Comparative Study of the IEEE 802.16 Random Access Mechanisms

Dirk Staehle and Rastin Pries
University of Würzburg, Inst. of Computer Science, Dept. of Distributed Systems
Am Hubland, 97074 Würzburg, Germany
{dstaehle,pries}@informatik.uni-wuerzburg.de

Abstract

The WiMAX technology based on the IEEE802.16 standard is currently the most prospective candidate for broadband wireless access networks. One of the key issues is the design of the MAC layer, in particular the multiple access scheme. The IEEE 802.16 standard specifies different scheduling services with individual mechanisms for accessing the channel on the uplink. The non-real-time polling service and the best-effort service mainly rely on a contention mechanism to submit bandwidth requests to the base station. These two services are currently used for all types of traffic with unknown characteristics, i.e. typically all traffic except for some special VoIP connection with known codec. In this paper, we evaluate and compare the performance of the contention mechanisms for fixed and mobile WiMAX.

1 Introduction

Fixed and mobile broadband wireless access networks are one of the key investments in the near future. For these investments to become profitable, wireless Internet access has to keep pace with the increasing data rates provided by wired Internet access technologies. The development of novel user-driven applications with users creating, distributing, and sharing their own content poses new challenges in particular to the uplink performance, and makes higher bandwidths and lower delays on the uplink necessary. Worldwide Interoperability for Microwave Access (WiMAX) currently presents the most recent development of wireless technology. Originally intended for Fixed Broadband Wireless Access (FBWA) networks and as a wireless competitor for wireline DSL and cable modem access in particular in rural and low-infrastructure areas, the most recent stage of WiMAX also provides mobility support mainly intended for nomadic users or users with little mobility. WiMAX is a consortium founded to enable the interoperability and foster the commercialization of products based on the IEEE 802.16 standard. The current IEEE 802.16-2004 [3] standard with the extensions for mobility support amended in the IEEE 802.16e-2005 [4] standard are the basis for two WiMAX certified products. The OFDM part of IEEE 802.16-2004 is known as Fixed WiMAX and the OFDMA part of IEEE802.16e-2005 is known as Mobile WiMAX.

One of the most critical and challenging parts for designing the MAC layer of a wireless technology is the uplink multiple access. On the downlink, the base station has full knowledge on the current bandwidth demand, i.e. on the packets stored in its buffers and is able to schedule the transmissions. On the uplink, the base station does not know the current buffer contents at the subscriber stations. There are two extreme solutions for uplink multiple access: The first one is to allocate a certain resource to every subscriber station oblivious of whether or not it has data to send. The other possibility is to grant resources to a subscriber station only when it has data ready to send and explicitly requests the bandwidth. The advantage of the first extreme is the short access delay, the disadvantage is the waste of resources if a subscriber station has nothing to transmit. The disadvantages of the second extreme are the access delay and the additional resources for transmitting bandwidth requests, the advantage is the good utilization of resources. The WiMAX standard specifies different variants for coordinating the access on the uplink called scheduling services. The unsolicited grant service (UGS) corresponds to the first extreme, the base station periodically grants resources to the subscriber station. The real-time polling service (rtPS) is located between the two extremes, the base station periodically grants opportunities to request bandwidth to the subscribed station. The non-real time polling service (nrtPS) and the best-effort service (BE) transmit bandwidth requests via random access or by piggybacking the requests to already granted data transmissions. For nrtPS connections the base station sporadically grants exclusive bandwidth request opportunities. The IEEE 802.16e-2005 standard additionally specifies the extended real-time polling service (ertPS) that is located between UGS and rtPS. It can be seen as UGS enhanced with the possibility to adapt the bandwidth.

The different scheduling services specified in the standard are well-designed for the different applications like VoIP, FTP, VoD, etc. and the respective traffic patterns and QoS requirements. One problem of the more enhanced scheduling services, UGS, ertPS, and rtPS, is that for using them efficiently, a good knowledge of the traffic characteristics is required. If e.g. a VoIP connection should be transported via UGS the voice codec, i.e. inter-arrival time of packets and packet size, need to be known in order to grant the resources to the subscriber station efficiently. Since the traffic characteristics of many applications like games, uploads from web cams, Skype calls, etc. that actually would fit well to one of the enhanced scheduling services, are not known or not signaled automatically to the CPE or base station, they are typically still transported via nrtPS or BE. This makes the simple random access one of the most important mechanisms for WiMAX performance.

Several papers have been published focusing on the random access phase for bandwidth requests in OFDM systems [7, 1, 6, 2]. In [7, 1], it is shown that the backoff window should be set to the number of transmission opportunities per frame or to a multiple of these transmission opportunities. A transmission opportunity is the time it takes to submit one bandwidth request. If the backoff is set to a lower value than the number of transmission opportunities, bandwidth will be wasted. [6, 2] also focus on the random access scheme in OFDM only. The difference to the papers referenced above is that not only the BE queue is simulated but all other service classes as well and their delay is compared.

The papers [5, 8] present both an analytical approach and a validation of the approach by simulation. However, in [5] it is claimed that the delay is less than one millisecond and it is not clear how the delay is measured and how the authors get to these small delays. In [8] the delay is measured for random access and unicast polling. However, unicast polling can result in a lot of wasted bandwidth, especially in scenarios with non periodic traffic like web browsing.

To the best of our knowledge, no paper has been published so far, with a comprehensive study of the random access phase for both OFDM and OFDMA systems in IEEE 802.16 networks. In this paper we compare the random access mechanisms of OFDM256 and OFDMA. We will further investigate the impact of different parameters defining the contention mechanism and partially identify optimal parameters.

In Section 2 we give an overview of the random access mechanism as specified in the WIMAX MAC layer. In Section 3 we describe the simulation model used for producing the results given in Section 4. In Section 5 we draw some conclusions.

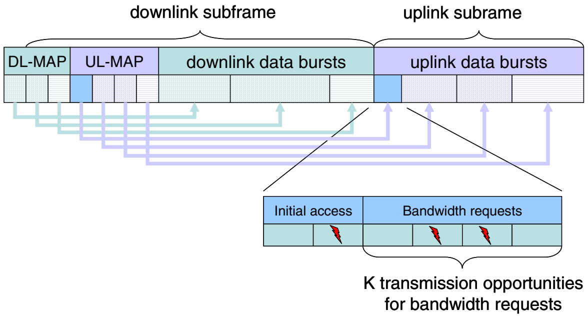

Figure 1. Structure of an IEEE 802.16 OFDM TDD frame

2 Short overview of IEEE802.16 MAC and PHY layer

The IEEE 802.16 standard specifies four physical layers namely SC and SCa for single carrier transmission in Line-of-sight and non-line-of sight environments, OFDM (also OFDM256) and OFDMA for multi-carrier transmission in non-line-of-sight environments. A common MAC layer is defined for all physical layers with only small adaptations to the different physical layers. The standard specifies two modes of operation, point-to-multi-point and mesh mode. In the following we focus on point-to-multi-point communication and the OFDM and OFDMA physical layers. The operation of the MAC layer is best explained by means of a graphic of the frame structure as shown in Fig. 1. The frame consists of a downlink subframe and an uplink subframe. The downlink subframe starts with a preamble and a frame control header not shown in the figure that mainly specifies the presence of control information within the downlink subframe. In particular it indicates changes with respect to the last frame. The DL-MAP and UL-MAP specify the usage of the rest of the frame. The DL-MAP defines the address and the burst profile (modulation and coding) of the data bursts within the data part of the downlink subframe. The UL-MAP allocates resources in the uplink subframe to the different subscriber stations. Additionally, the UL-MAP specifies resources to be used for the random access.

The way how random access is performed is different for OFDM and OFDMA so let us first discuss the main differences between the two physical layers. On very short terms, OFDM means that an OFDM symbol is entirely used by a single users while with OFDMA an OFDM symbol is separated into subchannels and multiple users may transmit in parallel. In the extreme, the smallest amount of data allocated to a single user is one subchannel for the duration of one OFDM symbol.

Now, back to the random access procedure for OFDM.

As described above, the whole frame contains a number of OFDM symbols and the UL-MAP specifies groups of OFDM sybmols within the uplink subframe called transmission opportunities to be used for transmitting data, for initial ranging, and for bandwidth requests. Initial ranging transmission opportunities are used by new subscriber stations to register to the base station. Bandwidth request opportunities are used by the subscriber stations to request bandwidth from the base station, i.e. they signal the amount of data they have to transfer and on reception of the bandwidth request the base station schedules grants in the UL-MAPs of the coming frames according to the QoS parameters of the respective connection. We distinguish unicast and broadcast bandwidth request transmission opportunities. Unicast transmission opportunities are dedicated to a single subscriber station and mainly used by rtPS and ertPS. Broadbast bandwidth request transmission opportunities are randomly accessed using a truncated binary exponential backoff mechanism. When a subscriber station receives data and intends to transmit a bandwidth request, it selects a backoff in terms of broadcast bandwidth request transmission opportunities between 0 and $2^{W_{min}}$. After counting down the backoff, it transmits the bandwidth request including all data present in its buffers, starts a timeout, and waits for a grant from the base station. If the timeout exceeds, the subscriber station selects a back-off between 0 and $2^{W_{min}+1}$ and repeats the whole procedure. The maximum upper bound for the backoff interval is $2^{W_{max}}$. A bandwidth request opportunity with full contention consists of a short preamble and an OFDM symbol what together makes up for two OFDM symbols per bandwidth request. While full contention is mandatory, subchannelization and region focused are alternative contention methods. Subchannelization means that a number of consecutive OFDM symbols may be subdivided in frequency and time into regions usable for sending a bandwidth request. Region focused means that the subscriber station sends a short contention code to the base station that after receiving this code grants a unicast bandwidth transmission opportunity to this code. Focused contention is not further considered in this paper since it is quite similar to the OFDMA contention mechanism.

OFDMA in the IEEE 802.16-2004 standard uses 2048 subcarriers. The IEEE 802.16e-2005 standard uses scalable OFDMA which supports 2048, 1024, 512, and 128 subcarriers depending on the channel bandwidths. For the rest of this paper we focus on OFDMA with 2048 subcarriers. The subcarriers are subdivided into subchannels using either PUSC (partial usage of subchannels) or FUSC (full usage of subchannels). We focus on FUSC in the following which means that we have 1440 data subcarriers subdivided into 60 subchannels with 24 subcarriers each. Now, let's come back to the question how a subscriber station requests bandwidths. The standard defines a single ranging channel for initial ranging, periodic ranging, and sending bandwidth requests. The ranging channel is allocated a region consisting of a multiple of N OFDMA symbols and a multiple of six adjacent subchannels. The minimum allocation for requesting bandwidth is six adjacent subchannels and one OFDMA symbol, i.e. 144 subcarriers. On these 144 subcarriers a ranging code is BPSK modulated, which is a pseudo-noise sequence consisting of 144 bit. Alternatively, the base station might signal to use N OFDM symbols for one request, the N consecutive ranging codes are modulated on the $N \times 144$ subcarriers. Using more than one OFDM symbol achieves a more robust transmission of ranging codes.

The standard defines a generator for 256 of these ranging codes, and every base station uses a group of codes for initial ranging, periodic ranging, handover ranging, and sending bandwidth requests. Sending a bandwidth request first involves sending a randomly chosen ranging code to the base station. On reception of the ranging code, the base station responds by granting a transmission opportunity for sending the actual bandwidth request to the respective ranging code.

As the transmission of ranging codes is uncoordinated, the following errors might occur: First, several subscriber stations might use the same ranging code on the same group of subchannels, the base station grants a request opportunity, and the actual bandwidth requests of the involved subscriber stations collide. Second, the ranging code might not be recognized either due to an erroneous channel which is rather improbable or due to multiple colliding different ranging codes. Third, a ranging code might be detected though it is not transmitted and a grant for a request remains unused.

Let us analyze the probability that a ranging code is not detected if K different ranging codes collide. We assume a perfect channel and equal received powers, i.e. the received ranging code sequence $r_1, ..., r_{144}$ is equal to the sum of the transmitted ranging codes $c_{k,1}, ..., c_{k,144}$ with $r_i = \sum_{k=1}^{K} c_{k,i}$ and $c_{k,i} \in \{-1, +1\}$. A ranging code is detected as transmitted if the scalar product of received sequence and ranging code exceeds a certain threshold T. The scalar product of r and c_k is

$$r \cdot c_k = \sum_{j=1}^{K} \sum_{i=1}^{144} c_{j,i} \cdot c_{k,i} = 144 + \sum_{j=1, j \neq k}^{K} \sum_{i=1}^{144} c_{j,i} \cdot c_{k,i}.$$

Assuming $c_{j,i}$ and $c_{k,i}$ to be independent, the product $c_{j,i} \cdot c_{k,i}$ assumes the values -1 and $+1$ with equal probability. Consequently, the product is equal to the random variable $2 * B_z - 1$ where B_z is a 0.5-Bernoulli random variable.

Accordingly we obtain,

$$r \cdot c_k = 144 + 2 \cdot \sum_{z=1}^{144\cdot(K-1)} B_z - 144 \cdot (K-1)$$
$$= 144 + 2 \cdot Bin(144(K-1), 0.5) - 144(K-1),$$

where $Bin(n,p)$ denotes a binomial random variable. Accordingly, we obtain the probability that a transmitted ranging code is not correctly detected as

$$p_{not} = P\left(Bin\big(144(K-1), 0.5\big) < \tfrac{T+144(K-2)}{2}\right) \quad (1)$$

and the probability that a not transmitted ranging code is erroneously detected as

$$p_{wrong} = P\left(Bin\big(144 \cdot K, 0.5\big) < \tfrac{T+144\cdot K}{2}\right). \quad (2)$$

In the following, we demonstrate the accuracy of the derivation by some simulation results. Therefore, we randomly choose K out of the 256 ranging codes and additionally select one special code among these K. We repeat this experiment for 1000 times and determine for every set of colliding codes whether the selected code is detected or not for different thresholds T=54, 72 and 90. Fig. 2 shows the probability that a transmitted ranging code is not detected depending on the number of colliding codes and the detection threshold. As expected, we observe that the failure probability increases both with the number of connections and the detection threshold. Good detection probabilities of 98% and more are obtained for less than 10 collided ranging codes and a threshold T=54. From this figure alone one could suggest to further decrease the theshold, however, this leads to an increasing probability of erroneously detecting not transmitted ranging codes. Fig. 3 shows the probability p_{wrong} for the same parameters. We can see that p_{wrong} behaves analogous to p_{not} if the thresholds T=90 and T=54 are switched. The reason for this is the symmetry of the binomial distribution. Let us further notice, that the anayltically derived values for p_{not} and p_{wrong} match quite well with the simulation.

3 Simulation model

In this section we give a short description of the simulation model. We consider only the uplink subframe and entirely neglect the downlink. A single base station has a fixed number of C connections to its subscriber stations and within every connection packets of size V arrive with inter-arrival time A.

We set the frame length to 4ms and the uplink subframe has a length of 2ms. The bandwidth is 10MHz leading to 80 OFDM symbols per frame for OFDM and 9 OFDM symbols for OFDMA. We choose 16QAM modulation with 1/2

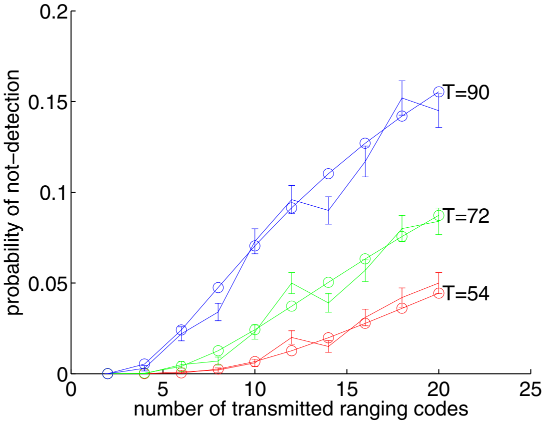

Figure 2. Probability that a transmitted ranging code is not detected.

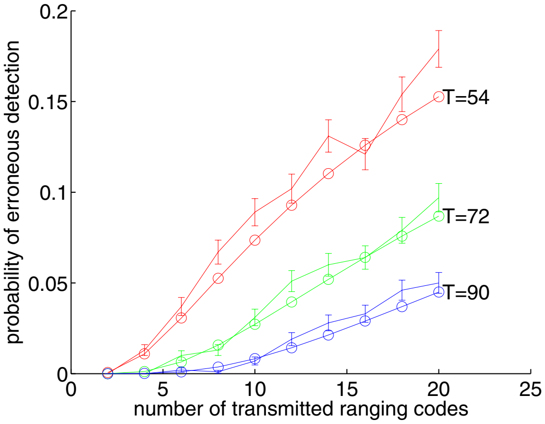

Figure 3. Probability that a not-transmitted ranging code is erroneously detected.

coding for all transmissions and furthermore assume error free communication. Accordingly, the smallest unit of data for OFDM is a single OFDM symbol with 384 bits and for OFDMA a 1×1-region, i.e. a single subchannel on a single OFDMA symbol with 48 bits.

Every connection has a buffer for 100 packets. A subscriber station immediately tries to transmit a bandwidth request after receiving new data. The minimum and maximum backoff values and also the maximum number R_{max} of retransmissions for a bandwidth request are set relative to the number S of bandwidth request or ranging opportunities, respectively.

$$W_{min} = \lfloor log_2(S) \rfloor, W_{max} = W_{min}+8, R_{max} = W_{max}+4$$

After sending a request, a subscriber station waits for $T_{out} = 2$ frames until retransmitting the request. The base station always schedules grants for new bandwidth requests first in order to avoid retransmissions for successfully received requests. The policy for scheduling grants is as follows: First, the base station tries to reduce the overhead due to the 48bit MAC header below 10%, i.e. is tries to grant blocks of at least two OFDM symbols for OFDM and at least ten 1×1-regions for ODFMA. This minimum data burst size limits the number of connections served in a single frame. The base station allocates grants of same sizes to all connections served in one frame. If somes connection are not able to utilize their share completely the excessive part is fairly distributed among the other connections. Finally, if the base station is not able to schedule all connections in a single frame, the serving order is determined by round-robin. A subscriber station might piggyback bandwidth request to already scheduled grants in order to refresh the amount of data waiting for transmission.

4 Simulation Study

In this section we study and compare the performance of the OFDM and OFDMA related MAC layers. Let us first investigate the capacity of the OFDM MAC layer under an optimized number S_{opt} of transmission opportunities. We define the capacity as the maximum traffic load $\rho_{max}(C, Q^*_{0.95,\tau})$ that yields a 95%-quantile $Q_{0.95,\tau}$ of the packet delay τ lower than a maximum value $Q^*_{0.95,\tau}$ when $C = 50$ connections are active. The traffic load is defined as

$$\rho = \frac{C \cdot \mathrm{E}[V] / \mathrm{E}[A]}{B},$$

where B is the theoretic traffic capacity with $B_{OFDM} = 80 \cdot 384 = 30720$ bits per uplink subframe and $B_{OFDMA} = 9 \cdot 60 \cdot 48 = 25920$ bits per uplink subframe. The lower capacity of OFDMA results from the unfavorable sampling rate and a higher number of pilot subcarriers. The maximum traffic load depends also on other parameters like the mean packet size $\mathrm{E}[V]$, and the coefficient of variation of packet size and inter-arrival time. In order to decrease the parameter space we keep these parameters constant as $\mathrm{E}[V] = 12000$ bit and $c_v[V] = c_v[A] = 1$, i.e. packet size and inter-arrival time are exponentially distributed. We considered upper bounds for the 95%-quantile of the packet delay between 20ms and 400ms correpsonding to 5 and 100 frames, respectively. We found the maximum acceptable load with a granularity of 0.05 while simultaneously optimizing the number of bandwidth request transmission opportunities with respect to the 95% packet delay quantile. Fig. 4 shows the maximum loads ρ_{max}, the optimum number S_{opt} of bandwidth request transmission opportunities, and the achieved 95%-quantile of the one way packet delay through the WiMAX network. The right graphic show the relationship between throughput or accepted traffic and QoS expressed as the delay quantile. Obviously, the highest load or throughput is achieved when the packet delay is not constrained. For achieving 95% packet delay of about 100ms the load must be reduced only to around 80% of the theoretical maximum. However, for achieving 95% delay quantiles below 100ms the load has to decrease considerably, delays below 20ms (the left most bar) allow only 25% of the theoretical maximum load. The middle graphic shows that achieving smaller packet delays requires a higher number of bandwidth request opportunities which obviously means that the data capacity of the uplink subframe shrinks.

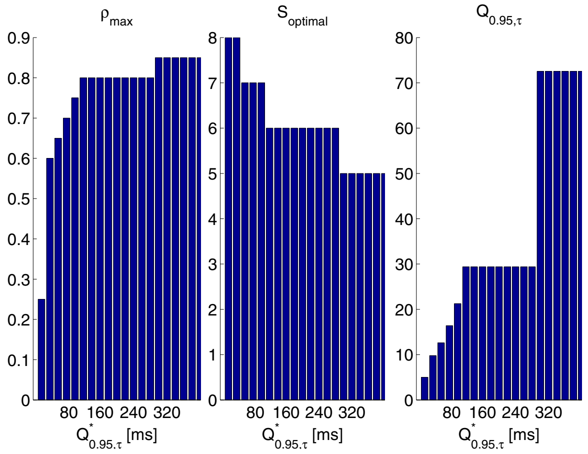

Figure 4. Maximum load ρ_{max} with optimum number S_{opt} of bandwidth request opportunities maintaining the target 95% packet delay quantile $Q^*_{0.95,\tau}$.

In the second study we compare the performance of OFDM, OFDMA, and OFDM with subchannelization. We again choose our default scenario with 50 connections and a mean packet size of 12000 bit. OFDM is set up with 4 trans-

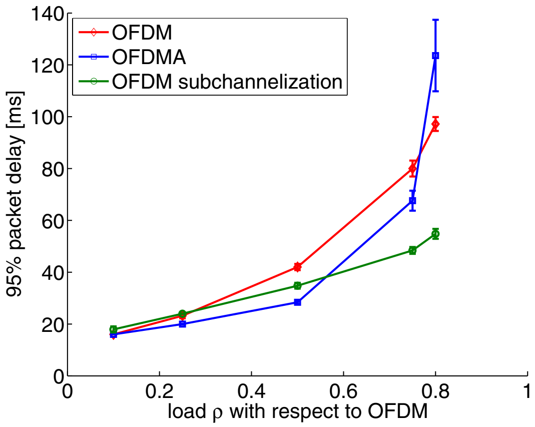

Figure 5. Performance of the different contention schemes.

mission opportunities, OFDMA with 50 ranging codes and 2 ranging code regions, OFDMA with subchannelization with 8 transmission opportunities covering three OFDM symbols. The performance is compared for the 95% quantile of the mean packet delay with a load increasing from 0.1 to 0.8. Note that the load is defined relative to the theoretic capacity of OFDM which is by about 20% larger than the theoretic OFDMA capacity. Fig. 5 shows the 95% packet delays. We can see that for medium load OFDMA performs best. Due to the lower theoretic capacity OFDMA is already close to its limits for a $\rho = 0.8$. For low loads all three schemes show almost equal performance. Subchannelization brings a clear benefit compared to OFDM without subchannelization in particular for the short frame sizes and high loads.

5 Conclusion

In this paper we have studied the performance of the different random access mechanisms present in the IEEE 802.16 standards. We have shown that the amount of resources that should be reserved for the random access strongly depend on the desired performance. If long access delays can be tolerated few resources are enough a higher throughput can be achieved. If however short delays a required as some users transport delay critical traffic over the best-effort connections a considerable part of the uplink subframe should be spent for the random access. Subchannelization and the three-way access through sending a randing code in OFDMA show a better performance and require less resources. In particular with OFDMA the random access mechanism works very efficiently and the resources required for the random access are almost negligible.

References

[1] B. Bhandari, R. Kumar, and S. Maskara. Uplink Performance of the IEEE 802.16 Medium Access Control (MAC) Layer Protocol. In *IEEE International Conference on Personal Wireless Communications 2005 (ICPWC 2005)*, pages 5–8, New Delhi, India, January 2005.

[2] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of Service Support in IEEE 802.16 Networks. *IEEE Network*, 20(2):50–55, April 2006.

[3] IEEE 802.16-2004. IEEE Standards for Local and Metropolitan Area Networks - Part16: Air Interface for Fixed Broadband Wireless Access Systems, 2004. IEEE 802.16-2004.

[4] IEEE 802.16e-2005. IEEE 802.16e-2005 Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems- Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1, February 2006.

[5] R. Iyengar, P. Iyer, and B. Sikdar. Delay Analysis of 802.16 based Last Mile Wireless Networks. In *IEEE Globecom*, St. Louis, MO, USA, November 2005.

[6] J. Sun, Y. Yao, and H. Zhu. Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems. In *Vehicular Technology Conference, 2006. VTC 2006-Spring*, pages 1221 – 1225, Melbourne, Australia, May 2006.

[7] A. Vinel, Y. Zhang, M. Lott, and A. Tiurlikov. Performance Analysis of the Random Access in IEEE 802.16. In *The 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC 2005)*, Berlin, Germany, September 2005.

[8] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov. Efficient Request Mechanism Usage of IEEE 802.16. In *Globecom 2006*, San Francisco, CA, USA, November 2006.