DOI: 10.1109/SPAWC.2011.5990461

PREEMPTION AND QOS MANAGEMENT ALGORITHMS FOR COORDINATED AND UNCOORDINATED BASE STATIONS

Olga Muñoz-Medina[1], *Antonio Pascual-Iserte*[1,2], *Pau Baquero*[1], *and Josep Vidal*[1]

[1]Dept. of Signal Theory and Communications - Universitat Politècnica de Catalunya (UPC), Spain
[2]Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Spain
Email: {olga.munoz,antonio.pascual,josep.vidal}@upc.edu

ABSTRACT

This paper addresses the problem of resource allocation and quality of service (QoS) management in a downlink communication system where the users are allowed to be attached to and receive data from one or several cooperating base stations (BSs) under orthogonal frequency division multiple access (OFDMA). Under the assumption that only information regarding the link path loss is available at each transmitter, the problem of power and bandwidth allocation to minimize transmission power subject to minimum rate (i.e., QoS) constraints is written as a convex optimization problem. Considering that the transmission power per BS is limited, the Lagrange multipliers of the optimization problem are exploited either to efficiently identify which are the candidates to be preempted whenever unaffordable power consumption arises, or to relax the target rate constraints for the ongoing connections. In the second case, the Lagrange multipliers are used in combination with the Newton-Raphson algorithm to adjust the QoS of the users in the system according to the available BS power.

Index Terms— Resource allocation, preemption, QoS management, BS cooperation, power and bandwidth allocation.

1. INTRODUCTION

Optimizing the allocation of available resources (i.e., power and bandwidth) to provide the best quality of service (QoS) to the maximum number of users is the goal of any communication system. This QoS can be formulated, for example, in terms of the provision of a target rate per user for real time applications such as voice or video.

Typically, the allocation of resources is done to maximize the system sum-rate. However, such an approach may yield solutions where some users receive a rate much greater than necessary, while others receive less than what they need. To ensure that all the users receive at least what they need, minimal rate constraints can be included in the problem. However, this may arise to an unfeasible problem and, even if the problem is feasible, it does not guarantee fairness. Note that user fairness could be achieved by following a max-min criterion. In such a case, all the users would receive the same rate, but this rate not necessarily fulfills the QoS requirements.

Taking into account previous considerations, we consider in this paper the minimization of the transmission power for a given target rate per user as in [1]. As the total transmission power is limited, it may happen that the demanded target rates cannot be provided due to the power limitations at the base stations (BSs). In this case, we need to:

- either preempt some services (possibly considering less important connections as candidate for preemption), and therefore reducing the number of users in the system,
- or reduce the rate of the users in the system (i.e., QoS management).

Preemption algorithms have been proposed for instance in routing [2, 3] to free up enough resources whenever the residual bandwidth of a link is lower than the bandwidth request of a higher priority connection. How should the selection of users be performed or how much should the target rate be reduced are not straightforward decisions. In this paper, two algorithms are proposed to answer both questions in a more efficient way than the pure "try and error" (for QoS reduction) or the "exhaustive search" (for users drop) approach. These algorithms, which have a low computational complexity, are based on the information provided by the optimum values of the Lagrange multipliers of the associated optimization problem. Note also that combining the rate readjustment and the users drop could be an attractive approach in scenarios where the users go into and out of the system randomly. By means of this combination, it could be used to adjust the trade-off between the call-drop rate due to congestion and the admissible reduction in rate.

In this paper, the algorithms for preemption and QoS management are proposed for a system where users are allowed to be attached to and receive data from one or several neighbor BSs. Attaching the mobile station (MS) to the BS with better signal-to-noise plus interference (SNIR) conditions is the solution that makes more sense with unlimited power resources. However, receiving data from more than one BS when the first one has exhausted its available power (or other resources) is a powerful approach particularly in asymmetric load situations.

The rest of the paper is organized as follows. In Section 2 the system and the resource allocation problem are described. Based on the previous problem, Section 3 presents two algorithms for the user preemption and the QoS readjustment problems. Finally, Section 4 provides some simulations results to prove the validity of the presented algorithms, whereas Section 5 concludes the paper.

2. PROBLEM DESCRIPTION

2.1. Problem formulation

We consider an orthogonal frequency division multiple access system (OFDMA) following a partial usage of subchannels (PUSC) or

This work was supported in part by the European Commission (FP7) through FREEDOM ICT-2007-4-248891 and NEWCOM++ grant no. 216715, by the Spanish Ministry of Education and Science, FEDER funds (TEC2006-06481 FBNIT, CONSOLIDER CSD2008-00010 COMONSENS and TEC2010-19171 MOSAIC) and the Catalan Government (2009SGR-01236, 2009SGR-891 AGAUR).

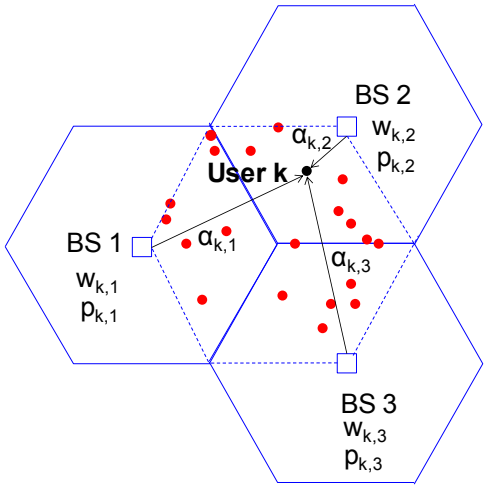

Fig. 1. Example of a deployment where 3 BSs are cooperating.

full usage of subchannels (FUSC) channelization as specified in the 802.16 standard [4]. In such a case, groups of non contiguous carriers are assigned to every user. Carriers within a given group are assumed to bear uncorrelated channel fading and codewords are interleaved among carriers. No detailed knowledge of the channel response per carrier is assumed and therefore the power allocation is uniform among carriers. Under this transmission strategy, the channel capacity is given by the ergodic (or average) capacity. This type of transmission is suitable for those user scenarios where the channel coherence time is in the range (or shorter) than the channel state update rate from the receiver to the transmitter.

Under these assumptions, we consider a system where K users are allowed to be attached to and receive data from N cooperating BSs to achieve a target rate r_k for each user $k = 1, \ldots, K$ (see Fig. 1 for an example of such a deployment). The total bandwidth assigned to the set of BSs is limited and orthogonal transmissions are considered, i.e., the frequency bands allocated to different BSs for the transmission towards different users are not overlapping. We assume that the symbols transmitted by the N base stations to a given user are uncorrelated and are transmitted simultaneously in time in non-overlapping frequency bands. In this framework, our objective is to minimize the maximum of the N BS transmitting powers. This problem can be formulated as follows:

$$\underset{t,\{p_{k,n}\},\{w_{k,n}\}}{\text{minimize}} \quad t \tag{1}$$

$$\text{subject to} \quad \sum_{k=1}^{K} p_{k,n} \leq t, \qquad n = 1, \ldots, N, \tag{2}$$

$$r_k \leq \sum_{n=1}^{N} w_{k,n} \log_2 \left(1 + \rho \alpha_{k,n} \frac{p_{k,n}}{w_{k,n}}\right), \quad k = 1, \ldots, K, \tag{3}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} w_{k,n} = 1, \tag{4}$$

where $p_{k,n}$ and $w_{k,n}$ correspond to the fractional power and bandwidth allocated to the k-th user in the n-th BS respectively, and the objective function t is the maximum fractional power transmitted within the set of BSs. Finally, $\alpha_{k,n}$ is defined as

$$\alpha_{k,n} = \frac{P_{BS}}{l_{k,n} N_o B} \tag{5}$$

with P_{BS} the available power per BS (for simplicity, all the BSs are assumed to have the same available power), $l_{k,n}$ the path-loss (including antenna pattern effects) between the k-th user and the n-th BS, B the total bandwidth for the set of BSs, and N_o the noise spectral density. Under the previous notation, the power and bandwidth allocated for the transmission from the n-th BS to the k-th user is $p_{k,n} P_{BS}$ and $w_{k,n} B$, respectively. Thus, if the resulting optimum value for t is greater than 1, this implies an unaffordable power consumption.

The right hand of eq. (3) is a lower bound to the ergodic capacity[1], with $\rho = 2^{\mathbb{E}[\log_2 |h|^2]}$ a constant that depends on the fading statistics of the channel ($|h|$ stands for the random channel amplitude assuming unitary variance). It was shown in [5] that this bound is extremely tight for Rayleigh fading. We shall therefore take (3) as a reasonable approximation of the maximum achievable rates.

To shorten notation, we shall refer to the solution to problem (1)-(4) as $t^\star(\mathbf{r})$, with $\mathbf{r}$ being a vector whose k-th component is the target rate of the k-th user, i.e. r_k ($\mathbf{r} = [r_1 \ \cdots \ r_K]^T$). Notice that the optimum solution of this problem will lead to a situation in which all the BSs will transmit the same power. In the particular case where N equals 1, every BS carries out the scheduling of its own users without coordination with the rest of BSs.

2.2. Relationship between the optimized objective function and the Lagrange multipliers

The problem defined in (1)-(4) is a convex optimization problem where the objective and all the inequality constraints functions are differentiable with respect to the optimization variables [6]. That enables to apply very powerful numerical primal-dual optimization techniques, such as the interior point method. These methods provide not only the optimum value of the primal optimization variables, but also the optimum value of the Lagrange multipliers or dual variables.

Let us denote the optimum value of the Lagrange multiplier corresponding to the inequality constraint (3) associated to the k-th user by $\lambda_k^\star(\mathbf{r})$. As shown in [6], there exists a relationship between the optimum value of the objective function (i.e., the minimum required transmitter power at each BS) and the target rate given by

$$\frac{\partial t^\star(\mathbf{r})}{\partial r_k} = \lambda_k^\star(\mathbf{r}), \qquad k = 1, \ldots, K. \tag{6}$$

In the particular case where all the user target rates are equal ($r_k = r, \ \ k = 1, \ldots, K$), we obtain [7]:

$$t^\star(r) = t^\star(\mathbf{r}), \quad \lambda_k^\star(r) = \lambda_k^\star(\mathbf{r}), \quad r_k = r, \quad k = 1, \ldots, K, \tag{7}$$

$$\frac{\partial t^\star(r)}{\partial r} = \sum_{k=1}^{K} \lambda_k^\star(r). \tag{8}$$

In addition, it can be proved that functions $t^\star(\mathbf{r})$ and $t^\star(r)$ are convex with respect to $\mathbf{r}$ and r respectively [6, 8].

[1] By ergodic capacity we understand that averaging is performed with respect to the fading distribution in the frequency domain assuming a sufficiently large number of independently fading carriers.This approximation is more accurate when the carriers allocated to one user are not contiguous, as happens in PUSC and FUSC permutation modes in WiMAX.

Users preemption algorithm

1.	**set** target rate and initial set of users: $\mathbf{r} = \mathbf{r}_{\text{target}}$ (e.g. $\mathbf{r}_{\text{target}} = 10 \cdot \mathbf{1}_{K\times 1}$), $\overline{K} = \{1, \ldots, K\}$
2.	**solve** optimization problem (1)-(4) and obtain: $p = t^\star(\mathbf{r})$ and $\lambda_k^\star(\mathbf{r})$ for $k = 1, \ldots, K$
3.	**while** $p > 1$ and $\overline{K} \neq \emptyset$ **do**
4.	**set** $\Delta p = 0$
5.	**while** $\Delta p < p - 1$ and $\overline{K} \neq \emptyset$ **do**
6.	$\widehat{k} = \max_{k\in\overline{K}} \lambda_k^\star(\mathbf{r})$
7.	$\Delta p = \Delta p + \lambda_{\widehat{k}}^\star(\mathbf{r}) r_{\widehat{k}}$
8.	$r_{\widehat{k}} = 0$
9.	$\overline{K} = \overline{K} - \{\widehat{k}\}$
10.	**end while**
11.	**solve** optimization problem (1)-(4) and obtain: $p = t^\star(\mathbf{r})$ and $\lambda_k^\star(\mathbf{r})$ for $k \in \overline{K}$
12.	**end while**

Table 1. Lagrange based users preemption algorithm.

The idea of using the Lagrange multipliers for terminal admission/removal was actually suggested in [9] for a different set up (i.e. without BS coordination and different performance criteria). The authors suggested in an ad-hoc way to remove, whenever necessary, the terminal with the greatest product between the Lagrange multiplier and the constraint.

Different from [9], we will exploit the relationships between the optimal dual variables and the minimum required transmission power at each BS that have been presented in this section. Indeed, eq. (6) and (8) and the fact that functions $t^\star(\mathbf{r})$ and $t^\star(r)$ are convex with respect to $\mathbf{r}$ and r respectively can be used to provide efficient algorithms for reducing the target rates and/or denying service to users, whenever the required transmission power is higher than the available power, i.e., $t^\star(\mathbf{r}) > 1$.

3. PREEMPTION AND QOS MANAGEMENT ALGORITHMS

3.1. Preemption strategy

Reducing the number of users admitted in the system allows for reducing the transmission power. The optimum solution for the users to be preempted requires a combinatorial search. For each possible combination, the optimization problem (1)-(4) needs to be solved, implying a high computational cost. On the other hand, eq. (6) suggests that eliminating the user with the highest Lagrange multiplier will have the highest impact on the reduction of the transmission power. The linear approximation for such a reduction, when dropping the k-th user, is:

$$\Delta t^\star(\mathbf{r}) \leq \lambda_k^\star r_k, \tag{9}$$

where the inequality comes from the fact that $t^\star(\mathbf{r})$ is convex with respect to $\mathbf{r}$ [6, 8]. This allows to know in advanced which is the minimum set of users that need to be dropped.

Table 1 summarizes the proposed preemption algorithm based on the information provided by the Lagrange multipliers. The accepted users will achieve the target rate.

QoS readjustment algorithm

1.	**set** $r = r_o$ (e.g. $r_o = 10$) such as $r_k = r$, $k = 1, \ldots, K$ and tolerance $\varepsilon > 0$ (e.g. $\varepsilon = 0.05$)
2.	**solve** optimization problem (1)-(4) and obtain: $p = t^\star(r)$ and $\lambda_k^\star(r)$ for $k = 1, \ldots, K$
3.	**while** $p > 1 + \varepsilon$ **do**
4.	$\Delta r = \frac{p-1}{\sum_{k=1}^{K} \lambda_k^\star(r)}$
5.	$r = r - \Delta r$
6.	**solve** optimization problem (1)-(4) and obtain: $p = t^\star(r)$ and $\lambda_k^\star(r)$ for $k = 1, \ldots, K$
7.	**end while**

Table 2. Lagrange based QoS readjustment algorithm.

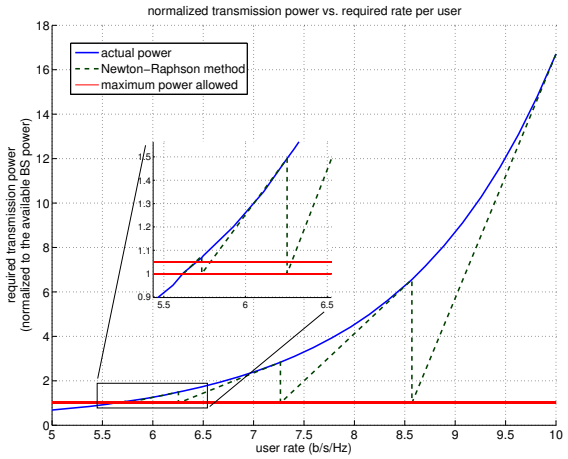

Fig. 2. Application of the Newton-Raphson's method for the rate readjustment.

3.2. QoS readjustment

Instead of reducing the number of active users in the system, an alternative is to keep them but applying a reduction in rate to fulfill the transmission power constraint per BS, i.e. $t^\star(\mathbf{r}) \leq 1$. Table 2 summarizes the algorithm based on the information given by the Lagrange multipliers that computes the maximum target rate without call-drops. Despite that for the sake of clarity the algorithm is written for equal user rates, it can be easily extended to the case of different rates for different users. This algorithm is based on the application of the well known Newton-Rapshon's method [10] in combination with the fact that $t^\star(r)$ is convex with respect to r and eq. (8). As $t^\star(r)$ is convex with respect to r, at each iteration the required power will be closer to 1 but greater than 1. To avoid infinite iterations, we define a stop criterion consisting in the fulfillment of the power constraint with a certain tolerance represented by ε. Thus, the computed required power will be equal to or smaller than[2] $(1 + \varepsilon)P_{BS}$.

An example for the application of such an algorithm is presented in Fig. 2, where we represent the required transmission power (normalized) vs. the target user rates. We allow a tolerance of $\varepsilon = 0.05$.

[2]Note that we can reduce the value of P_{BS} by a factor $(1 + \varepsilon)$ in the algorithm computations, so that the final computed power does not exceed the actual available power at the BSs.

	Optimum comb. search	Lagrange multipliers	Worst channels
% of coincidences	—	90 %	90 %
% of user acceptance	89.29 %	89.29 %	89.29 %
Computational complexity	502.75	2.45	3.1
Normalized trans. power	0.7779	0.7788	0.7790

Table 3. Comparison among optimum combinatorial search, Lagrange, and worst-channel based preemption algorithms.

The continuous line corresponds to the exact value of the required transmission power, i.e., $t^{\star}(r)$, whereas the dashed line results from the application of the Newton-Raphson's method. As can be seen observed in Fig. 2, starting from an initial target rate of 10, only 5 iterations are required to compute the rate readjustment.

4. SIMULATIONS

We consider a scenario with 3 coordinated sectors of 120^{o}, as shown in Fig. 1, each one with a radius of 1 km. 20 users are uniformly distributed within the sectors area. The BS power is 33 dBm. The power spectral density of the noise is -174 dBm/Hz and the bandwidth is 5 MHz. The pathloss is drawn according to the WINNER II channel model [11] at 2.5 GHz carrier frequency. Sector antennas [12] are considered at each BS.

The first set of results correspond to the preemption algorithm. Considering a user target rate of 7 bps/Hz, in most cases, the required transmission power is greater than the available one at each BS. The optimal preemption algorithm requires evaluating all the possible combinations of users to select the combination with the highest number of active users that fulfill the power constraint. To avoid unnecessary computations, combinations dropping 1 user are tested firstly, then combinations dropping 2 users and so on, until the required power at each BS is equal to or less than the available power.

The combinatorial approach is compared in Table 3 with the proposed preemption algorithm, and also with the approach of dropping the user with the worst channel until fulfilling the power constraint. For the comparison we consider four measurements:

- A coincidence is declared if the set of active (selected) users is equal to the set obtained by the optimum (combinatorial) approach.
- The percentage of user acceptance, i.e. the number of admitted users with respect to the total number of users.
- The computational complexity measured as the number of times that the optimization algorithm (1)-(4) is computed.
- The required power at each BS (normalized).

Table 3 shows the values of these figures computed over 20 independent scenarios. The values for the last three rows correspond to values averaged over the 20 scenarios. From the results in Table 3 we observe that proposed preemption algorithm reduces considerably the computational complexity with respect to the optimum approach (2.45 versus 502.75), at the expense of a slightly degradation in performance. This degradation is actually a very slight increase

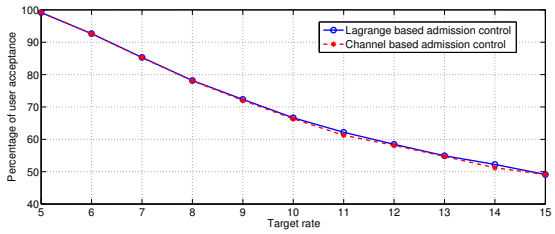

Fig. 3. Average admission percentage versus target rate for the Lagrange based preemption algorithm and worst-channel based preemption algorithm.

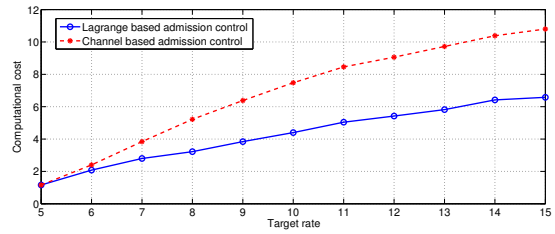

Fig. 4. Average computational cost versus target rate for the Lagrange based and worst-channel based preemption algorithms.

in the transmission power, while the average percentage of acceptance is the same (89.29%). The difference in power comes from the 10% of the cases where the users selected for dropping are not exactly the same. The approach of dropping the users with the worst channel presents a higher computational complexity with respect to the proposed algorithm (3.1 versus 2.45) and also a slight increase in transmitted power (0.7790 versus 0.7788).

Pursuing a greater target rate will require more resources and therefore the dropping of more users. Fig. 3 and 4 show the percentage of acceptance and the computational complexity respectively versus the target rate. Because of the unaffordable computational complexity of the combinatorial approach, results are shown only for the two suboptimal approaches. The figure depicts average values over 50 independent deployments. While the percentage of acceptance is practically equal for both approaches (Fig. 3), the computational complexity is smaller for the Lagrange based admission control (Fig. 4).

Finally, Fig. 5 corresponds to the Newton-Raphson's method for the rate readjustment, as an alternative to the pure 'try and error' approach. The figure depicts the computational complexity (averaged over 50 independent deployments) required for this algorithm versus the initial value of the rate. Three different available BS powers have been considered. The computational complexity is measured as in previous results. As expected, the computational complexity is greater when the starting rate is higher or the power is smaller, as the system requirements are harder. The computational complexity for the values considered is affordable as the number of times that the optimization algorithm (1)-(4) needs to be computed is less than 10.

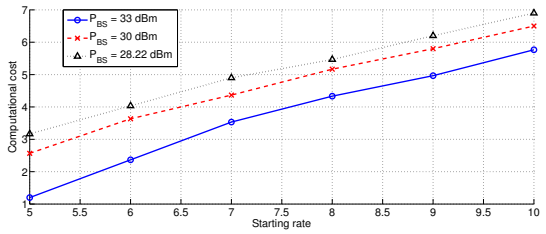

Fig. 5. Average computational cost to compute the rate readjustment versus the starting rate value.

5. CONCLUSIONS

This paper proposes two algorithms for preemption and QoS management respectively. The algorithms can be applied to both single BS and coordinated BSs set ups. The QoS management algorithm allows to perform the rate readjustment in a power limited system. On the other hand, for the preemption strategy the proposed Lagrange based approach has shown to be more efficient than the combinatorial search approach (optimum) without practically no degradation in terms of both the required transmitter power and the percentage of users acceptance. It is also more efficient than the solution consisting in dropping those users with the worst channel. Further research will consider the combination of both algorithms in order to achieve a trade-off between a small call-drop rate and an admissible reduction in rate.

6. REFERENCES

[1] A. Liu, H. Xiang, W. Luo, L. Ping, and Liu Y., "Power minimization of multiaccess MIMO systems with rate constraints and finite-rate feedback," in *Information Sciences and systems, CISS 2009*, pp. 488–493.

[2] F. Rafique Dogar, L. Aslam, Z. Afzal Uzmi, S. Abbasi, and Y-C. Kim, "Connection Preemption in Multi-Class Networks," in *Proc. IEEE Global Communications Conference (GLOBECOM'06)*, November 2006, pp. 1–6.

[3] S. Jeon, R.T. Abler, and A.E. Boulart, "The Optimal Connection Preemption Algorithm in a Multi-Class Network," in *Proc. IEEE International Conference on Communications (ICC'02)*, April 2002, pp. 2294–2298.

[4] IEEE Std. 802.16e 2005, *IEEE standard for local and metropolitan area networks, Part 16: Air interface for fixed and mobile broadband wireless access systems*, Tech. Rep. IEEE Standards Dept., 2005.

[5] E. Calvo, J. Vidal, and J. Fonollosa, "Optimal Resource Allocation in Relay-assisted Cellular Networks with Partial CSI," *IEEE Trans. on Signal Processing*, vol. 57, pp. 2809–2823, July 2009.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[7] C. Wrede and M. Spiegel, *Advanced Calculus*, McGraw-Hill, 2002.

[8] W. Yu and Y. Lui, "Dual Methods for Nonconvex Spectrum Optimization of Multicarrier Systems," *IEEE Trans. on Communications*, vol. 54, no. 7, pp. 1310–1322, July 2006.

[9] R. Stridh, M. Bengtsson, and B. Ottersten, "System Evaluation of Optimal Downlink Beamforming with Congestion Control in Wireless Communication," *IEEE Trans. on Wireless Communications*, vol. 5, no. 4, pp. 743–751, April 2006.

[10] E. Süli and D. Mayers, *An Introduction to Numerical Analysis*, Cambridge University Press, 2003.

[11] IST-4-027756 WINNER II, *WINNER II channel models Part II - Radio channel measurement and analysis results*, D1.1.2 V1.0, 2007.

[12] R. Srinivasan, *IEEE 802.16m Evaluation Methodology Document*, http://ieee802.org/16, March 2007.