Comparison of Strategies for Signaling of Scheduling Assignments in Wireless OFDMA

Reza Moosavi, Jonas Eriksson, Erik G. Larsson, Niclas Wiberg,
Pål Frenger and Fredrik Gunnarsson

Reza Moosavi, Jonas Eriksson, Erik G. Larsson, Niclas Wiberg, Pål Frenger and Fredrik Gunnarsson, Comparison of Strategies for Signaling of Scheduling Assignments in Wireless OFDMA, 2010, IEEE Transactions on Vehicular Technology.
http://dx.doi.org/10.1109/TVT.2010.2066589

Comparison of Strategies for Signaling of Scheduling Assignments in Wireless OFDMA

Reza Moosavi, Jonas Eriksson, Erik G. Larsson, Niclas Wiberg, Pål Frenger and Fredrik Gunnarsson

***Abstract*—This paper considers transmission of scheduling information in OFDMA-based cellular communication systems such as 3GPP long-term evolution (LTE). These systems provide efficient usage of radio resources by allowing users to be scheduled dynamically in both frequency and time. This requires considerable amounts of scheduling information to be sent to the users.**

The paper compares two basic transmission strategies: transmitting a separate scheduling message to each user versus broadcasting a joint scheduling message to all users. Different scheduling granularities are considered, as well as different scheduling algorithms. The schemes are evaluated in the context of the LTE downlink using multiuser system simulations, assuming a full-buffer situation.

The results show that separate transmission of the scheduling information requires a slightly lower overhead than joint broadcasting, when proportional fair scheduling is employed and the users are spread out over the cell area. The results also indicate that the scheduling granularity standardized for LTE provides a good trade-off between scheduling granularity and overhead.

I. Introduction

A. Background and Motivation

In this paper, we are interested in the problem of joint scheduling and transmission of the scheduling information for orthogonal-frequency division multiple access (OFDMA) systems. OFDMA is a powerful multiple-access technique that is used by many forthcoming wireless access systems such as *3GPP Long-Term Evolution* (LTE) and *Worldwide Interoperability for Microwave Access* (WiMax) [2]. OFDMA allows scheduling of users in both frequency and time by assigning them different OFDM subcarriers in different OFDM symbols. In order to obtain a high system throughput, users should be scheduled in the time/frequency slots in which their channel gains are large. Since the channels may change quickly in mobile wireless systems, the scheduling decisions must be made rapidly and frequently in order to best utilize the channel variations. The information about what time/frequency slots that are assigned to each user, along with pertinent information on the transmission parameters such as the modulation format, must then be sent to the users accordingly.

The scheduling of the users eventually results in a so-called *scheduling map* that describes the time/frequency slots that have been assigned to each user. Figures 2 and 3 illustrate such scheduling maps, with different colors representing different users (these figures will be discussed in more detail later). Sending the information contents of this map to all users may require a substantial amount of radio resources. We are interested in efficient ways of conveying the scheduling map along with possibly other relevant information that is associated with this map. Conveying the map entails first appropriately compressing it and then adding error protection. The overall goal is to keep the amount of channel resources required for transmission of the scheduling map small.

The issue of compression, encoding and transmission of the scheduling map leads to a number of intertwined problems. For example, the entire map may be encoded using a single channel code and broadcasted to all users at once. This requires the channel code to have low enough rate so that all users can decode the map without error or with a given (small) error probability. Alternatively, a binary sub-map associated with each specific user (describing whether or not the user is scheduled at a particular location in the time/frequency domain) may be separately transmitted to each user. This has the advantage that the code rate can be chosen on a per-user basis. However this scheme cannot exploit the fact that maps corresponding to different users are correlated. This correlation among individual scheduling maps comes from the fact that users are typically not scheduled in the same time/frequency slot. For example, if user 1 is allocated a certain slot, then it is impossible for user 2 to be scheduled in the same slot (assuming multiuser MIMO is not used).

B. Related Work and Contributions

There is a substantial body of literature on joint time/frequency resource allocation in OFDMA systems. Most of this work deals with algorithms for scheduling and with the inherent trade-off between system throughput (sum-rate) optimization and the notion of fairness. For example, in [3], [4] the resource allocation task is defined as a real-time scheduling problem in which quality-of-service (QoS) requirements are fixed by the application. Therein, QoS is defined in terms of data transmission rate and bit-error-rate (BER). The objective is to minimize the total transmit power by allocating subcarriers to the users and then determining the number of bits that should be transmitted on each subcarrier.

R. Moosavi, J. Ericsson and E. G. Larsson are with the Dept. of Electrical Engineering (ISY), Linköping University, Linköping, Sweden. Email:{reza,joner,egl}@isy.liu.se. N. Wiberg, P. Frenger and F. Gunnarsson are with Ericsson Research, Linköping, Sweden. Email:{niclas.wiberg,pal.frenger,fredrik.gunnarsson}@ericsson.com

This work was supported in part by Ericsson, VINNOVA, 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.

Parts of the material in this paper were presented at the IEEE VTC 2009 conference [1].

In most of the studies that deal with the resource allocation problem, the signaling overhead due to conveying of the scheduling assignments is ignored. We are only aware of relatively little literature addressing the signaling overhead problem. In [5], an algorithm for compression of the control information was proposed. The compression algorithm in [5] consists of a run-length encoder, followed by a universal variable-length code (UVLC). In [6], adaptive coding and modulation were proposed for the transmission of control information. The results show that adaptive coding and modulation can reduce the signaling overhead especially in systems with short-sized data services such as VoIP. Therein, an implicit assumption was that the channel is block fading (slow fading) and remains constant over one frame. However, in these studies the fundamental impact that the signaling overhead causes on the system performance was not considered. In [7], an analytical model for the performance evaluation of OFDMA systems was proposed. In this model, the signaling overhead associated with the control information was taken into account. This model, however, is suitable only for a specific application (VoIP services in IEEE 802.16e OFDMA systems). A solution to decrease the amount of control signaling overhead was proposed within the European-commission sponsored *wireless world initiative new radio* (FP6-WINNER) project [8]. In the proposed method, the users are grouped according to their channel gains, and all users in the same group use the same link adaptation parameters. The transmission parameters are then broadcasted to each group. In [9] an efficient algorithm for the transmission of the multicast sub-MAPs in the IEEE 802.16e systems was proposed. The idea therein is to use adaptive coding and modulation (AMC) for the multicast sub-MAPs without requiring information on the users' channel conditions. The base station adjusts the SNR threshold for the AMC levels based on the knowledge of the previous frame. If a user has decoded the data successfully, then the base station uses a higher AMC level for the current transmission. In [10], a method for scheduling under a constraint on the control signaling complexity was proposed. The method therein maximizes the throughput, taking into account the amount of signaling needed to transmit the scheduling maps to the users.

The paper that is most closely related to our work is [11]. Therein, the effect of the signaling overhead on the system throughput was studied. Reference [11] also proposed an idea to choose new scheduling assignments using the knowledge of the assignments in the previous frame and to change these assignments only if the gain in throughput is larger than the loss due to the signaling overhead caused by the reassignment. Therein, the transmission is done within frames and consists of a downlink transmission phase and an uplink transmission phase. In the case of a reassignment, the control information is broadcasted to all users within the downlink transmission phase. In our work, we are interested in a more general downlink system model. Specifically, the downlink transmission is done within frames consisting of several OFDM symbols. In contrast to the work in [11], the users' channel gains within a frame may change both with time and with frequency. In other words, the channel gains on each subcarrier are not necessarily constant for all OFDM symbols in each frame. In this regard, the presented model can represent radio environments that are changing rapidly. Furthermore, we consider different scheduling policies and also we study different methods for the transmission of the scheduling assignments.

The objective of this paper is to study the fundamental as well as practical limits that exist for the signaling of scheduling information in a general OFDMA system. The specific contributions of our work are:

- We formulate a model for the cost of the transmission of scheduling information, both in terms of spectral efficiency and in terms of actual channel resources spent on this signaling. The model is based on the performance of realistic error-control codes.
- We evaluate the cost of signaling of the scheduling information both for joint (broadcast) transmission and for separate transmission. In doing this, we study three different scheduling algorithms and four different scheduling granularities.
- We formulate a system simulation model that captures all relevant physical phenomena, including path-loss, log-normal shadow fading and fast fading, and perform numerical experiments under this model.

The paper extends our conference paper [1], in that we work with more realistic cost measures for the performance evaluation and for the error-correcting codes involved. Herein we also consider both proportional fair, round-robin and system-throughput maximizing schedulers.

II. System Model and Preliminaries

We consider the downlink of a cellular wireless system. The system consists of a base station surrounded by a random number of users that want to be scheduled for reception of payload data. The total number of users in the cell is N_t and we assume that N_u of these users are requesting service from the base station. For these N_u users, we assume that the data buffers are full, that is, the base station always has data to send to them. Therefore we can compare different scheduling strategies in terms of system sum-throughput.

The base station transmits data in frames. Each frame consists of N_s OFDM symbols with N_c subcarriers and a symbol duration of T_s [seconds]. We assume that each OFDM symbol includes a cyclic prefix of length T_{CP} [seconds]. Thus the total length of each frame is $T_f = N_s T_0$ [seconds] where $T_0 \triangleq T_s + T_{CP}$. The subcarrier spacing is $\Delta f = 1/T_s$ [Hz]. The channel resources in each frame can be thought of as an $(N_s \times N_c)$ grid of time/frequency slots. Each time/frequency slot is called a *resource element*.

We assume that each resource element is assigned to a single user. This creates a correlation between the scheduling information intended for different users. This is so because if a user, say user 1, is scheduled in a certain slot then it is impossible for other users to be scheduled in that slot. This correlation between the scheduling information is exploited by some of our schemes. (In so-called multi-user MIMO schemes, a resource element can be simultaneously assigned to several users. This considerably reduces the correlation between the scheduling information.)

The scheduling decisions are made before the transmission of a frame and the scheduling information is transmitted to the users at the beginning of each frame. This transmission expends radio resources starting with resource element $(1, 1)$, continuing along the frequency domain until the whole OFDM symbol is used up and then starts over with the next OFDM symbol and so on.

We assume that the propagation channels for all users remain constant over one resource element. Therefore, we can express the channel gain for user k in the (i, j)th resource element by a dimensionless complex scalar $h_{i,k}^{(k)}$. Let $\boldsymbol{H}^{(k)} \triangleq \left\{ h_{i,j}^{(k)} \right\}$, $i = 1, 2 \ldots, N_s$ and $j = 1, 2, \ldots, N_c$, be the $(N_s \times N_c)$-matrix that contains all channel gains for user k. We furthermore assume that the base station has a total transmit power budget of P [W], that is P/N_c [W] per subcarrier.

We define two quantities that will be frequently used throughout the paper:

- The *scheduling map (matrix)* $\boldsymbol{U}$ contains the identity numbers of the users that are scheduled in each of the $N_s \times N_c$ resource elements. Specifically, $\boldsymbol{U} = \{u_{i,j}\}$, where $u_{i,j}$ is an integer from the set $\{1, 2, \ldots, N_t\}$ representing the index of the user that has been granted the resource element (i, j).
- The *effective channel matrix* $\boldsymbol{S}$ is the effective channel to the scheduled users, as seen by the base station. More precisely, it is defined as

$$\boldsymbol{S} \triangleq \{s_{i,j}\} = \left\{ h_{i,j}^{(u_{i,j})} \right\}.$$

To compare the performances of different strategies that we explore in this study, we define two system performance measures:

(i) The **signaling overhead ratio** Σ is the number of resource elements that need to be set aside for conveying the scheduling map in relation to the total $N_s N_c$ resource elements.

(ii) The **system spectral efficiency** $C(\boldsymbol{S})$ is the total spectral efficiency of the transmission averaged over all the resource elements that are used for transmission of payload data. We define $C(\boldsymbol{S})$ in [bits/s/Hz] as

$$C(\boldsymbol{S}) \triangleq \frac{T_s}{T_0} \frac{1}{N_s N_c} \sum_{(i,j) \in \mathcal{T}} \log_2 \left(1 + \frac{|s_{i,j}|^2 P_{i,j}}{N_0 \Delta f} \right) \quad (1)$$

where N_0 is the noise power spectral density [W/Hz][1] and $P_{i,j}$ [W] is the transmit power used during resource element (i, j). Also $\mathcal{T}$ denotes the set of resource elements that are used for the transmission of payload data. Note that the resource elements that may be needed to convey the scheduling map are omitted from the summation in (1). The T_s/T_0 factor in (1) accounts for the loss in spectral efficiency due to the cyclic prefix. It is worth mentioning that, there is generally also a need for a few unused guard subcarriers which would additionally reduce the spectral efficiency. However we neglect this loss in the following analysis. Furthermore, since one part of the resource elements are set aside for the signaling of the scheduling map, the scheduling decision of the remaining slots may be incorrect given the scheduling strategy in question. The scheduling should actually be redone given the new set of payload data resource elements, resulting in a new scheduling map which would consume a different set of resource elements for its signaling, and so forth. However, in our analysis here we disregard the small discrepancies in the scheduling map that results from this effect.
In (1), $\log_2 \left(1 + \frac{|s_{i,j}|^2 P_{i,j}}{N_0 \Delta f} \right)$ represents the amount of mutual information that flows from the base station to the scheduled user during resource element (i, j). The averaging in (1) should be thought of as an approximation to the expectation operator that appears in the definition of ergodic capacity [12]. When referring to ergodic capacity, we make the implicit assumption that there exists a capacity-achieving coding scheme that codes across the resource elements. The rate $C(\boldsymbol{S})$ is not achievable in practice. However the $\log(1+\text{SNR})$ formula is often a useful measure of the system performance anyway, since the throughput of most adaptive coding and modulation schemes behaves as $\log(1 + \xi\text{SNR})$ for some ξ where ξ determines the performance gap to the capacity limit [13].

The choice of powers $P_{i,j}$ that maximize $C(\boldsymbol{S})$ can be formulated as an optimization problem subject to a total power constraint P. For a noise limited system (cf. footnote 1), the solution to this problem can be easily found via standard waterfilling [12]. In some systems such as LTE, the powers $P_{i,j}$ are taken to be equal for all (i, j) [2]. In fact, equal power allocation is nearly optimal provided that the scheduler always selects users that have reasonably high signal-to-noise-ratios (SNR) [14]. We have verified this via simulations and the performance gap between constant power allocation and the optimal power allocation using waterfilling is negligible in the cases of interest. For this reason, we will use an equal power distribution over all resource elements; that is we take $P_{i,j} = P/N_c$.

III. Scheduling Granularity

The *scheduling granularity* determines how small part of the channel resources that can be allocated to a specific user. There are in total $N_s N_c$ resource elements that may potentially be assigned to different users in each frame. As we will see later, assigning each resource element to the users individually can require a substantial amount of channel resources for conveying the scheduling maps. A common approach to keep the required amount of channel resources small is to lump resource elements together into bigger entities and assign each such scheduling entity to one user. We call such a scheduling entity a *scheduling block*. The granularity, therefore, determines how many resource elements that are aggregated into one scheduling block. We will consider four different scheduling granularities:

[1] We consider a noise-limited system throughout the paper. To include co-channel interference in the model, N_0 should be replaced with $N_0 + I_{i,j}$, where $I_{i,j}$ denotes the interference power over resource element (i, j).

- *Finest granularity:* Here each scheduling block consists of one single resource element. That is, each of the $N_c N_s$ resource elements can be assigned to a different user.
- *Frequency-only scheduling:* In this case, users are scheduled only in frequency. Each subcarrier is assigned to a single user during the whole frame. Since there is no scheduling in time, the scheduling matrix $\boldsymbol{U}$ reduces to a vector of length N_c. The corresponding scheduling block consists of N_s resource elements.
- *Frequency-aggregated granularity:* In order to further decrease the scheduling granularity, we aggregate N_f consecutive OFDM subcarriers in frequency and assign each such aggregated frequency block to one user during the entire frame. In other words, each scheduling block consists of $N_f N_s$ resource elements corresponding to N_f subcarriers in frequency and N_s OFDM symbols in time.
- *Frame-wise scheduling:* Here the entire frame is allocated to one single user. The resulting scheduling block consists of all $N_c N_s$ resource elements. The scheduling matrix $\boldsymbol{U}$ in this case consists of only one integer.

In order to support the scheduling strategies (see Section IV), the terminals need to provide the base station with channel-quality indicator (CQI) information. Therefore there is a need for feedback signaling of CQI information. We assume that for each scheduling block, each user sends one CQI value representing the throughput that she can obtain if she were scheduled in that scheduling block. More precisely, we model the ℓth CQI report (corresponding to the scheduling block ℓ) of user k as

$$r_\ell^{(k)} \triangleq \frac{1}{|\mathcal{B}_\ell|} \sum_{(i,j)\in\mathcal{B}_\ell} \log_2\left(1+\gamma_{i,j}^{(k)}\right) \tag{2}$$

where $\mathcal{B}_\ell$ denotes the set of all resource elements in scheduling block ℓ, $|\mathcal{B}_\ell|$ is the size of the corresponding scheduling block (which depends on the scheduling granularity) and

$$\gamma_{i,j}^{(k)} \triangleq \frac{\left|h_{i,j}^{(k)}\right|^2 P_{i,j}}{N_0 \Delta f}$$

is the instantaneous received SNR for the kth user in the resource element (i,j).

Table I summarizes the four different scheduling granularities along with the corresponding number of CQI values needed per frame for each user. Note that when we go to the coarser granularities, the amount of CQI transmitted over the feedback channel will decrease.

IV. Scheduling Strategies

In scenarios with full buffers, the scheduler must trade off between two conflicting objectives: to maximize the overall system throughput and to guarantee fairness among the users. We will study three different scheduling strategies: (i) system-throughput maximizing, (ii) round-robin and (iii) proportional fair. With the exception of round robin, the schedulers that we consider require the knowledge of the channel gains for all N_u active users, before making the scheduling decisions. In what follows we describe the scheduling strategies that we consider in the paper.

A. System-Throughput Maximizing Scheduler

The system-throughput maximizing scheduler (also known as the max-C/I or maximum sum-rate scheduler [2]) achieves the maximum sum-throughput by scheduling the user with the best channel in each scheduling block. Therefore the scheduling block ℓ is assigned to the user that supports the maximum throughput i.e. the user that has reported the highest CQI.

Since in a cellular environment the channel variations are typically independent between the users, there is almost always a user for which the fast fading is near its peak. This phenomenon is known as multiuser diversity in the literature [2], [12]. The larger the number of users in the cell, the more likely it is that one of the users has a good channel in a given scheduling block and consequently the larger the benefit from the multiuser diversity effect would be.

The max-C/I scheduler favors users with large average channel gains. In a cellular system these are the users that are located close to the base station. Users that are far from the base station are less likely on the average to be selected by this scheduler. Therefore the max-C/I scheduler is not fair in general which makes it unattractive for practical systems.

B. Round-Robin Scheduler

The round-robin scheduler lets the users take turns in using the channel resources, without taking the instantaneous channel gains into account [2]. Since the channel resources are evenly divided among the users, the round-robin scheduler is fair in the sense that all users get the same amount of the channel resources. However, it is not fair in the sense of providing the same average throughput to the users. The reason is that users at different distances from the base station have different average channel gains.

Since the round-robin scheduler ignores the instantaneous channel conditions, the effective radio link between the scheduled user and the base station will occasionally be poor. Thus the overall system throughput with the round-robin scheduler is lower than that of the max-C/I scheduler.

C. Proportional Fair Scheduler

The proportional fair (PF) scheduler [15], [16] provides a trade-off between maximizing the average sum-throughput and providing fairness to the users. The scheduling block ℓ is assigned to the user with the highest *priority*, where priority is defined as

$$\mathcal{P}_k \triangleq \frac{r_\ell^{(k)}}{T_k(t)}. \tag{3}$$

In (3), $T_k(t)$ represents the throughput of user k averaged over a time window in the past. Moreover, $r_\ell^{(k)}$ defined in (2) represents the instantaneous rate (mutual information) that user k can get in scheduling block ℓ. The average throughputs $T_k(t)$ are kept constant over all resource elements in the frame.[2]

[2] Note that this does not imply that one user gets all the resources in the frame.

		OFDM Symbols	
		1	N_s
OFDM Subcarriers	1	Finest Granularity ($N_s N_C$ CQI values)	Frequency-only (N_c CQI values)
	N_f	-	Frequency-aggregated (N_c/N_f CQI values)
	N_c	-	Frame-wise (1 CQI values)

TABLE I
SCHEDULING GRANULARITIES AND ASSOCIATED NUMBER OF CQI REPORTS PER FRAME

Once the scheduling decision has been made for the entire frame, the average throughputs are updated according to

$$T_k(t) = \left(1 - \frac{1}{t_f}\right) T_k(t-1) + \frac{1}{t_f} T_{k,\text{tot}} \tag{4}$$

where $T_{k,\text{tot}}$ is the total throughput that user k gets in the frame, that is

$$T_{k,tot} \triangleq \sum_{\ell \in \mathcal{T}_k} r_\ell^{(k)}. \tag{5}$$

In (5), $\mathcal{T}_k$ denotes the set of all scheduling blocks assigned to user k in this frame. Also $t_f \in \{1, 2, \ldots\}$ represents the length of the averaging window which should be chosen large enough so that the scheduler can exploit the variations in the instantaneous channel conditions but small enough not to starve users with poor channel conditions [16]. Choosing a large t_f exploits more multiuser diversity but requires some users to wait longer before they are scheduled, therefore increasing transmission delays. Using a small t_f yields a lower average system sum-throughput but shorter delays.

V. SIGNALING OF THE SCHEDULING ASSIGNMENTS

The scheduler produces a matrix $\boldsymbol{U}$ which must be conveyed to the users. This signaling of scheduling assignments consumes channel resources. A natural objective is to keep the amount of channel resources needed for this signaling as small as possible. In some systems such as LTE, the first few OFDM symbols in the frame are dedicated to a so-called *control region* which is used for transmission of the control signaling information [2]. In this paper we assume that the size of the control region can dynamically change. This assumption gives us the opportunity to compare the efficiency of different scheduling and signaling strategies. Herein, we only consider the part of the control signaling which concerns the resource allocation (scheduling) of different users.

In order to facilitate the decoding of the control information, we need to reduce the granularity of the control region (see Sections V-C and V-D). Therefore we aggregate several channel resources in frequency into bigger blocks and we call each such block a *control channel element* (CCE). Thus the control region consists of several CCEs.

We will study two methods for conveying the scheduling information. In the first approach, the scheduling information is first compressed and then broadcast to all users. This requires the channel code to have low enough rate so that all users can decode the map with a given (small) error probability. In the second approach, the information is sent to each of the scheduled users individually. While this scheme does not exploit the correlation among the individual scheduling maps, it has the advantage that the code rate can be chosen on a per-user basis. The choice of these two schemes for this study is motivated by an interest in understanding the fundamental aspects of the signaling problem. An optimal signaling strategy may consist of a combination of both these schemes.

Before we proceed to explore these two methods in more detail, we first present the method that we use to compress the scheduling maps, and a model for obtaining the transmission rate at a given SNR and for a given probability of error.

A. Compression of Scheduling Maps

We use run-length encoding [17] as the compression method. The main motivation for this is that run-length encoding does not require any statistics of the source. Thus it does not rely on any specific a priori assumptions or statistics that are hard to estimate from small amounts of data. Run-length encoding is a good compression tool for short data blocks and therefore it is especially powerful for the coarser granularity cases (frequency-only scheduling and frequency-aggregated granularity, in particular). Additionally, it has low implementation complexity.

Let $\boldsymbol{v}$ be a vector of length N consisting of the symbols s_m, $m = 1, 2 \ldots, N$ from the alphabet $\mathcal{S}$ of cardinality M.[3] Let ℓ_i be the length of the ith symbol-run and let q be the total number of symbol-runs in the vector $\boldsymbol{v}$. Thus,

$$\sum_{i=1}^{q} \ell_i = N.$$

There are M different symbols and therefore we use $\lceil \log_2(M) \rceil$ bits per symbol-run i in the run-length code to describe the symbol value. Since the length of the vector is N, the maximum possible length for the jth symbol-run is $N - \sum_{i=1}^{j-1} \ell_i$. Therefore, we represent the jth symbol-run with $\left\lceil \log_2 \left(N - \sum_{i=1}^{j-1} \ell_i \right) \right\rceil$ bits. Hence, by defining $\ell_0 = 0$, we can express the vector $\boldsymbol{v}$ with

$$N_b = q \lceil \log_2(M) \rceil + \sum_{j=1}^{q} \left\lceil \log_2 \left(N - \sum_{i=0}^{j-1} \ell_i \right) \right\rceil \tag{6}$$

[3] As we will see later, the choice of M depends on the control signaling strategy.

bits. The first term in (6) corresponds to the number of bits required to represent the symbols and the second term corresponds to the number of bits required to represent the lengths of the runs.

Note that for the frame-wise scheduling scheme, only one user is scheduled in the entire frame. Hence, in this case we only need to send the index of the scheduled user. That is, $\lceil \log_2(N_t) \rceil$ bits are needed to represent the scheduling map.

B. Model for Transmission of the Compressed Scheduling Information

We model the performance of the error correction scheme for the transmission of the scheduling information via a lookup table. This table gives the effective achievable rate as a function of the average SNR and of the number of information bits, at a given error probability. The lookup table was generated via extensive numerical simulations as explained in what follows.

Assume that we want to transmit a short block of information with a block-error-rate (BLER) below a target value P_e. A natural solution to meet this BLER requirement is to use an error-correcting code. When choosing the code, the overall goal is to keep the number of coded bits as small as possible. There are many possible error-correcting codes. We will use (punctured) convolutional codes with soft-decision Viterbi decoding [18]. While these codes are not capacity-achieving, they work well for the short block lengths encountered in our application. More sophisticated codes such as low-density-parity-check (LDPC) or turbo codes can operate much closer to the Shannon limit, but they are only suitable for long information blocks [19], [20].

The goal is to encode the signaling information into codewords that span over several resource elements. In practice, depending on the coherence time and the coherence bandwidth of the channel, the effective channel for this transmission will be block fading and it will exhibit more or less variations within a codeword. We will consider the two extreme cases of a stationary (AWGN) channel and of a fast Rayleigh fading channel. The first case corresponds to a stationary channel (infinite coherence time and bandwidth) and the second case corresponds to very rich frequency and time diversity (small coherence time and bandwidth). Since the channels in practice are wideband and offer a substantial amount of diversity, we will use the fast Rayleigh fading results for the system simulations in Section VII.

In LTE, the link adaptation for the control region is done through a rate-matching mechanism [2]. That is, the modulation scheme used for the transmission of the control information is fixed (QPSK), but the channel code is chosen based on the instantaneous channel condition. For simplicity, and in order to be consistent with the LTE standard, we will assume that the modulation scheme is QPSK and choose the channel code as a function of the channel condition (a code with high rate when the channel is good and vice versa).

For a given length of the information-bearing message, say N_b, and a given average SNR, we find the maximum achievable rate that meets the target BLER from a set of convolutional codes consisting of about 60 different codes with rates varying from 1/8 to 7/8 and constraint lengths K varying from 3 to 9. We then store the result in a lookup table. For trellis-based coding schemes, to ensure that the encoder returns to the all-zero state, we must append $(K-1)$ zeros to the information bits. Since the number of information bits is small, this may significantly affect the rate. Therefore we define the *effective rate* of the code as

$$\tilde{r}_c \triangleq r_c - \frac{K-1}{\nu} \tag{7}$$

where r_c is the rate of the base convolutional code and $\nu \triangleq \left\lceil \frac{N_b+K-1}{r_c} \right\rceil$ is the number of coded output bits.

In our model, the rates of the convolutional codes can vary from 1/8 up to 7/8. Let γ_0 and γ_t be the SNRs required to achieve the target BLER for the rate-1/8 and rate-7/8 codes, respectively. To extend the lookup table to arbitrary SNRs, we assume that code rate can never exceed 7/8, no matter how large the SNR is. That is, the rate is 7/8 for SNR $\geq \gamma_t$. For SNR γ below γ_0, we obtain the rate by extrapolating the table according to the linear formula

$$\tilde{r}_c = \eta\gamma \tag{8}$$

where the constant η is chosen such that (8) gives the same result as the simulation for rate 1/8 does. Equation (8) does not aspire to be an accurate model for the actual transmission, but we use this model in our numerical studies to obtain reasonable results for very low SNRs. The linear relation in (8) can be derived by assuming that the coding scheme consists of a concatenation of the convolutional coding with repetition coding at low SNRs.

Figure 1 illustrates the effective code rate versus SNR for block lengths $N_b = 10$ and $N_b = 100$. Since the average probability of a missed downlink scheduling grant is below 1% in the LTE standard [21], we assume a target BLER of $P_e = 0.01$ for the results presented in Figure 1. We can see that increasing N_b from 10 to 100 gives a small improvement in rate. Also, the effective code rate is higher for the AWGN channel than for the fast Rayleigh fading channel at a given SNR. We can also notice that for a block length of $N_b = 100$, the achievable effective rate is higher than when using a block length of $N_b = 10$ for the same base convolutional code. This can be readily understood from (7). In what follows, we only use the fast Rayleigh fading results, as discussed above.

*C. Joint Compression, Encoding and Broadcast (*JCEB*) Scenario*

The next step is to model the transmission of the scheduling information. As briefly discussed earlier, we study two methods namely joint compression, encoding and broadcast (JCEB) and separate compression, encoding and transmission (SCET). Figure 2 illustrates the JCEB scheme. Herein the scheduling information is encoded into one single codeword and broadcasted to all the users. The code used must be strong enough so that all scheduled users, especially the one with the weakest channel, can decode the information with low error probability. As discussed earlier, we assume that the modulation scheme used to transmit the scheduling

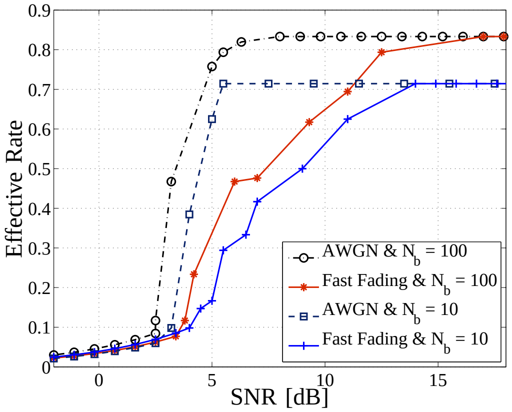


Fig. 1. Empirical effective code rates ($\tilde{r}_c$) for transmission of the control information using QPSK modulation and punctured convolutional codes.

information is QPSK, and that the lookup table described in Section V-B is used to determine the effective code rate for a given SNR. The choice of the code depends both on the SNR and on the number of information bits. Therefore, the code that is used for error correction is generally different in different frames. In order for the users to know what code that was used, we add a cyclic redundancy check (CRC) code with N_{CRC} bits to the codeword. This way all users can blindly decode the incoming scheduling information by trying different convolutional code candidates, and for each one check whether the CRC is satisfied. If the CRC passes, then the user determines that the corresponding channel code was actually used.

Recall that in the scheduling map $\boldsymbol{U}$, users are identified by a number from the set $\{1, 2, \ldots, N_t\}$. In order to decrease the number of bits required to represent the scheduling map, we first list the scheduled users and create a temporary identification number for each one consisting of their indexes in this list. Let N_{sched} denote the number of users that have been granted resources for transmission in the frame. Each user can then be identified by an integer from the set $\{1, 2, \ldots, N_{\text{sched}}\}$. Therefore, we need $\lceil \log_2(N_{\text{sched}}) \rceil$ bits to represent the temporary identity of the scheduled users. However, the list of identity numbers should first be broadcast to the users so that each user knows her temporary identifier. We assume that the length of the users' identity numbers is $N_\beta \triangleq \lceil \log_2(N_t) \rceil$. Note that in many real systems, users are assigned a long identity number. For example in LTE standard, the identity number is 16 bits long [22]. However it may be easily shortened to N_β bits where $N_\beta \ll 16$, for example by retaining only specific digits [23] or by using a predefined hash table.

Under these assumptions, we can now express the required number of resource elements for the signaling of the scheduling information in the JCEB scenario as

$$N_{\text{JCEB}} \triangleq \left\lceil \frac{N_{\text{sched}} N_\beta + N_{\text{map}} + N_{\text{CRC}}}{2\tilde{r}_c} \right\rceil . \tag{9}$$

In (9), $N_{\text{sched}} N_\beta$ is the number of bits required to transmit the user list, N_{map} is the number of bits required to represent the scheduling maps obtained from the compression scheme, and N_{CRC} is the length of the cell-specific CRC code. Also $\tilde{r}_c$ is the code rate which—as discussed earlier—is obtained from the lookup table. The factor two accounts for the fact that each QPSK symbol can carry two bits. Note that the code should be strong enough so that all users can decode the information with a given probability of error. Therefore we choose the rate based on the average SNR for the user with the weakest channel say $\tilde{\gamma}_k$:

$$\tilde{\gamma}_k \triangleq \min_k \left\{ \mathbb{E}_h \left[|h_{i,j}^{(k)}|^2 \right] \frac{P}{N_c N_0 \Delta f} \right\} \tag{10}$$

where $\mathbb{E}_h \left[|h_{i,j}^{(k)}|^2 \right]$ is the average channel gain for user k which is obtained in practice from a sample-mean estimator. In order to be precise, the average channel gain should be taken only over the resource elements that have been assigned to the control region. However, since the size of the control region can dynamically vary from frame to frame and since in the practical systems such as LTE [2], the control information is interleaved in the frequency domain, we will take the average over the entire frame. Finally the number of required CCE with the JCEB scheme is

$$N_{\text{CONT}} = \left\lceil \frac{N_{\text{JCEB}}}{N_{\text{CCE}}} \right\rceil \tag{11}$$

where N_{CCE} denotes the size of control channel elements.

*D. Separate Compression, Encoding and Transmission (*SCET*) Scenario*

In this approach, the scheduling information is sent to each user separately. See Figure 3. Since a user only needs to know which resource elements that have been assigned to her, there is no need to send the whole scheduling matrix $\boldsymbol{U}$ to the all users. Instead, for each scheduled user k, we first define a user-specific scheduling matrix $\boldsymbol{U}^{(k)} \triangleq \left\{ u_{ij}^{(k)} \right\}$ as

$$u_{ij}^{(k)} \triangleq \begin{cases} 1 & \text{if } u_{ij} = k \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The matrix $\boldsymbol{U}^{(k)}$ simply identifies the resource elements that have been assigned to the kth user. Each matrix $\boldsymbol{U}^{(k)}$ is then separately compressed and encoded, resulting in a codeword of $N_{\text{map}}^{(k)}$ bits. Since the matrices consist of only zeros and ones, the alphabet $\mathcal{S}$ in Section V-A is binary and consequently we only need to encode the lengths of the symbol-runs.

In order to find the required number of CCEs with the SCET scheme, we start with the first scheduled user and determine the number of CCEs needed for the transmission of $N_{\text{map}}^{(1)}$ bits corresponding to the scheduling matrix $\boldsymbol{U}^{(1)}$. In order to distinguish between the users, we add a user-specific CRC to each user's data. Therefore the number of required resource elements for user 1 is

$$N_{\text{SCET}}^{(1)} = \left\lceil \frac{N_{\text{map}}^{(1)} + N_{\text{CRC}}}{2\tilde{r}_{c_1}} \right\rceil$$

where $\tilde{r}_{c_1}$ is the code rate for user 1, which is obtained from the lookup table based on her average received SNR over the entire frame as in the JCEB scheme. This implies that we

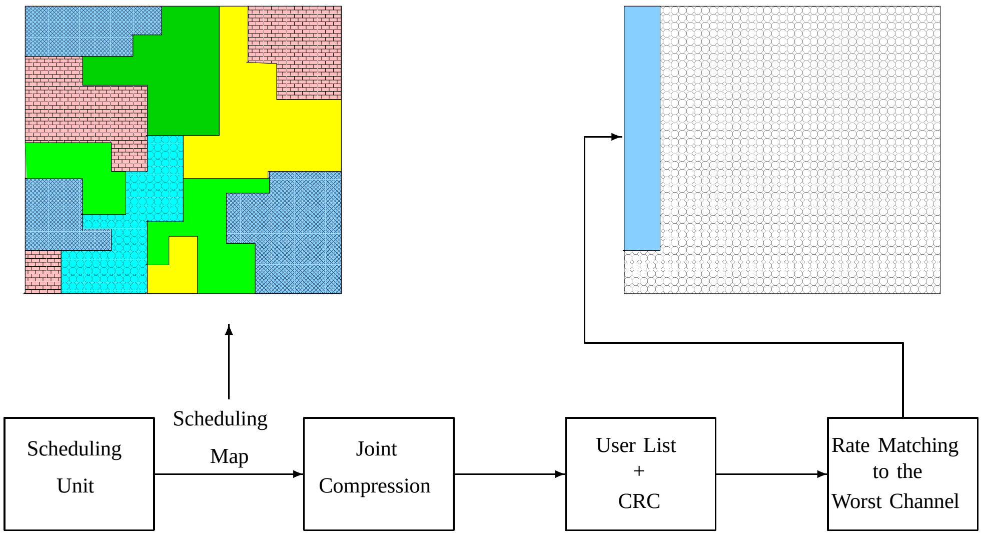


Fig. 2. Illustration of the JCEB scheme for the finest granularity case. After the scheduling decision has been made, the scheduling map is jointly compressed. Then scheduling list and the CRC are inserted and the resulting bits are protected by a channel code which is adapted to the user with the worst channel. The bits are then QPSK modulated and mapped to the first few OFDM symbols in the frame.

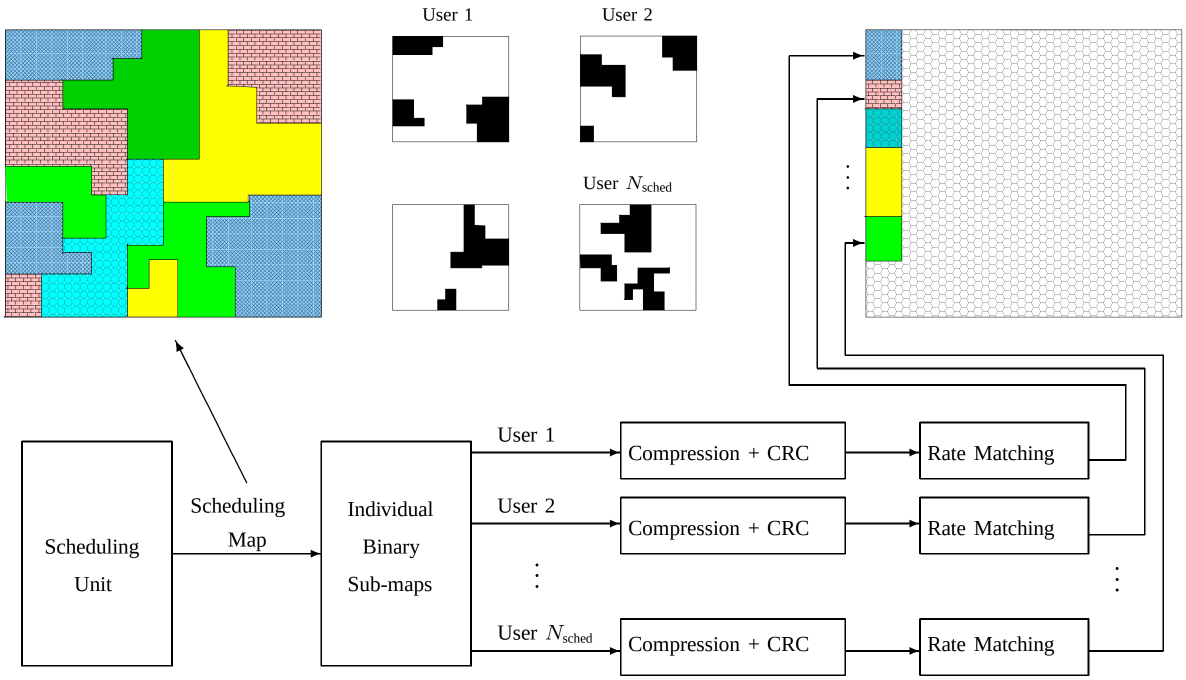


Fig. 3. Illustration of the SCET scheme for the finest granularity case. After the scheduling decision has been made, we find the binary sub-map associated with each user. The individual maps are then compressed and transmitted to the users separately.

need

$$N_{\text{CONT}}^{(1)} = \left\lceil \frac{N_{\text{SCET}}^{(1)}}{N_{\text{CCE}}} \right\rceil$$

CCEs to transmit the control information to user 1. We continue the same procedure for all the scheduled users. Hence the size of the control region (in terms of the required number of CCEs) is

$$N_{\text{CONT}} = \sum_{k=1}^{N_{\text{sched}}} N_{\text{CONT}}^{(k)}. \tag{13}$$

In order to find her own scheduling information, each user needs to blindly decode the incoming information with all possible codes and for all possible combinations of locations of the signaling data in the control region (different combination of the consecutive CCEs). The use of CCE makes the start and the end position of the control information subject to a certain structure. This will reduce the number of blind attempts by a terminal. A similar technique is used in LTE for the decoding of the downlink control information [2].

E. Remark on Error Probabilities of the Scheduling Information

Recall that the codes are chosen such that each user can decode the scheduling information with a given error probability P_e. Let $\mathcal{E}_i$ be the event that the ith user fails in this decoding. The probability that at least one user fails is

$$\Pr\{\mathcal{E}\} = \Pr\{\mathcal{E}_1 \cup \mathcal{E}_2 \cup \ldots \mathcal{E}_{N_{\text{sched}}}\} \tag{14}$$

which can be upper bounded by

$$\Pr\{\mathcal{E}\} \leq \sum_{i=1}^{N_{\text{sched}}} \Pr\{\mathcal{E}_i\}. \tag{15}$$

For the SCET scenario, the codes are chosen individually and independently for each user, so all $\Pr\{\mathcal{E}_i\}$ are equal to P_e. Therefore the probability of error for the entire system with SCET scenario is roughly

$$\Pr\{\mathcal{E}_{\text{SCET}}\} \leq N_{\text{sched}} P_e. \tag{16}$$

For the JCEB scenario, a single code is used for the encoding of the scheduling map. Since the code is chosen for the user with the poorest channel, $\Pr\{\mathcal{E}_i\}$ might be (much) smaller than P_e for some users. Therefore with the JCEB scenario, the chance that at least one user fails is generally smaller than in the SCET scenario and thus the system-throughput loss due to decoding failures of the scheduling information is lower for the JCEB scenario than for the SCET scenario at a given probability of error P_e. In the numerical results presented in Section VII we do not consider the effect of P_e on the system spectral efficiency.

VI. System Simulation Model

To simulate a realistic cellular environment, we create a so-called *scenarios* where for each scenario, we assume that the number of active users N_u is drawn from a binomial distribution with parameters N_t and p_u. That is, on the average $\mathbb{E}[N_u] = p_u N_t$ users out of the total N_t potential users are requesting service from the base station. The users are uniformly located in a circular cell area and this area is bounded by an inner radius R_0 and an outer radius R_c. The purpose of limiting the minimum distance to the base station to R_0 is to ensure that the path loss model (see below) is used only in a regime where it is valid.

We model the physical wireless channel in terms of path-loss, large-scale fading and small-scale (multipath) fading. The path-loss models the attenuation of the signal due to propagation distance. We model it via the multiplicative factor $(r/R_0)^{-\alpha}$, where r is the distance to the base station, α is the path-loss exponent and R_0 is a reference distance (that coincides with the inner radius of the cell). This path-loss model is valid for $r \geq R_0$. The large-scale fading models shadowing by large objects. We model it via a multiplicative factor $10^{\frac{\chi}{10}}$ where χ is a normally distributed random variable with zero mean and variance σ^2. We assume that the path-loss and the large-scale fading factors remain constant over time and frequency. The small-scale fading is due to the constructive and destructive interference between multiple signal paths between the base station and the user. We model small-scale fading by using a tapped-delayed line model for the channel impulse response

$$\sum_i a_i(t)\delta(t - \tau_i(t)). \tag{17}$$

The tap coefficients $a_i(t)$ are assumed to be independent Rayleigh fading stochastic processes with a Jakes Doppler spectrum. Thus, the overall channel frequency function for user k can be expressed as

$$\mathcal{H}^{(k)}(f,t) = \sqrt{(r_k/R_0)^{-\alpha}\, 10^{\frac{\chi}{10}}} \sum_i a_i^{(k)}(t) e^{-j2\pi f \tau_i^{(k)}(t)}. \tag{18}$$

Since the communication method is OFDM, we are interested in the channel coefficient for a specific resource element (i,j). For user k this coefficient is given by

$$h_{i,j}^{(k)} = \mathcal{H}^{(k)}\left(\left(j - \frac{N_c + 1}{2}\right)\Delta f, (i-1)T_0\right). \tag{19}$$

We next define the *system operating point*. Note first that the average (over the small-scale fading) SNR received at a given distance r from the base station is a random variable, since it depends on the shadow fading. Denote this random variable by $\text{SNR}(r)$. Next, let $\overline{\text{SNR}}(r) \triangleq E[\text{SNR}(r)]$ be its average, where the expectation is taken over the shadow fading. Also, let $\text{SNR}_\beta(r)$ be the βth percentile of $\text{SNR}(r)$. That is, on the average a fraction β of the users at distance r experience an SNR of at least $\text{SNR}_\beta(r)$. We take the system operating point to be the value of $\text{SNR}_\beta(r)$ at the cell border, i.e., $\text{SNR}_\beta(R_c)$. It can be easily shown that the average received SNR at distance r obeys the following relation

$$\begin{aligned} 10\log_{10}\left(\overline{\text{SNR}}(r)\right) =& 10\log_{10}\left(\text{SNR}_\beta\left(R_c\right)\right) + \\ & 10\alpha\log\left(R_c/r\right) - \sigma \mathrm{Q}^{-1}\left(\beta\right) \end{aligned} \tag{20}$$

where $\mathrm{Q}(x)$ is the Gaussian error integral (Q-) function, defined as

$$\mathrm{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t.$$

VII. Numerical Results

We performed Monte-Carlo simulations for different scheduling strategies and for different scheduling granularities. All presented results are obtained by averaging over N_{SCEN} independent scenarios. We considered three different system setups:

- **Model I:** In this model, all users are placed at the cell border and there is no shadow fading ($\sigma = 0$). We model all channel profiles using the Vehicular A tapped delay-line model defined by the ITU standard. With this model, the channels offer moderate frequency diversity, and all channels have the same long-term average.
- **Model II:** In this case the users are spread out uniformly at random over the entire cell area and there is log-normal shadow fading ($\sigma = 6$ dB). The users' channel profiles are the same as in Model I (Vehicular A). This model yields channels with moderate frequency diversity, and a large spread in long-term averages.

- **Model III:** Here the users are spread out in the cell with shadow fading as in Model II and we use the ITU Vehicular B channel delay profile model. This model gives channels with large frequency diversity and a large spread in long-term average gains. In this case the cyclic prefix does not entirely cover the maximum tap delay which means that we would experience inter-symbol interference. However, in our simulation model the effects of inter-block and inter-symbol interference are not included. We do not consider this as a problem for this study since the impulse response power outside the cyclic prefix is small (roughly 10% of the total power) and we do not take error events on the downlink channel into account when determining our throughput measure.

Model II probably represents the most interesting model from a practical perspective. Models I and III are more extreme and are included to demonstrate the effect of large/small frequency diversity and large/small spread in the average channel gains.

The parameters in the system simulation were chosen to resemble those of LTE with a 5 MHz system bandwidth. Table II shows the system parameters. In this table N_{fr} is the number of frames simulated in each scenario. N_{tap}, $P_{\text{tap}}^{(\cdot)}$, $\tau_{\text{tap}}^{(\cdot)}$ and d_j are the parameters of the tapped-delay line channel model representing the number of taps, the average tap power profile, the nominal tap delay profile and the per-user tap-delay jitter. Also, $f_d^{(\cdot)}$ represents the Doppler frequency and P_e is the BLER target used to obtain the code rates from the lookup table. In this table, the superscripts VA and VB correspond to the channel models Vehicular A and Vehicular B respectively. With these parameters, the difference in average SNR between a user on the cell border ($r = R_c$) and a user at the reference distance ($r = R_0$) is 40 dB. Hence, in Models II and III, the average channel gains differ by 40 dB plus the fluctuations induced by the shadow fading.

In the LTE standard, the smallest possible scheduling granularity consists of 12 consecutive OFDM subcarriers in frequency and 14 consecutive OFDM symbols in time [2]. These 14 symbols span the entire scheduling frame. This is the example of frequency-aggregated granularity we consider in this study and for the sake of clarity, we denote it with *LTE granularity* throughout this section. Furthermore, since the control channel elements in LTE consist of 36 resource elements [2], we assume $N_{\text{CCE}} = 36$ in the presented results.

A. Signaling Overhead Ratio

The signaling overhead ratio is defined as

$$\Sigma = \frac{N_{\text{CONT}} N_{\text{CCE}}}{N_S N_c} \tag{21}$$

where N_{CONT} is the number of control channel elements in the control region (cf. (11) and (13)).

Figures 4, 7 and 10 show the signaling overhead ratio in percent for Models I, II and III. The results for the max-C/I and the proportional fair scheduler are plotted separately in subfigures (a) and (b) respectively. For the round-robin scheduler, the users take turns in transmitting and we assume that no by-frame signaling is needed to support this mechanism, thus giving a signaling overhead ratio of zero. The implicit assumption is that the users know a priori their ordering at the start of the round-robin mechanism. In practice this requires some initial setup signaling in higher layers of the communication protocol, which we do not consider in this study.

N_{SCEN}	100	N_t	100
N_{fr}	1000	p_u	0.1
N_c	300	N_{CCE}	36
N_s	14	t_f	200
Δf	15 kHz	β	0.95
T_0	71.429 μs	N_{CRC}	16
R_0	150 m	P_e	0.01
R_c	1500 m	N_{tap}	6
α	4	d_j	0.3 μs

$P_{\text{tap}}^{(\text{VA})}$	[0,-1,-9,-10,-15,-20] dB
$P_{\text{tap}}^{(\text{VB})}$	[-2.5,0,-12.8,-10,-25.2,-16] dB
$\tau_{\text{tap}}^{(\text{VA})}$	[0,0.31,0.71,1.09,1.73,2.51] μs Coherence bandwidth(RMS)≈ 2.7 MHz
$\tau_{\text{tap}}^{(\text{VB})}$	[0,0.3,8.9,12.9,17.1,20] μs Coherence bandwidth(RMS)≈ 0.25 MHz
$f_d^{(\text{VA})}$	200 Hz Coherence time $\approx \frac{0.423}{f_d} \approx 2.1$ ms
$f_d^{(\text{VB})}$	300 Hz Coherence time $\approx \frac{0.423}{f_d} \approx 1.4$ ms

TABLE II
SYSTEM SIMULATION PARAMETERS

The signaling overhead ratio curves all show the same general structure. They tend to a nonzero limit at high SNR. This behavior is due to the fact that the average amount of scheduling data that we want to transmit for a given approach is the same regardless of the system operating point. The modus operandi of both the max-C/I scheduler and the PF scheduler ignores the overall system operating point and only regards the relative differences in channel quality between users. This constant amount of scheduling data together with our assumption of a highest possible code rate of $7/8$ and QPSK modulation, is the reason that the curves do not tend to zero as the SNR grows. Similarly, all curves display a knee when going towards lower SNRs. This knee indicates where we are forced to start using successively lower code rates in order to meet the requirements on probability of error on the control channel (see Section V-B).

From the graphs it is apparent that decreasing the scheduling granularity decreases the size of the control region. This result was expected. We also see that for a given scheduling granularity the JCEB method outperforms the SCET scheme in terms of signaling overhead for the max-C/I scheduler, for all system operating points. This is most likely so because the max-C/I scheduler only selects users with good channel conditions. Therefore, compressing the multiuser map and broadcasting it using a single error-correcting code adapted to the worst user's channel consumes a relatively small amount of channel resources.

The PF scheduler on the other hand typically schedules all N_u users regardless of their channel quality. Scheduling information must thus be sent also to the users with poor channels and doing so requires a code with low rate. Therefore, for the cases where the users' channel gains display large variations, we expect the SCET approach to perform better than JCEB. This can be seen from the graphs for Models II and III (Figures 7 and 10) where at low SNR, compressing and transmitting the scheduling information separately (SCET) gives a lower signaling overhead ratio.

There are two specific circumstances under which the scheduling maps tend to become complex, and therefore require a large amount of control signaling. The first is when the channel offers much frequency/time diversity so that the channel gain varies significantly between the scheduling blocks. This happens, for example, in Model III. The second circumstance is when the scheduler selects many users in the same frame. The PF scheduler generally does this. The max-C/I scheduler does so only when the users' average SNRs are similar, which happens in Model I. Therefore, we would expect that the signaling overhead is larger in the following two cases: (i) generally, with the PF scheduler, and (ii) with the max-C/I scheduler in Models I and III. These observations are in line with what we can see in Figures 4, 7 and 10. Note also that for both the max-C/I scheduler and the PF scheduler, the users' average channel qualities are better when the users are spread out in the cell (Models II and III), giving SCET better operating conditions than JCEB.

B. Spectral Efficiency

In addition to comparing the signaling overhead ratios of the different approaches we also study their spectral efficiencies. The motivation is that a performance advantage in signaling overhead ratio does not necessarily directly translate into a performance advantage in spectral efficiency. Figures 5, 8 and 11 illustrate the system spectral efficiency with the JCEB scenario for Models I, II and III respectively. Figures 6, 9 and 12 illustrate the system spectral efficiency for the SCET scenario. Again subfigures (a) concern the max-C/I scheduler and subfigures (b) concern the proportional fair scheduler. For comparison the performance curve for the round-robin scheduler and the *genie bound* where max-C/I with the finest granularity is deployed and no signaling for conveying scheduling assignments is assumed, are also included in all cases.

From the graphs, we see that for the max-C/I scheduler, the performance is nearly independent of the scheduling granularity for Model II (but not for Models I and III). The reason is the difference in the amount of signaling overhead that was discussed above. Furthermore, it is evident that the difference between using JCEB and SCET is small when there is not much channel variations and this is mostly pronounced at low SNR. However when there is high potential of diversity in the system (Model II), JCEB is slightly better than SCET approach. This is in concert with the findings in our previous work [1], where we indicated a substantial advantage for the JCEB strategy for a max-C/I scheduler.

For the PF scheduler, we can generally say that for the finest granularity case, the overall system spectral efficiency is low compared to that of the coarser granularity cases. This is so because the cost associated with conveying the scheduling decisions is high, and it indicates that a coarser granularity would be a better choice. Coarser granularities achieve a spectral efficiency close to that of the genie-bound. For these granularities the performance difference between JCEB and SCET is in general small, but in the low SNR region there is a slight advantage for the SCET approach.

VIII. Conclusions

We have studied the two intertwined problems of scheduling and signaling of the scheduling assignments in OFDMA systems. From the presented results we draw the following conclusions.

- The difference in spectral efficiency performance between the JCEB and SCET approaches is small when a dynamic control region is assumed.
- The performance of the system-throughput maximizing scheduler when the JCEB signaling approach is used, is slightly better than when SCET is used.
- Scheduling with the finest granularity, despite the fact that it provides the opportunity to exploit the most multiuser diversity both in time and in frequency, results in the worst performance for both the system-throughput maximizing scheduler and for the proportional fair scheduler. The signaling overhead due to the transmission of the scheduling assignments consumes a significant amount of channel resources.
- For coarser granularities the results show that SCET requires a slightly lower overhead than JCEB, when proportional fair scheduling is employed and the users are spread out over the cell area. This translates into a small advantage for SCET in terms of spectral efficiency in the low SNR region.
- The results also indicate that the scheduling granularity standardized for LTE provides a good trade-off between scheduling granularity and overhead.

In our investigation we did not consider the delays incurred by different approaches, an other important system performance measure. Such a quantitative investigation would lead too far outside the main scope of the paper. However, in brief we can note that the system-throughput maximizing scheduler can incur completely intolerable delays by heavily prioritizing users close to the base station and starving users at the cell edge. The proportional fair schedulers have an inherent fairness also in terms of delays. The delays incurred depend on the scheduling granularity and the averaging window used, which can be tuned for different demands. Frame-wise scheduling in this context will in the delay perspective perform worse than the finer granularities since it introduces a comparatively large time granularity between different scheduling blocks.

References

[1] J. Eriksson, R. Moosavi, E. G. Larsson, N. Wiberg, P. Frenger and F. Gunnarsson, "On coding of scheduling information in OFDM," *in Proc. of IEEE VTC*, pp. 1-5, Apr. 2009.

[2] E. Dahlman, S. Parkvall, J. Sköld and P. Beming, *3G Evolution HSPA and LTE for Mobile Broadband*. 2nd Edition Academic Press 2008.

[3] M. Ergen, S. Coleri and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, pp. 362-370, Dec. 2003.

[4] C. Y. Wong, R. S. Cheng, K. B. Letaief and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1747-1757, Oct. 1999.

[5] H. Nguyen, J. Brouet, V. Kumar and T. Lestable, "Compression of associated signaling for adaptive multicarrier systems," *in Proc. of IEEE VTC*, pp. 1916-1919, May 2004.

[6] T. Kwon and D. H. Cho, "Adaptive modulation and coding based transmission of control messages for resource allocation in mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 2769-2782, Jul. 2009.

[7] J. W. So, "Performance analysis of VoIP service in the IEEE 802.16e OFDMA system with inband signaling," *IEEE Trans. Veh. Technol.*, vol. 57, pp. 1876-1886, May 2008.

[8] M. Sternad, T. Svensson and M. Dottling, "Resource allocation and control signaling in the WINNER flexible MAC concept," *in Proc. of IEEE VTC*, pp. 1-5, Sep. 2008.

[9] J.-H. Yeom and Y.-H. Lee, "Efficient transmission of multicast MAPs in IEEE 802.16e," *IEICE Trans. Commun.*, vol. E91-B, no. 10, pp. 3157-3160, Oct. 2008.

[10] E. G. Larsson, "Optimal OFDMA downlink scheduling under a control signaling cost constraint", *IEEE Trans. Commun.*, To appear.

[11] J. Gross, H. F. Geerdes, H. Karl and A. Wolisz, "Performance analysis of dynamic OFDMA systems with inband signaling," *IEEE J. Select. Areas Commun.*, vol. 24, pp. 427-436, Mar. 2006.

[12] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[13] X. Qiu and K. Chawla "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, pp. 884-895, Jun. 1999.

[14] W. Yu and J. M. Cioffi, "On constant-power waterfilling: performance bound and low-complexity implementation," *IEEE Trans. Commun.*, vol. 54, no. 1, pp. 23-28, Jan. 2006.

[15] J. M. Holtzman, "CDMA forward link waterfilling power control," *in Proc. of IEEE VTC*, pp. 1663-1667, May, 2000.

[16] K. Ch. Beh, S. Armour and A. Doufexi, "Joint time-frequency domain proportional fair scheduler with HARQ for 3GPP LTE systems," *In Proc. of IEEE VTC*, pp. 1-5, Sep. 2008.

[17] K. Sayood, *Introduction to Data Compression*. 3th Edition Morgan Kaufmann, 2005.

[18] J. G Proakis and M. Salehi, *Digital Communication*. 5th Edition McGraw-Hill International Edition 2008.

[19] T. J. Richardson, M. A. Shokrollahi and R. L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 619-637, Feb. 2001.

[20] C. Berrou and A. Glavieux, " Near optimum error correcting coding and decoding: turbo-codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261-1271, Oct. 1996.

[21] 3GPP TS 36.101, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception", Mar. 2009.

[22] 3GPP TS 36.321, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification", Mar. 2009.

[23] M. P. Rinne, O. Tirkkonen, T. Kashima, S. P. W. Jarot and J. Kahtava, "Unified entry format for common control signaling". United States patent application, Mar. 2007.

Reza Moosavi received his B.Sc. in Electrical Engineering from Isfahan University of Technology, Isfahan, Iran in 2005 and his M.Sc. from Chalmers University of Technology, Gteborg, Sweden, in 2008. Since February 2009, he is a Ph.D student at the Communication Systems Division of the Department of Electrical Engineering, Linkping University, Sweden. His research interest include resource allocation and signaling protocols in cellular systems.

Jonas Eriksson received his M.Sc in Applied Physics and Electrical Engineering in 1995, from Linkping University, and thereafter held a position at Saab Dynamics in Sweden as a radar systems engineer until 1999. He then joined the Data Transmission division at the Department of Electrical Engineering at Linkping University and recieved his Ph.D in Electrical Engineering in 2006. He is currently a Research Associate at the Communication Systems division at the Department of Electrical Engineering at Linkping University. His current research interest include radio resource management in cellular systems, signaling protocols, hybrid ARQ systems and general coding theory.

Erik G. Larsson is Professor and Head of the Division for Communication Systems in the Department of Electrical Engineering (ISY) at Linköping University (LiU) in Linköping, Sweden. He joined LiU in September 2007. He has previously been Associate Professor (Docent) at the Royal Institute of Technology (KTH) in Stockholm, Sweden, and Assistant Professor at the University of Florida and the George Washington University, USA.

His main professional interests are within the areas of wireless communications and signal processing. He has published some 60 journal papers on these topics, he is co-author of the textbook *Space-Time Block Coding for Wireless Communications* (Cambridge Univ. Press, 2003) and he holds 10 patents on wireless technology.

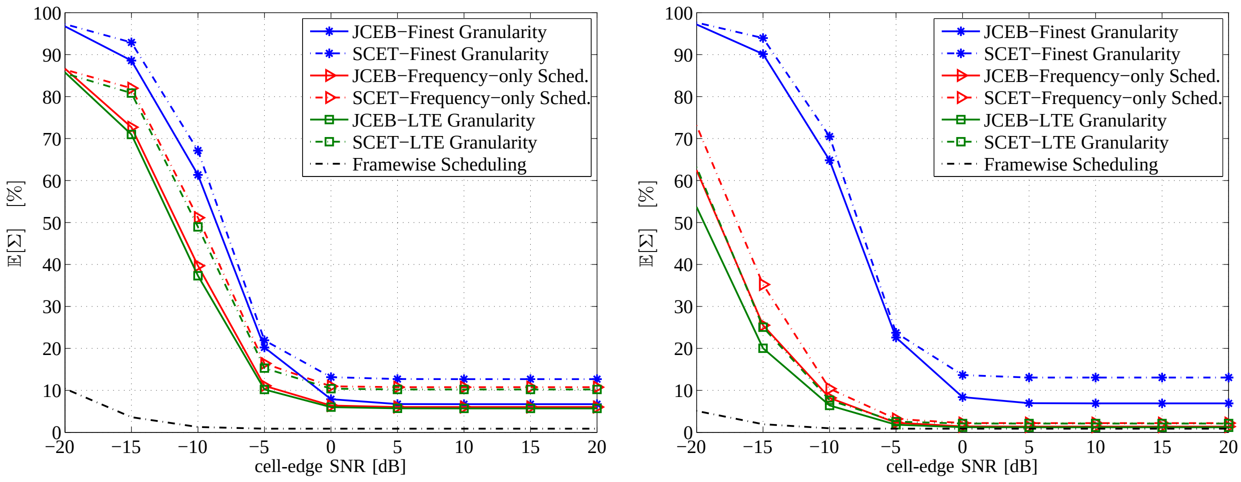
He has been Associate Editor for the *IEEE Transactions on Signal Processing*, the *IEEE Signal Processing Letters* and the *IEEE Transactions on Vehicular Technology*. He is a member of the IEEE Signal Processing Society SAM and SPCOM technical committees. He is active in conference organization, most recently as the technical area chair for communication systems of the Asilomar Conference on Signals, Systems and Computers 2010, and as general co-chair of CrownCom 2010.

Niclas Wiberg received his M.Sc in Computer Engineering in 1990, from Linkping University, and his Ph.D in Electrical Engineering in 1996, from the same university. He is currently an Expert at Ericsson Research in Sweden.

His research interest include radio protocols and radio resource management in cellular systems and radio network modeling and simulations.

Pål Frenger born in Alingsaas in 1968, received his M.Sc degree in Electrical Engineering in 1994 at Chalmers university of Technology, Gothenburg, Sweden. After having received his Ph.D degree at the same university he joined Ericsson Research in 1999. At Ericsson he has been working with both link level and system level research. In particular he has been heavily involved in the development of the 3GPP LTE system. Currently he is also working in the European EARTH project with radio network energy efficiency.

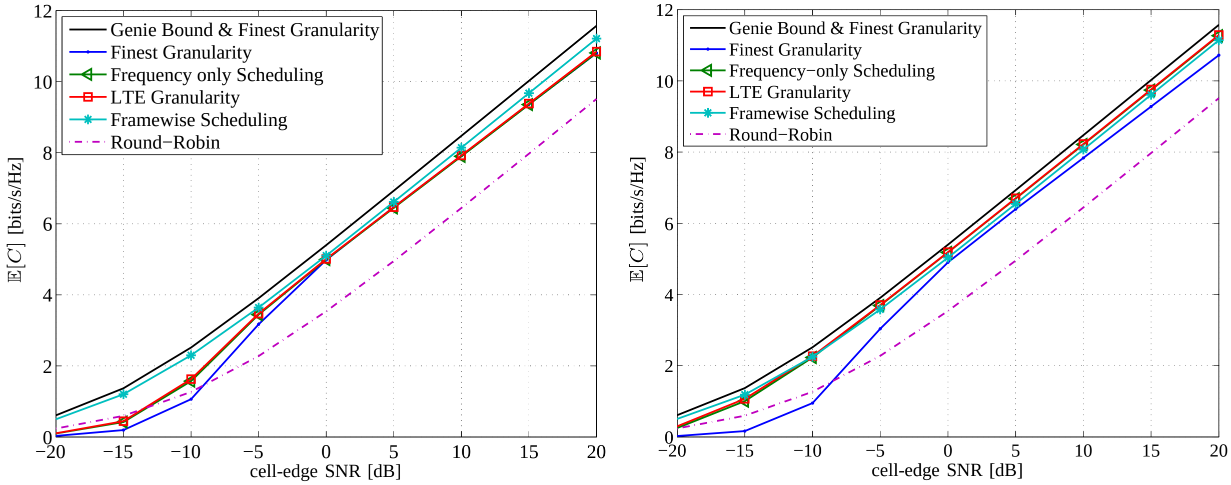
Fredrik Gunnarsson received his MSc in 1996 and his PhD in 2000, both in Electrical Engineering and from Linkping University, Sweden. Currently, he is with Ericsson Research, focusing on simplified radio network management, radio resource management, radio network modeling and simulations, primarily for 3GPP LTE. He is also a part-time Associate Professor in Automatic Control at Linkping University.



(a) Signaling Overhead Ratio for Max C/I

(b) Signaling Overhead Ratio for PF

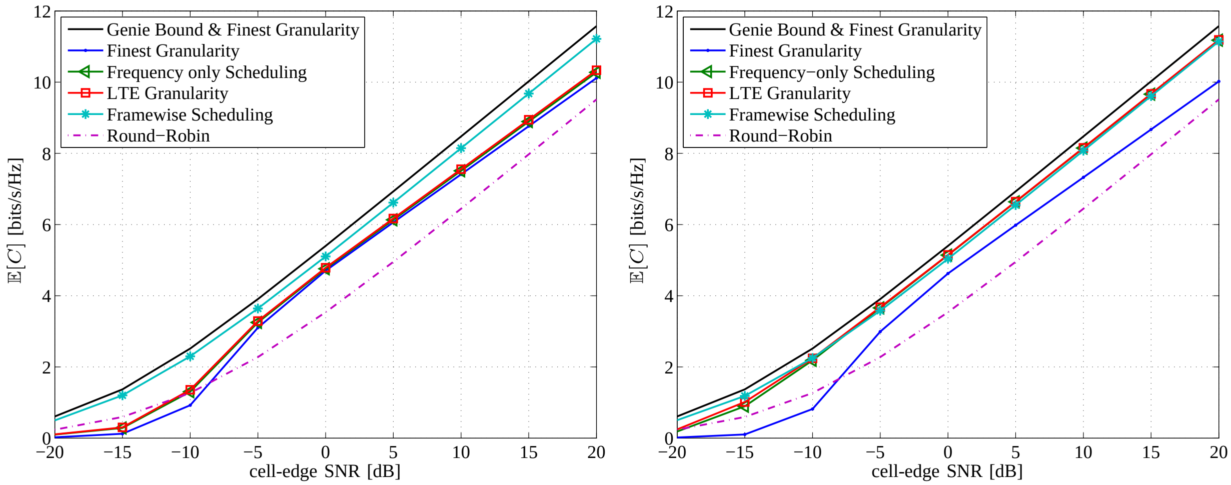
Fig. 4. Signaling overhead ratio in percent for Model I vs. cell-edge SNR 95%-percentile ($\mathrm{SNR}_{0.95}(R_c)$). Here all users are placed at the cell border and there is no shadow fading. Since there are no large channel variations (the users have equal average channel gains), JCEB is superior to SCET in this case (Model I).



(a) Spectral Efficiency for Max C/I with JCEB

(b) Spectral Efficiency for PF with JCEB

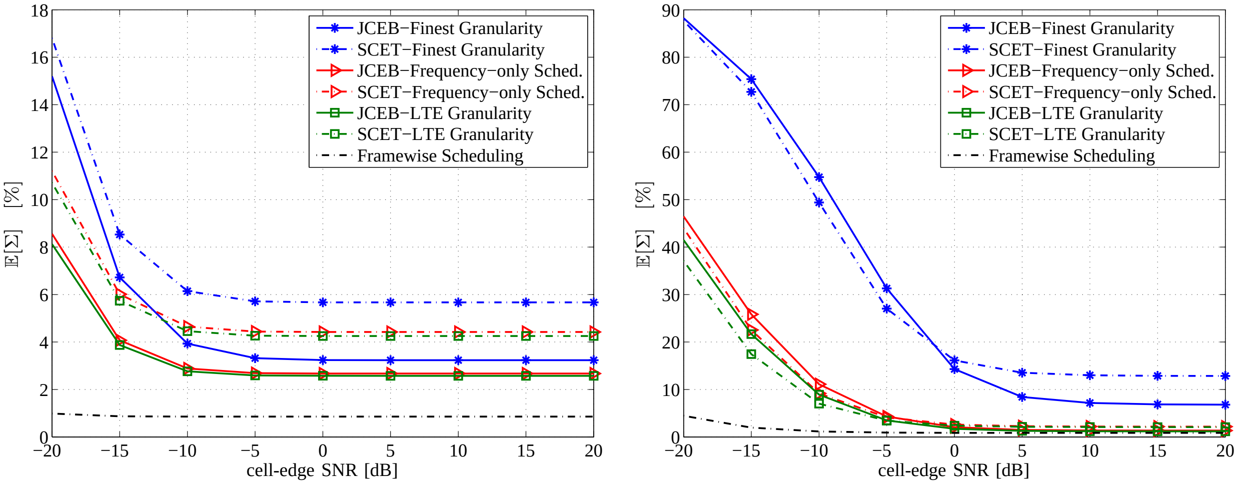
Fig. 5. Spectral efficiency for the joint compression, encoding and broadcast (JCEB) scenario and Model I vs. cell-edge SNR 95%-percentile ($\mathrm{SNR}_{0.95}(R_c)$).



(a) Spectral Efficiency for Max C/I with SCET

(b) Spectral Efficiency for PF with SCET

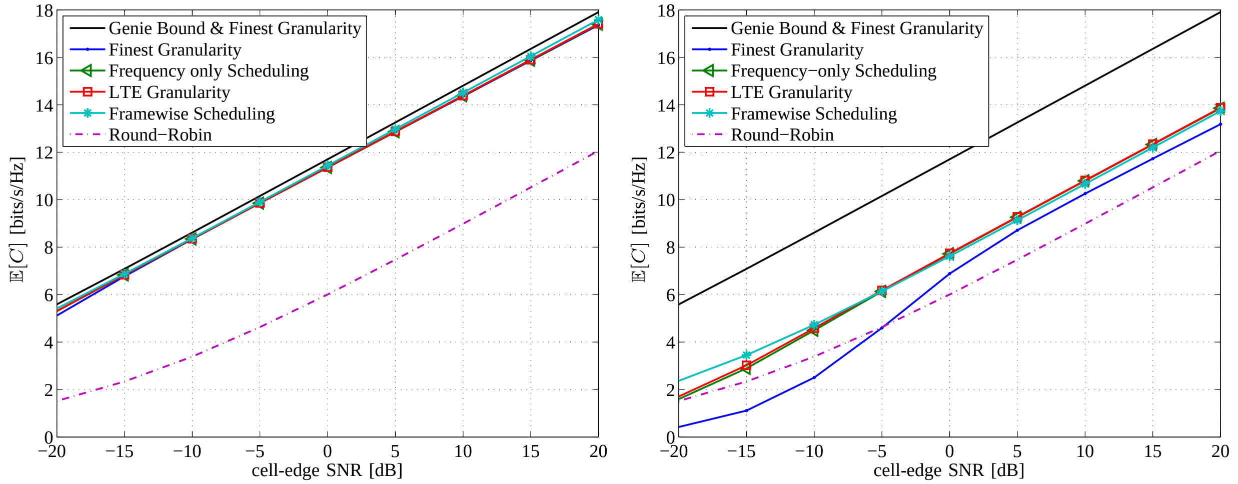
Fig. 6. Same as Figure 5 but for SCET scenario



(a) Signaling Overhead Ratio for Max C/I

(b) Signaling Overhead Ratio for PF

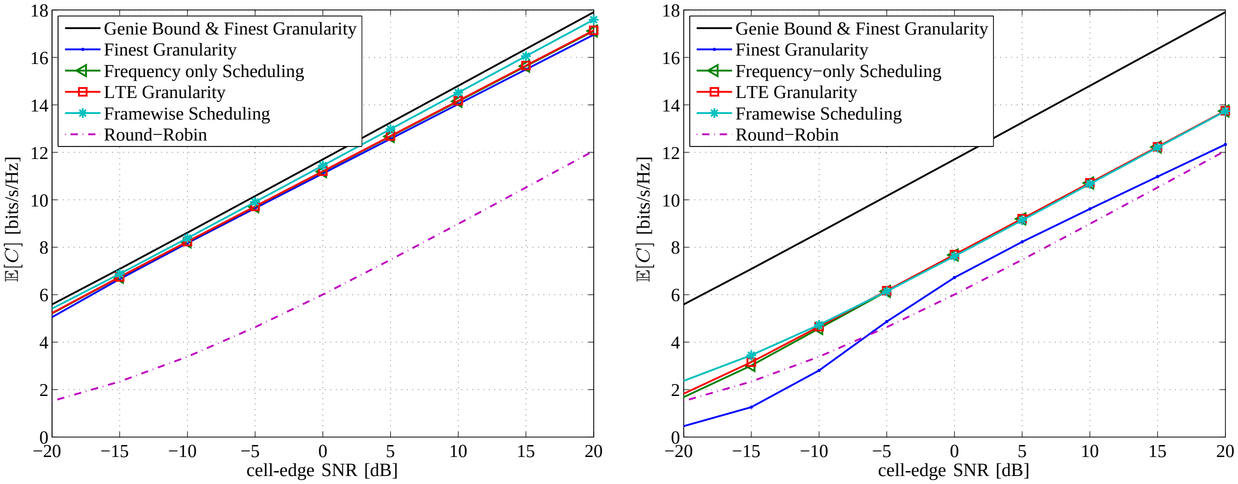
Fig. 7. Signaling overhead ratio for Model II vs. cell-edge SNR 95%-percentile ($\mathrm{SNR}_{0.95}(R_c)$).



(a) Spectral Efficiency for Max C/I with JCEB

(b) Spectral Efficiency for PF with JCEB

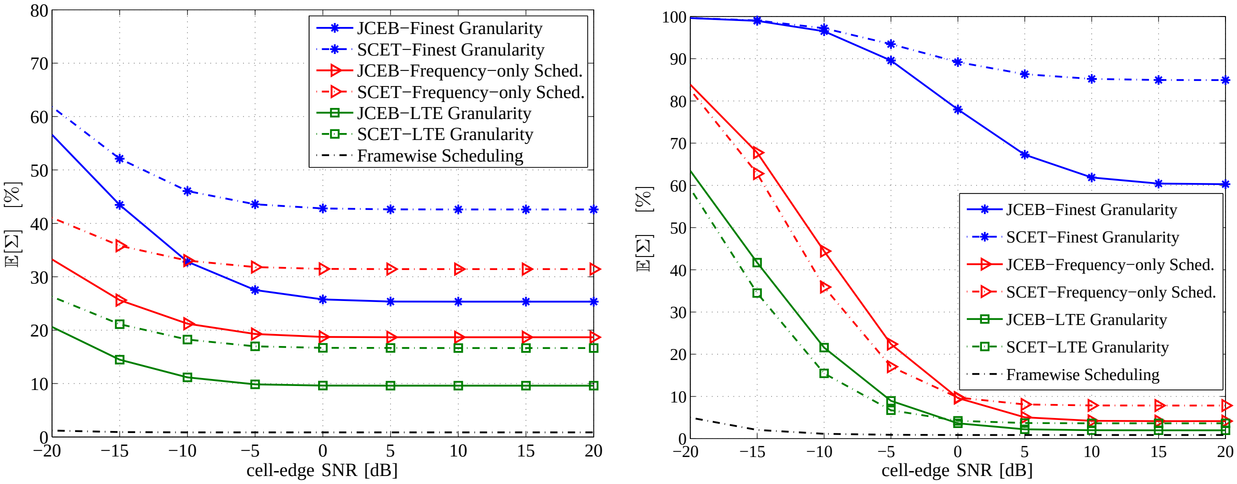
Fig. 8. Spectral efficiency for the JCEB scenario and Model II vs. cell-edge SNR 95%-percentile ($\mathrm{SNR}_{0.95}(R_c)$).



(a) Spectral Efficiency for Max C/I with SCET

(b) Spectral Efficiency for PF with SCET

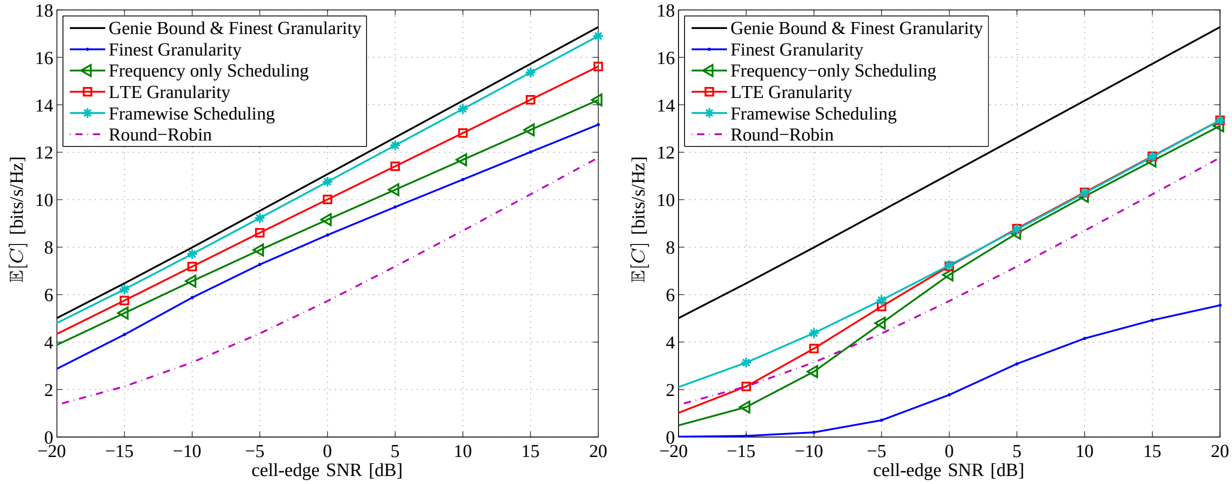
Fig. 9. Same as Figure 8 but for the SCET scenario.



(a) Signaling Overhead Ratio for Max C/I

(b) Signaling Overhead Ratio for PF

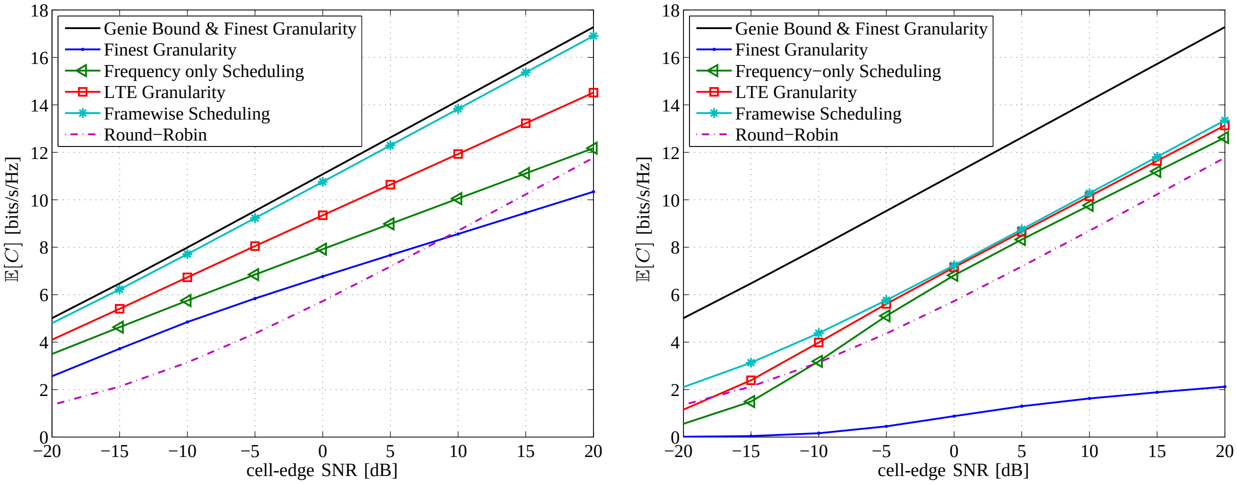
Fig. 10. Signaling overhead ratio for Model III (Vehicular B) vs. cell-edge SNR 95%-percentile ($\text{SNR}_{0.95}(R_c)$).



(a) Spectral Efficiency for Max C/I with JCEB

(b) Spectral Efficiency for PF with JCEB

Fig. 11. Spectral efficiency for the JCEB scenario and Model III vs. cell-edge SNR 95%-percentile ($\text{SNR}_{0.95}(R_c)$).



(a) Spectral Efficiency for Max C/I with SCET

(b) Spectral Efficiency for PF with SCET

Fig. 12. Same as Figure 11 but for the SCET scenario.