phi_i)\right|} e^{j\psi_\phi} \right] \right|^2 \quad (6)$$

where $P_r^{(k)}$ is the total received signal power by the antenna of UE k, η_o is the free-space wave impedance, β is the propagation constant, I is the total number of radio paths received by UE k, $E_{\theta,i}^k$ and $E_{\phi,i}^k$ are the theta and phi components of the electric field associated with the ith radio path received by UE k, $G_\theta^k(\theta_i,\phi_i)$ and $G_\phi^k(\theta_i,\phi_i)$ are the theta and phi components of UE k receiver antenna's gain for the ith path with a direction of arrival of θ_i and ϕ_i, and ψ_θ and ψ_ϕ are the relative phases of the theta and phi components of the far zone electric field.

Each of the E_θ and E_ϕ electric field components in (6) is further resolved into an appropriate pair of polarization components. One component is parallel (i.e., vertically polarized) to the plane of incidence at the reflection (or diffraction) point on an obstacle's surface intercepting the signal path. The other component is perpendicular (i.e., horizontally polarized). Each of the I paths might contain multiple reflection and diffraction points, or even a combination of them, throughout the radio path trip from the eNB antenna to the UE receiver's antenna. More details about the calculation of the polarization components at the reflection and diffraction points can be found in [22]. The RT engine that is capable of tracing the radio signal paths and evaluating their associated fields will be further illustrated later in regard to the proposed downlink scheduling system.

The signal power prediction, provided by the RT engine, is then utilized by the eNB's central scheduler to optimize the users' reception schedule in time (i.e., TTI) and frequency (i.e., RB) in terms of EE while meeting target QoS requirements. That is, knowing the received signal power over each RB during all TTIs within a single frame or across multiple frames will allow the scheduler to determine the received block size by each UE. This information then becomes available on the physical downlink shared channel in every TTI after setting the UE to a specific MCS (see [6, Tab. II]). Based on the received block size and the power consumed by the UE's receiver circuits to receive that block, the scheduler efficiently commands the UE, to turning its circuits ON or OFF during all of the observed TTI. This control scheme will be later formulated and explained in detail in Section IV.

It is also important to highlight that, in LTE, the UE is configured to report the channel quality indicator (CQI) feedback over the physical uplink shared channel [24] to assist the eNB in selecting an appropriate MCS to adopt for the downlink transmissions. However in our model, we assign this task to the RT engine, located in the central processing cloud (as explained in Section II-A), which predicts the downlink channel quality by tracing the radio paths to the user's geographical location. The study of how efficient the RT channel prediction replaces the traditional CQI reporting, in terms of offloading frequency resources and reducing the number of UE transmissions (i.e., feedback time), is beyond the scope of this paper.

III. Proposed Predictive Scheduling System

Advancements in today's mobile data services and applications continue to emerge and grow. The main challenge is that this growth is existing in a highly dynamic radio propagation environment. Consequently, the development of faster and more efficient propagation prediction platforms needed for designing optimized wireless networks in terms of spectral efficiency and EE is becoming more critical. In this context, RT prediction model have provided a promising agile solution with higher accuracy compared with traditional statistical models [25]. Due to its interactive nature that simulates the influence of the surrounding geographical environment on the propagation of radio waves, the RT model has been envisioned to enable real-time applications (e.g., vehicle-to-vehicle communication) that usually experience fast and dramatic change in the channel impulse response.

A. *Ray-Tracing Background*

RT is a deterministic approach that offers accurate modeling for predicting propagation effects in wireless communication channels based on the information of a geographic information systems (GIS) database. This database contains an accurate geometrical and morphological characterization of the objects existing in the propagation environment [22]. The basic mechanism of any RT algorithm is to search for all possible radio paths connecting the transmitter and receiver locations. The searching process must account for all combinations of propagation effects, such as direct line-of-sight (LOS), reflections, diffractions, and arising from the surrounding geographical environment. In the literature, two methods known as shooting and bouncing rays (SBR) method and image method are employed to determine the ray trajectory between the transmitter and the receiver. More details about the two methods can be found in [22]. Regardless of which method is employed in tracing the rays, a vector summation for the emitted field components associated with the received rays is then calculated to evaluate the total received power, as shown in (6).

It can be inferred from the previous paragraph that, in complicated dense environments with many scattering objects, the RT process becomes computationally intensive and time-consuming since it requires a huge number of ray intersection tests. Since the last decade [25], and even recently [26], many efforts have been made to accelerate the RT process mainly through simplifying the geometry of the RT environment. Use of these techniques leads to a fewer number of intersection tests and fast elimination for the redundant rays (i.e., rays which miss the receiver location).

B. *Proposed Ray-Tracing-Assisted Scheduling System*

In this paper, we visualize the RT technique as being a core part of an integrated cellular eNB platform that could be

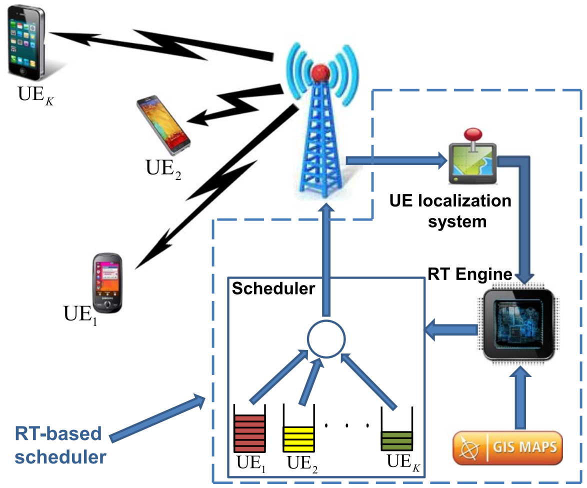

Fig. 4. RT-based downlink scheduler system.

utilized either in the current 4G or tomorrow's 5G networks. This platform is shown in Fig. 4. This promising platform is motivated by two important factors. As shown earlier, the first factor is the intensive research conducted (and still active) in the area of RT acceleration for efficient radio propagation modeling [26]. These efforts have produced numerous efficient acceleration techniques that make the implementation of ray tracers an attractive solution for modeling wireless channels. The second factor is the fast and continuing evolution of today's high-performance computing (HPC) platforms such as field-programmable gate arrays, graphics processing units, and advanced digital signal processors. These platforms have offered powerful solutions to build high-speed RT engines [27].

The predictive downlink scheduler system shown in Fig. 4 depends mainly on the channel's future information provided by the RT engine. The engine is designated for predicting the CSI for all UEs connected to the eNB for longer time intervals compared with traditional sounding of pilot signals [28] that provide short-term measurements. The long-term channel prediction for mobile users is assisted by an accurate localization system and GIS maps database. The function of the localization system is to interactively determine the geographical location of each UE within the cell's coverage GIS map. This ensures that the RT engine calculates an accurate channel SNR based on a real location of the UE within the cell propagation environment. Whether the localization strategy is UE-based, UE-assisted, or network-based [29], we consider the fact that eNB is capable of acquiring the geographical information of the UE along its trip route for a certain time interval in a periodic manner. This information is then utilized by the RT engine to predict the UE's CSI along its registered route (or route section). Hence, smart and seamless integration between the UE localization system and the GIS map with the RT engine is fundamental for building our predictive scheduling system.

To further elicit the relation between Figs. 2 and 4, it should be noted that the RT-based scheduler block highlighted in Fig. 4 represents the detailed structure of the RT-based scheduler located inside the centralized processing cloud of Fig. 2. Thus, the RT engine that predicts the CSI for each UE is a part of the shared architecture explained in Fig. 2. However, just like the BBU pool in the C-RAN model in Fig. 2, a pool of RT processors will also be available within the cloud to be efficiently shared between different cell towers. This way, there will be no need for a dedicated RT engine at each cell tower in the large-scale network.

It is important to highlight that one of the major challenges facing the C-RAN architecture that might affect the decision accuracy of the proposed predictive scheduler system shown in Fig. 4 is the fronthaul latency. The optical link between the eNB tower and the BBU shown in Fig. 2 (known as the fronthaul) introduces a transport network latency that did not originally exist in traditional RAN architecture that has the BBU and the radio tower collocated. Such latency is due to three major sources that are transmission, queuing, and processing of data. Therefore, to maintain the scheduling decision accuracy within one LTE frame of 10-ms duration given that light travels approximately 1 km in 5 μs in fiber, the maximum fiber distance allowed between the BBU and eNB tower should be less than 1000 km (typically less than or equal 20 km [18]) in order to have a round trip transmission delay less than 10 ms. Moreover, various promising solutions (e.g., compression techniques, single-fiber bidirection and wavelength division multiplexing) have been addressed in [18] to reduce the traffic volume over the fiber links and hence the queuing delay. In addition, optimizing the fronthaul queuing delay and its impact on the information flows has been recently addressed in [30]. Finally, the field trials carried out in [18] have showed that the processing delay could be practically less than 1 μs.

IV. Optimal Scheduler

A. General Formulation

Here, the optimal scheduling problem is formulated. The problem's objective as mentioned earlier is to minimize the UE's receiver energy consumption while maintaining the quasi-instantaneous rate for multiple connections per user terminal at a target value. As explained in Section II, the quasi-instantaneous rate constraint for each user connection, which has been designated by the system's central cloud, exclusively accounts for its QoS requirement. The problem constraints are divided into three sets as follows.

1) GBR constraint: Each connection for each user must accomplish fixed quasi-instantaneous transmission rate (i.e., effective bandwidth) throughout the requested connection duration T in time steps of τ (i.e., RT prediction range).
2) Interference constraint: To avoid intracell interference between users, a single RB must be exclusively allocated to a single user every TTI.
3) UE's circuits operation time constraint: To optimize the overall energy consumption for each UE, the scheduler is devoted to finding the optimal allocations, with respect to the energy consumed, in a minimum possible number of TTIs. This will ensure minimal base power consumption (i.e., $P_{\text{on}} + P_{\text{rx}}$) for the user's receiver circuit.

The optimal energy allocation is obtained by solving the following constrained sum-utility minimization:

$$\text{Min}$$

$$E_{\text{tot}} = T_s \sum_{k=1}^{K} w_k \sum_{h=1}^{H} \sum_{m=m_o}^{m_o+G-1} \times \Big(P_k\left(m, \mathcal{N}_{j,k}(m)\right) \Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) + P_c \Phi_k(m)\Big) \quad (7a)$$

$$\text{subject}$$

$$\sum_{m=m_o}^{m_o+G-1} B_k^h\left(m, \mathcal{N}_{j,k}(m)\right) \Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) \geq \tau\, R_D^{(k)}(h) \quad \forall k, h \quad (7b)$$

$$\bigcap_{k=1}^{K} \Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) = \phi \quad \forall m, h \quad (7c)$$

$$\Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) - \Phi_k(m) \leq 0 \quad \forall k, m, h \quad (7d)$$

where E_{tot} is the total energy consumed for all users over an observation period of G TTIs, w_k is a weighting factor for UE k, T_s is the TTI duration (i.e., 1 ms), $P_k(m, \mathcal{N}_{j,k}(m))$ is the total power consumption of the BB and RF receiver circuits for UE k during TTI m over the set of RBs $\mathcal{N}_{j,k}(m)$, $\mathcal{N}_{j,k}(m)$ is the set of RBs j assigned to UE k during TTI m, $\Psi^{h}_{\mathcal{N}_{j,k}(m)}(m)$ is a binary decision variable which indicates whether the set of RBs j is allocated to connection h of UE k during TTI m or not, P_c is the UE's receiver constant power consumption during each TTI that depends on the UE's operation state (i.e., P_{idle} for OFF state and $P_{\text{on}} + P_{\text{rx}}$ for ON state), $\Phi_k(m)$ is a binary indicator that determines whether UE k receiver circuit is in an active state during TTI m or not, $B_k^h(m, \mathcal{N}_{j,k}(m))$ is the number of scheduled transmitted bits in the downlink for connection h of user k over RBs set j during TTI m, τ is the scheduler's time granularity (or time step) that is related to the RT engine prediction range over G TTIs, and $R_D^{(k)}(h)$ is the data center's equivalent instantaneous transmission rate for connection h of user k.

The first decision variable $\Psi^{h}_{\mathcal{N}_{j,k}(m)}(m)$ in the cost function of (7a) allows the scheduler to optimally adjust the MCS of the RBs set allocated to each user in each TTI to minimize the UE's BB and RF circuits energy consumption (as explained in Section II-B). The second variable $\Phi_k(m)$ is devoted to minimizing the number of scheduled wake-up TTIs for each UE to receive its designated data bits. As could be inferred from (7a), this is achieved through penalizing the cost function by the UE's constant power consumption value P_c for each scheduled (i.e., wake up) TTI (per each user), irrespective of the number of RBs allocated within each TTI.

The constraint defined in (7b) resembles the quasi-instantaneous rate (i.e., effective bandwidth) constraint for each user connection that is assumed to exclusively satisfy its QoS requirements as described in Section II. The second constraint in (7c) ensures that each user is assigned to a unique set of RBs (i.e., not intersecting with other users' sets) during each TTI and hence avoid intracell interference between users. For a TTI m having N available RBs, the number of possible RB sets with sizes of $1, 2, \ldots, N$, from which the scheduler searches for each user, is equal to $\sum_{q=1}^{N-1} {}^{N}c_q + 1$. The constraint in (7d) is the well known IF–THEN constraint that is designed to ensure that the binary variable $\Phi_k(m)$ penalizes the cost function by P_c if a user is assigned to any set of RBs within TTI m.

B. Penalty Method-Based Formulation

In practice, the scheduler may not able to satisfy the rate constraint in (7b) for all users all the time, particularly in situations of deep fading (or outage) channel conditions. In other words, as a result of the time-varying nature (i.e., due to the multipath fading) of the users' channel, which temporarily limits its capacity to accommodate their rate constraints, the optimization problem in (7) could be unfeasible in different time intervals. Since we do not consider admission control procedures in this paper to handle the issue of unbalanced demand versus available resources within the duration of an admitted connection, the penalty method [31] is utilized to ensure feasible solutions for the optimization problem in (7) by relaxing the constraint (7b).

The penalty method is known to approximate the solution of constrained optimization problem by iteratively solving a series of dependent unconstrained problems whose solutions ideally converge to the original constrained problem. The dependence implies that the solution of each unconstrained problem in each iteration affects the following one. More specifically, each unconstrained problem adds a penalty term, also known as penalty function, to the objective function of the following problem in a successive manner until the penalty function converges to zero. The penalty function value represents how far the current solution is from that of the original constrained problem. In our problem, the penalty method is used to relax the constraint of (7b) (i.e., reversing the inequality sign) by adding a new term in the objective function. The new term role is to push the new unconstrained formulation to converge to the solution of the original constrained formulation as much as possible. The new unconstrained formulation can be illustrated as follows:

$$\text{Min}$$

$$Z_1 = E_{\text{tot}} + \sum_{k=1}^{K} \sum_{h=1}^{H} \alpha_{k,h} \times \left(\left(\tau R_D^{(k)}(h) + \ell(k,h)\big|_{m_o-G}^{m_o-1}\right) \Omega_k - \sum_{m=m_o}^{m_o+G-1} B_k^h\left(m, \mathcal{N}_{j,k}(m)\right) \Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) \right) \quad (8a)$$

$$\text{subject}$$

$$\sum_{m=m_o}^{m_o+G-1} B_k^h\left(m, \mathcal{N}_{j,k}(m)\right) \Psi^{h}_{\mathcal{N}_{j,k}(m)}(m) \leq \tau R_D^{(k)}(h) + \ell(k,h)\big|_{m_o-G}^{m_o-1} \quad \forall k, h \quad (8b)$$

$$(7c), (7d) \quad (8c)$$

$$\Omega_k > 0 \quad \forall k \quad (8d)$$

where Z_1 is the new unconstrained objective function, E_{tot} is the original constrained objective function that was given in (7a), $\alpha_{k,h}$ is a QoS optimization weighting factor for prioritizing traffic connection h of user k, $\ell(k,h)|_{m_o-G}^{m_o-1}$ is the leftover unscheduled bits from the previous G TTIs, and Ω_k is a new binary decision variable accounting for the total number of bits that user k requires to receive within the current G TTIs.

It is obvious in (8a) that the penalty function added to the constrained problem presented in (7) is the whole term added to E_{tot}. The key idea of the new formulation highlighted in (8) can be understood as follows: Any remaining bits for a certain user connection that has not been transmitted within the previous G TTIs will be accumulated as leftover bits (i.e., $\ell(k,h)|_{m_o-G}^{m_o-1}$) to the original bits (i.e., $\tau R_D^{(k)}(h)$) awaiting transmission in the current G TTIs. After a few iterations (corresponds to time delay), the leftover bits for the same connection are converged to zero. This is clearly understood from (8a) as minimizing the difference (i.e., for each user k) between what is demanded (i.e., $\tau R_D^{(k)}(h) + \ell(k,h)|_{m_o-G}^{m_o-1}$), and what is available and granted (i.e., $\sum_{m=m_o}^{m_o+G-1} B_k^h(m,\mathcal{N}_{j,k}(m))\Psi_{\mathcal{N}_{j,k}(m)}^h(m)$) will lead to a close solution of the constrained problem, albeit with a certain amount of delay. In addition, dynamically configuring the weighting factor $\alpha_{k,h}$ for each UE and each connection per UE enables class-based packet scheduling [20]. Thus, continuously prioritizing users (i.e., interscheduling) based on a certain fairness criteria, and prioritizing different traffic buffers (intrascheduling) for the same user based on their QCI priority index could be easily attained. From another perspective, the ratio of the weighting factors α to w for each user determines the amount of effort the scheduler spends to satisfy user's connection rate constraint (i.e., QoS) to that spent to minimize the amount of energy consumed (i.e., E_{tot}), respectively. Moreover, it has to be noted from (8d) that the decision variable Ω_k is constantly set to 1 to force its term—which resembles the total number of bits awaiting transmission for each user—to always appear in the cost function in (8a).

C. Practical ON–OFF Formulation

Looking back to (4) and (5), we found that changing the scheduled MCS, for any user, within a single TTI (i.e., 1 ms) has a marginal effect in terms of energy consumption compared with deciding whether to turn UE's circuits ON or OFF. To illustrate this comparison, we present the following numerical example: Assume a single cell operating with the full LTE bandwidth of 20 MHz with a total of 100 RBs available in each TTI. The 100 RBs are assumed allocated to a single user within a single TTI. The user's effective SNR is assumed 20 dB. With the aid of (4) and (5) and [6, Tab. II], the total power consumption (i.e., $P_k(m,\mathcal{N}_{j,k}(m))$) of the UE within the observed TTI when all RBs are configured to MCS index 1 is equal to 3.92 W. Whereas, in the case of MCS index 15, the total power is 4.11 W. Thus, there is a maximum of 4.62% reduction in the UE's power consumption if the scheduler coarsely changes the MCS index over the allocated RBs from index 15 to index 1. On the other hand, turning UE's circuits ON and OFF in each TTI affects the UE's power consumption budget by a value 1.45 W (i.e., $P_{\text{on}} + P_{\text{rx}} - P_{\text{idle}}$) that is almost equivalent to 35.28% of the total power consumed by the BB and RF circuits.

Based on the finding of the previous example and to simplify the scheduler's formulation, we further modify the formulation in (8) by removing the term $P_k(m,\mathcal{N}_{j,k}(m))$ from the cost function and letting the scheduler focuses only on minimizing the number of wake-up TTIs as it offers a remarkable reduction in the UE's energy consumption. Thus, the formulation in (8) can be rewritten as follows:

Min

$$Z_2 = T_s \sum_{k=1}^{K} \sum_{m=m_o}^{m_o+G-1} w_k P_c \Phi_k(m) + \sum_{k=1}^{K}\sum_{h=1}^{H} \alpha_{k,h} \times \left(\left(\tau R_D^{(k)}(h) + \ell(k,h)|_{m_o-G}^{m_o-1}\right)\Omega_k - \sum_{m=m_o}^{m_o+G-1} \sum_{n=1}^{N} B_k^h(m,n)\Psi_k^h(m,n) \right) \tag{9a}$$

subject

$$\sum_{m=m_o}^{m_o+G-1} \sum_{n=1}^{N} B_k^h(m,n)\Psi_k^h(m,n) \le \tau R_D^{(k)}(h) + \ell(k,h)|_{m_o-G}^{m_o-1} \quad \forall k,h \tag{9b}$$

$$\sum_{k=1}^{K} \Psi_k^h(m,n) \le 1 \quad \forall m,n,h \tag{9c}$$

$$\Psi_k^h(m,n) - \Phi_k(m) \le 0 \quad \forall k,h,m,n \tag{9d}$$

$$(8\text{d}). \tag{9e}$$

As explained earlier, the modified cost function in (9a) focuses mainly on minimizing the number of scheduled wake-up TTIs for each UE while maintaining the connection's target transmission rate throughout the connection duration T. The reader can also notice another difference between (9) and (8). This difference is changing the notations of the variables $\Psi_{\mathcal{N}_{j,k}(m)}^h(m)$ and $B_k^h(m,\mathcal{N}_{j,k}(m))$ that appeared in (8) to $\Psi_k^h(m,n)$ and $B_k^h(m,n)$ in (9). In particular, the scheduler's simplified formulation in (9) does not care about the sets of RBs allocated in every TTI for each user to optimize the UE's BB and RF power consumption (i.e., $P_k(m,\mathcal{N}_{j,k}(m))$) as was the case in (8). This is because the practical design for LTE scheduler suggests that, for each UE, all RBs within the same subframe are preferably adjusted to a fixed MCS [24]. Therefore, the new decision variable $\Psi_k^h(m,n)$ introduced in (9), which substantially reduces the solution space, is designated only to identify the number of scheduled transmitted bits over each individual allocated RB for each user (i.e., $B_k^h(m,n)$) to meet the user's connection target rate. Hence, we define $\Psi_k^h(m,n)$ as a binary decision variable that indicates whether connection h of user k has been assigned to RB n during TTI m or not and $B_k^h(m,n)$ as the number of allocated bits for connection h of user k over RB n during TTI m.

V. Heuristic Scheduler

Here, we design a heuristic algorithm for simplifying the solution of the optimization problem defined in (9). This is derived by the high complexity inherent in the optimal model particularly for large values of τ, which strictly determines the problem's solution space. To provide an approximate figure about the complexity of solving the problem in (9), we provide the following timing measurement for a small-scale problem. Consider three users each with a single connection and three available RBs at each TTI. In an attempt to solve problem (9) with a value of $\tau = 10$ ms (i.e., one frame of scheduling granularity) over a total number of frames $M = 5000$ (i.e., $T = 50$ s), the MATLAB Profiler recorded a total elapsed time of 565.45 h (i.e., 23.5 days). The MATLAB was running on an Intel Xeon CPU W3670 with six-core processors running at 3.2 GHz and 16 GB of RAM. On the other hand, our proposed heuristic was able to solve the same problem in just a few seconds.

A. Heuristic Algorithm

The proposed heuristic algorithm shown in Fig. 5 is fed by an initialization part, labeled by 1, which is created to set the parameters of the considered scenario. These parameters include: number of users K, number of connections for each user H, total number of TTIs considered for the users connections M, number of available RBs per TTI N, scheduler's granularity in TTIs G (i.e., taken in multiples of a frame), and a quasi-instantaneous target rate for each user connection $R_D^{(k)}(h)$. Furthermore, to calculate the total power consumed by each user (i.e., $P_t^{(k)}$) every TTI using the model described in (3). First P_r [i.e., in (5)] is calculated each TTI for each user by employing a reference channel model (i.e., discussed in detail in the following). Second, using [6, Tab. II] and based on the user's generated channel SNR, B_r [i.e., in (4)] is calculated accordingly each TTI. It should be noted that part 1 of Fig. 5 will be also used when considering solving the optimal problem in (9).

The second part of Fig. 5, which is labeled 2, represents the core heuristic algorithm. The algorithm is designed to run in a sequential manner on the requested user connections. In other words, the algorithm allocates resources for one user connection at a time. The connections scheduling sequence order is determined by the parameter $\alpha_{k,h}$ for each connection. In our heuristic, the parameter $\alpha_{k,h}$ is set to be proportional to the user's traffic queue length, which reflects the priority of scheduling that queue. Obviously, the length for each user queue is continuously changing in time based on the number of allocated resources and their respective capacities during each scheduling period τ. Therefore, large queue length is equivalent to large value of $\alpha_{k,h}$; hence, higher priority is allotted compared with smaller length queue.

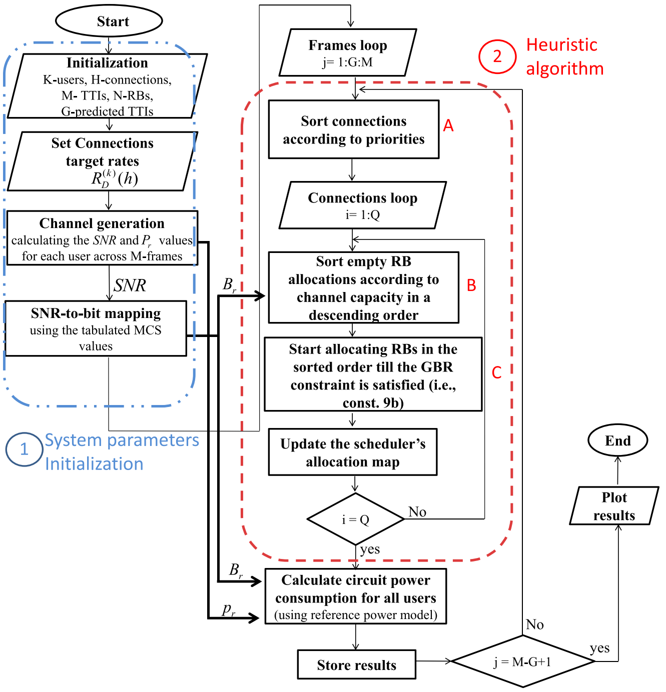

Fig. 5. Heuristic algorithm flowchart for the proposed scheduler.

The algorithm works as follows. For each G TTIs, that is equivalent to τ seconds from the total connection time T, all users connections are sorted according to their corresponding queue lengths. Then, for each connection according to the sorted order, the algorithm aims to allocate the minimum number of RBs within the smallest possible number of TTIs, which gives a total number of scheduled bits less than or equal to the target bits (i.e., equivalent to constraint 9b). This is done by sorting the unallocated RBs within the observed TTIs in descending order according to their capacities (i.e., number of received bits per RB). Sorting the RBs in this way results in both minimum numbers of allocations and TTIs. This is due to the fact that, for each UE, all RBs capacities (which correspond to different MCSs) are set to be equal to the lowest RB capacity (i.e., MCS index) within the same TTI. In the LTE standard, it is known to be highly efficient to reduce the signaling overhead by making the UE's receiver (or transmitter) circuit adjusted to a fixed MCS rather than using frequency-dependent MCS at a given subframe (see [24, Sec. 10.2]). In this case, the scheduler is strictly enforced to fully allocate RBs in each TTI before moving to another TTI and hence minimizing the overall number of wake-up TTIs and circuit energy consumption. The scheduler then updates the allocation map for all RBs across the observed time slots with the current connection allocations before proceeding with the next connection in the sorted listed. The algorithm continues until all connections in the sorted list are served. Based on the RB allocations for each user, the receiver's circuit power consumption is calculated as in (3) during the current scheduling period and then stored for later analysis. The whole algorithm continues in the same fashion for the next scheduling period (i.e., next G TTIs) until the last frame in the established connection time.

B. Complexity Evaluation

For the sake of assessing the algorithm complexity, we divide it into three major processing components labeled as: A, B, and

C as shown in Fig. 5. Those components hold the main operations constituting the algorithm. Component A, as explained previously, is responsible for sorting the admitted users' connections based on their queue lengths. Hence, the complexity of component A is equal to $O(Q\log(Q))$, where Q is the total number of connections for all admitted users. Component B sorts empty RB allocations to rank potential assignments to the observed connection. This requires first searching the indexes of unallocated RBs during the observed TTIs (i.e., of number G) then sorting them according to their capacities. Therefore, the worst-case complexity for component B, when allocating resources for the first connection in the sorted list, is equal to $O(NG) + O(NG\log(NG))$, where NG is the total number of RB allocations in G TTIs. Finally, component C keeps checking capacities for all unallocated RBs within the observed G TTIs to make final decisions about RB allocations, which satisfy the user's connection buffer requirement (i.e., equivalent to constraint 9b). This results in an upper bound complexity of $O(NG\log(NG))$. In summation, the asymptotic upper limit for the algorithm complexity when allocating resources for a total of Q connections can be approximated by $O(Q\log(Q))$ when $Q \gg NG$ or $O(NG\log(NG))$ when $NG \gg Q$.

VI. Numerical Results

Here, the performance of the proposed predictive scheduling scheme is compared with the GRA scheme proposed in [13] through MATLAB numerical simulations. To the best of our knowledge, the GRA is the only reported scheme that has been designed for optimizing the UE EE in downlink OFDMA systems and hence is exclusively considered in our comparison. The numerical simulations are conducted in two different scenarios. In the first scenario, the channel is modeled as a quasi-static block Rayleigh fading (QSBR) channel [32]. The channel is assumed constant within the 180-KHz band of each RB during each TTI. However, it changes randomly and independently from one subband to another (i.e., frequency selectivity) and from one TTI to another (i.e., time selectivity). Each UE is also assumed to experience independent fading. Considering the Rayleigh fading in this scenario, the distribution of the instantaneous (i.e., taken every subframe) received channel SNR over each RB follows the exponential distribution [33]. The second scenario, as shown in Fig. 6, utilizes a real RT measurements for the propagation channel of mobile UEs located in a part of Ottawa (i.e., north of centretown), Canada. This location represents an area in downtown Ottawa that includes Gloucester St., Laurier Ave W, Slater St., Albert St., Queen St., Sparks St., Lyon St. N, Kent St., Bank St., O'Connor St., Metcalfe St., and Rue Elgin St. Fig. 6(a) shows the actual RT experiment carried out using the Remcom's Wireless Insite tool [34].

The MATLAB model for both scenarios is implemented as shown in the flowchart of Fig. 5. The only two variable parts are the channel generation block located inside the initialization part labeled by 1 and the heuristic algorithm labeled by 2. The channel generation block is based on the simulation scenario (i.e., channel model) being considered, whereas the second part is based on the scheduler type (i.e., proposed optimal, proposed heuristic, GRA optimal, and GRA heuristic). In other words, in the case of scenario 1, the RT-based scheduler highlighted in Fig. 4 is assumed to have the same channel measurements as those generated by the QSBR model. In the case of scenario 2, the RT-based scheduler is fed by real RT measurements conducted by the Wireless Insite tool experiment shown in Fig. 6(a). MATLAB and the Insite tool are directly interfaced via a piece of MATLAB script.

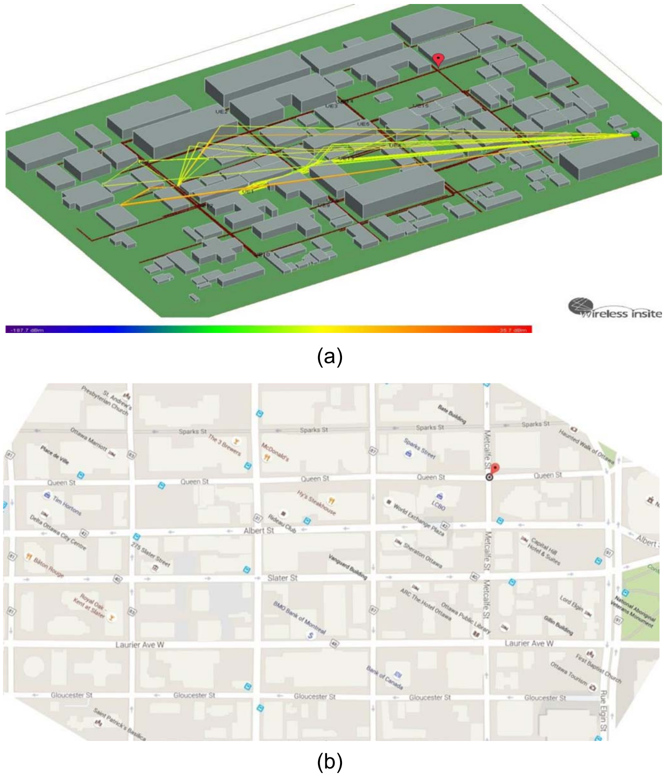

Fig. 6. Three-dimensional RT experiment for part of Ottawa City. (a) RT 3-D view. (b) Google Map top view.

For both scenarios of the simulation, all users are assumed located within a single-cell coverage and receiving downlink connections from its serving eNB. In scenario 1, two kinds of investigation were undertaken. The first focuses on the performance and complexity comparison of the optimal and heuristic versions of the proposed scheduler—described in (9) and Section V, respectively—to that of the GRA scheduler. For this and due to the high computational burden and latency of the optimal scheduler, a relatively small number of UEs (i.e., set to be equal or less to the number of available RBs) are admitted to request downlink connections from the eNB for a small number of frames (i.e., the simulation time). Having the proposed heuristic algorithm benchmarked, the following investigation is devoted to the study of the system capacity variations and users' buffers queue stability only for the heuristic versions of the proposed and the GRA schedulers. Thus, the number of admitted UEs and their requested connection duration are allowed to be increased while keeping the number of available resources constant. This increase can be easily handled due to the dramatic speed-up and simplicity of the heuristic scheduler compared with its optimal counterpart. The evaluation of the proposed scheduler takes place at different RT prediction ranges (or scheduling granularity) to show its impact on the scheduler's overall performance. For the sake of simplicity and not exceeding the maximum page allowance, only the second investigation has been carried out in scenario 2.

A. Scenario 1: QSBR Channel

As described earlier, scenario 1 utilizes the QSBR channel model for setting the UEs' propagation statistics. In MATLAB, we generate independent and identically distributed (i.i.d.) random variables. Each random variable, which models the downlink channel SNR values for a certain UE over a single RB across the selected M frames, has an exponential distribution with an assumed average of 10 dB. Since the adopted QSBR channel is known to model scattering environments with multiple path propagation, we use the Rayleigh channel model with the covariance function in the form of [32]

$$R_{\xi}(t_s) = J_0(2\pi f_d t_s) \tag{10}$$

where ξ is the channel Gaussian process, t_s is the channel sampling time, J_0 is the Bessel function of the first kind with order 0, and f_d is the Doppler frequency. To determine a proper value for the channel coherence time τ_{coh}, (10) should be evaluated at different speeds of the mobile terminal and the operating carrier frequency (i.e., taken 2.6 GHz). Taking 0.5 as a threshold value for the SNR correlation coefficient to determine the channel coherence time, one will find that any speed greater than or equal to 100 km/h leads to $\tau_{\text{coh}} = 1$ ms. As a result, 1 ms is assumed the sampling time for the generated random variables.

1) Performance and Complexity Comparison of the Proposed Scheme With the GRA Scheme: Here, we compare the EE performance and complexity of the optimal and heuristic versions for both of the proposed and GRA schemes. It is worth noting that the GRA scheme does not utilize the RT engine knowledge about the users' CSI, as is the case with the proposed scheme. As discussed earlier, the EE is compared in conjunction with satisfying a quasi-instantaneous target rate for each UE. Due to the inherent complexity of the optimal solution, for both schemes, we run the optimal schedulers for only 200 frames to find the global optimal allocations and compare them with those in the case of the heuristic. We also assume a small-size system that allows only three UEs, each with single downlink connection with a unique required rate and competing over three available RBs. The selected rates are 13.3, 64, and 128 kb/s that typically support VoIP, audio streaming, and FTP connections, respectively.

Fig. 7 shows how the proposed scheduling scheme performs and compares with the GRA scheme in terms of the achieved EE, particularly when increasing the scheduling time granularity (i.e., measured in frames). Having both schedulers satisfying the rate requirements as shown in Fig. 8, it can be seen that the proposed scheduler shows a significant EE improvement. This improvement is shown in Fig. 7,particularly when acquiring greater knowledge about the UE's CSI, which leads to increasing the optimization problem's solution space.

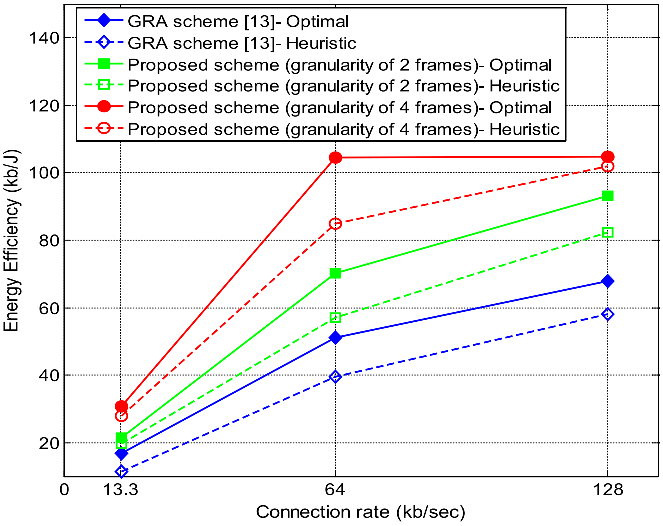

Fig. 7. Energy efficiency comparison for optimal versus heuristic schedulers.

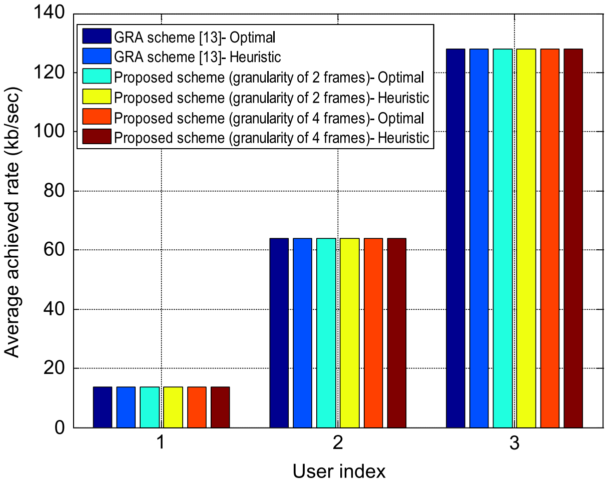

Fig. 8. Downlink rate satisfaction for optimal versus heuristic schedulers.

It is also noticed that the EE of the higher rate connections is greater than that of the lower rate. This is due to the fact that, for higher rate connections, the scheduler relatively allocates more resources within each TTI to support its high volume of data. Thus, the number of bits received for every joule consumed by the UE per TTI becomes larger on average. However, it should be noticed that the UE with the 128-kb/s connection have similar EE to that of the 64-kb/s UE in case of the proposed scheme at four frames of scheduling granularity (i.e., last two points on the red solid line). This can be attributed to the scheduler's increased ability to fully satisfy the connection rate requirement for the 64-kb/s UE at four frames of granularity while spending half the energy consumed by 128-kb/s UE.

From another perspective, Fig. 9 provides more insight on the increased capability, at larger time granularity, of the proposed scheduler for satisfying the quasi-instantaneous rate (i.e., measured every 80 ms) of the 13.6-kb/s connection.

Fig. 10 shows the increase in the lifetime of batteries in case of the proposed scheme (as function of the scheduling granularity) compared with the GRA scheme for different UEs.

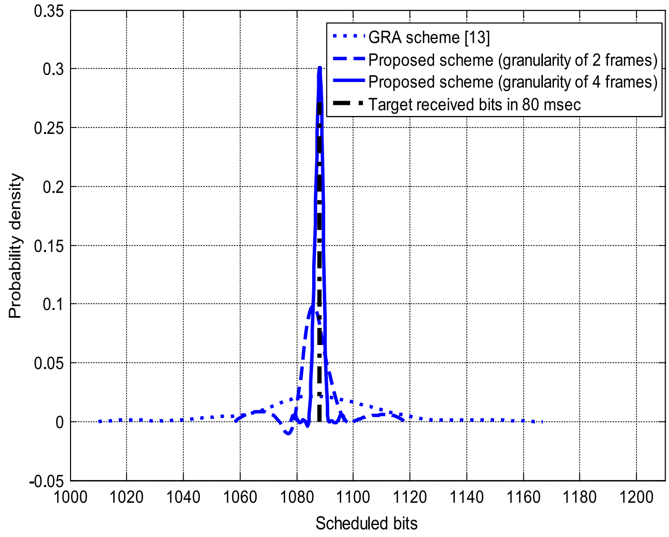

Fig. 9. Optimal scheduler capability for satisfying the 13.6-Kb/s connection at different scheduling granularities.

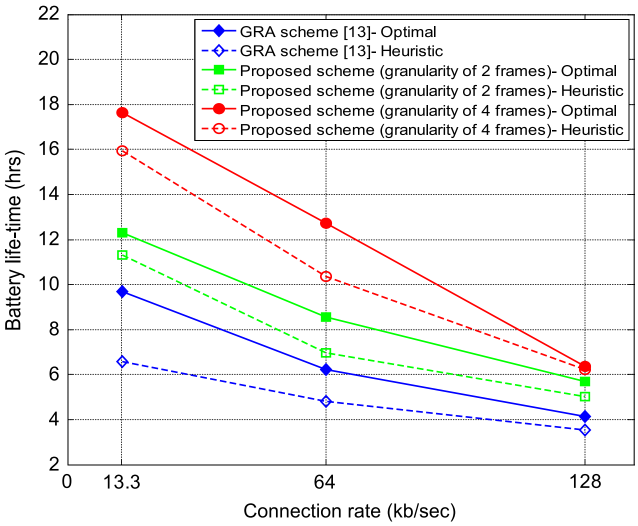

Fig. 10. UE battery lifetime.

TABLE I
COMPLEXITY COMPARISON

Scheme	Optimal time complexity	Heuristic time complexity
GRA	56.2 hrs	184 ms
Proposed	92.9 hrs @ granularity of 2 96.7 hrs @ granularity of 4	213.3 ms @ granularity of 2 267 ms @ granularity of 4

It is worth noting that the lifetime values obtained are strictly dependent on the energy consumption by the UE while receiving data in the downlink and an assumed battery capacity of 2915 mAh. Thus, the lifetime results shown in Fig. 10 do not account for any other factors that are known to deplete the cell phone battery (e.g., running any kind of applications, uplink transmission, synchronization with the eNB, etc).

On the other hand, the complexity of the proposed scheme is compared with the GRA scheme in terms of the computation time spent by the CPU to solve the allocation problem. The computation times, as recorded by the MATLAB Profiler, are shown in Table I.

2) Buffer Queue Stability With System Overloading in Case of Heuristic Schedulers: After benchmarking the heuristic schedulers for both of the proposed and the GRA schemes in part 1 of the results, here, we only consider the heuristic schedulers (for both schemes) for investigating the system stability when stressing the system by increasing the number of UEs. This is due to the substantial speed-up for the heuristic algorithm compared with solving the problem's optimal formulation. We also increase the value of M to 5000 frames (with the same 10-ms frame duration). For evaluating high-data-rate applications, the number of available resources per TTI is increased to 25 (i.e., equivalent to the 5-MHz LTE channel) instead of 3 as used in the previous part. The number of UEs is allowed to increase from 3 to 15 in a step of 1. In addition, a granularity of eight frames for the scheduler is added to the test to shed light on the bigger picture for the system's behavior. All UEs are assumed to have a single downlink connection with a rate of 400 kb/s (i.e., recommended bit rate for a 240p YouTube live stream). It is worth noting that the delay requirement for each connection is considered in this analysis as highlighted in the last part of Section II-A. In particular, the delay is bounded by satisfying the effective quasi-instantaneous rate for all the requested connections within a time horizon that does not exceed 100 ms (i.e., the typical packet delay budget for various services such as conversational voice, real-time video, and games [16]). In addition, the packet delay jitter that is a crucial QoS parameter for certain traffic types (e.g., VoIP and online gaming) is not considered in this paper. However, a more generalized system model and optimization framework is currently under development, which accounts for modeling and controlling the packet delay jitter for jitter sensitive applications.

The results shown in Fig. 11 show how the UE buffer queue length increases with the number of UEs for both of the proposed scheme and the GRA scheme [13]. This increase is a direct result of overloading the system available resources with a potentially increasing number of connections. The results confirm that the proposed scheme outperforms the GRA scheme particularly when increasing the scheduling time granularity. This increase leads to maintaining the average queue length per UE at an acceptable level for a larger number of admitted UEs. More specifically, for the GRA scheme, it is clear that the instability point (i.e., the point at which the UE buffer length grows without bound) appears at a smaller number of UEs compared with the proposed scheme. As a result, the proposed scheme is capable of increasing the system's resistance to instability and hence potentially increases the system's capacity to admit more users. This is obvious for the proposed scheme with eight frames of granularity that can support up to eight UEs compared with the GRA scheme that can only support up to five UEs while having an equal number of available resources. In addition, the results shown in Fig. 11 confirm the idea explained in Fig. 1. Our predictive scheduler is capable of meeting the QoS requirements while allocating a fewer number of resources when operating at higher time granularity, thus having higher UEs' admittance capacity.

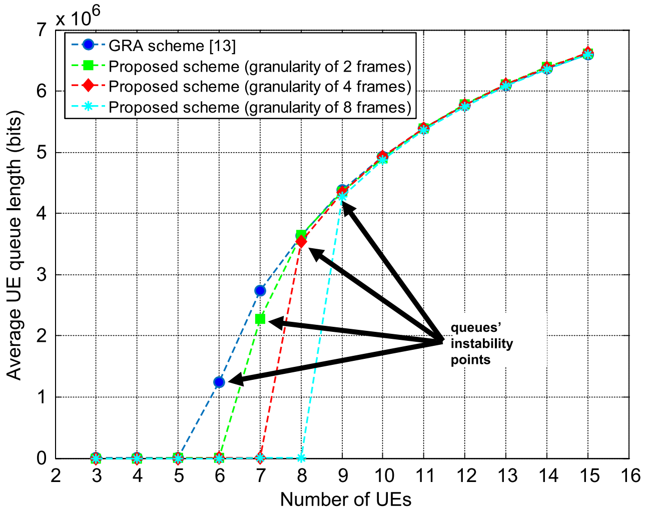

Fig. 11. UE's buffer queue stability.

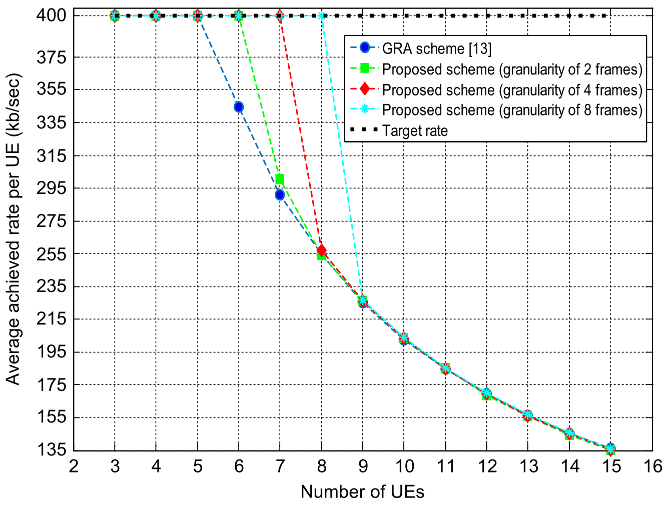

Fig. 12. UE's average achieved rate.

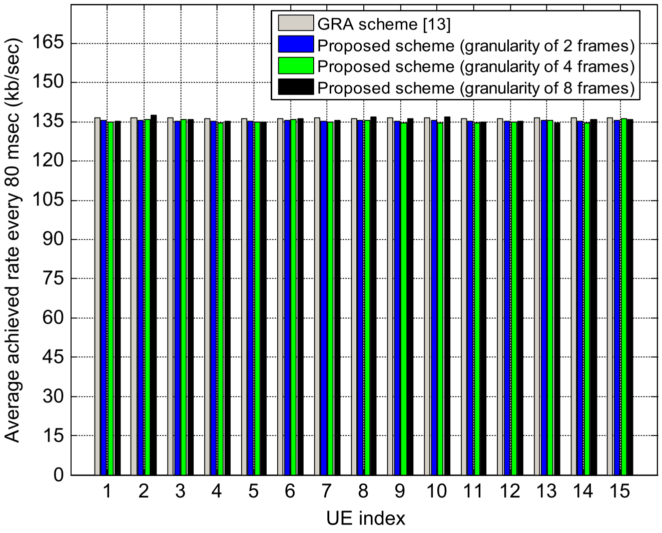

Fig. 13. Distribution of the total cell rate in case of 15 UEs.

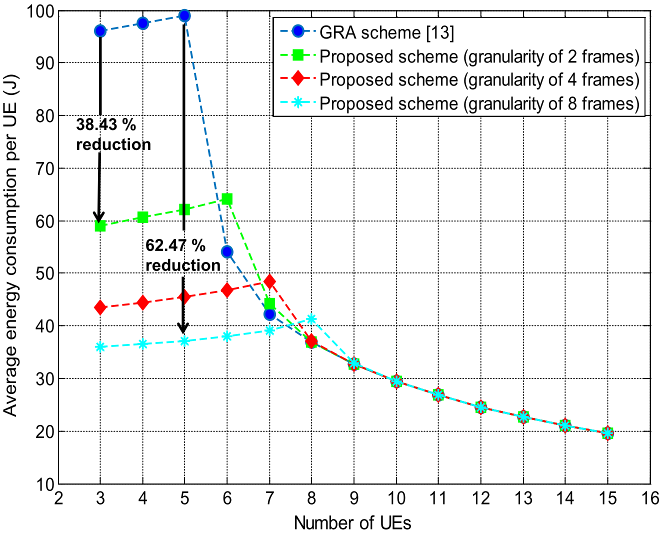

Fig. 14. Average energy consumption per UE.

As a consequence of the queue stability regions shown in Fig. 11, the average achieved rate per UE depicted in Fig. 12 could be directly justified. In other words, the deviation of the average achieved rate per UE from the target (i.e., 400 kb/s) as the number of UEs grows reflects the growth in the queue length noticed in Fig. 11. However, the system capacity is the same (i.e., the total cell rate), as it should be, although the average rate per UE is changing with the number of UEs. From another perspective, the distribution of the total cell rate among different UEs, as shown in Fig. 13 for the case of 15 UEs, evaluates the inherent fairness property of our proposed heuristic algorithm among different UEs with the same traffic connection requirements. This property is implicitly understood from the explanation of the algorithm provided in Section V. In particular, the uniform distribution of the total cell rate among all UEs (i.e., taken as a fairness indicator) shown in Fig. 13 is the result of continuously changing the scheduling priority among users having the same service requirements based on their varying queue lengths (i.e., equivalent to the throughput history). The same strategy is used in the well-known proportional-fairness scheduling policy. It is also worth noting that the fairness is considered in the optimal model described in (9) by setting equal values to the weighting factor $\alpha_{k,h}$ to connections having the same service requirements across different UEs.

In addition to its ability of increasing the number of serviced UEs, the EE improvement provided by our proposed scheme is noticed. This is shown in Figs. 12 and 14. We start by focusing on the common stability region for all curves where the number of UEs increases from 3 to 5. A significant increase in the EE is clearly noticed in Fig. 14 for the proposed scheme over the GRA scheme. This increase ranges from 38.43% at granularity of 2 frames and to 62.47% at granularity of 8 frames. This is due to a substantial drop in the energy consumption by the same percentages, as noticed in Fig. 14, while almost having the same average rate per UE as shown in Fig. 12.

Another conclusion could be drawn from Fig. 14. It could be noticed that, for each curve, the energy consumption slightly increases as the number of UEs grows until it reaches the

TABLE II
SIMULATION PARAMETERS

Parameter	Setting
Number of UEs	15
UE speed	50 Km/h
UE route length	690 m
Number of available RBs	3
Operating frequency	2.6 GHz
Available MCS	refer to Table II in [6]
Channel model	RT using SBR technique
Layout	urban with 1 microcell
Cell dimensions	1016 m (L) x 673 m (W)
Building walls relative permittivity (ε_r)	3 **[35]**
Building walls conductivity (σ)	0.005 S/m **[35]**
Building walls thickness	20 cm
Ground permittivity (ε_r)	15 **[35]**
Ground conductivity (σ)	7 S/m **[35]**
eNB antenna type	vertical isotropic
eNB antenna height (above the ground)	57 m
eNB transmit antenna input power	48 dBm
UE antenna height (above the ground)	2 m
Simulation time	50 sec (*i.e.*, 5000 frames)

maximum value before the corresponding instability point (e.g., six UEs for the GRA scheme curve). This is due the higher load experienced by the system's frequency resources. In other words, when the system is stable (or relaxed), increasing the number of UEs slightly limit the energy optimization due to the increasing load on the limited available resources. Hence, the average energy consumed per UE shows a slight increase (i.e., lower optimization efficiency). However, that effect becomes less pronounced in the case of the proposed scheme as the scheduling granularity increases, compared with the GRA scheme. This can be seen when comparing the rising slopes of the GRA scheme and the proposed scheme with granularity of eight frames curves in Fig. 14.

B. Scenario 2: RT-Based Channel

In this scenario, we examine a practical use case for measuring the performance of our scheduling scheme in comparison with the GRA scheme. We used a commercial radio propagation prediction software named Wireless Insite that is offered by Remcom Inc. [34] to build a realistic 3-D urban scenario, as shown in Fig. 6(a). The scenario detailed parameters are listed in Table II. The reason behind conducting this experiment with real RT measurements for the propagation channel is to provide insight on the performance bounds of our scheduling scheme with two different channel models, one of which is based on a practical scenario. Therefore, the results of scenario 2 should be looked at in comparison with that in scenario 1, which utilizes one of the common channel models used in the literature (i.e., QSBR model).

Unlike the results obtained in Fig. 11, both of the buffers' stability performance and the system's admittance capacity for the proposed scheme showed a slight improvement over the GRA scheme, as demonstrated in Fig. 15. We attribute this different behavior due to the slow fading channel of the chosen urban scenario where the UEs are moving with a speed of $V_{\mathrm{UE}} = 50$ km/h. This slow speed results in much greater channel coherence time (i.e., $\tau_{\mathrm{coh}} > 10$ ms) compared with that used in scenario 1 (i.e., $\tau_{\mathrm{coh}} = 1$ ms). Consequently, our proposed predictive scheduling scheme is not able to exploit better scheduling chances, as is the case in simulation scenario 1, particularly at the selected low granularities. However, it is still believed that arbitrarily increasing the scheduling granularity would be able to show better stability performance and increased system's admittance capacity than that shown in Fig. 15. However, this is beyond the scope of this paper due to the expected delay limitations that might appear in this case. In the same context, the drop in the average rate per UE that appears in Fig. 16 is consistent with the results shown in Fig. 15.

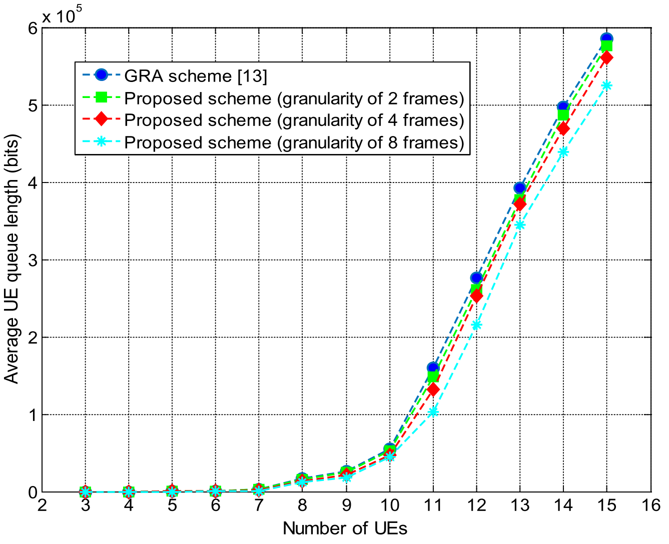

Fig. 15. UE's buffer queue stability.

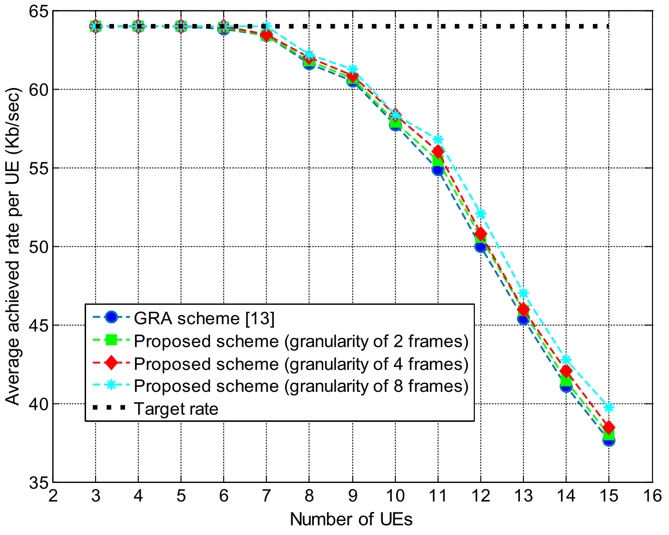

Fig. 16. UE's average achieved rate.

On the other hand, despite the stability performance observed in Fig. 15, a substantial energy reduction percentage per UE for the proposed scheme within the stability region (i.e., the number of UEs = 3–10) in Fig. 17 still exists. This reduction ranges from 25% to 56.6%, compared with the GRA scheme. Thus, although failing to boost the system's capacity, our proposed scheme is still able to improve the UE's EE by increasing the scheduler's time granularity in the presence of slow-varying channels.

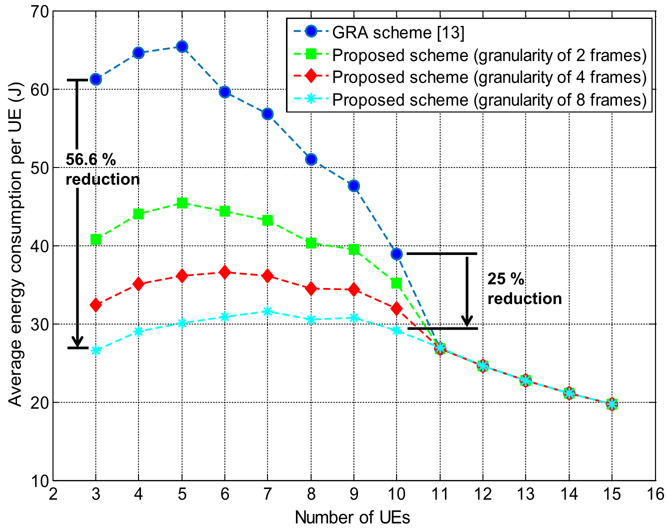

Fig. 17. Average energy consumption per UE.

VII. Conclusion

In this paper, we have developed a framework for implementing a QoS-aware energy-efficient predictive scheduling approach for the downlink in OFDMA-based cellular systems utilizing the well-known site-specific RT approach.

First, we proposed the downlink RT-based scheduling system. Second, based on a practical model for the LTE UE power consumption, we formulated a hard-constrained quasi-instantaneous rate problem with an objective to minimize the UE's receiving energy consumption in the downlink. Due to the natural channel capacity limitations, and for accounting only on the dominant components of the UE's receiver power consumption model, our problem formulation undergoes a series of modifications until we reach a practical formulation. Third, we designed a heuristic algorithm to relax the inherent computational burden in the optimal scheduler. To study the performance bounds, the proposed schedulers were comparatively evaluated with respect to the GRA scheme [13] twice, once in the presence of fast (i.e., QSBR model) and again in the slow (i.e., practical 3-D urban scenario) fading channels. In the presence of the fast fading channel, our proposed scheme was able to improve the EE and the scheduler's capacity to serve more UEs by up to 62.47% and 60%, respectively, compared with the GRA scheduler. On the other hand, in the presence of the slow fading channel, despite showing no effect on the scheduler's admittance capacity, the proposed scheme was still able to improve the UE's EE by up to 56.6% compared with the GRA scheme.

To summarize, it could be seen that our proposed scheduling scheme can work effectively and cooperatively with the current 3GPP LTE DRX power management scheme to prolong today's smartphone battery lifetime per charge.

References

[1] Statista. [Online]. Available: http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] PewResearchCenter. [Online]. Available: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

[3] D. Feng *et al.*, "A survey of energy-efficient wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 167–178, Feb. 2013.

[4] S. Mumtaz, K. Saidul Huq, and J. Rodriguez, "Direct mobile-to-mobile communication: Paradigm for 5G," *IEEE Wireless Commun.*, vol. 21, no. 5, pp. 14–23, Oct. 2014.

[5] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with BS coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[6] M. Kalil, A. Shami, and A. Al-Dweik, "QoS-aware power-efficient scheduler for LTE uplink," *IEEE Trans. Mobile Comput.*, vol. 14, no. 8, pp. 1672–1685, Aug. 2015.

[7] M. Kalil, A. Shami, A. Al-Dweik, and S. Muhaidat, "Low-complexity power-efficient schedulers for LTE uplink with delay-sensitive traffic," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4551–4564, Oct. 2015.

[8] D. Dechene and A. Shami, "Energy-aware resource allocation strategies for LTE uplink with synchronous HARQ constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, pp. 422–433, Feb. 2014.

[9] R. Gupta and E. Strinati, "Green scheduling to minimize base station transmit power and UE circuit power consumption," in *Proc. IEEE 22nd Int. Symp. Pers. Indoor Mobile Radio Commun.*, Sep. 2011, pp. 2424–2429.

[10] C. Xiong, G. Li, Y. Liu, Y. Chen, and S. Xu, "Energy-efficient design for downlink OFDMA with delay-sensitive traffic," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 3085–3095, Jun. 2013.

[11] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.

[12] S. Wang, W. Shi, and C. Wang, "Energy-efficient resource management in OFDM-based cognitive radio networks under channel uncertainty," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3092–3102, Sep. 2015.

[13] F. Chu, K. Chen, and G. Fettweis, "Green resource allocation to minimize receiving energy in OFDMA cellular systems," *IEEE Commun. Lett.*, vol. 16, no. 3, pp. 372–374, Mar. 2012.

[14] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Commun. Mag.*, vol. 47, no. 6, pp. 48–55, Jun. 2009.

[15] K. Hammad, M. Mirahmadi, S. Primak, and A. Shami, "On a throughput-efficient look-forward channel-aware scheduling," in *Proc. IEEE ICC*, Jun. 2015, pp. 6234–6239.

[16] C. Cox, *An Introduction to LTE*. New York, NY, USA: Wiley, 2012.

[17] D. J. Dechene and A. Shami, "Energy efficient QoS constrained scheduler for SC-FDMA uplink," *Phys. Commun.*, vol. 8, pp. 81–90, Sep. 2013.

[18] C.-L. I *et al.*, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[19] K. A. Meerja, A. Shami, and A. Refaey, "Hailing cloud empowered radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 122–129, Feb. 2015.

[20] B. Al-Manthari, H. Hassanein, N. Ali, and N. Nasser, "Fair class-based downlink scheduling with revenue considerations in next generation broadband wireless access systems," *IEEE Trans. Mobile Comput.*, vol. 8, no. 6, pp. 721–734, Jun. 2009.

[21] A. Jensen, M. Lauridsen, P. Mogensen, T. Sørensen, and P. Jensen, "LTE UE power consumption model: For system level energy and performance optimization," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2012, pp. 1–5.

[22] M. F. Catedra and J. Perez, *Cell Planning for Wireless Communications*, 1st ed. Boston, MA, USA: Artech House, 1999.

[23] C. A. Balanis, *Antenna Theory: Analysis and Design*, 3rd ed. Hoboken, NJ, USA: Wiley-Interscience, 2005.

[24] S. Stefania, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution*. New York, NY, USA: Wiley, 2011.

[25] M. Iskander and Z. Yun, "Propagation prediction models for wireless communication systems," *IEEE Trans. Microw. Theory Techn.*, vol. 50, no. 3, pp. 662–673, Mar. 2002.

[26] H. Azodi, U. Siart, and T. Eibert, "A fast 3-D deterministic ray tracing coverage simulator including creeping rays based on geometry voxelization technique," *IEEE Trans. Antennas Propag.*, vol. 63, no. 1, pp. 210–220, Jan. 2015.

[27] H.-Y. Kim, Y.-J. Kim, J.-H. Oh, and L.-S. Kim, "A reconfigurable SIMT processor for mobile ray tracing with contention reduction in shared memory," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 4, pp. 938–950, Apr. 2013.

[28] M. Morelli and U. Mengali, "A comparison of pilot-aided channel estimation methods for OFDM systems," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3065–3073, Dec. 2001.

[29] J. Del Peral-Rosado, J. Lopez-Salcedo, G. Seco-Granados, F. Zanier, and M. Crisci, "Achievable localization accuracy of the positioning reference signal of 3GPP LTE," in *Proc. Int. Conf. Localization GNSS*, Jun. 2012, pp. 1–6.

[30] W. Wang, V. Lau, and M. Peng, "Delay-optimal fronthaul allocation via perturbation analysis in cloud radio access networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 3999–4004.

[31] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Belmont, MA, USA: Athena Scientific, 1996.

[32] J. Proakis, *Digital Communications*, 4th ed. New York, NY, USA: McGraw-Hill, 2001.

[33] S. Primak, V. Kontorovich, and V. Lyandres, *Stochastic Methods and Their Applications to Communications: SDE Approach*. Chichester, U.K.: Wiley, 2004.

[34] Remcom Inc.. [Online]. Available: http://www.remcom.com/wireless-insite

[35] A. Kanatas, I. Kountouris, G. Kostaras, and P. Constantinou, "A UTD propagation model in urban microcellular environments," *IEEE Trans. Veh. Technol.*, vol. 46, no. 1, pp. 185–193, Feb. 1997.

Karim Hammad (S'15) received the B.Sc. and M.Sc. degrees in electronics and commuincations engineering from Arab Academy For Science, Technology & Maritime Transport, Alexandria, Egypt, in 2005 and 2009, respectively. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Western Ontario, London, ON, Canada.

His current research interests include wireless communications, data networking, and digital circuit design.

Serguei L. Primak (S'94–M'97) was born in Mozdok, Russia, in 1967. He received the M.S.E.E. degree from St. Petersburg University of Telecommunications, St. Petersburg, Russia, in 1991 and the Ph.D. degree in electrical engineering from Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 1996.

He is currently a Lecturer and a PostDoctoral Fellow with the University of Western Ontario, London, ON, Canada. His current research interests include ultrawideband radar applications, random signal generations, modeling of wave propagation in a city, time–frequency analysis, and inverse problems of electromagnetic.

Mohamad Kalil (M'15) received the B.Sc. and M.Sc. degrees in electrical engineering from Jordan University of Science and Technology, Irbid, Jordan, in 2009 and 2011, respectively. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Western Ontario, London, ON, Canada.

His research interests include cross-layer design, radio resource management, wireless network virtualization, and cloud radio access networks.

Abdallah Shami (SM'09) received the B.E. degree in electrical and computer engineering from the Lebanese University, Beirut, Lebanon, in 1997 and the Ph.D. degree in electrical engineering from the City University of New York, New York, NY, USA in 2002.

Since July 2004, he has been with the University of Western Ontario, London, ON, Canada, where he is currently a Professor with the Department of Electrical and Computer Engineering. His current research interests include network-based cloud computing and wireless/data networking.

Dr. Shami is the Chair of the IEEE Communications Society Technical Committee on Communications Software. He has served as the Chair for the key symposia of the IEEE Global Communications Conference; the IEEE International Conference on Communications; the IEEE International Conference on Computing, Networking, and Communications; and the International Conference on Computer and Information Technology. He currently serves as an Associate Editor for IEEE COMMUNICATIONS SURVEY AND TUTORIALS, *IET Communications Journal*, and the *Wiley Journal of Wireless Communications and Mobile Computing*.