Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2010, Article ID 375410, 20 pages
doi:10.1155/2010/375410

Research Article

A Dynamic Utility Adaptation Framework for Efficient Multimedia Service Support in CDMA Wireless Networks

Timotheos Kastrinogiannis and Symeon Papavassiliou

Network Management and Optimal Design Laboratory (NETMODE), Institute of Communications and Computer Systems (ICCS), School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), 9 Iroon Polytechniou Street, Zografou 157 73, Athens, Greece

Correspondence should be addressed to Symeon Papavassiliou, papavass@mail.ntua.gr

Received 30 March 2010; Revised 14 July 2010; Accepted 24 August 2010

Academic Editor: Liang Zhou

In this paper, the problem of channel-aware opportunistic resource allocation for the downlink in code division multiple access wireless networks supporting simultaneously real-time multimedia and non-real-time data services is addressed. In order to treat different types of services with diverse QoS prerequisites through common optimization formulation a utility-based power and rate allocation framework is adopted. Emphasis is placed on real-time services' strict short-term QoS prerequisites, the fulfillment of which requires a significantly more different treatment than the use of static utility functions, traditionally used to address long-term QoS or fairness prerequisites of delay-tolerant data services. To that end, we introduce a novel generic framework that enables the dynamic adaptation of real-time multimedia users' utilities as the system evolves, with respect to the corresponding short-term throughput service performance variations. The corresponding nonconvex network utility maximization (NUM) problem is then formulated and solved to obtain optimal downlink power and rate allocation. Via simulation and analysis, it is demonstrated that significant performance improvements are achieved in terms of real-time user's short-term throughput requirement satisfaction, without any considerable loss in the total system throughput. Finally, an essential tradeoff between efficiently fulfilling real-time services' short-term QoS prerequisites and maximizing overall system performance, under an opportunistic scheduling wireless environment, is revealed and quantified.

1. Introduction

With the growing demand for high data rate and support of multiple services with various quality of service (QoS) requirements, the scheduling policy plays a key role in the efficient resource allocation process in future wireless networks. Moreover, users' time and location-dependent channel conditions limit the system's available resources and hence its ability to satisfy their QoS properties. Therefore, a flexible power and rate allocation scheme is essential for optimizing the system's performance.

Considerable research efforts have been devoted to the combined problem of power and rate allocation for the downlink of a code division multiple access (CDMA) system (e.g., [1–3]) aiming at the exploitation of multiuser diversity (i.e., users' time-varying channel conditions) towards optimizing the system's performance, while satisfying various QoS constraints [4–9]. Moreover, due to the heterogeneity of the wireless environment and the need for the support of diverse QoS requirements, the concept of utilities from the field of economics has been adopted for devising proficient opportunistic resource allocation algorithms. A utility function reflects a users' degree of satisfaction with respect to their service performance in a normalized and transparent way, allowing services with assorted QoS prerequisites to be represented, by forming appropriate utilities [10–15], under a common utility-based optimization framework. Hence, Network Utility Maximization (NUM) theory provides the foundations and mathematical tools for setting and treating such problems.

In a typical NUM formulation, user's utilities are static, predetermined functions, associated to specific services or

service classes, emphasizing mainly on the support of non-real-time users' long-term requirements. Therefore, users' utilities mainly define a continuous relationship between their service performance and their actual achieved throughput (i.e., goodput that reflects the number of reliable bits transmitted) over the wireless opportunistic CDMA paradigm, while considering long-term user's fairness issues [16], minimum performance requirements [17] and/or appropriate constraints imposed by the devices' physical limitations [18].

On the other hand, the delay-sensitive nature of real-time multimedia services poses additional demands on accessing the system resourses within short time intervals. Therefore, the adoption of probabilistic short-term delay [19, 20] or throughput [21, 22] constraints have been proposed towards efficiently expressing real-time services' QoS prerequisites over a time-varying wireless environment. However, the conventional use of static predefined users utilities does not permit the efficient integration of the latter probabilistic short-term constraints within a NUM problem formulation and thus in the system's resource scheduling policy.

In this paper, we study the problem of jointly scheduling multiple services, that is, delay-sensitive, real-time, and delay-tolerant high-throughput non-real-time services, over a heterogeneous CDMA wireless system via NUM optimization. Towards achieving our goal, this paper makes the following contributions.

(a) We design real-time users' utilities that dynamically adjust with respect to users' service short-term QoS satisfaction levels fulfillment, enabling them to efficiently reflect real-time services strict, instantaneous resource demands at the scheduling policy. We refer to the above novel approach as Dynamic Utility Adaptation (DUA) framework. DUA serves as an extension to current NUM theory in order to mainly treat and overcome issues that arise from the use of static utilities, when aiming at introducing users' short-term goals or prerequisites under a NUM optimization setting.

(b) We adopt and exploit probabilistic short-term throughput constraints, instead of myopic probabilistic delay constraints, in order to introduce the essential requirements of real-time users (requesting multimedia services) in the resource allocation process of a CDMA wireless network.

(c) Through the proper use of static and dynamic utilities according to the respective service types, we aim at:(a) meeting various types of user services' QoS requirements, namely, real-time and non-real-time, under a common optimization framework and (b) exploiting the benefits emerging by the scheduler's opportunistic character, not only individually per type of users but also in a collaborative manner as well. In this way, a scheduling policy is devised that avoids the problem where the optimization of the performance of users of a specific type of service leads to the corresponding degradation of the performance of other types of services. Thus, several inherent system and users' limitations in satisfying services' short-term QoS requirements, caused by the corresponding physical hardware constraints, under an interference limited opportunistic wireless environment are highlighted and discussed.

(d) Finally, two simple iterative algorithms are proposed. The first one, residing at the base station, attains an asymptotically optimal (in the number of users) power and rate allocation of systems' non-convex optimization problem, which is continuously reset at the beginning of each time slot with respect to users' utilities adaptation. The second one, residing at the mobile node, dynamically adapts a real-time user's utility by realizing a control loop which: (a) constantly monitors a user's service performance, (b) analyzes its current status with respect to QoS requirements, and (c) reacts to QoS triggering events via the dynamic alteration of the user's utility. It is demonstrated via modeling and simulation that our proposed scheme achieves to the fulfillment of real-time users' short-term prerequisites without any considerable loss in the system's total achieved throughput. The obtained results allow to reveal and quantify the inherent tradeoff between efficiently fulfilling real-time services' demanding short-term QoS prerequisites and maximizing overall system performance, under an opportunistic wireless environment.

The rest of the paper is organized as follows. In Section 2, the system model and definitions are presented. In Section 3, the proposed dynamic users' utility adaptation framework is first analyzed, and its application on real-time services is presented. Then, the corresponding utility-based optimization problem is formulated, and its solution is derived. In Section 4, real-time users' self-adaptation mechanism in QoS-triggered events is described, and an enhanced power and rate allocation scheme is proposed. Numerical results and relevant discussions are provided in Section 5, while Section 6 concludes the paper.

2. System Model and Definitions

In this paper, we consider the downlink of a single cell time-slotted CDMA wireless system with N continuously backlogged users at time slot t. A time slot is a fixed interval of time and could consist of one or several packets. User-channel conditions, which are affected by shadow fading and long-time scale variations, are assumed to be fixed within the duration of a time slot. The scheduler is assumed to resign at the base station, and hence it can make decisions on users' power and rate allocation at the beginning of each time slot. Let us denote by $R_i(t)$ the downlink transmission rate at which the base station transmits to user i in the slot under consideration and by $R_i^{\max}$ the maximum rate at which they can receive data (due to physical hardware limitations). Let us also denote by $\gamma_i(t) = E_b(t)/I_o(t)$ the bit energy-to-interference density ratio for user i at their mobile device receiver, by $G_i(t)$ the path gain from the base station to

mobile user i, and by $P_i(t)$ the transmission power allocated at a given slot to user i, which, however, is limited by the base station's maximum downlink power $P_{\max}$. The received $\gamma_i(t)$ for each user i is given [16–18] by

$$\gamma_i\left(R_i(t),\overline{P}(t)\right) = \frac{W}{R_i(t)}\frac{G_i(t)P_i(t)}{\theta G_i(t)\sum_{j=1}^{N}P_j(t)-\theta G_i(t)P_i(t)+I_i(t)} = \frac{W}{R_i(t)}\frac{P_i(t)}{\theta\sum_{j=1}^{N}P_j(t)-\theta P_i(t)+A_i(t)}, \quad (1)$$

where θ denotes the orthogonality factor, W is the system's spreading bandwidth, $\overline{P}(t)$ denotes the users' power allocation vector, $I_i(t)$ includes the background noise and intercell interference at user i, $G_i(t)\sum_{j=1}^{N}P_j(t) - G_i(t)P_i(t)$ determines the intracell interference at user i and $A_i(t) = I_i(t)/G_i(t)$ denotes the transmission environment between user i and the base station.

In our system, we consider two basic types of users, namely, non-real-time users (NRT) requesting delay-tolerant high-throughput services and real-time (RT) users with strict short-term QoS constraints. Throughout the rest of the paper we denote by N_{NRT} (N_{RT}) the number of non-real-time users (real-time users) and by S_{NRT} (S_{RT}) the corresponding set. Due to the variety of the supported services' QoS prerequisites, each mobile user is associated with a proper utility function U_i^* which represents his degree of satisfaction in accordance to his actual expected downlink throughput and can be expressed as

$$U_i^*\left(R_i(t),\overline{P}(t),a_i,b_i\right) = R_i(t)f_i(\gamma_i(t),a_i,b_i), \quad i=1,2,...,\mathrm{N}, \quad (2)$$

where f_i represents a function for the probability of a successful packet transmission for user i and is an increasing function of their bit energy to interference ratio $\gamma_i(t)$ at any time slot. A user's function for the probability of a successful packet transmission at fixed data rates depends on the transmission scheme (modulation and coding) being used and can be represented by a sigmoidal-like function of their power allocation for various modulation schemes [18]. Therefore, a user i efficiency function f_i has the following properties.

(1) f_i is an increasing function of $\gamma_i(t)$.

(2) f_i is a continuous, twice differentiable sigmoidal function with respect to $\gamma_i(t)$.

(3) $f_i(0) = 0$ to ensure that $U_i = 0$ when $P_i(\mathrm{t})= 0$.

(4) $f_i(\infty) = 1$.

Moreover, we define as a_i, b_i the two tunable parameters of the sigmoidal function f_i that determine function's f_i steepness and unique inflection point, respectively [24] (generic definition: $f(\gamma,a,b) = c\{1/(1+e^{-a(\gamma-b)}) - d\}$, where $\mathrm{c} = (1+e^{ab})/e^{ab}$ and $d = 1/(1+e^{ab})$. Intuitively, since parameter a controls the slope of the sigmoidal function, it determines a user's tolerance in power deviations (in the region of functions f inflection point), while parameter b, controls the relative place of the inflection point of function f (at the access of P), and thus the power level upon which a user's successful packet receive probability increases rapidly (for small deviations of the allocated power), following a concave form [18, 23]). Without loss of generality, we assume that all users have the same value for their parameter a_i,(i.e., $a_i = a$ for $i = 1,\ldots,N$). The validity of the above properties has been demonstrated in several practical scenarios with reasonably large packet sizes M (i.e., $M \geq 100$ bits) [25, 26].

Observing a user's utility as defined in (2), we can point out that the main factors that affect its values are a user's transmission environment (A_i), transmission rate (R_i) and transmission scheme (parameter b_i of function f_i). For delay-tolerant non-real-time users, the maximization of their utility corresponds to their desired goodput maximization, and as a result, the corresponding utility is suitable for reflecting their desired throughput maximization at the system's resource allocation optimization problem. On the other hand, real-time users' degree of satisfaction does not increase in a linear or concave way along with their throughput maximization (as in the case of NRT users), but according to their fixed data rate expectation fulfillment, as well as their short-term QoS requirements satisfaction due to their delay sensitive nature (e.g., sigmoidal form).

2.1. Real-Time Services' QoS Requirements. A real-time user's requirements consist mainly of a constant downlink rate and short-term delay and throughput guaranties [21, 22]. Therefore, we consider as a real-time user's performance indicator, the achieved probability of receiving an amount of service, in terms of data units, smaller than a predefined threshold within successive short observation time intervals, which is expressed as follows:

$$\Pr\left[\hat{\beta}_{\mathrm{RT},i}(t) \leq B_{\mathrm{RT},i}\right]_{W_i} \quad \forall t(\mathrm{slot})\ \forall i \in S_{\mathrm{RT}}, \quad (3)$$

where W_i denotes a RT user i observation time interval in terms of slots, $B_{\mathrm{RT},i}$ his predefined data units threshold, and $\hat{\beta}_{\mathrm{RT},i}(t)$ the amount of data they received within a specific time interval from slot $(t - W_i + 1)$ to slot t. The smaller the achieved value of an RT user's short-term throughput probability (3), the greater is their degree of satisfaction. Given a real-time user i requiring downlink rate, $R_{\mathrm{RT},i}$, we can estimate their data units threshold as

$$B_{\mathrm{RT},i} = R_{\mathrm{RT},i}\cdot W_i \cdot t_s, \quad (4)$$

where t_s denotes the duration of a time slot. It has been shown in [22] that short-term throughput constraints can more efficiently and comprehensively reflect the essential requirements of RT users (i.e., both delay and throughput expectations) compared to myopic probabilistic delay constraints. This is due to the fact that the adoption of the latter over a time-varying wireless environment may often cause RT users' throughput rates dissatisfaction, within either small or long time intervals, due to their potentially bad

channel conditions and variations, leading to their service QoS-aware performance degradation.

The previous RT users' QoS quarantines, as defined in (3) and (4), are suitable for Constant Bit Rate (CBR) real-time traffic (e.g., video conferencing, telephony (voice services), etc.). To incorporate the QoS prerequisites of other types of real-time services such as real-time Variable Bit Rate (VBR) traffic (e.g., compressed video streams) in the proposed probabilistic short-term throughput framework, the ability to dynamically adjust the requested downlink data rate, $R_{\mathrm{RT},i}(t)$, and thus their data units threshold $B_{\mathrm{RT},i}(t)$, at the corresponding RT user i should be provided. Therefore, in this case, we define $B_{\mathrm{RT},i}(t) = R_{\mathrm{RT},i}(t) \cdot W_i \cdot t_s$. The adopted short-term throughout prerequisites inherent attribute of fulfilling the requested data rate of an RT user within short-term time intervals, instead of converging to it within long-term intervals (as in [16–18]), allows the efficient support of both CBR and VBR traffic.

In order to guarantee short-term throughput requirements satisfaction for all RT users (i.e., achieve small values for their probabilities defined in (3)), we aim at providing them with the flexibility of dynamically affecting the priorities of being selected for receiving service according to their corresponding short-term throughput performance, through the introduction of an appropriate user-centric dynamic behavior which drives their ability to dynamically adapt their utility functions, as detailed in the following section.

3. Dynamic Utility Adaptation (DUA) Frame Work—Problem Formulation and Solution

In this section, we first detail and analyze a novel framework for reflecting users' short-term QoS requirements at their utility functions under a NUM problem formulation. This is achieved via the dynamic alteration of the utilities' properties in accordance to generic short-term time-varying QoS performance metrics—we refer to this framework as Dynamic Utility Adaptation (DUA). Emphasizing on multimedia services and their corresponding QoS prerequisites, a methodology for dynamically adapting RT users' utility parameters in accordance to their short-term throughput requirements is examined. Then, the overall utility-based optimization problem is formulated, considering both NRT and RT users' performance expectations, and its solution is derived. Finally, following a pure optimization theoretic analysis, the design properties of the proposed DUA framework are examined by determining the way users' utilities parameters deviations affect their priority of accessing system resources.

3.1. Dynamic Utility Adaptation Framework. Towards optimizing system's performance, a scheduling policy should allocate wireless network resources, in terms of transmission powers and corresponding rates, in a way that not only maximizes users' utilities and hence their degree of satisfaction in each time slot, but also satisfies their QoS prerequisites. The use of fixed predefined utility functions enables the reflection of users' long-term performance expectations at the scheduler and is in line with its opportunistic channel-aware nature [24]. On the other hand, RT users' short-term QoS demands require the scheduler's response within short-time intervals in the light of short-term QoS violations; therefore, the latter should also be reflected in their utilities.

With respect to the previous discussion and analysis, we introduce the dynamic adaptation of RT users' utilities $U_i^*(R_i(t), \overline{P}(t), a, b_i)$ for all $i \in S_{\mathrm{RT}}$ by allowing them to properly and dynamically adjust the values of their utility parameter b_i for all $i \in S_{\mathrm{RT}}$. Moreover, we redefine RT users' utility function as follows:

$$U_i^*\left(R_i(t), \overline{P}(t), a, \widehat{b}_{\mathrm{RT},i}(t)\right) \quad \forall i \in S_{\mathrm{RT}}, \tag{5}$$

where $\widehat{b}_{\mathrm{RT},i}(t)$ denotes a user's utility tunable parameter b_i at time slot t and is defined as

$$\begin{aligned} \widehat{b}_{\mathrm{RT},i}(t) &= b_{\mathrm{RT},i} + b_{\mathrm{RT},i}(t) \quad \forall i \in S_{\mathrm{RT}} \\ \widehat{b}_{\mathrm{RT},i}(t) &\in [b_{\min,i}(t), b_{\max,i}(t)], \\ \text{where } 0 \le b_{\min,i}(t) &\le b_{\mathrm{RT},i} \le b_{\max,i}(t) \le \infty, \end{aligned} \tag{6}$$

where $b_{\mathrm{RT},i}$ is RT user i proper parameter in accordance to his transmission scheme (i.e., function's f_i initial fixed b_i parameter) and $b_{\mathrm{RT},i}(t)$ is the factor that dynamically adjusts parameter's b_i overall value in accordance to user's short-term performance. Thus, $b_{\mathrm{RT},i}(t)$ is fixed within the duration of a time-slot. Let us underline that when RT users adjust their $\widehat{b}_{\mathrm{RT},i}(t)$ parameter does not actually select a different modulation scheme, defined only by the fixed part of (6) (i.e., $b_{\mathrm{RT},i}$), but aim at reflecting in the scheduling policy (via their utility function) their expectations in system resources with respect to their current short-term QoS performance and thus, affecting their priority in accessing system's resources. In general, as $\widehat{b}_{\mathrm{RT},i}(t)$ decreases a user's lack of resources is mirrored to his utility and consequently their need for having high priority in accessing system resources is revealed, a desirable property that justifies its selection, as it is shown via the solution of the corresponding utility-based system optimization problem.

Parameters $b_{\min,i}(t)$ and $b_{\max,i}(t)$ are the upper and lower bounds of a RT user's parameter $\widehat{b}_{\mathrm{RT},i}(t)$ in each time slot t, respectively. As it is analyzed later in this paper (Appendix A), the existence of these bounding parameters restricts a user's ability to self-optimize their QoS performance over a time-varying wireless environment, due to the potentially bad channel conditions or lack of available system radio resources.

3.2. Adjusting the Properties of Real-Time Users' Utilities . In order real-time users to efficiently adjust their utility parameter $\widehat{b}_{\mathrm{RT},i}(t)$ for all $i \in S_{\mathrm{RT}}$ at the beginning of each time slot t, according to their short-term throughput requirements, the introduction of their actual short-term throughput performance information into the tuning procedure of their utility $\widehat{b}_{\mathrm{RT},i}(t)$ parameter is essential. Therefore, let us define the actual amount of data units that a real-time

user i received within his observation time interval W_i, from slot $(t - W_i + 1)$ to slot $(t - 1)$ as follows:

$$B_{\mathrm{RT},i}^{W_i-1}(t) = \sum_{k=1}^{W_i-1} \beta_{\mathrm{RT},i}(t-k) \quad \forall\ i \in S_{\mathrm{RT}}, \tag{7}$$

where $\beta_{\mathrm{RT},i}(t) = U_i^*(R_i(t), P_i(t), a, b_{\mathrm{RT},i}) \cdot t_s$ denotes the actual amount of data that a real-time users i received at time slot t and $R_i(t)$,$P_i(t)$ denote his corresponding transmission rate and power allocation at the under consideration time slot, respectively. By using the above information and comparing it with a portion of his predefined short-term data units threshold $B_{\mathrm{RT},i}$, an RT users can adjust their utility $\hat{b}_{\mathrm{RT},i}(t)$ parameter as

$$\hat{b}_{\mathrm{RT},i}(t) = \begin{cases} b_{\mathrm{RT},i} - I_{\mathrm{RT},i}(t)\left[b_{\mathrm{RT},i} - b_{\min,i}(t)\right] \\ \quad \text{if } \mathrm{B}_{\mathrm{RT},i}^{W_i-1}(t) \le (\mathrm{Tr}+1)B_{\mathrm{RT},i}, \\ b_{\mathrm{RT},i} + G_{\mathrm{RT},i}(t)\left[b_{\max,i}(t) - b_{\mathrm{RT},i}\right] \\ \quad \text{if } \mathrm{B}_{\mathrm{RT},i}^{W_i-1}(t) > (\mathrm{Tr}+1)B_{\mathrm{RT},i}, \end{cases} \tag{8}$$

where $I_{\mathrm{RT},i}(t) \in [0,1]$ and $G_{\mathrm{RT},i}(t) \in [0,1]$ for all $i \in S_{\mathrm{RT}}$ are two normalized indicators that reflect a real-time user's need for accessing the system resources at time slot t, when they have shortage or excess of data units received within their current short observation time interval, respectively. Furthermore, parameter Tr ($\mathrm{Tr} \ge 0$), referred as the system's triggering parameter, determines the system's degree of preemption. Large values for the system's triggering parameter will make the scheduling policy react in a more preemptive way to real-time users' short-term throughput performance deviations, and therefore the achieved probabilities of not satisfying their short-term QoS requirements values will decrease.

In accordance to (8), when a real-time user i has received till time slot $(t - 1)$ less amount of data units than his predefined threshold, then in order to accomplish his short-term throughput QoS requirements satisfaction, the value of his utility $\hat{b}_{\mathrm{RT},i}(t)$ parameter decreases and thus, his probability of being selected at current slot t increases. Furthermore, the reduction of a real-time user's i utility $\hat{b}_{\mathrm{RT},i}(t)$ parameter from its corresponding value $b_{\mathrm{RT},i}$ is determined by his normalized indicator $I_{\mathrm{RT},i}(t)$ at that time slot. A RT user's $I_{\mathrm{RT},i}(t)$ indicator reflects his need of accessing the system resources, according to the weighted distribution of his received data within his observation time interval, and therefore is defined as

$$I_{\mathrm{RT},i}(t) = 1 - \sum_{k=1}^{W_i-1} \frac{\hat{w}_{\mathrm{RT},i}(k, W_i)\beta_{\mathrm{RT},i}(t-k)}{(\mathrm{Tr}+1)B_{\mathrm{RT},i}} \quad \text{if } \mathrm{B}_{\mathrm{RT},i}^{W_i-1}(t) \le (\mathrm{Tr}+1)B_{\mathrm{RT},i}, \tag{9}$$

$$\text{where } \hat{w}_{\mathrm{RT},i}(k, W_i) = \frac{W_i - k}{W_i - 1} \quad \text{for } k = 1, \ldots, W_i \ \forall i \in S_{\mathrm{RT}}.$$

It is noted that $\hat{w}_{\mathrm{RT},i}(k, W_i)$ represents a weight related to each time slot within the last $(W_i - 1)$ successive slots of an RT user i observation interval W_i that determines the importance of the user i received amount of service at that slot (i.e., time slot $t-k$) on his estimated indicator. Moreover, the values of an RT user slots weights, as well as the importance of his information, are linearly inversely proportional to his distance k from the current slot t, since we want the information of the most distant slots to play a more important role on the degree of his need in accessing the system's resources. For instance, even if two real-time users i and j have received the same amount of data within the same observation time intervals (i.e., $B_{\mathrm{RT},i}^{W_i-1}(t) = B_{\mathrm{RT},j}^{W_j-1}(t)$ when $W_i = W_j$, $B_{\mathrm{RT},i} = B_{\mathrm{RT},j}$, $b_{\mathrm{RT},i} = b_{\mathrm{RT},j}$ and $b_{\min,i}(t) = b_{\min,j}(t)$), but user i has received service in slots more recent to the current than user j, then their indicator's $I_{\mathrm{RT},i}(t)$ value will be smaller than user j indicator value $I_{\mathrm{RT},j}(t)$ according to his slots' weights, since his tolerance for not accessing the system's resources is greater. Thus, user j utility $\hat{b}_{\mathrm{RT},j}(t)$ parameter will be smaller than user i corresponding parameter, and therefore they will have higher priority on accessing the system's resources at the current slot.

On the other hand, when a real-time user has received a larger amount of data units than the predefined threshold within the last $(W_i - 1)$ successive time slots of his observation interval W_i, then their utility $\hat{b}_{\mathrm{RT},i}(t)$ parameter increases according to (8), and their priority in being served decreases. The larger RT user i received amount of data within their observation interval W_i is, the lower their selection priority should be, and therefore their normalized indicator can be defined as follows:

$$G_{\mathrm{RT},i}(t) = \frac{\sum_{k=1}^{W_i-1} \hat{w}_{\mathrm{RT},i}(k, W_i)\beta_{\mathrm{RT},i}(t-k)}{\sum_{k=1}^{W_i-1} \hat{w}_{\mathrm{RT},i}(k, W_i)R_i^{\max}t_s} \quad \text{if } \mathrm{B}_{\mathrm{RT},i}^{W_i-1}(t) > (\mathrm{Tr}+1)B_{\mathrm{RT},i}, \tag{10}$$

where the denominant $\sum_{k=1}^{W_i-1} \hat{w}_{\mathrm{RT},i}(k, W_i)R_i^{\max}t_s$ denotes the maximum weighted amount of data units an RT user i can receive within any time interval of $(W_i - 1)$ time slots, due to their downlink rate limitation $R_i^{\max}$. Such a design attribute allows the reallocation of excess system resources to NRT users towards the desirable optimization of their throughput performance [21, 22].

Concluding this section's analysis, let us underline that the methodology expressed via (8), (9), and (10) applies in the most demanding case where the objective is to minimize RT users' probabilistic throughput constraints (i.e., $\min \Pr[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}]_{W_i}$ for all $i \in S_{\mathrm{RT}}$). Moreover, in the special case where an upper bound is set for RT users probabilistic prerequisites, that is, $\Pr[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}]_{W_i} \le q_{\mathrm{RT},i}$ for all $i \in S_{\mathrm{RT}}$, then the two normalized parameters $I_{\mathrm{RT},i}$, *and* $B_{\mathrm{RT},i}$, are defined as

$$I_{\mathrm{RT},i}(t) = 1 - \frac{\Pr\left[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}\right]_{W_i}}{q_{\mathrm{RT},i}} \quad \text{if } \Pr\left[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}\right]_{W_i} \le q_{\mathrm{RT},i}, \tag{11}$$

$$G_{\mathrm{RT},i}(t) = 1 - \frac{q_{\mathrm{RT},i}}{\Pr\left[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}\right]_{W_i}} \quad \text{if } \Pr\left[\hat{\beta}_{\mathrm{RT},i}(t) \le B_{\mathrm{RT},i}\right]_{W_i} > q_{\mathrm{RT},i}, \tag{12}$$

towards reflecting RT user's need for accessing the system resources at time slot t, when they are accomplishing or not the requested bound $q_{\mathrm{RT},i}$, respectively, under the assumption of the system's feasibility (i.e., there always exists at least one power and rate vector that leads to the satisfaction of all RT users' probabilistic throughput constraints).

3.3. Problem's Formulation, Transformation, and Solution. In order to optimize the overall system performance as well as users' degree of satisfaction, the following utility-based power and rate allocation optimization problem must be solved at the scheduler at every time slot

$$\begin{aligned} \max_{\overline{R}(t),\overline{P}(t)} \quad & \sum_{i=1}^{N} U_i^*\left(R_i(t), \overline{P}(t), a, b_i(t)\right) \\ \text{s.t.} \quad & \sum_{i=1}^{N} P_i(t) \le P_{\max} \\ & 0 \le P_i(t) \le P_{\max} \quad i = 1, 2, ..., N \\ & 0 \le R_i(t) \le R_i^{\max} \quad i = 1, 2, ..., N \\ & b_i(t) = \hat{b}_{\mathrm{RT},i}(t) \quad \forall i \in S_{\mathrm{RT}}, \\ & b_i(t) = b_{\mathrm{NRT},i} \quad \forall i \in S_{\mathrm{NRT}}, \end{aligned} \tag{13}$$

where $b_{\mathrm{NRT},i}$ denotes NRT users' fixed parameter of their f_i function in accordance to the used modulation and coding scheme and $\hat{b}_{\mathrm{RT},i}(t)$ is obtained via (8). Intuitively, (13) aims at jointly fulfilling and optimizing both NRT and RT users' QoS-aware degree of satisfaction, via maximizing the actual achieved throughput of the first (i.e., expressed via utility (2)) and via fulfilling the probabilistic short-term throughput prerequisites of the second (i.e., expressed via utility (5) and (6)). In the rest of the paper, for simplicity in the presentation, we omit the notation of the specific slot t in the notations of the system's and users' variables that remain fixed within the duration of a time slot.

Following the approach in [24], the optimal solution of the above problem is achieved when the base station transmits with its maximum power level $P_{\max}$ (i.e., $\sum_{i=1}^{N} P_i = P_{\max}$), and hence a user i utility defined in (2) or (5) is adjusted according to the following expression:

$$\begin{aligned} & U_i^{R_i^*}\left(R_i^*(P_i), P_i, a, b_i\right) \\ & = U_i^{R_i^*}(P_i, a, b_i) \\ & = \begin{cases} \dfrac{WP_i}{\gamma_i^*(\theta P_{\max} - \theta P_i + A_i)} f_i(\gamma_i^*), & \text{if } P_i \le \dfrac{R_i^{\max}\gamma_i^*(\theta P_{\max} + A_i)}{W + \theta R_i^{\max}\gamma_i^*} \\ & \quad = P_i^{\mathrm{LIM}}(\gamma_i^*, A_i) \\ R_i^{\max} f_i(\gamma_i(R_i^{\max}, P_i)), & \text{otherwise,} \end{cases} \end{aligned} \tag{14}$$

where $\gamma_i^* = \mathrm{argmax}_{\gamma \ge 1}\{(1/\gamma) f_i(\gamma)\}$, $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)$ is the break point of function $U_i^{R_i^*}(P_i, a, b_i)$ and $R_i^* = WP_i/\gamma_i^*(\theta P_{\max} - \theta P_i + A_i)$ when $P_i \le P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)$, or $R_i^* = R_i^{\max}$ otherwise. For $P_i \le P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)$, $U_i^{R_i^*}$ is a convex function of P_i and for $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i) < P_i \le P_{\max}$ is a sigmoidal function. Therefore, $U_i^{R_i^*}$ is a sigmoidal function of P_i at his maximum transmission rate $R_i^{\max}$, with inflection point denoted as P_i^0 (specifically, $\partial^2 U_i^{R_i^*}(P_i, a_i, b_i)/\partial P_i^2|_{P_i = P_i^0} = 0$, $U_i^{R_i^*}(P_i, a_i, b_i)/\partial P_i^2|_{P_i < P_i^0} > 0$ and $\partial^2 U_i^{R_i^*}(P_i, a_i, b_i)/\partial P_i^2|_{P_i > P_i^0} < 0$). Furthermore, the optimization problem (13) can be transformed to the following:

$$\begin{aligned} \max_{\overline{P}} \quad & \sum_{i=1}^{N} U_i^{R_i^*}(P_i, a, b_i), \\ \text{s.t.} \quad & \sum_{i=1}^{N} P_i \le P_{\max}, \\ & 0 \le P_i \le P_{\max} \quad i = 1, 2, ..., N. \end{aligned} \tag{15}$$

Towards solving the non-convex optimization (15), a pricing-based algorithm was developed in [18], and its asymptotic optimality, when the number of users is large, has been proven. Initially, the scheduler selects users to which nonzero power will be allocated by using the information of their parameters' $\lambda_i^{\max}$ values. Parameter $\lambda_i^{\max}$ represents user i maximum willingness to pay per unit power

$$\lambda_i^{\max} = \min\left\{\lambda \ge 0 \mid \max_{0 \le P \le P_{\max}} \left\{U_i^{R_i^*}(P_i, a, b_i) - \lambda P_i\right\} = 0\right\}. \tag{16}$$

In other words, $\lambda_i^{\max}$ is the price λ that maximizes user i net utility $P(\lambda) = \mathrm{argmax}_{0 \le P \le P_{\max}}\{U_i^{R_i^*}(P, a, b_i) - \lambda P\}$ (i.e., the utility minus the cost) and can be calculated as follows:

$$\lambda_i^{\max} \begin{cases} \left.\dfrac{\partial U_i^{R_i^*}(P_i, a, b_i)}{\partial P_i}\right|_{P = P^*} & \text{if } U_i^{R_i^*} \text{ is a sigmoidal} - \text{like function and } P^* \text{exists} \\ \dfrac{U_i^{R_i^*}(P_{\max}, a, b_i)}{P_{\max}} & \text{otherwise,} \end{cases} \tag{17}$$

where P_i^* is the unique solution of

$$U_i^{R_i^*}(P_i, a, b_i) - P_i \frac{\partial U_i^{R_i^*}(P_i, a, b_i)}{\partial P_i} = 0 \quad \text{for } P_i^0 \le P_i \le P_{\max}. \tag{18}$$

Moreover, if for any two users i and j $U_i^{R_i^*} \ge U_j^{R_j^*}$ for $0 \le P \le P_{\max}$, then $\lambda_i^{\max} \ge \lambda_j^{\max}$, and therefore user i is more likely to be selected than user j. Hence, the scheduler selects users in a decreasing order of their maximum willingness to pay from 1 to T satisfying

$$T = \max\left\{1 \le j \le N \mid \sum_{i=1}^{j} P_i\left(\lambda_j^{\max}\right) \le P_T\right\} \tag{19}$$

Finally, for the selected users the base station updates and broadcasts $\lambda_T^{\max}$ till finding a unique equilibrium price λ^* that satisfies $\sum_{i=1}^{T} P_i(\lambda^*) = P_{\max}$ [18]. Knowing λ^*, selected users' transmission powers and rates can be easily derived.

In accordance to the previous analysis, the price of a users' willingness to pay $\lambda_i^{\max}$ plays a key role in their selection priority and, moreover, in the portion of total system's resources a user will finally occupy in the subsequent time slot. The following proposition shows that by allowing RT user to adapt their utility properties via adjusting his b_i parameter, they gain the enhanced flexibility of controlling the priority of being selected in accessing system resources among the others, towards optimizing their service performance.

Proposition 1. *If $A_i = A_j$ and $R_i^{\max} = R_j^{\max}$, then if $b_i < b_j$ $\lambda_i^{\max} > \lambda_j^{\max}$.*

Proof. see Appendix B. □

Proposition 1 asserts that if all other conditions are equivalent, a user i with smaller parameter b_i has a higher priority in being selected than a user j with larger value for his parameter b_j. Moreover, if smaller values of an RT user's expected throughput within their observation time interval are observed, then lower values of his b_i parameter will result to higher probability in being selected, and vise versa. Essentially, the above proposition can be generalized for more than two users, revealing not only a relational dependency among users' utilities properties and his allocated resources, when the latter are derived through the solution of the system's utility-based optimization problem, defined in (13), but also the validity of the proposed DUA methodology expressed via (8), (9), and (10).

4. Proposed Scheduling Policy—Towards Node's QoS-Aware Self-Optimization

Nodes' QoS-aware self-optimization refers to the ability of sensing his service performance variations as well as his environment changes, and then reacting to QoS triggering events towards optimizing his service performance. Such a behavior is revealed through the solution of the corresponding power and rate allocation optimization problem in CDMA networks when both NRT and RT services require access at system resources. Users requesting real-time services can monitor his services' performance, analyze and compute their resource expectations in a normalized way according to their short-term QoS prerequisites, and then adapt their utility functions' properties in order to affect their selection priority in the scheduling policy as well as the amount of anticipated resources. Moreover, at the base station, the system scheduler interacts with the mobile nodes towards solving the corresponding optimization problem, as defined in (13).

In the rest of this section, we present a Dynamic Utility Adaptation-based Users' Power and Rate Allocation (DUA_UPRA) scheme, which is realized by the efficient collaboration of two low complexity algorithms residing at the mobile nodes and base station, respectively. From mobile nodes perspective, DUA_UPRA introduces a control loop towards enabling their QoS-aware self-optimization, while at the base station, DUA_UPRA realizes a flexible algorithm, executed at the end of each time slot, to obtain optimal users' power and rate vectors for the subsequent time slot via obtaining the solution of (13).

DUA_UPRA Scheme

At Mobile Nodes [A Control Loop]

Step 1 (Information Monitoring). A user computes the actual amount of data units that has received within his current observation timeinterval according to (7).

Step 2 (Information Analysis). Determines his need for accessing system resources with respect to his QoS prerequisites (3), in accordance to (9) or (10).

Step 3 (Decision Making Towards Self-Optimization). Reflects his QoS requirements and resources expectation at the scheduler by adjusting his utility function following (8) and then, disseminates this information at the base station.

At Base Station (A Resource Scheduler)

Step 1. The scheduler requests users' utility functions.

Step 2. The non-convex power and rate optimization problem (13) is redefined with respect to the current users' utilities (i.e., $U_{\mathrm{RT},i}^*(P, a, \widehat{b}_{\mathrm{RT},i})$ for all $i, \in, S_{\mathrm{RT}}$, and $U_{\mathrm{NRT},j}^*(P, a, b_{\mathrm{NRT},j})$ for all $j \in S_{\mathrm{NRT}}$.

Step 3. Users' selection is performed for the current optimization problem, according to the mobile selection procedure in (17)–(19).

Step 4. Users' downlink allocated paower and throughput are estimated for non-real-time users from $\widehat{R}_{\mathrm{NRT},i}^{R_i^*} = \widehat{U}_{\mathrm{NRT},i}^{R_i^*}(P_i, a, b_{\mathrm{NRT},i})$ for all $i \in S_{\mathrm{NRT}}$ and for real-time users from $\widehat{R}_{\mathrm{RT},i}^{R_i^*} = \widehat{U}_{\mathrm{RT},i}^{R_i^*}(P_i, a, b_{\mathrm{RT},i})$ for all $i \in S_{\mathrm{RT}}$, according to the power and rate allocation algorithm (PAA) in [18]. Let us underline, that a real-time user i actual downlink power and rate estimation is a function only of $b_{\mathrm{RT},i}$ parameter in Step 4 and hence of his corresponding transmission scheme.

In the following, the complexity of DUA_UPRA scheme is discussed. We initially place emphasis on DUA_UPRA scheme at the mobile node, due to the low computational power of mobile devices. Specifically, the proposed control loop needs to perform the following computations to obtain: (a) a finite summation (7), (b) one normalized real number via (9) or (10) (a summation and a deviation), and then (c) an additional sum in (8). The latter requires the computation of the upper and lower bounds of $\widehat{b}_{\mathrm{RT},i}$ via the algorithms provided in Appendix A. The maximum upper-bounded number $\hat{v}$ of iterations required to obtain the above

bounds is also justified in Appendix A. Apart from the time complexity, due to mobile devise hardware limitations, space considerations are also important. For the implementation of our proposed approach, the mobile device needs to store $2^* W_i + 1$ (i.e., maintained on its memory) real numbers in order DUA_UPRA scheme at the mobile node to operate (i.e., W_i real numbers for computing its short-term throughput performance (7), W_i real numbers for the corresponding slots weights and *1* for maintaining the value of $\hat{b}_{\mathrm{RT},i}$). Finally, concerning DUA_UPRA scheme at the base station, we adopt the low complexity algorithms provided in [18] towards obtaining the solution of non-convex optimization problem (14) (a simple shorting and a simple bisection algorithms with overall upper-bounded number of iterations to convergence).

5. Numerical Results and Discussions

In this section, the operation and performance of the proposed dynamic utility adaptation-based users' power and rate allocation scheme DUA_UPRA is evaluated via modeling and simulation. In order to better illustrate the performance and the efficacy of the proposed scheme, in terms of average achieved actual downlink throughput and RT users' short-term throughput constraints satisfaction, we compare it against the performance of a fundamental utility-based power and rate allocation scheme [24] (in the following, we refer to it as UPRA algorithm) which only aims at optimizing users' actual throughput performance, without considering RT users' QoS prerequisites; therefore serving the purpose of system's performance benchmark.

Throughout our study, we considered a single cell time-slotted CDMA system. The duration of a slot is assumed to be 1.67 msec and the simulation lasts for *10,000* slots. We assume that the base station is located at the cell's center and that its maximum transmission power is $P_{\max} = 10$ (Watts). We model the path gain from the base station to user i, G_i as $G_i = K_i/d_i^{a'}$ (Rayleigh channels), where d_i is the distance of user i from the base station, a' is the distance loss exponent ($a' = 4$), and K_i is the log-normal distributed random variable with mean *0* and variance $\sigma^2 = 8$ (dB), which represents the shadowing [27].

We assumed that the system's spreading bandwidth is $W = 10^8$ and that the maximum downlink rate for all users is $R_i^{\max} = 2 \cdot 10^3$ Kbps. The total number of continuously backlogged users in the system is $N = 30$, and we considered two types of users, namely, non-real-time users (N_{NRT}) and real-time users (N_{RT}). Unless otherwise explicitly indicated, in the following, we consider that real-time users require constant downlink rates of $R_{\mathrm{RT},i} = 512$ kbps for all $i \in S_{\mathrm{RT}}$ (i.e., CBR real-time traffic) while their corresponding observation time intervals were set to $W_i = 20$ slots for all $i \in S_{\mathrm{RT}}$, and therefore an RT user's short-term data units threshold is set equal to $B_{\mathrm{RT},i} = 17002$ bits. We consider saturated NRT users requesting best effort NRT services, aiming at maximizing the achieved actual downlink throughput. The system's triggering parameter is $Tr = 0.3$. Both types of users are assumed to have the same transmission scheme. Therefore, we considered that their $f_i(\gamma)$ functions' parameters are $a = 2$ [18] and $b_{\mathrm{NRT},j} = b_{\mathrm{RT},i} = 3$ for all $i \in S_{\mathrm{RT}}$, for all $j \in S_{\mathrm{NRT}}$. In order to compute real-time users' minimum and maximum values for their parameter $\hat{b}_{\mathrm{RT},i}(t)$ in each time slot t, according to the algorithms proposed in Appendix A, we considered that $\varepsilon = 10^{-5}$ and $L_{\max} = 10^5$.

With the objective of better evaluating the performance of the proposed DUA_UPRA scheme, we considered four basic scheduling scenarios. In the first scenario, referred to as SC1, in order to explore our scheme's behavior in terms of satisfying RT and NRT service QoS requirements and to gratify that the proposed dynamic users' utilities adaptation framework DUA reflects correctly their corresponding degree of satisfaction, we assumed that all users have the same average channel conditions. In the second and the third scenarios (SC2 and SC3), we evaluate the performance of our proposed scheduler when users with different average channel conditions are served, considering, respectively, only RT users (SC2) and both NRT and RT users (SC3) at the system in order to demonstrate our schemes' flexibility in adapting the resource allocation process according not only to users' various QoS requirements but also to their average channel conditions, aiming at reducing the drawbacks emerging from the users' "near-far" effect. Finally, the fourth scheduling scenario (SC4) aims at demonstrating and revealing DUA_UPRA algorithms efficacy in supporting variable rate real-time traffic users.

5.1. Scheduling Scenario 1 (SC1). Figure 1 illustrates the total system's actual average throughput as a function of the number of RT users in the system (i.e., $N_{\mathrm{RT}} = 5, 10, \ldots, 30$, and therefore RT users' percentage in the system ranges from 16.67% *to* 100%, resp.), while Figure 2 presents RT users' probabilities of not satisfying their short-term throughput requirements (i.e., $\Pr[\hat{\beta}_{\mathrm{RT},i}(t) \leq B_{\mathrm{RT},i}]_{W_i}$) as a function of their number in the system under UPRA algorithm (black columns) and DUA_UPRA scheme (blue columns). All users average channel conditions are similar (i.e., are placed at same distances from the cell's center), therefore, only their fast fading attribute affects their instantaneous values.

We can clearly observe from Figure 2 that RT users' probabilities of not satisfying their short-term throughput requirements are significantly reduced under DUA_UPRA scheme, compared to the UPRA algorithm, even for large numbers of RT users in the system. Furthermore, our scheduling scheme's efficacy in satisfying RT users' QoS requirements is obtained without any considerable loss in the system's average (per user) achieved throughput, since as shown in Figure 1 system's average achieved throughput under DUA_UPRA scheme remains very close to the optimal one achieved by a pure opportunistic utility-based algorithm (UPRA). The observed loss in overall system's average throughput under DUA_UPRA is due to RT users slight overprovisioning of available resources towards maintaining their strict short-term throughout prerequisites (i.e., fixed amount of data per short-term windows). On the other hand, the latter resources are allocated to NRT users under UPRA, which are purely opportunistically served and, therefore, obtain increased average actual throughout

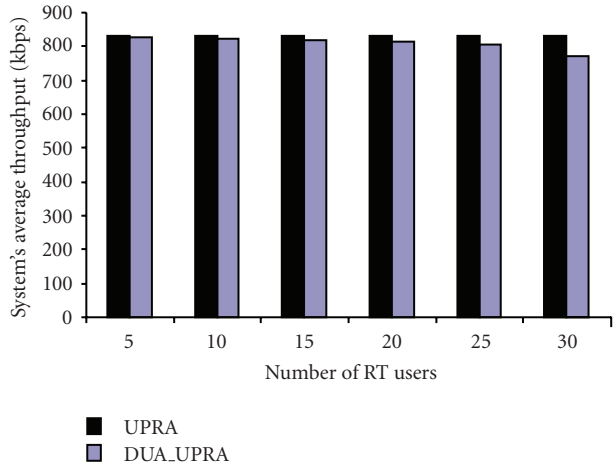

Figure 1: System's average throughput in SC1.

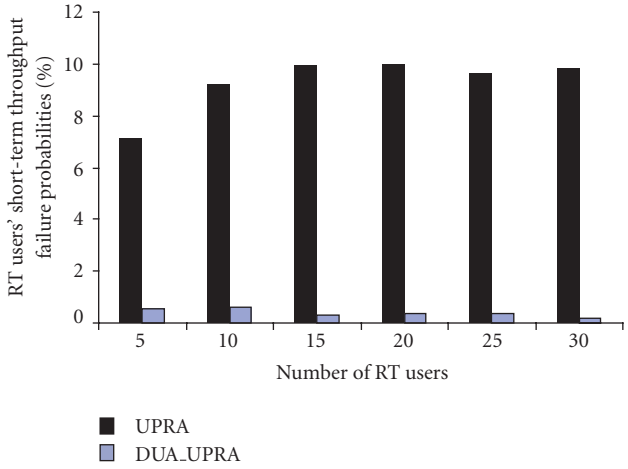

Figure 2: RT users' short-term throughput failure probabilities in SC1.

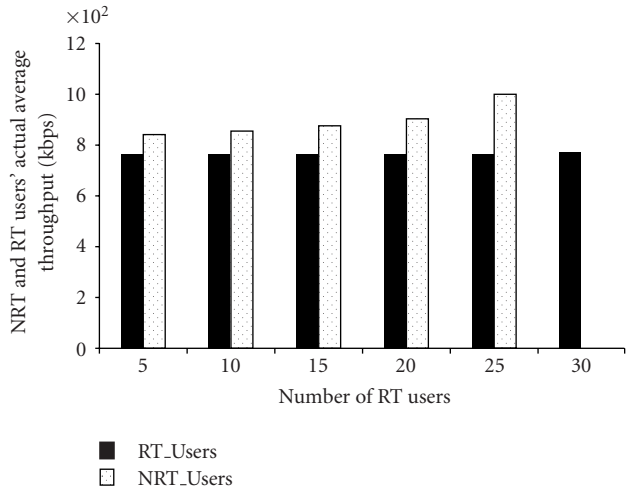

Figure 3: RT and NRT user's actual average received throughput under DUA_UPRA scheme in SC1.

(leading to better overall system throughput), at the expense of high RT user's short-term throughput failure probabilities (i.e., high RT users' performance degradation). The latter tradeoff is revealed in more detail in the following scenarios as well.

Moreover, by closely observing the allocation of system resources, in terms of actual average throughput for each one of the considered types of users individually, we can further see our scheme's property of exploiting the opportunistic nature individually for each type of users in order to optimize their diverse QoS requirements. Therefore, Figure 3 illustrates RT (NRT) users' actual average throughput as a function of their number in the system under DUA_UPRA. Specifically, it can be observed that an RT user's average received throughput remains almost constant, independent of their number in the system, due to DUA_UPRA scheme's property of allocating system resources to RT users up to the point where their required streaming throughput is satisfied. It is noted that, as observed in Figure 3, RT users' average achieved throughput is slightly larger than their predefined fixed downlink transmission rate, due to the system's preemptive nature when supporting RT services (determined by the values of triggering parameter Tr). On the other hand, an NRT user's average received throughput increases as the number of NRT users in the system decreases because the degree of competition among them for the excess system resources decreases as well, which is an inherent characteristic of any opportunistic scheduler.

With the presentation of the following two figures (Figures 4 and 5), we focus on DUA_UPRA scheme's performance under the most demanding case in *SC*1, in terms of RT users' short-term throughput QoS requirements satisfaction, where all the users in the system are RT users (i.e., $N_{RT} = N = 30$). We aim at demonstrating that DUA_UPRA scheme's enhanced performance, with respect to RT services QoS properties, asserts and affects all RT users and not only a portion of them, despite the large fluctuations on their channel conditions (due to fast fading). Specifically in Figure 4, we present each RT user's average actual achieved throughput, for all thirty users in the system, while in Figure 5 their corresponding short-term throughput failure probabilities under DUA_UPRA scheme (black dots) and UPRA algorithm (gray square) are depicted.

We observe that all users' probabilities of not satisfying their short-term throughput constraints are very small (maximum value: 0.7% average: 0.19%) under DUA_UPRA, while under UPRA they are high and diverse (maximum value: 20.7% average: 9.8%). Furthermore, under DUA_UPRA real-time users' average achieved throughput remains very high compared to the one achieved under UPRA that exploits optimally system's throughput abilities without, however, providing short-term throughput constraints. Thus, all real-time users' average actual received throughput is almost the same under DUA_UPRA.

5.2. Scheduling Scenario 2 (SC2). In the second scheduling scenario SC2, we also considered a system with 30 RT users (i.e., $N = N_{RT} = 30$), however, separated into

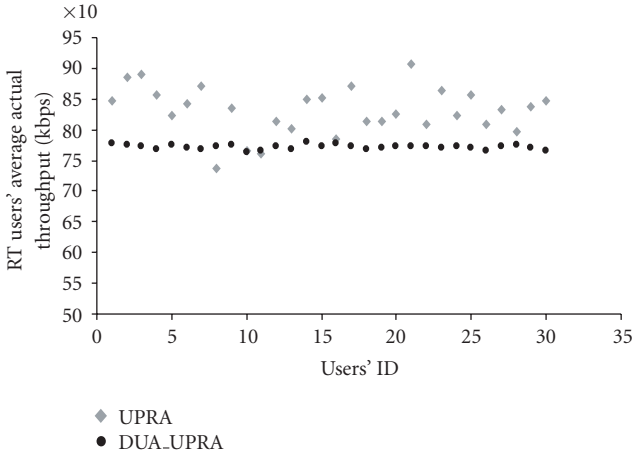

Figure 4: RT users' average throughput under UPRA and DUA_UPRA algorithms in SC1 (when $N = N_{RT} = 30$).

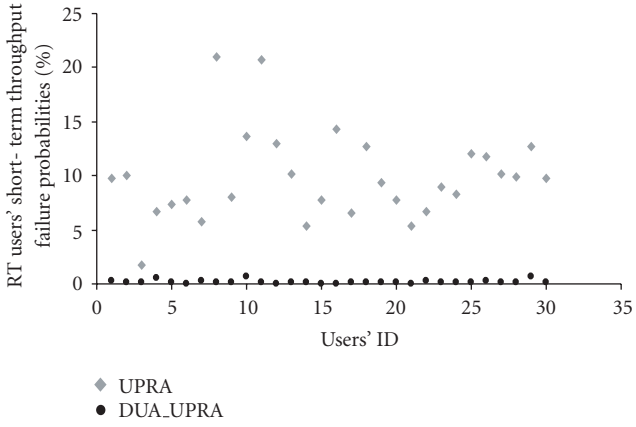

Figure 5: RT users' short-term throughput constraints failure probabilities under UPRA and DUA_UPRA algorithms in SC1 (when $N = N_{RT} = 30$).

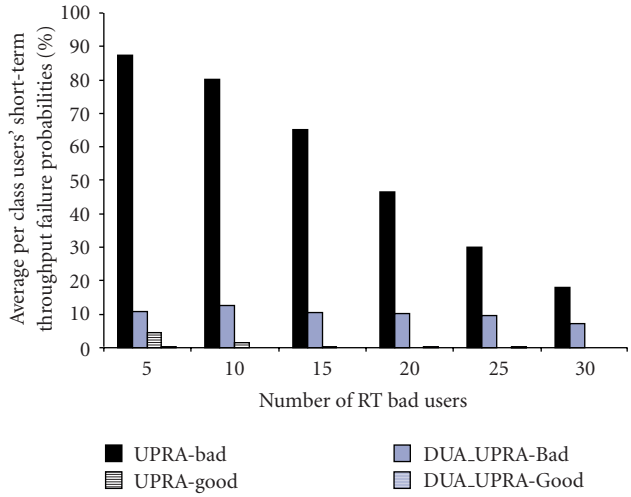

Figure 6: Users' average short-term throughput failure probabilities in SC2.

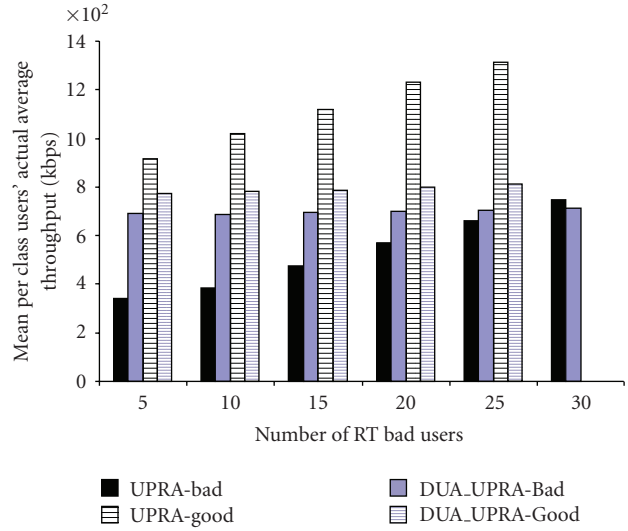

Figure 7: Users' average achieved throughput in SC2.

two classes: good users and bad users with good and bad average channel conditions, respectively. Good users' average channel conditions are assumed to be 7 dB larger than bad users. For each type of users in the system, we evaluate their average probabilities of not satisfying their short-term throughput constraints (illustrated in Figure 6), as well as their throughput performance (presented in Figure 7), as a function of the number of RT bad users in the system, under DUA_UPRA (blue columns, solid for bad users and stripes for good users) and UPRA (black columns, solid for bad users and stripes for good users) algorithms.

The corresponding results demonstrate that under UPRA algorithm (black solid columns) bad users are strongly unfavored, not only in terms of their short-term throughput constraints dissatisfaction (Figure 6) but also due to their low throughput performance (Figure 7), especially when their number in the system is low. This mainly occurs due to UPRA goal of maximizing the system's total utility. Bad users' contribution to the maximization of the system's total utility is very low (i.e., they practically contribute only when their channels are very good compared to good users' average channel conditions), and therefore they are rarely selected by UPRA algorithm, which leads to their throughput performance degradation. On the other hand, under DUA_UPRA scheme bad users' short-term throughput performance is highly improved (Figure 7, solid blue columns). Especially, when their number in the system is small, the percentage of their short-term throughput dissatisfaction decreases even 75% compared with the corresponding one achieved under UPRA (solid black columns). Moreover, we observe that bad users' average downlink throughput takes the same values independently of their number in the system under DUA_UPRA scheme (Figure 7, solid blue columns).

Observing good users' performance metrics, we notice that their probabilities of not satisfying their short-term throughput constraints are highly improved under DUA_UPRA scheme (Figure 6 striped blue columns (last))

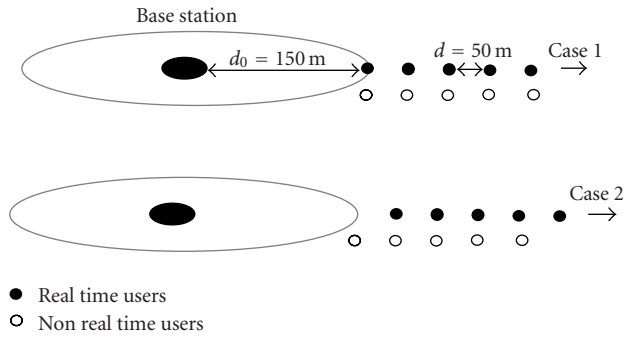

Figure 8: Network varying topology (per *Case #*) under Scenario 3 (SC3).

(i.e., always smaller than 0.18%), especially when their number in the system is high (number of bad users is small). On the other hand, under UPRA algorithm, due to the high competition, their short-term throughput constraints are still not satisfied (Figure 6 striped black columns). Finally, the downlink throughput performance reduction of good RT users under DUA_UPRA scheme, when compared to the one achieved under UPRA algorithm, is not only harmless (Figure 7 striped blue and black columns) since good RT users' required downlink rate is still achieved and satisfied, but rather desired since the excess system resources can be efficiently allocated to bad RT users in order to improve their short-term throughput requirements, as well as to other NRT users.

5.3. Scheduling Scenario 3 (SC3). With this scenario (SC3), we aim at studying DUA_UPRA scheme's ability to efficiently treat "near-far" effect in a more pragmatic wireless setting, as well as quantifying the tradeoff between RT users short-term throughput satisfaction fulfillment and system's overall achieved throughout. To that end, we consider $N = 30$ active users in the system, where five ($N_{\mathrm{RT}} = 5$) request real-time traffic ($R_{\mathrm{RT},i}$ =512 Kbps, $W_i = 20$ slots, $B_{\mathrm{RT},i} = 17002$ bits) for all $i \in S_{\mathrm{RT}}$, while the rest are considered as NRT users ($N_{\mathrm{NRT}} = 25$). NRT users constantly maintain their position with respect to cell's center, placed in groups of five NRT users in the following distances $d_0 = 150$ (m) $d_{p+1} = d_p + 50$ (m) for $p = 0, \ldots, 3$ (Figure 8). On the other hand, the set of RT users in the system is gradually moving away from cell's center (per case), as shown in Figure 8, in order to better simulate the fact that RT users experience Rayleigh fast fading channels with various average quality conditions, due to their corresponding distance to cell's base station. Thus, RT users' distances from cell's centre per case are provided in Table 1.

Figure 9, illustrates overall system downlink average throughput (black columns), as well RT and NRT users' average throughput (gray and dotted columns, resp.) for each one of the simulated cases (horizontal axis). Furthermore, the corresponding RT users' average short-term throughput failure probabilities are presented in Figure 10. The results show that the proposed DUA_UPRA scheme efficiently overcomes the problem of "near-far" effect, since RT users QoS

Table 1: RT users' distances in *(m)* for cell' center per *Case #* under Scenario 3 (SC3).

Case #	RT User ID				
	1	2	3	4	5
0	150	200	250	300	350
1	200	250	300	350	400
2	250	300	350	400	450
3	300	350	400	450	500
4	350	400	450	500	550
5	400	450	500	550	600

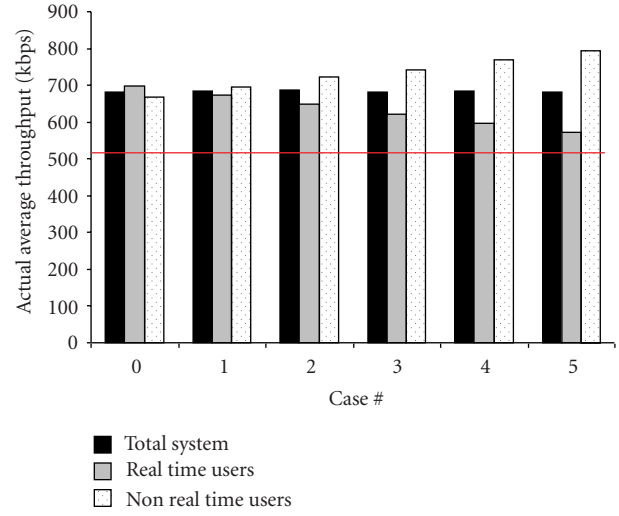

Figure 9: Total systems, RT and NRT users' average achieved actual throughput in SC3.

prerequisites are fulfilled, in terms of achieved average actual throughout larger than *512* kbps and short-term throughput failure probabilities less than *0.8*%, even under the most demanding scenario that is, *Case* = 5. On the other hand, a significant tradeoff is revealed. As the average channel quality of RT users decreases (i.e., RT users are moving away from cell's center) then, the system increases the number of slots allocated to them, in order to balance between their short-term throughput requirements fulfillment and their inevitable actual throughput degradation (due to their bad channel quality). Therefore, RT users' average throughput deceases (Figure 9, grey columns) but their short-term throughput failure probabilities remain very low (Figure 10). At the end, RT users QoS prerequisites are preserved, but, at the cost of low NRT users' throughput as well as overall system throughout. That is due to the small number of system slots allocated to NRT users. The latter, leads only to a small increment of NRT users' average throughput (as RT users are moving away from cell's center), even if their instantaneous achieved rates are increased, due to their good channel conditions (Figure 9, dotted columns).

5.4. Scheduling Scenario 4 (SC4). In this final set of simulations (SC4), we explore the service performance of a RT user requesting variable rate traffic under the proposed

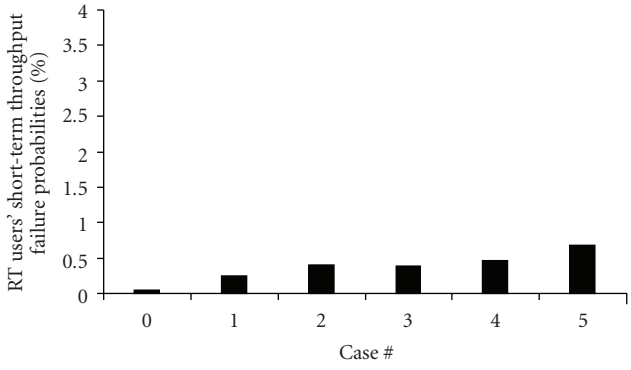

Figure 10: RT users' average short-term throughput failure probabilities in SC3.

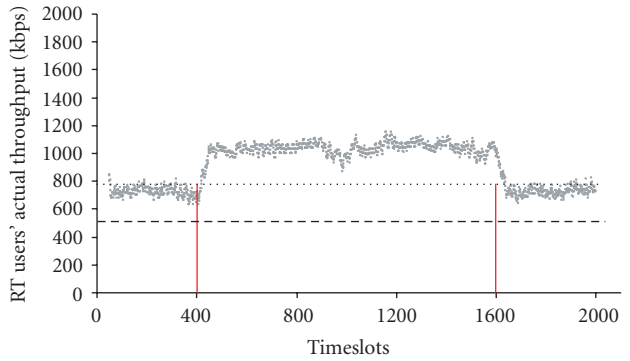

Figure 11: Actual achieved throughput as the system evolves in SC4 for a user with variable rate real-time traffic.

DUA_UPRA scheme, as the system evolves. To that end, we consider a scenario with $N = N_{RT} = 30$ real-time users, with the same average channel conditions (with fast-fading Rayleigh channels). All users except one user (user j) are assumed to have the same QoS prerequisites (i.e., CBR traffic of $R_{RT,i} = 512$ Kbps, $W_i = 20$ slots, $B_{RT,i} = 17002$ bits for all $i \in S_{RT}, i \neq j$). User j traffic is considered to be of variable rate as follows: from $t = 0$ *to* $t = 400$ timeslot, $R_{RT,j}(t) = 512$ (Kbps), from $t = 401$ *to* $t = 1600$ timeslot, $R_{RT,j}(t) = 768$ (Kbps) and from $t = 1600$ to $t = 2000$ timeslot, $R_{RT,j}(t) = 512$ (Kbps).

Figure 11 illustrates the variable rate RT user's instantaneous actual throughput as a function of time, while Figure 12 their corresponding short-term throughput probability (i.e., $\Pr[\widehat{\beta}_{RT,i}(t) \leq B_{RT,i}]_{W_i}$, at timeslot t) as a function of time, under DUA_UPRA scheme. In both figures, the timeslots at which user j required throughput alters are marked with vertical red lines, while the corresponding requested rates are presented with gray horizontal lines in Figure 11. The results show that the dynamic adaptation of the under consideration user's requested actual throughput is fulfilled under DUA_UPRA, and thus the timeframe required to complete a new request is less than *50* timeslots (i.e., less than 8.3 msec). Moreover, during the latter transition period (i.e., after a change of the requested throughput), the RT user's short-term throughput failure probabilities

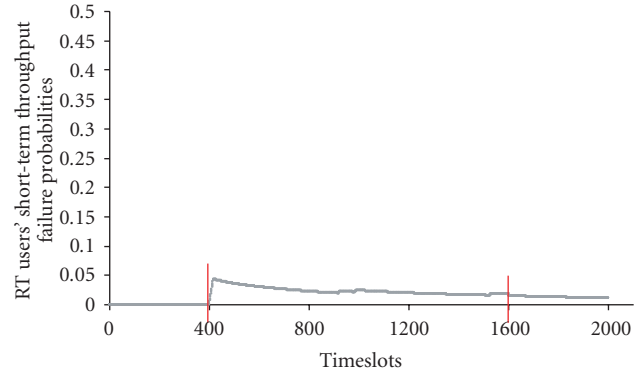

Figure 12: Short-term throughput failure probability (%) as the system evolves in SC4 for a user with variable rate real-time traffic.

slightly increases (maximum value 0.04 = 4%) and then starts dropping again, due to algorithms adaptation to the new request (Figure 12). Finally, we can observe that the user's actual throughput always remains slightly higher that the required (Figure 11), since in order to always fulfill their short-term throughput requirements (per time slot), the system slightly overprovides them with resources.

6. Concluding Remarks

In this paper, we studied the combined problem of allocating system resources, in terms of power assignment and transmission rate, in the downlink of a CDMA wireless system, where multiple services with various QoS requirements are simultaneously requested. We expressed users' degree of satisfaction with respect to their QoS demands fulfillment (non-real-time and real-time services) through a common utility-based framework which provides us with the enhanced flexibility of effectively influencing the opportunistic scheduler to meet their various QoS prerequisites.

Emphasis was placed on RT users' probabilistic short-term throughput requirements satisfaction. Specifically, in order to dynamically and accurately affect their selection priority with respect to their QoS prerequisites satisfaction, we introduced the information of their short-term received data distribution into the proposed methodology of tuning their utility functions' properties. Through modeling and simulation under various scheduling scenarios, we demonstrated that significant performance improvements are achieved in terms of real-time user's short-term throughput requirement satisfaction and non-real-time users actual throughput maximization, without any considerable loss in the total system throughput.

It should be noted that in this work, we considered linear relationship between users' assigned data rates and their corresponding degree of satisfaction. However, the degree of a user's satisfaction with respect to their service quality can be more efficiently expressed by applying other than linear utility functions of their actual throughput rates [11]. The mathematical formulation and the analytical solution of the above utility optimization problem provide a first

step towards the realization of autonomic wireless network where hybrid data flows are simultaneously supported, and therefore is an issue of a great importance and part of our current research work.

Appendices

A. Limitations on Controlling Users' Selection Priorities

In the following, we rigorously define the lower and upper bounds of real-time users' $\hat{b}_{\mathrm{RT},i}$ for all $i \in S_{\mathrm{RT}}$ parameter (i.e., $b_{i,\min}$ and $b_{i,\max}$, resp.) and justify their role in the proposed scheduling policy as well as the way they restrict users' ability to self-optimize their services' QoS performance. It is shown that even if RT users' introduced self-optimization behavior enhances their ability to self-optimize their QoS-aware performance, mobile nodes' potential bad channel conditions, the system's lack of available resources, as well as their physical limitations may prevent the fulfillment of their short-term QoS requirements. Moreover, we provide low complexity algorithms for computing the above boundary values.

As analyzed in the previous sections, a real-time user's utility function parameter $\hat{b}_{\mathrm{RT},i}$ plays a key role in the selection priority of accessing system's resources at each time slot, since it affects the corresponding value of their willingness to pay $\lambda_i^{\max}$. On the other hand, the appropriate values for $b_{\min,i}$ and $b_{\max,i}$ should be such that for any further decrement or increment in the value of $\hat{b}_{\mathrm{RT},i}$ (with respect to these bounds) the corresponding value of parameter $\lambda_i^{\max}$ is not affected. Consequently, parameter $\hat{b}_{\mathrm{RT},i}$ should be bounded among them. Therefore, we use the following condition for identifying an RT user i bounds of their $\hat{b}_{\mathrm{RT},i}$ parameter

$$\left.\frac{\partial \lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})}{\partial \hat{b}_{\mathrm{RT},i}}\right|_{\substack{\hat{b}_{\mathrm{RT},i}<b_{\min,i}\\ \hat{b}_{\mathrm{RT},i}>b_{\max,i}}} = 0, \tag{A.1}$$

where $\lambda_i^{\max} \equiv \lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})$ represents RT users' maximum willingness as a function of his parameter $\hat{b}_{\mathrm{RT},i}$.

Definitions of $b_{\min,i}$ and $b_{\max,i}$ for all $i \in S_{\mathrm{RT}}$. Following a pure functions theoretic analysis, the lower bound of an RT user's utility function parameter $\hat{b}_{\mathrm{RT},i}$ can be formally defined as follows.

Proposition 2 (Definition of $b_{\min,i}$ for all $i \in S_{\mathrm{RT}}$). *We defines as $b_{\min,i}$ for all $i \in S_{\mathrm{RT}}$ the maximum value of a RT user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ at slot t, such that*

$$b_{\min,i} = \max\left\{ \hat{b}_{\mathrm{RT},i} : P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)\Big|_{\hat{b}_{\mathrm{RT},i}} > P_i^0\Big|_{\hat{b}_{\mathrm{RT},i}} \quad \text{where} \left.\frac{\partial^2 U_i^{R_i^*}(P_i)}{\partial {P_i}^2}\right|_{P=P_i^0|_{\hat{b}_{\mathrm{RT},i}}} = 0, \right.$$

$$\left. \left[U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right) - P\frac{\partial U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right)}{\partial P} \right]_{P=P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}}} > 0 \right\} \tag{A.2}$$

where $b_{\min,i} \in [0, b_{\mathrm{RT},i})$. Thus, when $\hat{b}_{\mathrm{RT},i} < b_{\min,i}$ then $\partial\lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})/\partial\hat{b}_{\mathrm{RT},i}|_{\hat{b}_{\mathrm{RT},i}<b_{\min,i}} = 0$.

Proof. See Appendix C. □

The existence of such a lower bound reveals an inherent users' limitation on controlling their services' short-term QoS requirements when operating over a time-varying wireless environment; the reason is twofold. On the one hand, due to its opportunistic nature, the scheduler in the sight of low available resources may potentially unfavor some user's towards optimizing overall system's welfare. On the other hand, even when plethora of system resources is available, RT users' potential bad instantaneous channel condition and their physical limitations may make the goal of their short-term QoS prerequisites fulfillment unreachable.

The following proposition defines an RT user's utility function parameter $\hat{b}_{\mathrm{RT},i}$ upper bound.

Proposition 3 (Definition of $b_{\max,i}$ for all $i \in S_{\mathrm{RT}}$). *One defines as $b_{\max,i}$ for all $i \in S_{\mathrm{RT}}$ the maximum value of a real-time user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ at time slot t such that*

$$b_{\max,i} = \left\{ \hat{b}_{\mathrm{RT},i} : \min\left(\left.\frac{\partial \hat{b}_{\mathrm{RT},i}}{\partial \lambda_i^{\max}}\right|_{\hat{b}_{\mathrm{RT},i}=b_{\max,i}} \le -L_{\max}, B_{\mathrm{MAX},i}(A_i) \right) \right\}, \tag{A.3}$$

where $L_{\max}$ is a large positive number and $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$. Thus,

$$B_{\mathrm{MAX},i}(A_i) = \frac{1}{a}\left\{ \ln\left(\frac{c_i}{1 + (A_i/aW(P_{\max} + A_i)) + c_i d_i} - 1 \right) + \frac{aWP_{\max}}{R_i^{\max} A_i} \right\}. \tag{A.4}$$

Proof. See Appendix D. □

From the definition of $B_{\mathrm{MAX},i}(A_i)$ we can observe that the worse RT user's channel conditions are (i.e., parameter A_i increases), the smaller is our ability of influencing their selection priority, since the range of their utility function parameter $\hat{b}_{\mathrm{RT},i}$ decreases as well.

Algorithms for Computing $b_{\min,i}$ and $b_{\max,i}$. We conclude this section by introducing two low complexity algorithms for computing RT users' parameters $\hat{b}_{\mathrm{RT},i}$ lower and upper

bounds at each time slot. Initially, by using Proposition 2, we provide a "divide and conquer"-based algorithm for computing a real-time user's parameter $\hat{b}_{\mathrm{RT},i}$ lower bound $b_{\min,i}$.

Algorithm for Computing ($b_{\min,i}$ for all $i \in S_{\mathrm{RT}}$). Let us refer to $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}(t)=b_{\min,i}(t)^{(\nu)}} > P_i^0|_{\hat{b}_{\mathrm{RT},i}(t)=b_{\min,i}(t)^{(\nu)}}$ and $[U_i^{R_i^*}(P, \hat{b}_{\mathrm{RT},i}) - P(\partial U_i^{R_i^*}(P, \hat{b}_{\mathrm{RT},i})/\partial P)]_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)|_{\hat{b}_{\mathrm{RT},i}=b_{\min,i}^{(\nu)}}} > 0$ as conditions A and B, respectively.

(i) Set $\nu = 1$, $l^{(\nu)} = 0$, $r^{(\nu)} = b_{\mathrm{RT},i}$ and $b_{\min,i^{(\nu)}} = b_{\mathrm{RT},i}$.

(ii) If A AND B are true then $l^{(\nu+1)} = (l^{(\nu)} + r^{(\nu)})/2$, $r^{(\nu+1)} = r^{(\nu)}$ and $b_{\min,i^{(\nu+1)}} = l^{(\nu+1)}$ else $l^{(\nu+1)} = l^{(\nu)}$, $r^{(\nu+1)} = (\, l^{(\nu)} + r^{(\nu)})/2$ and $b_{\min,i^{(\nu+1)}} = r^{(\nu+1)}$.

(iii) If $|l^{(\nu+1)} - l^{(\nu)}| > \varepsilon OR\ |r^{(\nu+1)} - r^{(\nu)}| > \varepsilon$ then $\nu = \nu + 1$, go to (ii).

(iv) If NOT ($AANDB$) is true then $b_{\min,i^{(\nu+1)}} = b_{\min,i^{(\nu+1)}} - \varepsilon$.

The maximum number of algorithm's iterations ν^* is $\nu^* \leq \log_2(b_{\mathrm{RT},i}/\varepsilon) + 1$, where ε is a small positive constant. Finally, with respect to Proposition 3, we propose the adoption of a "Steepest Descent"-based algorithm for computing at every time slot t, a real-time user i parameter $\hat{b}_{\mathrm{RT},i}$ upper bound, $b_{\max,i}$.

Algorithm for Computing ($b_{\max,i}$ for all $i \in S_{\mathrm{RT}}$).

(i) Let $L_{\max}$ be a large positive constant.

(ii) $n = 0$, $b_{\max,i}^{(n=0)} = b_{\mathrm{RT},i}$.

(iii) For $n > 0$ then $b_{\max,i}^{(n+1)} = b_{\max,i}^{(n)} + 10^{\lfloor D^{(n)}-1 \rfloor}$ where: $D^{(n)} = \partial\hat{b}_{\mathrm{RT},i}/\partial\lambda_i^{\max}|_{\hat{b}_{\mathrm{RT},i}=b^{(n)}{}_{\max,i}}$.

(iv) If $\partial\hat{b}_{\mathrm{RT},i}/\partial\lambda_i^{\max}|_{\hat{b}_{\mathrm{RT},i}=b^{(n+1)}{}_{\max,i}} < -L_{\max}$, go to (vii)

(iv) If $b_{\max,i}^{(n+1)} > B_{\mathrm{MAX},i}(A_i)$, then go to (vii)

(vi) $n = n + 1$ and go to (iii).

(vii) Stop.

The previous algorithm is a modified "Steepest Descent"-based algorithm, adapted to the needs of our problem (Proposition 3). Specifically, in our case, we are not interested in finding the minimum of the function $\hat{b}_{\mathrm{RT},i}(\lambda_i^{\max})$, but in accordance to (A.1), in finding a large value of b at which the absolute values of gradient $|\partial\hat{b}_{\mathrm{RT},i}/\partial\lambda_i^{\max}|_{\hat{b}_{\mathrm{RT},i}=b_{\max,i}}|$ are very large (for practical considerations, we approximate infinite with a large number $L_{\max}$). Therefore, to improve convergence time, we use as a step the corresponding power of *10* (in step *iii*). It is easy to show, considering the complexity of the "Steepest Descent" [28], that the maximum number of iterations required for convergence are ν^{**}, where $\nu^{**} \leq \lceil (1/2)\ln(L_{\max}) \rceil$.

Finally, the total maximum number of iterations required to compute the minimum and the maximum values of $\hat{b}_{\mathrm{RT},i}$ is $\hat{\nu} \leq \log_2(b_{\mathrm{RT},i}/\varepsilon) + 1 + \lceil (1/2)\ln(L_{\max}) \rceil$, where ε and $L_{\max}$ are a small and a large positive constants, respectively.

B. Proof of Proposition 1

From (14), we can observe that when for two users i, j, $A_i = A_j$ and $R_i^{\max} = R_j^{\max}$, then the only parameters that are affected by variations in the value of parameters b_i, b_j, and hence determine the properties of their utility functions $U_i^{R_i^*}(P, a, b_i)$ and $U_j^{R_j^*}(P, a, b_j)$ are γ_i^*, γ_j^* and the corresponding values of their f_i *and* f_j functions, respectively. Therefore, we first provide the following two lemmas that determine the way that a user i, parameter b_i affects his f_i function and his γ_i^* parameter.

Lemma 4. *If $b_i < b_j$ then $f_i(\gamma, a, b_i) > f_j(\gamma, a, b_j)$ for all $\gamma \in (0, \infty)$.*

Proof. If one sets as $x_i = e^{ab_i}$ in the generic definition of a sigmoidal function:

$$f_i\left(\gamma_i\left(R_i, \overline{P}\right), a, b_i\right) = c_i\left\{\frac{1}{1 + e^{-a(\gamma_i - b_i)}} - d_i\right\}, \tag{B.1}$$

where $c_i = (1 + e^{ab_i})/e^{ab_i}$ and $d_i = 1/(1 + e^{ab_i})$, one can rewrite itas $f_i(\gamma, a, b_i) = (1 - e^{-a\gamma})/(1 + x_i e^{-a\gamma})$. Thus, if $b_i < b_j$ then $1 + x_i e^{-a\gamma} < 1 + x_j e^{-a\gamma}$ and $f_i(\gamma, a, b_i) > f_j(\gamma, a, b_j)$ for all $\gamma \in (0, \infty)$. □

Lemma 5. *If $A_i = A_j$ and $R_i^{\max} = R_j^{\max}$, if $b_i < b_j$, then $\gamma_i^* < \gamma_j^*$.*

Proof. From (14), one can compute a user i, parameter γ_i^* value as follows: $\gamma_i^* = \max_{\gamma \geq 1}\{(1/\gamma) f_i(\gamma, a, b_i)\} = \max_{\gamma \geq 1}\{(1/\gamma)(e^{a\gamma} - 1)/(e^{a\gamma} + x_i)\}$, where $x_i = e^{ab_i}$. Let one defines as $g(x_i, \gamma) = (1/\gamma)(e^{a\gamma} - 1)/(e^{a\gamma} + x_i)$ and as $h(x_i, \gamma) = \partial g(x_i, \gamma)/\partial\gamma = e^{a\gamma}(a\gamma x_i - e^{a\gamma} - x_i + 1 + a\gamma) + x_i/(e^{a\gamma} + x)^2\gamma^2$. When $\gamma_i = \gamma_i^*$, then $\partial g(x_i, \gamma)/\partial\gamma = 0$ since γ_i^* is maximum and

$$\begin{aligned} h(x_i, \gamma) &> 0 \quad \forall\, \gamma < \gamma_i^*, \\ h(x_i, \gamma) &< 0 \quad \forall\, \gamma > \gamma_i^*. \end{aligned} \tag{B.2}$$

Thus, if $x_i > x_j$, then $h(x_i, \gamma) > h(x_j, \gamma)$ since

$$\frac{\partial h(x_i, \gamma)}{\partial x_i} = \frac{e^{a\gamma} a\gamma (e^{a\gamma} - 1)}{(e^{a\gamma} + x_i)^3 \gamma^2} > 0 \quad \forall\, a \geq 0,\ \gamma \geq 0,\ b_i \geq 0. \tag{B.3}$$

If there exists x_i, γ_i^* such that $h(x_i, \gamma_i^*) = 0$, and x_j, γ_j^* such that $h(x_j, \gamma_j^*) = 0$, then if $x_j > x_i$, from (B.3), one has $h(x_j, \gamma_i^*) > h(x_i, \gamma_i^*)b$, and hence $h(x_j, \gamma_i^*) > 0$. Thus, since $h(x_j, \gamma_j^*) = 0$ and $h(x_j, \gamma_i^*) > 0$, from (B.2) one has $\gamma_i^* < \gamma_j^*$.

Finally, since if $x_j > x_i$ then $\gamma_i^* < \gamma_j^*$, and when $x_j > x_i$ then $b_i < b_j$, one concludes that if $b_i < b_j$ then $\gamma_i^* < \gamma_j^*$. Finally, based on Lemmas 4 and 5 one proves Proposition 1. From Remark 1 (Proposition 3, [24]) one has seen that if for any two users i, j $U_i^{R_i^*}(P, a, b_i) > U_j^{R_j^*}(P, a, b_j)$ for $0 \leq P \leq P_{\max}$, then $\lambda_i^{\max} > \lambda_j^{\max}$. Therefore one has to prove that if $b_i < b_j$, then $U_i^{R_i^*}(P, a, b_i) > U_j^{R_j^*}(P, a, b_j)$ for $0 \leq P \leq P_{\max}$.

By the definition of a user i utility in (14), and Lemma 5, let one defines:

$$K_i = \frac{R_i^{\max}\gamma_i^*(\theta \mathrm{P}_{\max} + A_i)}{W + \theta R_i^{\max}\gamma_i^*} \le \frac{R_j^{\max}\gamma_j^*\left(\theta P_{\max} + A_j\right)}{W + \theta R_j^{\max}\gamma_j^*} = K_2. \tag{B.4}$$

Furthermore, there are three possible cases for the value of parameter P, when $0 \le P \le P_{\max}$.

Case 1. If $P \le K_1 < K_2$, then from (14) if $b_i < b_,$ then $f_i(\gamma_i^*, a, b_i)/\gamma_i^* \ge f_j(\gamma_j^*, a, b_j)/\gamma_j^*$, and hence it easily follows that $U_i^{R_i^*}(P, a, b_i) \ge U_j^{R_j^*}(P, a, b_j)$.

Case 2. If $P \ge K_1$ and $P \ge K_2$, then from Lemma 4 if $b_i < b_j$ we have $f_i(\gamma_i^*, a, b_i) \ge f_j(\gamma_j^*, a, b_j)$, and it follows that $U_i^{R_i^*}(P, a, b_i) \ge U_j^{R_j^*}(P, a, b_j)$.

Case 3. If $P \ge K_1$ and $P \le K_2$, then since $R_i^{\max} f_i(\gamma_i(R_i^{\max}, P), a, b_i) \ge WP/(\gamma_i^*(\theta P_{\max} - \theta P + A_i))f_i(\gamma_i^*, a, b_i)$ and (as in Case 1) $R_i^{\max} f_i(\gamma_i(R_i^{\max}, P), a, b_i) \ge WP/(\gamma_j^*(\theta P_{\max} - \theta P + A_j))f_j(\gamma_j^*, a, b_j)$, we can conclude that $U_i^{R_i^*}(P, a, b_i) \ge U_j^{R_j^*}(P, a, b_j)$.

Finally, since when $b_i < b_j, U_i^{R_i^*}(P, a, b_i) > U_j^{R_j^*}(P, a, b_j)$ for $0 \le P \le P_{\max}$, the proof is completed. □

C. Proof of Proposition 2

In order to determine a lower bound of an RT user's parameter $\hat{b}_{\mathrm{RT},i}$, we first identify some of the main properties of his utility function with respect to $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$. The following lemma defines the relationship between the inflection point of a user's sigmoidal-like utility function and its parameter $\hat{b}_{\mathrm{RT},i}$.

Lemma 6. *For any two values $b'_{\mathrm{RT},i}$, $b''_{\mathrm{RT},i}$, of user i utility parameter $\hat{b}_{\mathrm{RT},i}$ such that $b'_{\mathrm{RT},i} < b''_{\mathrm{RT},i}$, it holds that $P_i^{0'} < P_i^{0''}$ where $P_i^{0'}$, $P_i^{0''}$ are the inflection points of their corresponding utilities $U_i^{R_i^*}(P_i, b'_{\mathrm{RT},i})$ and $U_i^{R_i^*}(P_i, b''_{\mathrm{RT},i})$, respectively.*

Proof. See Appendix E. □

Furthermore, we can also prove that when $\hat{b}_{\mathrm{RT},i} = 0$, then a user's utility function inflection point has always smaller value than his utility separation point, $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)$.

Lemma 7. *When $\hat{b}_{\mathrm{RT},i} = 0$, $P_i^0|_{\hat{b}_{\mathrm{RT},i}=0} < P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=0}$.*

Proof. See Appendix F. □

Based on the two previous lemmas, we can see that there always exists a value $\hat{b}^*_{\mathrm{RT},i}$ for a user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$, such that for smaller than $\hat{b}^*_{\mathrm{RT},i}$ values of $\hat{b}_{\mathrm{RT},i}$, the inflection point of function $U_i^{R_i^*}(P_i, \hat{b}_{\mathrm{RT},i})$ is always smaller that its separation point (i.e., $P_i^0 < P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)$). Therefore, we can provide the following proposition.

Proposition 8. *There always exists a value of a real-time user i utility function parameter $\hat{b}_{RT,i}$ when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$, denoted as $\hat{b}^*_{\mathrm{RT},i}$, such that $P_i^{LIM}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=\hat{b}^*_{\mathrm{RT},i}} = P_i^0|_{\hat{b}_{\mathrm{RT},i}=\hat{b}^*_{\mathrm{RT},i}}$, and hence $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}} > P_i^0|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}}$ when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^*_{\mathrm{RT},i})$.*

Proof. See Appendix H. □

We can now proceed to prove that when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^*_{\mathrm{RT},i})$, there always exists a value for a user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ denoted as $b_{\min,i}$ such that when $\hat{b}_{\mathrm{RT},i} < b_{\min,i}$ then their maximum willingness to pay is calculated by the second part of (17), because P_i^* in (18) does not exist and hence, condition (A.1) is fulfilled. Towards that, we first provide the following proposition.

Proposition 9. *There always exists a value for a real-time user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^*_{RT,i})$, indicated as $\hat{b}^{**}_{RT,i}$ such that when $\hat{b}_{\mathrm{RT},i}$ $\hat{b}^{**}_{\mathrm{RT},i}$ then there is no $P_i^* \in [P_i^0, P_{\max}]$ such that*

$$U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right) - P\frac{\partial U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right)}{\partial P} = 0 \quad \text{for } P_i^0 \le P \le P_{\max}. \tag{C.1}$$

Proof. See Appendix I. □

According to Proposition 8, there always exists a value for a real-time user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ (i.e., $\hat{b}^*_{\mathrm{RT},i}$) when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$ such that for $0 \le \hat{b}_{\mathrm{RT},i} < \hat{b}^*_{\mathrm{RT},i} < b_{\mathrm{RT},i}$ then

$$P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)\Big|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}} > P_i^0\Big|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}}, \tag{C.2}$$

and in accordance to Proposition 9, there always exists a value for a real-time user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ (i.e., $\hat{b}^{**}_{\mathrm{RT},i}$) when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$ such that for $0 \le \hat{b}_{\mathrm{RT},i} < \hat{b}^{**}_{\mathrm{RT},i} \le \hat{b}^*_{\mathrm{RT},i} < b_{\mathrm{RT},i}$ then

$$\left[U_i^{R_i^*}(P) - P\frac{\partial U_i^{R_i^*}(P)}{\partial P}\right]_{P=P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}(t)<\hat{b}^{**}_{\mathrm{RT},i}(t)}} > 0, \tag{C.3}$$

and hence (18) has no solution. Therefore, there always exist a value for parameter $\hat{b}_{\mathrm{RT},i}$ when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i})$, denoted as $b_{\min,i}$, such that both (C.2) and (C.3) are satisfied. Specifically, $b_{\min,i} = \hat{b}^{**}_{\mathrm{RT},i}$. Moreover, since when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^{**}_{\mathrm{RT},i} = b_{\min,i})$ then, (18) has no solution, and by (17), we have $\lambda_i^{\max} = U_i^{R_i^*}(P_{\max})/P_{\max} = (WP_{\max}/A_i)((1 - e^{-(aN_iP_{\max}/A_i)})/(1 + e^{a(\hat{b}_{\mathrm{RT},i} - (N_iP_{\max}/A_i))}))$, where $N_i = W/R_i^{\max}$. Furthermore, since $\hat{b}_{\mathrm{RT},i} \ll N_iP_{\max}/A_i$, we conclude that

when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\min,i})$, then $\lambda_i^{\max} = WP_{\max}/A_i$ and hence $\partial\lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})/\partial\hat{b}_{\mathrm{RT},i}|_{\hat{b}_{\mathrm{RT},i}<b_{\min,i}} = 0$.

Moreover, the proof is completed.

D. Proof of Proposition 3

In the following, we determine the sufficient conditions that will allow us to define an upper bound of an RT user's utility parameter $\hat{b}_{\mathrm{RT},i}$ when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ and to justify its purpose with respect to condition (A.1). Moreover, since $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$, then without loss of generality in the rest of this subsection, we assume that a user i function f_i parameters $c_i(\hat{b}_{\mathrm{RT},i}) = (1 + e^{a\hat{b}_{\mathrm{RT},i}})/e^{a\hat{b}_{\mathrm{RT},i}} \equiv c_i$ and $d_i(\hat{b}_{\mathrm{RT},i}) = 1/(1 + e^{a\hat{b}_{\mathrm{RT},i}}) \equiv d_i$ have fixed values independent from the value of parameter $\hat{b}_{\mathrm{RT},i}$.

Initially, we determine an upper bound of a real-time user's parameter $\hat{b}_{\mathrm{RT},i}$, denoted as $B_{\mathrm{MAX},i}(A_i)$ for all $i \in S_{\mathrm{RT}}$, without considering the satisfaction of condition (A.1). The existence of such an upper bound is based on the base station's total downlink power limitation, as explained in the following lemma.

Lemma 10. *There always exists an upper bound of a real-time user's $i \in S_{\mathrm{RT}}$ utility function parameter $\hat{b}_{\mathrm{RT},i}$, when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$, denoted as $B_{\mathrm{MAX},i}(A_i)$, where*

$$B_{\mathrm{MAX},i}(A_i) = \frac{1}{a}\left\{\ln\left(\frac{c_i}{1 + (A_i/aW(P_{\max} + A_i)) + c_i d_i} - 1\right) + \frac{aWP_{\max}}{R_i^{\max} A_i}\right\}, \quad \text{(D.1)}$$

due to the power limitations of the base station ($P_{\max}$) and the corresponding user's channel conditions per time slot (i.e., A_i). Specifically, if $\hat{b}_{\mathrm{RT},i} > B_{\mathrm{MAX},i}(A_i)$ then $P_i^* > P_{\max}$ where P_i^* is the solution of (18).

Proof. See Appendix I □

Moreover, we have already proven that $\lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})$ is a continuous and decreasing function of $\hat{b}_{\mathrm{RT},i}$ (in Proposition 1) and vice versa. Furthermore, the following lemma states that when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$,$\hat{b}_{\mathrm{RT},i}$ is also a concave up function of $\lambda_i^{\max}$.

Lemma 11. *When $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ an* RT *user's $i \in S_{\mathrm{RT}}$ utility function parameter $\hat{b}_{\mathrm{RT},i}$ is a concave up function of his parameter $\lambda_i^{\max}$, since $\partial^2\hat{b}_{\mathrm{RT},i}/\partial(\lambda_i^{\max})^2 > 0$.*

Proof. See Appendix J. □

Based on the previous lemmas, we can now formally define an upper bound for a real-time user's parameter $\hat{b}_{\mathrm{RT},i}$, with respect to (A.1), as follows. Let $L_{\max}$ be a positive constant such as when $\partial\hat{b}_{\mathrm{RT},i}/\partial\lambda_i^{\max}|_{\lambda_i^{\max}\to 0} \le -L_{\max}$, $\partial\lambda_i^{\max}(\hat{b}_{\mathrm{RT},i})/\partial\hat{b}_{\mathrm{RT},i}|_{\hat{b}_{\mathrm{RT},i}\to\infty} \simeq 0$. Moreover, for a user's parameter $\hat{b}_{\mathrm{RT},i}$, which is a function of their maximum willingness to pay $\lambda_i^{\max}$, we have proven the following properties:

(1) $\hat{b}_{\mathrm{RT},i} < B_{\mathrm{MAX},i}(A_i)$ (in Lemma 10).

(2) For $b_{\mathrm{RT},i} < \hat{b}_{\mathrm{RT},i} \le B_{\mathrm{MAX},i}(A_i)$:

(a) $\hat{b}_{\mathrm{RT},i}$ is a continuously decreasing function of parameter $\lambda_i^{\max}$ (in Proposition 1).

(b) $\hat{b}_{\mathrm{RT},i}$ is a concave up function of parameter λ_i^{max} (in Lemma 11).

Therefore, we can conclude that there always exists a $\hat{b}_{\mathrm{RT},i}$ for all $i \in S_{\mathrm{RT}}$, denoted as $b_{\max,i}$, such that

$$b_{\max,i} = \left\{\hat{b}_{\mathrm{RT},i} : \min\left(\left.\frac{\partial\hat{b}_{\mathrm{RT},i}}{\partial\lambda_i^{\max}}\right|_{\hat{b}_{\mathrm{RT},i}=b_{\max,i}} \le -L_{\max}, B_{\mathrm{MAX},i}(A_i)\right)\right\}, \quad \text{(D.2)}$$

which completes the proof.

E. Proof of Lemma 6

If P_i^0 is the inflection point of the sigmoid like function $U_i^{R_i^*}(P_i, \hat{b}_{\mathrm{RT},i}) \equiv U_i^{R_i^*}(P_i, b_i) \equiv U_i^{R_i^*}(P_i)$ (for simplicity in the presentation, in this proof, we denote $\hat{b}_{\mathrm{RT},i}$ as b_i) the following equation must be satisfied

$$\left.\frac{\partial^2 U_i^{R_i^*}(P_i)}{\partial {P_i}^2}\right|_{P_i=P_i^0} = 0. \quad \text{(E.1)}$$

Moreover, we can compute the second derivative of a users' utility function in accordance to the partial derivatives chain rule as follows, $\partial^2 U_i^{R_i^*}(P_i)/\partial {P_i}^2 = (\partial^2 U_i^{R_i^*}(\gamma_i)/\partial {\gamma_i}^2)(d\gamma_i/dP_i)^2 + (\partial U_i^{R_i^*}(\gamma_i)/\partial\gamma_i)(d^2\gamma_i/d{P_i}^2)$ and after mathematical manipulations, we can conclude that

$$\frac{\partial^2 U_i^{R_i^*}(P_i)}{\partial {P_i}^2} = \frac{N_i(P_T + A_i)}{(P_T - P_i + A_i)^3} \times \left[\frac{\partial^2 U_i^{R_i^*}(\gamma_i)}{\partial {\gamma_i}^2}(\gamma_i + N_i) + 2 \cdot \frac{\partial U_i^{R_i^*}(\gamma_i)}{\partial\gamma_i}\right], \quad \text{(E.2)}$$

where $N_i = W/R_i^{\max}$. Furthermore, with respect to (B.1) we can easily derive that

$$\frac{\partial^2 U_i^{R_i^*}(P_i)}{\partial {P_i}^2} = \frac{N_i(P_T + A_i)}{(P_T - P_i + A_i)^3} \times \left[\frac{\left(1 + e^{ab_i}\right)}{e^{ab_i}} a_i R_i^{\max} e^{a(b_i - \gamma_i)} \times \frac{e^{a(b_i-\gamma_i)}\left[a(\gamma_i + N_i) + 2\right] - \left[a(\gamma_i + N_i) - 2\right]}{\left(1 + e^{a(b_i-\gamma_i)}\right)^3} \right]. \tag{E.3}$$

Let us further define as

$$F_i(\gamma_i(P_i), b_i) = e^{a(b_i-\gamma_i)}\left[a(\gamma_i + N_i) + 2\right] - \left[a(\gamma_i + N_i) - 2\right], \tag{E.4}$$

where $F_i(\gamma_i(P_i), b_i)|_{P_i = P_i^0} = 0$ in order (E.1) to be satisfied when $P_i = P_i^0$. Function $F_i(\gamma_i(P_i), b_i)$ is an increasing function of variable $b_i \in (-\infty, \infty)$ since

$$\frac{\partial F_i(\gamma_i(P_i), b_i)}{\partial b_i} = a e^{a(b_i-\gamma_i)}\left[a(\gamma_i + N_i) + 2\right] > 0 \quad \forall b_i \in (-\infty, \infty), \tag{E.5}$$

and a decreasing function of variable $\gamma_i \in (-\infty, \infty)$ since

$$\frac{\partial F_i(\gamma_i(P_i), b_i)}{\partial \gamma_i} = (-a)\left\{e^{a(b_i-\gamma_i)}\left[a(\gamma_i + N_i) - 1\right] + 1\right\} < 0 \quad \forall \gamma_i \in (-\infty, \infty). \tag{E.6}$$

Now, let us consider the case where $b_i = b_i'$, $\gamma_i(P_i^{0\prime}) \equiv \gamma_i'$ and with respect to (E.1) and (E.4), $F_i(\gamma_i(P_i), b_i')|_{P_i = P_i^{0\prime}} = 0$ since $P_i^{0\prime}$ is the inflection point of user i utility function $U_i^{R_i^*}(P_i, b_i)$ when $b_i = b_i'$ (i.e., $U_i^{R_i^*}(P_i, b_i)$). Then, if we increase the value of user i parameter b_i from b_i' to b_i'', where $b_i' < b_i''$, then

$$F_i\left(\gamma_i(P_i), b''_i\right)\Big|_{P_i = P_i^{0\prime}} > 0, \tag{E.7}$$

since F_i is an increasing function of parameter b_i and the value of parameter γ_i is fixed (i.e., $\gamma_i(P_i^{0\prime}) \equiv \gamma_i'$). Furthermore, with respect to (E.1) and (E.4), there must also exist a value of parameter P_i^0 regarding the inflection point of function $U_i^{R_i^*}(P_i, b_i'')$, where $\partial^2 U_i^{R_i^*}(P_i, b_i'')/2P_i^2|_{P_i = P_i^{0\prime\prime}} = 0$, and hence

$$F_i(\gamma_i(P_i), b_i'')\big|_{P_i = P_i^{0\prime\prime}} = 0. \tag{E.8}$$

According to (E.7) and (E.8), and since F_i is a decreasing function of parameter γ_i, we can easily conclude that, $\gamma_i(P_i^{0\prime\prime}) > \gamma_i(P_i^{0\prime})$ when $b_i' < b_i''$. Finally, since γ_i is an increasing function of P_i we proved that if $b_i' < b_i''$, $P_i^{0\prime} < P_i^{0\prime\prime}$.

F. Proof of Lemma 7

It can be easily shown that

$$\text{when} \quad \hat{b}_{\mathrm{RT},i} = 0 \quad \text{then } \gamma_i^* = 1 \tag{F.1}$$

Moreover, we can see from (14) that when $\hat{b}_{\mathrm{RT},i}(t) = 0$ then

$$P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)\Big|_{\hat{b}_{\mathrm{RT},i}=0} = \frac{P_T + A_i}{N_i + 1} \tag{F.2}$$

Thus, if $\partial^2 U_i^{R_i^*}(P_i)/\partial {P_i}^2|_{P = P_T + A_i/N_i + 1} < 0$, then $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=0}$ belongs to the concave part of the sigmoid-like function $U_i^{R_i^*}(\gamma(P_i), P_i)$, and hence $P_i^0|_{\hat{b}_{\mathrm{RT},i}=0} < P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=0}$. Towards that and in accordance to (E.3) in Lemma 6, when $\hat{b}_{\mathrm{RT},i}(t) = 0$ we have

$$\left.\frac{\partial^2 U_i^{R_i^*}(P_i)}{\partial {P_i}^2}\right|_{\hat{b}_{\mathrm{RT},i}=0,\ P = P_T + A_i/N_i + 1} = \frac{2aR_i^{\max} e^{-\gamma_i a} e^{-2a}(N_i + 1)^3}{N_i^2 (P_T + A_i)^2 (1 + e^{-\gamma_i a})^3} \cdot \left\{e^{-\gamma_i a}\left[a(\gamma_i + N_i) + 2\right] - \left[a(\gamma_i + N_i) - 2\right]\right\}. \tag{F.3}$$

In order (F.3) to have negative values for $\gamma \geq 1$ and $\alpha_i \geq 1$ the following inequality must be asserted, $e^{-\gamma_i a}[a(\gamma_i + N_i) + 2] - [a(\gamma_i + N_i) - 2] \leq 0$, which is true.

G. Proof of Proposition 8

In accordance to Lemma 6 the inflection point P_i^0 of a user i sigmoidal like utility function is an increasing function of his utility parameter $\hat{b}_{\mathrm{RT},i}$ when $\hat{b}_{\mathrm{RT},i} \in [0, b_{\mathrm{RT},i}]$. Therefore, since when $\hat{b}_{\mathrm{RT},i} = 0$, then $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=0} > P_i^0|_{\hat{b}_{\mathrm{RT},i}=0}$ according to Lemma 7, and when $\hat{b}_{\mathrm{RT},i} = b_{\mathrm{RT},i}$, then $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=0} < P_i^0|_{\hat{b}_{\mathrm{RT},i}=0}$ according to the definition of a user's utility function, we can conclude that there always exists a value of a real-time user i utility function parameter $\hat{b}_{\mathrm{RT},i}$ where $\hat{b}_{\mathrm{RT},i} \in (0, b_{\mathrm{RT},i})$, (i.e., $\hat{b}_{\mathrm{RT},i}^*$) such that $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=\hat{b}_{\mathrm{RT},i}^*} = P_i^0|_{\hat{b}_{\mathrm{RT},i}=\hat{b}_{\mathrm{RT},i}^*}$, and hence when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}_{\mathrm{RT},i}^*)$, then $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}<\hat{b}_{\mathrm{RT},i}^*} > P_i^0|_{\hat{b}_{\mathrm{RT},i}<\hat{b}_{\mathrm{RT},i}^*}$.

H. Proof of Proposition 9

Let us define

$$h_i(P) = U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right) - P\frac{\partial U_i^{R_i^*}\left(P, \hat{b}_{\mathrm{RT},i}\right)}{\partial P} \equiv U_i^{R_i^{\max}}(P) - P\frac{\partial U_i^{R_i^*}(P)}{\partial P}. \tag{H.1}$$

Moreover, as it has been proved in [24], Lemma 6, a user i net utility $P_i(\lambda)$ is maximized only when their power allocation value is in the concave part of their utility function

(i.e., $P \in [P_i^0, P_{\max}]$). Therefore, we search for the solution of $h_i(P)|_{P=P_i^*} = 0$ only within the range $P_i^* \in [P_i^0, P_{\max}]$, as it further can be observe from (18). Furthermore, since from Proposition 2 when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^*_{\mathrm{RT},i})$, then $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}} > P_i^0|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}}$, the concave part of a user's utility function is for $\hat{b}_{\mathrm{RT},i} < \hat{b}^*_{\mathrm{RT},i}$ within the range of $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}=\hat{b}^*_{\mathrm{RT},i}}$ to $P_{\max}$ and if $\hat{b}_{\mathrm{RT},i} < \hat{b}^*_{\mathrm{RT},i}$, then

$$P_i^* \in \left[P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)\Big|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}}, P_{\max} \right]. \tag{H.2}$$

When $P_{\max} \geq P \geq P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}\leq\hat{b}^*_{\mathrm{RT},i}}$, then $\partial h_i(P)/\partial P = -P(\partial^2 U_i^{R_i^*}(P)/\partial P^2) \geq 0$ because $\partial^2 U_i^{R_i^*}(P)/\partial P^2 < 0$ for all $P > \mathrm{P}_i^0$ therefore is an increasing function of P and $P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}\leq\hat{b}^*_{\mathrm{RT},i}} \geq P_i^0|_{\hat{b}_{\mathrm{RT},i}\leq\hat{b}^*_{\mathrm{RT},i}}$. Furthermore, if we prove that there exists $\hat{b}^{**}_{\mathrm{RT},i} \in [0, \hat{b}^*_{\mathrm{RT},i})$ such that $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^{**}_{\mathrm{RT},i}}} > 0$ when $\hat{b}_{\mathrm{RT},i} \in [0, \hat{b}^{**}_{\mathrm{RT},i})$, we will have conclude the proof of the proposition, since P_i^* will not exist. After some algebra, we have from (H.1) that

$$\begin{aligned} &h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^*_{\mathrm{RT},i}}} \\ &= R_i^{\max} \cdot \frac{N_i + N_i e^{a(b-\gamma_i^*)} - N_i e^{-a\gamma_i^*} - N_i e^{a(b-2\gamma_i^*)} - N_i a\gamma_i^* e^{-a_i\gamma_i^*}}{N_i\left(1+e^{a(b-\gamma_i^*)}\right)^2} \\ &\quad \frac{-N_i a\gamma_i^* e^{a(b-\gamma_i^*)} - a(\gamma_i^*)^2 e^{-a\gamma_i^*} - a(\gamma_i^*)^2 e^{a(b-\gamma_i^*)}}{N_i\left(1+e^{a(b-\gamma_i^*)}\right)^2}, \end{aligned} \tag{H.3}$$

where we have denote $\hat{b}_{\mathrm{RT},i} \equiv b$ for presentation purposes. Since the denominator in (H.3) takes no negative values, we must examine the properties of the numerator. Therefore, let us define the following function

$$\begin{aligned} &H_i(P,b) \\ &= N_i + N_i e^{a(b-\gamma_i^*)} - N_i e^{-a\gamma_i^*} - N_i e^{a(b-2\gamma_i^*)} - N_i a\gamma_i^* e^{-a\gamma_i^*} \\ &\quad - N_i a\gamma_i^* e^{a(b-\gamma_i^*)} - a_i(\gamma_i^*)^2 e^{-a\gamma_i^*} - a(\gamma_i^*)^2 e^{a(b-\gamma_i^*)} \end{aligned} \tag{H.4}$$

Thus,

$$\begin{aligned} &\frac{\partial H_i(P,b)}{\partial b} \\ &= (1-a\gamma_i^*) N_i a e^{a(b-\gamma_i^*)} - N_i a e^{a(b-2\gamma_i^*)} - (a\gamma_i^*)^2 e^{a(b-\gamma_i^*)} < 0, \\ &H_i(P,b)|_{\substack{\hat{b}_{\mathrm{RT},i}=0 \\ \gamma_i^*=1}} = N_i + N_i\left(e^{-2a} - 2ae^{-a}\right) - 2ae^{-a} > 0. \end{aligned} \tag{H.5}$$

By (H.5), we can prove that since the numerator of $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{RT,i}<\hat{b}^{**}_{\mathrm{RT},i}}}$ is an decreasing function of b, $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}=0}} > 0$ and the denominator of $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^{**}_{\mathrm{RT},i}}}$ is always positive, then if we decrease a real time user's utility function parameter $\hat{b}_{\mathrm{RT},i}$ from $\hat{b}^*_{\mathrm{RT},i}$ to *0*, then there always exists a value for parameter $\hat{b}_{\mathrm{RT},i}$, namely, $\hat{b}^{**}_{\mathrm{RT},i}$, such that $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}=\hat{b}^{**}_{\mathrm{RT},i}}} = 0$ and when $\hat{b}_{\mathrm{RT},i} < \hat{b}^{**}_{\mathrm{RT},i}$ then $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^{**}_{\mathrm{RT},i}}} > 0$, therefore, when $\hat{b}_{\mathrm{RT},i} < \hat{b}^{**}_{\mathrm{RT},i}$ there is no $P_i^* \in [P_i^{\mathrm{LIM}}(\gamma_i^*, A_i)|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^{**}_{\mathrm{RT},i}}, P_{\max}]$ such $h_i(P)|_{P=P_i^{\mathrm{LIM}}(\gamma_i^*,A_i)^{(+)}|_{\hat{b}_{\mathrm{RT},i}<\hat{b}^{**}_{\mathrm{RT},i}}} > 0$.

I. Proof of Lemma 10

In accordance to (17) and (18), a real-time user's willingness to pay can be estimated when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ as

$$\lambda_i^{\max} = \left.\frac{\partial U_i^{R_i^*}(P)}{\partial P}\right|_{P=P_i^*}. \tag{I.1}$$

Moreover, we have that $U_i^{R_i^*}(P_i^*) - P_i^*\lambda_i^{\max} = 0$ and after some mathematical manipulations, we conclude

$$e^{-a(\gamma(P_i^*)-\hat{b}_{\mathrm{RT},i}(t))} = \frac{R_i^{\max} c_i}{P_i^*\lambda_i^{\max} + R_i^{\max} c_i d_i} - 1, \tag{I.2}$$

and finally that

$$\hat{b}_{\mathrm{RT},i} = \frac{1}{a}\left\{ \ln\left(\frac{R_i^{\max} c_i}{P_i^*\lambda_i^{\max} + R_i^{\max} c_i d_i} - 1 \right) + a\gamma(P_i^*) \right\}. \tag{I.3}$$

Moreover, we have that

$$\begin{aligned} \lambda_i^{\max} &= \left.\frac{U_i^{R_i^*}(P)}{P}\right|_{P=P_i^*} \\ &= \frac{Wa(P_{\max}+A_i)\left[R_i^{\max} c_i(1-d_i) - P_i^*\lambda_i^{\max}\right]}{(P_{\max}-P_i^*+A_i)^2 (R_i^{\max})^2 c_i} \\ &\quad \times \frac{(P_i^*\lambda_i^{\max} + R_i^{\max} c_i d_i)}{(P_{\max}-P_i^*+A_i)^2 (R_i^{\max})^2 c_i}, \end{aligned} \tag{I.4}$$

and hence after some algebra with respect to $U_i^{R_i^*}(P_i^*) - P_i^*\lambda_i^{\max} = 0$, we can see that

$$A'(P_i^*)^2 + B'(P_i^*) + C' = 0 \tag{I.5}$$

$$\begin{aligned} A' &= k_i(\lambda_i^{\max})^2 + \lambda_i^{\max} c_i (R_i^{\max})^2, \\ B' &= R_i^{\max} c_i \lambda_i^{\max}\left[-k_i(1-2d_i) - 2R_i^{\max}(P_{\max}+A_i)\right], \\ C' &= (R_i^{\max})^2 c_i\left[\lambda_i^{\max}(P_{\max}+A_i)^2 - c_i d_i(1-d_i)\right], \text{where} \\ k_i &= Wa_i(P_{\max}+A_i). \end{aligned}$$

In order (I.5) to have a real solution, P_i^*, when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$, and thus $P_i^* \leq P_{\max}$, the following two conditions must be satisfied (without loss of generality and for simplicity in the presentation since $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ in the following, we set $c_i = 1$, $d_i = 0$) :

(1) $\Delta = (B')^2 - 4A'C' \geq 0$ and after some algebra we can conclude that the following inequality has to be satisfied

$$\lambda_i^{\max} \leq \frac{aW + 4R_i^{\max}}{A_i + 4P_{\max}}. \tag{I.6}$$

(2) $P_i^* \leq P_{\max}$, and thus, in accordance to (I.5)

$$P_i^* = \frac{-B' \pm \sqrt{\Delta}}{2\mathrm{A}'} = R_i^{\max}(P_{\max} + \mathrm{A}_i) \times \frac{[aW + 2R_i^{\max}] \pm \sqrt{aW[aW + 4R_i^{\max} - 4\lambda_i^{\max}(P_{\max} + \mathrm{A}_i)]}}{2\left[k_i\lambda_i^{\max} + (R_i^{\max})^2\right]}. \tag{I.7}$$

Furthermore, since $P_i^* > 0$ and $aW \gg 4R_i^{\max} - 4\lambda_i^{\max}(P_{\max} + A_i)$, in order $P_i^* \leq P_{\max}$ the following inequality must be satisfied

$$\frac{R_i^{\max}(P_{\max} + A_i)(aW + R_i^{\max})}{k_i\lambda_i^{\max} + (R_i^{\max})^2} \leq P_{\max} \tag{I.8}$$

and after some algebra we conclude that

$$\lambda_i^{\max} \geq \frac{R_i^{\max}}{P_{\max}} + \frac{A_i R_i^{\max}}{aWP_{\max}(P_{\max} + A_i)}. \tag{I.9}$$

Finally, from (I.3) and (I.9), we can determine the upper bound of a real-time $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ user's $i \in S_{\mathrm{RT}}$ utility function parameter $\hat{b}_{\mathrm{RT},i}$ as follows:

$$\hat{b}_{\mathrm{RT},i} \leq \frac{1}{a}\left\{\ln\left(\frac{c_i}{1 + A_i/aW(P_{\max} + A_i) + c_i d_i} - 1\right) + \frac{aWP_{\max}}{R_i^{\max}A_i}\right\} \triangleq B_{\mathrm{MAX},i}(A_i). \tag{I.10}$$

J. Proof of Lemma 11

In accordance to (I.3) and (I.7) in Lemma 10, we have that when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ then

$$\hat{b}_{\mathrm{RT},i} = \frac{1}{a}\left\{\ln\left(\frac{\lambda_i^{\max}aW(P_{\max} + A_i) + (R_i^{\max})^2}{\lambda_i^{\max}(P_{\max} + A_i)[aW + 2R_i^{\max}]} - 1\right) + \frac{aW + R_i^{\max}}{\lambda_i^{\max}(P_{\max} + A_i) - R_i^{\max}}\right\}. \tag{J.1}$$

Finally, we can also easily compute $\partial^2\hat{b}_{\mathrm{RT},i}/\partial(\lambda_i^{\max})^2 = (P_{\max} + A_i)\{(P_{\max} + A_i)[2aW + 2R_i^{\max} - R_i^{\max}\lambda_i^{\max}] + (R_i^{\max})^2\}/[\lambda_i^{\max}(P_{\max} + A_i) - R_i^{\max}]^3 > 0$, since $W \gg R_i^{\max}$ and hence $\lambda_i^{\max} > R_i^{\max}/P_{\max} + A_i$ (since from (I.9) in Lemma 10, we proved that when $\hat{b}_{\mathrm{RT},i} \gg b_{\mathrm{RT},i}$ then $\lambda_i^{\max} \geq R_i^{\max}/P_{\max} > R_i^{\max}/(P_{\max} + A_i)$) which concludes the proof.

Acknowledgment

This work has been partially supported by EC FP7 EFIPSANS Project (INFSO-ICT-215549).

References

[1] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473–489, 1999.

[2] T. Lee, J. Lin, and Y. T. Su, "Downlink power control algorithms for cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 89–94, 1995.

[3] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, no. 7, pp. 70–77, 2000.

[4] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, pp. 976–985, April 2001.

[5] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, pp. 2415–2424, March 2005.

[6] F. Berggren and R. Jäntti, "Asymptotically fair transmission scheduling over fading channels," *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 326–336, 2004.

[7] Y. Liu, S. Gruhl, and E. W. Knightly, "WCFQ: an opportunistic wireless scheduler with statistical fairness bounds," *IEEE Transactions on Wireless Communications*, vol. 2, no. 5, pp. 1017–1028, 2003.

[8] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451–474, 2003.

[9] J. B. Kim and M. L. Honig, "Resource allocation for multiple classes of DS-CDMA traffic," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 2, pp. 506–519, 2000.

[10] F. Meshkati, A. J. Goldsmith, H. Vincent Poor, and S. C. Schwartz, "A game-theoretic approach to energy-efficient modulation in CDMA networks with delay constraints," in *Proceedings of the IEEE Radio and Wireless Symposium (RWS '07)*, pp. 11–14, January 2007.

[11] T. Harks, "Utility proportional fair bandwidth allocation: an optimization oriented approach," in *Proceedings of the 3rd International Workshop on QoS in Multiservice IP Networks*, vol. 3375 of *Lecture Notes in Computer Science*, pp. 61–74, Springer, Catania, Italy, 2005.

[12] P. Hande, S. Zhang, and M. Chiang, "Distributed rate allocation for inelastic flows," *IEEE/ACM Transactions on Networking*, vol. 15, no. 6, pp. 1240–1253, 2007.

[13] M. Dianati, X. Shen, and K. Naik, "Cooperative fair scheduling for the downlink of CDMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 4 I, pp. 1749–1760, 2007.

[14] X. Duan, Z. Niu, and J. Zheng, "Utility optimization and fairness guarantees for multimedia traffic in the downlink of DS-CDMA systems," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, pp. 940–944, December 2003.

[15] X. Duan, Z. Niu, and J. Zheng, "A dynamic utility-based radio resource management scheme for mobile multimedia DS-CDMA systems," in *Proceedings of the of IEEE Global Telecommunications Conference (GLOBECOM '02)*, Taipei, Taiwan, November 2002.

[16] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic power scheduling for dynamic multi-server wireless systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 6, Article ID 1638671, pp. 1506–1515, 2006.

[17] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic power scheduling for multi-server wireless systems with minimum performance constraints," in *Proceedings of the 23rd Conference of the IEEE Communications Society (INFOCOM '04)*, pp. 1067–1077, March 2004.

[18] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Downlink power allocation for multi-class wireless systems," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 854–867, 2005.

[19] S. Shakkotai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *American Mathematical Society Translations*, vol. 207, 2002.

[20] T. Kastrinogiannis and S. Papavassiliou, "Satisfying elastic short term fairness in high throughput wireless communication systems with multimedia services," in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 5064–5069, June 2007.

[21] T. Kastrinogiannis, S. Papavassiliou, K. Kastrinogiannis, and D. Soulios, "A utility-based resource allocation approach for the downlink in CDMA wireless networks with multimedia services," in *Proceedings of the 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, September 2007.

[22] T. Kastrinogiannis and S. Papavassiliou, "Utility based short-term throughput driven scheduling approach for efficient resource allocation in CDMA wireless networks," *Wireless Personal Communications*, vol. 52, no. 3, pp. 517–535, 2010.

[23] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Transactions on Networking*, vol. 11, no. 2, pp. 210–221, 2003.

[24] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Joint resource allocation and base-station assignment for the downlink in CDMA networks," *IEEE/ACM Transactions on Networking*, vol. 14, no. 1, pp. 1–14, 2006.

[25] F. Meshkati, H. V. Poor, and S. C. Schwartz, "A non-cooperative power control game in delay-constrained multiple-access networks," in *Proceedings of IEEE International Symp. on Information Theory (ISIT '05)*, Adelaide, Australia, September 2005.

[26] F. Meshkati, A. J. Goldsmith, H. Vincent Poor, and S. C. Schwartz, "A game-theoretic approach to energy-efficient modulation in CDMA networks with delay constraints," in *Proceedings of the IEEE Radio and Wireless Symposium (RWS '07)*, pp. 11–14, January 2007.

[27] G. Stuber, *Principles of Mobile Communication*, Kluwer Academic, Norwell, Mass, USA, 1996.

[28] J. R. Shewchuck, *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*, School of Computer Science, Carnegie Mellon University, Pittsburg, Pa, USA, 1.25 edition, 1994.