Towards a QoS-Aware IEEE 802.16 Downlink Sub-Frame Mapping Scheme

Panagiotis G. Sarigiannidis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
sarpan@csd.auth.gr

Malamati Louta
Department of Informatics and Telecommunications Engineering
University of Western Macedonia
Kozani, Greece
louta@uowm.gr

Dimitrios G. Stratogiannis and Georgios I. Tsiropoulos
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
dstratog@mail.ntua.gr and gitsirop@mail.ntua.gr

Abstract— **IEEE 802.16 (WiMAX) is considered as one of the most promising wireless access technologies supporting high-capacity and long-distance communications as well as user mobility. A problem that should be addressed in the context of multi-access communications is the efficient accommodation of traffic requests to the downlink subframe. The allocation operation in downlink sub-frame is not standardized, while only restrictions on the usage of downlink sub-frame's bandwidth are defined. The most noticeable restriction is the rectangular restriction, requiring all downlink allocations to be mapped in a two-dimensions rectangular shaping. This study is a first step towards defining a QoS-aware mapping scheme, prioritizing traffic requests in accordance with the strict delay requirements they pose. The mapping scheme applies horizon scheduling, permitting bursts to be scheduled efficiently and in a simple way, following the horizons as pilots. The QoS-aware mapping scheme is evaluated by means of simulation experiments, which indicate that the proposed scheme operates effectively and efficiently, by reducing the number of unserviced users and traffic requests, and the portion of the dropped real-time traffic.**

Keywords-IEEE 802.16; Downlink mapping; OFDMA; QoS; Simulation;

I. INTRODUCTION

Wireless communications and networking technologies have been expected to contribute successfully to the ubiquitous and seamless service provisioning with adequate QoS support. Thus, they have attracted the interest of the research community and industry in recent years. IEEE 802.16 standard (WiMAX) constitutes one of the most promising broadband wireless access technologies, supporting high-capacity, long-distance communications and user mobility [1]. Today, the IEEE 802.16e standard, known as mobile WiMAX, forms the basis for the WiMAX solution for nomadic and mobile applications.

Our focus is laid on efficient services' multiplexing, considered as one of the most interesting technical design challenges of the current IEEE 802.16e standard. Mobile WiMAX, exploits the orthogonal frequency division multiple access (OFDMA), which, in order to meet the multi-user communication problem, applies a multi-access approach with subscribers sharing both subcarriers and timeslots, making, thus, efficient use of the bandwidth available [2].

In the context of the communication process, the WiMAX standard considers a base station (BS) and a number of mobile stations (MSs), where the participants exchange data within specific time periods. Time is organized into fixed frame periods, while the frame is divided into uplink and downlink sub-frames. The bi-directional communication can be realized by applying either frequency division duplexing (FDD) or time division duplexing (TDD). In the FDD technique, uplink and downlink periods use different frequency bands, allowing the simultaneous transmission of both downlink and uplink sub-frames. The TDD technique offers a flexible bandwidth allocation, by allowing a downlink frame followed by an uplink frame after a small guard interval. TDD is favored by a majority of applications, due to a) its flexibility in choosing uplink-to-downlink data rate ratios, b) its ability to exploit channel reciprocity, c) its ability to implement in nonpaired spectrum and d) less complex transceiver design demanded [2] In the current version of this study the authors adopt TDD as well for bi-directional communication support.

The BS has undertaken the responsibility for accommodating all MSs, in both uplink and downlink sub-frames. For the downlink sub-frame, the BS can allocate bandwidth to each MS, based on the needs of the incoming traffic, without involving the MS. For the uplink sub-frame, allocations are based on requests from the MS, by involving several mechanisms and polling schemes by which an MS can send bandwidth requests.

978-1-4244-8865-0/10/$26.00 ©2010 IEEE

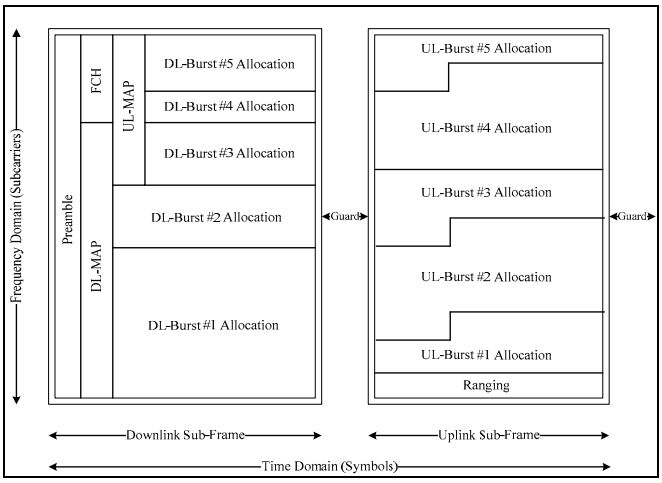

Figure 1. A TDD frame structure of IEEE 802.16 standard

The WiMAX is quite flexible regarding the accommodation of multiple subscribers on a single frame. However, it only defines restrictions on the usage of downlink sub-frame's bandwidth. One of the most noticeable limitations on the downlink allocation procedure is the rectangular restriction. According to this, the MSs' requests have to be formed in a 2D-rectangular fashion, which is called a burst, with one dimension associated with the time domain and the other associated with the frequency domain. Hence, in this context, a problem that should be addressed is the efficient accommodation of the traffic requests on the downlink sub-frame, since this procedure is not standardized and it is left on the WiMAX implementation.

Recent allocation schemes and mapping algorithms have presented various techniques, trying to accommodate 2D-rectangular requests into the downlink sub-frame. To the best of our knowledge, in the related research literature there has not been proposed an allocation scheme that takes into account QoS requirements of the traffic requests. Thus, our study aims to contribute to this research direction by presenting a QoS-aware mapping scheme for the IEEE 802.16 down-link sub-frame. Specifically, a low complexity allocation burst scheme is introduced, applying horizon scheduling, which divides the sub-frame capacity into simple horizons, permitting bursts to be scheduled efficiently and in a simple way, following the horizons as pilots. Traffic requests are served in priority on the basis of strict delay requirements they present. Our QoS aware burst mapping scheme achieves better performance compared to current leading one, as the simulation results indicate under various traffic load scenarios.

The rest of the paper is organized as follows. Section II describes the OFDMA multi access background and Section III briefly revisits past research efforts on the solution of the downlink sub-frame mapping problem. Section IV presents and explains in detail the proposed QoS-aware accommodation scheme and Section V presents a set of simulation results and discusses on the performance of our proposed scheme compared to recent mapping algorithms. Finally, in Section VI conclusions are drawn and our directions for future work are given.

II. OFDMA BASICS

The main feature of OFDMA technique is the multi-access provisioning, enabling the concurrent of multiple subscribers realized by the transmission of numerous subcarriers. In order to achieve this, the OFDMA technique combines time and frequency division multiple access, providing multiple timeslots at different frequencies to multiple users.

The slot represents the minimum allocation structure based on OFDMA. Each slot consists of one subchannel over one, two, or three OFDM symbols, depending on the particular subchannelization scheme used [2]. The TDD operation mode is depicted in Fig. 1.

The length of the frame is predefined and may support multiple time durations from 2 to 20 ms. In TDD format the downstream and the upstream are distinguished in time domain, defining two separate subframes that follow on each other in time. The two sub-frames are separated by a TTG (Transmit to Transmit Gap) in order for the BS to switch from transmit to receive mode and by a Receive to Transmit Gap (RTG) in order for the BS to switch from receive to transmit mode. A preamble is used for time synchronization. Then control information follows, such as the frame control header (FCH), which defines MAP lengths, the downlink map (DL-MAP) along with the uplink map (UL-MAP), which define the burst-start and burst-end time for each subframe and modulation type and forward error control (FEC) for each MS. The FCH field defines the duration of MAP messages and usable subcarriers. The DL-MAP and UL-MAP messages store the bandwidth allocations along with the DL and UL time and frequency organization. DL and UL allocations follow, structured into a downlink and an uplink transmission period respectively.

In regards to the uplink, usually there is one burst per subscriber. For the downlink case, the standard allows more than one burst per subscriber, packing multiple connections into one burst. However, the packing increases the mandatory stored control information in the DL-MAP field. Using or not the packing solution is an open research issue along with the mapping problem, in which the bursts should be treated as rectangular objects, filling the available allocation space, known as bin. Obviously, a rigorous mapping scheme is required to support an effective communication, by allocating the subscribers' burst efficiently.

III. RELATED DOWNLINK MAPPING SCHEMES

The simplest mapping approach is the static predefined burst mapping, according to which requests are allocated in a predefined way. Specifically, S fixed rectangles are accommodated, where S is taken equal to the number of the connected MSs. Beyond the simplicity of this approach, the static mapping scheme leads to inefficient performance, since the predefined allocations are static and may notably vary compared to the subscribers' requests. Thus, valuable bandwidth portion may be wasted. Additionally, a control mechanism is needed to provide knowledge on the incoming requests prior to the mapping procedure.

In [3], the downlink sub-frame is thoroughly scanned in a slot-by-slot basis and the incoming requests are accommodated in one line after the other. This approach seems simple with minimal complexity. However, it suffers in terms of flexibility. Given the fact that the number of incoming requests per frame is not known, each request is accommodated as a single burst and the size of control information increases, yielding, thus, to a reduction in the available space for data allocation.

The scheme presented in [4] introduces full-search mapping tries, until the optimal one is found. A binary-tree full search operation is applied to exhaustively calculate the total possibilities. Such an effort demands for its execution crucial operational time, hence, the authors limit the number of accommodated subscribers to eight per frame.

Efforts in [5] and [6] present similar techniques, which include construction of buckets, where multiple requests are gathered in allocation columns. Each bucket collects requests in the frequency domain with common transmission characteristics. Even though these schemes present better performance compared to the static mapping scheme, the logic underlying the accommodation algorithm leaves significant number of unused slots in case the bucket fails to cover the entire column in the frequency domain.

Finally, in [7, 8] two simple heuristic mapping schemes have been proposed, aiming to keep the mapping operational complexity low. eOCSA [8] mapping scheme schedules each subscriber's request into the downlink sub-frame as an individual downlink burst, resulting in a reduced DL-MAP overhead. As a first step, the algorithm sorts the incoming bursts in a descending order. During the second step, known as vertical mapping, the sorted bursts are accommodated on the basis of a suggested mapping strategy from bottom to top and from left to right. The remaining unallocated space is handled in the third step, during which the horizontal mapping takes place, and where the eOCSA tries to assign the unallocated space to the next largest request that can be accommodated in.

It is worth mentioning that in general, the target of the mapping operation is to accommodate as many as possible subscribers' requests, since the restrictions regarding the capacity of the downlink frame may violate QoS guarantees, resulting in high delays and high packet loss. However, all previous schemes do not consider specific QoS requirements of the traffic requests. This study is a first step towards defining a QoS-aware mapping scheme, prioritizing traffic requests in accordance with the strict delay requirements they pose. The mapping scheme applies horizon scheduling, permitting bursts to be scheduled efficiently and in a simple way, following the horizons as pilots.

IV. Proposed QoS-Aware Mapping Scheme

The proposed mapping scheme has two major targets. Firstly, it aims at providing QoS-aware mapping policy, giving emphasis to real-time traffic streams and secondly allocating efficiently the subscribers' requests to the available allocation bin, using horizon-based accommodation technique. IEEE 802.16 defines five QoS service classes: Unsolicited Grant Scheme (UGS), Real Time Polling Services (rtPS), Non Real Time Polling Service (nrtPS), Best Effort Service (BE), and Extended Real-Time Variable Rate (ertVR) [2]. Each class has different QoS parameters. For instance, rtPS, UGS and ertVR are treated as real-time traffic, following specific guidelines considering the delay, the throughput and the jitter performance metrics. Real-time traffic is considered as more sensitive compared to non real-time traffic, such as the nrtPS and BE. Hence, the introduced mapping scheme prioritizes the servicing of the sensitive real-time traffic streams.

Latency is the most sensitive QoS related issue in real-time traffic. Demanding data as VoIP, Video Conference or interactive gaming are carried by packets with strict delay requirements. Packets received after the deadline are dropped, leading to performance degradation. In this manner, the proposed QoS-aware mapping scheme takes into account the time requirements. If the deadline is shorter than the current frame length then this request is prioritized over all other and it is mapped directly.

Concurrently, the proposed mapping scheme applies Horizon-based allocation, by creating initial pilots for the forthcoming requests. The pilots are created by allocating large requests with minimum remaining idle space. This goal is achieved by defining appropriate dimensions of the 2D-rectangular shaped requests from right to left and bottom to top. Then the second phase follows, in which the remaining requests are mapped based on pilots. This logic offers two benefits: (a) the complexity of the whole process is reduced due to pilots, since the remaining requests follow pre-defined allocation Horizons, receiving specific width based on the selected pilot and (b) the dimensions of each request are selected based on the wasted space that the request creates upon the allocation.

Initially, the introduced scheme detects the sensitive real-time requests with short deadlines. Then, the first phase begins by applying a descending order sorting, giving the opportunity to large real-time requests to be accommodated firstly. Upon the completion of the real-time requests, the mapping algorithm undertakes the non real-time requests and once again applies a descending order sorting to ensure that large requests can find available allocation space. Each accommodation is processed based on the idle (wasted) space that each request leaves behind.

For each request the available bin is examined and the request dimensions are determined. The bin has initial dimensions $H \times W$ and apparently the first request for each frame can be accommodated in an available space equal to $H \times W$. According to standard restrictions the slots of each request should be shaped in a rectangular form, hence the next step is the determination of request dimensions (rectangle's width and height). Algorithm1 shows the mapping procedure during the first phase:

Algorithm 1 Defining Horizons and right-to-left Mapping

- Set initial available bin width equal to W.
- Set initial available bin height equal to H.
- **DO**
- Define the collection set of incoming downlink real-

time requests, which have shorter deadline than the frame length: $S = \{A_1, A_2 ...\}$.

- Find and select the request A with largest slots. Let A_{slots} stands for the number of slots that A requests.
- Set its width, denoted by A_{width}, equal to

$$\arg\min_{A_{width} \in [1,W]} (A_{width} \times \left\lceil A_{slots} / A_{width} \right\rceil - A_{slots}).$$

- Set its height, denoted by A_{height} equal to

$$A_{height} = \left\lceil A_{slots} / A_{width} \right\rceil.$$

- Allocate the slots of request A into the bin with upper left point coordinates equal to: $A_x = W - A_{width}, A_y = H - \left\lceil A_{slots} / A_{width} \right\rceil)$.
- Request A forms an Horizon. Set Horizon's height, denoted by HZ_{height}, equal to:

$$HZ_{height} = H - \left\lceil A_{slots} / A_{width} \right\rceil.$$

- Set Horizon's width, denoted by HZ_{width}, equal to:

$$HZ_{width} = A_{width}.$$

- Update the available bin width: $W = W - A_{width}$
- Remove request A from the S set.
- Find and select the (next) request A with shorter deadline.
- **WHILE** ($H \times W \geq A_{slots}$ **AND** $S \neq \varnothing$)
- Define the collection set of incoming downlink non real-time requests $S' = \{A'_1, A'_2 ...\}$ and repeat the do-while loop for *S'*.

The first phase ends and a set of Horizon pilots have been defined from either real-time or non real-time requests. By defining width and height, each Horizon indicates an unallocated 2D-rectangular region. Then the second phase begins, in which the algorithm accommodates the remaining requests into the Horizon regions (Algorithm 2).

Algorithm 2 Bottom-to-top mapping of the remaining bursts

- Define the collection set of remaining unmapped real-time requests, which have shorter deadline that the frame length $S = \{A_1, A_2 ...\}$.
- **DO**
- Find and select the request A with largest slots. Let A_{slots} stands for the number of slots that A requests.
- Find the appropriate Horizon to accommodate request A: For each Horizon that it is large enough to enclose the request A select the $HZ^{selected}$ that minimizes the remaining wasted slots:

$$\arg\min_{selected} (HZ^{selected}_{width} \times \left\lceil A_{slots} / HZ^{selected}_{width} \right\rceil MODA_{slots})$$

- Set request's width equal to:

$$A_{height} = \left\lceil A_{slots} / HZ^{selected}_{width} \right\rceil$$

- Set request's height equal to: $A_{width} = HZ^{selected}_{width}$
- Update $HZ^{selected}$ dimensions.
- Remove request A from the S set.
- Find and select the (next) request A with shorter deadline.
- **WHILE** (there are available Horizons with $HZ_{width} \times HZ_{height} \geq A_{slots}$ **AND** $S \neq \varnothing$)
- Define the collection set of incoming downlink non real-time requests $S' = \{A'_1, A'_2 ...\}$ and repeat the do-while loop for *S'*.

V. Performance Evaluation and Simulation Results

The proposed QoS-aware scheme has been evaluated by a set of simulation experiments. Additionally, the performance of the proposed scheme is compared with the performance of the eOCSA mapping scheme [8], which has been chosen due to its low complexity, its common burst construction technique, and its efficient performance compared to past mapping schemes. The considered simulation assumptions are shown in Table I.

Assuming that the partially used sub-channelization (PUSC) mode is considered, the downlink sub-frame defines 30 channels. The downlink-to-uplink sub-frame ratio is fixed and equal to 2:1. The frame size is also fixed and set to 10 ms, allowing 95 symbols to attach to the downlink and the uplink sub-frames. Three symbols are destined to control information (Preamble, MAP and FCH fields) and are excluded from the available slots for allocation needs.

The downlink traffic load follows a Poisson process. In order to ascertain a realistic environment the subscribers are divided into four equal groups. Each group includes 25% of total subscribers. The first group produces very light traffic, so the parameter λ of the Poisson process is set to 40.

The second group creates light traffic with λ equal to 70, the third group produces medium load and the λ parameter is set to 100 and finally group four presents high load and the λ stands equal to 130. Each request has 25% possibility to be

TABLE I. Simulation Assumptions

Channel Mode	PUSC
Frame Size	10 ms
Preamble Size	1 Symbol
MAP, FCH Sizes	2 Symbols
Downlink sub-frame Symbols	27 (2:1 downlink-to-uplink ratio)
Total sub-frame capacity	810 slots
Frame Iterations (Trials)	2000

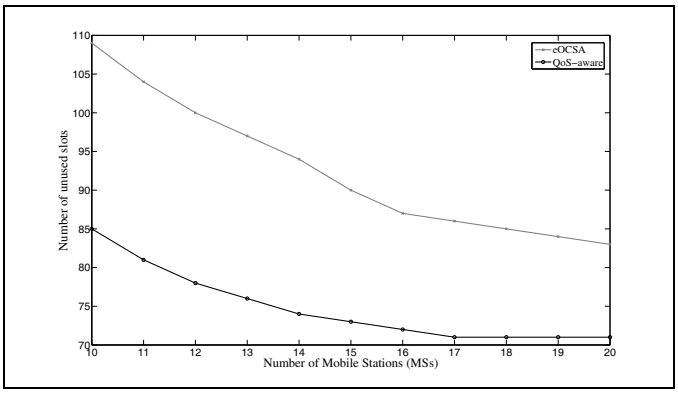

Figure 2. Mean number of unused slots per frame vs. number of MSs

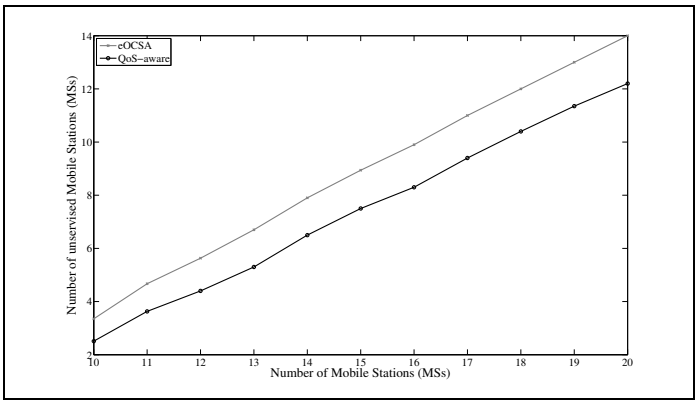

Figure 3. Mean number of unserviced MSs per frame vs. number of MSs

real-time with sensitive deadline. Each experiment has been conducted for 2,000 frames.

In the first set of experiments the amount of wasted bandwidth is examined in terms of the number of unused slots. The number of unused slots is measured by calculating the number of slots being idle in the downlink sub-frame allocation space. Figure 2 shows the performance of the two compared schemes, in terms of unused slots. Obviously, the proposed scheme reduces the number of unused slots, due to its horizon-based nature.

The mean number of unserved MSs per frame is shown in Figure 3. Again, the number of connected MSs varies from 10 to 20. It is clear that the QoS-aware scheme succeeds to serve more users than eOCSA, as the mapping logic is more efficient and permits larger number of requests to be mapped.

Finally, the drop ratio of the real-time traffic is depicted in Figure 4. Once more the suggested scheme presents better performance compared to the eOCSA, minimizing the number

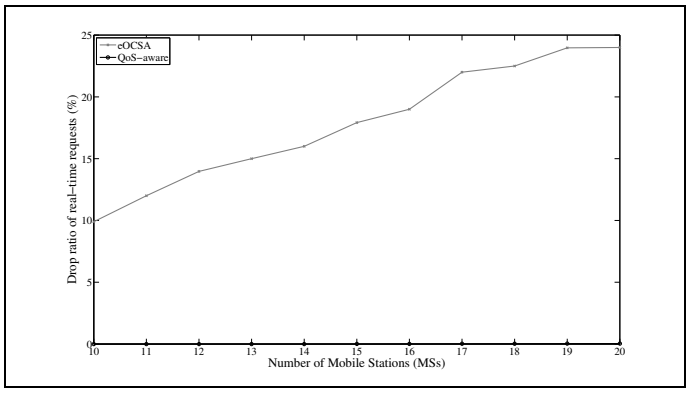

Figure 4. Drop ratio of real-time traffic vs. number of MSs

of dropped sensitive traffic. In this manner, it offers guaranteed QoS, allowing sensitive applications to be transferred adequately.

VI. CONCLUSION

A novel QoS-aware mapping scheme has been presented in this paper. The proposed scheme introduced a new horizon-based allocation technique, while it prioritizes the sensitive real-time traffic. Simulation results indicate that the introduced scheme presents better performance than the current leading one, in terms of unused slots, unserved users and real-time traffic drop ratio. Our next step focuses on the fairness issue, which is one of the most important factors, considering the selection of the telecommunication company features.

REFERENCES

[1] M. Katz and F. Fitzek, WiMAX Evolution: Emerging Technologies and Applications, Wiley Publishing, 2009.

[2] J. Andrews, A. Ghosh, R. Muhamed, Fundamentals of WiMAX, Understanding Broadband Wireless Networking, Prentice Hall, 2007.

[3] Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, "Two-Dimensional Mapping for Wireless OFDMA Systems," *IEEE Trans. Broadcast.*, vol. 52, no. 3, pp. 388-396, Sep. 2006.

[4] C. Desset, E. B. de Lima Filho, and G. Lenoir, "WiMAX Downlink OFDMA Burst Placement for Otimized Receiver Duty-Cycling," in Proc. *IEEE ICC*, 2007, pp. 5149-5154.

[5] T. Ohseki, M. Morita, and T. Inoue, "Burst Construction and Packet Mapping Scheme for OFDMA Downlinks in IEEE 802.16 Systems," in Proc. *IEEE Globecom*, 2007, pp. 4307-4311.

[6] T. Wang, H. Feng, and B. Hu, "Two-Dimensional Resource Allocation for OFDMA System," in Proc. *IEEE ICC*, 2008, pp. 1-5.

[7] C. So-In, R. Jain, and A. Al-Tamimi, "OCSA: An Algorithm for Burst Mapping in IEEE 802.16e Mobile WiMAX Networks," in Proc *APCC*, 2009, pp. 52-58.

[8] C. So-In, R. Jain, and A. Al-Tamimi, "eOCSA: An Algorithm for Burst Mapping with Strict QoS Requirements in IEEE 802.16e Mobile WiMAX Networks," in Proc. *WD*, 2009, pp. 1-5.