A Survey on Next Generation Mobile WiMAX Networks: Objectives, Features and Technical Challenges

Ioannis Papapanagiotou, *Graduate Student Member*, Dimitris Toumpakaris, *Member*, Jungwon Lee, *Member*, Michael Devetsikiotis, *Senior Member*

***Abstract*—In order to meet the requirements of 4G mobile networks targeted by the cellular layer of IMT-Advanced, Next Generation Mobile WiMAX devices based on IEEE 802.16m will incorporate sophisticated signal processing, seamless handover functionalities between heterogeneous technologies and advanced mobility mechanisms. This survey provides a description of key projected features of the physical (PHY) and medium access control (MAC) layers of 802.16m, as a major candidate for providing aggregate rates at the range of Gbps to high-speed mobile users. Moreover, a new unified method for simulation modeling, namely the Evaluation Methodology (EVM), introduced in 802.16m, is also presented.**

***Index Terms*—IEEE 802.16m, Mobile WiMAX Networks, Next Generation Wireless Networks, Broadband Wireless Access, Evaluation Methodology.**

I. INTRODUCTION

DESPITE the challenges faced when transmitting data through varying wireless channels, broadband metropolitan area wireless systems are becoming a reality, partly thanks to the increasingly sophisticated designs that are being employed. Such designs have been made possible by theoretical advances and also by improvements in technology that have led to faster and cheaper implementations compared to older systems. Currently, the focus is on developing 4G systems in the framework of IMT-Advanced [1], an ITU platform on which the next generation of wireless systems will be built. This paper is a survey of some of the state-of-the-art characteristics of the physical (PHY) and medium access control (MAC) layers of 802.16m, one of the major candidates for Next Generation Wireless Systems. The concepts and algorithms that are implemented in 802.16m will very likely be used in IMT-Advanced systems, either as they will appear in the 802.16m standard or in similar, and possibly more sophisticated versions. Therefore, it is of interest not only to present these concepts and the ways that they address the needs of future systems, but also to identify technical challenges that will have to be tackled in order for Next Generation Wireless Systems to meet the objectives set by IMT-Advanced.

Manuscript received 9 October 2008; revised 8 February 2009.

I. Papapanagiotou and M. Devetsikiotis are with the Department of Electrical & Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA. e-mail: {ipapapa, mdevets}@ncsu.edu.

D. Toumpakaris is with the Wireless Telecommunications Laboratory, Department of Electrical & Computer Engineering, University of Patras, 265 00 Rio Achaias, Greece. e-mail: dtouba@upatras.gr.

J. Lee is with Marvell Semiconductor, Santa Clara, CA 95054, USA. e-mail: jungwon@stanfordalumni.org

Digital Object Identifier 10.1109/SURV.2009.090402.

IMT-Advanced is the continuation of IMT-2000, the global standard for 3G wireless communications. The goal of IMT-2000 was "to provide a framework for worldwide wireless access by linking the diverse systems of terrestrial and/or satellite based networks" [2]. IMT-2000 comprised a range of activities both inside and outside the ITU by partnerships such as 3GPP and 3GPP2. The main ambition of IMT-2000 was to combat fragmentation and to unify different services (such as voice and multimedia) over a common platform. This way, operators should be able to provide seamless connectivity to users anytime and anywhere. Compared to earlier, 2G systems, IMT-2000 systems aimed at higher transmission rates for both mobile and fixed users. Therefore, IMT-2000 standards needed to combine flexibility, affordability, modular design, and be backwards compatible with existing systems. Five IMT-2000 radio interfaces were approved in 1999. A sixth one, namely IP-OFDMA TDD WMAN was added in 2007. As will be explained below, IP-OFDMA TDD WMAN is a subset of the IEEE 802.16 standard and the WiMAX specification.

Similar to IMT-2000, IMT-Advanced aims at providing the platform on which 4G wireless systems will be built, although it is not guaranteed that all 4G systems will be compliant with IMT-Advanced. Moreover, it is possible that the capabilities of 4G systems extend beyond IMT-Advanced. Naturally, the requirements of IMT-Advanced are more stringent compared to IMT-2000, and new services have been provisioned. 4G systems are expected to provide higher rates (up to 100 Mbps and 1 Gbps aggregate rates for mobile and fixed users, respectively), and a wide range of services and service classes in order to meet Quality of Service (QoS) requirements. Another notable feature of 4G systems is that an all-IP network architecture will be used, at least in the near future.

Currently, both 3GPP and the IEEE aim at developing standards to fulfill the goals of IMT-Advanced. The 3GPP partnership has started work on LTE-Advanced, and Release 10 will likely target IMT-Advanced [3]. The IEEE has formed the 802.16m Task Group for the development of the next amendment to the 802.16 standard [4]. However, 3GPP has as a *modus operandi* to provide a closed standard, whereas the IEEE 802.16m task group is developing a free standard. This survey focuses on the evolution of the IEEE WirelessMAN standard, although several of the principles will likely also be applied to the design of LTE-Advanced systems.

Several articles that focus on Mobile WiMAX networks complying to the IEEE 802.16e standard have appeared recently [6]–[10]. This survey shifts the focus to IEEE 802.16m-based Next Generation Mobile WiMAX, the recent design advances and the requirements as set by IMT-Advanced. An attempt is made to identify the new challenges of Next Generation wireless systems, and provide the reader with a unifying overview of the most important design improvements and changes in 802.16m compared to 802.16e. For this reason, an effort has been made to

- provide a survey of some of the state-of-the-art technologies used in Wireless Next Generation Networks,
- identify current bottlenecks and technical challenges and propose solutions or present already proposed approaches,
- decompose the Broadband Wireless Access technology to its parameters and present their correlations, and
- present the IEEE 802.16m Evaluation Methodology that aims at facilitating the collaboration of partners involved in the standardization of a complex system such as IEEE 802.16m.

The remainder of this survey is organized as follows: Section II contains a brief overview of WirelessMAN standards culminating to the development of IEEE 802.16m that is currently under way. Section III discusses some of the changes and enhancements that are expected in the OFDMA PHY layer of IEEE 802.16 systems in order for Next Generation 802.16m systems to meet IMT-Advanced requirements, whereas Section IV focuses on the Medium Access Control (MAC) layer. The Evaluation Methodology that is being developed for 802.16m systems is presented separately in Section V. Finally, Section VI contains concluding remarks.

II. Brief Overview of previous IEEE802.16 standards and requirements for 802.16m

The IEEE 802.16 task group has been developing a family of standards for Wireless Metropolitan Area Networks (WMANs). WiMAX systems are based on IEEE 802.16. However, strictly speaking, a WiMAX system is certified by the WiMAX forum, an industry-led organization [11]. Certified systems should conform to specified parts of the 802.16 standard and pass specific performance tests. That said, the terms IEEE 802.16 and WiMAX are often used interchangeably. The first 802.16 standard was approved in 2001. It employs Single-Carrier (SC) modulation in the 10-66 GHz band and targets Line-of-Sight (LOS) scenarios. 802.16a, the first amendment, was ratified in 2003. It added support for non-LOS environments in order to support last-mile fixed broadband access. For this reason, Orthogonal Frequency Division Multiplexing (OFDM) and Orthogonal Frequency Division Multiple Access (OFDMA) were introduced as options for the implementation of the physical (PHY) layer. 802.16d that followed 802.16c, a minor amendment, superseded all previous 802.16 standards (in the form of 802.16-2004) and is frequently referred to as Fixed WiMAX [12]. The success of OFDM-based WLANs and the gradual appearance of 802.16d-based products triggered work on a new amendment that would support constant mobility in cellular networks that resulted in 802.16e-2005, commonly referred to as Mobile WiMAX [7], [13].

Similar to previous amendments, 802.16e-2005 comprises different options for the implementation of the physical layer, depending on the radio frequency and the deployment environment. However, the main focus of 802.16e in the physical layer is OFDMA (the IP-OFDMA TDD WMAN part of the standard). More specifically, compared to 802.16d, more Fast Fourier Transform (FFT) sizes are supported and the FFT size is variable (Scalable OFDMA) in order for systems to be able to adapt to different conditions and requirements. 802.16e was the first 802.16 amendment to use MIMO spatial multiplexing (in addition to Alamouti transmit precoding that had appeared in 802.16d), therefore increasing the maximum spectral efficiency. The use of multiple antennas coupled with channel feedback also makes possible the use of beamforming as a means of focusing to specific users while, at the same time, protecting others from interference. This capability is expected to be exploited more in the future to implement interference avoidance schemes, not only in a specific cell, but also among users belonging to different cells. Moreover, Hybrid ARQ, a retransmission scheme that can trade off delay for improved link reliability was included in the standard. Such improvements augment the capacity of systems in terms of users and allow 802.16e systems to support mobile users moving at speeds up to 120 km/h. Finally, in each OFDMA frame of 802.16e systems, different modulation can be used in different groups of subcarriers (subchannels) allocated to different users. This way, the available system resources can be utilized more efficiently. As mentioned previously, the OFDMA implementation option of the 802.16e standard is now the sixth approved standard of IMT-2000 of the ITU.

In order to meet IMT-Advanced goals, several enhancements and new capabilities are being studied for inclusion in the Physical Layer of 802.16m. The aim is to use resources more efficiently, increase capacity, improve reliability in highly mobile environments, and accommodate users with different requirements. One change that is being studied is the decoupling of the subchannel permutations from the transmission modes, and also the use of a common basic unit for the permutations. This is explained in more detail in the following section. The goal is to reduce overhead, improve flexibility and facilitate channel estimation. In 802.16e systems, MIMO has mainly been addressed from the single-user perspective, and subcarriers are allocated to a given base station - mobile station pair. However, when many antennas are available at the base station, they can be used to transmit to more than one users simultaneously in each subcarrier. Use of beamforming and other Multi-user (MU-MIMO) techniques, such as Dirty Paper Coding [14], can increase the capacity of the system and/or improve reliability. Hence, algorithms are needed that allocate subcarriers and users and calculate the transmit vectors depending on the Quality of Service (QoS) requirements of each session. Design of such algorithms requires a good theoretical background, but also depends on practical considerations, the most important being the feedback of channel information to the base station. In high doppler speeds channel information may not be reliable. Moreover, even when channel variations are not an issue, the amount of overhead that is required for channel feedback should be kept to reasonable levels. The performance of WiMAX systems can be enhanced further by performing the allocation of frequency and spatial resources at the inter-cell rather than the intra-cell level. This way, adaptive and flexible frequency reuse can be implemented and the available spectrum can be redistributed depending on the requirements and the locations of the users. Using cooperation among neighboring cells, interference avoidance techniques can also be employed, especially for users near the cell boundaries. Finally, the reliability of transmission can be improved by devising and incorporating HARQ designs that take into account the performance of HARQ in practical MIMO scenarios. Such possible enhancements and additions to the Physical Layer of the 802.16 standard are discussed in Section III.

Regarding the MAC Layer, in order to achieve energy conservation and support mobility in rural environments, an efficient handover mechanism and power saving method has been standardized in 802.16e, which allows discontinuous reception of data from the base station. Moreover, in order to accommodate the new application demands, Quality of Service has been an essential part of all new WMAN standards. The IEEE 802.16-2004 standard mainly included QoS functionalities from the DOCSIS standard (multiple QoS classes) [5]. In IEEE 802.16e an additional QoS class has been introduced that would support real-time applications with variable bit rate. Many other mobility features, such as those related to paging and location update, were also introduced by the IEEE 802.16e workgroup [6].

In order to comply with 4G standards, IEEE 802.16m allows handover with service continuity for Radio Access Technologies (specific interest being paid on IEEE 802.11, 3GPP GSM/EDGE, UTRA, E-UTRA and 3GPP2 CDMA2000) and support of IEEE 802.21 Media Independent Handover (MIH) services. The IEEE 802.21 standard allows optimal wireless network selection, seamless roaming to maintain data connections and lower power operation for multi-radio devices. Some other operational requirements include multi-hop relaying (as introduced in IEEE 802.16j), synchronization among all base stations and mobile stations and self-organizing mechanisms. Moreover, IEEE 802.16m focuses on supporting legacy devices and cooperation among other 802.16 standards.

IEEE 802.16m will comprise three documents: A System Requirements Document (SRD) [15], a System Description Document (SDD) [16], and an Evaluation Methodology Document (EMD) [17]. As its name suggests, the System Requirements Document contains high-level requirements for 802.16m-compliant systems, including, among others, rates, throughput, coverage, mobility support, operating bandwidths and frequencies, QoS, latency, handover and security. The main goal is to meet the cellular layer demands of IMT-Advanced, and, at the same time, support legacy WirelessMAN-OFDMA equipment. A list of some of the main requirements for IEEE 802.16m is given in Table I. These requirements are compared with IEEE 802.16e, the legacy standard for Mobile WiMAX.

The System Description Document contains all the details on the implementation of 802.16m. It should be noted that the focus is on the definition of the signals emitted from the transmitter. The implementation details of the transmitter, as well as the design of the receiver are left to the manufacturers. The SDD specifies the Physical (PHY) layer and the Medium Access Control (MAC) layer and is currently in a preliminary stage.

A new element in 802.16m compared to previous amendments is the Evaluation Methodology Document. The EMD was introduced in order to provide a common baseline that will enable the evaluation and the comparison of different technology proposals. For this reason, both link-level and system-level simulation models and metrics are defined. Because of the projected complexity of 802.16m systems and the diverse environments and scenarios in which they will be deployed, it could be argued that the development of a methodology that will lead to reliable evaluation is itself a challenge. In this survey, a separate section is dedicated to the EMD, in order to provide more details on the work in this area.

III. Physical Layer enhancements

This section contains an overview of some Physical Layer enhancements that are currently being considered for inclusion in future systems. Because the development of the 802.16m standard is still in a relatively early stage, the focus is on presenting the concepts and the principles on which the proposed enhancements will be based, rather than on providing specific implementation details. Although the exact degree of sophistication of the new additions to the standard cannot be safely predicted, it is expected that the additions will make some use of the concepts described below.

A. Flexibility enhancements to support heterogeneous users

Because the goal of future wireless systems is to cater to needs of different users, efficient and flexible designs are needed. For some users (such as streaming low-rate applications) link reliability may be more important than high data rates, whereas others may be interested in achieving the maximum data rate even if a retransmission, and, therefore, additional delay, may be required. Moreover, the co-existence of different users should be achieved with relatively low control overhead. For these reasons, the frame format, the subcarrier mapping schemes and the pilot structure are being modified for 802.16m with respect to 802.16e.

Each 802.16e frame consists of a downlink (DL) and an uplink (UL) part separated in time by an OFDMA symbol and is of variable size. The (downlink or uplink) frame begins by control information that all users employ to synchronize and to determine if and when they should receive or transmit in the given frame. Control information is followed by data transmission by the base station (in the downlink subframe) or the mobile stations (in the uplink subframe). For each mobile station, transmission or reception happens in blocks that are constructed from basic units called *slots*. Each slot can be thought of as a two-dimensional block, one dimension being the time, the other dimension being the frequency. In general, a slot extends over one *subchannel* in the frequency direction and over 1 to 3 OFDMA symbols in the time direction, depending on the permutation scheme. The subchannels are

TABLE I
MOST IMPORTANT FEATURES AND SYSTEM REQUIREMENTS OF MOBILE WIMAX STANDARDS

Requirement	IEEE 802.16e	IEEE802.16m
Aggregate Data Rate	63 Mbps	100 Mbps for mobile stations, 1 Gbps for fixed
Operating Radio Frequency	2.3 GHz, 2.5-2.7 GHz, 3.5 GHz	$<$ 6 GHz.
Duplexing Schemes	TDD and FDD	TDD and FDD
MIMO support	up to 4 streams, no limit on antennas	4 or 8 streams, no limit on antennas
Coverage	10 km	3 km, 5-30 km and 30-100 km, depending on scenario
Handover Inter-frequency Interruption Time	35-50 ms	30 ms
Handover Intra-frequency Interruption Time	Not Specified	100 ms
Handover between 802.16 standards (for corresponding mobile station)	From 802.16e serving BS to 802.16e target BS	From legacy serving BS to legacy target BS From 802.16m serving BS to legacy target BS From legacy serving BS to 802.16m target BS From 802.16m serving BS to 802.16m target BS
Handover with other technologies	Not Specified	IEEE 802.11, 3GPP2, GSM/EDGE, (E-)UTRA (LTE TDD) Using IEEE 802.21 Media Independent Handover (MIH)
Mobility Speed	Vehicular: 120 km/h	Indoor: 10 km/h Basic Coverage Urban: 120 km/h High Speed: 350 km/h
Position accuracy	Not Specified	Location Determination Latency: 30 s Handset based: 50 m (67-percentile), 150 m (95-percentile) Network based: 100 m (67-percentile), 300 m (95-percentile)
IDLE to ACTIVE state transition	390 ms	50 ms
Quality of Service Classes	UGS, nrtPS, ertPS, rtPs, BE	UGS, nrtPS, ertPS, rtPs, BE

groups of OFDMA subcarriers. The number of subcarriers per subchannel and the distribution of the subcarriers that make up a subchannel in the OFDMA symbol are determined based on the permutation scheme. As explained in more detail below, the subcarriers of a given subchannel are not always consecutive in frequency. Downlink and uplink subframes can be divided into different zones where different permutation schemes are used.

In the Partial Usage of Subchannels (PUSC) zone that is mandatory, the priority is to improve diversity and to spread out the effect of inter-cell interference. Each slot extends over 2 OFDMA symbols, and a subchannel consists of 24 data subcarriers that are distributed over the entire signal bandwidth (OFDMA symbol). Thus, each subchannel has approximately the same channel quality in terms of the channel gain and the inter-cell interference. To reduce the effect of the inter-cell interference, when PUSC is used, the available subchannels are distributed among base stations so that adjacent base stations not use the same subchannels.

When the inter-cell interference is not significant, as in the case of mobile stations located closely to a base station, it may be advantageous to employ Full Usage of Subchannels (FUSC). The goal of the FUSC permutation scheme is similar to PUSC, *i.e,* to improve diversity and to spread out the effect of inter-cell interference. However, as the name suggests, in the FUSC zone all subchannels are used by a base station. For this reason, the design of the pilot pattern for the FUSC zone is slightly more efficient compared to PUSC. A subchannel in the FUSC permutation zone consists of 48 data subcarriers and the slot only comprises one OFDMA symbol.

For users with high rate requirements, the Adaptive Modulation and Coding (AMC) zone is employed instead of PUSC or FUSC. AMC makes it easier to exploit multiuser diversity by using adjacent subcarriers to form a subchannel. Subchannels made of adjacent subcarriers vary in quality across the frequency spectrum. Therefore, the system can employ opportunistic schemes that do not perform well when the transmission bandwidth is averaged (as is the case when subcarriers are spread out). A subchannel in the AMC zone consists of 16 data subcarriers. For the 2×3 AMC mode, each slot extends over 3 OFDMA symbols.

Therefore, in addition to requiring overhead to transition between zones inside a frame, in IEEE 802.16e the size of the basic data unit (the slot) depends on the permutation zone.

The frame structure is being redefined in 802.16m to make the allocation of transmission resources between the downlink and the uplink more flexible [16]. 802.16m consists of a 20-ms superframe, divided in equally sized 5-ms radio frames using either Time-Division Duplexing (TDD) or Frequency-Division Duplexing (FDD). Each 5-ms frame is further divided in 8 subframes when OFDMA is used. Each subframe is assigned to either downlink or uplink transmission, the allocation decision being based on QoS. If more capacity is required for the downlink, then the scheduler allocates more subframes to the downlink. Moreover, latency requirements can be satisfied via proper subframe allocation between the downlink and the uplink. Although some additional control overhead will be needed to signal the transitions between downlink and uplink and, vice versa, overall the system can benefit from the improved flexibility in rate allocation and latency guarantees.

Moreover, in order to reduce the overhead that is required for the placement of different zones in a frame, an effort is under way in 802.16m to define the same basic unit for all permutation schemes and to improve the flexibility of the system. This will be achieved by separating the subcarrier allocation mode from the transmission scheme. More specifically, a localized (contiguous) and a distributed (non-contiguous) resource unit permutation mode are defined for 802.16m [16]. As the name suggests, the localized (contiguous) permutation mode employs groups of contiguous subcarriers, whereas for the distributed (non-contiguous) permutation mode the subcarriers of each group are spread out. A given transmission scheme (such as SISO, beamforming or space-time coding using many transmit antennas etc.) can be used with both localized and distributed permutation modes. Having only two modes will also simplify channel estimation. Clearly, some control overhead is still needed to signal the subchannels that are assigned to each user and the subframe allocation. However, it is expected that this overhead will be reduced compared to 802.16e.

The possible modification of the permutation modes will also have an impact on the position of the pilot subcarriers that are used for channel estimation. In fact, this is a good opportunity to devise good subcarrier placements for more than two transmit antennas, especially because the goal of Next Generation WiMAX seems to be to support up to at least 4 downlink streams, possibly up to 8 streams. In 802.16e, the number of pilots per subchannel is constant. Therefore, when two transmit antennas are employed, the pilots are typically divided between the two transmit antennas. This means that the channel estimates may be less accurate when two transmit antennas are used at the base station instead of one. For more than two antennas, the estimation quality drops further, since the available pilots will need to be distributed to more antennas. Although the 802.16e standard contains provision for more than two antennas at the base station, Mobile WiMAX systems currently employ two antennas. The design of pilots for 802.16m is already under way. The current proposal uses the same pattern for localized and distributed resource units. It has been reported that the required overhead is smaller compared to 802.16e and that the estimation performance is better. In order to improve support for two downlink streams, it is proposed that the number of pilots be twice as large compared to when only one stream is used (12 within each resource unit *i.e.,* 6 pilots per stream) [16]. For 4 streams, 16 pilots per resource unit are being proposed (or, equivalently, 4 per stream). Although the quality of channel estimation is expected to be inferior to the 2-stream case, 4 streams will be used in better channel conditions where the quality of the estimation may be less crucial to the system performance.

The flexibility of 802.16m systems can also be improved by allowing use of more than one Radio Frequency (RF) bands when such bands are available. By including *multi-carrier* support, the available bandwidth, and, consequently, the system capacity, can increase. The RF bands do not need to be adjacent. In order to support multiple carriers without lowering the efficiency of the system, the distance between carriers will need to be an integer multiple of the subcarrier spacing to eliminate the need for additional guard bands. Moreover, the MAC should be modified in order to provide multi-carrier support.

Finally, it is important that 802.16m systems be backwards compatible with legacy standards so that other 802.16 devices be supported and coexist with new 802.16m-compliant equipment. 802.16m will contain provisions for backwards compatibility, coexistence and also handover with IMT-2000 and IMT-Advanced systems (such as LTE). This is discussed in more detail in Section III-D.

B. Extending use of MIMO transmission

Multiple-Input Multiple-Output (MIMO) communication is already a reality in wireless systems. It will be supported by the IEEE 802.11n amendment to the 802.11 WLAN standards that is expected to be ratified in the near future. Similarly, 802.16e includes support for MIMO downlink and uplink transmission. As MIMO technology matures and implementation issues are being resolved, it is expected that MIMO will be widely used for wireless communication.

Current Mobile WiMAX profiles include support for up to 2 transmit antennas even though the IEEE 802.16e standard does not restrict the number of antennas, and allows up to 4 spatial streams. The current aim for Next Generation WiMAX systems is to support at least up to 8 transmit antennas at the base station, 4 streams and Space-Time Coding.

Moreover, although some other MIMO features of 802.16e, such as closed-loop MIMO, have not appeared in Mobile WiMAX profiles yet, it is expected that they will be included in new 802.16m-based systems. More specifically, it has been already decided to support closed-loop MIMO using Channel Quality Information, Precoding Matrix Index and rank feedback in future systems [18].

In 802.11 systems, as well as in the 802.16e standard, MIMO transmission is used to increase the data rate of the communication between a given transmitter-receiver pair and/or improve the reliability of the link. It is expected that 802.16m and future 3GPP systems will extend MIMO support to Multi-user (MU-) MIMO. More specifically, use of multiple antennas can improve the achievable rates of users in a network with given frequency resources. In information-theoretic terms, the capacity region of the uplink and the downlink increases, in general, when MIMO transmission is employed [14]. In many cases, a large portion of this capacity increase can be achieved using relatively simple linear schemes (transmit beamforming at the downlink and linear equalizers at the uplink). Therefore, the achievable rates can be increased without the need for sophisticated channel coding. If larger complexity can be afforded, even higher gains can be attained using successive decoding at the uplink and Dirty Paper Coding schemes at the downlink.

An overview of the projected MIMO architecture for the downlink of 802.16m systems is given in the System Description Document (SDD), and is repeated in Fig. 1 for convenience.

One of the major theoretical advances of recent years was the characterization of the Gaussian MIMO uplink (a Multiple Access Channel) and the Gaussian MIMO downlink (a Broadcast Channel). More specifically, for the uplink, the capacity is achieved by the receiver decoding users successively, whereas, in the downlink, capacity-achieving codes are created by superimposing encoded streams destined for each mobile station. The encoded streams are created successively using Dirty Paper Coding (DPC) techniques [19]. In the uplink, the optimality of successive decoding at the receiver holds when the transmitted signals are Gaussian. Moreover, the codes should be sufficiently long in order for the probability of error of each user that is decoded successively to be arbitrarily close to zero. Similarly, in the downlink, capacity is achieved

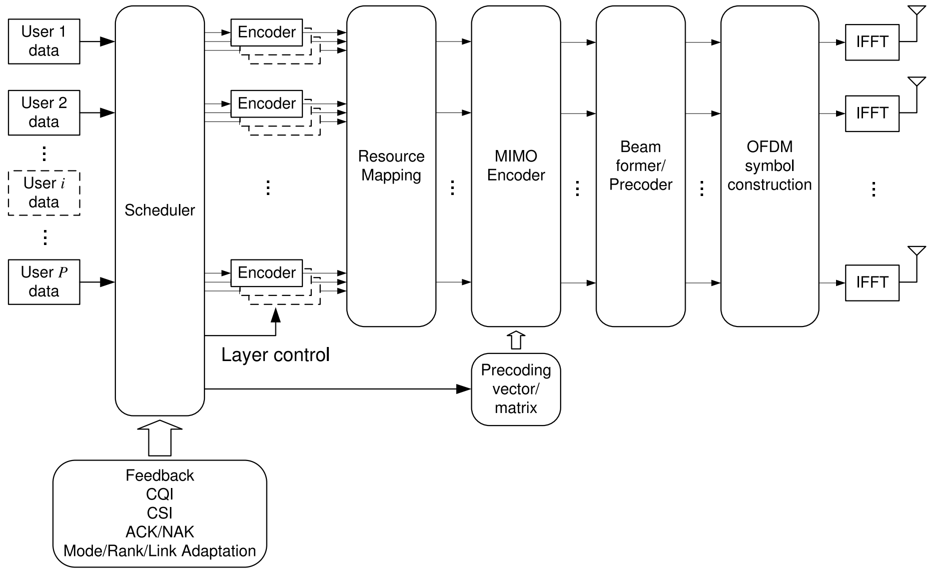


Fig. 1. MIMO architecture for the downlink of 802.16m systems.

by employing sufficiently long codes and encoded signals with Gaussian distribution. In practice, codes of reasonable encoding and decoding complexity are needed, both for the uplink and for the downlink. The encoding schemes may also affect the implementation of the receiver. Linear beamforming schemes [20], [21] that avoid the use of DPC have been shown to perform reasonably well in practical scenarios and could be a good candidate for 802.16m systems.

WiMAX and 3GPP networks employing MU-MIMO will need to calculate which users should transmit and receive during each frame, as well as the best achievable rate that corresponds to each user based on their QoS requirements, the number of users in each cell and their position. Although the information-theoretic capacity has been characterized, this is not an easy task, even for narrowband systems, and it is even more challenging when all subcarriers of the OFDMA system are considered. Therefore, efficient algorithms will be needed at the base station for *user selection* that will also determine the beamforming filters for the downlink, the receiver filters for the uplink and the required power allocation at the base station and each mobile station.

The performance of MU-MIMO also depends on what is known about the channel at the transmitter. When the doppler speeds are high, the channels may be changing faster than the speed with which the transmitter can track the channel. Even in low speeds, obtaining perfect channel knowledge may not be desirable from a practical point of view, especially in MIMO systems. For example, a base station with N_t antennas wishing to obtain channel information corresponding to K mobile stations each with N_r antennas will need to collect information for $K \cdot N_t \cdot N_r$ channel coefficients. This consumes part of the bandwidth that would otherwise be allocated to data communication. Moreover, as was mentioned previously, the pilot structure should be such that the channel measurement be obtained with sufficient reliability. Therefore, in addition to well-designed pilot patterns, it is very likely that base stations of future systems will have to rely on imperfect and partial channel information. 802.16e already includes five different mechanisms for channel feedback ranging from the mobile station only reporting to the base station which antennas should be used (Antenna Selection) to the mobile station sending to the base station exact information about the MIMO channel (Channel Sounding). The channel feedback mechanisms will need to be extended to the case of MU-MIMO and algorithms will be needed in order for the mobile stations to determine what information to send back to the base station. For this reason, the performance of MU-MIMO with partial channel information at the transmitter, and the mechanisms to obtain channel information at the receiver and feed it back to the transmitter are currently topics of considerable theoretical and practical interest [18], [22]–[24].

C. Resource allocation and multi-cell MIMO

In cellular networks careful frequency planning is required in order to achieve communication with small outage probability and, at the same time, minimize interference among users of neighboring cells. Users near the cell edges are particularly vulnerable, because they receive signals of comparable strength from more than one base stations. For this reason, different parts of the frequency spectrum are typically assigned to neighboring cells.[1] The assignment in current systems is static and can only be changed by manual re-configuration of the system. Changes to the frequency allocation can only be performed periodically and careful cell planning is required in order not to affect other parts of the system. Frequencies are reused by cells that are sufficiently far away so that the interference caused by transmissions on the same frequencies is small enough to guarantee satisfactory Signal-to-Interference and Noise Ratios (SINRs). Although static frequency reuse schemes greatly simplify the design of cellular systems, they incur loss in efficiency because parts of the spectrum in some cells may remain unused while, at the same

[1]CDMA systems are an exception to this scenario although the system still needs to guarantee that the chip sequence assigned to a given user is not employed by any other user that communicates with the same base stations.

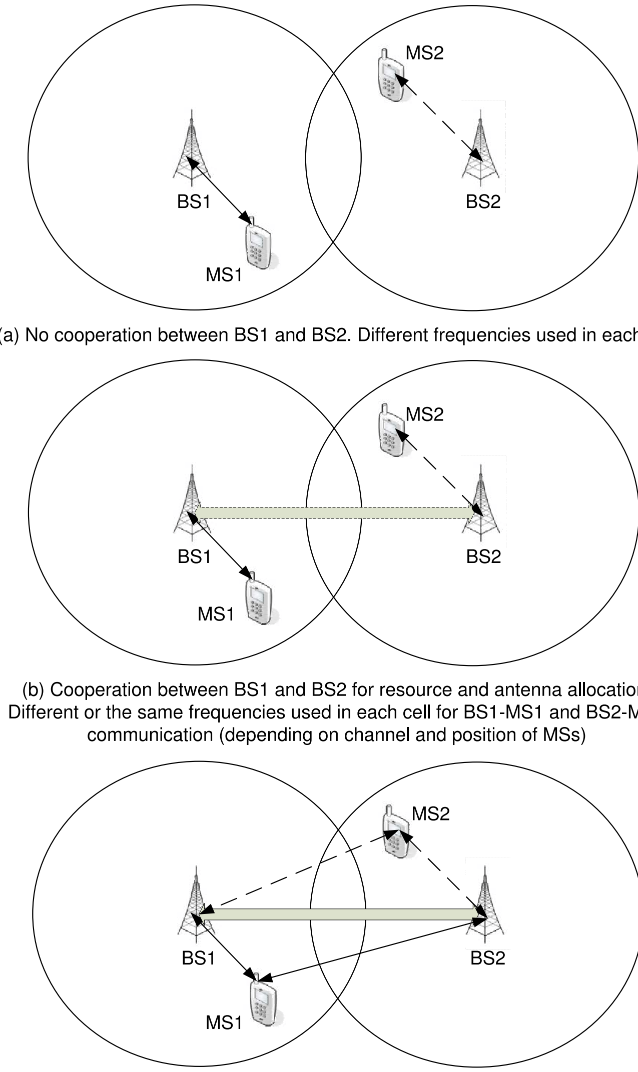
time, other cells may be restricting the rates of their mobile stations or even denying admission to new users. Moreover, the handover process is more complicated for mobile stations since communication in more than one frequencies is required.

In order for future wireless systems to attain the high rate and reliability requirements of IMT-Advanced, efficient management of the system resources is needed. Ideally, systems should be able to re-allocate dynamically the available spectrum among different cells. As discussed previously, by using beamforming, MIMO systems can transmit to more than one user in a given frequency, in general. So, it is possible, and it may actually also be desirable, to allow more than one users of the same or of different cells to transmit or receive on the same frequency. In general, MU-MIMO transmission and resource allocation are closely linked to each other, as can also be seen in Fig. 1. For optimal performance, resource allocation (frequency, power, antennas and users) should be made by a central controller that receives information from all base stations of all cells. Clearly, this is very complicated, since the amount of information that needs to be transferred to the central controller and the complexity involved in performing resource allocation can be very large.

In order to improve the efficiency of resource allocation with reasonable complexity, distributed approaches can be used. For example, resource allocation can be seen as a non-cooperative game among different base stations. Each base station attempts to allocate resources to satisfy the QoS requirements of the users in the cell it serves, but, at the same time, gets penalized for the amount of resources that are being used [25]. If information on the interference caused to users of other cells can be conveyed to the base station, the cost function could include the effect of the choices of the base station on the transmissions occurring in neighboring cells. Simpler approaches can also be used. As an example, a dynamic scheme is being developed for 802.16m that divides the frequency band into 4 sub-bands [16]. Three sub-bands use a frequency reuse factor of 1/3, whereas the frequency reuse factor in the fourth sub-band is equal to 1. The size of the sub-bands can be changed dynamically, thus changing the overall frequency reuse factor of the system. This dynamic Fractional Frequency Reuse scheme could be a first step towards fully dynamic frequency reuse schemes.

Finally, as the infrastructure improves, cellular systems could also take advantage of communication between the base stations. As mentioned above, exchange of information regarding the interference caused by neighboring cells can improve the efficiency of resource allocation. If user data can also be communicated, in addition to channel measurements, the system performance at the cell edge could be potentially improved by processing jointly the data of users that are received by more than one base stations. Other open- and closed-loop multi-cell MIMO transmission schemes are also being considered for 802.16m, but the development is still in a very early stage for safe predictions to be made on what will be included in the standard.

An example of transmission depending on the level of inter-cell cooperation is given in Fig. 2 where, in each cell i, a base station BSi communicates with a mobile station MSi. In Fig. 2(a), a traditional system with static frequency reuse and no inter-cell cooperation is considered. Even if the base stations have many antennas, different parts of the available spectrum are used by each cell for transmission because BS1 does not have access to information regarding the frequencies used by BS2 to transmit to MS2, and vice-versa. In Fig. 2(b), BS1 and BS2 exchange information on the beamforming vectors and the frequencies that are used for transmission. This way, they can avoid interfering with each other. This can be achieved by choosing different frequencies from a shared pool of frequencies, or, if this is possible, use the same frequency, but pick beamforming vectors that minimize interference between the two links. When the interference between the mobile stations is weak, or when the rates requested by the corresponding users are sufficiently small, a given frequency may be shared even if no beamforming is used (for example at the uplink, when each mobile station has a single antenna). Finally, in Fig. 2(c), the base stations can also share the received symbols (or, equivalently, they can send them to a central processor). This way, an equivalent, virtual



(a) No cooperation between BS1 and BS2. Different frequencies used in each cell

(b) Cooperation between BS1 and BS2 for resource and antenna allocation. Different or the same frequencies used in each cell for BS1-MS1 and BS2-MS2 communication (depending on channel and position of MSs)

(c) Full cooperation and data exchange between BS1 and BS2 (Virtual MIMO). Cells can be treated as one cell.

Fig. 2. Communication and resource allocation depending on level of inter-cell cooperation.

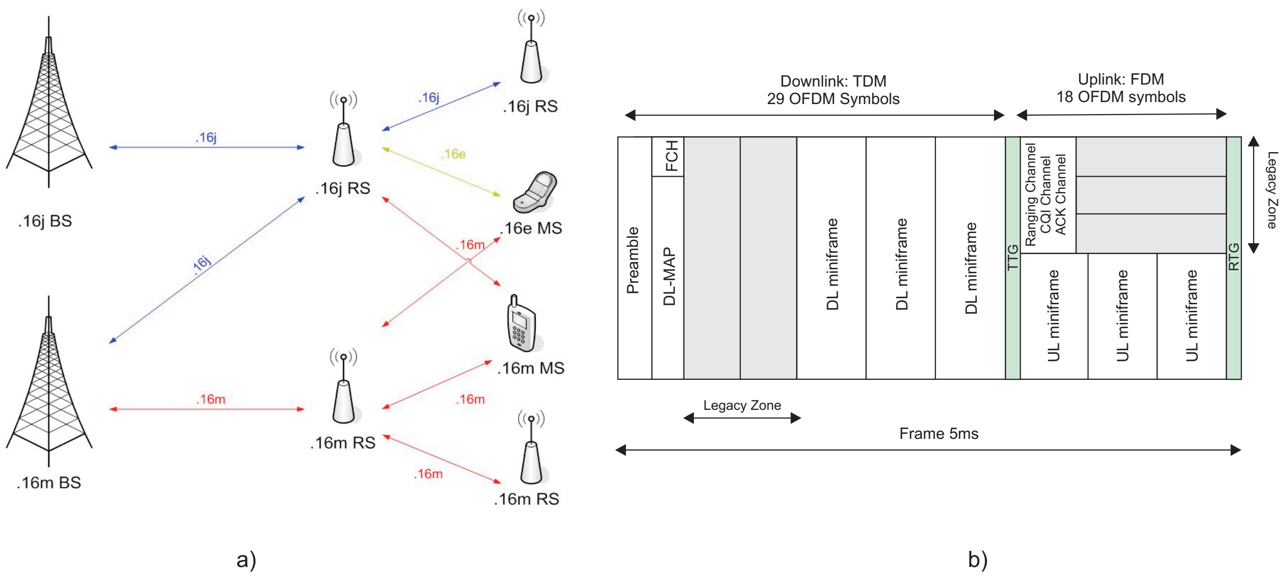


Fig. 3. Supported 802.16 connections and IEEE 802.16m frame structure with TDM Downlink and FDM Uplink.

Multiuser-MIMO system is created between the combination of the antennas of BS1 and BS2 and the mobile stations of both cells. Clearly, this last architecture is the most efficient in terms of resource usage, but also the most complicated, especially as the number of cells that are coupled together increases.

It is important to note that, most likely, the future WiMAX standards will not describe in detail how resource allocation should be performed, and it will be up to the manufacturers to devise efficient and practical schemes. For example, 802.16e does not specify how subcarriers and transmit matrices are allocated to users in a cell. The base station performs resource allocation based on information that is received from mobile stations and then informs mobile stations when each should transmit and receive. What is important is to ensure that future amendments contain the mechanisms that will enable the transmission and exchange of the necessary information for the implementation of resource allocation mechanisms. Changes in the MAC layer may also be needed if mobile stations need to communicate simultaneously with more than one base stations.

D. Interoperability and coexistence.

In order for the standard to be able to support either legacy base and mobile stations or other technologies (e.g. LTE), the concept of the time zone, an integer number (greater than 0) of consecutive subframes, is introduced.

Interoperability among IEEE 802.16 standards: The 802.16m Network Reference Model permits interoperability of IEEE 802.16m Layer 1 and Layer 2 with legacy 802.16 standards. The motivation for ensuring interoperability comes from the fact that WiMAX networks have already been deployed, and it is more realistic to require interoperability instead of an update of the entire network. Another advantage is that each 802.16 standard provides specific functionalities in a WiMAX network. The goal in 802.16m is to enable coexistence of all these functionalities in a network without the need to create a new standard that contains all of them. The supported connections and frame structure are summarized in Fig. 3. The legacy standard can transmit during the legacy zones (also called LZones), whereas 802.16m-capable stations can transmit during the new zones. The Uplink (UL) portion shall start with the legacy UL zone, because legacy base stations, mobile stations or relays expect IEEE 802.16e UL control information to be sent in this region. When no stations using a legacy 802.16 standard are present, the corresponding zone is removed. The zones are multiplexed using TDM in the downlink, whereas both TDM and FDM can used in the uplink. In each connection, the standard that is in charge is showcased. The Access Service Network can be connected with other network infrastructures (e.g. 802.11, 3GPP etc.) or to the Connectivity Service Network in order to provide Internet to the clients.

Adjacent Channel Coexistence with LTE TDD: 802.16m-compliant devices may coexist with LTE devices (E-ULTRA TDD) in adjacent channels. This is accomplished by inserting either idle symbols within the 802.16m frame or idle subframes. Thus, the 802.16m system shall be capable of applying an operator configurable delay or offset such that the 802.16m UL shall start at the same subframe as the LTE frame. Hence, both standards can coexist and be synchronized. Similarly, 802.16m-compliant devices can coexist with other technologies such as ULTRA LCR (TD-SCDMA) etc.

IV. MEDIUM ACCESS CONTROL LAYER

In contrast to contention-based WiFi networks where each station needs to compete for access to the medium, the IEEE 802.16 standard is a connection-oriented wireless network. A scheduling algorithm allocates resources (either in specific grants or contention periods) to each service flow (or connection) taking into account the Quality of Service needs. Further extensions to the initial design of IEEE 802.16-2004 included mobility (*i.e.,* functionalities associated with handover and energy-efficient algorithms), relaying [26], efficient

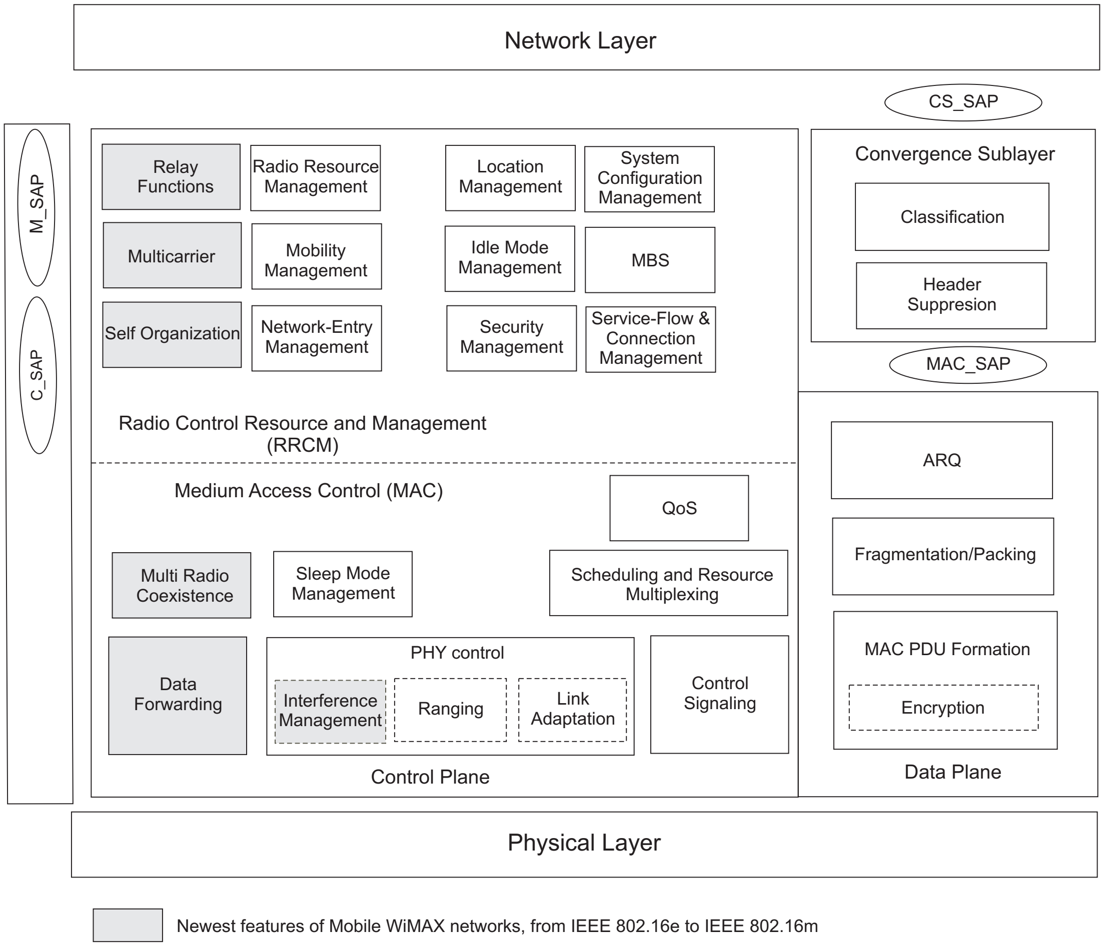


Fig. 4. MAC layer architecture of IEEE 802.16m.

localization and several other MAC layer features. In this section, the MAC layer enhancements of IEEE 802.16m that aim at meeting the IMT-Advanced requirements are presented. A basic understanding of connection-oriented and WiMAX network architectures is assumed in the following. For more information on this subject, see, for example, [28].

The MAC of 802.16m is divided into three sublayers:

- Convergence sublayer (CS)
- Radio Resource Control and Management (RRCM) sublayer
- Medium Access Control (MAC) sublayer

The general architecture of the MAC layer of 802.16m is shown in Fig. 4. The purpose of the sublayers is to encapsulate wireline technologies such as ETHERNET, ATM and IP on the air interface, and to introduce the state-of-the-art connection-oriented features. In the following these three sublayers are discussed in more detail.

A. Convergence sublayer

The Convergence sublayer (CS) resides on top of the MAC CPS (Common Part Sublayer). The Packet CS is responsible for accepting data units from higher layers, classifying, processing based on classification, and delivering CS Protocol Data Units (PDU) to the appropriate MAC SAP (Service Access Point). In reality, it is the sublayer that unites the IEEE 802.16 MAC with the network layer. The MAC CPS creates its Protocol Control Information (MAC header) and is responsible for the delivery of MAC PDUs to its peer MAC-CPS according to the Quality of Service (QoS) requirements of the particular Service Flow (SF). Those functionalities, as provided by the current form of the CS of IEEE 802.16m, are similar to the ones of legacy 802.16 systems.

B. RRCM sublayer

The RRCM sublayer includes functionalities that determine the performance of the service flow. The functionalities are indicated in the boxes of Fig. 4. New functionalities are shown in grey boxes. The functionalities of RRCM are the following.

- *Relay Functions*. They are described in detail in [26] and include methodologies referred to as “Mobile Multihop Relay” (MMR). The basic idea behind MMR is to allow WiMAX base stations that do not have a backhaul connection to communicate with base stations that do (in order to increase network coverage). A technical survey for the nonexpert can be found in [27].
- *Radio Resource Management*, that adjusts radio network parameters related to the traffic load, and includes some other QoS-related functionalities, such as load, admission and interference control.
- *Mobility Management*, that is related to handover.

- *Network Management*, that is in charge of initialization procedures.
- *Location and Idle Mode Management*, responsible for supporting location-based services and for controlling Idle Mode operation, respectively. The system requirements for location-based Services (LBS) are identified in Table I.
- Since security is one of the most important issues in wireless networks, *Security Management* provides secure communication between the mobile station and the base station by handling encryption and authentication.
- Broadcast control messages, such as the downlink/uplink channel descriptor (DCD/UCD), are generated from the *System Configuration Management* block.
- MBS (Multicast and Broadcasting Service) controls management messages and coordinates with the base station to schedule the transmission related to MBS data.
- *Connection Management* allocates connection identifiers (CID) during initialization, handover and service flow creation procedures.
- *Self Organization* includes the procedures to request that mobile stations report measurements for self-configuration and optimization.
- A *Multi-carrier* block, that enables a common MAC entity to control a PHY spanning over multiple frequencies.

The way that a specific service flow will be handled by the MAC layer is based on the performance of the scheduler. Contrary to the 802.16e standard, the IEEE 802.16m workgroup does not provide amendments on the service flow management part. Five QoS categories are defined to handle different types of applications.

- *UGS* (Unsolicited Grant Service) is designed to support real-time uplink service flows that generate fixed-size packets on a periodic basis. For this reason, the base station generates periodically fixed-size data grants to minimize the overhead and the latency. It includes VoIP applications without silence suppression and also handles T1/E1.
- *rtPS* (Real-time Polling Service) supports variable-size data packets, such as MPEG video. The base station offers real-time periodic unicast request opportunities and allows the mobile station to specify the size of the desired grant.
- *Extended rtPS* (Extended Non-Real-time Polling Service) provides unicast grants with variable size in an unsolicited manner. It is supposed to be the service flow that combines the advantages of both UGS and rtPS.
- *nrtPS* (Non Real-time Polling Service) is designed to support non real-time uplink service flows, where the base station issues unicast request opportunities on a regular basis. During the request interval, the mobile stations contend and the base station allocates resources accordingly. It may also support unicast request opportunities. This service flow may provide access to applications that do not require strict QoS guarantees, even during network congestion.
- *BE* (Best Effort) is based on a Request/Transmission policy where the mobile stations are contending for request opportunities. It is the only scheduling service where no mandatory QoS parameters are set because it handles only best-effort traffic.

C. MAC sublayer

The functionalities of the MAC sublayer are related to PHY control (cross-layer functionalities, such as HARQ ACK/NACK etc). The Control Signaling block is responsible for allocating resources by exchanging messages such as DL-MAP and UL-MAP. The QoS block allocates the input traffic to different traffic classes based on the scheduling and resource block, according to the SLA guarantees. The name of other blocks, such as fragmentation/packing, multi-radio coexistence and MAC PDU formation, clearly describes their function.

The MAC sublayer also deploys state-of-the-art power saving and handover mechanisms in order to enable mobility and make connections available to speeds up to 350 km/h. Since newer mobile devices tend to incorporate an increasing number of functionalities, in WiMAX networks the power saving implementation incorporates service differentiation on power classes. A natural consequence of any sleeping mechanism is the increase of the delay. Thus, delay-prone and non delay-prone applications are allocated to different classes, such that the energy savings be optimized, while satisfying the appropriate QoS (e.g those that support web page downloading or emails).

MAC addresses play the role of identification of individual stations. IEEE 802.16m introduces two different types of addresses in the MAC sublayer. 1) The IEEE 802 MAC address that has the generic 48-bit format and 2) two MAC logical addresses that are assigned to the mobile station by management messages from the base station. These addresses are used for resource allocation and management of the mobile station and are called “Station Identifiers” (assigned during network entry) and “Flow Identifiers” (assigned for QoS purposes).

1) MS state: The mobile station can be in four different states based on the MAC functionalities, namely Initialization, Access, Connected and Idle State. The states are shown in Fig. 5. The Initialization State is where the mobile station performs the scanning and synchronization based on the base station preamble. As soon as it acquires the system configuration information, it is ready to perform the ranging process and transition to the Access State. If the mobile station cannot perform BCH information decoding (that usually involves the evaluation of the error locating polynomial) and cell selection, it goes back to scanning and DL synchronization.

In the Access State, MOB_RNG-REQ and MOB_RSP-REG MAC messages[2] are exchanged in order to initiate the ranging process and Uplink (UL) synchronization, respectively. For security purposes, Authentication and Authorization is performed before mobile station registration to the target

[2]Since 802.16 networks are connection-oriented, the base station and mobile station communicate with management messages. Starting from 802.16e these messages are of the form MOB_XXX-YYY, where MOB stands for Mobility and the rest for each functionality (e.g. XXX can be REG: Registration, RNG: Ranging, SCN: Scanning, SLP: Sleep mode functionality message and YYY REQ: Request Message, RSP: Response message, ADV: Advertisement).

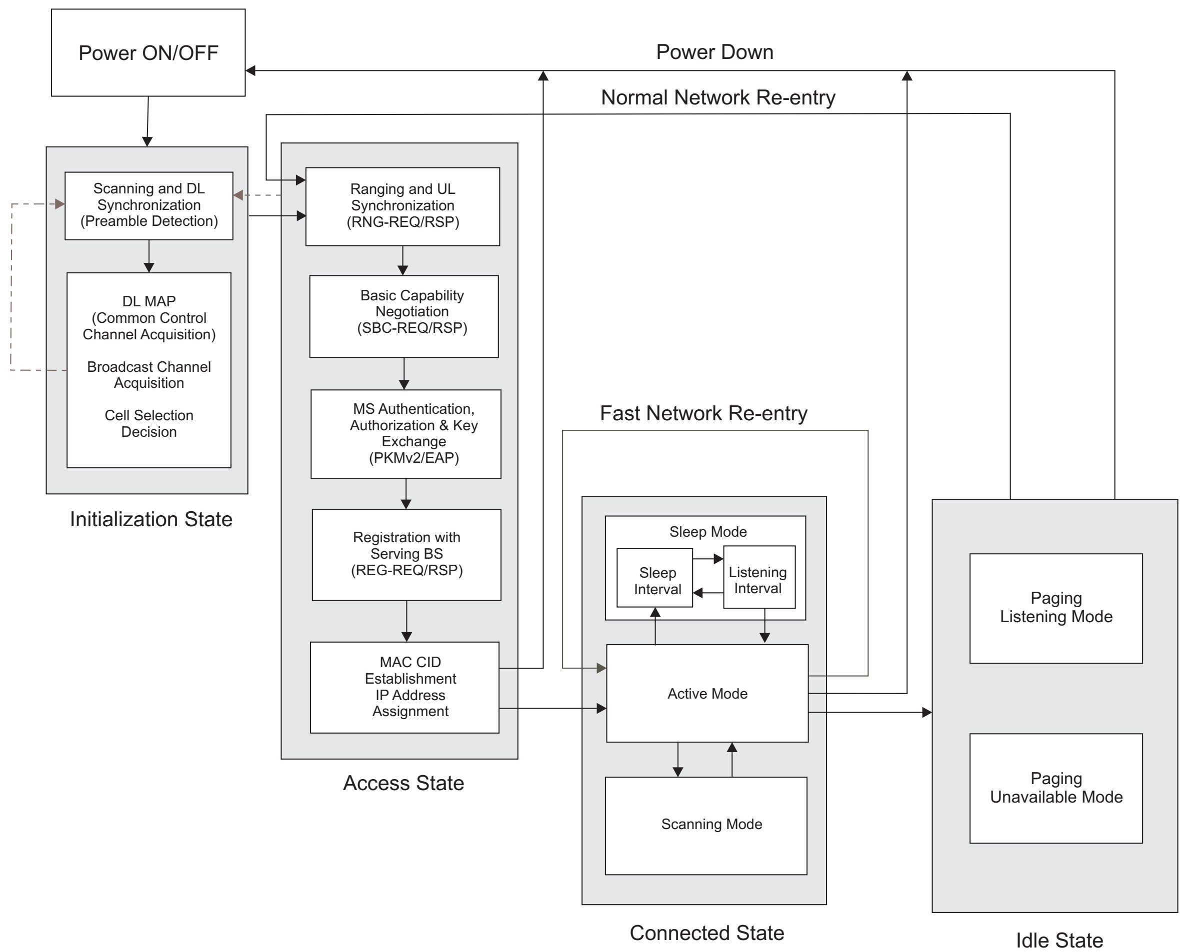


Fig. 5. IEEE 802.16m mobile station state transition diagram.

base station. During the registration, the mobile station is associated with a MAC Connection Identifier (CID) and an IP address. The dashed arrows in Fig. 5 denote the case where the mobile station moves back to the Initialization step because of failure.

The Connected State comprises three different modes. Traffic is transmitted and handover is performed when on Active Mode. For moving from Active to Sleep Mode, the standard defines specific messages that are exchanged between the mobile station and the base station. The mobile station also needs to scan for other available base stations, and for that it must move to the Scanning Mode (MOB_SCN-REQ and MOB_SCN-RSP messages are exchanged for this transaction), thus the Scanning Mode State.

As mentioned before, in order to conserve energy and to prolong the battery lifetime of the mobile station, a Sleep Mode is defined. The Sleep Mode is further divided into a sleep interval and a listen interval. This mechanism takes advantage of the idle transmission intervals of the BS when there are no packets destined to the MS in the output buffer of the BS. Therefore, the sleep interval is the amount of time during which the radio interfaces are periodically shut down. After that, a listen interval follows, in which the mobile station synchronizes with the serving base station and receives a small amount of data or a traffic indication message (called TIM; it indicates the existence of a packet in the buffer of the base station). However, because packets are buffered when the radio is shut down multiple times, an optional Idle State was introduced in IEEE 802.16e. When in Idle State, the mobile station can monitor the link channel, transition to the Active Mode and perform handover. This can be done in the Paging Listening Mode using MOB_PAG-ADV messages. However, monitoring is not performed in Paging Unavailable Mode. The advantage of this implementation is that battery life can be prolonged, and the delay is maintained below the agreed levels. For higher QoS classes, such as constant bit rate VoIP, the sleep interval is usually in-between the packets, whereas for best-effort traffic classes an exponential algorithm is implemented for the sleep intervals [33].

2) Hybrid ARQ: Hybrid Automatic Repeat reQuest (HARQ) is an extension of ARQ. When the receiver detects a corrupted packet, a retransmission is requested from the receiver. The standard retransmission strategies of ARQ (such as stop-and-wait, go-back-N and selective-repeat) can also be used for HARQ. The difference from ARQ is that HARQ uses all received copies of a given packet for decoding. HARQ transmission can improve the throughput of a communication link, and for this reason it is supported in 802.16e. Two variants of HARQ have been developed. When Chase-Combining HARQ (CC- or Type I HARQ) is employed, the same packet is sent after each request. Incremental Redundancy HARQ (IR- or Type II/III HARQ) is more general: retransmitted packets may contain different bits of an encoded stream, or a different encoder may be used. In general, the use of different encoders and/or different bits provides a coding gain, and, for this reason, IR-HARQ performs better than CC-HARQ. However, there are some cases of fading channels where CC-HARQ may outperform IR-HARQ [29]. Moreover, the architecture of transceivers that employ CC-HARQ is simpler. The profiles of Mobile WiMAX only support CC-HARQ, although IR-HARQ was included in the 802.16e standard. IR-HARQ is now being considered for Next Generation systems and code designs are being developed.

In the following, some issues related to the application of HARQ in 802.16m systems are discussed.

- *MIMO HARQ*: When HARQ is used in MIMO systems, the implementation of the receiver may have substantial impact on the performance of HARQ. Transmission is affected by Gaussian noise and the fading channel, but also from the interference among streams received at different antennas. When Maximum-Likelihood (ML) decoding is too complex to implement in terms of operations and/or required storage, simpler architectures need to be employed. When CC-HARQ is used, the decoder can be simplified without loss in performance by combining the received modulated symbols at each antenna. However, in general, the bit-to-symbol mapping may not be constant in IR-HARQ. Suboptimal approaches such as decoding after each retransmission and combining of the decoded soft bits may result in significant performance loss. In some cases, for schemes used in 802.16e, use of a practical receiver may result in IR-HARQ transmission performing worse than CC-HARQ [30]. Therefore, it is important to develop codes that will exhibit good coding gains and will be robust to suboptimal operations at the receiver. One approach to obtaining a good complexity-performance tradeoff is to include mechanisms that enable IR-HARQ that preserves the alignment between bits and the symbol vectors that are transmitted at the base station [31]. This allows reduction of complexity and storage without a penalty in performance. However, care has to be taken in the design of the IR-HARQ scheme to combine constant bit-to-symbol vector alignment with sufficient coding gains so that IR-HARQ still have an advantage over CC-HARQ. It should be noted that, although constant bit-to-symbol vector alignment can lead to considerable simplification of the receiver, optimal receivers for IR-HARQ will be more complex than their counterparts for CC-HARQ. However, for a future WiMAX system where efficiency is very important, the additional gain provided by IR-HARQ may be required, and, therefore, simplification of a receiver capable of supporting IR-HARQ will be very beneficial. Another approach is to design IR-HARQ coding schemes that take the MIMO nature of the system into account [32]. Again, the complexity of the receiver will depend on the scheme, and, inversely, will impact the performance of HARQ.
- *Synchronous and Asynchronous HARQ*: One of the latest developments on the HARQ technology is to classify the HARQ operations in terms of retransmission timing. However, it is still being debated whether synchronous or asynchronous HARQ or a combination of both will be used. In synchronous HARQ the retransmission timing accurately defines the exact period in which HARQ retransmission processes will be exchanged between the mobile station and the base station. The advantage of this mechanism is that less signaling overhead is required during retransmissions. In asynchronous HARQ, the retransmissions can take place at any time. Because of the usage of explicit signaling, the overhead is increased, but this allows more flexibility on the decisions of the scheduler.
- *Adaptive/Non-adaptive HARQ*: Adaptive HARQ provides more flexibility in scheduling, because transmission attributes (such as modulation order, code rate etc) may be changed during retransmissions. However, adaptive HARQ also requires more signaling overhead in order to inform the receiver of any changes on the transmission attributes every time a retransmission occurs. Non-adaptive HARQ does not allow any changes on the transmission attributes, and, therefore, requires less signaling overhead. However, the gain is smaller compared to adaptive HARQ.

3) Handover: Next Generation Mobile WiMAX networks will include advanced handover functionalities (low transition delay for fast moving mobile stations in combination with sleep mode operations). Handover occurs when the received signal quality at the mobile station is not adequate to provide the Quality of Service required by an application, in which case it is handed over to another base station. The mechanism is divided into two phases, Network Topology Acquisition, during which the mobile station obtains information about the neighboring base stations through several advertisement messages and scanning processes, and the Handover Process. An excellent description of the functionalities required for handover in Mobile WiMAX systems is given in [34].

Apart from the fast handover requirements [15], the IEEE 802.16m standard includes handover support between femto base stations and base stations, and legacy Mobile WiMAX systems and Inter-RAT handover procedures. Femto cells are designed for residential or business environments and may enhance the coverage of indoor locations. Although the specific functionalities of femto base station handover support are not yet defined, it is required that the mobile station be allowed to cache the information when making a handover to the specific femto cell. The handover support for legacy systems includes a new method for Network Topology Acquisition. The YBS[3] advertises the system information to its neighbor YBSs and to the LZones of its neighbor ABSs.[4] The ABS does the same for its neighbors' YBSs in both its LZone and its 802.16m zone. Finally, the ABS advertises the system information to its neighbor ABSs in the 802.16m zone only. In the Network Topology Advertisements, the ABS may indicate its support for 802.16m mobile stations. LZones and 802.16m zones are used in a similar way during the handover from 802.16m to 802.16e (and from 802.16e to 802.16m).

One of the goals of IMT-Advanced, as described in the introduction, is seamless handover between Radio Access

[3]Yardstick BS: Base station compliant with the WirelesMAN-OFDMA reference system. Similarly, in IEEE 802.16m, new abbreviations have been introduced for YMS (Yardstick MS): Mobile station compliant with the WirelessMAN-OFDMA reference system, AMS (Advanced MS): Mobile station capable of acting as a YMS and additionally implementing the protocol defined in IEEE 802.16m, and ABS (Advanced BS), the base station equivalent of AMS.

[4]Details on the LZones are not repeated here because they have been discussed in Section III-D.

Technologies. Therefore, Network Topology Acquisition for Inter-RAT handover includes the advertisement of information about other RATs to assist the mobile station with network discovery and selection. IEEE 802.16m systems provide a mechanism for the mobile station to obtain information from a base station about other access networks in its vicinity, either by making a query or by listening to information broadcasts. Moreover, other mechanisms are introduced for conducting inter-RAT measurements and reporting. Based on whether the mobile station supports a dual transmitter/dual receiver, it may connect to both an ABS and a base station operating on another RAT simultaneously during handover.

V. Evaluation methodology

The Evaluation Methodology of IEEE 802.16m defines a unified method of simulation models and associated parameters that can be used when introducing new proposals for IEEE 802.16m or when presenting new research results. As shown in Fig. 6, the simulation components can introduce results both from the link-level perspective, when only one base station and one mobile station exist in the network topology scenario, and from the system-level perspective when multiple base stations and mobile stations communicate. Some aspects of the document that are related to the physical layer are discussed first, followed by a description of the handover evaluation. More details on other components of the evaluation methodology (such as antenna and traffic patterns) can be found in the Evaluation Methodology Document (EMD) of the IEEE 802.16m workgroup [17].

A. System-level evaluation methodology and link-to-system mapping using Effective SINR

Because WiMAX systems can employ various constellation sizes, subchannel permutation schemes, coding methods and MIMO modes, and may be based on different transmitter and receiver hardware and implementations, a set of representative and widely acceptable scenarios needs to be defined. This is expected to enable the developers present their systems using a common language. Moreover, the scientific and business communities will be able to assess independently the submitted proposals. Some of the most important sets of parameters defined in the Evaluation Methodology Document (EMD) for system simulation are the following.

1) *System-level simulation assumptions for the downlink and the uplink*, including the constellation size, the duplexing scheme, the subchannelization method, the pilot structure, the number of transmit and receive antennas, the interference cancelation method at the receiver, the channel coding scheme, the scheduling method and the frequency reuse pattern.
2) *Test scenarios*, including the carrier frequency, the site-to-site distance, the power transmitted by the base station and the mobile stations, the path loss and shadowing model and the mobile speeds.
3) *Base station and mobile station equipment model parameters*, including the number of transmit and receive antennas and the number of sectors, the spacing among antennas, the transmit power and the noise figure.
4) *OFDMA parameters*, including the FFT size, the sampling frequency, the subcarrier spacing, the total bandwidth, the length of the Cyclic Prefix and the frame length.

Moreover, the simulations should be performed over agreed channel models. The channel models need to be representative of the different radio environments and propagation scenarios in which 802.16m systems will be deployed [17]. In addition to the so-called “Baseline Test Scenario” that is mandatory and part of the system-level simulations specification as described above, seven additional optional test scenarios are defined for 802.16m: Urban Macrocell, Suburban Macrocell, Urban Microcell, Indoor Small Office, Outdoor to Indoor, Indoor Hotspot and Open Rural Macrocell, each corresponding to different cell sizes, propagation conditions, and number and density of mobile stations. Each model is associated with specific path loss, shadowing, and Cluster-Delay-Line (CDL) models.

An additional goal of the EMD is to create a PHY abstraction that will allow accurate prediction of the performance of the link layer in a computationally simple way. The abstraction needs to accurately reflect the fact that the PHY layer becomes increasingly dynamic and adapts to the changing channel conditions. Therefore, evaluating the performance based on the expected topology and average Signal-to-Noise ratios may not be sufficient. One methodology that is used in the EMD is Effective SINR mapping (ESM). ESM consists of calculating an *effective SINR* $SINR_{eff}$ based on the SINR of each subchannel where a given coded block is transmitted. Then, the calculation of the expected Block Error Rate (BLER) is based on a single value $SINR_{eff}$. There are several approaches for the exact mapping of the subchannel SINRs to $SINR_{eff}$, some of them taking into account the case of MIMO transmission. ESM can also be applied when HARQ (CC- or IR-) or repetition coding is used. The PHY link-to-system mapping procedure as depicted in the EMD and is given in Fig. 7 for convenience.

The BLER that is calculated based on $SINR_{eff}$ is then used to compute the Packet Error Rate (PER), based on the number of blocks that comprise a packet.

B. System Simulation of Handover

High Speed Mobility scenarios require low delay handover so as not to degrade the end-user experience when moving from one cell to another. For the comparison of different handover schemes, a specific evaluation methodology was developed in IEEE 802.16m, where latency and data loss rate are taken into account as key metrics, since both have a direct impact on the QoS as perceived by the user. Thus, two handover evaluation methods are introduced. The first method is based on a single moving mobile station with all other users at fixed locations, whereas in the second method all users move randomly.

1) Single Moving MS: In order to reduce complexity and provide simple handover results, a single moving mobile station scenario can be used. Mobility speeds can be chosen among three different scenarios: Stationary-Pedestrian 0-10 km/h (optimized), Vehicular 10-200 km/h (graceful degradation) and High Speed Vehicular 120-350 km/h (maintain

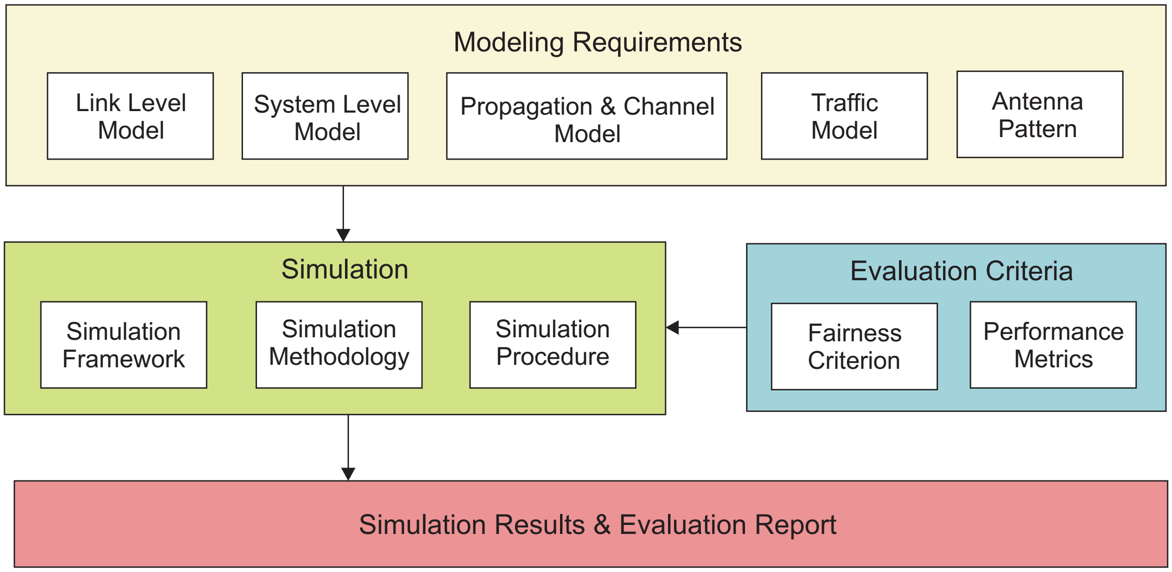


Fig. 6. Simulation Components for Evaluation Methodology.

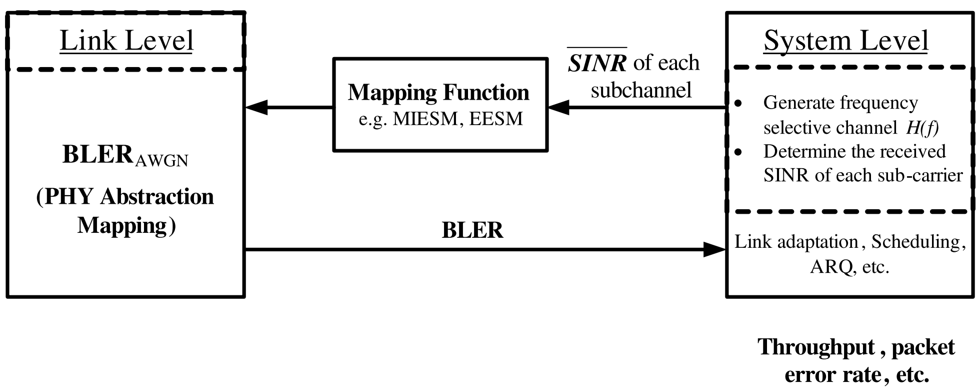


Fig. 7. PHY link-to-system mapping procedure.

connection). The simulation scenario may include either a two-cell topology were the mobile station is moving from Cell 1 to Cell 2 following the trajectories of Fig. 8, or a single moving mobile station in a 10-cell scenario. Each cell has three sectors and frequency reuse is modeled by planning frequency allocation in different sectors. Path loss and fading are updated (based on location and velocity) and the simulation statistics are measured on the moving mobile station.

The trajectories in the single moving mobile station scenario are the following:

- *Trajectory 1*: Begins from the center of cell 1 and ends at the center of cell 2. The purpose is to evaluate a case where the mobile station is moving from a high signal strength position with respect to cell 1 (low with respect to cell 2) to a low signal strength position (high with respect to cell 2).
- *Trajectory 2*: The mobile station is moving along the boundary of two adjacent sectors and until it reaches the midpoint of the cells boundary.
- *Trajectory 3*: The mobile station is now moving from the center of cell 2, along the boundary of two adjacent sectors and towards the center of cell 1.

2) Multiple Moving MS: In the Multiple Moving MS scenario the mobile stations are uniformly distributed over a 19-cell topology. Each initial angle trajectory is again chosen randomly, at the beginning of the call, where angle 0 corresponds to the North of the simulation environment. The mobility speed can take one of the values described in the single moving mobile station, and the mobile station users remain at that speed and direction for the duration of the simulation drop.

In order to model the interference from other cells, the total network topology requires 7 clusters, each cluster consisting of 19 cells. Depending on the configuration, the impact of the outer six cells may be neglected and only the central cluster is modeled. Thus, the simulation parameters and distribution of the nodes are set for the central cluster (cluster-0). The six outer clusters are exact copies of cluster-0, so that the corresponding cells have the same antenna configurations, traffic, fading, etc. except for the location.

Since each mobile station is associated with a serving cell, a two-step algorithm is employed to identify the sector/cell to which the mobile station belongs. First, the 19 shortest distance cells for each mobile station from all seven logical cell clusters are determined. Then the mobile station is associated

$$\text{Average Radio Layer Latency} = \frac{\sum_{i=1}^{N_{HO_{success}}} (T_{2,i} - T_{1,i})}{N_{HO_{success}}}, \tag{1}$$

$$\text{Average Network Entry and Connection Setup Time} = \frac{\sum_{i=1}^{N_{HO_{success}}} (T_{3,i} - T_{2,i})}{N_{HO_{success}}}, \text{ and} \tag{2}$$

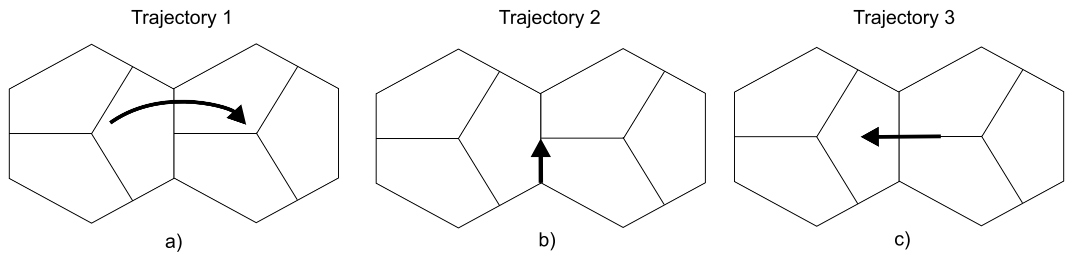


Fig. 8. EVM Handover Trajectories.

with the cell/sector that has the best link quality based on path loss and shadowing.

3) Handover Performance Metrics: For the evaluation of the performance of different handover schemes the following statistics/measurements shall be kept, for the entire simulation time, and are calculated based on the number of successful handovers $N_{HO_{success}}$, the number of failed handovers $N_{HO_{fail}}$, and the number of handover attempts $N_{attempt} = N_{HO_{success}} + N_{HO_{fail}}$ [17]. See Equations (1) and (2) above.

$$\text{Handover Interruption Time} = T_{3,i} - T_{2,i}, \tag{3}$$

where $T_{1,i}$ is the time instant that a mobile station transmits to the serving base station its commitment to handover (the time the mobile station disconnects from the serving base station for Hard Handover); $T_{2,i}$ is the time instant that the mobile station successfully synchronizes with the target base station; and $T_{3,i}$ is the time instant of the start of transmission of the first data packet between the mobile station and the target base station. The Handover Interruption Time represents the time interval during which a mobile station cannot receive service from any base station during handover. Two additional performance metrics are

$$\text{Data Loss} = \frac{\sum_{i=1}^{N_{HO_{success}}} (D_{TX,i} - D_{RX,i})}{N_{HO_{success}}} \text{ and} \tag{4}$$

$$\text{Handover Failure Rate} = \frac{N_{HO_{fail}}}{N_{attempt}}, \tag{5}$$

where $D_{TX,i}$ and $D_{RX,i}$ correspond to the number of bits transmitted and received during the handover process i. It is useful to mention that a handover failure will happen if the handover is executed while the reception conditions are inadequate on either the downlink or the uplink such that the mobile station would have to move to a Network Entry State.

VI. Conclusion

The recent proliferation of quadruple-play services (High Quality Voice and Video, the Internet and mobility) has led not only to an increase in the demand for bandwidth for broadband access networks, but also to the requirement of flexible Next Generation Wireless Systems. To address these needs, IEEE 802.16m appears as a strong candidate for providing aggregate rates to high-speed mobile users at the range of Gbps, while guaranteeing flexibility and backwards compatibility with existing systems. Although these issues have been addressed by several scientific groups, this is typically done either from the research or the engineering perspective. In this survey an attempt was made to combine these approaches into a unified view.

The main focus of this survey was to outline the evolution of several parameters and the way that they can be unified in order to build the structure of IEEE 802.16m systems. From the PHY layer perspective, the focus was on flexibility enhancements, aiming at supporting heterogeneous users, on extensions of MIMO transmission and resource allocation techniques, and on ways to guarantee interoperability with other access technologies. A collection of MAC layer functionalities was also analyzed whose goal is to combine energy conservation and security algorithms, as well as handover and relaying mechanisms with Quality of Service requirements. Because the combination and interdependence of multiple parameters increases the complexity when introducing research results or when submitting new proposals, the fundamentals of a unified way of simulation modeling and corresponding parameters, namely the IEEE 802.16m Evaluation Methodology, were also discussed.

Clearly, several unresolved issues remain to this date and, at this stage, it is not possible to predict accurately when the 802.16m standard will be ratified and to what extent the new approaches will be incorporated in its final form. Therefore, it is hoped that this survey will serve as a reference on the most relevant technical issues and that it will assist the interested reader in identifying the most challenging and interesting parts of the algorithmic design of Next Generation Wireless Systems.

References

[1] http://www.itu.int/md/R07-IMT.ADV-C
[2] http://www.itu.int/home/imt.html
[3] http://www.3gpp.org/article/lte-advanced
[4] http://wirelessman.org/tgm/

[5] Data Over Cable Service Interface Specification (DOCSIS 3.0), Third-generation transmission systems for interactive cable television services - IP cable modems, ITU-T Recommendation J.222

[6] B. Li, Y. Qin, C. P. Low, and C. L. Gwee, "A Survey on Mobile WiMAX," in *IEEE Commun. Mag.,* vol. 45, no. 12, pp. 70–75, Dec. 2007.

[7] K. Etemad, "Overview of Mobile WiMAX Technology and Evolution," in *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 31–40, Oct. 2008

[8] P. Taaghol, A. K. Salkintzis and J. Iyer, "Seamless Integration of Mobile WiMAX in 3GPP Networks," in *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 74–85, Oct. 2008

[9] M. Riegel, "Ethernet Services over Mobile WiMAX," in *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 86–93, Oct. 2008

[10] F. Wang, A. Ghosh, C. Sankaran, P. J. Flemming, F. Hsieh, and S. J. Benes, "Mobile WiMAX Systems: Performance and Evolution," in *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 41–49, Oct. 2008

[11] `http://www.wimaxforum.org/`

[12] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband Wireless Access with WiMAX/802.16: Current performance benchmarks and future potential," in *IEEE Commun. Mag.,* vol. 43, no. 2, pp. 129–136, Feb. 2005.

[13] IEEE Std 802.16e-2005, "IEEE Standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems, amendment 2: physical and medium access control layers for combined fixed and mobile operation in licensed bands," Feb. 2006.

[14] D. Tse, and P. Viswanath, "Fundamentals for Wireless Communication," Cambridge University Press, 2005.

[15] M. Cudak, ed., "IEEE 802.16m System Requirements Document, draft IEEE 802.16m-07/002r9," Sep. 2009.

[16] S. Hamiti, ed., "IEEE 802.16m System Description Document, draft IEEE 802.16m-09/0034r2," Sep. 2009.

[17] R. Srinivasan, ed., "IEEE 802.16m Evaluation Methodology Document (EMD), draft IEEE 802.16m-08/004r5," Jan. 2009.

[18] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An Overview of Limited Feedback in Wireless Communication Systems," in *IEEE J. Select. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[19] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel," in *IEEE Trans. Inform. Theory,* vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[20] G. Dimic, and N.D. Sidiropoulos, "On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm," in *IEEE Trans. Signal Processing,* pp. 3857–3868, vol. 53, no. 10, Oct. 2005.

[21] T. Yoo, and A. Goldsmith, "On the Optimality of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming," in *IEEE J. Select. Areas Commun.,* pp. 528–541, vol. 24, no. 3, Mar. 2006.

[22] Samsung Corp., "Downlink MIMO for EUTRA, R1-060335," 3GPP TSG RAN WG1 Meeting #44, Denver, USA, Feb 2006.

[23] S. Saur, H. Halbauer, A. Rueegg, and F. Schaich, "Grid-of-Beams (GoB) Based Downlink Multi-User MIMO, IEEE C802.16m-08/487," May 2008.

[24] W. C. Wong, and S. Talwar, "Interference mitigation using downlink transmit beamforming with nulling techniques, IEEE C802.16m-08/653r2," Jul. 2008.

[25] H. Kwon, and B. G. Lee, "Distributed Resource Allocation through Noncooperative Game Approach in Multi-Cell OFDMA Systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Istanbul, June 2006.

[26] IEEE Std 802.16j-2006, "IEEE Standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems, Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Multihop Relay Specification," March 2006.

[27] S. W. Peters and R. W. Health Jr. "The future of WiMAX: Multihop Relaying with IEEE 802.16j," in *IEEE Commun. Mag.,* vol. 47, no. 1, pp. 104–111, Jan. 2009.

[28] S. Ahson and M. Ilyas, "The WiMAX Handbook. Vol: Standards and Security," CRC Press, Taylor & Francis Group, 2007.

[29] J.-F. Cheng, "Coding Performance of Hybrid ARQ schemes,", in *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1017–1029, June 2006.

[30] D. Toumpakaris, J. Lee, A. Matache, and H.-L. Lou, "Performance of MIMO HARQ Under Receiver Complexity Constraints," in *Proc. IEEE Globecom 2008*, New Orleans, USA, Dec. 2008.

[31] J. Lee, D. Toumpakaris, H-L. Lou, E. W. Jang, and J. M. Cioffi, "Transceiver Design for MIMO Wireless Systems Incorporating Hybrid ARQ," in *IEEE Commun. Mag.,* vol. 47, no. 1, pp. 32–40, Jan. 2009.

[32] Z. Ding, and M. Rice, "Hybrid-ARQ code combining for MIMO using multidimensional space-time Trellis codes," in in *Proc. IEEE ISIT '07*, Glasgow, Scotland, UK, Jun. 2007.

[33] G.S. Paschos, I. Papapanagiotou and T.M. Bohnert, "Power Saving Mechanisms in IEEE 802.16e Wireless Networks," Chapter in "WiMAX Networks Planning and Optimization," Auerbach Publications, CRC Press, 2009.

[34] H-H. Choi and D-H. Cho, "Mobility Support for IEEE 802.16e System," Chapter in "WiMAX Standards and Security," pp. 103–128, CRC Press, 2008.

Ioannis Papapanagiotou [S 05] received his Diploma in Electrical & Computer Engineering from the University of Patras, Greece in 2006, and his MS in Electrical Engineering from North Carolina State University, in 2009, where he is currently a PhD student. His main research interests lie in the area of Network Design & Dimensioning, network pricing and Wireless NGNs. Ioannis has received the Best Paper Award, the NSF Student Travel Grant Award in IEEE GLOBECOM 2007, and is supported by the Academy of Athens.

Dimitris Toumpakaris [S'98, M'04] received his Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 1997, and his MS and PhD degrees in Electrical Engineering from Stanford University in 1999 and 2003, respectively. He was a Senior Design Engineer in Marvell Semiconductor Inc., Santa Clara, California from 2003 to 2006. He has also worked as an intern for Bell-Labs, CERN and France Telecom, and as a consultant for Ikanos Communications and Marvell Semiconductor Inc. He is currently an Assistant Professor in the Wireless Telecommunications Laboratory, Department of Electrical & Computer Engineering, University of Patras, Greece. His current research interests include information theory with emphasis on multi-user communications systems, digital communication, synchronization and estimation, and cross-layer optimization.

Jungwon Lee [S00, M05] received his BS degree in Electrical Engineering from Seoul National University in 1999, and he received his MS and PhD degrees in Electrical Engineering from Stanford University in 2001 and 2005, respectively. From 2000 to 2003, he worked as an intern for National Semiconductor, Telcordia Technologies, and AT&T Shannon Labs Research and as a consultant for Ikanos Communications. Since 2003, he has worked for Marvell Semiconductor Inc., Santa Clara, California, where he is now a Senior Manager/Principal Engineer leading algorithm development and system architecture design for the next generation wireless communication network. His research interests lie in wireless and wireline communication theory with emphasis on multi-input multi-output (MIMO) orthogonal frequency division multiplexing (OFDM) system design.

Michael Devetsikiotis [S 1985, M 1993, SM 2004] received his Diploma in Electrical Engineering from the Aristotle University of Thessaloniki Greece, in 1988, and the M.Sc. and Ph.D. degrees in Electrical Engineering from North Carolina State University, Raleigh, in 1990 and 1993, respectively. He became a faculty in the Department of Systems and Computer Engineering at Carleton University in April 1995, and then joined the Department of Electrical and Computer Engineering at NC State where he is now a full Professor. He is also an active member of the Operations Research faculty, and an associate member of the faculty of Computer Science at NC State. He recently finished his stint as Chairman of the IEEE Communications Society Technical Committee on Communication Systems Integration and Modeling. He is an Area Editor of the ACM Transactions on Modeling and Computer Simulation and a member of the editorial board of the Journal of Internet Engineering, the IEEE Communication Surveys and Tutorials and the International Journal of Simulation and Process Modeling. Michael was recently selected to become an IEEE/Communications Society Distinguished Lecturer for the period 2008-2009.