Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2010, Article ID 819729, 16 pages
doi:10.1155/2010/819729

Research Article

Scheduling for Improving System Capacity in Multiservice 3GPP LTE

Francisco Rafael Marques Lima,[1] Stefan Wänstedt,[2] Francisco Rodrigo Porto Cavalcanti,[1] and Walter Cruz Freitas Junior[1]

[1] *GTEL-Wireless Telecom Research Group, Department of Teleinformatics Engineering, Federal University of Ceará, Campus do Pici, 60455-760 Fortaleza, Brazil*
[2] *Ericsson AB, Laboratoriegränd 11, 97128 Luleå, Sweden*

Correspondence should be addressed to Francisco Rafael Marques Lima, rafaelm@gtel.ufc.br

Received 1 February 2010; Revised 3 May 2010; Accepted 26 June 2010

Academic Editor: Raymond Kwan

Copyright © 2010 Francisco Rafael Marques Lima et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We study the impact of scheduling algorithms on the provision of multiple services in the long term evolution (LTE) system. In order to measure how well the services are provided by the system, we use the definition of joint system capacity. In this context, we claim that scheduling strategies should consider the current satisfaction level of each service and the offered load to the system by each service. We propose a downlink-scheduling strategy according to these ideas named capacity-driven resource allocation (CRA). The CRA scheduler dynamically controls the resource sharing among flows of different services such as delay-sensitive and rate demanding ones. Moreover, CRA scheduler exploits the channel-quality knowledge to utilize the system resources efficiently. Simulation results in a multicell scenario show that the CRA scheduler is robust regarding channel quality knowledge and that it provides significant gains in joint system capacity in single and mixed service scenarios.

1. Introduction

The cellular networks have allowed us to communicate with people who are at the most remote places in the world through mobile phone calls. Furthermore, we are used to searching for information and entertainment by utilizing fixed broadband access in our homes. With the introduction of third generation (3G) networks, besides phone calls the mobile phones are now capable of accessing data services such as Web browsing and e-mail. However, the increased demand for new multimedia services, lower costs and improved quality of service (QoS) provision continuously stimulate the evolution of mobile communications. Consequently, 3rd generation partnership project (3GPP) and other standardization bodies have been working with the specification of the next steps in mobile communications: the long term evolution (LTE) and LTE-Advanced.

LTE will bring advantages for subscribers with new applications such as interactive TV and user-generated videos, and for operators with backward compatibility with legacy networks and simpler architecture. Among the main features of LTE we can highlight the utilization of orthogonal frequency division multiple access (OFDMA) as the radio access technology in the downlink and a pure packet-based all-internet protocol (IP) architecture.

OFDMA is a multiple access scheme based on orthogonal frequency division multiplexing (OFDM) digital modulation scheme where multiple user equipments (UEs) get assigned subcarriers or subsets of them in order to be served simultaneously. One of the advantages of an OFDMA-based system is the opportunity to benefit from frequency and multiuser diversities. Due to the frequency diversity, it is unlikely that all frequency resources in a link have the same channel quality. The multiuser diversity comes from the fact that UEs located at different positions within a cell experience almost independent channel qualities [1].

All-IP is a broad concept which means that the core network will be completely packet-switched and based on IP [2]. The main advantages of an All-IP architecture are reduced costs and efficient support to mass-market usage of

any IP-based service. On the other hand, a packet-switched network imposes some challenges on the provision of QoS guarantees for delay-sensitive services such as voice that was traditionally provided over circuit-switched networks and now must share the system resources with other services. In this multiservice scenario, system operators expect to achieve high system capacity by fulfilling the heterogeneous QoS requirements of the multiple flows in the system. Although in LTE a UE can bear multiple service flows, without loss of generality, here, only one service flow is considered per UE. Consequently, flow and UE are interchangeable throughout the text.

In order to exploit the advantages of OFDMA multiple access scheme and guarantee the QoS of different services with distinct traffic patterns and requirements, scheduling is of utmost importance in LTE. Scheduling algorithms are responsible for selecting which UEs will have access to the system resources and with which configuration. Therefore, in this paper, we deal with downlink scheduling algorithms for capacity maximization in multiservice scenarios. Our main contributions in this paper are:

(i) review a reasonable definition of system capacity suitable for multiservice scenarios and discuss how different scheduling strategies impact on it;

(ii) propose a scheduling strategy with the objective of improving the joint system capacity of the LTE system. Specifically, our proposal takes into account many requirements and limitations imposed by the LTE architecture; and

(iii) present a performance evaluation by using a detailed computational simulator in order to analyze the possible benefits of our proposed scheduler when applied to the LTE system. In the performance evaluation, we study some relevant aspects in multiservice scenarios.

The remainder of this paper is organized as follows. In Section 2, we present a brief overview about scheduling algorithms in the literature and contextualize our proposal. Section 3 is devoted to the problem formulation and presentation of the assumed system modeling. Specifically, in the problem formulation we define the joint system capacity and discuss how scheduling algorithms impact on it. After that, we present the proposed scheduling algorithm in Section 4. Then, in Section 5, we present a performance evaluation of our proposal by using a computational simulator that models the main aspects and restrictions of the LTE system. Finally, we provide the main conclusions and perspectives of this work.

2. Related Works

Studies in scheduling algorithms for wireless networks have acquired emphasis with the introduction of packet-switched single-carrier networks, such as high speed downlink packet access (HSDPA) and code-division multiple access 2000 evolution-data only (CDMA2000 EV-DO), where the system resources are no longer dedicated to the flows but shared among them. One of the first approaches followed by the community research was to generalize the concepts of queuing theory employed in wireline schedulers for the wireless setting [3–6]. As an example, in [7] the authors provide important bounds on queue backlog for different scheduling algorithms in OFDMA networks. Although the theoretical results in these works are conditioned to strong requirements regarding information availability and involved stochastic processes, the achieved results have paved the path towards the design of wireless schedulers. Important insights such as the relevance of the use of channel quality information, fairness in wireless environment and QoS-related aspects (e.g., delay) were obtained.

Scheduling algorithms in general are designed to deal with (RT) and/or non-real time (NRT) services. RT services are characterized by the short time response between the communicating parts. These services have strict requirements regarding packet delay and jitter. As an example of this kind of service we can mention voice over IP (VoIP). On the other hand, NRT services do not have tight requirements concerning packet delay although high packet delays are unacceptable. In fact, when transmitting NRT services the major constraint is the information integrity, that is, information loss is not tolerable. Therefore, applications of this type must have error-correction or recovery mechanisms. Web browsing is an example of an NRT service.

Contributions on scheduling algorithms can be categorized according to the multiple access method (single- and multicarrier networks) and provided services by the network (single- and multiservice scenarios). The interested reader can refer to [8–11] for single-carrier schedulers developed for NRT and RT services, respectively. In [12–15] the reader can find multicarrier schedulers for single-service scenarios. Our focus on this paper is on multicarrier schedulers for multiservice scenarios.

One of the first works that studied the QoS provision in multiservice scenarios was [16] that focused on single-carrier system. The proposed scheduler is based on proportional fair (PF) [8] with an additional weight that depends on packet delay for RT services and on a token bucket control algorithm for NRT services. Another work that also studies PF-like schedulers for multiservice scenarios is [17] where the flows are differentiated by service-dependent weights that are either fixed or dependent on packet delay.

Many works have been published focusing on the LTE system and multiservice scenarios and a complete survey is out of the scope of this section. The objective is to show some guidelines and how our work innovates compared to the state of art. In [18], the authors propose a downlink scheduling algorithm based on PF that takes into account in its formulation the per-flow amount of data awaiting transmission. In the scheduling process, flows belonging to a high priority service (delay-sensitive service) have an explicit priority over the others. In [19, 20], the authors consider a mixed service scenario; however, the main concern is with delay-sensitive services such as VoIP and video. In [21], the authors propose a scheduling algorithm that has an inter and intraservice part. In the former, the scheduler defines which service will have the flows scheduled. In the latter, the scheduler selects

the flows that will be scheduled from the service defined in the interservice part. The results shown in that article were spectral efficiency, fairness and average throughput. Although the scheduler presented in [21] can be applied in multiservice scenario, the authors neither provide results with mixed service scenarios nor consider delay-sensitive services.

The paper [22] highlights the importance of using channel- and buffer size/delay-aware schedulers in achieving high performance in multicarrier systems. In this paper, we do follow this approach by using information about channel-quality and packet delay in our proposed solution. Also in [22], the authors conclude that strict prioritization of a specific service (as it is done in some of the commented previous articles in this section) is not suitable for multiservice scenarios.

In summary, the main objectives of the multiservice schedulers designed for LTE described in this section are either to provide better spectral efficiency while keeping fairness among flows or protecting the QoS of high priority services such as delay-sensitive ones. In this paper, we propose a scheduling algorithm to improve the joint system capacity of the LTE system. This approach has not been followed by the previous articles on LTE to the best of our knowledge. In the next section, we formally define the joint system capacity.

3. Problem Formulation and System Modeling

In this section, we formally define the joint system capacity and relate it with other existing system-capacity definitions. Moreover, we present the main aspects about LTE that are relevant to this work.

3.1. Capacity Definitions. There are many ways to measure the capacity of wireless systems. A well-known definition of capacity is the one provided by Shannon which consists in the maximum achievable set of rates in multiple access channels with an arbitrarily small probability of error [23]. As this metric represents a bound in performance, in practice, the sum of the transmitted data rates (downlink) or aggregated data rate is used. Usually, this metric is also normalized by the system bandwidth and expressed in b/s/Hz.

However, with the increased availability of new services in wireless networks, the user perceived quality or QoS should also be included in the capacity measures. In this sense, the system capacity could be defined as the maximum aggregated data rate subject to the constraint that the average experienced quality of all flows in the system should be fulfilled according to a given target. As average experienced quality, we can mention the average delay of all transmitted packets or the average packet throughput, for example.

As in the wireless systems the perceived QoS can significantly vary among different flows, we believe that fairness related aspects should be taken into account when defining the capacity measure. This can be accomplished by the joint system capacity that is shown in Section 3.2.

3.2. Joint System Capacity. The joint system capacity used in this paper was first defined in [24] and used as performance metric in many papers including [17, 25–27].

Consider a multiservice system where the offered load is measured as the number of connected flows in the system. Let us identify the service types by indices that compose the service set Ψ. In a wireless system, flows start and finish their data sessions in a dynamic process. Considering that the system is in stable state where the statistics can be considered stationary, the mean number of connected flows in the system or offered load from service s is ρ_s. The total offered load to the system by all services is given by

$$\rho^{\text{total}} = \sum_{s \in \Psi} \rho_s. \tag{1}$$

The fraction of the total load offered by service s is given by

$$f_s = \frac{\rho_s}{\rho^{\text{total}}}. \tag{2}$$

We define the service mix, $\mathbf{f}$, as a vector composed of the elements f_s. In order to measure the quality provided to the flows of a specific service $s \in \Psi$ when a scheduling strategy named SCHED is used, we consider the user satisfaction ratio $q_s^{\text{SCHED}}(\rho^{\text{total}}, \mathbf{f})$. The user satisfaction ratio for a given service is defined as the fraction of flows from this service whose data sessions ended with the QoS requirements fulfilled. The user satisfaction ratio is a nonincreasing function of the total offered load. Furthermore, the user satisfaction ratio depends on the service mix since the load imposed by the flows of each service to the system is different due to distinct QoS demands.

We consider that the individual capacity for a service is the maximum total offered load in which the majority of the ended data sessions of this service achieve the QoS requirements. More specifically, the individual capacity of service $s \in \Psi$ (measured in number of connected flows) with the scheduling strategy SCHED is defined as

$$i_s^{\text{SCHED}}(\mathbf{f}) = \max\left(\rho^{\text{total}} \mid q_s^{\text{SCHED}}\left(\rho^{\text{total}}, \mathbf{f}\right) \geq Q_s^{\text{thres}}\right), \tag{3}$$

where Q_s^{thres} consists in the user satisfaction ratio threshold for service s, that is, the minimum acceptable user satisfaction ratio for service s defined by the system operator.

In a mixed service scenario, we have to take into account the QoS provided to the flows of all services. The joint system capacity is the maximum total offered load in which all provided services fulfill the user satisfaction ratio threshold. Therefore, the joint system capacity when a scheduling algorithm SCHED is used, $c^{\text{SCHED}}(\mathbf{f})$, is defined as

$$c^{\text{SCHED}}(\mathbf{f}) = \min\left(i_s^{\text{SCHED}}(\mathbf{f}), \forall s \in \Psi\right). \tag{4}$$

The joint system capacity is able to capture relevant aspects of multiservice wireless systems such as user and service specific quality requirements. The scheduling algorithm proposed in this paper in Section 4 aims at improving the joint system capacity of LTE.

3.3. Impact of Scheduling Strategies on the Joint System Capacity. Once we have defined the joint system capacity, we will discuss the effects of some scheduling strategies on the system capacity. Consider, for example, a scheduling algorithm that gives explicit priority to a given service such as the approaches of some of the papers described in Section 2. In this case, the flows of the service with higher priority tend to be scheduled more often than the flows of other services. As a consequence, the individual capacity of the prioritized service will be higher than the other services. However, as defined in (4), the joint system capacity is limited by the service with lower individual capacity.

In Figure 1, we illustrate this issue in an example with a two-service case. In this figure, "Prio" is a scheduling strategy that tends to allocate most of the system resources to the flows of the second service that has high priority. On the other hand, "Balanced" is an example of another scheduling algorithm that is capable of balancing the QoS experienced by the flows of the different services in order to improve the joint system capacity. As can be seen in this figure, the joint system capacity with the first scheduler is i_1^{Prio} and for the second scheduler is i_1^{Balanced} with $i_1^{\text{Balanced}} > i_1^{\text{Prio}}$. In order to improve the joint system capacity, the scheduler "Balanced" had to degrade the individual capacity of service 2 in order to improve the individual capacity of service 1 (direction of the arrows in the figure). Although the degradation of the QoS provided to the flows of a specific service can seem odd at first, this is supported by the system operator's point-of-view. When offering wireless services the system operator is interested in fulfilling each per-service minimum user satisfaction ratio threshold (Q_s^{thres}) and, therefore, quality overprovision for a specific service will not bring additional benefit for the system operator.

In general, a condition to the joint system capacity maximization is that all provided services achieve the same individual system capacity [24]. As a conclusion, scheduling strategies with explicit priority for flows of a specific service are not able to maximize the joint system capacity.

In a stationary environment where aspects such as traffic mix proportions and channel conditions are kept statistically unchanged, a suitable set of weights of the PF-based schedulers such as in [17] could be found in order to provide QoS balancing among services. However, in real networks, these aspects are rather unpredictable and time-variant. Consequently, a fixed set of weights would not lead to an improved joint system capacity in this scenario.

In order to improve the joint system capacity, the scheduling strategy should perform a controlled resource sharing among flows of different services. Aspects such as the load of each service in the system (traffic mix proportion, $\mathbf{f}$) and the satisfaction level of each flow should be addressed. These are the main ideas of our proposed scheduling algorithm described in Section 4.

3.4. System Modeling. In this section, we point out the relevant aspects of the LTE system that impact on our work. For a complete description of LTE the interested reader can refer to the 3GPP's standards and the articles [28–30], for example. In Sections 3.4.1 and 3.4.2 we present the physical

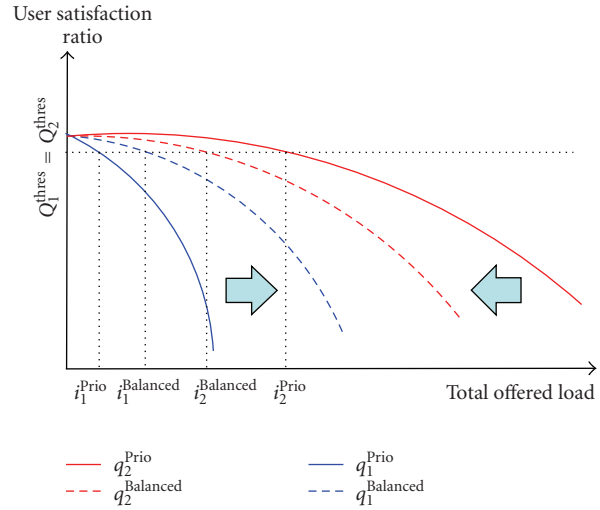


Figure 1: Illustration of the effects of scheduling strategies in the joint system capacity.

[31] and medium access control (MAC) [32] layers and in Section 3.4.3 we show the services types and user satisfaction model.

3.4.1. Physical Layer. The time domain structure of LTE is composed of radio frames of 10 ms. Each radio frame has 10 equally-sized subframes of length 1 ms. Subframes, in turn, consist of two slots of length 0.5 ms. The scheduling takes place in a subframe basis.

The default subcarrier spacing is 15 kHz and all subcarriers are grouped in sets of 12 subcarriers. A resource block in LTE is defined as a two-dimensional grid with 12 subcarriers in frequency and 0.5 ms in time that corresponds to 6 or 7 OFDM symbols depending on cyclic prefix length. The Resource Unit (RU) in the system is composed of two resource blocks concatenated in the time domain, that is, 12 subcarriers and 1 ms.

The resources are utilized by physical channels and signals. Physical channels are utilized for transmission of data and/or control information from the MAC layer. The physical signals are used to support physical-layer functionality and do not carry any information from the MAC layer [31].

Among the physical channels, we emphasize the function of physical downlink shared channel (PDSCH) and physical downlink control channel (PDCCH). The former is utilized for transmission of data traffic while the latter is used for downlink layer 1/layer 2 control signaling. Specifically, PDCCH is used to carry uplink scheduling grants and downlink scheduling assignments, such as PDSCH resource indication, transport format, hybrid automatic repeat request (HARQ) information and transport block size. Depending on the time-variant PDCCH capacity, different number of UEs can be scheduled in a given transmission time interval (TTI). In this work, we consider that there is a fixed limit in the number of scheduled UEs in a TTI. Although this is an important aspect of LTE system, this issue has not been

considered in the majority of the works about scheduling in the literature.

In this study, we consider that the allocated power per RU is fixed and is equal to the ratio between the available power and the number of RUs.

3.4.2. Medium Access Control. When a connection (or bearer) is established between the UE and the LTE core network a QoS class identifier (QCI) is specified. This defines whether the bearer is guaranteed bit-rate or not, target delay and loss requirements, for example. The enhanced node B (eNB) translates the QCI attributes into requirements for the air interface. The scheduling should allocate resources according to these requirements.

The HARQ comprises a number of processes where each process uses a simple stop-and-wait protocol. HARQ for downlink, that is the focus in this study, is asynchronous and adaptive. By asynchronous we mean that the scheduler has the freedom to choose the subframe for retransmission dynamically. In adaptive HARQ, the scheduler can use a different resource configuration for retransmission compared to the previous (re)transmission. In case the data is a retransmission of a previously stored data, the received data is soft combined with the data stored in the soft buffer.

3.4.3. Service Types and Model for User Satisfaction. The concept of user satisfaction is very important when interpreting the system performance. There are many parameters to consider such as the service type, technical parameters (e.g., delay and throughput), and even economical issues (such as the price to use the wireless service) [33, 34]. However, in this study we consider only technical aspects concerning the perceived quality by the end user.

In this work, we consider two classes of services that have been used as reference in the research community: RT and NRT services that were described in Section 2. Note that we can directly map the QCIs attributes to the service classes used in this study [35].

In the area of RT services, quite extensive models have been obtained that relate the frame erasure rate (FER) with perceived quality [36]. Therefore, we consider that a RT flow, j, is satisfied when

$$\gamma_j[k] = \frac{n_j^{\text{lost}}[k]}{n_j^{\text{lost}}[k] + n_j^{\text{succ}}[k]} \leq \gamma_j^{\text{req}}, \tag{5}$$

where $\gamma_j[k]$ is the accumulated FER for flow j at TTI k, and γ_j^{req} is the FER requirement of flow j. The variables $n_j^{\text{succ}}[k]$ and $n_j^{\text{lost}}[k]$ are the number of successfully transmitted and lost packets from flow j at TTI k since the session initialization, respectively. If the FER of a flow is higher than the requirement this flow is considered unsatisfied.

The satisfaction model for NRT flows is based on the average data rate and is suitable for services with bursty traffic [37]. A flow j that belongs to an NRT service is satisfied when

$$\bar{r}_j[k] = \frac{l_j[k]}{\left(t_j[k] \cdot a\right)} \geq \bar{r}_j^{\text{req}}, \tag{6}$$

where $\bar{r}_j[k]$ is the average data rate of flow j at TTI k computed from the session initialization, and $\bar{r}_j^{\text{req}}$ is the average data rate requirement of flow j. The variable $l_j[k]$ is the number of correctly transmitted bits from flow j at TTI k since the session initialization, $t_j[k]$ is the total active time of flow j at TTI k since the session initialization and a is the duration of one TTI. By active time we mean the total time that a flow has data to transmit. If the average data rate of a flow is lower than the requirement this flow is considered unsatisfied.

4. Capacity-Driven Resource Allocation (CRA)

Based on the previous discussion about joint system capacity and the main aspects/restrictions of the LTE architecture, we propose the scheduling algorithm CRA. The main objective of CRA is to improve the joint system capacity of the LTE system in multiservice scenarios.

4.1. CRA Overview. We have followed a common approach when designing schedulers for multicarrier systems that is to split the scheduling functionality into two parts: Resource Allocation and Resource Assignment. The Resource Allocation part is responsible for defining which flows will be scheduled and determining their required data rate at the current TTI, while the Resource Assignment part defines which resources will be assigned to the selected flows in the Resource Assignment part.

In Figure 2, we illustrate the main building blocks of the CRA scheduler. In the Resource Allocation part, the CRA scheduler firstly builds priority lists for each existing service. These services can be either NRT or RT. In the priority list, the flows that belong to a specific service are ordered according to a priority based on the satisfaction level. The prioritization intends to give transmission opportunities to the flows that can be easily satisfied. Besides, for each flow CRA also calculates the required data rate that this flow needs to transmit in the current TTI.

The last step of Resource Allocation part consists in the selection, according to the load imposed by each service, of the flows that will receive RUs in the Resource Assignment part. As the flows are arranged in priority lists, the task is to define how many flows of each service will be chosen to get resources in the Resource Assignment part. In this last part, the selected flows get assigned RUs in a channel opportunistic fashion.

4.2. CRA: Resource Allocation Part. As described previously, for each flow we determine the required data rate to transmit at the current TTI. For an NRT flow, this data rate is calculated as follows:

$$\Delta r_j[k] = \bar{r}_j^{\text{req}} \cdot \left(t_j[k] + 1\right) - \bar{r}_j[k-1] \cdot t_j[k-1]. \tag{7}$$

This required data rate represents the data rate that should be allocated to an unsatisfied flow in order to this flow stay satisfied even if it does not have transmission opportunities in the next TTI (see Appendix A for demonstration). Note that if a flow is already satisfied the required rate would

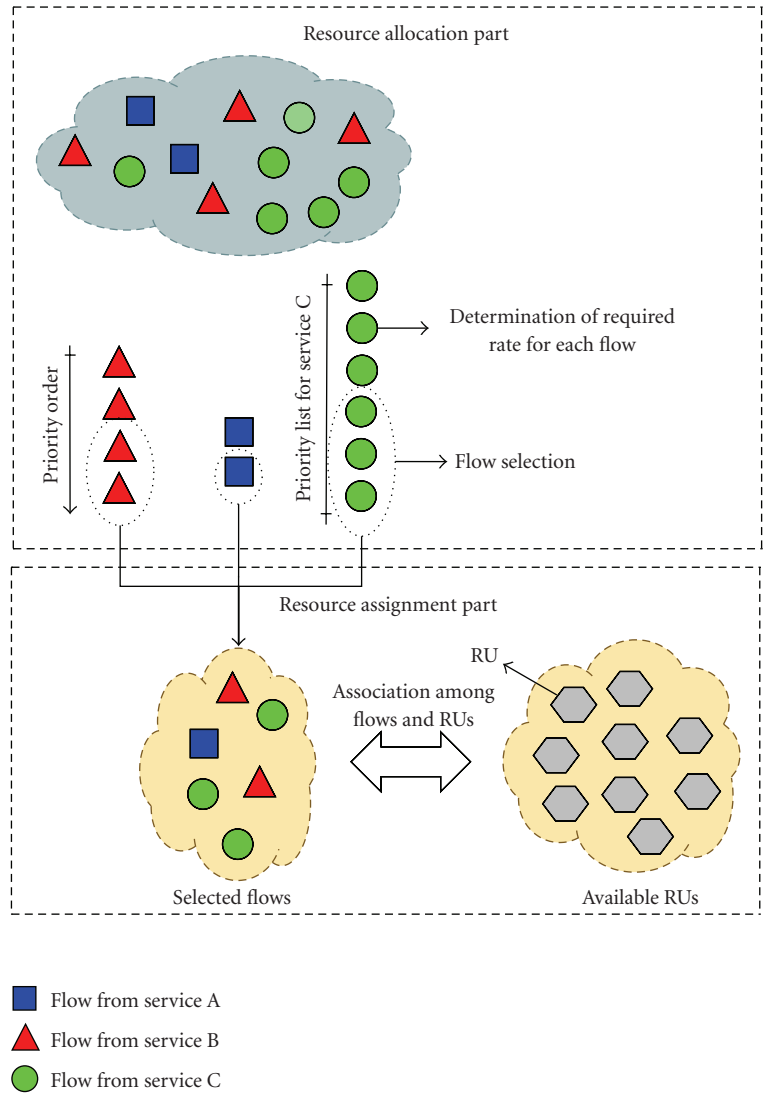


Figure 2: Building blocks of CRA. Illustration of the Resource Allocation and Resource Assignment parts.

be negative. Therefore the modulus (or absolute value) of this required rate can be seen as the "distance" that the average data rate of a flow ($\bar{r}_j[k]$) is from the average data rate requirement ($\bar{r}_j^{\text{req}}$).

For RT flows we define the required data rate as follows:

$$\Delta r_j[k] = \frac{b_j^{\text{oldest}}[k]}{a}, \tag{8}$$

where $b_j^{\text{oldest}}[k]$ represents the number of bits of the oldest packet in the transmit buffer of the eNB corresponding to the flow j at TTI k, that is, the packet that is waiting for transmission for the longer period of time. The choice of this required rate for RT flows is based on the fact that, in general, the upper protocol layers from this service split the data to transmit in small packets with short delay requirements. In this way, when these flows get transmit opportunities the complete packet should be transmitted in order to avoid packet discard.

Once the data rate demanded by each flow is determined, the next step is to build priority lists for each service. The priority lists are built according to the service classes. The main idea is to prioritize the flows that can be easily satisfied.

In the priority list for any service, the flows with retransmissions have the highest priority. Concerning NRT services, the flows that are currently unsatisfied have precedence over the ones that are satisfied. The prioritization is the opposite for RT flows. The reason for this strategy is the fact that users of NRT services tolerate temporary QoS fluctuations during the data session if in average the QoS is fulfilled. On the other hand, due to the quick response characteristic of RT services, a temporary oscillation in the experienced QoS compromises the whole session.

Besides the prioritization based on the satisfaction status, we assign a priority to each flow, p_j, in order to sort the flows within the set of satisfied and unsatisfied flows. An illustration of the process to build the priority list is shown in Figure 3. The priority for NRT flows is given by

$$p_j = \frac{\overline{\alpha_j}[k]}{\left\|\Delta r_j[k]\right\|}, \tag{9}$$

where $\overline{\alpha_j}[k]$ is the ratio between the transmit data rate of a flow j at TTI k in case it gets assigned all available RUs and the number of available RUs, and the operator $\|\cdot\|$ returns the absolute value. Therefore, within the group of unsatisfied flows, the flow that is in good channel condition and requires lower data rate to become satisfied than the other unsatisfied flows is more prioritized. This is a reasonable strategy in order to increase the number of satisfied flows in the system. In the group of satisfied NRT flows, the ones that are in good channel conditions and that are near to the unsatisfaction have precedence over the other satisfied flows. In Appendix B, we show that this prioritization is an optimum strategy to solve the problem of maximizing the number of satisfied flows when the flow's requirements are represented in terms of number of RUs.

Before defining the priority of an RT flow we define the concept of "distance" to the requirement for RT flows as in the following

$$w_j[k] = \begin{cases} \left\lfloor \dfrac{\left(n_j^{\text{succ}}[k] + n_j^{\text{lost}}[k]\right)\cdot \gamma_j^{\text{req}} - n_j^{\text{lost}}[k]}{1-\gamma_j^{\text{req}}} \right\rfloor, & \\ \qquad\qquad \text{if } \gamma_j[k] \le \gamma_j^{\text{req}} & \\ \left\lceil \dfrac{n_j^{\text{lost}}[k] - \left(n_j^{\text{succ}}[k] + n_j^{\text{lost}}[k]\right)\cdot \gamma_j^{\text{req}}}{\gamma_j^{\text{req}}} \right\rceil, & \\ \qquad\qquad \text{otherwise,} & \end{cases} \tag{10}$$

where the operators $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ return the first integer greater than or equal to and the first integer lower than or equal to a real number, respectively.

The variable $w_j[k]$ represents how many consecutive packets an unsatisfied flow j at TTI k has to successfully transmit to become satisfied (with $\gamma_j[k] \le \gamma_j^{\text{req}}$). For a satisfied flow j, $w_j[k]$ means the maximum number of packets that this flow can lose successively and still be satisfied. In other words, $w_j[k]$ defines how close (or far) the FER of a given flow is from the required FER. See Appendix B for the demonstration of (10).

In this way, the priority for an RT flow is given by

$$p_j = \frac{1}{\left(d_j^{\text{req}} - d_j^{\text{oldest}}\right)\cdot\left(w_j + 1\right)}, \tag{11}$$

where d_j^{req} and d_j^{oldest} are the packet delay requirement and current packet delay of the oldest packet of flow j, respectively. In Figure 4, we plot p_j in function of d_j^{oldest} and $w_j[k]$ for RT flows considering d_j^{req} equal to 80 ms. From this figure we can see that the flows that have packets with delays close to their deadlines and shorter "distance" between the current and required FER than the other flows are more prioritized. The rule is the same for satisfied and unsatisfied flows. By prioritizing flows with packet delays close to the deadline we take advantage of the fact that RT applications tolerate a certain packet delay without compromising the end user perceived quality. Consequently, more RT connected flows can be multiplexed in order to increase capacity.

The last part of Resource Allocation is the definition of which flows of each priority list will be chosen to get resources in the Resource Assignment part. Consider that set Ω_s is composed of the active flows (flows that have data to transmit) from service $s \in \Psi$. The number of flows that will be selected is constrained by the conditions below

$$y_s \approx \mu \cdot \frac{|\Omega_s|}{\sum_{p\in\Psi}\left|\Omega_p\right|}, \qquad \sum_{s\in\Psi} y_s \le \mu, \tag{12}$$

where y_s represents the number of selected flows from service s to be scheduled, $|\cdot|$ represents the cardinality of a set and μ is the maximum possible number of scheduled terminals in a TTI.

The first part in (12) states that the number of selected flows must be almost proportional to the number of active flows from each existing service in the cell. The second part has the objective of guaranteeing that the number of scheduled flows is equal to or smaller than the maximum number of terminals that can be scheduled in a TTI (as presented in Section 3.4.1). The main objective of these constraints is to provide a better resource distribution among the different services even when the offered load per service is unbalanced.

4.3. CRA: Resource Assignment Part. The main idea in the Resource Assignment is to distribute the RUs in a fair and opportunistic way among the selected flows in the Resource Allocation part. The Resource Assignment part is executed in phases. In each phase, all the flows get assigned one RU. However, the flow that will choose its RU first is the one that has the RU in better channel conditions among all other flows. The process continues until all flows get assigned one RU in the current phase. The flows compete for resources until receiving the number of RUs to transmit with the required data rate, $\Delta r_j[k]$, defined in the Resource Allocation part. In this case, they are taken out of the process. If all the selected flows have RUs enough to fulfill the required data rate, the remaining RUs are equally divided among all selected flows.

5. Performance Results

In this section, we present a case study where we apply the proposed scheduling strategy in the LTE system in a multiservice scenario composed of VoIP (RT service) and Web (NRT service). We firstly present the simulation setup

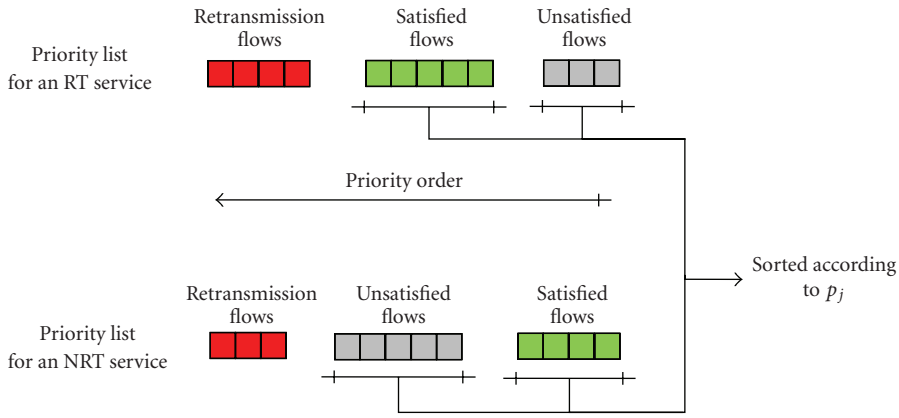


Figure 3: Priority lists for RT and NRT services. The prioritization takes into account the retransmission status, satisfaction level and a per-flow priority.

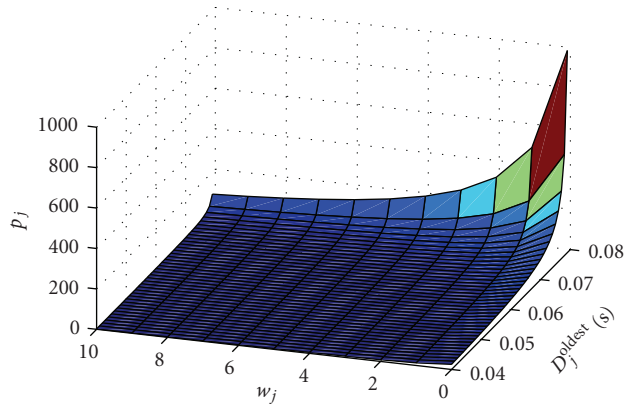


Figure 4: Illustration of p_j for RT flows considering D_j^{req} equal to 80 ms.

where we provide the simulation parameters and define reference schedulers used for comparison. The results are divided into two parts: sensitivity analysis and performance evaluation in single and mixed service scenarios.

5.1. Simulation Setup. The results presented in this section are drawn from a dynamic system-level simulator that models the LTE system according to 3GPP specifications detailed in Section 3.4. The simulator model includes multiple cells, intercell interference and propagation phenomena such as path loss, shadowing and fast fading. Moreover, the main aspects related to both radio interface layers and upper layers (Transport control protocol (TCP), user datagram protocol (UDP), IP and applications) were taken into account in the simulation models.

The Web traffic model is characterized as request-response traffic: a client, that utilizes a mobile station and is located in the radio network, requests one or more web pages within a session, that is, hypertext transfer protocol (HTTP) requests. The server generates and returns the web pages. Once a web page is received at the client, this user reads the web page for some seconds and then requests another web page. We consider that the Web pages have a fixed length and that the reading time follows an exponential distribution. The Web service can be mapped to QCI1, for example.

The VoIP packets are generated by a speech coder that mimics the adaptive multirate (AMR) codec. This coder produces voice frames every 20 ms during speech periods and small packets, named silence insertion description (SID) packets, to simulate background noise during silence periods. In this study we consider that there is a conversation between two clients, one out of the LTE network utilizing a computer (client A) and another client utilizing a UE (client B) in the radio access network. As the downlink is focused, the performance is measured in the client located in the radio network utilizing the UE. The model for conversation has three states: Client A talking, client B talking and mutual silence. The model switches between this three states with a time period drawn from exponential distributions. The VoIP service can be mapped to QCI8, for example. The main simulation parameters are shown in Table 1.

The reference schedulers used in the simulation are: delay scheduler (DS) and maximum rate (MR). All these reference schedulers give transmission opportunities to the flow with higher priority. The selected flow gets assigned its RUs in better channel condition until the data rate necessary to transmit all backlogged data is achieved. If the selected flow does not utilize all available RUs, the next more prioritized flow is selected to get resources and so on. Note that there is the limitation in the maximum number of scheduled flows as described in Section 3.4.1.

The difference between the reference schedulers is the prioritization. DS assigns the best RUs of the flow (VoIP or Web in this case) whose headline radio link control (RLC) SDU has the current greatest delay. Therefore, DS scheduler is a channel- and QoS-aware scheduler. The MR scheduler chooses the flow that can transmit more information bits when using the available bandwidth (better channel condition). In this way, as reference schedulers we have a strategy that takes into account channel and QoS aspects (DS) and another that is only channel-aware (MR).

Table 1: Main simulation parameters considered in this work. The parameters are classified in general parameters of LTE, propagation, deployment, and service-specific ones.

Parameter	Value	Unit
General		
Bandwidth	3	MHz
Carrier frequency	2	GHz
Number of RUs	15	—
Total cell power	20	W
Transport network packet delay (including Internet and Core Network (CN))	14	ms
PDCCH capacity (number of scheduling grants per TTI)	5 (static modeling)	—
Number of HARQ processes	8	—
Maximum number of HARQ retransmissions	10	—
VoIP/Web user satisfaction ratio thresholds	95/90	%
Propagation		
Path gain at 1 meter distance	−29.03	dB
Path gain per dB distance	−3.52	dB
Shadowing standard deviation	8	dB
Antenna type	SCM 3GPP [38]	—
Deployment		
Number of eNB/cells per eNB	3/3	—
Number of antennas in the UEs/cell	2/1	—
Cell radius	500	m
Frequency reuse	full	—
UE speed	3	km/h
Voip Flows		
RLC service data unit SDU discard period	80	ms
Mean talk period time	5	s
Voice activity factor	0.5	—
Frame size	264	bits
Frame period	20	ms
Maximum end-to-end VoIP frame delay	140	ms
SID frame size	39	bits
Required FER	1	%
Web Flows		
Web page size (fixed)	10,000	bytes
Mean reading time	1.5	s
Average data rate requirement	128	kbps

In the next sections, we will evaluate how these two strategies impacts on the joint system capacity in a multiservice scenario.

5.2. Sensitivity Analysis. CRA is a channel-aware scheduler, that is, it relies on channel quality measurements. Consequently, before analyzing simulation results in mixed service scenarios we devote this section to the sensitivity analysis of our proposal regarding channel state reporting. In Figure 5, we show the user satisfaction ratio versus the offered load (measured in number of flows in a cell) in the Web-only scenario when the channel state reporting period is increased. The CRA scheduler utilizes channel quality measurements for Web flows in two parts: in the priority calculation of Web flows and in the resource assignment.

In the priority calculation, an average channel quality measurement is considered. Therefore, it is expected that the dependence of this part on channel measurements is not critical. In the Resource Assignment, a per-RU channel quality measurement is utilized in order to assign the best resources to the UEs. Consequently, this part must be more affected by higher channel reporting periods. However, as it can be seen in Figure 5, the degradation in capacity of CRA considering a user satisfaction ratio threshold of 90% is of only 2 UEs, which represents a capacity loss of approximately 2% when the channel reporting period is changed from 10 ms to 25 ms. DS and MR also suffer a capacity loss of approximately 2 UEs considering the same user satisfaction ratio threshold. This similar performance among the schedulers points to a degradation in the link

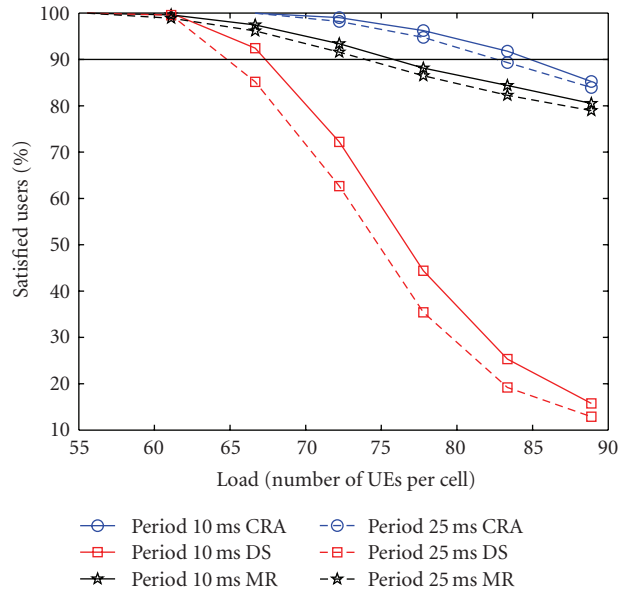


Figure 5: User satisfaction ratio in the Web-only scenario with variable channel state reporting periods for CRA, DS and MR schedulers.

adaptation as the main reason for the capacity loss. Link adaptation, which is common for any scheduler, also utilizes channel quality measurements.

In Figure 6, we show the user satisfaction ratio with different channel state reporting periods in a VoIP-only scenario. The channel quality measurements are only utilized in the Resource Assignment part of CRA when VoIP flows are concerned, that is a common part for any service type. The degradation observed in the VoIP service is similar to the one observed in the Web-only scenario in Figure 5. Changing the channel reporting period from 10 ms to 25 ms caused similar capacity decreases of approximately 2% for CRA and 3% for DS considering the user satisfaction ratio threshold of 95%. The capacity loss for MR scheduler cannot be measured in the user satisfaction ratio threshold of 95% because of its poor performance in the simulated load range. However, the degradation in user satisfaction ratio is similar to the visualized in the other schedulers. The performance loss was mainly caused by a degradation of link adaptation, as in the previous scenario.

5.3. Performance Evaluation. In Figure 7, we show the user satisfaction ratio for the simulated schedulers in three mixed service scenarios: 25% of VoIP and 75% of Web flows (v25w75), 50% of VoIP and 50% of Web flows (v50w50) and 75% of VoIP and 25% of Web flows (v75w25). Note that these proportions are related to the number of connected flows in the system and not active flows. In Figure 7(a), we show the user satisfaction ratio for VoIP service and in Figure 7(b) the user satisfaction ratio for Web service. In the axis of abscissas, we present the system load measured in number of flows in a cell (the sum of VoIP and Web flows).

In general, we can observe that the user satisfaction ratio for the simulated schedulers and services is improved when the percentage of VoIP flows in the system is increased. The reason for this behavior is the fact that a VoIP flow demands lower data rates and consequently resources of LTE system. In order to measure the individual service capacity, as defined in (3), the user satisfaction ratio thresholds of 95% for VoIP and 90% for Web should be considered as depicted in the figures.

In all simulated mixed service scenarios with the MR scheduler, the user satisfaction ratio for Web service is better than the VoIP one. This good performance for Web service is due to the burst nature of Web traffic and a more flexible QoS requirement based on average data rate. Because of the burst traffic pattern of Web, during the inactive periods of the flows in better channel conditions the MR scheduler can select the other flows. This works as a statistical time multiplexing mechanism that is not present in low-rate and regular VoIP traffic. Furthermore, when MR schedules VoIP flows the scheduling process is limited by the maximum number of scheduled UEs instead of the number of available RUs. This leads to a low resource usage.

When DS scheduler is concerned, we can observe in Figure 7 that the Web service experiences a lower individual capacity than VoIP in the simulated mixed service scenarios. Consequently, the former limits the joint system capacity of DS scheduler. The packet delay is one important measurement when scheduling RT services because it directly affects the FER that determines the user satisfaction for RT flows. Moreover, another reason to prioritize flows with high headline packet delays is that these flows usually have more than one buffered packet to transmit. The transport block size in LTE utilizing one RU can, depending on the modulation order and code rate, be greater than one RLC SDU that is mapped one to one with VoIP frames. As a result, scheduling flows with high headline packet delays increases the efficiency by reducing the protocol layer overheads and padding rate per sent VoIP packet [39]. This explains the good performance of DS for VoIP service. However, scheduling Web traffic based on packet delay usually grants flows with transmission opportunities that have too much buffered data due to poor channel conditions.

Despite the differences among the reference schedulers they all have in common one aspect: they do not consider satisfaction status and load per service in their formulation. Although the joint system capacity in the simulated mixed service scenarios with the CRA scheduler is limited by the Web service, the user satisfaction ratios provided to the VoIP and Web services are improved. In Figure 7, we can observe that the CRA scheduler achieves higher individual capacities for VoIP and Web services in the user satisfaction ratio thresholds of 95% and 90%, respectively. The scheduling in CRA is accomplished in a more intelligent way by considering the number of active flows from each service and the current QoS conditions of each flow. The Resource Assignment part in the CRA scheduler was designed to give equal opportunities to each selected flow get its resources in better channel quality. In the following, we show that the CRA scheduler provides improved joint system capacity also in the single-service scenarios.

In Figure 8, we show the system capacity region that is built from the user satisfaction ratio curves for several

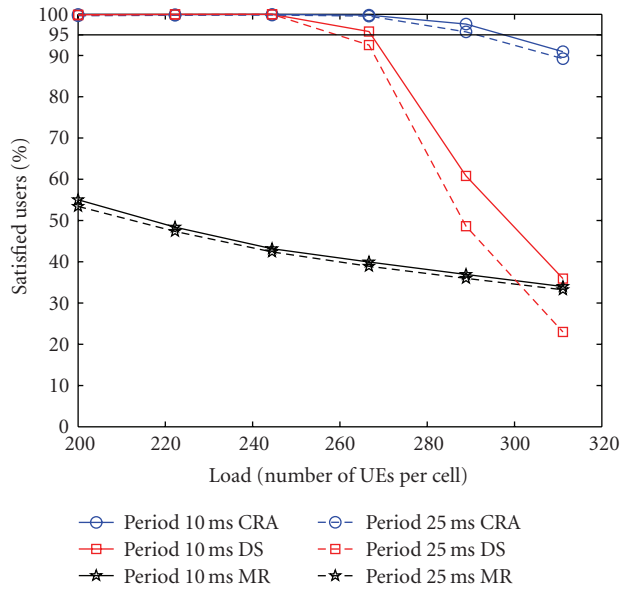
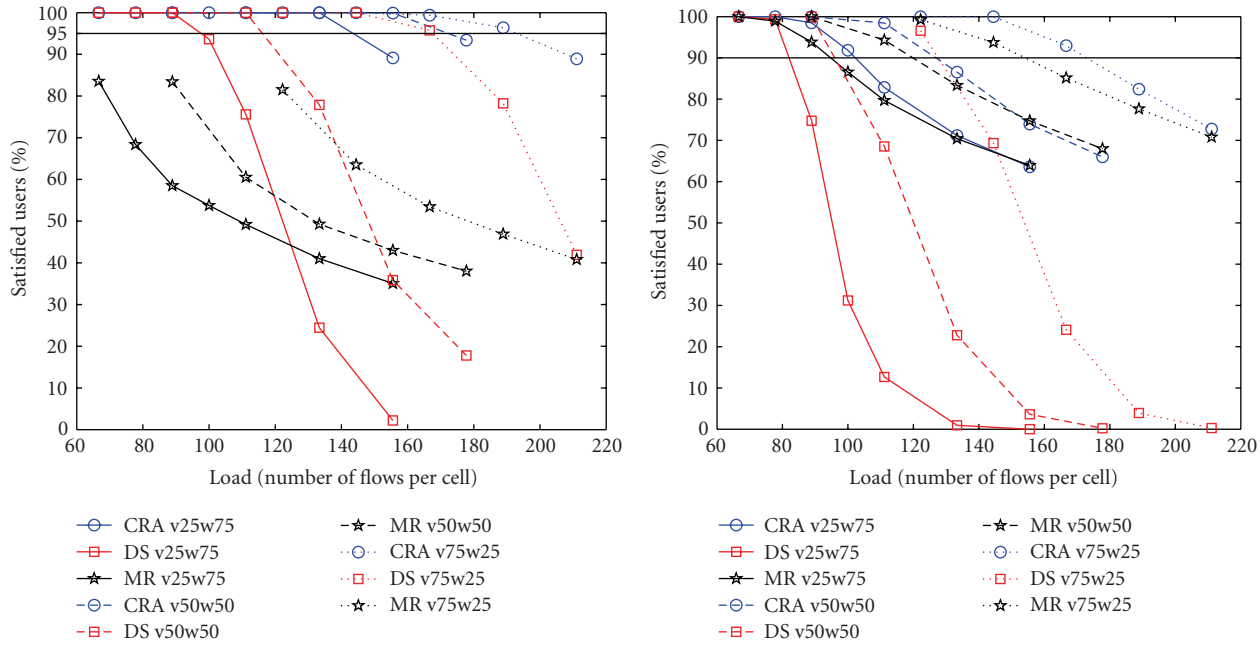


Figure 6: User satisfaction ratio in the VoIP-only scenario with variable channel state reporting periods for CRA, DS and MR schedulers.



(a) User satisfaction ratio for VoIP service

(b) User satisfaction ratio for Web service

Figure 7: User satisfaction ratio in the mixed service scenarios 25% VoIP and 75% Web, 50% VoIP and 50% Web and 75% VoIP and 25% Web for CRA, DS and MR schedulers.

traffic mixes. This is an important result since it allows us to assess the performance of schedulers when the network is submitted to different traffic mixes. The MR scheduler is not included in this figure because its user satisfaction ratio for VoIP service was lower than the VoIP satisfaction threshold for the simulated offered load.

By increasing the user satisfaction ratio for each service, the CRA scheduler also provides greater capacity region. The gain of the CRA scheduler over DS in joint capacity can be quantified by the larger area (below the curves) in the system capacity region. The gain of the CRA scheduler over DS is of approximately 37%. This gain may not be completed realized in real deployments since there are many other aspects in real networks that are not feasible to model even in a detailed simulator such as the one used in this study. However, we expect that even in real deployments our proposal is able to overperform the reference schedulers concerning the joint system capacity. Therefore, we believe that the main ideas of

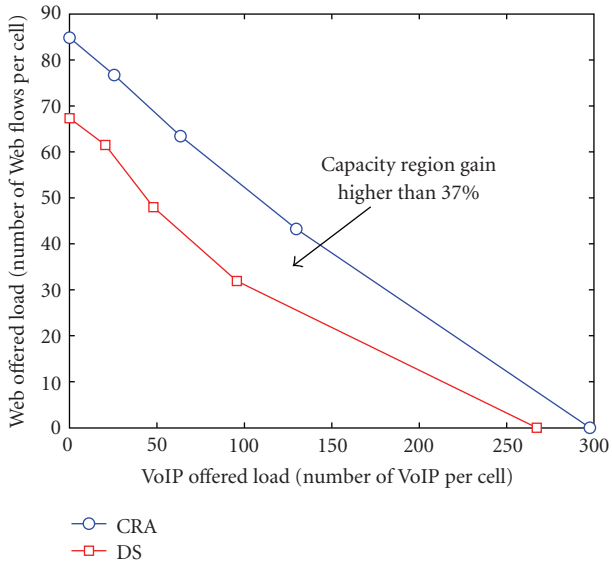


Figure 8: Capacity region.

our proposed scheduler should be considered in the design of scheduling algorithms for the LTE system in order to improve the joint system capacity in multiservice scenarios.

6. Conclusions

The provision of sustainable quality of service (QoS) to different flows with heterogeneous requisites is an important issue in multiservice wireless networks such as long term evolution (LTE). In this scenario, the system capacity should not only take into account the flows of a specific service. In fact, the system capacity should measure how well all the services are provided by the system. Therefore, in order to improve the system capacity, scheduling algorithms are of utmost importance due to their intrinsic task of distributing resources to the flows of different services in a short transmission time interval (TTI).

Some approaches in scheduling design found in the literature are not capable of improving the system capacity in a general setting. The main reason for that is the disregard of two aspects: the offered load to the system by each service and satisfaction level of the connected flows. In this context we proposed a downlink scheduling algorithm named capacity-driven resource allocation (CRA) whose main objective is to improve the system capacity in multiservice wireless networks. In addition, CRA was designed according to the main restrictions and characteristics of the LTE architecture.

By the performance results of a sensitivity analysis regarding the periodicity of channel state report, we concluded that the performance of the CRA scheduler is as dependent of this metric as reference schedulers are. Morevoer, in mixed service scenarios the simulation results have shown that the CRA scheduler is able to provide an overall gain in joint capacity higher than 37% over reference schedulers.

Although the presented performance evaluation considers a scenario with two services, the ideas in our proposed scheduler are general enough to deal with many concurrent services in the same packet-switched network. In this multiservice scenario, CRA is capable of providing a better QoS balancing independently of time varying aspects such as channel conditions and service mix proportions.

Appendices

A. Demonstration of (7)

Without loss of generality consider an NRT flow j that is currently unsatisfied at TTI k, that is, $\bar{r}_j[k] < \bar{r}_j^{\text{req}}$. Therefore, we would like to know which data rate should be allocated at the current TTI k in order to the flow j become satisfied for the next λ TTIs even if no resource is assigned to it, that is, $\bar{r}_j[k+\lambda] = \bar{r}_j^{\text{req}}$. Expanding $\bar{r}_j[k+\lambda]$ we have

$$\begin{aligned}\bar{r}_j[k+\lambda] &= \frac{l_j[k+\lambda]}{a \cdot t_j[k+\lambda]} = \frac{l_j[k-1]+\mu}{a \cdot \left(t_j[k]+\lambda\right)} \\ &= \frac{a \cdot \left(\bar{r}_j[k-1]\right) \cdot \left(t_j[k-1]\right)+\mu}{a \cdot \left(t_j[k]+\lambda\right)},\end{aligned} \tag{A.1}$$

where μ is the amount of data (bits) that should be transmitted at TTI k.

Therefore, we have to find μ by solving the following equation

$$\frac{a \cdot \left(\bar{r}_j[k-1]\right) \cdot \left(t_j[k-1]\right)+\mu}{a \cdot \left(t_j[k]+\lambda\right)} = \bar{r}_j^{\text{req}}. \tag{A.2}$$

The solution of this equation is

$$\mu = a \cdot \left(t_j[k]+\lambda\right) \cdot \bar{r}_j^{\text{req}} - a \cdot \bar{r}_j[k-1] \cdot t_j[k-1]. \tag{A.3}$$

Finally, the current data rate that should be allocated to the flow j at TTI k in order to this flow become satisfied for the next λ TTIs even if no resource is assigned to it, $\Delta r_j[k]$, is given by

$$\Delta r_j[k] = \frac{\mu}{a} = \left(t_j[k]+\lambda\right) \cdot \bar{r}_j^{\text{req}} - \bar{r}_j[k-1] \cdot t_j[k-1]. \tag{A.4}$$

Note that in (7), we considered $\lambda = 1$. The choice of the parameter λ depends on the satisfaction level that we intend to provide to the scheduled flows. When the required rate ($\Delta r_j[k]$) is calculated using high values of λ the scheduled flows will stay satisfied for several TTIs. On the other hand, the required rate increases with the parameter λ. Therefore, the scheduled flows will get more resources in order to fulfill their required rate. In this way, few flows could be scheduled simultaneously. Therefore, we have chosen $\lambda = 1$ in order to allow for better resource distribution among flows.

B. Maximization of Satisfied Flows

In this appendix, we show the intuition behind the flow prioritization in (9). The objective of this flow prioritization is to decrease the number of unsatisfied NRT flows (or increase the number of satisfied flows) at the current TTI. The task to be performed in the Resource Allocation part is to define the flows that should get system resources at the Resource Assignment part of the CRA algorithm. In the Resource Assignment part, the scheduled flows are associated with the available resources.

The flows' requirements are represented by the required data rate $\Delta r_j[k]$. As different flows experience different channel states in each individual RU, the task of defining the scheduled flows in order to achieve the stated objective is a hard problem to solve. Indeed, our problem would become easier to solve if the flows' requirements were directly represented in number of required RUs instead of data rate.

The representation of flow's requirement in number of RUs can be performed by the following relation

$$\overline{m}_j = \left\lceil \frac{\Delta r_j[k]}{\overline{\alpha}_j} \right\rceil, \tag{B.1}$$

where $\overline{m}_j$ is the required number of resources demanded by flow j at the current TTI considering the average transmit data rate among all RUs and $\lceil \cdot \rceil$ returns the first integer greater than or equal to a real number. Obviously, some information about individual channel quality of each RU is lost when we consider this representation. However, as point out in [40] that used a similar approach to solve a subproblem, this procedure is justified by the low performance degradation and reduced computational complexity when an opportunistic resource assignment is performed.

According to this, we can formulate our problem as

$$\begin{aligned} \max_{\mathbf{m}} \quad & \sum_{j=1}^{J} u\left(m_j - \overline{m}_j\right) \\ \text{subject to} \quad & \sum_{j=1}^{J} m_j \le N, \\ & m_j \in \mathbb{N}, \end{aligned} \tag{B.2}$$

where m_j is the number of resources allocated to flow j, $\mathbf{m}$ is a $J \times 1$ column vector composed of m_j, N is the number of available RUs, J is the number of active flows at the current TTI and, finally, $u(\cdot)$ is the step function that assumes the value 0 when its argument is negative and 1 otherwise. In summary, this is an optimization problem to define the number of RUs that should be assigned to the flows so as to maximize the number of satisfied flows (with $m_j \ge \overline{m}_j$) constrained to the limited number of RU. Problem (B.2) is a combinatorial optimization problem with potentially multiple optimum solutions.

Algorithm 1 is able to find one of the optimum solutions. Basically, the resources are allocated to the flows with lower required number of resources in order to become satisfied. In order to prove the optimality of this algorithm consider a solution given by this algorithm, $\mathbf{m}^* = [m_1^* \; m_2^* \; \cdots \; m_J^*]$, that leads to L satisfied flows with $L < J$. Suppose also that there is another solution, $\mathbf{m}' = [m_1' m_2' \; \cdots \; m_J']$, with H satisfied flows where $L < H < J$. For the sake of this proof, consider that $\mathbf{o} = [o_1 \; o_2 \; \cdots \; o_J]$ is the vector with the index of the flows sorted in the ascending order of $\overline{m}_j$ according to the line 2 of Algorithm 1, that is, $\overline{m}_{o_1} \le \overline{m}_{o_2} \le \cdots \le \overline{m}_{o_J}$. Consider also that $\mathbf{x} = [x_1 \; x_2 \; \cdots \; x_H]$ is a vector of length H with the indices (disposed in any order) of the satisfied flows given by solution $\mathbf{m}'$.

In order to make the solutions $\mathbf{m}^*$ and $\mathbf{m}'$ lead to L and H satisfied flows, respectively, the following constraints should be fulfilled

$$\begin{aligned} m_{o_1}^* &\ge \overline{m}_{o_1} & m_{x_1}' &\ge \overline{m}_{x_1} \\ &\vdots & &\vdots \\ m_{o_L}^* &\ge \overline{m}_{o_L} & m_{x_H}^* &\ge \overline{m}_{x_H} \\ \sum_{j=o_1}^{o_L} m_j^* &\le N & \sum_{j=x_1}^{x_H} m_j' &\le N, \end{aligned} \tag{B.3}$$

By adding these constraints we have that

$$N \ge \sum_{j=o_1}^{o_L} m_j^* \ge \sum_{j=o_1}^{o_L} \overline{m}_j, \tag{B.4}$$

$$N \ge \sum_{j=x_1}^{x_H} m_j' \ge \sum_{j=x_1}^{x_H} \overline{m}_j. \tag{B.5}$$

Particularly, the following equation derived from (B.5) should also hold

$$N \ge \sum_{j=x_1}^{x_{L+1}} m_j' \ge \sum_{j=x_1}^{x_{L+1}} \overline{m}_j. \tag{B.6}$$

As in Algorithm 1, the flows with lower required number of RUs are selected firstly we have that

$$\begin{aligned} \overline{m}_{x_1} &\ge \overline{m}_{o_1}, \\ \overline{m}_{x_1} + \overline{m}_{x_2} &\ge \overline{m}_{o_1} + \overline{m}_{o_2}, \\ &\cdots \qquad\qquad \cdots, \\ \overline{m}_{x_1} + \overline{m}_{x_2} + \cdots + \overline{m}_{x_L} &\ge \overline{m}_{o_1} + \overline{m}_{o_2} + \cdots + \overline{m}_{o_L}, \\ \overline{m}_{x_1} + \overline{m}_{x_2} + \cdots + \overline{m}_{x_L} + \overline{m}_{x_{L+1}} &\ge \overline{m}_{o_1} + \overline{m}_{o_2} \\ &\quad + \cdots + \overline{m}_{o_L} + \overline{m}_{o_{L+1}}. \end{aligned} \tag{B.7}$$

However, as the solution found by the Algorithm 1 was able to satisfy only L flows we have that

$$\overline{m}_{o_1} + \overline{m}_{o_2} + \cdots + \overline{m}_{o_L} + \overline{m}_{o_{L+1}} > N, \tag{B.8}$$

and consequently by the last constraint in (B.7) we have that

$$\overline{m}_{x_1} + \overline{m}_{x_2} + \cdots + \overline{m}_{x_L} + \overline{m}_{x_{L+1}} > N. \tag{B.9}$$

(1) $m_j \leftarrow 0\ \forall j$
(2) $\mathbf{o} \leftarrow \text{sort}_{\forall j}(\overline{m}_j)$ {Sort in the ascending order}
(3) $\theta \leftarrow N$
(4) $i \leftarrow 1$
(5) **while** $(\theta > 0)$ AND $(i \leq J)$ **do**
(6) $\quad j^* \leftarrow o_i$
(7) $\quad$ **If** $(\theta - \overline{m}_{j^*}) > 0$ **then**
(8) $\quad\quad m_{j^*} \leftarrow \overline{m}_{j^*}$
(9) $\quad\quad \theta \leftarrow \theta - \overline{m}_{j^*}$
(10) $\quad$ **else**
(11) $\quad\quad m_{j^*} \leftarrow \theta$
(12) $\quad\quad \theta \leftarrow 0$
(13) $\quad$ **end if**
(14) $\quad i \leftarrow i + 1$
(15) **end while**
(16) **If** $\theta > 0$ **then**
(17) $\quad i \leftarrow 1$
(18) $\quad$ **while** $\theta > 0$ **do**
(19) $\quad\quad j^* \leftarrow o_i$
(20) $\quad\quad m_{j^*} \leftarrow m_{j^*} + 1$
(21) $\quad\quad \theta \leftarrow \theta - 1$
(22) $\quad\quad i \leftarrow i + 1$
(23) $\quad\quad$ **if** $i > J$ **then**
(24) $\quad\quad\quad i \leftarrow 1$
(25) $\quad\quad$ **end if**
(26) $\quad$ **end while**
(27) **end if**
(28) Output of the algorithm: **m**

Algorithm 1: Algorithm to solve the problem (B.2).

As a consequence, (B.9) contradicts (B.5), (B.6) and (B.3). Therefore, the solution given by Algorithm 1 provides an optimum solution of problem (B.2).

As selecting the flows with higher priority p_j (given by (9)) is similar to the steps of Algorithm 1 where the flows with lower $\overline{m}_j$ are chosen first, we can conclude that the employed prioritization is a reasonable strategy to increase the number of satisfied flows.

B. Demonstration of (10)

Consider an RT flow j that is currently unsatisfied at TTI k, that is, $\gamma_j[k] > \gamma_j^{\text{req}}$. Therefore, we would like to know how many packets, ν, this flow has to successfully transmit in a row so as to become satisfied, that is, $\gamma_j[k'] = \gamma_j^{\text{req}}$ where $k' > k$. In this way, the FER at TTI k' is given by

$$\gamma_j[k'] = \frac{n_j^{\text{lost}}[k']}{n_j^{\text{lost}}[k'] + n_j^{\text{succ}}[k']} = \frac{n_j^{\text{lost}}[k]}{n_j^{\text{lost}}[k] + \left(n_j^{\text{succ}}[k] + \nu\right)}. \tag{C.1}$$

Therefore, we have to solve the following equation

$$\frac{n_j^{\text{lost}}[k]}{n_j^{\text{lost}}[k] + \left(n_j^{\text{succ}}[k] + \nu\right)} = \gamma_j^{\text{req}}. \tag{C.2}$$

The solution of this equation is

$$\nu = \frac{n_j^{\text{lost}}[k] - \left(n_j^{\text{succ}}[k] + n_j^{\text{lost}}[k]\right) \cdot \gamma_j^{\text{req}}}{\gamma_j^{\text{req}}}. \tag{C.3}$$

If the RT flow j is currently satisfied, that is, $\gamma_j[k] \leq \gamma_j^{\text{req}}$, we need to know the maximum number of packets, ϵ, that this flow can lose successively and still be satisfied, that is, $\gamma_j[k'] = \gamma_j^{\text{req}}$ where $k' > k$.

The FER at TTI k' is given by

$$\gamma_j[k'] = \frac{n_j^{\text{lost}}[k']}{n_j^{\text{lost}}[k'] + n_j^{\text{succ}}[k']} = \frac{\left(n_j^{\text{lost}}[k] + \epsilon\right)}{\left(n_j^{\text{lost}}[k] + \epsilon\right) + n_j^{\text{succ}}[k]}. \tag{C.4}$$

The equation to be solved is as follows

$$\frac{\left(n_j^{\text{lost}}[k] + \epsilon\right)}{\left(n_j^{\text{lost}}[k] + \epsilon\right) + n_j^{\text{succ}}[k]} = \gamma_j^{\text{req}}. \tag{C.5}$$

This equation is solved by setting ϵ as follows

$$\epsilon = \frac{\left(n_j^{\text{succ}}[k] + n_j^{\text{lost}}[k]\right) \cdot \gamma_j^{\text{req}} - n_j^{\text{lost}}[k]}{1 - \gamma_j^{\text{req}}}. \tag{C.6}$$

Acknowledgment

This work was supported by the Research and Development Center, Ericsson Telecomunicações S.A., Brazil, under EDB/UFC.22 Technical Cooperation Contract.

References

[1] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of the IEEE International Conference on Communications (ICC '95)*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[2] 3GPP, "All-IP Network (AIPN) feasibility study," Tech. Rep. TR22.978 V8.0.0—Release 8, 3rd Generation Partnership Project, December 2008.

[3] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: issues and approaches," *IEEE Personal Communications*, vol. 6, no. 1, pp. 44–53, 1999.

[4] Y. Cao and V. O. K. Li, "Efficient algorithms for broadband space-time coded wireless communication," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 76–87.

[5] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76–83, 2002.

[6] S. Shakkottai and T. S. Rappaport, "Research challenges in wireless networks: a technical overview," in *Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications*, vol. 1, pp. 12–18, October 2002.

[7] G. Wunder and C. Zhou, "Queueing analysis for the OFDMA downlink: throughput regions, delay and exponential backlog bounds," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 871–881, 2009.

[8] F. P. Kelly, A. K. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[9] S. Wänstedt, F. Rui, M. Ericsson, and M. Nordberg, "Providing reliableand efficient VoIP over cellular networks," in *Proceedings of the Future Telecommunications Conference*, October 2005.

[10] B. Wang, K. I. Pedersen, T. E. Kolding, and P. E. Mogensen, "Performance of VoIP on HSDPA," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 4, pp. 2335–2339, Stockholm, Sweden, June 2005.

[11] P. Hosein, "Scheduling of VoIP traffic over a time-shared wireless packet data channel," in *Proceedings of the 7th IEEE International Conference on Personal Wireless Communications (ICPWC '05)*, pp. 38–41, January 2005.

[12] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in UTRA long term evolution downlink," in *Proceedings of the 3rd IEEE International Symposium of Wireless Pervasive Computing (ISWPC '08)*, pp. 308–313, May 2008.

[13] A. Pokhariyal, K. I. Pedersen, G. Monghal, et al., "HARQ aware frequency domain packet scheduler with different degrees of fairness for the UTRAN long term evolution," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 2761–2765, April 2007.

[14] G. Mongha, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, "QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2532–2536, May 2008.

[15] R. Kwan, C. Leung, and J. Zhang, "Multiuser scheduling on the downlink of an lte cellular system," *Research Letters in Communication*, vol. 2008, Article ID 323048, 4 pages, 2008.

[16] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–153, 2001.

[17] A. R. Braga, E. B. Rodrigues, and F. R. P. Cavalcanti, "Packet scheduling for VoIP over HSDPA in mixed traffic scenarios," in *Proceedings of the 17th IEEE International Symposium Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, pp. 1–5, Helsink, Finland, September 2006.

[18] B. Chen, H. Hu, B. Wang, and H. Wang, "A novel multi-service scheduling scheme for E-UTRA," in *Proceedings of the 3rd IEEE/IFIP International Conference in Central Asia on Internet (ICI '07)*, pp. 1–5, September 2007.

[19] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "MAC scheduling scheme for VoIP traffic service in 3G LTE," in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1441–1445, Baltimore, Md, USA, October 2007.

[20] S. Wang, Y. Gao, X. Gu, H. Tian, and P. Zhang, "Packet scheduling for multimedia traffics in downlink multi-user OFDM systems," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '06)*, pp. 1–4, September 2006.

[21] M. Gidlund and J.-C. Laneri, "Scheduling algorithms for 3GPP long-term evolution systems: from a quality of service perspective," in *Proceedings of the 10th IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '08)*, pp. 114–117, August 2008.

[22] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 510617, 18 pages, 2009.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1991.

[24] A. Furuskär, *Radio resource sharing and bearer service allocation for multi-bearer service, multi-access wireless networks*, Ph.D. thesis, Royal Institute of Technology (KTH), Radio Communication Systems, April 2003.

[25] M. Ericson and S. Wänstedt, "Mixed traffic HSDPA scheduling—impact on VoIP capacity," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1282–1286, Dublin, Ireland, April 2007.

[26] A. Furuskär and J. Zander, "Multiservice allocation for multiaccess wireless systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 174–183, 2005.

[27] E. B. Rodrigues, F. R. P. Cavalcanti, and S. Wänstedt, "QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 100–107, 2008.

[28] D. M. Sacristán, J. F. Monserrat, J. Cabrejas-Penuelas, D. Calabuig, S. Garrigas, and N. Cardona, "On the way towards fourth-generation mobile: 3GPP LTE and LTE-advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 354089, 10 pages, 2009.

[29] S. Parkvall, E. Dahlman, A. Furuskär et al., "LTE-advanced—evolving LTE towards IMT-advanced," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 1–5, September 2008.

[30] D. Astély, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, 2009.

[31] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); long term evolution (LTE) physical layer; general description," Tech. Rep. TS 36.201 V8.2.0—Release 8, 3rd Generation Partnership Project, December 2008.

[32] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," Tech. Rep. TS36.321 V8.4.0—Release 8, 3rd Generation Partnership Project, December 2008.

[33] N. Enderlé and X. Lagrange, "User satisfaction models and scheduling algorithms for packet-switched services in UMTS," in *Proceedings of the 57th IEEE Semiannual Vehicular Technology Conference (VTC '03)*, vol. 3, pp. 1704–1709, Jeju, South Korea, April 2003.

[34] L. Badia, M. Boaretto, and M. Zorzi, "A users' satisfaction driven scheduling strategy for wireless multimedia QoS," in *Proceedings of the Quality of Future Internet Services (QoFIS '03)*, vol. 2811, pp. 203–213, Stockholm, Sweden, October 2003.

[35] 3GPP, "Technical Specification Group Services and System Aspects; policy and charging control architecture," Tech. Rep. TS 23.203V9.4.0—Release 9, 3rd Generation Partnership Project, March 2010.

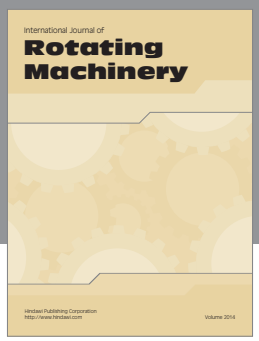
[36] 3GPP, "TSG-SA Codec Working Group: Mandatory speech codec; AMR speech codec; interface to Iu and Uu," Tech. Rep. TS26.102, 3rd Generation Partnership Project, 1999.

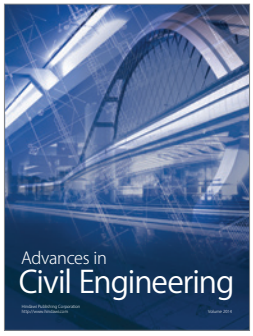
[37] 3GPP, "Physical layer aspects for evolved universal terrestrial radio access (UTRA)," Tech. Rep. TR 25.814 V7.1.0—Release 7, 3rd Generation Partnership Project, September 2006.

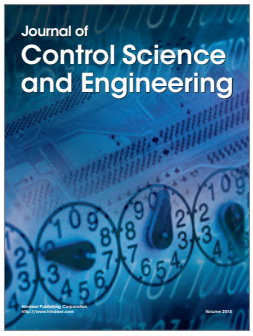
[38] 3GPP, "Spatial channel model for multiple input multiple output (MIMO) simulations," Tech. Rep. TR 25.996 V8.0.0, 3rd Generation Partnership Project, December 2008.

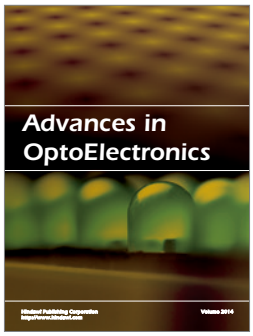
[39] F. Persson, "Voice over IP realized for the 3GPP long term evolution," in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1436–1440, October 2007.

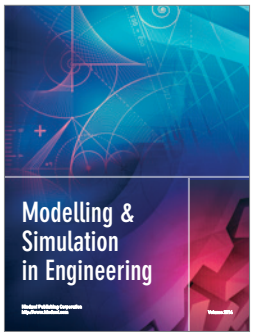
[40] D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1150–1158, 2003.

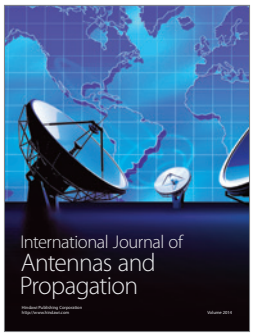


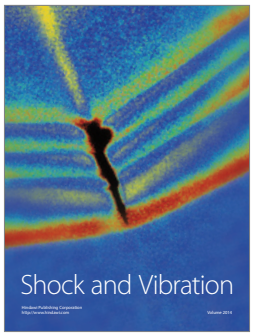


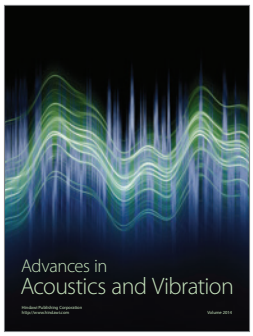












Hindawi

Submit your manuscripts at
http://www.hindawi.com