Linköping University Pre-Print

Reducing Downlink Signaling Traffic in Wireless Systems Using Mobile-Assisted Scheduling

Reza Moosavi and Erik G. Larsson

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

Reza Moosavi and Erik G. Larsson, Reducing Downlink Signaling Traffic in Wireless Systems Using Mobile-Assisted Scheduling, 2010, Proceedings of the IEEE Global Communications Conference 2010 (GLOBECOM 2010).

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-57629

Reducing Downlink Signaling Traffic in Wireless Systems Using Mobile-Assisted Scheduling

Reza Moosavi and Erik G. Larsson
Dept. of Electrical Engineering (ISY), Linköping University, Linköping, Sweden. Email:{reza,egl}@isy.liu.se

***Abstract*—We present an idea to reduce the part of the downlink signaling traffic in wireless multiple access systems that contains scheduling information. The theoretical basis of the scheme is that the scheduling decisions made by the base station are correlated with the CSI reports from the mobiles. This correlation can be exploited by the source coding scheme that is used to compress the scheduling maps before they are sent to the mobiles. In the proposed scheme, this idea is implemented by letting the mobiles make tentative scheduling decisions themselves, and then letting the base station transmit "agreement maps" instead of raw scheduling maps to the mobiles. The agreement maps have lower entropy and they require less resources to be transmitted than the original scheduling maps do. The improvement can be substantial.**

I. Introduction

A. Background and Motivation

In this paper we present an idea that can improve the performance (in the sense of resources required) of control signaling in wireless multiple access systems. The idea is motivated by the fact that the *scheduling map* which describes how resources are allocated to different users, is correlated with the channel state information (CSI). More precisely, in many wireless multiple access systems the scheduling decisions are made based on users' instantaneous channel conditions, which in turn, are obtained from CSI reports. Each terminal in the cell measures her received signal-to-noise ratio (SNR) or signal-to-interference-plus-noise ratio (SINR) and reports it back to the base station. The base station uses the received reports from all terminals to make the scheduling decisions. For instance, with proportional fair scheduling, each resource is allocated to the user with the highest priority, where the priority for each user is a known, monotonously increasing function of her instantaneous channel gain [1]. Since the terminals know their own reported CSI, one can exploit it to compress the scheduling maps more efficiently. We propose a scheme that implements this idea.

B. Related Work

Various techniques have been proposed to reduce the control signaling overhead caused by the transmission of scheduling assignments. A common approach that is also used in 3GPP Long-Term Evolution (LTE) is to reduce the scheduling granularity. More specifically, in LTE, OFDM is deployed as the multiple access scheme and the smallest possible scheduling granularity consists of 12 consecutive OFDM subcarriers in frequency and 14 consecutive OFDM symbols in time [2]. Another technique to reduce the control signaling was proposed in [3]. The idea is to choose new scheduling assignments using 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. In [4] a solution to decrease the amount of control signaling overhead was proposed. Therein, the users were grouped according to their channel gains, and all users in the same group used the same link adaptation parameters. A compression scheme to encode the control information was proposed in [5]. The compression algorithm therein consists of a run-length encoder, followed by a universal variable-length code (UVLC). In [6], a method for scheduling under a constraint on the control signaling cost was proposed. In [7] two different schemes for encoding of the control information were considered and their performances in terms of system throughput were evaluated.

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.

II. System Model

Assume that there are N_s *resource blocks* (resources, for short) that can be assigned to the users. Each such resource may represent a subdivision of the time, frequency, code or spatial domain or any combination of those, depending on the application. Corresponding to each of these N_s resources, we assume that the terminals can report a CSI message of length L bits. Hence, each report from a terminal is $N_s L$ bits long. Let N_u denote the total number of users in the cell.

III. Conventional approach

Let us first review the conventional approach of conveying scheduling maps to the users. Since there are N_s resources, the scheduling map is a vector of length N_s. For each terminal, we first find the corresponding *binary bitmap* which determines the resources that have been assigned to her. We compress each such binary bitmap individually. We use run-length encoding as the compression method [8]. However, for small N_s, in some cases run-length encoding can be more costly than sending uncompressed maps. Whenever this occurs, we send the data as an uncompressed binary map and we use run-length

encoding only if it achieves compression. Thus an extra flag bit is necessary to indicate whether run-length encoding is used.

Let $N_{\text{map}}^{(i)}$ denote the number of bits required to represent the binary bitmap associated with the ith user. The total number of required bits in this case is $N_{\text{Conv}} = \sum_{i=1}^{N_u} N_{\text{map}}^{(i)}$.

IV. Proposed scheme

We now present our new scheme. Herein, we let the users compute their priorities themselves. Since each terminal keeps track of both her channel states and her throughput, she is able to compute her priority vector. She then quantizes her priority vector using an L-level quantizer and reports it back to the base station. The base station collects these reports from all terminals and makes the scheduling decision. The basic rule when making the scheduling decision is that each resource is allocated to the user with the highest reported priority. However, sometimes more than one user will report the same priority for a given resource and a conflict occurs. This conflict must be resolved, see the discussion in the next paragraph. Once the scheduling decisions have been made, the base station sends an *agreement map* $\mathcal{M}_i$ and a *threshold* τ_i to each terminal i. The agreement map along with the threshold indicate the resources that have been assigned to the terminal. The agreement map consists of an N_s long binary vector per user whose elements correspond to one of the N_s resources. Once the terminal receives her associated agreement map, she finds her scheduling map as follows. If the corresponding field in the agreement map is "1", the terminal knows that the base station has agreed with her *proposal* with respect to the given threshold. Then she compares her reported value with the threshold and if her reported value is greater than the threshold, she knows that she has been allocated in that resource. If the corresponding field in the agreement map is "0", then she knows that a collision has occurred and that her proposal has not been accepted. In this case, if her reported value is smaller than the threshold, she knows that the corresponding resource is assigned to her. Figure 1 summarizes this procedure.

A special treatment is needed when a *collision* occurs and two or more users report the same priority for a certain resource. In the case of a collision, the base station assigns the resource to one of the candidate users in such a way that the resulting scheduling map has the shortest possible length (after appropriate compression, whenever that is effective). Note that finding the best assignment is a combinatorial optimization problem that can be solved exactly via dynamic programming or approximately by greedy techniques. For small N_s, N_u and L, it can be solved by a brute-force search.

The threshold τ_i is set individually for each user and there are L^{N_u} possible choices of thresholds. In the proposed scheme, the base station selects the set of thresholds that minimizes the combined size of the compressed agreement maps. This is a combinatorial problem as well. In the examples in Section VI, we used a brute-force search to select the optimum thresholds. The total number of bits required to represent the scheduling assignments with the proposed method is:

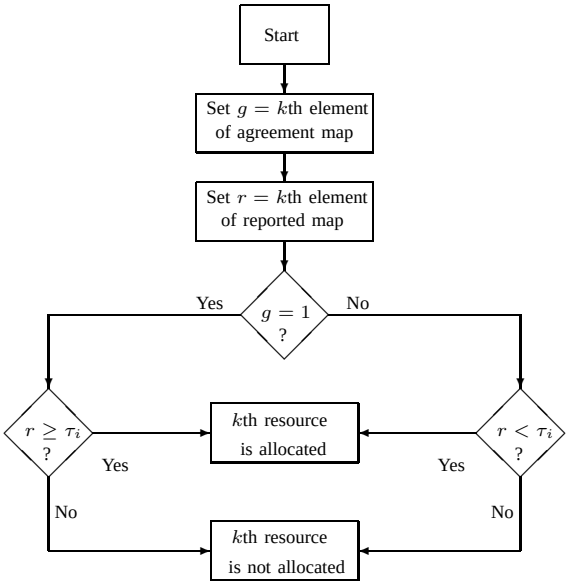

Fig. 1. Flowchart of the algorithm used by terminal i in order to find her scheduling maps

$$N_{\text{Prop}} = \sum_{i=1}^{N_u} \left(M_{\text{map}}^{(i)} + T \right) \tag{1}$$

where $M_{\text{map}}^{(i)}$ and T are the number of bits required to represent the agreement map $\mathcal{M}_i$ and the thresholds respectively.

To illustrate how the proposed scheme works, we provide an example. Let there be $N_u = 4$ users, $N_s = 16$ resources, and $L = 4$ CSI/priority quantization levels. Assume that the vectors of priorities received from the terminals are:

$$\begin{aligned}
\mathcal{P}_1 &= [1,\ 4,\ 3,\ 3,\ 4,\ 1,\ 1,\ 1,\ 3,\ 3,\ 4,\ 4,\ 2,\ 3,\ 3,\ 1] \\
\mathcal{P}_2 &= [2,\ 1,\ 2,\ 1,\ 2,\ 4,\ 4,\ 1,\ 3,\ 2,\ 2,\ 4,\ 1,\ 2,\ 2,\ 4] \\
\mathcal{P}_3 &= [1,\ 1,\ 1,\ 3,\ 4,\ 3,\ 1,\ 1,\ 4,\ 2,\ 4,\ 2,\ 3,\ 1,\ 4,\ 2] \\
\mathcal{P}_4 &= [4,\ 1,\ 4,\ 4,\ 1,\ 4,\ 3,\ 2,\ 3,\ 2,\ 4,\ 1,\ 4,\ 2,\ 2,\ 2]\,.
\end{aligned}$$

Each element in each vector corresponds to one resource block. The base station makes the decision based on the reports from the terminals. Each resource is allocated to the user who has reported the largest value. We see that for some resources, there is more than one choice and we have a collision. Table I illustrates the procedures of finding the optimal map, for a sample choice of threshold values $[4, 4, 4, 4]$.

V. Theoretical Justification of the Proposed Scheme

Here we provide some more fundamental motivation for the scheme proposed in Section IV. We compare the conventional and proposed schemes for compressing scheduling maps in terms of the entropy of their associated binary scheduling maps that are transmitted to the users. Thereby the implicit assumption is that we can use an entropy achieving compression scheme. This generally requires the maps to be infinitely long [9]. While this might not be practical due to short block lengths and tight delay requirements, it provides an insight on the ultimate improvement that we can achieve.

Reported Vectors				Candidates for Resources	Optimal Assignment
$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$		
1	2	1	**4**	4	4
4	1	1	1	1	1
3	2	1	**4**	4	4
3	1	3	**4**	4	4
4	2	**4**	1	1 , 3	3
1	**4**	3	**4**	2 , 4	2
1	**4**	1	3	2	2
1	1	1	**2**	4	4
3	3	**4**	3	3	3
3	2	2	2	1	1
4	2	**4**	**4**	1 , 3 , 4	3
4	**4**	2	1	1 , 2	2
2	1	3	**4**	4	4
3	2	1	2	1	1
3	2	**4**	2	3	3
1	**4**	2	2	2	2

(a) **Conventional Approach**

Binary Bitmaps			
$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$	$\mathcal{B}_4$
0	0	0	1
1	0	0	0
0	0	0	1
0	0	0	1
0	0	1	0
0	1	0	0
0	1	0	0
0	0	0	1
0	0	1	0
1	0	0	0
0	0	1	0
0	1	0	0
0	0	0	1
1	0	0	0
0	0	1	0
0	1	0	0

Required Bits			
$N_{\text{map}}^{(1)}$	$N_{\text{map}}^{(2)}$	$N_{\text{map}}^{(3)}$	$N_{\text{map}}^{(4)}$
17	17	17	17

Total Cost

N_{Conv}: 68

(b) **Proposed Scheme**

Agreement Maps			
$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
1	1	1	0
1	1	1	1
1	1	1	0
1	1	1	1
0	1	1	1
0	1	1	0
1	1	1	1
1	1	1	1
0	1	1	1
1	1	1	1
1	1	1	1

Required Bits			
$M_{\text{map}}^{(1)}$	$M_{\text{map}}^{(2)}$	$M_{\text{map}}^{(3)}$	$M_{\text{map}}^{(4)}$
17	6	6	17

Total Cost

N_{Prop}: 54

TABLE I

ILLUSTRATION OF THE PROPOSED SCHEME, FOR THE THRESHOLD VALUES $[4, 4, 4, 4]$.

For the analysis to follow, we assume that the reports from the terminals are i.i.d. with a uniform distribution over $[0,1]$. The assumption that the reports are i.i.d. is critical. However, the assumption on a uniform distribution will not affect the result because a deterministic function (transformation) can be applied to the received reports at the scheduler to make the received reports look like they have a uniform distribution (after the transformation). The threshold τ would then need to be transformed in the same way, too. To compute the entropy of the binary maps in each part, we first need to find the probability of having a "1" in each position. Without loss of generality, consider the event that the jth slot is assigned to the first user, and denote this event by X. Clearly, by symmetry the probability of this event is $\Pr\{X\} = 1/N_u$, since any user is equally likely to "win" if all users reports are independent and have the same statistical distribution. We provide an alternative derivation of $\Pr\{X\}$ in order to illustrate some principles of calculation that we will need later on. Let R_i be the report from the ith terminal. Since we assign the slot to the first user if her reports is greater than the rest of the received reports from other terminals, we have

$$\begin{aligned}\Pr\{X\} &= \Pr\{R_1 > R_i, \forall i = 2,3,\dots N_u\} \\ &= \int_{r_1} p(R_1 > R_i, \forall i = 2,3,\dots N_u | R_1) p_{R_1}(r_1) \mathrm{d}r_1 \\ &= \int_0^1 \left(\prod_{i=2}^{N_u} p_{R_i}(r_i < r_1) \right) \mathrm{d}r_1 = \int_0^1 r_1^{N_u - 1} \mathrm{d}r_1 = \frac{1}{N_u} \quad (2)\end{aligned}$$

Now let us compute the probability that the jth bit in the agreement map for the first user is "1". First recall that we agree with a user either if (i) her report is greater than the threshold τ and the slot has been assigned to her or if (ii) her report is less than the threshold and the slot has not been assigned to her. Therefore we have

$$\Pr\left\{\mathcal{M}_1^{(j)} = 1\right\} = \Pr\{R_1 > \tau, X\} + \Pr\{R_1 < \tau, X^c\} \quad (3)$$

where X, as before, is the event that the jth slot is assigned to the first user. Let us compute each term in (3) separately.

$$\begin{aligned}\Pr\{R_1 > \tau, X\} &= \int_{r_1} p(R_1 > \tau, X | R_1) p_{R_1}(r_1) \mathrm{d}r_1 \\ &= \int_\tau^1 p(R_i < r_1, \forall i = 2,3,\dots N_u) \mathrm{d}r_1 \\ &= \int_\tau^1 r_1^{N_u - 1} \mathrm{d}r_1 = \frac{1 - \tau^{N_u}}{N_u}, \quad (4) \\ \Pr\{R_1 < \tau, X^c\} &= \int_{r_1} p(R_1 < \tau, X^c | R_1) p_{R_1}(r_1) \mathrm{d}r_1 \\ &= \int_0^\tau p(R_i > r_1, \text{ for some } i = 2,3,\dots N_u) \mathrm{d}r_1 \\ &= \int_0^\tau \Big(1 - p(R_i < r_1, \forall i = 2,3,\dots N_u)\Big) \mathrm{d}r_1 \\ &= \int_0^\tau \left(1 - r_1^{N_u - 1}\right) \mathrm{d}r_1 = \tau - \frac{\tau^{N_u}}{N_u}. \quad (5)\end{aligned}$$

Therefore we have

$$\Pr\left\{\mathcal{M}_1^{(j)} = 1\right\} = \tau + \frac{1 - 2\tau^{N_u}}{N_u}. \quad (6)$$

The entropy for a binary source is given by

$$H(p) = -p \log_2(p) - (1-p) \log_2(1-p) \quad (7)$$

where p is the probability of observing a "1". We note that for the proposed scheme the threshold can be chosen such that the entropy of the agreement map is as small as possible. Figure 2 illustrates the entropy of the binary maps for the two schemes and the best choice of the threshold (in terms of achieving the minimum entropy), as a function of the number of users. As we see the entropy of the agreement map is smaller than the entropy of the conventional approach. This means that after appropriate compression, the map should require less resources to be transmitted. As the number of users grows, the probability of assigning an slot to a user decreases (cf. (2)). Therefore on the average, when there are many users in the cell, in order for a user to be scheduled her report should be relatively larger compared to the case with a small number of users. Thus the optimum threshold should be an increasing function of the number of users. This explains the behavior of the optimum threshold curve in Figure 2.

VI. Numerical Results

We compare the average performance of the proposed approach with that of the conventional approach in terms of the total number bits required to convey the scheduling assignments. For each method, this is the sum of the number of bits required to represent the individual maps. We assume that the reports from the terminals are independent of each other and that each report is uniformly distributed over the L possible levels. However we have seen in experiments not reported here that the choice of the distribution does not have a significant effect on the presented results. Figure 3 compares the two methods for a scenario with $L = 4$ reporting levels and $N_s = 20$ resources. As we see, with 6 users a gain of 15 bits is achieved by using the proposed scheme. We expect a larger gain when the number of resources is larger. This can be verified from the Figure 4 where the number of reporting levels is the same as Figure 3 but the number of resources is increased to $N = 64$. We used a *greedy search* algorithm to find a sub-optimal solution to the combinatorial optimization problem. It is not clear how far this solution is from the optimum one. With 4 users, we can achieve a gain of 30 bits.

VII. Improvements

So far we have not considered the actual cost associated with the transmission of the scheduling assignments. Generally, in practical systems there will always be users that have poor channel conditions. For instance, there are users that are located far from the base station or users that are temporarily in deep fade. Transmission of the scheduling information to those users can be very costly in terms of the resources needed to be set aside for the signaling. Taking the transmission cost into

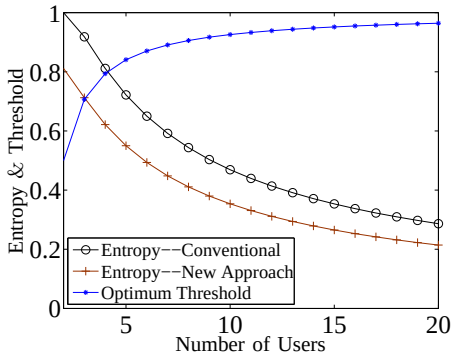

Fig. 2. The entropy of the binary maps and the optimum threshold as a function of the number of users.

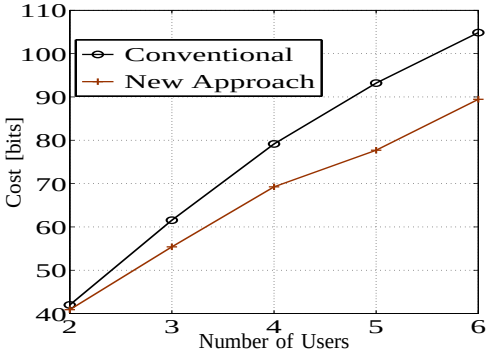

Fig. 3. Comparison between two methods with $L = 4$ quantization levels in the CSI/priority report and $N_s = 20$ resources per frame.

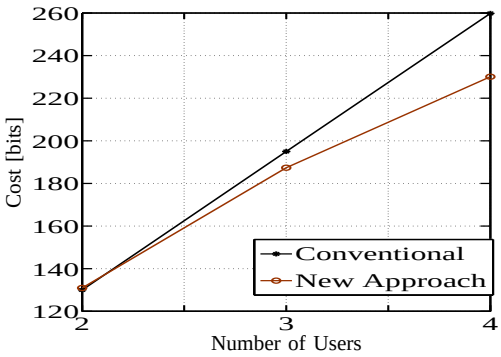

Fig. 4. Comparison between two methods with $L = 4$ quantization levels in the CSI/priority report and $N_s = 64$ resources per frame.

account, it is occasionally more beneficial to grant more of the proposals from the users who have poor channel conditions. For example, by simply agreeing to the entire proposal from a user with a poor channel, the base station does not need to send any scheduling (nor agreement) map[1] to them and thus the scheduling cost can be significantly reduced. However in doing so, the base station may need to compromise the throughput-optimality of the resulting scheduling assignments.

To illustrate this concept, consider an example with $N_s = 4$ and $L = 4$. Assume that we want to schedule two users. Let

$$\mathcal{P}_1 = [4,\ 3,\ 3,\ 1] \quad , \quad \mathcal{P}_2 = [3,\ 4,\ 1,\ 2]$$

be the reports from the two users and assume that the second user has a poor channel. Following the conventional scheme of assigning each resource to the user that has reported the highest value, the throughput-optimal scheduling map will be $[1,\ 2,\ 1,\ 2]$. On the other hand, by using the proposed scheme described in Section IV, the optimal thresholds for the two users are 4 and the corresponding agreement maps are:

$$\mathcal{M}_1 = [1,\ 1,\ 0,\ 1] \quad , \quad \mathcal{M}_2 = [1,\ 1,\ 1,\ 0]\,.$$

As we see, the agreement maps are "easier" to compress than the original binary bitmaps. However we can further reduce the signaling cost by compromising the scheduling map. Note that the base station has not agreed with the second user at the 4th resource. This problem arises from the fact that the second user has reported an small number (2) and that it therefore does not expect to receive any data in that slot. However, since the other user has also reported an small number (1), this slot is allocated to the second user. Now if we let the base station compromise the scheduling map and allocate the last resource to the first user, then the agreement maps would become:

$$\tilde{\mathcal{M}}_1 = [1,\ 1,\ 0,\ 0] \quad , \quad \tilde{\mathcal{M}}_2 = [1,\ 1,\ 1,\ 1]\,.$$

We see that for the second user, the associated agreement map consists of only "1"s. We refer to this situation as *full agreement*. In such cases, it is enough to send a flag bit along with the threshold. Since the transmission of the scheduling information to the second user is relatively costly, one can save a large amount of channel resources by the proposed method.

VIII. Conclusions

We have presented a new scheme for reducing the downlink signaling traffic in wireless multiple access systems. The proposed scheme can achieve a reduction of 15 bits for 6 users when there are 20 resource blocks and a reduction of at least 30 bits for 4 users and 64 resource blocks. We showed that finding the optimum solution (the optimum assignments and the set of thresholds) is a combinatorial problem. Future work should include both an investigation on finding efficient algorithms for solving the problem and also a study of the proposed improvements discussed in Section VII.

[1] Except for a flag bit needed to indicate the acceptance of their proposals, and an associated value of the threshold.

References

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

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

[3] 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, March 2006.

[4] 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, Sept. 2008.

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

[7] 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, April 2009.

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

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition A Wiley-Interscience publication 2005.