On the Integration of Scheduling and Allocation in OFDMA-Based WiMAX Networks

Cristiano Both[1,2], Rafael Kunst[1], Enzo Mingozzi[3], Lisandro Granville[1], Juergen Rochol[1]
[1]Federal University of Rio Grande do Sul (UFRGS). Porto Alegre, Brazil
[2]University of Santa Cruz do Sul (UNISC). Santa Cruz do Sul, Brazil
[3]University of Pisa. Pisa, Italy
Email: {cbboth, rkunst, granville, juergen}@inf.ufrgs.br, e.mingozzi@iet.unipi.it

***Abstract*— Mobile devices and real-time applications are increasingly demanding metropolitan network access with guaranteed quality of service. Current research points to the need of two main stages to provide quality of service in OFDMA-based WiMAX networks. These stages, called scheduling and allocation are deeply investigated on the literature. However, details on the integration of these mechanisms are not provided. Therefore, in this paper, we propose the integration of scheduling and allocation through a threshold criteria that considers the diversity of the traffic, limiting the amount of data and mapping overhead sent to allocation. The results show that our proposal is fundamental to integrate scheduling and allocation and is able to improve quality of service parameters, such as data allocation and delay.**

***Keywords*— Scheduling, allocation, quality of service, WiMAX**

I. INTRODUCTION

A Wireless Metropolitan Area Network (WMAN) typically serves a large number of mobile and nomadic users, which are distributed along a wide geographical area. Because of the increasing number of mobile devices and real-time applications, modern WMANs are required to provide strict guarantees of Quality of Service (QoS). A considerable amount of network resources (*e.g.*, channel bandwidth) have to be compromised to improve the network throughput so that the required QoS guarantees are achieved.

IEEE 802.16 [1], also known as Worldwide Interoperability for Microwave Access (WiMAX), is a WMAN standard that defines the employment of sophisticated transmission technologies, such as Orthogonal Frequency Division Multiple Access (OFDMA). The implementation of OFDMA is mandated in mobile WiMAX networks compliant with the IEEE 802.16e amendment, also known as mobile WiMAX. OFDMA has been considered a proper solution because of its resilience to physical impairments, resulting from the division of frequencies into groups of orthogonal subcarries called subchannels.

OFDMA technique optimized the use of the available network resources by supporting the transmission of multiple users in the same symbol. User's traffic is encapsulated within a frame that can be viewed as a bidimensional matrix divided in uplink (UL) and downlink (DL) subframes. The IEEE 802.16 standard specifies details on how traffic must be organized in UL subframe, while in DL, some details are left open on purpose to allow technological advancements and vendors specific solutions. Therefore, a two-stages approach may be considered, where scheduling and allocation algorithms must be designed for organizing DL transmissions. These algorithms should work in an integrated way to avoid QoS degradation.

In the recent past, investigations on scheduling and allocation algorithms for mobile WiMAX networks were mostly based on optimizing the use of the available resources, such as bandwidth, considering signal propagation conditions [2]. More recently, studies have analyzed scheduling and allocation algorithms also considering the OFDMA frame structure, although still in a non-integrated way, *e.g.*, focusing solely on the scheduler [3] or proposing solutions for the allocation problem [4] without considering the scheduler. To the best of our knowledge, the only solution, which is called Micro and Macro scheduling, proposed an integrated architecture to solve the scheduling and allocation problems [5]. However, in such a proposal, details of the integration process, the amount of traffic exchanged between scheduling and allocation stages, and the overhead caused by the integration processes have not been addressed.

In this paper we introduce a solution where the WiMAX scheduling and allocation algorithms operate in an integrated manner. We propose a threshold criteria to consider the traffic diversity by limiting the amount of data and mapping overhead sent from the scheduling to the allocation stage. The main contribution of our solution is to calculate the optimal amount of traffic to improve the efficiency of allocation algorithms in arranging data in the OFDMA DL subframe. As a consequence, the implementation of the proposed threshold criteria improves the overall network QoS by optimizing bandwidth usage for all classes of service and guaranteeing delay requirements of real time. To evaluate the proposed integration, we simulate transmissions of voice, video, and data traffics, following the specification of the system evaluation methodology defined by the WiMAX forum [6].

The remainder of this paper is organized as follows. In Section II we review the related work on scheduling and allocation algorithms considering the OFDMA frame structure defined in the IEEE 802.16 standard. In Section III we present our proposal of integrating the scheduling and allocation stages. In Section IV we evaluate our solution and discuss the associated results. Finally, we close this paper in Section

V, where conclusions and future work are discussed.

II. Related Work

Current research show the need of scheduling and allocation to provide guaranteed QoS in WMANs. Although the IEEE 802.16 standard [1] defines the role of each stage, it does not specify details about how these stages must be designed and integrated. In this section, we review related work on the scheduling and allocation, as well as their integration.

A. Scheduling proposals

Scheduling stage can be either channel-unaware or channel-aware [3]: the former does not take the Radio Frequency (RF) channel propagation conditions into account, while the latter explicitly considers these conditions. In the case of mobile WiMAX, channel-aware scheduling is more adequate because it explores multiuser diversity to increase the overall throughput and supports a high number of users, what results in a better performance.

This stage can be characterized according to the number of scheduling levels. Typically, an one-level structure, with a single queuing discipline that attends all scheduling services, is used [7]. However, such structure is usually devoted to a specific problem and ignores the assembly of the OFDMA frame. Scheduling stage can present a hierarchical structure of two levels, called Intra-class and Inter-class. The first organizes the traffic within the queues of each class of service using traditional scheduling algorithms, such as Round Robin (RR), and Weighted Fair Queuing (WFQ). The second defines the order in which Intra-class scheduling queues should be served applying standard queuing-disciplines as Priority Fair (PF) and Deficit Round Robin (DRR).

Currently, the proposals on channel-aware scheduling are focused on developing algorithms that optimize the provisioning of QoS guarantees. However, the scheduling stage alone cannot guarantee the QoS requirements for different classes of service because the user's traffic must be allocated within the OFDMA frame. Therefore, these guarantees are also influenced by the allocation stage.

B. Allocation proposals

Data must be allocated in slots that compose the DL subframe. These slots are filled using bursts with rectangular shape that are grouped according to their Modulation and Coding Scheme (MCS). Cicconetti *et al.* [4] classify the allocation stage as (i) sequential and (ii) non-sequential. In sequential allocation, data bursts are allocated in the same order that are received from the scheduling stage. The allocation starts at the top of the frame and continues from the left to the right, as shown in Fig. 1 (a). Non-sequential allocation chooses the data bursts that best fit in the subframe, as show in Fig. 1 (b).

Mobile WiMAX networks operates with heterogeneous traffic and high variability of RF, which lead to bursts of variable sizes. Due to this diversity, non-sequential allocation has better performance than the sequential one because of the possibility

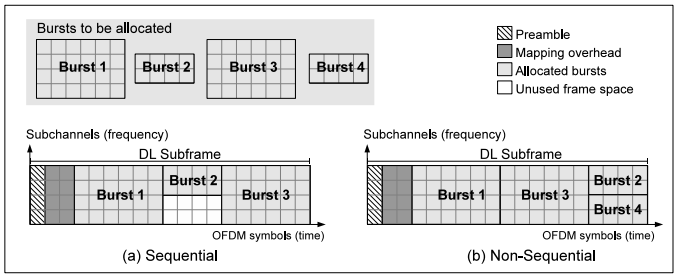

Fig. 1. Classifications of the Allocator Component

of freely arranging these bursts within the subframe. Three main non-sequential proposals have been presented in the literature. The Micro Scheduling Problem (MiSP) [5] uses a group of bin-packing algorithms to present a solution which amount of allocated information is at least as high as that of the optimal solution. The One Column Striping with non-increasing Area first mapping (OCSA) [8] is a bidimensional rectangular burst mapping algorithm composed of three steps. First, the set of data bursts to be allocated is sorted in descending order with respect to its size. Second, the largest element is allocated in a vertical manner. Third, the algorithm allocates the largest data burst that fits in the left-over area of the DL subframe. The last non-sequential proposal is Recursive Tiles and Stripes (RTS) [4], which is composed of two phases. First, an iterative selection of a subset of slots (S) used by n bursts not exceeding the capacity of the frame (L). Second, an inner packing algorithm allocates S into a set of sub-bursts.

Although many proposals consider scheduling and allocation in isolated approaches, the integration between these stages is fundamental to increase the performance of the overall IEEE 802.16 network. The main integration proposal is thus reviewed in the following subsection.

C. Integration between scheduling and allocation proposal

Cohen and Katzir presented a solution [5] where the OFDMA scheduling problem is addressed assuming that the BS determines in advance the physical profile to be used for each Protocol Data Unit (PDU), as well as the gain obtained by transmitting this PDU using its selected physical profile. That is performed in two phases. In the first phase, referred to as Macro scheduling, the BS determines which of the PDUs will be selected for transmission. In the second phase, called Micro scheduling, the scheduling algorithm determines how to build the OFDMA frame using the selected PDUs. The Micro scheduling is considered an allocator because of the functionality of this second phase. The algorithms performance is analyzed separately. Moreover, the mapping overhead is not considered, what may lead to overestimation of available network resources.

The related work shows that some important features of the IEEE 802.16 technology are ignored in the design of an integration between scheduling and allocation stages, such as (i) diversity on both traffics and RF conditions, (ii) overflows in the data flow between these stages, and (iii) overall QoS guarantees, considering the classes of service defined by the standard. In this context, an integration of scheduling

and allocation stages in OFDMA-based WiMAX networks is presented in the following section.

III. Integration of scheduling and allocation in OFDMA-Based

The scheduling stage is responsible for selecting a set of PDUs ($Npdu$) of different traffics to send to the allocation stage. However, an integration module needs to be proposed to define a threshold criteria that allows variations on the size of the set $Npdu$. In other words, it is necessary to allow more variability of bursts size to the arrangement algorithm to enable a better allocation efficiency. To design the threshold criteria it is first necessary to calculate the overhead of the OFDMA frame. This overhead cannot be ignored when calculating the amount of the PDUs to send to allocation, because otherwise the scheduling may overflow the allocation.

A. Mapping Overhead

The size of the DL_{MAP} is proportional to the number of connections currently active and to the MCS, *i.e.* the amount of burst profiles carried within the OFDMA DL subframe. The structure of the DL_{MAP} is composed of a header with fixed length, that we call DL_{head} and several Information Element (IE) headers, that we name DL_{IE}. A DL_{IE} is composed of multiple Connection Identifier (CID) fields, belonging to connections associated with a MCS level. All the connections belonging to a given IE are part of a same burst profile. In our approach, the information regarding the CID is called DL_{CID}.

Considering the structure of the DL_{MAP}, we define equation (1). This equation permits us to obtain the size of the DL_{MAP} for a given frame, taking into account that a DL_{IE} is composed of a variable number of DL_{CID}.

$$DL_{MAP} = DL_{head} + \sum_{j=1}^{N} \left(DL_{IE(j)} + \sum_{k=1}^{Cj} DL_{CID(j,k)} \right) \quad (1)$$

Where, $j \in$ {1, 2, ... N} is the MCS level index, N is the number of MCS levels, $k \in$ {1, 2, ... C} is the number of DL connections scheduled in a DL_{IE}, and C is the number of CIDs. DL burst is a set of $Npdu$, which MCS levels are the same. Therefore, the number of DL_{IE} in a DL_{MAP} is the number of different MCS levels used for transmitting the scheduled $Npdu$.

The size of the UL_{MAP} depends only on the number of CIDs. Each CID is associated with a burst and its specific MCS. This association, called UL_{UIUC} is made by a UL_{IE}, that carries information about resources allocated. To calculate the size of the UL_{MAP} we define equation (2).

$$UL_{MAP} = UL_{head} + \sum_{p=1}^{R} \left(UL_{IE(p)} + UL_{UIUC(p)} \right) \quad (2)$$

Where, $p \in$ {1, 2, ... R} is the index of the allocated resource, and R is the number of resources allocated with an UIUC association. The UIUC should be used to define the type of UL access, *i.e.* Code Division Multiple Access (CDMA) bandwidth request, CDMA ranging, CDMA allocation IE, different burst profiles, among others. Therefore, each type of UL_{UIUC} has variable size.

The mapping overhead is composed of other structures, such as DCD and UCD. The organization of DCD and UCD is similar to the organization of DL_{MAP} and UL_{MAP}. Although we consider these structures in our proposal, we do not present details on the structure to avoid repetitions.

To transmit PDUs, maps must be inserted into the frame. Therefore, the bits of each PDU must be arranged in slots. The transmission capacity of the each slot is defined according to the number of OFDMA symbols of each subframe, which is calculated considering the total duration of a symbol and the duration of the subframe. In mobile WiMAX networks, the mapping overhead is formed by two kinds of information. The first information, in our approach, is named static map overhead (Ψ), and is composed of Preamble, Frame Control Header (FCH), and the DL and UL headers, as presented in equation (3).

$$\Psi = Preamble + FCH + DL_{head} + UL_{head} \quad (3)$$

The static map overhead has fixed length, because its value depends on network parameters that do not change dynamically. The second kind of information, called dynamic map overhead (Φ), varies in time. This variation happens due to the MCS level and to the number of connections scheduled in a DL subframe. The variation is affected by the amount of resources allocated in the UL subframe. In equation (4) we consider the channel descriptors and the headers of DL and UL IEs, as well as their relationship with the CIDs and UIUC.

$$\Phi = DL_{MAP} - DL_{head} + UL_{MAP} - UL_{head} + \sum_{j=1}^{N} \left(DCD_{(j)} + UCD_{(j)} \right) \quad (4)$$

B. Threshold Criteria

After calculating the static and dynamic overhead, the threshold criteria must convert the mapping overhead, and PDUs to the allocation unit, *i.e.*, slots. The number of slots (N_{slots}) associated with the amount of data that should be transmitted in a subframe is defined in equation (5). The N_{slots} value is obtained considering the quantity of bits per modulation symbols (M) and the FEC encoding (K) for each m^{th} and s^{th} slots of the subframe, which is able to carry mapping control information, and PDUs. The calculation of N_{slots} must consider the amount of symbols per slots (S_{slot}) and the number of subcarries per subchannel (N_{sc}).

$$N_{slots} = \left\lceil \frac{(\Psi + \Phi).M_m.k_m + \sum_{s=1}^{Q} \left(\sum_{i=0}^{P_s} PDU_{(s,i)} \right).M_s.K_s}{S_{slot}.N_{sc}} \right\rceil \quad (5)$$

Where, m and $s \in$ {1, 2, ... Q} are the indexes of Q slots in the OFDMA subframe, $i \in$ {0, 1, ... P_s} is the index of a PDU in a slot s.

After converting the amount of bits to the number of slots of the subframe, the threshold criteria must satisfy inequation (6), that selects a set of PDUs ordered by the scheduler component to be allocated considering the capacity of the DL subframe (L) in units of slots. The allocator component, receives the set of PDUs, and is responsible for assembling the frame to be transmitted. PDUs that are not allocated due to arrangement algorithm in the OFDMA frame remain in the first position of scheduling queues to be served in the next frame, *i.e.* the integration module assures that the QoS requirements are regarded.

$$N_{slots} \leq L + \Lambda \tag{6}$$

We use an adjustment factor (Λ) that allows to vary the amount of PDUs, in unit of slots, to be sent from the scheduler to the allocator. This factor provides a bigger variability of bursts size to the arrangement algorithm and to enable a better allocation efficiency. For example, let us consider L equal to 420 slots. When Λ equal to 0% means that the threshold criteria associates N_{slots} equal to L. In this case, the algorithm should not use all available space within OFDMA subframe due to little variability of bursts size for the arrangement. On the other hand, Λ equal to 10% means that the threshold criteria allows the scheduling to send 42 slots exceeding L to the allocation. In this case, the algorithm receives a higher diversity of burst size and, consequently should have a better allocation efficiency.

IV. Simulation Results

In this section, we analyze the performance of the integration between scheduling and allocation. Three traffic models were developed to represent VoIP, video, and HTTP applications. These models and the physical configurations of the simulation scenario are based on the *System Evaluation Methodology* document, published by the WiMAX Forum [6]. Furthermore, we have considered that each class of service generates traffics in same proportion. All the results were obtained with a confidence interval of 95%.

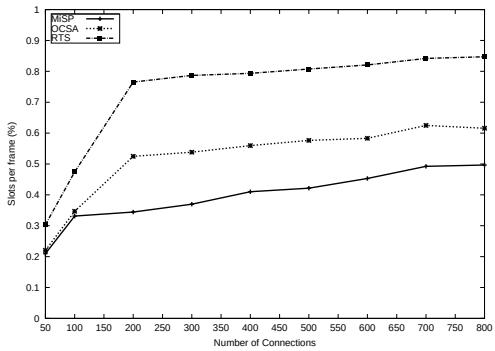

Fig. 2. Performance of the allocation algorithms

The first investigation shows the efficiency of three allocation algorithms: MISP, OCSA, and RTS in a scenario composed of the PF algorithm in the Inter-class scheduling, using no Λ was considered. PF algorithm was chosen to prioritize VoIP traffic in detriment to others. Fig. 2 shows that in scenarios with more than 200 connections, RTS allocates approximately 25% more data than OCSA algorithm and around 18% more data than MiSP algorithm. Otherwise, in scenarios with less than 200 connections, the total amount of traffic is smaller than L. Thus, the allocation algorithms present a linear increase in terms of allocated data before reaching their maximum efficiency.

In the previously described simulation scenario, DL and UL mapping overhead were also analyzed. The overall overhead was of approximately 10%, where the UL mapping generates about 3% more overhead than the DL mapping. This behavior occurs because UL transmissions cannot associate a set of connections with a same burst profile. These associations are not possible because the medium access is shared among the MSs in UL transmissions. The mapping overhead depends on the number of ongoing connections and the burst profiles. However, due to L, the average amount of connections and burst profiles are constant within OFDMA frames. Consequently, the average overhead also is constant. In this context, the next investigations consider the mapping overhead.

The goal of the second analysis is to investigate whether the variation of Λ improves the efficiency of the allocation algorithm. We chose RTS algorithm for this analysis because it had the best data allocation performance. Fig. 3 shows the RTS efficiency with three Λ values, considering that eventually unallocated bursts are re-inserted into the scheduling queues.

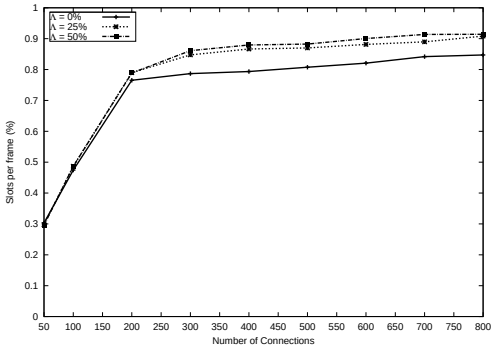

Fig. 3. Adjust factor performance

The variation of Λ does not represent any changes on RTS algorithm's performance in scenarios with less than 200 connections because the amount of traffic generated in this situation is smaller than L. Otherwise, when there are more than 200 active connections, the scheduling queues receive amounts of data that exceed L, thus changes in the Λ value improves the performance of the RTS algorithm. We can observe gains of approximately 8% due to bigger variability of bursts size granted to the RTS algorithm. However, Λ equal to 25% e 50% do not present a significant gain because the efficiency is not proportional to the increase on the Λ value, since RTS algorithm presents a limit on the data size that can be arranged within the DL subframe.

To deeper investigate the Λ performance, another simulation scenario with VoIP as predominant traffic was analyzed. The proportion of the traffic configured is of 70% VoIP, 15% Video, 15% HTTP. The predominant VoIP traffic is considered in this investigation because it is composed of small PDUs that can

be better arranged by the allocation algorithm within OFDMA subframe. Figure 4 presents a gain of about 11% when the allocation algorithm received a bigger variability of the data to be arranged within the DL subframe. Furthermore, it is possible to observe that the RTS algorithm used approximately all L in the scenario with 800 connections, *i.e.* the RTS algorithm almost reached the ideal efficiency in the data allocation within OFDMA frame.

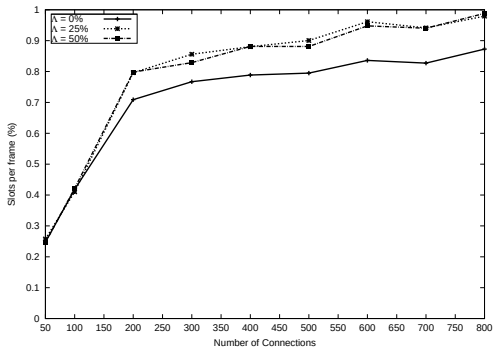

Fig. 4. Adjust factor performance

Another important investigation is the delay caused by the integration between scheduling and allocation stages considering VoIP transmissions. Fig. 5 shows that RTS algorithm normally allocates the VoIP PDUs in the first frame after the PDU arrives. This behavior is observed with Λ equal to 50%. On the other hand, OCSA and MiSP algorithms tend to increase the delay after 600 connections. This number of connections indicates a threshold until which the maximum delay efficiency is guaranteed.

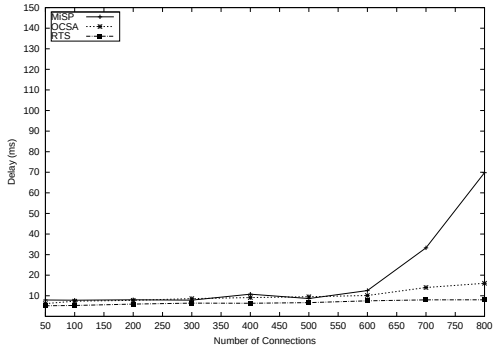

Fig. 5. Delay analysis for VoIP traffic

Our last investigation presents the delay of video traffic, also considering Λ equal to 50%. Fig. 6 shows that if compared with VoIP transmission, video presents a higher delay, since it has lower priority. Furthermore, we can observe that RTS algorithm has the best performance. The higher difference between the best and the worse allocation algorithm is observed when there are 300 active connections, where RTS serves the video queue about 20ms before MiSP, what leads to a gain of approximately 4 frames.

V. Conclusions and Future Work

In this paper, we investigated the integration between scheduling and allocation stages, designed to provide QoS guarantees in WMAN networks. The results showed the performance of the proposed integration, considering the threshold criteria that allows a higher diversity of the traffic sent to the allocation stage. Our analysis permits to conclude that the proposed integration is fundamental to provide QoS. We conclude that the adjustment factor of the threshold criteria improves the efficiency of the allocation algorithms and consequently guarantees the delay requirements defined by the WiMAX Forum. Directions for future investigations include considering also the connection admission control system that limits the number of accepted connections.

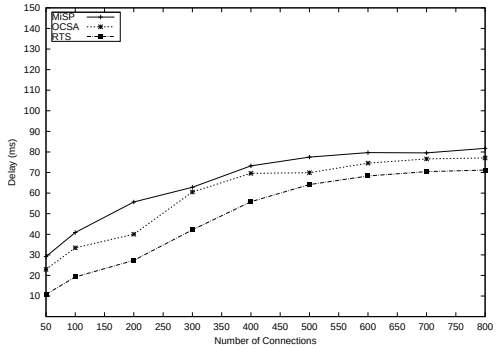

Fig. 6. Delay analysis for video traffic

Referências

[1] IEEE, "IEEE standard for local and metropolitan area networks, part 16 - air interface for broadband wireless access systems," May 2009, IEEE 802.16-2009. in http://standards.ieee.org/getieee802/download/802.16-2009.pdf. Accessed in March 2010.

[2] J. Lu and M. Ma, "A cross-layer elastic CAC and holistic opportunistic scheduling for QoS support in WiMAX," *The International Journal of Computer and Telecommunications Networking*, vol. 54, no. 7, pp. 1155–1168, 2010.

[3] C. So-In, R. Jain, and A.-K. Tamimi, "Scheduling in IEEE 802.16e mobile WiMAX networks: Key issues and a survey," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 2, pp. 156–171, February 2009.

[4] C. Cicconetti, L. Lenzini, A. Lodi, S. Martello, E. Mingozzi, and M. Monaci, "Efficient two-dimensional data allocation in IEEE 802.16 OFDMA," in *Proceedings. IEEE INFOCOM - Conference on Computer Communications*, San Diego, March 2010, pp. 1–9.

[5] R. Cohen and L. Katzir, "Computational analysis and efficient algorithms for Micro and Macro OFDMA downlink scheduling," *IEEE/ACM Transactions on Networking*, vol. 18, no. 1, pp. 15–26, February 2010.

[6] WiMAX Forum, "WiMAX system evaluation methodology version 2.1," July 2008, WiMAX Forum. in http://www.wimaxforum.org/documents. Accessed in March 2010.

[7] I. C. Msadaa, D. Câmara, and F. Filali, "Scheduling and CAC in IEEE 802.16 fixed BWNs: A comprehensive survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 4, pp. 459–487, 2010.

[8] C. So-In, R. Jain, and A.-K. A. Tamimi, "OCSA: An algorithm for burst mapping in IEEE 802.16e mobile WiMAX networks," in *Proceedings. (APCC) 15th Asia-Pacific Conference on Communications*, Shanghai, October 2009, pp. 52–58.