

Linköping University Post Print

**Optimal OFDMA Downlink Scheduling Under
a Control Signaling Cost Constraint**

Erik G. Larsson

N.B.: When citing this work, cite the original article.

©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Erik G. Larsson, Optimal OFDMA Downlink Scheduling Under a Control Signaling Cost Constraint, 2010, IEEE Transactions on Communications, (58), 10.

<http://dx.doi.org/10.1109/TCOMM.2010.090215>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-58553>

Optimal OFDMA Downlink Scheduling Under a Control Signaling Cost Constraint

Erik G. Larsson

Abstract—This paper proposes a new algorithm for downlink scheduling in OFDMA systems. The method maximizes the throughput, taking into account the amount of signaling needed to transmit scheduling maps to the users. A combinatorial problem is formulated and solved via a dynamic programming approach reminiscent of the Viterbi algorithm. The total computational complexity of the algorithm is upper bounded by $O(K^4N)$ where K is the number of users that are being considered for scheduling in a frame and N is the number of resource blocks per frame.

Index Terms—OFDMA, scheduling, control signaling, optimization.

I. BACKGROUND

OFDMA is a common access method in many systems contemplated for the future, such as LTE [1]. An advantage of OFDMA is that users can be scheduled for transmission/reception precisely in the specific time/frequency slots where their channel is good. The resulting improvement in system capacity and robustness is often referred to as “multiuser diversity” [2]. The basic benefit of multiuser diversity is that the channel statistics of individual users become less important. What matters, instead, is that at any given point in time and frequency, there is at least one user who desires to transmit or receive data, and who has a good channel.

There is a rather substantial literature on scheduling in OFDMA systems, see, e.g., [3] for a review. The basic philosophy behind the design of schedulers is to maximize the system sum-throughput, subject to some constraints. It is common, for example, to impose constraints on fairness [4], or on the quality-of-service experienced on the application layer [5]. It is also common to design schedulers that explicitly take into account delay requirements of data packets [6]. A basic difficulty with OFDMA is that all receiving users have to be told in what time/frequency-slots their data are located. This requires signaling of control information, consisting of scheduling maps. The overhead associated with the transmission of such control information can be significant [7]–[9] and may constitute a limiting factor in some systems [10].

In this paper we design a scheduler for the OFDMA downlink that maximizes the throughput by taking into account

the cost of conveying the scheduling decisions. To compress the scheduling information we use runlength encoding. We formulate the scheduling as a combinatorial problem that can be solved as a sequence of dynamic programming problems. As an integral component, the method contains a user selection mechanism. That is, herein “scheduling” means both choosing what users to transmit to, and assigning them time/frequency resources. The computational cost of our proposed algorithm is proportional to the number of subcarriers.

The problem that we consider here has received very little attention previously. We are aware of one related sequence of work: [7], [8], that formulated quantitative models for the signaling cost and proposed some optimized schedulers. The main novelties over [7], [8] are: (i) we formulate and solve the problem both for the case that the same scheduling map is broadcast to all users, and for the case that the users receive individual scheduling maps; (ii) we integrate user selection and scheduling, i.e., we do not assume that the set of users selected for transmission is fixed beforehand; (iii) we formulate the problem so that it can be computationally efficiently solved; and (iv) our formulation includes a single scalar parameter that models the relative cost of transmitting the scheduling information: by changing this parameter, more or less complex maps are obtained. A fundamental difference to [8] is that we do not force the scheduler to assign the same number of subcarriers to all users. This constraint was imposed to obtain “fair” solutions. By contrast, our formulation yields solutions where some users are allocated many resource blocks, some are allocated a few, and some are allocated none. We allow for fairness constraints via the per-resource-block capacity expressions that go into the optimization problem.

II. NOTATION, ASSUMPTIONS AND PRELIMINARIES

We consider the downlink of an OFDMA system. We introduce the following terminology:

- A *frame* is a set of consecutively transmitted OFDM symbols. Typically, the control signaling data that contain the scheduling information occupies parts of the first few OFDM symbol(s).
- A *resource block* is an entity representing the smallest subdivision of the time/frequency domain that can be allocated to a user. We assume that the channel is approximately constant over one resource block. There are N resource blocks in each frame.
- K is the largest number of users that can be scheduled in the frame.
- $C_n(k)$ is the “capacity” conveyed by resource block n , if this block is assigned to user k . With adaptive modulation, $C_n(k)$ may be the modulation-constrained mutual

Paper approved by Y. Fang, the Editor for Wireless Networks of the IEEE Communications Society. Manuscript received October 10, 2009; revised March 29, 2010.

The author is with Linköping University, Dept. of Electrical Engineering (ISY), Division of Communication Systems, 581 83 Linköping, Sweden (e-mail: erik.larsson@isy.liu.se).

This work was supported in part by Ericsson, the Swedish Research Council (VR) and the Swedish Foundation for Strategic Research (SSF). E. Larsson is a Royal Swedish Academy of Sciences (KVA) Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation.

Digital Object Identifier 10.1109/TCOMM.2010.082010.090215

information, or just a function that assigns the number of bits per resource block that the modulation scheme can support if user k is scheduled in resource block n . $C_n(k)$ may also be a priority-adjusted throughput measure, for example that defined by a proportional-fair mechanism [2].

- \mathcal{U} stands for the set of users that are scheduled for transmission in the frame. (\mathcal{U} does not include the exact user–resource block association; this is contained in \mathcal{S} below.)
- \mathcal{S} denotes the resource block assignment for the users scheduled for transmission in the frame. If \mathcal{S} is known, then naturally \mathcal{U} is known too. Also, \mathcal{S}_n stands for the index of the user assigned to resource block n , under the assignment \mathcal{S} .

The scheduling problem is a combinatorial optimization problem with K^N candidate solutions. If transmission of the scheduling maps were for free, then the optimal assignment \mathcal{S} that maximizes the sum-throughput could be directly found from

$$\max_{\mathcal{S}} \sum_{n=1}^N C_n(\mathcal{S}_n). \quad (1)$$

Here $\sum_{n=1}^N C_n(\mathcal{S}_n)$ has the interpretation of ergodic capacity in the sense that for N large and with coding across the resource blocks, the achievable rate approaches $N \cdot E[C_n(\mathcal{S}_n)]$. The solution to (1) is the system-capacity maximizing scheduler, that simply takes

$$\mathcal{S}_n = \operatorname{argmax}_k C_n(k).$$

In practice the transmitter must also send the resource block assignment \mathcal{S} to the users. \mathcal{S} can be represented by a scheduling map \mathcal{M} (illustrated in Fig. 1) that identifies which user \mathcal{S}_n that is assigned to each resource block n . This map must be transmitted to the users within the same frame as the payload data. This can be done in at least two ways. The first possibility is to generate a binary map \mathcal{M}_k for each user k . The n th entry of the binary map \mathcal{M}_k is “1” if user k is scheduled at resource block n ($\mathcal{S}_n = k$), and “0” otherwise. The set of binary maps $\{\mathcal{M}_k\}$ for $k \in \mathcal{U}$ are then sent separately to the users. This is the way things are done in LTE. Alternatively, the joint scheduling map \mathcal{M} that represents the entire assignment \mathcal{S} may be broadcast to all users. That is, all users get information about the resource assignments of all other users.

The two possibilities of individual transmission of binary maps $\{\mathcal{M}_k\}$, and broadcasting of a joint map \mathcal{M} , both have their advantages and disadvantages. Which is better will heavily depend on the statistics of the channels in the system, and on the distribution of the users’ signal-to-noise-ratio (SNR) in particular [9]. The main insight is that since the individual maps \mathcal{M}_k are correlated, transmission of \mathcal{M} in lieu of $\{\mathcal{M}_k\}$ can save bandwidth. On the other hand, if \mathcal{M} is broadcast, then it must be encoded with a strong enough code so that all users, including the one with the smallest SNR, can decode it. This renders the broadcast method inefficient due to the excessive amount of error protection needed in some cases.

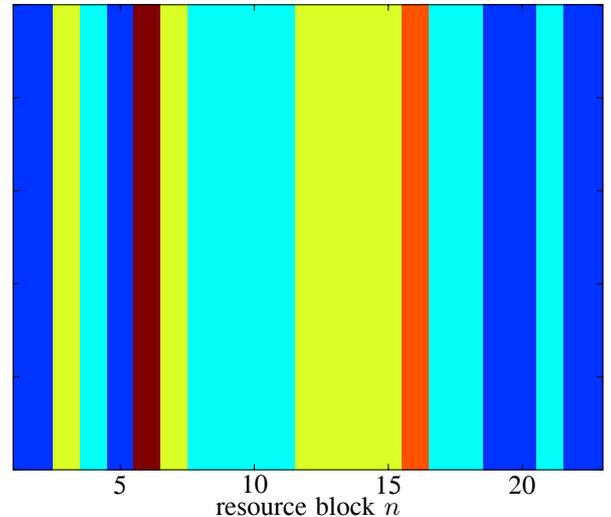


Fig. 1. Illustration of a joint scheduling map \mathcal{M} . The user indices are color coded.

III. PROBLEM FORMULATION

The problem is to select \mathcal{U} and \mathcal{S} under a constraint on the signaling cost. The optimization problems that we formulate in what follows are applicable both to the individual transmit strategy and to the broadcasting strategy. For both strategies, we assume that standard runlength encoding is used to encode the scheduling maps ($\{\mathcal{M}_k\}$ and \mathcal{M} respectively). Runlength encoding was used for similar purposes, e.g., in [9], [11]. Runlength encoding is suboptimal, but it is computationally inexpensive, and it does not require knowledge of the map statistics.

A. Optimization problem with individual transmission of scheduling maps $\{\mathcal{M}_k\}$

Here an individual scheduling map \mathcal{M}_k is associated with each scheduled user k . This map is encoded by runlength encoding. A new entry in the runlength table for \mathcal{M}_k is created precisely when a switch to user k from another user (or conversely, a switch from user k to another user) is performed at some resource block n . For example, for the green user in Fig. 1, the map \mathcal{M}_k is 00100010000111100000000 and the output of the runlength encoder is 2131448.

Generally, the size of \mathcal{M}_k after the runlength encoding is

$$R_k + \sum_{n=2}^N f_k(\mathcal{S}_n, \mathcal{S}_{n-1}, n)$$

where R_k is a fixed overhead for user k (to be defined later on) and $\sum_{n=2}^N f_k(\mathcal{S}_n, \mathcal{S}_{n-1}, n)$ is the number of bits added to the table when switching between the users occurs. When $\mathcal{S}_n = \mathcal{S}_{n-1}$, no switching occurs. When $\mathcal{S}_n \neq \mathcal{S}_{n-1}$ and $\mathcal{S}_n = k$ or $\mathcal{S}_{n-1} = k$, we must add P_n , say, bits to describe the length of the next active interval. Hence:

$$f_k(i, j, n) = \begin{cases} 0, & i = j \\ P_n, & i \neq j \cap (i = k \cup j = k) \end{cases}$$

We will take

$$P_n = \lceil \log_2(N - n) \rceil,$$

because the length of the next interval cannot exceed $N - n$.¹ Strictly speaking, with optimal map encoding, P_n might be taken to be smaller, because a data compression method could be used to encode the runlengths in the table. However, in the applications we consider, the runlengths are likely to have a nearly uniform distribution so the proposed model should be accurate. The total size of the compressed scheduling maps is

$$\sum_{k \in \mathcal{U}} \left(R_k + \sum_{n=2}^N f_k(\mathcal{S}_n, \mathcal{S}_{n-1}, n) \right). \quad (2)$$

The overhead R_k represents extra bits needed for initialization. Precise models for R_k can be defined for specific systems. For illustration, we will take

$$R_k = N_{\text{fec}} + 1$$

bits. The first term represents the overhead associated with error protection/detection of the scheduling map.² The second term accounts for one bit needed to tell the runlength decoder whether the user is assigned the first resource block ($n = 1$) or not.

We propose to maximize a throughput measure where the control signaling cost is penalized:

$$\max_{\mathcal{S}} \tilde{C}(\mathcal{S}) \quad (3)$$

where

$$\tilde{C}(\mathcal{S}) \triangleq \sum_{n=1}^N C_n(\mathcal{S}_n) - \sum_{k \in \mathcal{U}} \rho_k \left(R_k + \sum_{n=2}^N f_k(\mathcal{S}_n, \mathcal{S}_{n-1}, n) \right). \quad (4)$$

The first term of $\tilde{C}(\mathcal{S})$ in (4) is the capacity (throughput) that the frame can offer, measured in number of bits. The second term is a penalty term that accounts for the control signaling. Note that $\tilde{C}(\mathcal{S})$ in (4) is also a function of \mathcal{U} , since \mathcal{U} is implicit in \mathcal{S} .

In (4), the constants ρ_k are fixed weights that reflect how much more expensive it is to transmit one bit of control information, relative to losing one bit of payload data capacity (“capacity” in the sense of how $C_n(k)$). Ideally, adding one extra bit of control signaling means losing one bit of throughput in the frame, so $\rho_k = 1$. However, ρ_k may be increased above 1 to penalize control signaling more. One reason for doing so may be that the control signaling is encoded with much heavier error control coding than what is used for the payload data. Ultimately, this depends on what goes into $C_n(k)$. For generality, we allow ρ_k to depend on k .

¹In LTE a resource block has bandwidth 180 kHz and there is no scheduling in time. With a 20 MHz bandwidth, there are at most 110 resource blocks [1, p. 321]. Since $110 < 2^7 = 128$, the first value of P_n is $P_1 = 7$.

²In LTE this consists of a CRC check that uses 24 extra bits, so $N_{\text{fec}} = 24$ [1, p. 361]. In LTE this CRC check is also used to identify which scheduling map that is intended for which user. This works by letting all users try blindly to decode all maps. Whenever a user succeeds, she assumes that the corresponding maps was aimed at her. The probability of correctly decoding a map that was aimed for someone else is vanishingly small.

B. Optimization problem with broadcasting of joint scheduling map \mathcal{M}

Here there will be only one scheduling map \mathcal{M} for all users. See Fig. 1 for an example. In this example, assuming that the figure uses a rainbow color mapping of the user indices, the map is given by 11321532222333342211211. Runlength encoding yields 2111114412212, 1321532342121, where the first vector describes the runlengths, and the second vector represents the indices of the users that are assigned the corresponding resource blocks.

Each new entry in the runlength table requires $P_n + Q$ bits. P_n is the number of bits that are needed to represent the number of consecutive resource blocks given to each user after a switching point in the map occurs. Q is the number of bits needed to identify which user that is being assigned the following resource blocks. We take P_n as in Section III-A and

$$Q = \lceil \log_2(|\mathcal{U}|) \rceil.$$

There will be a fixed overhead of R bits for the entire map. For illustration purposes, we take

$$R \triangleq |\mathcal{U}| \cdot L_{\text{id}} + N_{\text{fec}} + \lceil \log_2(K) \rceil.$$

The first term represents the need to assign a short-ID to the \mathcal{U} users, assuming that each user has a long ID number of L_{id} bits.³ The second term models the CRC check (one for the entire map). The last term accounts for the number of bits needed to quantify how many users there are in total in the map.

The total size of the compressed scheduling map will be

$$R + \sum_{n=2}^N f(\mathcal{S}_n, \mathcal{S}_{n-1}, n), \quad (5)$$

where

$$f(i, j, n) = \begin{cases} 0, & i = j \\ P_n + Q, & i \neq j \end{cases}. \quad (6)$$

The optimization problem becomes

$$\max_{\mathcal{S}} \tilde{C}(\mathcal{S}) \quad (7)$$

where

$$\tilde{C}(\mathcal{S}) \triangleq \sum_{n=1}^N C_n(\mathcal{S}_n) - \rho \left(R + \sum_{n=2}^N f(\mathcal{S}_n, \mathcal{S}_{n-1}, n) \right). \quad (8)$$

Here we have just one single weight ρ .

IV. EFFICIENT ALGORITHM FOR SOLUTION OF (3) AND (7)

Problems (3) and (7) have the same structure. We provide an algorithm that solves both problems. The proposed algorithm consists of two components. The first component (see Section IV-A) finds the optimal resource assignment \mathcal{S} as a function of a hypothetical user selection \mathcal{U} . This procedure is based on dynamic programming, and it is both exact (non-approximate) and computationally efficient. The second component of our algorithm (Section IV-B) finds the optimal user selection \mathcal{U} . The scheme in Section IV-B is an approximate method based on greedy optimization, and it uses the procedure in Section IV-A as a subroutine.

³In LTE, for example, each user has a long ID number of $L_{\text{id}} = 16$ bits length.

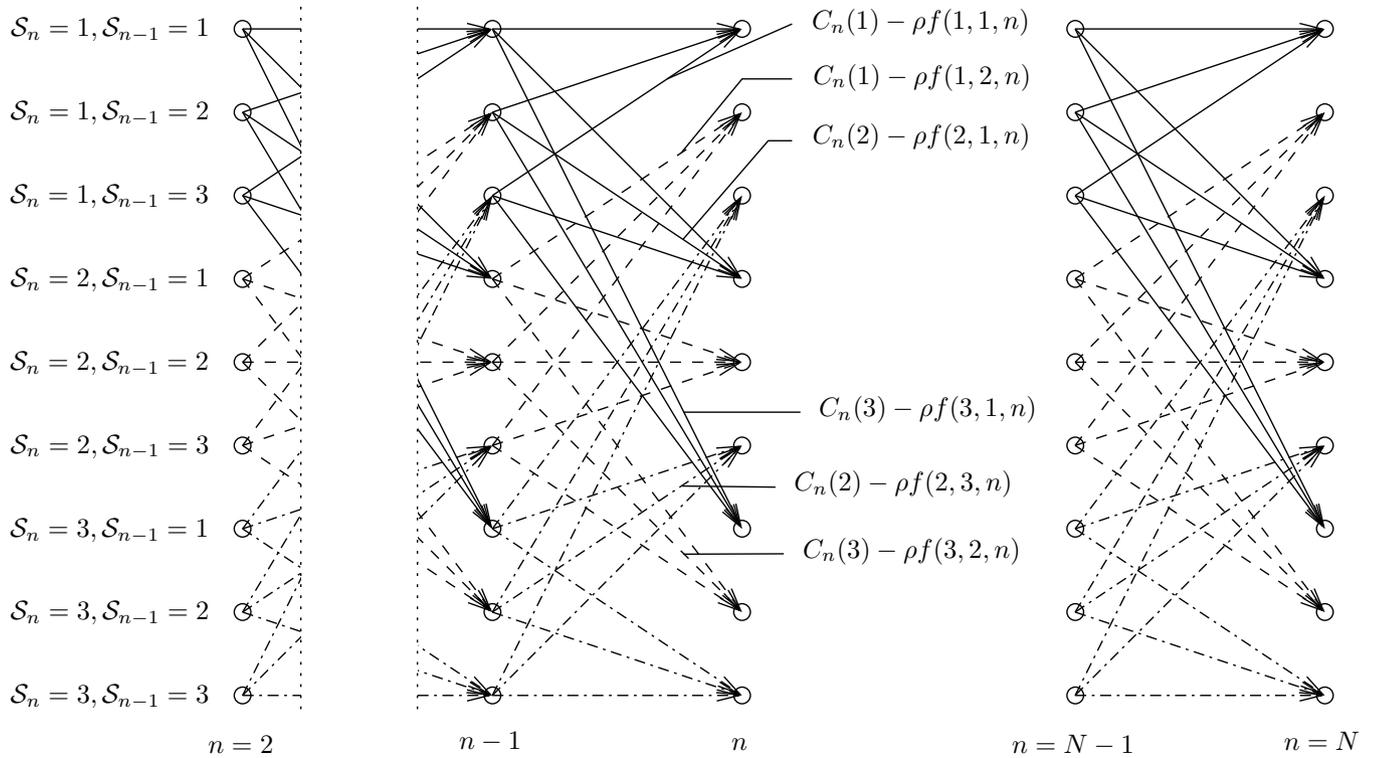


Fig. 2. Illustration of the trellis for the algorithm in Section IV-A, for $|\mathcal{U}| = 3$. The figure only shows excerpts of the trellis. Some randomly chosen branch metrics are also illustrated, using the notation for broadcast transmission mode. Note that the trellis is not terminated at any of its ends.

A. Finding the optimal assignment \mathcal{S} for a given user set \mathcal{U}

Here we show that solving (3) and (7) for given \mathcal{U} (i.e., finding the solution as a function of \mathcal{U}) is a dynamic programming problem, and hence falls into a well-understood area of optimization theory [12]. The key observation is that if several candidate assignments \mathcal{S} exist for which a switch occurs between two users k and k' at resource block n , then only the best of these two solutions needs to be considered for the following resource blocks n' , $n' > n$. The process is reminiscent of maximum-likelihood (in a most-likely sequence sense) decoding of a convolutional code. More precisely one obtains a trellis with $|\mathcal{U}|^2$ states and N instances in “time”, see Fig. 2. Each state corresponds to the pair $\{\mathcal{S}_n, \mathcal{S}_{n-1}\}$ at time n . There are $|\mathcal{U}|$ branches emanating from each state. The metric associated with each branch is

$$C_n(\mathcal{S}_n) - \sum_{k \in \mathcal{U}} \rho_k f_k(\mathcal{S}_n, \mathcal{S}_{n-1}, n)$$

in the individual map transmission scenario and

$$C_n(\mathcal{S}_n) - \rho f(\mathcal{S}_n, \mathcal{S}_{n-1}, n)$$

in the broadcast mode. The fixed penalties R_k and R can be taken care of outside the trellis. Once the trellis has been defined, a backtracing (Viterbi-like) algorithm can be used to find the shortest path through the trellis. The computational complexity of finding the solution is $O(|\mathcal{U}|^2 N)$.

B. Selecting the optimum set of active users \mathcal{U}

The problem of selecting what users \mathcal{U} that should be scheduled is fundamentally difficult. There are 2^K possible sets

\mathcal{U} . If K is very small, then an exhaustive search through all possibilities is feasible. For large K , we propose a suboptimal method, based on greedy optimization [13], as follows. The basic idea is to start with the “best” user, and then keep on adding users as long as this increases the total offered capacity in the frame, taking into account the signaling cost under the assumption that the best resource block assignment is used.

Let $\mathcal{S}^*(\mathcal{U})$ and $\tilde{C}^*(\mathcal{U})$ be the optimum resource block assignment and the resulting penalized capacity, respectively, for a given user set \mathcal{U} . These two quantities are given by the optimum values of $\tilde{C}(\mathcal{S})$ in (4) and (8), and they are computed by the algorithm of Section IV-A. Then the proposed algorithm for selecting \mathcal{U} consists of the following steps:

- 1) Find the “best” user overall:

$$\hat{k} = \operatorname{argmax}_{k \in \{1, \dots, K\}} \tilde{C}^*({k}).$$

Set $m := 1$ and $\mathcal{U}^{(1)} := \{\hat{k}\}$.

- 2) Find the user that would contribute the most to the overall capacity, if it were scheduled:

$$\hat{k} = \operatorname{argmax}_{k \in \{1, \dots, K\}, k \notin \mathcal{U}^{(m)}} \tilde{C}^*(\mathcal{U}^{(m)} \cup k). \quad (9)$$

When computing $\tilde{C}^*(\mathcal{U}^{(m)} \cup k)$ in (9), the actual number of scheduled users, that is $|\mathcal{U}^{(m)} \cup k| = m + 1$, is used to determine the signaling overhead parameters R and Q .

- 3) Check whether

$$\tilde{C}^*(\mathcal{U}^{(m)} \cup \hat{k}) > \tilde{C}^*(\mathcal{U}^{(m)}).$$

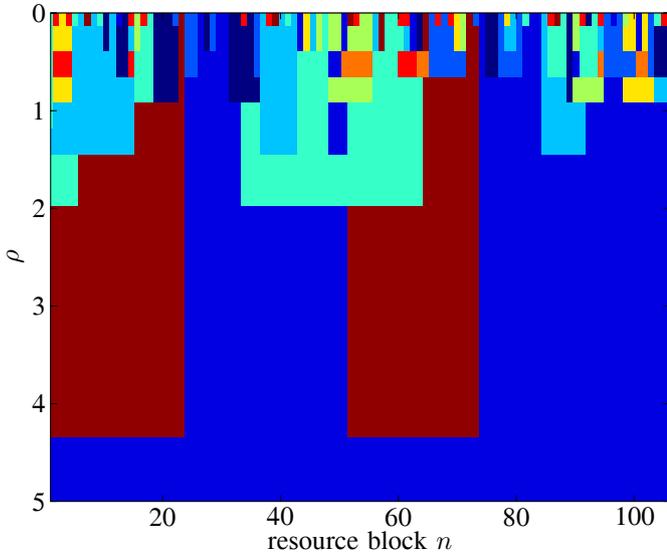


Fig. 3. Illustration of the algorithm operation. The figure shows the scheduling assignments \mathcal{S} as function of ρ , for a setup with $K = 10$ active users, $N = 100$ resource blocks per frame, and randomly chosen channels.

If yes, then the capacity will be increased by adding user \hat{k} . Set $m := m + 1$ and let

$$\mathcal{U}^{(m)} := \{\mathcal{U}^{(m-1)} \cup \hat{k}\}.$$

Then repeat from step 2. If no, then scheduling more users does not help. Then return

$$\{\mathcal{U}^{(m)}, \mathcal{S}^*(\mathcal{U}^{(m)}), \tilde{C}^*(\mathcal{U}^{(m)})\}$$

as the solution and terminate.

V. ILLUSTRATION

Figure 3 illustrates the operation of the algorithm for a setup with $K = 10$ active users, $N = 100$ resource blocks per frame, and randomly chosen channels. The users are color coded and the resulting assignment \mathcal{S} is shown as a function of ρ . The capacity measure $C_n(k)$ is the mutual information assuming Gaussian signaling, at an average SNR of 5 dB. The parameters are set for broadcast mode, for simplicity of the exposition. For ρ near zero, control signaling is not significantly penalized and the scheduler yields a very fragmented map. In this regime, control signaling is “for free”, and even slight variations in the channel are exploited to let users with slightly better channels be scheduled for brief periods in the frequency domain. As ρ increases, the cost of control signaling increases and the scheduling map becomes less and less fragmented. Eventually, the scheduling map is very smooth and very few or even just one single user per frame is scheduled.

Figures 4–6 show some more quantitative results for an LTE-reminiscent OFDMA system with 18 MHz bandwidth and 15 kHz subcarrier separation (1200 subcarriers), and ten OFDM symbols per frame. All results are averages over 5000 random channel realizations. In Fig. 4, the channel is i.i.d. Rayleigh fading. Scheduling is done over resource blocks consisting of 12 neighboring subcarriers. Figure 5 shows the

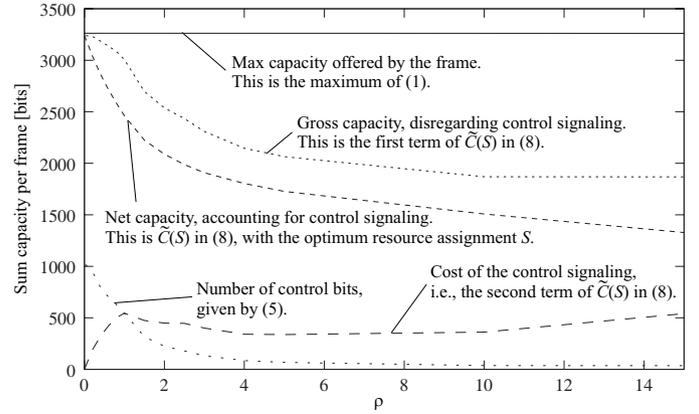


Fig. 4. Average throughput measures (cost functions) as function of ρ , for a sample LTE-reminiscent OFDMA system with 18 MHz bandwidth and 15 kHz subcarrier separation (1200 subcarriers), and ten OFDM symbols per frame. The channel is i.i.d. Rayleigh fading and scheduling is done over resource blocks consisting of 12 neighboring subcarriers. Hence, each resource block has bandwidth $12 \times 15 = 180$ kHz.

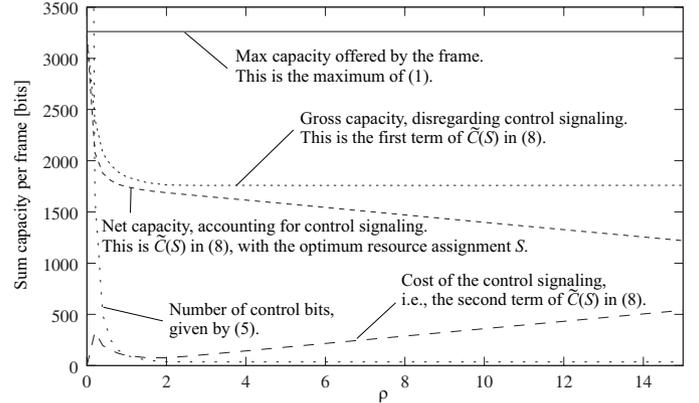


Fig. 5. Same as Fig. 4, but here the scheduling is done over the individual subcarriers, instead of over 12 subcarriers at a time. Hence, each resource block here is equal to one subcarrier and has bandwidth 15 kHz.

same result but here the scheduling is done over the 1200 individual subcarriers, instead of over 12 subcarriers at a time. Compared to Fig. 4, the signaling maps in this scenario are much larger. This is natural. Also, the net and gross capacities fall much faster when ρ increases. This is so because the cost of signaling is much larger, which results in more suboptimal maps. The asymptotic ($\rho \rightarrow \infty$) value of the gross capacity is very slightly smaller than in Fig. 4, since in this regime the scheduler selects the user with the largest average capacity (average over the channel), and the fluctuations around the expected capacity (obtained as $N \rightarrow \infty$) are somewhat less here ($N = 1200$) than in Fig. 4 ($N = 100$), hence resulting in a slight multiuser diversity effect.

Figure 6 shows the same as Fig. 4, but with a Rayleigh fading channel that follows the 3GPP Extended Vehicular A power delay profile. Compared to Fig. 4, the net and gross throughputs here are higher. This is so because the channel is correlated across frequency, so that the users who are scheduled are likely to have a large channel gain at many frequencies. Hence, given that only a small number of users can be simultaneously scheduled (except for at very small ρ), this effect will improve the sum capacity.

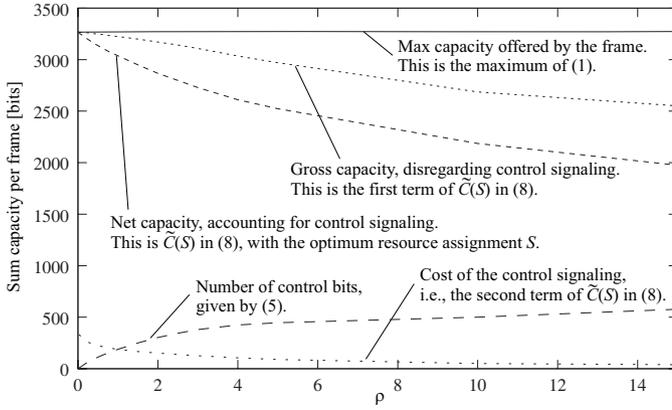


Fig. 6. Same as Fig. 4, but here the channel is Rayleigh fading and follows the 3GPP Extended Vehicular A power delay profile.

Generally, for small ρ , many bits are required for the control signaling but the cost of transmitting them is small. With $\rho = 0$, we obtain the maximum-system throughput scheduler that does not consider the cost of transmitting scheduling information. By contrast, when ρ increases, transmitting scheduling information becomes relatively very expensive, and the maps become simpler and simpler. As $\rho \rightarrow \infty$, the optimal solution is to schedule only one user per frame. The net capacity (accounting for the signaling) is a decreasing function of ρ , since signaling becomes more and more expensive. The gross capacity (not accounting for the signaling) approaches a floor when $\rho \rightarrow \infty$, because it must decrease (the scheduling decisions become less and less optimal from the perspective of maximizing (1)), but it cannot go to zero.

VI. DISCUSSION AND CONCLUSIONS

Our algorithm offers a structured, systematic solution to the problem which can be efficiently computed. In practice, if the total size of the user pool that may come in question for scheduling (K) is large, one may a priori define a hard cutoff threshold, say K^* , on the number of users that may be simultaneously scheduled in one frame. The algorithm in Section IV-B may then be terminated if $m \geq K^*$. Then the procedure in Section IV-A must be run at most KK^* times. The total operation count of the entire algorithm is then of the order

$$K \sum_{k=1}^{K^*} k^2 N \sim O(KK^*{}^3 N).$$

In practice, the appropriate value of the parameter ρ_k (ρ) probably must be chosen by performing system simulations that include precise models for error control of the scheduling information. The point is that once an appropriate value of ρ_k (ρ) is found and agreed on, the proposed algorithm gives an efficient tool for online optimization of the scheduling decisions. Alternatively, if one knows precisely how many bits of signaling that can be afforded, then one can adaptively select ρ_k (ρ) to achieve that target.

For future work, it may be of some interest to extend the model presented here to the case that the scheduling information includes information on power and adaptive modulation. One may also consider the possibility of including a semi-persistent scheduling mode into the model, although it is not obvious how to do that.

REFERENCES

- [1] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution: HSPA & LTE for Mobile Broadband*. Academic Press, 2008.
- [2] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1277-1294, June 2002.
- [3] M. Sternad, T. Svensson, T. Ottosson, A. Svensson, and A. Brunström, "Towards systems beyond 3G based on adaptive OFDM transmission," *Proc. IEEE*, vol. 95, pp. 2432-2455, Dec. 2007.
- [4] Y. Ma and D. I. Kim, "Rate-maximization scheduling schemes for uplink OFDMA," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 3193-3205, June 2009.
- [5] J. Huang, V. G. Subramaniam, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 226-234, Feb. 2009.
- [6] G. Song, Y. Li, and L. J. Cimini, "Joint channel- and queue-aware scheduling multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, pp. 2109-2121, July 2009.
- [7] J. Gross, H. F. Geerdes, H. Karl, and A. Wolisz, "Performance analysis of dynamic OFDMA systems with inband signaling," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 427-436, Mar. 2006.
- [8] J. Gross, P. Alvarez, and A. Wolisz, "The signaling overhead in dynamic OFDMA systems: reduction by exploiting frequency correlation," in *Proc. IEEE International Conf. Commun. (ICC)*, June 2007.
- [9] R. Moosavi, J. Eriksson, E. G. Larsson, N. Wiberg, P. Frenger, and F. Gunnarsson, "Comparison of strategies for signaling of scheduling assignments in wireless OFDMA," *IEEE Trans. Veh. Technol.*, to appear.
- [10] T. Henttonen, K. Aschan, J. Puttonen, N. Kolehmainen, P. Kela, M. Moisio, and J. Ojala, "Performance of VoIP with mobility in UTRA long term evolution," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, May 2008.
- [11] H. Nguyen, J. Brouet, V. Kumar, and T. Lestable, "Compression of associated signaling for adaptive multicarrier systems," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, May 2004.
- [12] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific Publ., 2007.
- [13] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 2009.