Soft Capacity of OFDMA Networks Is Suitable for Soft QoS Multimedia Traffic

Davide Chiarotto[*], **Leonardo Badia**[†‡] and **Michele Zorzi**[†‡]

[*] New Vision Group, 35016 Piazzola sul Brenta (PD), Italy
[†] Department of Information Engineering, University of Padova, 35131 Padova, Italy
[‡] Consorzio Ferrara Ricerche, 44122 Ferrara, Italy
email: davide.chiarotto@newvision.it, {badia,zorzi}@dei.unipd.it

Abstract—**Multimedia traffic is expected to be widespread in next generation wireless networks, which will be likely based on Orthogonal Frequency Division Multiple Access. While multimedia content is heavily demanding in terms of network resources, it is also inherently adaptable at the application layer, thereby imposing soft QoS constraints, rather than strict requirements on a specific data rate. In this paper, we specifically investigate the suitability of such a medium access control rationale for this kind of traffic. It turns out that, if properly managed, next generation networks can accommodate several multimedia users, thanks to a proper exploitation of user and frequency diversity. However, on the application side a great deal of attention should be paid to take advantage of scalability of the video flows and adaptability of this kind of traffic, to exploit the network capacity at its fullest.**

Index Terms—**Multimedia traffic, resource allocation, medium access control, wireless networks.**

I. Introduction

NOWADAYS, a growing number of wireless network users are equipped with powerful portable devices (e.g., smart phones, tablets), which make it possible to experience innovative multimedia communications. This compels mobile operators to support video traffic delivery in their next generation wireless networks with very good quality and in an efficient way, offering multimedia services such as Personal Broadcast, Mobile TV and Video on Demand. To satisfy the user expectations, the Evolved Packet System has been standardized by the Third Generation Partnership Project (3GPP) [1]. The new release is characterized by a new core network based on the Internet Protocol (IP) and a new radio access technology, the Long Term Evolution (LTE) air interface. A further advancement of LTE is represented by LTE-Advanced (LTE-A), which represents the latest 3GPP standardization for the radio access [2]. At the access level, LTE and LTE-A employ Orthogonal Frequency Division Multiple Access (OFDMA) technology in the downlink, to serve end users in the network by handling traffic on a subcarrier-by-subcarrier basis for a specified number of symbol periods [3].

Typically, video users have certain Quality-of-Service (QoS) requirements that the network operator has to address. However, most models inappropriately regard video traffic as composed of *constant* bit rate flows, which leads to *guaranteed* rate assignments. That is, the scheduler must grant a *fixed* transmission rate to all users, with the aim to satisfy strict QoS constraints. We deem such a strategy to be inappropriate and to lead to a too conservative usage of the available capacity. In the literature, many papers showed that the efficiency of the radio resource allocation can be highly improved if the network traffic allows rate adaptation, a scenario known as *elastic* traffic (as opposed to rigid traffic, strictly requiring a minimum guaranteed rate and more difficult to handle). The main inspiration for our work is [4], which shows that the capacity of a cellular system (in that case, a CDMA network) is increased if the elastic traffic can tolerate a slower rate.

Actually, video content is highly dynamic and consists of packets with different priority, which permits to adapt video flows to the available transmission resources and adjust their data rate according to the network conditions, as enabled by video compression formats such as H.264/AVC [5]. Our primary idea is instead to directly exploit this adaptability of multimedia traffic, that we dub "soft QoS," in a similar way to what is done for the elastic traffic. Moreover, we will show that soft QoS becomes particularly useful when the network capacity, instead of being fixed a priori, is determined by interference and channel conditions, a situation known as "soft capacity." This wording also reflects that, if radio resource management is performed following our approach, network capacity and user perceived QoS can be fitted together, so that the efficiency of the allocation is improved.

Our first aim is to quantify the Shannon capacity limit [6] of the downlink of a cellular system where multiple users are served by a single scheduler. This shows the impact of multiuser diversity on the total capacity, hinting that more users than the guaranteed approach can be served by exploiting a "soft QoS" principle, which can be described as an intermediate step between hard QoS constraints and a pure best effort approach. In our allocation rationale, video users are guaranteed very loose minimal rates, which are sufficient to maintain the connection if the application quality is properly scaled down. Conversely, at the data link layer, we show that the "soft capacity" of the OFDMA access control (due to user diversity and frequency diversity) permits the allocation of many more users with such a soft QoS approach, as opposed to the limitations encountered by allocations satisfying hard QoS constraints. In other words, with our approach the system can

This work was partially supported by MEDIEVAL (MultiMEDia transport for mobIlE Video AppLications), a medium-scale focused EU research project of the 7th Framework Programme of the European Commission, grant agreement no. FP7-258053. D. Chiarotto was with Consorzio Ferrara Ricerche when this paper was written.

sustain the same number of users with the same rate as in the hard QoS approach and, in addition, another set of users with a lower rate which is however above a minimal guarantee. Such a guaranteed rate is below that of the hard QoS, but may still be sufficient if video scalability or other application features (e.g., terminals with limited capabilities) are properly exploited. The resulting allocation represents a definite improvement because a higher number of users are admitted to the system, yet the guarantees of the hard QoS case are still met.

The rest of this paper is organized as follows. In Section II we review the related literature. In Section III we describe our proposed system model. The possible allocation algorithms investigated are discussed in Section IV. Section V discusses and evaluates the impact of our proposal in a cellular scenario and finally Section VI concludes the paper.

II. Related Work

Scheduling and resource allocation are well investigated problems in the literature related to multimedia traffic in wireless networks [7]. The introduction of the OFDMA technique in cellular systems has triggered many authors to find an efficient manner to assign a portion of a shared resource to each user. To perform this in an *optimal* way, the problem is often converted into the optimization of a set of parameters, such as packet delay and transmission rate. For example, the problem of allocating users is considered in [8] with a two-level scheduling procedure. In the first step, the allocator establishes the amount of data to transmit, in order to respect the delay constraint of each video flow; then, in the second step, the flows are accommodated in the time-frequency grid considering fairness and system throughput as constraints. Another interesting approach is proposed in [9], where the problem of assigning portions of the OFDMA grid to each user is seen as a knapsack problem. In addition, the authors classified other papers in the literature, grouping them together according to the parameter that is optimized (e.g., instantaneous rate, fairness or inter–cell interference). A different perspective is adopted in [10], where a Lagrange dual decomposition has been exploited to solve the sum rate and weighted sum power maximization problems. Basically, the authors proposed an efficient algorithm to solve both allocation problems, knowing that their optimal solution consists in performing multilevel water-filling. A remarkable interpretation of the cross–layer philosophy is given in [11], where a joint optimization of the application, datalink, and physical layers is developed, but without considering the LTE specifications. Similarly, [12] proposed a new resource allocator for video streaming. Using cross-layer information, e.g., packet delay, signal to noise ratio, and buffer occupancy, an algorithm determines the resources needed to support real time flows for the users.

A parallel line of research studies how to accommodate in the OFDMA time-frequency grid both the *elastic* traffic, which imposes no hard constraint on QoS, and the *constant bit rate* traffic. In [13], the simultaneous presence of both types of traffic is modeled by a continuous time Markov chain for the OFDMA-based WiMAX system. Assuming that all real time applications have constant bit rate, the paper proposes an analysis to derive the call blocking probability, mean delay and mean throughput performance of an admission control system that guarantees a minimum rate to each user. In the literature, the coexistence of the two types of traffic is handled by reserving part of the resources to the constant bit rate traffic, and allocating in the remaining part the elastic traffic, which has more degrees of freedom. This approach has been used in several papers, such as [14] and [15], which have derived algorithms to maximize the overall throughput of a cellular system. Resource allocation is explored in [16] through an information theoretic approach, so as to account for constraints on the blocking probability and the mean throughput for both kinds of traffic (constant bit rate and elastic).

With respect to this literature, focused on the proposal of new algorithms with different objective functions, the goal of our work is to assess the impact of user and frequency diversity on the performance of a realistic cellular network scenario and show that a different rationale is possible: besides hard QoS constraint imposed by the guaranteed approach, it is also possible to further serve new users with a soft QoS, which is not a best effort criteria, but rather a constraint with a lower requirement in terms of rate with respect to the hard QoS.

III. System Model

The transmission scheme in LTE-A follows the OFDMA shared-channel transmission principle, i.e., the time-frequency resource is dynamically shared between users. The eNodeB (eNB) performs the resource allocation at each transmission time interval (TTI), whose duration is the same as a subframe, i.e., 1 ms. Each subframe contains 2 slots and each slot consists of 7 OFDM symbols. The number of available subcarriers changes depending on the transmission bandwidth, but subcarrier spacing is fixed to $\Delta\mathrm{f} = 15$ kHz. The eNB controls to which users the shared resources are assigned by giving a combination of Resource Blocks (RBs). A RB is the smallest element that can be allocated to a user; it has the duration of a single subframe ($\mathrm{T}_{\mathrm{RB}} = 14$ OFDM symbols) and consists of $\mathrm{K}_{\mathrm{RB}} = 12$ subcarriers, for a nominal bandwidth of 180 kHz.

We consider a single cell with Q_{TOT} stationary users potentially served by a single eNB, which has N_{RB} RBs available in the downlink physical resource at each TTI. Moreover, we suppose that the interference across the RBs has been removed [17]. Let $a(m,n)$ be the m–th resource element (RE), which is the smallest modulation structure in LTE, or rather is the (complex) modulation symbol transmitted by the eNB in a given subcarrier of the nth RB, which is composed of $\mathrm{M}_{\mathrm{RE}} = \mathrm{K}_{\mathrm{RB}}\,\mathrm{T}_{\mathrm{RB}} = 168$ REs, since the duration of a RB is equal to $\mathrm{T}_{\mathrm{RB}} = 14$ OFDM symbols. The received RE as seen by the q–th user, $q = 1, \ldots, Q_{\mathrm{TOT}}$, is expressed by

$$\widetilde{a}_q(m,n) = A^{1/2} d_q^{-\eta/2} h_q(m,n) a(m,n) + z_q(m,n) \qquad (1)$$

with $m = 1, \ldots, \mathrm{M}_{\mathrm{RE}}$ and $n = 1, \ldots, N_{\mathrm{RB}}$. The path loss between the eNB and user q is $(A d_q^{-\eta})^{(1/2)}$, where η is the path loss exponent and A is a unitless constant [6]; the term $z_q(m,n)$ denotes the complex white Gaussian noise term for the m–th RE of the n–th RB with zero mean and power $\mathbb{E}[|z_q(m,n)|^2] = \mathrm{N}_0\,\Delta\mathrm{f}$. The channel coefficient of user q, RE m and RB n, $h_q(m,n)$, is modeled by a quasi-static Rayleigh fading distribution, i.e., it is assumed to be constant within each RE and it is a complex Gaussian random variable with zero mean and power equal to σ^2. The quasi-static assumption defines the behavior of the channel coefficient inside each RE. We also utilize two different models to describe the channel coefficients between different REs and RBs. In the *RE-iid* model, we assume that $h_q(m,n)$ varies independently and identically distributed (iid) between different REs; conversely, in the *RE-const* model, $h_q(m,n)$ is constant for all the REs belonging to the same RB, and varies independently between different RBs. We observe that the two different models here presented permit to study the main phenomena involved in the resource allocation, such as user diversity and frequency diversity. For the sake of simplicity, we do not consider further features of the LTE-A standard, such as QoS profiles [1].

Let γ_q define the received signal-to-noise ratio (SNR) of a RB between the eNB and user q, that is

$$\gamma_q = \frac{\mathrm{P} A d_q^{-\eta}}{\mathrm{N}_0\,\Delta\mathrm{f}\,\mathrm{K}_{\mathrm{RB}}} \tag{2}$$

where P is the power allocated by the eNB to a single RB.

Based on the Shannon capacity limit [6], the maximum achievable spectral efficiency between the eNB and the q-th user (and RE m, RB n) is given by $C_{a_q}(m,n) = \log_2(1 + |h_q(m,n)|^2\gamma_q)$, evaluated in b/s/Hz. Similarly, the spectral efficiency of the nth RB of user q can be derived as

$$C_{\mathrm{RB}_q}(n) = \sum_{m=1}^{\mathrm{M}_{\mathrm{RE}}} C_{a_q}(m,n) = \sum_{m=1}^{\mathrm{M}_{\mathrm{RE}}} \log_2(1 + |h_q(m,n)|^2\gamma_q) \tag{3}$$

The eNB can decide how to reserve the spectrum (i.e., the RBs at each TTI) to each user $q{=}1,\ldots,Q_{\mathrm{TOT}}$ according to the spectral efficiency $C_{\mathrm{RB}_q}(n)$, with $n{=}1,\ldots,N_{\mathrm{RB}}$. This procedure permits to derive the total spectral efficiency as

$$C_{\mathrm{TOT}} = \sum_{n=1}^{N_{\mathrm{RB}}} C_{\mathrm{RB}_{q^*(n)}}(n) \tag{4}$$

where $q^*(n)$ represents the user to be allocated in the nth RB. We also note that equation (3) does not change with the *RE-iid* assumption, while in the *RE-const* case it can be simplified to $C_{\mathrm{RB}_q}(n) = \mathrm{M}_{\mathrm{RE}} \log_2(1 + |h_q(n)|^2\gamma_q)$. Even though the practical values that the system can sustain during a real LTE transmission may be different, this evaluation can provide a qualitatively effective way to estimate the performance of the cellular scenario that can be used for performance assessment purposes. To quantify the global performance of the cellular system, also including the presence of multiple users, we need to resort to other algorithms, which try, in different ways, to allocate the resource of the eNB to the users.

IV. Algorithms for Resource Allocation

In this section, we investigate algorithms for allocating the time-frequency resource to multiple users and evaluate the resulting capacity. The main goal of all the algorithms is to achieve a value C_{TOT} for the spectral efficiency of the cellular system that is as high as possible. This is realized by also imposing at the same time, under different forms, a fairness constraint, i.e., a lower limit on $N_{\mathrm{RB}}/Q_{\mathrm{TOT}}$, which establishes the number of minimum RBs for each user. This division could have a non-zero remainder, therefore we need to introduce the threshold $N_{\mathrm{MIN}} = \lfloor N_{\mathrm{RB}}/Q_{\mathrm{TOT}} \rfloor$. If the remainder of $N_{\mathrm{RB}}/Q_{\mathrm{TOT}}$ is 0, all the RBs are equally divided between the users giving N_{MIN} RBs to each of them; otherwise, we have a reduced set of RBs, with cardinality $N_{\mathrm{REM}}{=}N_{\mathrm{RB}}{-}Q_{\mathrm{TOT}}N_{\mathrm{MIN}}$, which cannot be equally divided. In this case, the fairness condition is met by giving N_{MIN} RBs to $Q_{\mathrm{TOT}}{-}N_{\mathrm{REM}}$ users and $N_{\mathrm{MIN}}{+}1$ RBs to the remaining N_{REM} users. The algorithms described below take into account these differences. In the following, we indicate with $\mathcal{Q}{=}\{1,\ldots,Q_{\mathrm{TOT}}\}$ the set of users that can be potentially served and with $\mathcal{N}{=}\{1,\ldots,N_{\mathrm{RB}}\}$ the set of available RBs. A numerical comparison of the proposed strategies will be carried out at the end of this section and the most representative of them will be used to show the impact of soft QoS on the performance, see Section V.

a) THR-MAX: for each RB, select the user with the best spectral efficiency (fairness is neglected). Let $q^*(n)$ be the selected user for the nth RB through the following relationship:

$$q^*(n) = \arg\max_{q\in\mathcal{Q}} C_{\mathrm{RB}_q}(n) \quad \forall\ n \in \mathcal{N} \tag{5}$$

and C_{TOT} in (4) becomes

$$C_{\mathrm{TOT}}^{(\textit{RE-iid})} = \sum_{n=1}^{N_{\mathrm{RB}}} \sum_{m=1}^{\mathrm{M}_{\mathrm{RE}}} \log_2(1 + |h_{q^*(n)}(m,n)|^2\gamma_{a_{q^*(n)}}) \tag{6}$$

$$C_{\mathrm{TOT}}^{(\textit{RE-const})} = \mathrm{M}_{\mathrm{RE}} \sum_{n=1}^{N_{\mathrm{RB}}} \log_2(1 + |h_{q^*(n)}(n)|^2\gamma_{a_{q^*(n)}}) \tag{7}$$

for *RE-iid* and *RE-const*, respectively.

b) THR-FAIR: the previous approach does not fulfill fairness requirements, but uses all the available N_{RB} RBs. In a somewhat dual fashion to THR-MAX, the THR-FAIR approach allocates to each user its most favorable N_{MIN} RBs within the pool. However, this policy is left free to violate the constraint on the RB allocation by assigning the same RB to more than one user; thus, it is a non-achievable upper bound. After the first assignment, we may need to further allocate the remaining N_{REM} RBs, as seen before. Again, these are assigned so that each user selects the best RB within $\mathcal{N}_{\mathrm{REM}}$ as

$$n_q^{\mathrm{REM}} = \arg\max_{n\in\mathcal{N}_{\mathrm{REM}}} C_{\mathrm{RB}_q}(n) \quad \forall\ q \in \mathcal{Q} \tag{8}$$

and the corresponding value of spectral efficiency is grouped into the set $\{C_{\mathrm{RB}_1}(n_1^{\mathrm{REM}}),\ldots,C_{\mathrm{RB}_{Q_{\mathrm{TOT}}}}(n_{Q_{\mathrm{TOT}}}^{\mathrm{REM}})\}$. Regardless of the RBs overlap, the algorithm completes the allocation procedure by selecting the N_{REM} users with the highest spectral efficiency from the previous set.

c) FAIR-MAX: this strategy respects the constraints of both RBs and fairness. The algorithm starts by computing

$$q^*(n^*) = \arg\max_{q\in\mathcal{Q},n\in\mathcal{N}} C_{\mathrm{RB}_q}(n) \tag{9}$$

and continues by removing the allocated RB n^* from set $\mathcal{N}$ and also the user q^* from $\mathcal{Q}$ only if it satisfies the fairness condition N_{MIN} (i.e., with RB n^*, user q^* has received the N_{MIN}–th RB); the algorithm derives again (9) with the updated sets until $\mathcal{Q}$ is empty. In a very similar way, the remaining N_{REM} RBs are allocated, with the constraint that each user can reserve only 1 extra RB over the limit imposed by N_{MIN}.

d) FAIR-VAR: the allocator exploits the channel variability for each user, by deriving, before selecting the RB with the highest spectral efficiency, the variance of the set $\{C_q(1), \ldots, C_q(N_{\text{RB}})\}$ for each user $q \in \mathcal{Q}$. This permits to assess the sensitivity of each user: the lower the variance of the user, the higher the probability that all the RBs have a similar value. Therefore, users with higher variance are allocated first. Once the most sensitive user is selected, the algorithm selects the RB with the best channel quality, which is removed from the pool of available RBs; the same happens to the user if it reaches the maximum value N_{MIN} imposed by fairness requirements. This procedure stops when all users have N_{MIN} RBs each. The remaining N_{REM} RBs are accommodated in a similar way, with the constraint that each user can reserve only 1 extra RB with respect to the limit imposed by N_{MIN}.

e) FAIR-BIP-joint: the optimal feasible allocation is

$$\max_{\{x_q(n), \forall q, n\}} \sum_{q=1}^{Q_{\text{TOT}}} \sum_{n=1}^{N_{\text{RB}}} x_q(n) C_{\text{RB}_q}(n) \tag{10}$$

$$\text{subject to} \sum_{q=1}^{Q_{\text{TOT}}} x_q(n) \le 1 \qquad \forall\, n \in \mathcal{N} \tag{11}$$

$$\sum_{n=1}^{N_{\text{RB}}} x_q(n) \le N_{\text{MIN}} + \mathbf{1}_{\text{REM}} \qquad \forall\, q \in \mathcal{Q} \tag{12}$$

which is a binary integer program (BIP) and can be solved using a branch-and-bound algorithm. Variable $x_q(n) \in \{0, 1\}$ is set to 1 if user q uses RB n and 0 otherwise. Eqs. (11) and (12) impose maximum RB and fairness constraints, respectively. In particular, (12) already includes N_{REM}. In fact, $\mathbf{1}_{\text{REM}}$ is equal to 1 only if $N_{\text{REM}} \neq 0$ and 0 otherwise.

f) FAIR-BIP-2steps: this strategy can be used only if $N_{\text{REM}} \neq 0$. It exploits BIP as the FAIR-BIP-joint algorithm, but proceeds in two steps. First, the algorithm optimally allocates N_{MIN} RBs to each user, as per (10), except for (12) becoming $\sum_{n=1}^{N_{\text{RB}}} x_q(n) \le N_{\text{MIN}}$. Secondly, the remaining N_{REM} RBs have to be allocated between the users. In this case, condition (12) reduces to $\sum_{n=1}^{N_{\text{RB}}} x_q(n) \le 1$, in fact at most 1 extra RB can be allocated to each user. Thus, if $N_{\text{REM}} = 0$, the first step of this approach is equal to the previous problem (10) and the second step is ignored.

g) FAIR-EXACT-LP: the last algorithm defines the problem as the following linear program (LP)

$$\max_{\{y_q(n), \forall q, n\}} \sum_{q=1}^{Q_{\text{TOT}}} \sum_{n=1}^{N_{\text{RB}}} y_q(n) C_{\text{RB}_q}(n) \tag{13}$$

$$\text{subject to} \sum_{q=1}^{Q_{\text{TOT}}} y_q(n) \le 1 \qquad \forall\, n \in \mathcal{N} \tag{14}$$

$$\sum_{n=1}^{N_{\text{RB}}} y_q(n) \le \frac{N_{\text{RB}}}{Q_{\text{TOT}}} \qquad \forall\, q \in \mathcal{Q} \tag{15}$$

where $0 \le y_q(n) \le 1$. Compared to FAIR-BIP-joint, this algorithm derives the exact value of the total spectral efficiency for the considered scenario supposing that the fairness constraint holds with equality. This approach gives non-integer partitions, although RBs are atomic and cannot be split. Its fractional allocation coincides with that of FAIR-BIP-joint if $N_{\text{REM}} = 0$.

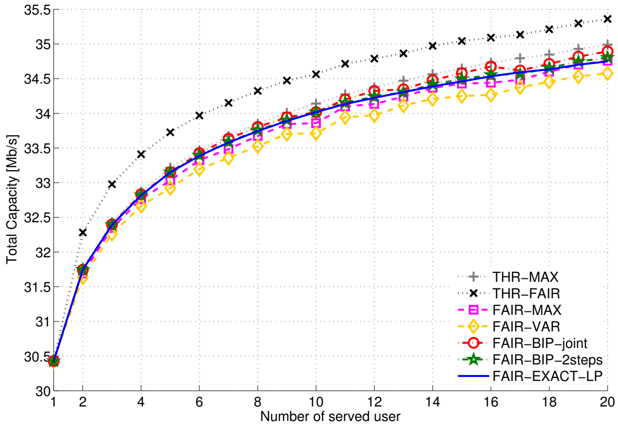

Fig. 1. Total capacity in Mb/s versus the number of served users for bandwidth equal to 10 MHz, 50 RBs, $\sigma^2 = 1$ and in the *RE-iid* case.

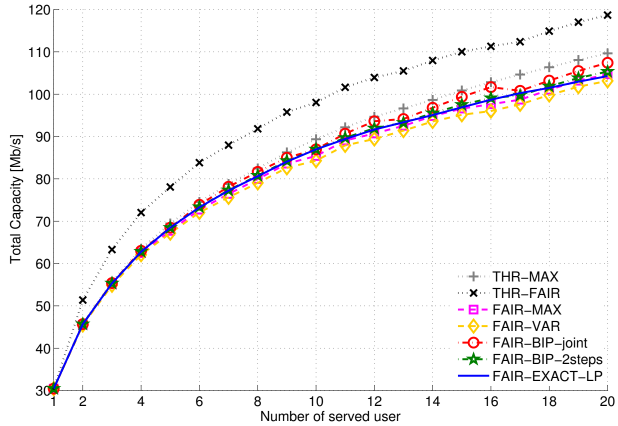

Fig. 2. Total capacity (in Mb/s) versus the number of served users for bandwidth equal to 10 MHz, 50 RBs, $\sigma^2 = 1$ and in the *RE-const* case.

A. Quantitative Algorithms Comparison

To quantify the impact of spatial diversity, we assume that all users $q \in \mathcal{Q}$ are stationary and placed at the same distance d from the eNB. Thus, in the formulas of Section III we set $d_q = d$, $\forall\, q \in \mathcal{Q}$; the only difference among the users is the realization of the channel coefficients of the REs, for which we used either the *RE-iid* or the *RE-const* assumptions. The received SNR of a RB, γ_q, defined in (2), is determined by considering an eNB transmit power of $\text{P} = 26.98$ dBm for each RB, a noise power spectral density of $\text{N}_0 = -174$ dBm/Hz, $d = 1$ km, $\eta = 4$ and carrier frequency equal to 2 GHz; A is set to the free space path gain at distance $d_0 = 1$ m [6]. Therefore, the received SNR for a RB is equal to $\gamma = 8.6$ dB. Finally, the channel bandwidth is $B = 10$ MHz, hence $N_{\text{RB}} = 50$ [1] and the total capacity, in b/s, is given by BC_{TOT}, with C_{TOT} defined in (4) as the total spectral efficiency.

The total capacity obtained by each resource allocator with σ^2=1 for *RE-iid* and *RE-const*, is reported in Figs. 1 and 2, respectively. In spite of their different scales, the trend of the

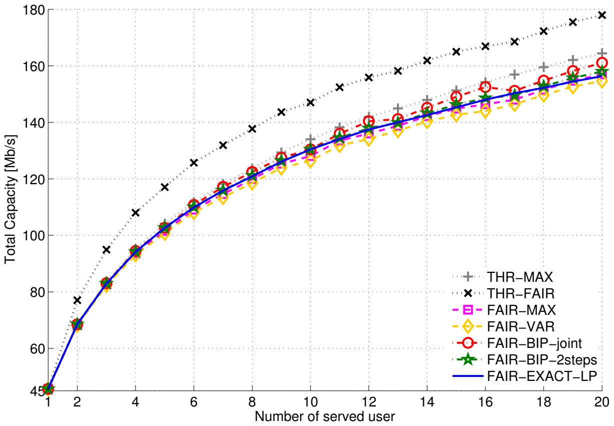

Fig. 3. Total capacity in Mb/s versus the number of served users for bandwidth equal to 10 MHz, 50 RBs, $\sigma^2 = 1.5$ and in the *RE-const* case.

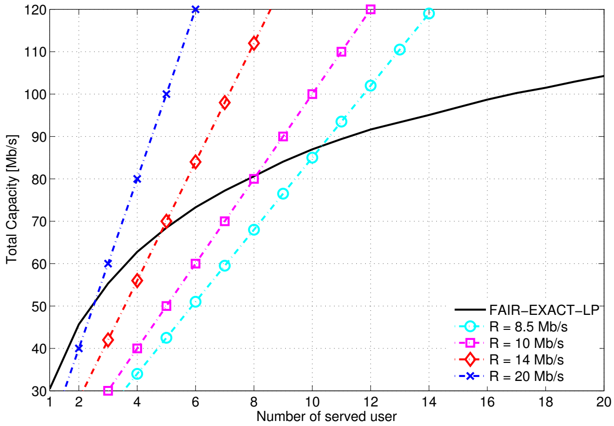

Fig. 4. Total capacity versus number of served users with channel bandwidth equal to 10 MHz (50 available RBs), $\sigma^2 = 1$ and *RE-const* assumption.

curves is similar. The larger capacity values of Fig. 2 are due to the contribution to capacity of a given RB in *RE-iid* being very similar for all the users, since it results as the sum of iid contributions of each OFDMA symbol. Thus, multi-user diversity is low. In the *RE-const*, instead, a single channel coefficient determines the spectral efficiency of the RB for each user, and multi-user diversity is higher. Thus, Fig. 1 can be seen as an overall lower bound, which is however very loose. In fact, we tested other intermediate assumptions about the correlation of RBs and Fig. 2 appears to be much more representative even when only mild correlation is present. Therefore, we focus only on the *RE-const* in the following.

We highlight that all the algorithms exhibit proper comparative performance. First of all, the non-achievable upper bound THR-FAIR is always the topmost curve. The second top most curve is THR-MAX, which gives an achievable allocation and can be seen as an upper bound (though not exactly, because it is a greedy policy). Instead, the lowest curves are related to the approaches that take fairness into account: FAIR-MAX and FAIR-VAR. Finally, the BIP and LP approaches achieve intermediate values and can be regarded as slightly more accurate, since they solve constrained maximization problems. However, the span of all the achievable curves is pretty narrow. Hence, regardless of the actual algorithm used, any of these curves can be used as a good indicator of the total capacity.

In Fig. 3 we also verified the impact of the channel variability by changing σ^2 to 1.5; other choices would yield identical results. Again, the values change quantitatively but not qualitatively (the curves scale almost perfectly), thereby confirming the generality of our reasonings. Thus, although we are well aware of the different theoretical rationale behind each algorithm, we believe that these numerical results support general conclusions about Soft QoS allocation, even by limiting the attention to some specific algorithms only.

V. Soft QoS Allocation

To see how soft QoS can be provided to multimedia users, it does not seem restrictive to take a specific curve of Fig. 2, as they all exhibit similar behavior. In particular, in the following we look at the FAIR-EXACT-LP curve, although the same reasoning applies to any other plot or algorithm. A notable feature of the curve is that it is not flat but increases steadily even when the number of users is high. This is evidence of the multiuser diversity phenomenon in OFDM networks and justifies the term "soft capacity" for such systems.

In the following, we analyze our capacity estimations for system dimensioning purposes, i.e., we use the curves derived previously to pursue a rough quantification of the number of users that can be admitted into the system. This evaluation is to be meant just as on average and with no pretense of optimality, even though it would be possible to apply these results also from a more rigorous point of view. However, more than seeking the actual optimization of how many users can be admitted into the system, which can be subject of further investigations, we just want to point out how the soft capacity of the system is better matched by exploiting at the same time the soft QoS requirements imposed by multimedia traffic.

The objective of fully exploiting the system capacity can be seen as the maximization of the number of users in the system. In the classic "hard QoS" approach, this would mean to allocate as many users as possible by providing all of them with a minimum guaranteed rate R.[1] With a conceptual simplification, we may assume that, when the system capacity is C, we allocate on average $\lfloor C/R \rfloor$ users. Note that the capacity value C is the sum of the individual rates achievable by the users; thus, in reality if these values are highly diverse, one cannot trade the allocation of a user to another. However, since we consider users at the same distance from the eNB, our argument still holds on average. Moreover, it would be straightforward to extend to more complex scenarios, the only difference being that the capacity curve is harder to quantify (but it would still have the same "soft" increasing behavior).

To follow the "hard QoS" rationale, in Fig. 4 we evaluate C as given by the FAIR-EXACT-LP approach and plot it versus the number of users N. Let $y = C(N)$ represent the behavior of the solid curve in Fig. 4. Moreover, we take different values of $R \in \{8.5, 10, 14, 20\}$ Mb/s and we plot dashed lines $y{=}RN$ that cross the solid curve in the maximum number of users that can be allocated respecting the hard QoS constraint. E.g., with

[1] The same requirement for all the users keeps the analysis simple; considering different requirements is possible but would be out of scope here.

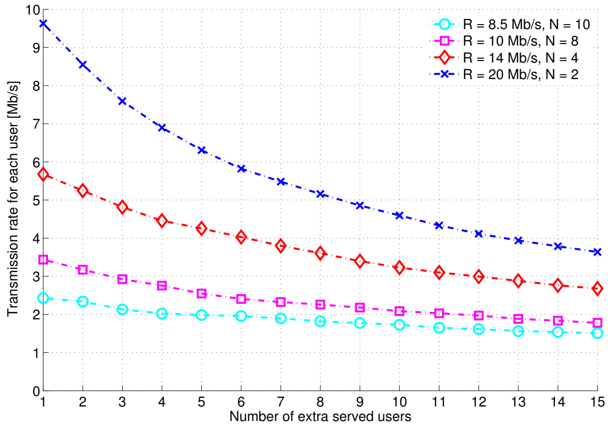

Fig. 5. Transmission rate for each extra user beyond the N allocated with a hard QoS approach. Bandwidth is 10 MHz, $\sigma^2 = 1$, *RE-const* assumption.

R=10 Mb/s, the system can admit N=8 users with hard QoS. Similarly, the system can sustain N=5 users with R=14 Mb/s.

Yet, the curve keeps increasing, thus there would be room for additional users. We consider multimedia traffic to be elastic and fill the remaining capacity in a softer manner. However, allocating traffic without rate guarantees would be unsuitable for multimedia traffic, which is highly demanding and requires a minimal rate. Thus, we take an intermediate approach between the hard QoS and a totally unregulated best effort allocation. We want to see if there is room to allocate $\widetilde{N}$ extra users at a given transmission rate $\widetilde{R}$, which is lower than R, but still provides a loose guarantee.

As per the previous example, we guarantee $N = 8$ users to be served at $R = 10$ Mb/s. However, in our soft QoS allocation we can additionally introduce $\widetilde{N} = \{0, 1, 2, \ldots\}$ users with a lower guarantee $\widetilde{R}$ where $C(N + \widetilde{N}) = C(N) + \widetilde{R}\widetilde{N}$ can be reformulated as $\widetilde{R} = (C(N + \widetilde{N}) - C(N))/\widetilde{N}$ and imposes an upper limit to the utilization of the available capacity. Fig. 5 plots the resulting $\widetilde{R}$ versus the number $\widetilde{N}$ of extra users served, showing the role of spatial diversity in the definition of the soft capacity: e.g., with $N = 8$ users served at a guaranteed rate of $R = 10$ Mb/s, one may further allocate $\widetilde{N} = 2$ users with soft QoS guarantee of $\widetilde{R} = 3$ Mb/s and even 11 extra soft users if the soft QoS guarantee is set to 2 Mb/s. In conclusion, the total capacity for 10 MHz of channel bandwidth is equal to 80 Mb/s with hard QoS requirement fixed at $R = 10$ Mb/s and 8 served users; however, it can reach 86 or 102 Mb/s in case of soft QoS fixing the guaranteed transmission rate to 3 or 2 Mb/s and serving 10 or 19 users, respectively.

Depending on the kind of terminal, even lower guarantees can be acceptable for multimedia users as they may provide a sufficient QoS in many cases. Actually, the key point of the allocation is to carefully select the users, and possibly scale down the rate provided to the users that allow for adaptation. In this sense, our "soft QoS" approach can perform even better than what envisioned by this discussion. The average values can be improved if a proper user selection mechanism is adopted. Moreover, to push adaptability of multimedia traffic even further, the rate guarantee R can be adjusted and capacity can be increased by jointly optimizing both R and $\widetilde{R}$. These points represent possible further developments of this work.

VI. Concluding Remarks

We derived the capacity of an OFDMA cellular network under different approaches, summarizing their most important qualitative aspects. Regardless of the specific methodology, an important general conclusion is that the OFDMA capacity is inherently soft, due to multiuser and frequency diversities. This represents an advantage to exploit for soft QoS traffic such as that of multimedia users. Thus, we proposed a soft QoS allocation, as opposed to the hard QoS resource assignment with fixed rates, providing the same guarantee to an identical number of users but also a looser rate guarantee to additional users. We believe that such a model can have influential consequences on the quality provision paradigm of multimedia traffic in next generation networks. This whole reasoning was derived from aggregate capacity evaluations; the extension of this idea to either optimization frameworks or practical allocators in LTE networks will be considered in future work.

References

[1] 3GPP, "The mobile broadband standard." [Online]. Available: www.3gpp.org

[2] T. Nakamura, "Proposal for candidate radio interface technologies for IMT-Advanced based on LTE release 10 and beyond (LTE-Advanced)," 3GPP IMT-Advanced Evaluation Workshop, Dec. 2009.

[3] H. Yin and S. Alamouti, "OFDMA: A broadband wireless access technology," in *Proc. of IEEE Sarnoff Symposium*, Princeton, NJ, Mar. 2006, pp. 848–852.

[4] E. Altman, "Capacity of multi-service cellular networks with transmission-rate control: a queueing analysis," in *Proc. of ACM MobiCom*, Atlanta, GA, Sep. 2002, pp. 205–214.

[5] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[6] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Pr., 2005.

[7] J. Huang, V. Subramanian, R. Berry, and R. Agrawal, *Scheduling and Resource Allocation in OFDMA Wireless Communication Systems*. Auerbach Publications, Taylor and Francis Group, 2010.

[8] L. Badia, A. Baiocchi, A. Todini, S. Merlin, S. Pupolin, A. Zanella, and M. Zorzi, "On the impact of physical layer awareness on scheduling and resource allocation in broadband multicellular IEEE 802.16 systems," *IEEE Wireless Commun. Mag.*, vol. 14, no. 1, pp. 36–43, Feb. 2007.

[9] G. Bartoli, A. Tassi, D. Marabissi, D. Tarchi, and R. Fantacci, "An optimized resource allocation scheme based on a multidimensional multiple-choice approach with reduced complexity," in *Proc. of IEEE ICC*, Kyoto, JP, Jun. 2011.

[10] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. of IEEE ISIT*, Seattle, WA, Jul. 2006, pp. 1394–1398.

[11] S. Karachontzitis, T. Dagiuklas, and L. Dounis, "Novel cross-layer scheme for video transmission over LTE-based wireless systems," in *Proc. of IEEE ICME*, Barcelona, ES, Jul. 2011.

[12] H. A. Mohd Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, X. Minjie, and C.-C. Lin, "Resource allocation technique for video streaming applications in the LTE system," in *Proc. of WOCC*, Shanghai, CN, May 2010, pp. 1–5.

[13] C. Tarhini and T. Chahed, "System capacity in OFDMA-based WiMAX," in *Proc. of ICSNC*, Tahiti, PF, Oct. 2006.

[14] C. Koutsimanis and G. Fodor, "A dynamic resource allocation scheme for guaranteed bit rate services in OFDMA networks," in *Proc. of IEEE ICC*, Beijing, CN, May 2008, pp. 2524–2530.

[15] N. Mokari, M. Javan, and K. Navaie, "Resource allocation based on channel distribution information for elastic and streaming traffic in OFDMA networks: A heuristic algorithm," in *Proc. of IEEE VTC-Spring*, Barcelona, ES, Sep. 2009.

[16] M. Karray, "Analytical evaluation of QoS in the downlink of OFDMA wireless cellular networks serving streaming and elastic traffic," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1799–1807, May 2010.

[17] D. Huang and K. B. Letaief, "An interference-cancellation scheme for carrier frequency offsets correction in OFDMA systems," *IEEE Trans. Commun.*, vol. 7, no. 7, pp. 1155–1165, Jul. 2005.