Journal of Computer Systems, Networks, and Communications

WiMAX, LTE, and WiFi Interworking

Guest Editors: Rashid A. Saeed, Ahmed A. M. Hassan Mabrouk, Amitava Mukherjee, Francisco Falcone, and K. Daniel Wong

WiMAX, LTE, and WiFi Interworking

Journal of Computer Systems, Networks,
and Communications

WiMAX, LTE, and WiFi Interworking

Guest Editors: Rashid A. Saeed, Ahmed A. M. Hassan Mabrouk,
Amitava Mukherjee, Francisco Falcone, and K. Daniel Wong

Contents

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 754187, 2 pages
doi:10.1155/2010/754187

Editorial
WiMAX, LTE, and WiFi Interworking

Rashid A. Saeed,[1] Ahmed A. M. Hassan Mabrouk,[2] Amitava Mukherjee,[3] Francisco Falcone,[4] and K. Daniel Wong[5]

[1] *Department of Electrical and Computer Engineering, Engineering Faculty, IIUM, Kuala Lumpur, Malaysia*
[2] *Information and Communication Technology Faculty, IIUM, Kuala Lumpur, Malaysia*
[3] *IBM India Private Limited, Salt Lake, Calcutta 700 091, India*
[4] *EE Department, Universidad Pública de Navarra, Campus de Arrosadía, Pamplona, 31006 Navarre, Spain*
[5] *Daniel Wireless LLC, Palo Alto, CA 94306, USA*

Correspondence should be addressed to Rashid A. Saeed, eng_rashid@ieee.org

Received 28 April 2010; Accepted 28 April 2010

Recently wireless network interworking has become an important area of research in academia and industry. This is due to the huge diversity of wireless network types, which range from wireless body area network (WBAN) covering areas up to a few inches to wireless regional area networks (WRANs) covering up to several miles. All these types of networks have been developed separately with different usage and applications scenarios, which make interworking between them a challenging task.

The main challenges in wireless interworking of connecting the cellular network with the other wireless networks include issues like security, seamless handover, location and emergency services, cooperation, and QoS. The developed interworking mechanisms, that is, unlicensed mobile access (UMA), IP Multimedia Subsystem (IMS), and Media independent handover (MIH), due to the characteristics of wireless channel, need to be analyzed and tested under various circumstances.

The aim of this special issue in Journal of Computer Systems, Networks, and Communications (JCSNC) is to highlight the problems and emphasize and analyze the solutions in this area, which can give a guideline to telecom industry for new techniques and business opportunities. Many researchers from different parts of the world and different background have participated in the issue. The accepted papers are diverse at different interworking levels, spanning from network layer down to link level for example, the paper entitled "*Technology-integration framework for fast and low cost handovers, case study: WiFi-WiMAX network*" by M. Kassab, et al.where the end-to-end delay is optimized with minimum management signaling cost.

On the other hand, in "*WiFi and WiMAX secure deployments*" by P. Trimintzios and G. Georgiou, the security intrusion that may occur during handover is discussed. In the paper "*Seamless video session handover between WLANs*" by C. C. Monteiro, et al. an architecture for session proxy (SP) with video streaming quality preservation has been developed.

At the physical layer, the paper "*Investigation of cooperation technologies in heterogeneous wireless networks*" by Z. Sun and W. Wang discussed the radio access technology (RAT) for various standards, where issues like multiradio resource management (MRRM) and generic link layer (GLL) were proposed. In the paper entitled "*A multi-standard frequency offset synchronization scheme for 802.11n, 802.16d, LTE and DVB-T/H systems*" by J. González-Bayón et al. carrier frequency offset in OFDM systems is discussed where common synchronization structure for all these systems is proposed.

C. So-In et al. in "*Capacity Evaluation for IEEE 802.16e Mobile WiMAX*" emphasize on the overhead of the WiMAX protocol and its effect on the link capacity. Many applications have been tested that is, Mobile TV and VOIP. In the same area P. T. T. Pham and T. Wada's paper "*Effective scheme of channel tracking and estimation for mobile WiMAX DL-PUSC System*" discussed the packet error rate (PER) and user throughput in various channels.

K.-P. Lin and H.-Y. Wei discussed a new random walk mobility model in "*Paging and location management in IEEE*

802.16j multihop relay network". The proposed model is suitable for multihop relay network, where the handover process is frequently performed.

Finally, "*Multi mode flex-interleaver core for baseband processor platform*" by R. Asghar and D Liu introduces a new flexible interleaver architecture supporting many standards like WLAN, WiMAX, HSPA+, LTE, and DVB at the system level. Both maximum flexibility and fast switchability were examined during run time.

This special issue would not have come true without the tight guidelines and support from the Editor-in-Chief Professor Hsiao-Hwa Chen and Mariam Albert the editorial staff in Hindawi Publishing Corporation.

Rashid A. Saeed
Ahmed A. M. Hassan Mabrouk
Amitava Mukherjee
Francisco Falcone
K. Daniel Wong

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 205786, 21 pages
doi:10.1155/2010/205786

Research Article

Technology Integration Framework for Fast and Low Cost Handovers—Case Study: WiFi-WiMAX Network

Mohamed Kassab,[1] Jean-Marie Bonnin,[1] and Abdelfettah Belghith[2]

[1] *Telecom Institute/Telecom Bretagne/RSM Department, Université Européenne de Bretagne, 35510 Cesson Sevigné, France*
[2] *ENSI/CRISTAL Lab/HANA Research Group, University of Manouba, 2010 Manouba, Tunisia*

Correspondence should be addressed to Mohamed Kassab, mohamed.kassab@gmail.com

Received 1 October 2009; Revised 14 February 2010; Accepted 18 April 2010

Academic Editor: K. Daniel Wong

The Next Generation Wireless Networks (NGWNs) are seemed to be heterogeneous networks based on the integration of several wireless technologies. These networks are required to achieve performances equivalent to classic wireless networks by ensuring the continuity of communications and the homogeneity of network management during horizontal and vertical handovers. This task is even more important when management services, like security and quality of service (QoS), are deployed at access technology level. In this paper, we propose a framework for heterogeneous wireless technology integration based on network architecture skeleton and a handover management mechanism. This framework optimizes the layer-2 handover procedure to achieve performances required by sensitive applications while ensuring the minimization of signaling overhead required for operated networks. As an application example, we make use of this framework to propose a heterogeneous network based on WiFi and WiMAX technologies. We present an application example of the framework using the specification of a WiFi-WiMAX network. We propose several performance evaluations based on simulation tests based on this application. The latter confirm the efficiency of handover delay optimization and the minimization of management signaling costs.

1. Introduction

The growth of wireless communication has been, in a few years, important thanks to the advantages they offer such as deployment flexibility and user mobility during communications. Several wireless technologies have emerged. These technologies have been designed independently and intended to cover specific service types, user categories, and usability domains. Among these technologies, there is not one good and generic enough to replace all the others; each technology has its own merit, advantages, and development possibilities. For example, 3G technologies, for example, UMTS and CDMA2000, propose network access associated to telephony services. WMAN technologies, for example, WiMAX and HyperMAN, are used to deploy outdoor metropolitan networks. WLAN technologies, for example, WiFi, have been developed to be an extension of already existing wired LANs; they are also used to deploy local public wireless networks. In addition, user categories and usability domains have converged so that terminals and communication means have evolved to integrate multiple technologies.

The result of this evolution is a multitechnology environment that can be exploited to offer an enhanced connectivity to users. The Next Generation Wireless Networks (NGWNs) appear to be the integration of already existing and newly developed wireless technologies that offers a heterogeneous access to the same global core network. A multi-technology terminal will be able to change its access technology each time its environment changes. For example, it will be connected to a WiFi access point when it is in the mall; it will handover to the WiMAX when it will move to the street and it will use UMTS in the train. This could be a great advance depending on the adequate mechanisms which are available to ensure a seamless mobility.

On the other hand, wireless technologies are no longer limited to be a basic communication medium. They evaluate by integrating several management services such as

user authentication, data exchange confidentiality, and QoS management. However, the integration of these services at the access technology level with specific designs will affect the handover performances in NGWNs. In fact, the change of the serving Point of Attachment (PoA) requires the renegotiation of management services between the terminal and the network in addition to the redirection of data traffic to the new terminal location. As a result, the HO execution time may increase significantly, which should induce significant latency to exchanged data and even the break of the ongoing session.

Public wireless networks have to guarantee a good level of service while insuring the transparency of management to users. The deployment of such networks using heterogeneous technologies will require a good connectivity during handovers, by reducing latency, and the homogeneity of management services such as authentication and QoS. This is possible by deploying anticipation mechanisms that reduce negotiation exchanges between the terminals and the network, such as context transfer and proactive negotiation [1], and accelerate the redirection data traffic during the execution of the HO.

Researchers have been interested in this problem and several papers have proposed models for efficient technology-integration solutions that deal with network access provider requirements. However, the mobility management offered by these solutions does not ensure yet seamless handovers during heterogeneous mobility. Indeed, most solutions offer roaming possibilities based on the sharing of user databases. At best, the integration architectures offer to graft one technology to another and to manage heterogeneous mobility based on Mobile IP and extensions. These solutions enable the optimization of the network reattachment (i.e., the layer-3 HO) by limiting the heterogeneous handover to the reattachment to the new PoA (i.e., layer-2 HO). This does not solve the connectivity disruption due to the re-establishment of network services defined at the technology level. On the other hand, the structure of these technology-integration solutions is not suited to heterogeneous mobility. Indeed, the organization of the PoAs in the core network is based on the access technology they offer rather than the closeness of radio coverage while the executed HOs will be based on the latter closeness. As a consequence, the HO management mechanisms based on exchanges between heterogeneous entities will result in a nonnegligible overhead that could disrupt the network performances.

In this work, we propose a technology-integration framework that provides a new approach to deploy next generation wireless networks. This framework offers a heterogeneous access to a global network with optimized mobility performances regarding HO execution time and signaling cost. The idea is to optimize the layer-2 HO execution in a heterogeneous and homogeneous mobility and to adapt the network architecture so that this optimization yields to a minimum signaling surplus. The framework defines a network architecture skeleton and HO management mechanisms. They tend to optimize the layer-2 HO execution while ensuring the continuity of management services defined at the technology-level. In addition, we propose an application of this framework to an actual wireless network based on the WiFi and WiMAX technologies. We make use of this application to demonstrate the ability of the proposed framework to enable the enhancement of HO performances while ensuring a reduced signaling overhead.

This paper is organized as follows. In Section 2, we propose an overview of solutions adopted for wireless technology integration. In Section 3, we detail the specification of the technology-integration framework. We propose, in Section 4, the specification of wireless network based on the WiFi and WiMAX technologies. We demonstrate the advantages offered by this architecture based on performances evaluations in Section 5. We detail how the proposed framework can get along with layer-3 mobility management mechanisms in Section 6. We propose, in Section 7, a discussion about heterogeneous technology integration. We draw up main conclusions and propose future trends of our work in Section 8.

2. Technology Integration in the Literature

Heterogeneous-technology integration has been studied by several researches. Most studies focused on networks integrating UMTS and data wireless technologies, that is, WiFi [2–6] and WiMAX [7–9]. Two inter working architectures have been proposed: loosely and tightly coupled architectures [2, 10].

With loosely coupled architecture, the interconnected technologies are considered as independent networks concerning the handling of data traffic and the management of network services such as authentication and QoS. Each technology has a separate user subscription and profile management systems. Roaming privileges are assigned to subscriptions related to one network. This helps to minimize session disruption based on the cooperation of accounting entities. The tightly coupled architecture proposes the integration of wireless technologies in the same network architecture. This integration may be performed in different levels of the management architectures of the considered technologies. User subscriptions and profiles are management based on common centralized entities. In all cases, user mobility is managed using Mobile IP and its extensions [11].

The main advantage of loosely coupled architectures is the few modifications to technologies and their core network architectures. However, due to the high level of integration, the mobility management mechanisms are not able to optimize significantly the performance of layer-3 handover. Thus, the roaming mechanisms are not able to reduce sufficiently the session disruption to deal with requirements of sensitive applications.

The tightly coupled architectures propose integration at lower level of network architecture. The complexity of the implementation increases, and more modifications must be operated to technologies and core network architecture. Nevertheless, the lower level of integration ensures a very interesting enhancement of HO performances [4, 5]. This is due to the fact that the inter-working takes place at a point of the management architecture closer to the mobile terminal.

The tightly coupled architecture can significantly improve the performance of heterogeneous handovers. This can be even more enhanced by using the ConteXt Transfer Protocol (CXTP) [12] in addition to MIP. The CXTP proposes a protocol to transfer mobile terminal contexts between Access Routers managing the access control of a wireless network. CXTP has been designed as a generic protocol that can accommodate a wide range of services. The context transfer can be reactive, during the HO execution, or proactive from the serving AR to a possible target AR. CXTP can be useful if some network services such as user authentication and QoS are integrated to the layer-3 level in wireless networks [13]. Consequently, several management exchanges between a terminal and the Access Router (AR), which controls the access to the network, are required during the network entry. Thus, the CXTP enables the reduction of exchanged messages between mobile terminal and target AR during the HO execution.

However, the latter optimization limits only the effects of sub network change during terminal mobility (layer-3 HO optimization). Indeed, all the negotiation exchanges and the service establishment procedures defined at access-technology level must be performed during heterogeneous handover executions.

A solution could be the association of the tightly coupled architectures to an optimization of the terminal to technology association procedure. This optimization will take into account the possible resemblances between the definition of services and user profiles of technologies to prevent the execution of the negotiations and procedures during handover executions. This may be based on management mechanisms like context transfer or proactive execution of exchanges.

3. Technology-Integration Framework

This framework aims at defining an optimization of the handover performances as part of a heterogeneous mobility.

We consider an operator network that offers a reliable network access, to mobile terminals, based on several wireless technologies. Network services, such as user authentication, QoS management, and billing, have to work properly and seamlessly while terminals are moving over the network. We define the network architecture and the position of management entities that are involved in the handover management procedure.

The proposed framework specifies the skeleton of the network architecture, the definition of mobility context and the L2-HO management mechanisms. The latter proposes the enhancement of L2-HO performances based on mobility-context exchanges.

3.1. Network Architecture Skeleton. The global wireless network is organized into *access subnetworks*, each one gathering a set of PoAs. We do away with the classic organization of wireless networks that separates each technology in an autonomous network. PoAs can be gathered in access sub networks based on the closeness of their wireless coverage or based on common management requirements. It also remains possible to gather PoAs offering the same wireless access technology. We define new management entities: *the Layer 2 Access Managers (L2-Acc-Mgrs)* that manage terminal mobility over the network. To each access subnetwork is associated an L2-Acc-Mgr. Figure 1 shows this architecture.

The L2-Acc-Mgr integrates several functions to manage terminal mobility. It acts as a *service proxy* regarding exchanges between terminals and core network entities during the network entry procedure. For example, terminal authentication is supported by the L2-Acc-Mgr that acts as AAA-proxy between the terminal and the AAA server in the core network. At the end of this procedure, the L2-Acc-Mgr maintains the terminal authentication profile (authentication keys) to use it for future purposes.

The L2-Acc-Mgr supports the *Neighborhood management function* that maintains the PoAs' neighborhood. It provides a list of PoAs to which a terminal may move while being associated with a particular PoA.

The *L2-HO management function* integrates the intelligence related to the L2-HO management, that is, the triggering of HO management exchanges, the execution of exchanges and the management of terminal contexts.

3.2. L2-HO Management Mechanisms. During the network entry, a terminal associates itself with the network and activates a set of services and functionalities. The *terminal context* includes the parameters negotiated during the network entry and states related to network services used by the terminal [1]. The acceleration of the establishment of this context is required, at the time of handover, to reduce the delay that results from the HO execution phase. The establishment of the terminal context on the target PoA, based on already available information, is the solution.

The nature of information elements included in the terminal context defines how it can be exploited to perform a context re-establishment. This defines values of information elements to be established, when and how they will be established, and the network entities that have to manage these information elements [1]. Authors in [14] propose a study that define the latter points based on the characteristics of information elements and particularly:

(i) the scope of the information element,

(ii) the transferability of the information element,

(iii) and the stability of the information element value over the time.

In the following part, we identify the network entities that will manage the context establishment, the values to be established, the mechanisms that establish contexts, and finally when the establishment has to be performed (i.e., before, during, or after the HO execution), while taking into account the network architecture decided upon and the nature of information elements that may be included in terminal contexts.

3.2.1. Management of Terminal Contexts. Regarding the scope, a terminal context consists of *global session* and *local*

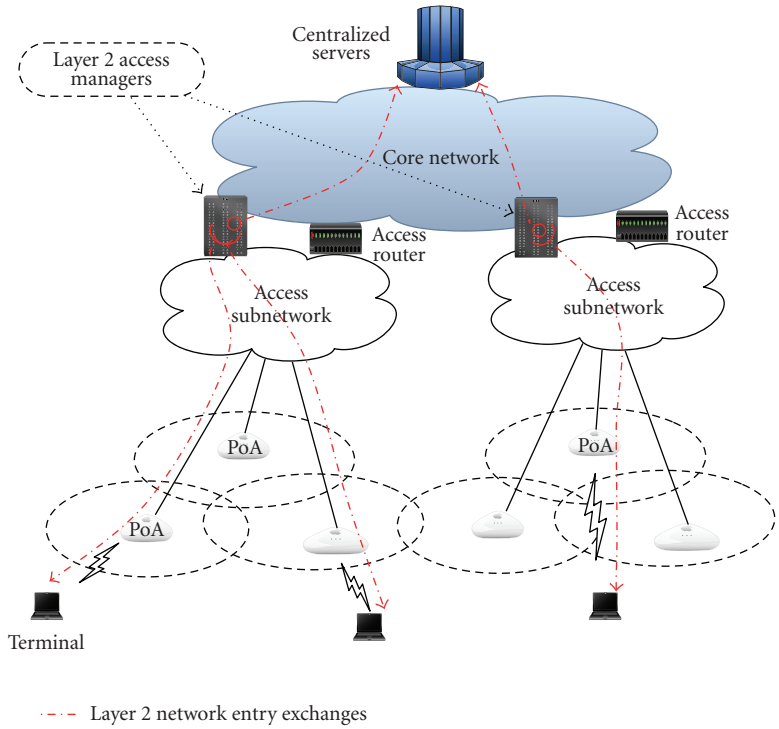


Figure 1: The L2-Acc-Mgr in the network architecture.

association information elements. The *global session* information elements are related to the association established between the terminal and the core network entities such as AAA servers. The *local association* information elements are related to the association established between the terminal and the serving PoA. When a terminal executes a HO without performing a new network entry, it maintains its *global session* while re-establishing the *local association* with the new serving PoA.

Then, a context information element is *transferable information* when it remains valid while the terminal changes its serving PoA. Such information element can be reused with target PoA to avoid renegotiation during HO execution. Other elements are *nontransferable context information*, their current value, associated to a serving PoA, cannot be exploited to avoid negotiations between the terminal and target PoA to establish a new association. This type of information has to be re-established through regular exchanges during the HO execution. Finally, an information element can be *conditionally transferable* if the value associated to the serving AP is not valid for transfer; however, it can be used to define a new value associated to target PoAs. It is possible to define *translation rules* for this specific set of information elements so as to enable their establishment while avoiding negotiations during HO execution.

Based on these two classifications we define the content of terminal contexts and the entities that have to manage these contexts, following the recommendation proposed in [1].

The L2-Acc-Mgr is the most entitled entity to manage the greater part of the terminal context. First, the global session information elements are held by the L2-Acc-Mgr thanks to the *service proxy function*. Second, local information elements that are conditionally transferable may require centralized information related to the neighbor PoAs or the terminal to be translated for re-establishment. The latter information is held by the L2-Acc-Mgr, so it is the better able to manage conditionally transferable local information elements. The *HO management function* of the L2-Acc-Mgr is responsible of managing the latter information elements, of the terminal context.

The *HO management function* defines the values for information elements to be established by the L2-Acc-Mgr. The latter values will be derived based on the ones used with the current association, cached information elements or terminal accounting profile. A *Translation function* is defined as a part of the *HO management function*. It is responsible of defining values to be established for information elements constituting the context terminal.

This case can be illustrated over a heterogeneous wireless network offering access to multi-technology terminals. A mobile terminal can switch between two PoAs offering heterogeneous technologies. In this case, QoS parameters can be transferred to re-establish the new association since the two wireless technologies do not necessarily use the same QoS representation. A QoS translation function can solve the conformity problem as most QoS management mechanisms have common bases.

The definition of new values for a context information element may result into a synchronization problem between the terminal and the network. Indeed, the terminal must be able to integrate the translation subfunction used by the

L2-Acc-Mgr to define the new information element value. Therefore, *the translation rules* are defined so that both the terminal and the L2-Acc-Mgr can compute a value that corresponds to the new association without performing any exchange.

The local information elements that have values valid for different local associations (transferable information), are managed by the PoAs. A serving PoA is responsible for redistributing them to target PoAs and caching them.

Finally, there is a set of information elements that current values cannot be exploited to avoid management exchanges between a mobile terminal and the network to establish a new association. We name this category: *non transferable context information*. This type of information has to be re-established through regular exchanges during the handover execution. We can mention connection parameters used with a terminal, for example, data rate. These parameters depend on the position of the terminal in the cell and the serving AP capacity, and so they have to be negotiated during the association.

3.2.2. Context Establishment Exchanges. Two options are available for context establishment: the context transfer and the proactive negotiation [1].

The context transfer is an adequate establishment solution for transferable information elements. It is performed between the entity managing the information element and one or a set of PoAs. In the same way, conditionally transferable information element re-establishment can be based on a context transfer mechanism. After being translated, an information element is transferred to target PoAs.

The context transfer is not the appropriate solution for the re-establishment of non-transferable information elements. An information element might require to be re-established over standard exchanges or the involvement of the terminal in the negotiation or generation process. It remains possible to establish non transferable information elements using *proactive negotiations*. The latter are based on the standard exchanges usually performed during the network entry procedure to generate information elements.

The adequate time to perform a context establishment depends on the stability of the information element value during the time. There are static information elements that values do not change during the local association and dynamic information elements that values change during a local association based on network conditions, terminal behaviors, accounting constraints, and so forth. Proactive context establishment can be performed with static information elements so that it will be available immediately at the HO execution. However, proactive establishment is not excluded with dynamic context. This depends on the frequency of information element update. If an information element is known not to be frequently updated, it remains possible to perform a *conditional proactive establishment*. The information element shall be associated to a *validity condition*. At the time of the handover, the information element is used only if the validity condition is verified. In other cases, the information element is established reactively during HO execution based on its last update.

3.2.3. HO Establishment Exchanges. Regarding our specification, the context transfer is suitable for information elements managed by the L2-Acc-Mgr. Proactive and reactive exchanges are combined to manage static and dynamic information elements. The exchange (a) of Figure 2 shows the proactive establishment procedure involving the L2-Acc-Mgr and two neighbor PoAs. The target PoA may execute a reactive exchange to obtain values related to dynamic information elements from the L2-Acc-Mgr as shown in Figure 2(b).

The establishment of local association information elements managed by serving PoA can be based on *context transfer* and/or *proactive negotiation*. These mechanisms may be combined to establish one or more information elements in the same procedure or used as alternatives for the same information element to define different procedures since they have different properties [1]. Figure 3 shows exchanges based on the two mechanisms.

The context transfer can be proactive and/or reactive. For the proactive one, the establishment exchanges are initiated by the serving PoA with a list of neighbor PoAs indicated by the L2-Acc-Mgr. During HO execution, a target PoA may require additional information elements from the serving PoA. As such, it can engage reactive context transfers with the previous serving PoA.

Proactive negotiations are engaged between the terminal and neighbor PoAs through the current association (established with the serving PoA). It is mostly used for information elements managed by PoAs that cannot be established through context transfer.

The L2-Acc-Mgr is responsible of managing L2-HO management exchanges with entities associated to its access subnetwork (i.e., PoAs and terminals) and L2-Acc-Mgrs from other access subnetworks. Consequently, the L2-HO management exchanges are limited to the access subnetwork during intrasubnet mobility. Intersubnetworks exchanges are relayed by L2-Acc-Mgrs during inter-subnetwork mobility. A target L2-Acc-Mgr converses with the serving L2-Acc-Mgr for centralized establishment exchanges as shown in Figure 4.

In a nonoptimized architecture, the HO management exchanges between PoAs are routed through the core network from one access subnetwork to another during intersubnet mobility. The HO management exchanges between PoAs and centralized entities, during an intra-subnet mobility event, are engaged through the core network while the terminal mobility is restricted to the access network. Thus, the use of L2-Acc-Mgrs restricts as much as possible the HO management operations to intra-access subnetwork exchanges. This may ensure the efficiency of these exchanges and reduce the signaling overhead over the core network.

4. WiFi-WiMAX Network

As an application of the technology-integration framework, we propose the integration of the WiFi and WiMAX

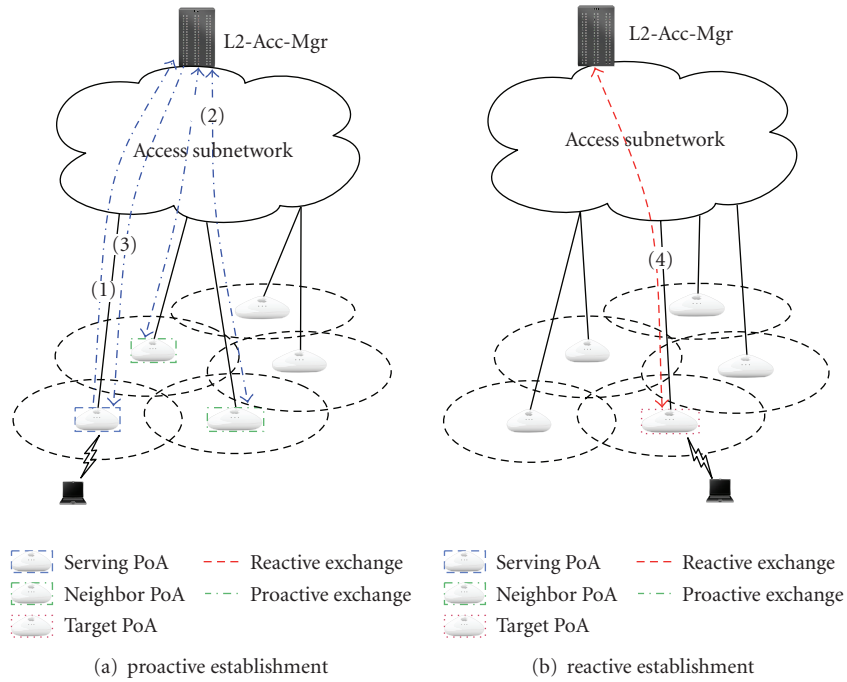


Figure 2: Centralized establishment.

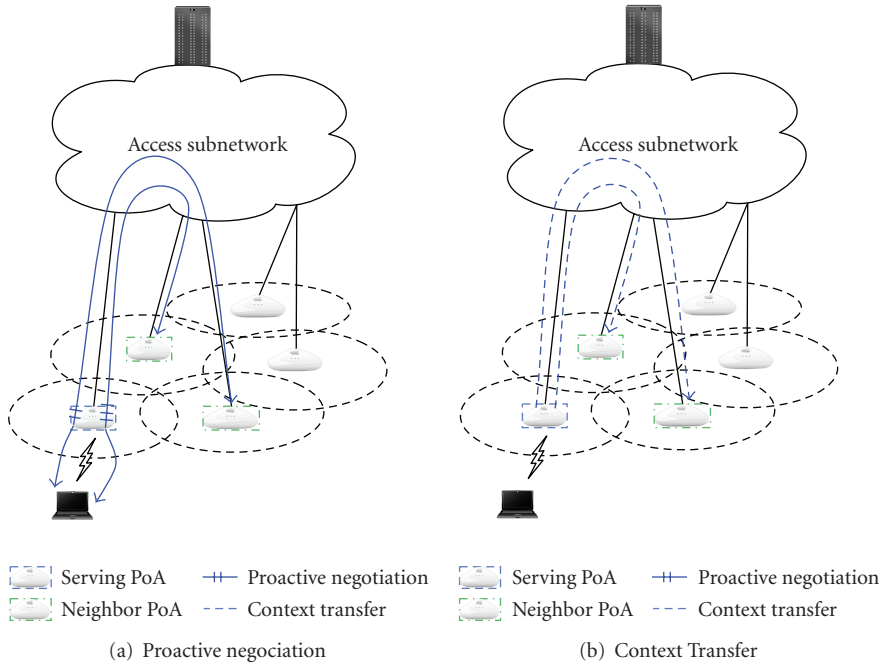


Figure 3: Distributed establishment.

technologies in a heterogeneous wireless network. This network offers to terminals a wireless connectivity adapted to their location. The WiMAX is deployed for an outdoor access and the WiFi in building for indoor access. Terminals will roam from one technology to another according to their movements while being attached to the same global network.

4.1. WiFi-WiMAX Integration in the Literature. Some researches were interested in the collaboration between WiFi and WiMAX technologies. Most of these researches have proposed to use the WiMAX technologies as backhaul support for WiFi hotspot [7, 15, 16]. Therefore, the designed networks did not fall within the category of 4G networks, and the two technologies do not cooperate to offer the wireless

access to mobile users. More recent research studies were interested in the inter-working of the WiFi and the WiMAX as access technologies in the same heterogeneous network. However, the majority of these studies were limited to the enhancement of the HO decision mechanism between the two technologies and did not discuss the problems related to the integration and the collaboration of these technologies in the same network architecture [17–19].

In [20], authors were interested in inter-working of the WiFi and the WiMAX technologies. They proposed a solution to ensure a continuity of QoS management through the heterogeneous wireless access. The solution proposes a mapping between the QoS management parameters of each technology to ensure seamless change of technologies. To fix the context of their work, authors tried to define an interconnection architecture for the network. They proposed the interconnection of separate WiFi and WiMAX access networks through a core network and to manage the layer-3 HO using Mobile IP. However, no additional management arrangements were proposed (e.g., collaboration between QoS accounting, context transfer between BSs and APs) to enable the use of the QoS mapping through the deployed access network.

Thus, at the best of our knowledge, there is no serious work that offers a design of a heterogeneous network integrating the WiFi and the WiMAX technologies.

4.2. Technologies' Overview. We propose an overview of the WiFi [21] and WiMAX technologies [22]. We focus particularly on the network architecture and the layer-2 network service defined by each technology and the manners in which they interact with mobility management.

4.2.1. WiFi. The WiFi technology is based on the IEEE 802.11 standard that defines the PHY and MAC layers for the wireless medium. This standard has been completed by several extensions that define services such as the QoS management and user authentication. The proposed specification is limited to the management of these services through the wireless part of the network and has not defined operations that involve centralized entities.

User authentication is proposed by IEEE 802.11i extension [23] that defines a robust securing mechanism offering a privacy equivalent to wired network. It proposes a complete security framework defining the security architecture, the key hierarchy, and the cryptographic mechanisms. The 802.11i authentication is based on an authentication key hierarchy and key generation exchanges. They establish mutual authentication between peers and generate cryptographic suite to secure data exchanges.

The basic IEEE 802.11 standard offered only a best effort service to an application flow. The QoS management for the WiFi technology has been defined by the IEEE 802.11e extension [24]. Two operation modes have been defined:

(i) a per-packet QoS management, *the prioritized QoS*, based on priorities associated to transmission queues with different channel access priorities,

Table 1: User priority to traffic class mapping.

User Priority	Traffic Type	Description
1	Background	Bulk transfers, games, etc.
2	Spare	
0	Best Effort	Ordinary LAN priority
3	Excellent Effort	Best Effort for important users
4	Controlled Load	Some important applications
5	Video	Less then 100 millisecond delay
6	Voice	less than 10 millisecond delay
7	Network Control	High requirements

(ii) a per-flow QoS management, *the parameterized QoS*, based on QoS parameters associated to virtual traffic stream. The latter are a set of data packets to be transferred in accordance with the QoS requirements of an application flow.

The WiFi equipments and deployed networks are followed by particular evolution. Indeed, the QoS management proposed by IEEE 802.11e was not adopted in network deployments. The enhancements of the communication performances were based on the evolution of the PHY layer performances.

With the WiFi-WiMAX integration, the WiFi technology will coexist with the WiMAX technology, which offers a strong service differentiation between categories of data traffics based on user profiling (c.f. the next subsection). So as to offer a homogenous network access service to users over the network, we propose to adopt a QoS-enabled WiFi access in our specification. We consider the *Parameterized QoS* as it most closely matches the QoS management defined by WiMAX [25].

The Parameterized QoS proposes a QoS management based on virtual connections: the Traffic Streams (TSs). The latter are sets of data packets to be transferred in accordance with the QoS requirements of an application flow. A terminal specifies TS requirements to the Access Point (AP) using the admission control exchange. The requirements can be data rate, packet size, service interval, and so forth. An AP may accept or reject new Traffic Specification requests based on the network conditions, terminal profile, and so forth. The traffic differentiation is based on traffic specification (TSPEC) associated to TSs. The TSPEC element contains a set of QoS parameters that define the characteristics and the QoS expectations of a traffic flow. In addition User Priorities (UP) are used to indicate the traffic class of the TS. Table 1 presents the mapping between UP values and traffic class.

The WiFi technology was developed to be an extension of wired networks and not as an operator technology such as WiMAX or UMTS. Thus, the IEEE 802.11 standard and its extensions have not specified the core network architectures and mechanisms. The deployment of RSN security and parameterized QoS requires an AAA server that manages the identities and the profiles of authorized users.

The negotiations defined by the WiFi authentication and the parameterized QoS, during the network entry, require considerable time, which turns into a connection

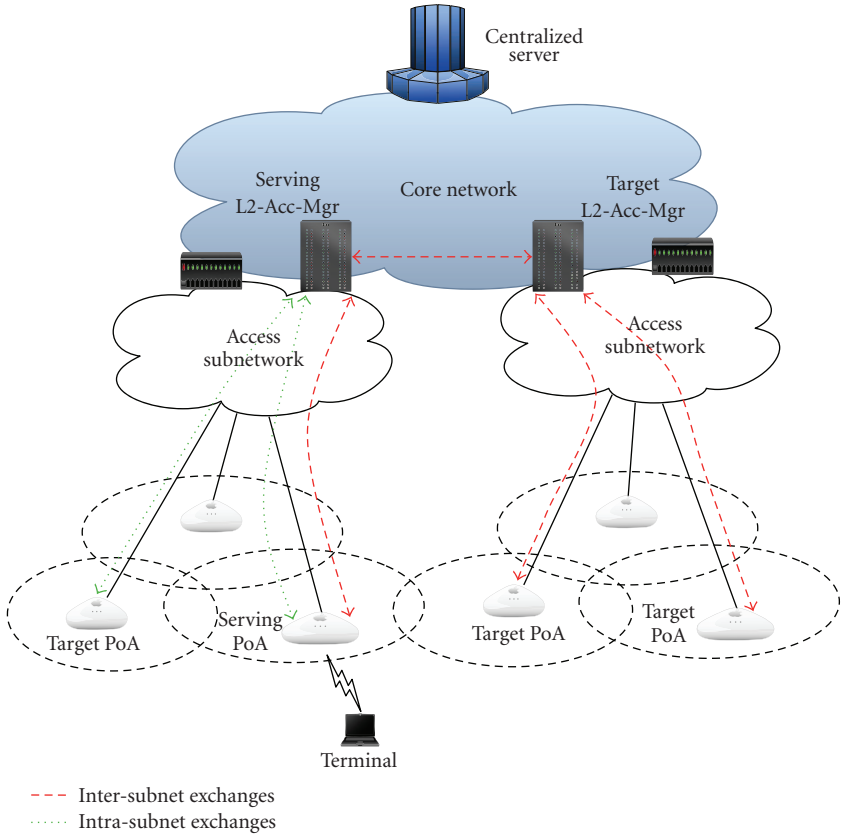


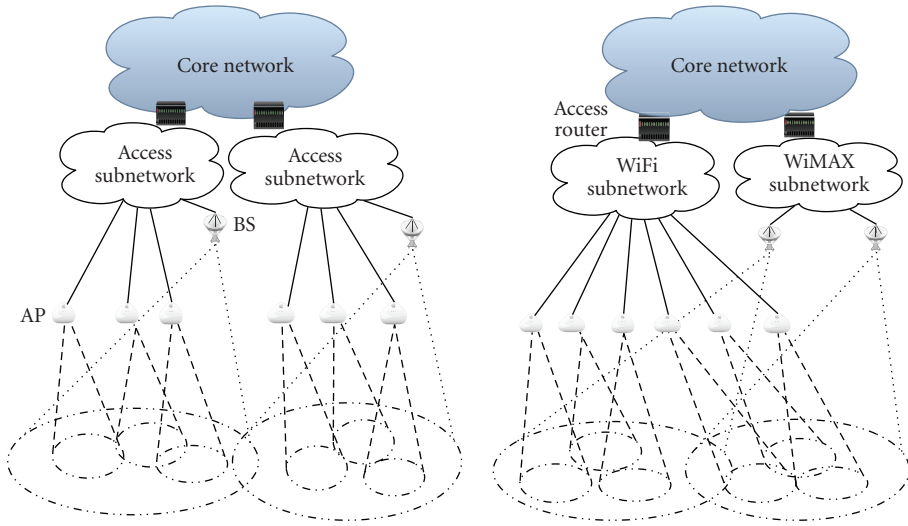
Figure 4: HO management exchanges.

interruption during a handover. The authentication process can last up to 1 s [26]. Several solutions are available to ensure reduced authentication delays during horizontal HO less than 25 milliseconds (ms) [27]. However, these solutions are not effective for a heterogeneous HO management, which will be the current architecture results to a new network entry for the target technology.

4.2.2. WiMAX. The WiMAX technology offers a last mile wireless broadband access as an alternative to cable and DSL. It defines the physical layer design and the wireless medium access mechanism and network services such as the QoS management, mobility management, user authentication, and accounting for wireless part of the network based on the IEEE 80216 standards [28, 29]. In addition, an end-to-end network specification is proposed by the WiMAX forum [30–33]. It includes the core network architecture reference models, protocols for end-to-end aspects, procedures for QoS management, and user authentication.

The reference model defines a logical modeling of the network architecture. The Access Service network (ASN) is defined as a set of network functions providing radio access to mobile stations. The Connectivity Service Network (CSN) is a set of network functions that provides IP connectivity services to Mobile Stations such as IP parameters allocation, Policy and Admission Control, and Inter-ASN mobility management. CSN includes network elements such as routers, AAA proxy/servers, and user databases. The QoS management is defined by the NWG specification [30–33] and the IEEE 802.16e-2005 standard [29]. It defines the data traffic differentiation mechanism over the wireless link and associated management functions included in the core network entities, that is, ANS-GWs and Authorization and Accounting servers.

A terminal is associated with a number of service flows characterized by QoS parameters. This information is provisioned in a subscriber management system or in a policy server, typically a AAA server. A service flow is a MAC transport service that provides unidirectional transport of packets (uplink or downlink). IEEE 802.16 specifies five Data Delivery services in order to meet the QoS requirement of multimedia applications: *Unsolicited Grant service (UGS)*, *Real-Time Polling Service (rtPS)*, *Non-Real Time Polling Service (nrtPS)*, *Extended Real-Time Variable Rate (ERT-VR) service,* and *Best Effort (BE)*. Each Data Delivery Service is associated with a predefined set of QoS-related service flow parameters. The QoS profile, which is a set resource-access authorizations and preprovisioned service flows, is downloaded from the AAA server to the ASN-GW at the network entry as a part of the authentication and authorization procedure. Service flows creation is initiated based on negotiation exchanges engaged by the terminal, the BS, and the ASN-GW.



(a) heterogeneous access subnetworks (b) homogeneous access subnetworks

Figure 5: WiFi-WiMAX network.

Security in WiMAX network is based on Key management protocol (PKM). The latter defines mutual authentication exchanges between the terminal and the network entities, that is, the BSs and the ANS-GWs. These exchanges result in the generation of a hierarchical sequence of authentication keys. Each key is related to the authentication of the terminal with a level of the access network: BS, ASN-GW, and AAA server. After the authentication, the terminal negotiates with the serving BS a cryptographic suite for each provisioned service flows.

The WiMAX network entry procedure requires, as with WiFi, several exchanges for the authentication and the establishment of provisioned service flows. The technology defines an HO management mechanism based on proactive and reactive terminal context transfers from the ASN-GW and the serving BS to target BSs while attempting to ensure minimal delay and data loss during the HO procedure. The terminal context includes authentication parameters, service flow parameters (QoS information, cryptographic information, classification rules, etc.), and PHY capabilities of the terminal. Having these information elements, a target BS will be able to associate the terminal during the HO procedure with the minimum of negotiation exchanges. However, such as the HO management mechanism defined for the WiFi, this optimization is restricted to horizontal HOs.

4.3. WiFi-WiMAX Integration

4.3.1. Network Architecture. We propose a flexible deployment schema for the network architecture. The access subnetworks may offer a *homogeneous deployment* that gathers PoAs offering the same technology: WiMAX subnetworks including Base Stations (BSs) and WiFi subnetworks including Access Points (APs). It is also possible to offer a *heterogeneous deployment* that gathers PoAs according to the wireless coverage neighborhood apart from their technologies. In all types of deployment, a mobile terminal may execute vertical HOs (BS to AP and AP to BS) and horizontal HOs (AP to AP and BS to BS). Figure 5 shows the two deployments.

4.3.2. The L2-Acc-Mgr. L2-Acc-Mgrs, associated to access subnetworks, manage the L2-HO for both vertical and horizontal HOs. They support *WiFi and WiMAX specific functions* that manage authentication and accounting exchanges with terminals during network entries. An L2-Acc-Mgr acts as an ASN-GW for the WiMAX terminals and as an AAA proxy for the WiFi terminal during the network entries. These functions allow the L2-Acc-mgr to support *layer-2 service proxy* function.

This specification defines management exchanges between L2-Acc-Mgr and PoAs (APs and BSs), the intelligence related the triggering of exchanges, and the management of context information elements. We limit the description of the neighborhood management function to the definition of *Recommended PoA lists*. The actual content is to be defined by the network operator that can define the neighborhood management function based on wireless cell load, network topology, PoA geographic neighborhood, link status, and mobility behaviors.

The translation functions define the information element values to be established during HO procedures for both vertical and horizontal HOs. This specification considers the user authentication, the QoS management and WiMAX PHY layer enhancement as the services to be managed during the L2-HO preparation procedure. In the next subsection, we detail the specification of this function.

4.3.3. Terminal Context Translation. For horizontal HOs, the translation function provides context information elements based on the ones used during actual association. The

Table 2: QoS mapping between IEEE 802.11e and IEEE 802.16e-2005 classes.

802.16e-2005 Data Delivery service	802.11e UPs	Application
UGS	6,7	Voice
ERT-VR	5	Voice with silence suppression
RT-VR	4	Video
NRT-VR	3	FTP
BE	1,2,0	Email,Web

computation is based on what is defined by each technology for internal HO optimization.

When the context establishment is executed to prepare a vertical HO (serving PoA and target PoA with different technologies), the computation of values of context information elements is less obvious than with horizontal HOs. However, we have found a similitude between the QoS and authentication management of WiMAX and WiFi. Therefore, we define a mapping between the terminal context of the WiFi and WiMAX that enables the translation function to define values for WiFi context information-elements (resp., for WiMAX context information-elements) based on values related to a WiMAX association (resp., for WiFi association).

(a) QoS Information Elements. Regarding QoS management, the traffic differentiation defined by *IEEE 802.11e parameterized QoS* mechanism and the WiMAX QoS management are very similar, particularly *Traffic Stream* and *Service Flow* concepts.

We specify an association between User Priorities used in IEEE 802.11e and IEEE 802.16e-2005 Data Delivery services. These two types of information are used to characterize in each technology the class of the traffic flow. We suggest the static association between class of services of both technologies shown in Table 2. Classes are mapped according to the key QoS requirement for each Data Delivery Service. As shown in the mapping table, more than one User Priority correspond to UGS and BE data delivery service. Therefore, when the IEEE 802.16e-2005 is the serving technology, we propose to map Service Flows with data delivery service corresponding to UGS into TSs with UP equal to 6 and those with data delivery service corresponding to BE into TSs with UP equal to 1.

In addition, we propose a mapping between QoS parameters associated to each IEEE 802.16e-2005 Data Delivery service and IEEE 802.11e QoS parameters defined in the TSPEC information element. The IEEE 802.16e-2005 defines specific QoS parameters for each Data Delivery Service. However, IEEE 802.11e defines a list of parameters used for QoS characterization that may be more extensive than needed or available for any particular instance of parameterized traffic. The specification does not define a correspondence between traffic categories (defined using UPs) and possible lists of associated parameters. To be able to ensure a mapping between QoS parameters, we propose to consider the matching defined by the IEEE 802.16e-2005 between Scheduling services and QoS parameters as a reference in the translation procedure. The parameters associated to a traffic flow depend on the traffic class associated to it in both IEEE 802.11 and IEEE 802.16e-2005. We propose a static translation procedure between QoS parameters to be used by the Translation Function. The translation process depends on the QoS information related to the current terminal association, that is, the serving technology.

(i) *Terminal associated to a IEEE 802.11 PoA*: in this case, the Parameter Translation Function translates the TSPEC list into an SF info list.

Firstly, the UP related to the TS is translated into a Data Delivery Service in accordance to mapping proposed in Table 2. The retained Data Delivery Service indicates the IEEE 802.11e QoS parameters to be determined using the translation. Secondly, the Parameter Translation Function defines values related to the Data Delivery Service parameters based on the mapping in Table 3.

(ii) *Terminal associated to IEEE 802.16 PoA*: in this case, the Parameter Translation Function translates the SF info list into a TSPEC list.

SF info includes the Data Delivery Service and related QoS parameters. The Parameter Translation Function translates the Data Delivery Service into a UP based on mapping defined in Table 2. Then, it defines which parameters to be included in the TSPEC and their values.

Table 3 presents the mapping used to compute IEEE 802.16e-2005 QoS parameters based on the IEEE 802.11e parameters.

We now discuss some translation choices and difference with mapping used in the reverse translation (i.e., from 802.16e-2005 parameters to 802.11e ones).

(a) Unsolicited Grant Interval parameter indicates the nominal interval between successive grant opportunities for UGS and ERT-VR flows. Unsolicited Polling Interval parameter indicates the same QoS characteristic for RT-VR flows. These parameters do not have an equivalent in 802.11e QoS parameters. However, the TSPEC include Maximum Service Interval and Minimum Service Interval that defines, respectively, maximum and minimum of the interval between the start of two successive transmission opportunities. Thus, we use these two parameters to define a mean value corresponding to the IEEE 802.16e-2005 parameter: (*MinimumServiceInterval* + *MaximumServiceInterval*)/2. When the current serving technology is the 802.16e-2005, we may allocate the same value to Maximum and Minimum Service Interval 802.11 parameters. This value tallies to Unsolicited Grant Interval or Unsolicited Polling Interval value depending on Data Delivery Service.

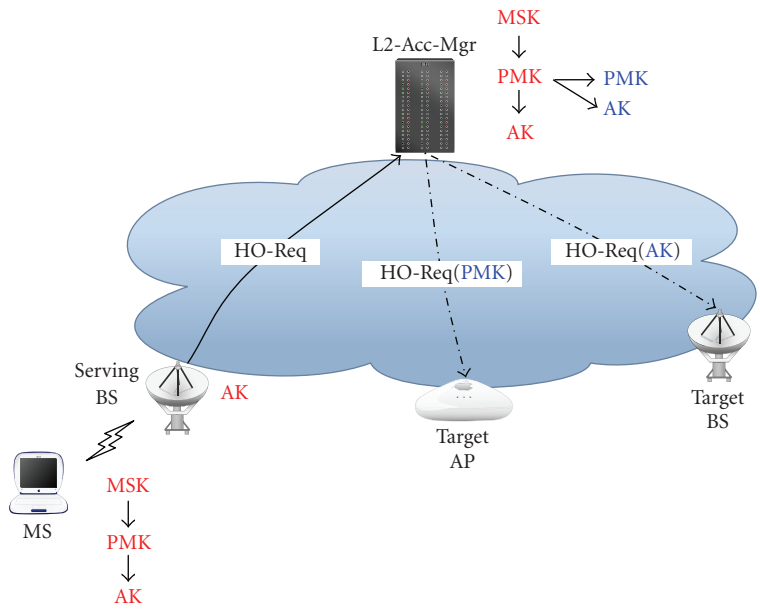


Figure 6: Proactive key distribution, Scenario 1.

(b) The correspondence between Traffic Priority and User Priority is defined only for mapping from 802.11 specification to the 802.16 one. In the reverse case, the value of the User Priority parameter is obtained based on the Data Delivery Service as previously indicated.

(c) The Tolerated Jitter parameter do not have an equivalent in 802.11e QoS specification. However, we propose to compute a corresponding value based on available parameters. The jitter value is defined as $J = \max(D) - \min(D)$ where D is the delay imposed to exchanged data packets. We have $D = D_l + D_n$, where D_l is local delay due to buffering and scheduling and D_n is the network delay due to the transmission of the packet. We suppose that D_l is negligible compared to D_n, and thus the latter equation will be $D = D_n$. Thus, $\max(D)$ corresponds to the *Delay Bound* 802.11 parameter. Additionally, $\min(D)$ can be computed based on the data rate perceived by the 802.11 station. The Parameter Translation Function can obtain a *Mean Data Rate* value based on information gathered by the L2-Acc-Mgr about mobile connectivity and cell states.

(b) Authentication Information Elements. The authentication procedures defined by the WiFi and the WiMAX are both based on negotiation exchanges that result to the generation of hierarchical sequences of authentication keys. The two key sequences are similar and have a common root key, the Master Session Key (MSK), negotiated between the AAA server, and the terminal for WiFi and WiMAX. Thus, it is possible to define a mapping between levels of two key sequences.

The WiMAX authentication procedure results to the establishment of the MSK transferred from the AAA server to the authenticator. The authenticator computes a Pairwise Master Key (PMK) and an Authorization Key (AK); it transfers the AK to the Base Station. A 3-way-handshake exchange is performed between the terminal and the BS based on the AK. The exchange results in the generation of Traffic Encryption Keys (TEK).

The IEEE 802.11i authentication results to an MSK negotiated between the terminal and the AAA server. The latter generates a PMK key, based on the identity of the serving AP, that it transfers to the AP. This key is used to perform the 4-way-handshake between the terminal and the serving AP. This exchange computes the Pairwise Transient Key (PTK) used to secure data transfer.

Conforming to the WiMAX specification, the AK is generated by the L2-Acc-Mgr, which acts as an ASN-GW, and delivered to the BS. Similarly, the 802.11 PMK is generated by the L2-Acc-Mgr (the 802.11 AAA proxy) and delivered to the AP. The 802.16 AK and the 802.11 PMK have the same functionality in authentication procedures. We consider these two keys as the starting point to define the inter technology translation for security parameters.

When the terminal is associated with a BS, it shares an 802.16 PMK with the L2-Acc-Mgr. This key is used to compute the AK that the L2-Acc-Mgr transfers to the BS. During the HO preparation procedure, the L2-Acc-Mgr uses the 802.16 PMK to generate keys for target PoAs. 802.16 AKs are generated for BSs, and 802.11 PMK are generated for APs. Figure 6 details related exchanges.

When the terminal is associated with an 802.11 AP, it shares an 802.11 PMK with the L2-Acc-Mg.During the HO preparation procedure, the L2-Acc-Mgr uses the 802.11 PMK to generate keys for target PoAs. 802.16 AKs are generated for BSs, and 802.11 PMK are generated for APs. Figure 7 details related exchanges.

Table 3: QoS mapping between IEEE 802.11e and IEEE 802.16e-2005 classes.

IEEE 802.16e-2005 parameter	IEEE 802.11e parameter	Description
Maximum Sustained Traffic Rate	Peak Data Rate	The peak information rate in bit per second
Maximum Latency	Delay Bound	The latency period starting at the arrival of a packet at the MAC till its successful transmission to the destination
Minimum reserved Traffic rate	Minimum Data Rate	The minimum data rate required by the traffic flow
Maximum Traffic Burst	Burst Size	The maximum continuous burst the system should accommodate for the traffic flow
SDU size	Nominal MSDU size	Number of bytes in a fixed size packet
Unsolicited Polling Interval	(**a**)	The maximum nominal interval between successive polling grant opportunities for the traffic flow
Unsolicited Grant Interval	(**a**)	The nominal interval between successive grant opportunities for the traffic flow
Traffic Priority	User Priority (**b**)	The priority among two IEEE 802.16e-2005 service flows identical in all QoS parameters.
Tolerated Jitter	(**c**)	The maximum delay variation (jitter) (in milliseconds)

(c) WiMAX PHY Information Elements. The WiMAX technology defines parameters related to PHY-layer capabilities of terminal. These parameters have no equivalent in the WiFi specification. Thus, we maintain a caching mechanism for PHY-layer capabilities managed by the translation function. PHY-layer capabilities of terminals are maintained during the ongoing session. When preparing an HO with target BSs, if a terminal has never been attached to a BS in previous associations, the L2-Acc-Mgr sends an HO-Req to target BSs without these parameters. Additionally, it indicates to the terminal, in the recommended Candidate PoA List, to execute proactive exchanges to negotiate these parameters with target BSs.

4.3.4. Context Establishment Procedure. The L2-HO optimization is based on the establishment of terminal contexts on target PoAs to avoid their re-negotiation and consequently reduce the HO delay. The context establishment procedure is mainly proactive. The neighborhood management function provides the *Recommended PoA List* to which the establishment is initiated. The QoS parameters, the authentication keys, and the WiMAX PHY profiles are established based on a context transfer managed by the L2-Acc-Mgr. The cryptographic suites are established based on a context transfer between the serving PoA and target PoAs (preparation of a horizontal HO) or proactive negotiation between the terminal and target PoAs (preparation of a vertical HO). The translation function computes values for the information elements to be established based on the available terminal context.

In addition to proactive establishment, the specification defines reactive establishment exchanges that may be engaged by the target PoA during the HO execution.

Figure 8 shows an example of the proactive phase of the context establishment procedure. The terminal is associated with a serving AP. The context establishment is performed with an AP and a BS. When a mobile terminal associates itself through an AP, the context establishment is started using an HO-Request, which includes QoS information elements sent by the serving AP to the L2-Acc-Mgr. The translation function builds the contexts related to PoAs in the *Recommended PoA List*. The HO management function initiates context transfer to PoAs using HO Request messages that includes terminal contexts. Based on target PoA responses, which indicates the support of terminal requirements, the HO management function builds the *PoA List* that is forwarded to the serving AP. The serving AP transfers the list to the terminal. The cryptographic suites are established, with available PoAs, using a context transfer with target APs and a proactive negotiation with the target BSs.

The previous example describes a preparation procedure performed with target PoAs in the same access network as the serving PoA. The HO messages are exchanged between PoAs, and the L2-Acc-Mgr managing the subnetwork and context messages are exchanged between involved PoAs. When a target PoA is located in an access network different from the serving PoA one, the HO management exchanges are relayed between the serving L2-Acc-Mgr and the target L2-Acc-Mgr to reach the involved entities. The serving L2-Acc-Mgr is the manager of the preparation procedure while the target L2-Acc-Mgr relays the messages between the latter entity and the target PoA. Figure 9 shows the exchange.

Regarding context transfers between PoAs and proactive negotiations between the terminal and the target PoAs, we make the choice not to execute these exchanges during the inter-subnet preparation procedure. Therefore, the preparation will be limited to centralized exchanges performed between the L2-Acc-Mgr and the PoAs. This is justified by results we have obtained in work related to HO preparation mechanisms proposed for the IEEE 802.11 networks regarding velocity support and signaling cost [34]. The evaluation has shown that exchanges performed between PoAs and particularly proactive negotiations are not adapted to inter-subnet mobility. In fact, they increase the signaling cost of the preparation procedure and reduce the HO performance in high mobility environments.

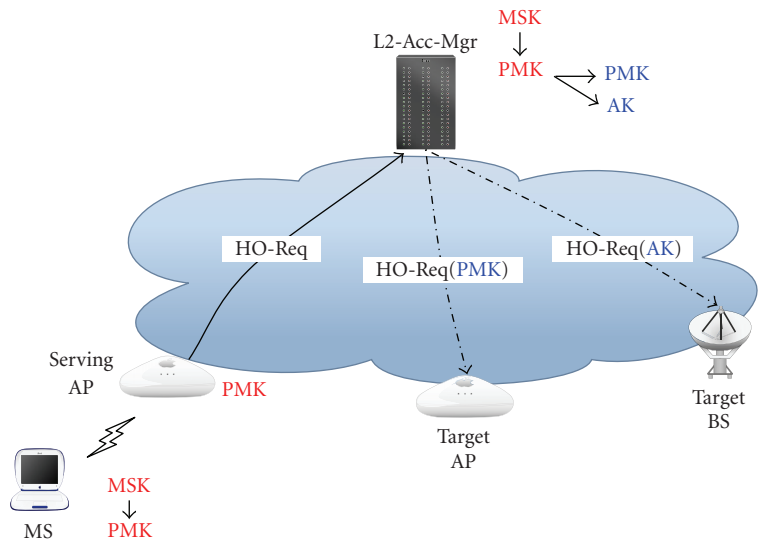


Figure 7: Proactive key distribution, Scenario 2.

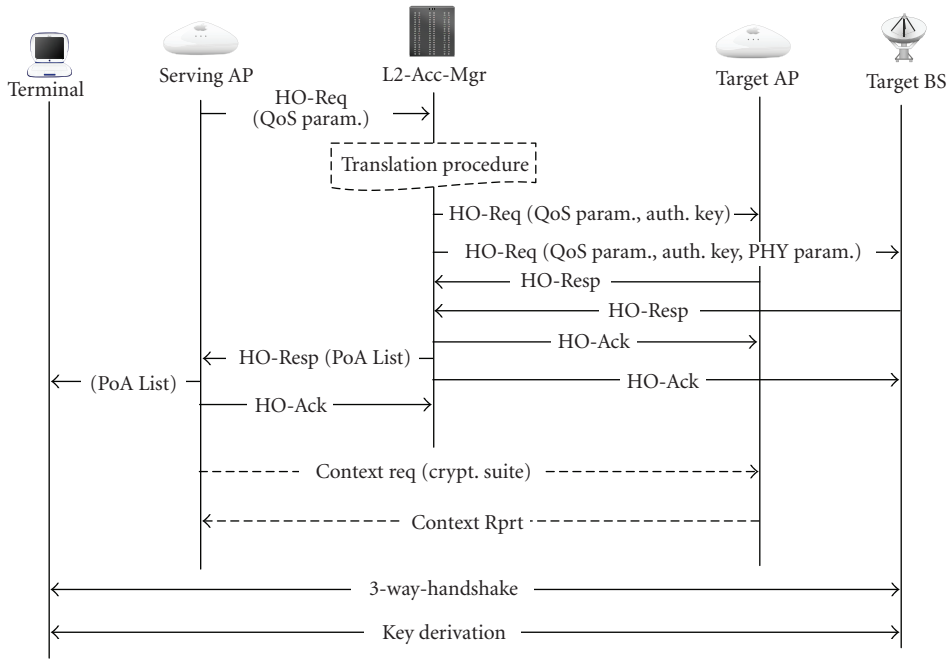


Figure 8: Example of context establishment.

4.3.5. HO Execution Optimization. The HO preparation procedure, presented in previous sections, establishes a set of context information elements and parameters in target PoAs. The exchanges engaged during the HO execution depend on the information elements that were established proactively during the HO preparation procedure or requested reactively during the HO execution. We present in the following paragraphs possible HO execution scenarios for both WiMAX and WiFi technologies. We consider optimal scenarios where target PoAs were able to acquire all context information elements.

The establishment of the terminal context results in an important optimization of the L2-HO execution procedure for both vertical and horizontal HOs. The terminal no longer needs to reauthenticate itself and to renegotiate QoS parameters and PHY profile (when the WiMAX is the target technology) during the L2-HO execution.

Figure 10 presents a regular WiFi network entry that may be executed during a first network association and an optimized reassociation procedure that may be executed during HO with an AP. In the first case, the terminal performs a regular 802.11i authentication (2, 3, 4, and 5),

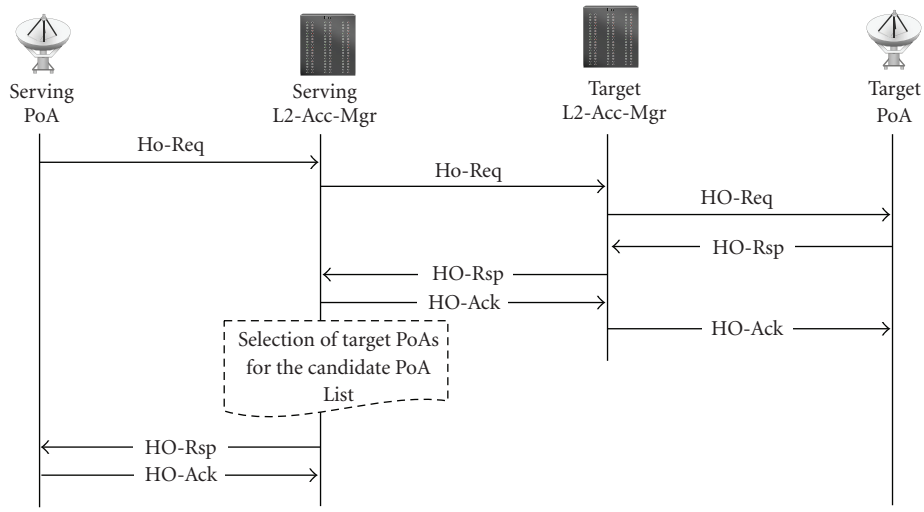


Figure 9: Inter-subnet HO preparation exchanges.

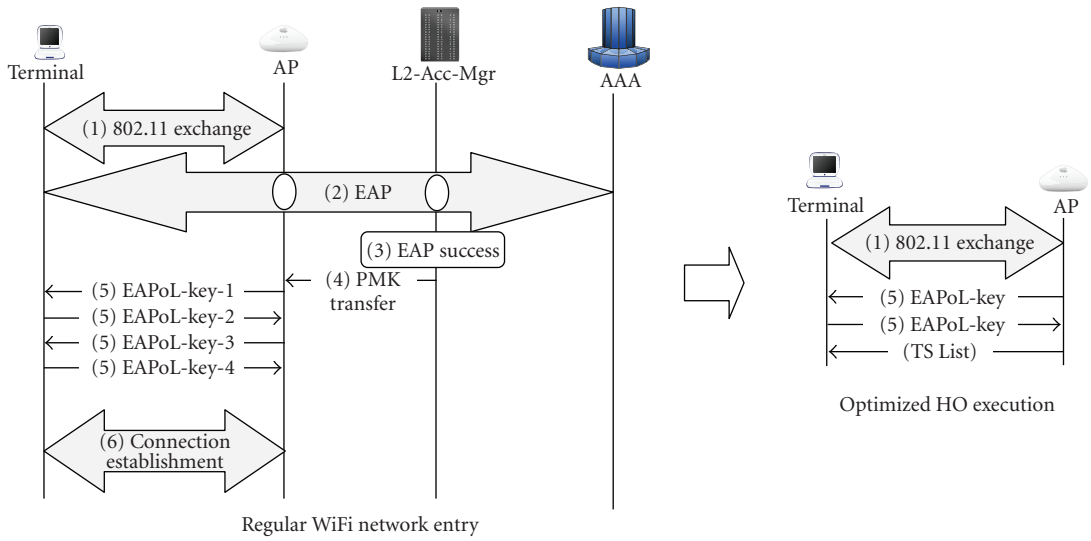


Figure 10: Association versus Re-association with a WiFi Access Point.

including exchanges with the AAA server, and the 802.11e traffic streams' establishment (6).

During a HO preparation, a target AP may acquire the Traffic Stream (TS) list and the PMK during the first phase of the procedure based on exchanges performed with the serving L2-Acc-Mgr. The target AP acquires also the PTK based on a context transfer or computes this key with a proactive negotiation performed with the AP. Therefore, in the second case of Figure 10, the terminal starts the HO execution with the legal IEEE 802.11 re-association and authentication. Over Authentication Req/Resp, the terminal and the target AP inform each other about the preestablished keys. Then, they engage a key-handshake to exchange the Group Temporal Key (GTK). If this part of the authentication exchange succeeds, the new serving AP sends to the terminal the TS List (including TSPECs), and the latter can start data exchange.

Figure 11 presents a regular WiMAX network entry that is executed during a first network association and an optimized re-association procedure that have to be executed during an HO with a BS. In the first case, the terminal performs all steps of regular WiMAX association: synchronization (1), ranging (2), basic capabilities negotiation (3), authentication (4,5, and 6), cryptographic key negotiation (7,8), and connection establishment (10,11) [29].

During handover preparation, a target BS may acquire proactively the authentication key AK, the encryption key list TEK list, the SF list, and the WiMAX PHY capabilities of the terminal. So in the second case of Figure 11, The HO execution starts with a Ranging exchange between the terminal and the target BS. The Ranging Response (RNG-Rsp) indicates the re-entry steps that are omitted thanks to the availability of terminal context information elements obtained during HO execution. Then, the target BS sends an

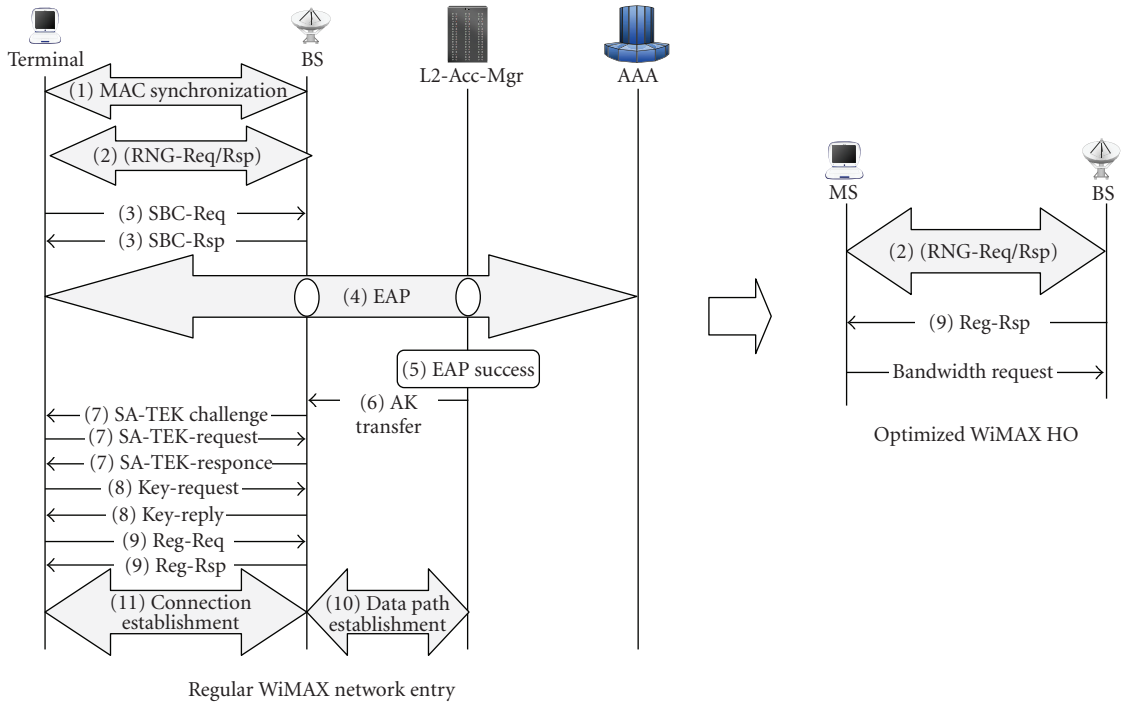


Figure 11: Association versus Re-association with a WiMAX Base station.

unsolicited Registration Response (REG-Rsp) that includes information about connections. Finally, the terminal sends a Bandwidth Request header with zero BR field to the target BS that regards this message as a confirmation of successful re-entry registration.

As shown in Figures 10 and 11, the handover execution is significantly reduced for both WiFi and WiMAX.

5. Performance Evaluation

In this section, we evaluate the performances of the L2-HO management for WiFi-WiMAX network. This evaluation requires the definition of parameters and metrics that will constitute the reference of the evaluation. The evaluation criteria will highlight both the contributions of new mechanisms and the limits of their application.

5.1. Handover Delay. The most obvious criterion that must be evaluated is the HO delay. The latter is defined as the time during which the station is not connected to any PoA. Therefore, the HO delay includes the time required to detect the need to perform a handover, to choose a target PoA, and to perform re-association exchanges.

We adopt the network simulator SimulX [35] that supports features that enable the design and the evaluation of future communication protocols like cross-layer interactions, multi-interface inter-working in terminals, and heterogeneous network environments. We have integrated to SimulX the IEEE 802.11 architecture [14] and the WiMAX architecture [36]. Both have been validated through simulation tests that result in well-known performances of both technologies. The WiFi-WiMAX architecture and the L2-HO optimization mechanism proposed in this researches have been implemented in the simulator based on the latter architectures [25].

Table 4: Handover delay.

Target technology	Opt. HO (ms)	Non-opt. HO (ms)
WiFi	24, 67	1000
WiMAX	23, 16	700

In the first scenario, we evaluate the HO delay performed when we use the L2-HO optimization mechanism. We consider a wireless network with a single access subnetwork that includes all the PoAs (two BSs and two APs). A terminal moves with a straight path to cross the wireless coverage of all PoAs of the network. We measure the delay involved by the executed L2-HOs. To show the contribution of L2-HO optimization mechanism, we can compare the inter-technology HO delay to the network entry delay of the WiFi and WiMAX technologies, which correspond to non-optimized HOs.

Table 4 lists HO delay values obtained with different types of HOs. The delay due to non-optimized HOs is evaluated to 700 ms when the WiMAX is the target technology and 1000 ms when the WiFi is the target technology. Let's note that the WiFi handover delay is larger than the WiMAX handover delay although that nonoptimization handover execution of WiMAX seems to engage even more exchanges than the WiFi handover execution (c.f. Figures 10 and 11). Actually, the detection and the search phases contribute largely to the delay induced to traffic during the handover

procedure of WiFi. However, these phases are well optimized in handover procedure of WiMAX. For example, there is no search phase at the time of HO as the serving BS sends a recommended neighbor list to terminal. As a consequence, the overall HO delay of WiFi network entry during HO is larger that of the WiMAX.

The L2-HO management mechanisms ensure a uniform execution time for both intratechnology and intertechnology HOs limited to a mean value of 24,63 ms. This is obtained thanks to the context establishment mechanism that ensures the same optimization of the HO execution regardless of the target PoA type.

In a second phase of this evaluation, we study the effect of wireless cell conditions on the performances of the L2-HO optimization performances. We consider a network topology integrating six BSs with six APs in each WiMAX cell. The PoAs are attached to two access subnetworks: a WiFi subnetwork and a WiMAX subnetwork relayed through a core network, which hosts also the AAA server. A terminal moves with a straight path and a velocity of 10 m/s. We measure the HO delay for WiFi to WiMAX and WiMAX to WiMAX handovers.

In WiFi networks, the performance of terminal exchanges depends on the cell load because of the contention-based medium access [27]. In a previous research, we were interested in the evaluation of HO performances in WiFi networks. We showed that the wireless cell load has non-negligible effects on the HO execution performances. We evaluated a management mechanism that ensures the same optimization of HO execution for WiFi terminals. Results demonstrated that such optimization ensures a limited execution time (lower than 50 ms) even with high loads.

The performance of WiMAX wireless access is not sensitive to the cell load as the medium access is managed by the BS that allows transmission opportunities to the medium modeled by transmission frame [28]. However, two parameters can have an influence on the performances of HO execution: the IEEE 802.16 frame duration and the contention-based transmission period defined for network entry.

The duration of the IEEE 802.16 frame, which is configurable, has an effect on the delay between two transmission opportunities for one terminal, which impacts on the delays for exchange between the terminals and the BS. In a previous research, we have evaluated the variation of the regular WiMAX network entry as a function of the frame duration. Results have shown that the network entry duration vary from 700 ms to 1 s with frame duration that varies from 3 ms to 12 ms.

We evaluate the effect of the frame duration of the optimized WiMAX handover. Figure 12 plots the delay due to optimized WiMAX handover as a function of the 802.16 frame duration. This curve shows that the handover delay increases when the lEEE 802.16 frame duration increases. However, even with frame duration of 12 ms the handover delay remains reasonable and does not exceed the value of 50 ms (tolerable threshold of real-time applications).

The second parameter considered for WiMAX cells is the contention-based transmission period. It is used by a terminal that starts an HO procedure or an association

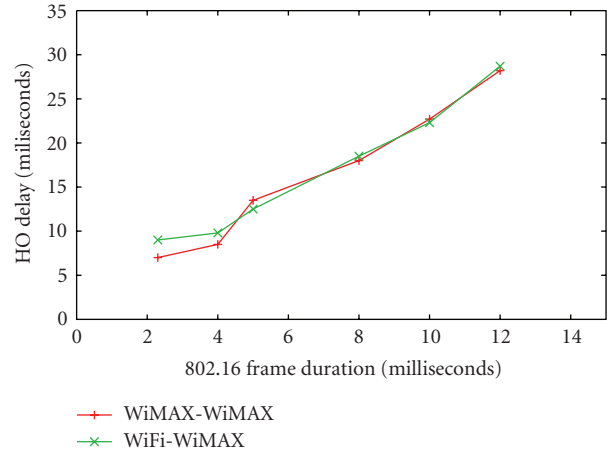


Figure 12: Effect of the 802.16 frame duration on optimized HO performances.

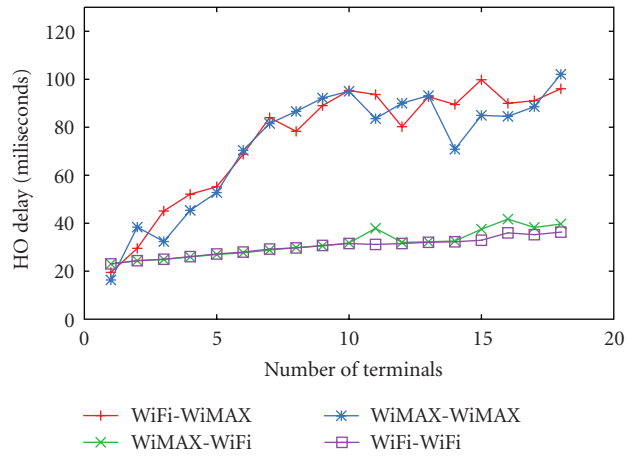


Figure 13: Effect of number of terminals on optimized HO performances.

procedure with a BS. This period has a limited duration during a single frame. The exchanges over it will be impeded by the number of terminals trying to communicate.

To evaluate the effect of the number of terminals executing a network entry on the HO delay, we define a simulation scenario that varies the number of terminals executing HOs in the same contention-based transmission period of a cell, and we measure the average of HO delays. The simulation scenario defines a set of terminal moving at the same velocity, over similar trajectories, and neighbor starting points. The network topology includes six BSs with six APs in each WiMAX cell.

Figure 13 plots the evolution of the HO delay as a function of the number of terminals. The curves show an increase of the HO execution time (WiMAX to WiMAX HOs and WiFi to WiMAX HOs) with the increase of the number of terminals. This parameter exceeds 50 ms as soon as the number of terminals that try to associate exceeds 5.

5.2. Signaling Cost. We propose to evaluate the signaling overhead of the HO management mechanism associated to the WiFi-WiMAX integration network. This evaluation aims to compare the new architecture with alternative network deployments under the same conditions.

We consider a realistic deployment of the WiMAX and WiFi technologies over a city. The WiMAX is used to offer an outdoor access while the WiFi is used to offer indoor accesses. As shown in Figure 14, the WiMAX access is offered to user over a continuous coverage. The WiFi access is offered via scattered areas over the WiMAX coverage.

We compare the performances of the integration architecture (optimized architecture) to an architecture that does not integrate an L2-Acc-Mgr (non-optimized architecture). In the latter architecture, we suppose that the HO management functions, for example, neighborhood management and context establishment, are supported by centralized network servers. In addition, we evaluate the influence of the design of access subnetworks (*homogeneous deployment* versus *heterogeneous deployment*) on the HO management signaling cost performances. Four network architectures are considered: non-optimized architecture with homogeneous deployment, non-optimized architecture with heterogeneous deployment, optimized architecture with homogeneous deployment, and optimized architecture with heterogeneous deployment.

The signaling cost of a management mechanism is the transmission cost of management messages over the network links. We define a signaling cost formula that models the signaling overhead generated by one HO. This formula takes into account the proactive exchanges with neighbor PoAs during the HO preparation and the execution exchanges with a target PoA at the time of HO as shown in (1):

$$S_{\mathrm{HO}} = S_{\mathrm{HOpreparation}} + S_{\mathrm{HOexecution}}. \quad (1)$$

We consider three types of network links: the local links (between entities in the same access subnetwork), the core network links, and the wireless links. To each link we associate a *weight* that models the cost of transmitting of one byte over this link. These weights allow to quantify link transmission costs relatively rather than define absolute values. A signaling cost formula is the sum of subformulas that are products of the messages' size into the crossed links' weight.

The sub-formula $S_{\mathrm{HOpreparation}}$ of (1) (resp., $S_{\mathrm{HOexecution}}$) is different as the HO preparation is engaged from a serving AP or a serving BS (resp., the HO execution is engaged with a target AP or a target BS).

We make use of the *VanetMobiSim* software to emulate the terminal mobility over the considered wireless deployment [37]. This software offers the list of executed HOs considering a wireless deployment and a mobility model. The combination of the signaling cost formulas and the mobility statistics allow us to evaluate the signaling cost average of the HO management over the considered deployment [25]. We assume a mix of three types of mobility model: walking users, slow cars, and fast cars. We consider one hop neighborhood definition. The Recommended PoA list integrates PoAs whose coverage areas are tangent to the serving PoA one.

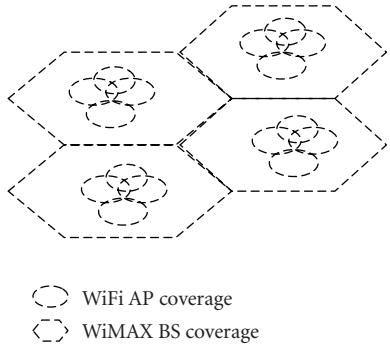


Figure 14: WiFi-WiMAX wireless coverage.

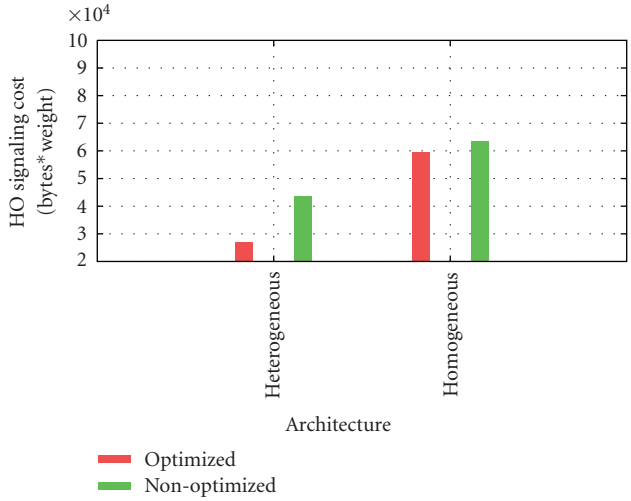


Figure 15: Basic configuration signaling cost.

In a first evaluation, we consider an arbitrary configuration with fixed value for link weight. These values indicate that the transmission cost of a management message over the core links is twice the transmission cost over the local links. The transmission cost over the wireless links is fourfold the transmission cost over local links. With this configuration, Figure 15 plots the measured HO signaling costs related to network architectures.

Both the optimized architecture and the heterogeneous deployment reduce the signaling cost of an HO. Particularly, a combination of these strategies in the same network offers a significant reduction of the HO signaling cost. The optimized architecture allows the confining of establishment exchanges at best to an access network and at worst to a connection between two L2-Acc-Mgrs. As a result, there is no more exchanges with centralized servers for HO management. On the other hand, the heterogeneous deployment allows to gather neighbor PoAs in the same access network. The use of the latter deployment with a non-optimized architecture enables to reduce inter-PoAs exchanges to the intra-access networks exchanges, which reduces significantly the HO management signaling cost. With an optimized architecture,

the heterogeneous deployment enables, as well, to confine centralized exchanges to into one access network.

In a second step, we study the effect of architecture parameters on the HO management signaling cost. We consider the core-link weight and the neighborhood definition.

Figure 16 plots the evolution of the handover signaling cost as a function of the core-link weight. Both the optimized architecture and the heterogeneous deployment reduce the effect of core link cost on the HO signaling cost. The combination of an optimized architecture and a heterogeneous deployment offers the better optimization. These results confirm that the design of a network architecture based on this combination reduces the consumption of the core network resources by HO management signaling overhead. In fact, the signaling exchanges related to a mobile terminal will be enclosed in the wireless cells and access subnetworks in its mobility areas. Thus, the proposed designs ensure the enhancement of HO performances while reducing the core network resources.

The enlargement of neighborhood definition is important to ensure a better mobility support. Indeed, a multiple-hop neighborhood should ensure a good support of fast moving terminals. However, this neighborhood definition may result to an increase of the signaling cost of HOs. To study the effect of the neighbor list size, we assume a second neighborhood definition including PoAs that are reachable within two hops. The neighbors of an AP are the APs that surround within two hops and the BS that covers the area if it is reachable by a terminal on two hops. The neighbors of a BS are the APs on its coverage zone reachable at most with two hops and the BSs in its immediate wireless neighborhood.

We compare the HO signaling costs of this neighborhood definition to those obtained with the one-hop neighborhood definition proposed in the basic network configuration. The results are shown in Figure 17. Both the optimized architecture and the heterogeneous deployment reduce the effect of the growth of the neighbor-list size on the HO signaling cost. As in the previous evaluation, the combination of these network designs offers the better results regarding HO management signaling cost. This combination allows the operator to design wireless network with better mobility support without increasing the HO management signaling overhead.

6. Interaction with Layer-3 Handover Management Mechanisms

In this study, we are interested in optimization of HO performances in heterogeneous networks. Our proposals have been limited to the management of layer-2 handovers (L2-HO). Thus, it seemed interesting to study the interaction of this framework with additional HO management mechanisms, proposed in the literature, that may be deployed in heterogeneous networks. We consider in particular the mobility management based on FMIPv6 and the Media Independent Handover (MIH) mechanism proposed by the IEEE 802.21 standard to optimize vertical HOs.

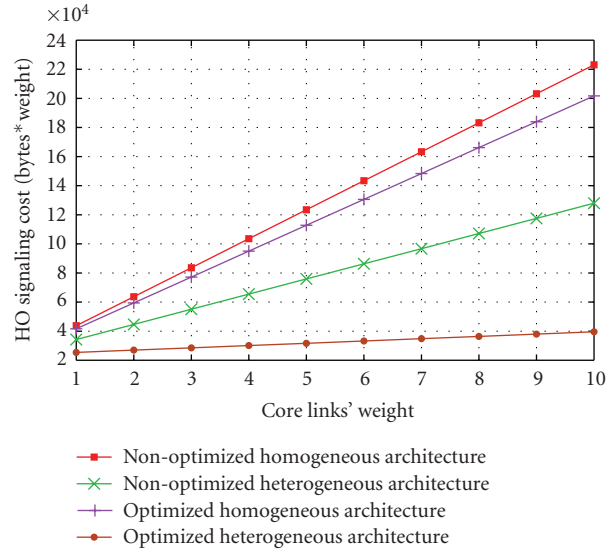


Figure 16: Core Link weight effect on HO signaling cost.

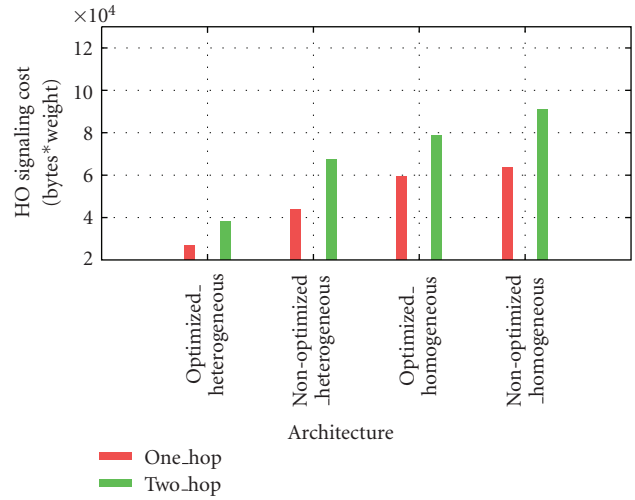


Figure 17: Neighborhood definition effect on HO signaling cost.

6.1. Collaboration with FMIP. The Fast handover for Mobile IPv6 (FMIPv6) [38] proposes an improvement to the MIPv6 that reduces the layer-3 handover latency. FMIPv6 defines a collaboration between access routers (ARs) to accelerate the acquisition of link configuration parameters and the forwarding of data traffic when a terminal executes a handover from a previous AR (PAR) to a new AR (NAR). It enables the mobile terminal to learn the IPv6 link configuration parameters (IP subnet) related to links, that it detects, before it starts effectively the HO execution. The terminal may request information, about all wireless links, to the current router. The reply can be received on the old link or on the new link (reactive HO). During the HO execution, the terminal sends a message to the NAR to inform it about the movement.

The framework, proposed in this research, enables two possible configurations regarding L3-HOs. In the first case,

access subnetworks offers heterogeneous access technologies, which allow having several technologies on the same IP subnetwork (with the same prefix). This approach avoids the need to define a relation between the L2-HO mechanisms and a possible L3-HO, since the latter is no longer necessary. With the other possible configuration, each access subnetwork offers a single access technology, that is, WiFi access subnetworks and WiMAX access subnetworks. With this architecture, a vertical HO leads to a L2-HO associated to an L3-HO. Therefore, in addition to the L2-HO management mechanism we have defined, there is a need to ensure a management of the L3-HO. This can be possible by defining an interaction between the latter mechanism and FMIPv6. The L2-HO management mechanism defines the reception of neighboring PoAs list with which the HO preparation has been performed. This list may be used, by the FMIPv6 module, to engage the management procedure defined previously with ARs attached to PoAs in the list. Upon receiving an indication of the imminent HO execution, the terminal knows its next AR; so it can prepare the configuration of its interface with new IP parameters and wait for the indication of the L2-HO handover execution success. The latter HO execution is optimized thanks to the preparation procedure of the L2-HO management mechanism. The link availability indication may also be used to trigger the preparation of following handovers.

6.2. Collaboration with the MIH. The Media Independent Handover (MIH), proposed by the IEEE 802.21 [39], defines tools to manage multiple interfaces in the same terminal. Particularly, it manages exchange of information elements between the terminal and the network to enhance the decision and search phases of the handover procedure. It also helps the preparation of the HO execution between heterogeneous technologies. For example, the MIH provides to upper layers, link-layer triggers based on reactive and predictive local link state changes and network information (load balancing information, operator preferences) that enhance the HO detection. It also supports the transfer of global network information (list of available networks, neighbor maps and higher layer network services) from network servers to the terminal to help it on the HO preparation procedure. However, the handover execution optimization is not part of the MIH functions.

The mechanisms, proposed by the MIH, are complementary to the solution we have proposed. Indeed, it is possible to make use of the MIH with our solution. Its role will be to manage exchanges between the terminal and the network entities during the HO preparation procedure and to interact with heterogeneous interfaces for the optimization of HO execution based on context information elements established proactively.

In the integration example we have proposed in IV, we use mechanisms offered by WiFi and WiMAX to perform actions related to the heterogeneous HO management. The IEEE 802.21 proposes media-dependent interfaces and primitives to be used with the WiFi and the WiMAX technologies. This will make easier the integration of the MIH to the specification we have proposed. MIH functions can be used, for example to, transfer the Recommended PoA list to the terminal during HO preparation.

7. Discussions about Heterogeneous Technology Integration

It is obvious that the mobility management in the heterogeneous wireless networks is more complex than classic wireless networks. Indeed, the more we try to optimize the HO at a low level (to ensure better performances), the more proposed solutions are dependent on the specificities of technologies. This makes difficult the optimization of the L2-HO between heterogeneous technologies, particularly when their designs are based on different principles, for example, the network accesses (connected mode or shared access mode), core network organization, and so forth. In this research, we have been able, as well, to propose a layer-2 handover optimization solution based on general and technology-agnostic framework. This framework offers mechanisms that optimize the L2-HO delay independently of the engaged mobility type (homogeneous or heterogeneous), which is a novel idea.

Another interesting point related to this framework is the ability of the proposed architecture to facilitate the extension of heterogeneous networks based on additional technologies. In fact, the location of HO management functions at L2-Acc-Mgr allows avoiding the modification of technology specific network entities, for example, PoAs, and functions, for example, authentication and accounting during these possible extensions. Modifications are restricted to the adaptation of the L2-Acc-Mgr and their functions. Let us consider the extension of the WiFi-WiMAX network, we have proposed in Section 4, based on a UMTS access. This will require, first, to define the possible associations between the QoS and security parameters in UMTS, WiFi, and WiMAX to include adequate translation rules at the *Translation function*. Second, we have to define at UMTS core network entities that manage terminal active contexts, for example, Radio Network Controllers (RNCs) or Serving GPRS Support Node (SGNC), a context exchange with L2-Acc-Mgrs. Therefore, the latter will be able to execute translation rules and to engage context establishment over WiMAX BSs and/or WiFi AP.

Based on this framework, it is possible to propose a new organization of heterogeneous networks where heterogeneous PoAs are gathered in the same access subnetwork based on the neighbor of their wireless coverage. Although, this organization remains far from current deployments' organization, it is very interesting to consider these aspects for future network deployments as we have demonstrated that such a configuration enables optimized heterogeneous HOs with very low singling overhead, which is not the case with classic network configuration. At least, network providers have to retain that with the growth of heterogeneous mobility there is a need to consider wireless coverage neighborhood between heterogeneous PoAs to ensure a reasonable signaling overhead above the core network.

Finally, we return to the fact that the use of this framework remains interesting with classic architectures and that this configuration does not have as many constraints as is believed. In fact, we can use this framework to propose the interconnection of local and restricted wireless networks, for example, a WiFi hotspot or a private WLAN, to a larger network such as a WWAN or a WMAN. The L2-Acc-Mgrs will connect the hotspot to the core network router of the WWAN that manages PoAs with coverage close to the hotspot.

8. Conclusion

In this work, we have been interested in the integration of heterogeneous wireless technologies in the same network. We have defined a technology-integration framework that defines an optimization of both horizontal and vertical HOs based on context establishment mechanisms in heterogeneous environments. We have proposed an application of this general framework to the deployment of a WiFi-WiMAX network. This application demonstrates the utility of this framework based on a practical network deployment and enables the performance of evaluation tests. The latter shows an efficient optimization of handover delays associated to a minimization of management signaling costs.

We have shown the interest for network access providers to upside the conventional network architecture by merging the backbones of heterogeneous wireless access networks. Thus, PoAs will be gathered based on the closeness of wireless coverage, which ensures an efficient optimization of HO performances with minor signaling overhead. Such network deployments are more adapted to Next Generation Wireless Networks where vertical HOs will be more frequent and trivialized.

In future work, we are interested in proposing an application of this framework for the deployment of communication systems for transport context and especially rail transport. The latter are required to operate in extremely varied environments, such as urban and suburban environments, countryside, sparsely or very low populated, tunnels, and railway stations. In addition, transport systems have very high constraints regarding transmission delays, robustness, and reliability. On the other hand, the fact that trajectories are easily predictable offers interesting perspectives for the context management, which raises the interest of adapting our solution to this particular context.

References

[1] M. Kassab, J.-M. Bonnin, and A. Belghith, "General strategies for context re-establishment in IEEE 802.11 networks," in *Proceedings of the 8th International Conference on Intelligent Transport System Telecommunications (ITST '08)*, pp. 72–77, October 2008.

[2] G. Lampropoulos, N. Passas, L. Merakos, and A. Kaloxylos, "Handover management architectures in integrated wlan/cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 7, no. 4, pp. 30–44, 2005.

[3] C. Makaya and S. Pierre, "An interworking architecture for heterogeneous ip wireless networks," in *Proceedings of the 3rd International Conference on Wireless and Mobile Communications (ICWMC '07)*, p. 16, March 2007.

[4] R. Samarasinghe, V. Friderikos, and A. Aghvami, "Analysis of Intersystem Handover: UMTS FDD & WLAN," Centre for Telecommunications Research.

[5] S.-L. Tsao and C.-C. Lin, "Design and evaluation of UMTS-WLAN interworking strategies," in *Proceedings of the 56th Vehicular Technology Conference*, vol. 2, pp. 777–781, September 2002.

[6] N. Vulic, I. Niemegeers, and S. H. De Groot, "Architectural options for the WLAN integration at the UMTS radio access level," in *Proceedings of the 59th IEEE Vehicular Technology Conference (VTC '04)*, vol. 5, pp. 3009–3013, May 2004.

[7] Y.-T. Chen, "Achieve user authentication and seamless connectivity on wifi and wimax interworked wireless city," in *IFIP International Conference on Wireless and Optical Communications Networks (WOCN '07)*, pp. 1–5, July 2007.

[8] S. Khan, S. Khan, S. A. Mahmud, and H. Al-Raweshidy, "Supplementary interworking architecture for hybrid data networks (UMTS-WiMAX)," in *International Multi-Conference on Computing in the Global Information Technology (ICCGI '06)*, August 2006.

[9] Q. Nguyen-Vuong, L. Fiat, and N. Agoulmine, "An architecture for umts-wimax interworking," in *Proceedings of the 1st International Workshop on Broadband Convergence Networks (BcN '06)*, pp. 1–10, April 2006.

[10] G. TS, "3GPP System to WLAN Interworking: System Description (Release6)," Tech. Rep., 3GPP TS, March 2004.

[11] C. Perkins, "IP Mobility Support for IPv4," IETF, August 2002.

[12] J. Loughney, M. Nakhjiri, C. Perkins, and R. Koodli, "Rfc context transfer protocol," *Internet Draft, draft-ietf-seamoby-ctp-11.txt*, February 2005.

[13] D. Forsberg, Y. Ohba, B. Patil, H. Tschofenig, and A. Yegin, "Protocol for Carrying Authentication for Network Access (PANA)," Draft IETF (Work in progress), December 2006.

[14] M. Kassab, S. Hachana, J.-M. Bonnin, and A. Belghith, "High-mobility effects on WLAN fast re-authentication efficiency," in *FTDA-DN Workshop, Held in Conjunction with Qshine*, July 2008.

[15] K. Gakhar, A. Gravey, and A. Leroy, "IROISE: a new QoS architecture for IEEE 802.16 and IEEE 802.11e interworking," in *Proceedings of the 2nd International Conference on Broadband Networks (BROADNETS '05)*, pp. 607–612, October 2005.

[16] D. Niyato and E. Hossain, "Wireless broadband access: WiMax and beyond—integration of WiMAX and WiFi: optimal pricing for bandwidth sharing," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 140–146, 2007.

[17] Z. Dai, R. Fracchia, J. Gosteau, P. Pellati, and G. Vivier, "Vertical handover criteria and algorithm in IEEE 802.11 and 802.16 hybrid networks," in *IEEE International Conference on Communications (ICC '08)*, pp. 2480–2484, May 2008.

[18] J. Nie, J. Wen, Q. Dong, and Z. Zhou, "A seamless handoff in IEEE 802.16a and IEEE 802.1 in hybrid networks," in *International Conference on Communications, Circuits and Systems*, pp. 383–387, May 2005.

[19] S.-F. Yang and J.-S. Wu, "Handoff management schemes across hybrid WiMAX™ and Wi-Fi™ networks," in *IEEE Region 10 Conference (TENCON '07)*, pp. 1–4, November 2007.

[20] T. Ali-Yahiya, K. Sethom, and G. Pujolle, "Seamless continuity of service across WLAN and WiMAN networks: challenges and performance evaluation," in *Proceedings of the 2nd IEEE/IFIP International Workshop on Broadband Convergence Networks (BcN '07)*, pp. 1–12, May 2007.

[21] "Part II: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," EEE Computer Society, Standard, 1999.

[22] WiMAX Forum, "WiMAX Forum Web page," September 2008, http://www.wimaxforum.org/.

[23] LAN/MAN Standards Committee, "IEEE 802.11i: Amendment 6: Medium Access Control (MAC) Security Enhancements," IEEE Computer Society, Standard, April 2004.

[24] LAN/MAN Standards Committee, "IEEE 802.11e Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements," IEEE Computer Society, Standard, November 2005.

[25] M. Kassab and J.-M. Bonnin, "Optimized layer-2 handover inWiFi-WiMAX networks," Research Report, Telecom Bretagne, 2009.

[26] M. Kassab, A. Belghith, J.-M. Bonnin, and S. Sassi, "Fast and secure hanfoffs for 802.11 infrastructures networks," *NetCon05 Lannion France*, november 2005.

[27] M. Kassab, A. Belghith, J.-M. Bonnin, and S. Sassi, "Fast pre-authentication based on proactive key distribution for 802.11 infrastructure networks," in *Proceedings of the 1st ACM International Workshop on Wireless Multimedia Networking and Performance Modeling (WMuNeP '05)*, pp. 46–53, October 2005.

[28] I. L. S. Committee, "Part 16: Air interface for fixed broadband wireless access systems," IEEE Computer Society, Standard, June 2004.

[29] I. L. S. Committee, "Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands," IEEE Computer Society, Standard, February 2006.

[30] N. WG, "Wimax forum network architecture stage 2: architecture tenets, reference model and reference points, part 0," WiMAX Forum, Wimax End-to-End Network Systems Architecture, August 2007.

[31] N. WG, "Wimax forum network architecture stage 3: detailed protocols and procedures," WiMAX Forum, Wimax End-to-End Network Systems Architecture, March 2007.

[32] N. WG, "Wimax forum network architecture stage 2: architecture tenets, reference model and reference points, part 1," WiMAX Forum, Wimax End-to-End Network Systems Architecture, August 2007.

[33] N. WG, "Wimax forum network architecture stage 2: architecture tenets, reference model and reference points, part 2," WiMAX Forum, Wimax End-to-End Network Systems Architecture, August 2007.

[34] M. Kassab and J.-M. Bonnin, "HO preparation based on network-entry parameter pre-establishement: a signaling cost study," Research Report, Telecom Bretagne, October 2007.

[35] N. Montavont, J. Montavont, and S. Hachana, "Wireless IPv6 simulator: SimulX," in *Proceedings of the 40th Annual Simulation Symposium, Part of the Spring Simulation Multiconference*, Norfolk, Va, USA, March 2007.

[36] M. Kassab, J.-M. Bonnin, and M. Mahdi, "WiMAX Simulation module with management architecture and signaling exchanges," in *International Workshop on Network Simulation Tools (NSTOOLS)*, October 2009.

[37] M. Fiore, "Vanetmobisim," February 2007, http://vanet.eurecom.fr/.

[38] E. R. Koodli, "Mobile IPv6 Fast Handovers," IETF, RFC 5268, June 2008.

[39] LAN/MAN Standards Committee, "IEEE Standard for Local and Metropolitan Area Networks- Part 21: Media Independent Handover," *IEEE Std 802.21-2008*, pp. c1-301, January 2009.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 423281, 28 pages
doi:10.1155/2010/423281

Review Article

WiFi and WiMAX Secure Deployments

Panagiotis Trimintzios[1] and George Georgiou[2]

[1] *Technical Competence Department, European Network and Information Security Agency (ENISA), P.O. Box 1309, GR-71001 Heraklion Crete, Greece*
[2] *Thermal Construction and Engineering Department, Public Power Corporation (PLC), Chalkokondili 30, GR-10432 Athens, Greece*

Correspondence should be addressed to Panagiotis Trimintzios, panagiotis.trimintzios@enisa.europa.eu

Received 30 September 2009; Accepted 23 December 2009

Academic Editor: Francisco Falcone

Wireless Broadband offers incredibly fast, "always on" Internet similar to ADSL and sets the user free from the fixed access areas. In order to achieve these features standardisation was achieved for Wireless LAN (WLANs) and Wireless Metropolitan Area Networks (WMANs) with the advent of IEEE802.11 and IEEE802.16 family of standards, respectively. One serious concern in the rapidly developing wireless networking market has been the security of the deployments since the information is delivered freely in the air and therefore privacy and integrity of the transmitted information, along with the user-authentication procedures, become a very important issue. In this article, we present the security characteristics for the WiFi and the WiMAX networks. We thoroughly present the security mechanisms along with a threat analysis for both IEEE 802.11 and the 802.16 as well as their amendments. We summarise in a comparative manner the security characteristics and the possible residual threats for both standards. Finally focus on the necessary actions and configurations that are needed in order to deploy WiFi and WiMAX with increased levels of security and privacy.

1. Introduction

In 1997, the initial form of the 802.11 protocol was presented [1]. Since then, various amended protocols have been added. The reason was the demand for higher data rates, different modulations and frequency transmissions, improved Quality of Service (QoS), enhanced security and authentication mechanisms. When the technology was brought to the market, there were concerns if products from different vendors could meet interoperability.

This issue was addressed with the formation of an industry consortium named Wireless Fidelity Alliance (WiFi). WiFi Alliance implemented a test suite to certify interoperability for the adopted 802.11b products. The 802.11b protocol [2], an amendment of the initial 802.11, operates in the ISM band with data rates up to 11 Mbps, in infrastructure and in ad-hoc mode for client-to-client connections.

Later on, the IEEE 802.11g was introduced and certified as a continuity and extension of the 802.11b. 802.11g operates in the same frequency range with data rates up to 54 Mbps [3], providing compatibility with 802.11b devices. The higher data rates achieved with the usage of a wider range of modulation options. Another important amendment was the IEEE 802.11i protocol [4], in which, newer and stronger security and authentication mechanisms were added in order to address security deficiencies that were presented in WiFi.

After the commercial success of the standard-based equipment and the thriving demand for broadband wireless access, the vision of networks covering larger areas and extended services was the next undertake of the IEEE. As a consequence in 2001, the 802.16 standard was introduced; initially its scope was to solve the "last mile" problem. While the 802.11 protocol offers service for few hundred meters range and only for a few users, the new IEEE 802.16 standard was designed for deploying Wireless Metropolitan Area Networks (WMAN) and thereby it can provide services to hundreds or thousands of users, in a point-to-point (PP) or point-to-multipoint (PMP) setting.

In June 2004, the standard was ratified under the title "IEEE 802.16-2004 Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Broadband Wireless Access Systems" [5]. This protocol was an amendment of the earliest version 802.16-2001

with the integration of the 802.16a-2003 and the 802.16c-2002 standards. In 2005, the IEEE introduced the 802.16e-2005 amendment and the 802.16-2004 Corrigendum [6], which provide mobility along with enhanced security and authentication mechanisms. The initial specification was for fixed users, designed to operate in the 10–66 GHz frequency range. The new modifications for fixed and nomadic users include mesh and Non-Line-Of-Sight (NLOS) by adding coverage in the 2–11 GHz range.

The inherent QoS parameters in the standard include minimum traffic rate, maximum latency and tolerated jitter, helping thus the usage of low-tolerant services such voice and streaming video. Additionally, the standard provides services to support both Asynchronous Transfer Mode (ATM) and packet services. ATM is important because of its role in telecom carrier infrastructure since it is often used to support Digital Subscriber Line (DSL) services. ATM is also widely used to support voice transmissions. The packet operation in the 802.16 standard supports the IPv4, IPv6, Ethernet, and Virtual LAN (VLAN) services.

The IEEE 802.16 currently employs the most sophisticated technology solutions in the wireless world, and correspondingly it guarantees performance in terms of covered area, bit-rate, and QoS. In order to spread the use of the 802.16 standard solutions, verify the interoperability of 802.16 devices built by different manufacturers and certify interoperable devices, an analogous to WiFi consortium of wireless device manufacturers was created named as Worldwide Interoperability for Microwave Access (WiMAX) [7]. As wireless broadband technology has become very popular, the introduction of WiMAX will increase the demand for wireless broadband access in the fixed and the mobile devices. This development makes wireless security a very serious concern.

Although the functional characteristics of the 802.11 and the 802.16 are different, they do have some similarities in their architecture structure. One of them, the basis of the protocol functionality, is the mechanism of the Wireless Medium Access Control (MAC) and the Physical Layer (PHY) specification. The similarity in the structure of the MAC and the PHY layer will derive substantial results from the comparison of the two standards.

This article is organized as follows. In Sections 2 and 2.1 we provide a thorough description of the security mechanisms for the IEEE 802.11, the 802.16 and their amendments. Section 2.2 we summarise the security overview for WiFi and WiMAX is provided. In Section 3, we analyse the residual threats for the two standards. Due to the fact that the 802.11 protocol has many years of operation, an analytical description of the already known vulnerabilities is provided. On the other hand, the security mechanisms of the IEEE 802.16 and its amendments have not been tested in actual conditions for a substantial amount of time, as it is a relatively new technology, not deployed widely to determine possible serious threats and vulnerability issues. Therefore, the IEEE 802.16 threat analysis will be based on the already registered threats from the 802.11 and any possible operational weaknesses that might come up after the scrutinized analysis of the 802.16 security mechanisms. Section 3.1 summarizes in a nutshell the possible threats for both standards along with their amendments and Section 3.2 of this article we provide guidelines for usage and deployment of infrastructure design and optimal configuration for WiFi and WiMAX. Finally, in Section 5.1 we conclude and discuss the related open research challenges and the work that should be done in the future.

2. WiFi Security Mechanisms

Every security mechanism for wireless transmission is built to provide three basic functions: (i) Authentication to verify the identity of the authorized communicating client stations; (ii) confidentiality (Privacy) to secure that the wirelessly conveyed information will remain private and protected; (iii) integrity to secure that the transmitted MPDU from a source will arrive at its destination intact, without being modified. Authentication operates at the Link Level between WiFi stations. Confidentiality and Integrity is implemented in the MAC security sublayer, just a level higher from the PHY layer.

2.1. Wired Equivalent Privacy (WEP). The first security mechanism was the Wired Equivalent Privacy or Wireless Encryption Protocol (WEP). WEP has the following functions to implement the aforementioned security functions.

2.1.1. Confidentiality (Privacy). WEP uses the RC4 encryption algorithm. RC4 is a stream cipher that operates by expanding a short key into an infinite pseudo-random key stream. The station XORs the key stream with the plaintext and produces the cipher text. The first definition was the WEP-40 due to the use of a 40-bit shared key. Many vendors increased the key size to 104 bits providing the WEP-104.

To avoid encrypting two texts with the same key-stream, an Initialization Vector (IV) is used to enhance the shared secret key and create a different key (WEP seed) for each packet. The IV field is 32 bits long and contains three subfields. The first contains the 24 bit IV, the second a 2-bit Key Identifier and the third a 6-bit Pad subfield. The 24-bit IV size gives a total of 64 or 124 bits key. The encryption-decryption task remains the same despite the key size (see Figure 1). RC4 receives the payload concatenated with the Integrity Check Value (ICV) (Analysis for the WEP-ICV follows in "WEP Integrity" session) at the end, and encrypts it with the 64 or the 124 bit key described earlier. At its destination the message firstly gets decrypted. The receiver with the shared key that it possesses and with the IV from the received MPDU will decrypt the encrypted payload and ICV.

2.1.2. Integrity. To ensure the integrity of the MPDU data, WEP uses the Integrity Check Value (ICV) mechanism. ICV implements a 32 bit Cyclic Redundancy Check (CRC-32). For each transmitted MPDU payload, the CRC checksum is computed and concatenated at the end of the MPDU. Both the payload and the ICV are encrypted with the RC4 cipher. At its destination the message is decrypted and the CRC

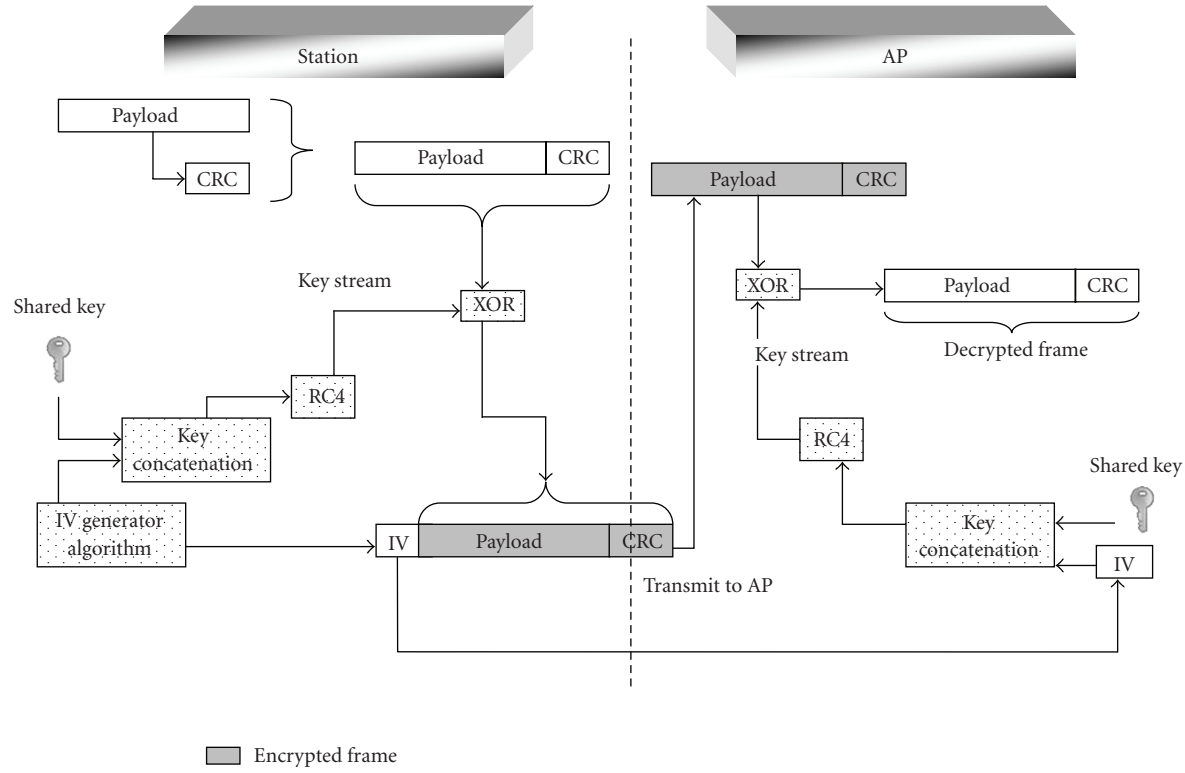


Figure 1: WEP confidentiality and integrity procedure.

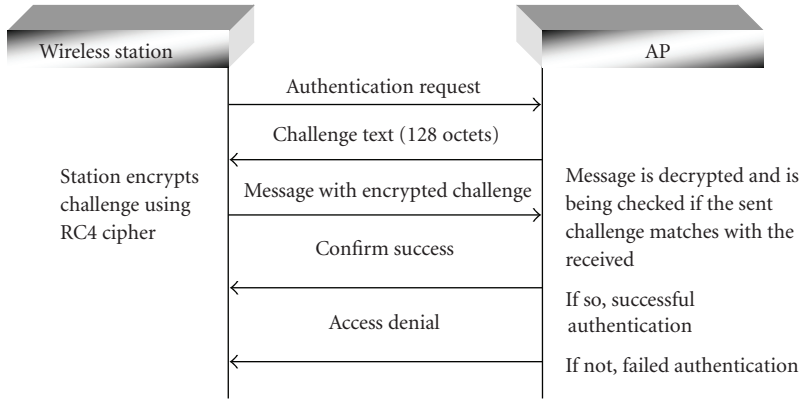


Figure 2: 4-way message authentication.

of the arrived payload is computed. If the CRC, which was produced by the source and it was sent with the message, is the same with the recomputed CRC, the message is valid and is forwarded to the Link Layer; otherwise, the message indicates integrity violation and it is discarded.

2.1.3. Authentication. WEP has two types of authentication: Open and Shared key. Open authentication actually is a non-authentication procedure since the AP accepts every station without identity verification. Thus, the station in a two-message exchange with the AP provides its identity and the request to authenticate. The AP responds with a message confirming successful authentication.

Shared key authentication (see Figure 2) requires the knowledge of a secret key to join the network. The key knowledge implies that the station is a trustful entity, and therefore authorized. The way that the key is obtained from a client station is not an issue for WEP. Another secure way must be implemented to ensure that only trusted entities will have this key. If the station possesses the key, it begins a four-way message exchange to achieve authentication. The first message from a station declares its MAC address and the authentication request. AP replies with a generated string, fixed at 128 octets, as a challenge text. The third message from station will send this challenge back to AP encrypted with an RC4 encryption, along with the ICV. The AP de-encapsulates

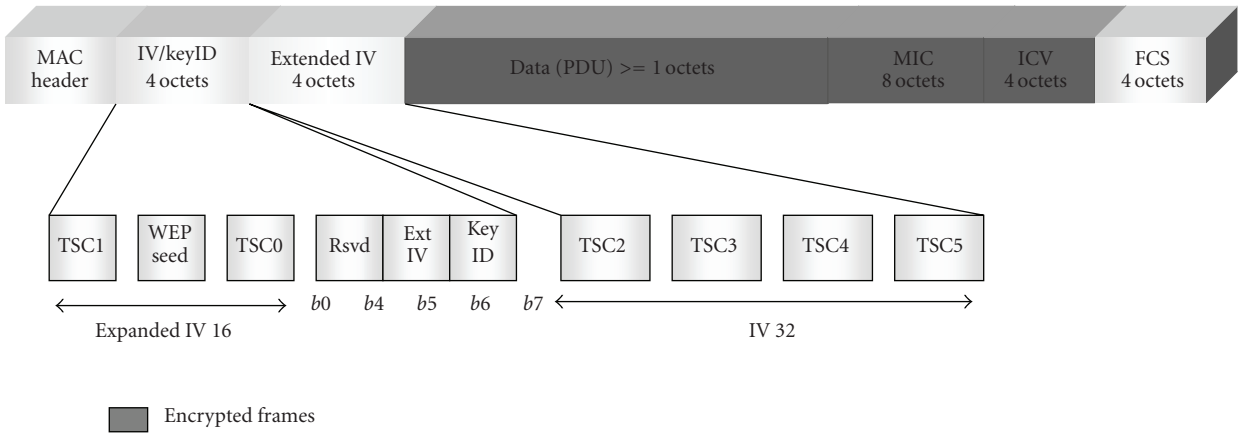


Figure 3: WPA MPDU Format.

the encrypted frame, checks the decrypted ICV, and if it is successful, the AP compares the received decrypted challenge text with the 128-byte message that was sent from it with the second message. If the two texts are the same, AP sends the last message for successful authentication. In any other case where ICV does not match or the challenge comparison is different, the AP notifies for unsuccessful authentication and rejects the station.

2.2. WiFi Protected Access (WPA). It was proved that WEP does not provide adequate security. Some of the WEP weaknesses are the following

(i) RC4 has a weak key schedule [8].

(ii) The cryptographic key and the IV are short and cannot be automatically and frequently updated.

(iii) CRC-32 is not capable of providing integrity as linear codes are susceptible to attacks on data integrity.

For the aforementioned reasons WiFi introduced the WiFi Protected Access (WPA) to enhance WEP. WPA is a part of the 802.11i standard, and it is designed to allow legacy equipment with WEP security to upgrade their firmware. WPA uses the Temporal Key Integrity Protocol (TKIP) for confidentiality and integrity while for authentication it additionally uses the 802.1X authentication protocol mechanism.

2.2.1. TKIP Confidentiality. TKIP like WEP uses the RC4 cipher for encryption-decryption. To reinforce security, TKIP doubles the IV field to 48 bits. This 48-bit field is used as a per-MPDU TKIP Sequence Counter (TSC), to create a packet sequence during transmission. If the receiver detects that a MPDU does not follow the increasing reception sequence, it drops the packet. This mechanism enhances security to replay attacks. The key mixing function is more complicated and it strengthens encryption. It generates a unique encryption key for each MPDU frame by combining the Temporal Key (TK), the Transmit Address (TA), and the TSC for the WEP seed. The WEP seed, which produced from the aforementioned parameters, operates just like the WEP IV, and with the RC4 key it creates the key stream. The encrypted parts of the MPDU are the payload, the MIC (analysis of MIC follows in TKIP Integrity) and the ICV (see Figure 3).

When the message arrives at its destination, the TSC number is checked to verify that the packet follows the increasing reception sequence. If so, the key forms the RC4 key-stream and decrypts the encrypted parts. The next step is the ICV check; if it is successful, the WPA integrity check follows.

2.2.2. TKIP Integrity. TKIP uses the Message Integrity Code (MIC) called "Michael". MIC enhances security against forgery attacks compared to the ICV usage in WEP. This time MIC is applied to MSDUs, and the MIC comparison is implemented in the MSDU-level as well. The reason is the increase of the implementation flexibility with re-existing WEP hardware. Michael with a 64-bit key is implemented on the MSDU Sender and Destination Address (SA, DA), the MSDU Priority, and the MSDU payload. MIC is 64-bit long and it is placed at the end of the MSDU payload. Knowing that a MSDU could be partitioned into more than one MPDU, the integrity check for each MPDU takes place with ICV. Then, with the concatenation of all the MSDU parts, each MSDU is checked with Michael. If the comparison of the decrypted MIC from the arrived MPDU, and the MIC which is created from the receiver, are the same, the message is valid. If not, the MSDU is discarded and measures are taking place.

2.2.3. Authentication. WPA uses the authentication methods described in WEP. Additionally, the 802.11i standard introduces the 802.1X authentication mechanism which is implemented when the WPA suite is used. A thorough analysis of the 801.1X authentication along with the Extensible Authentication Protocol (EAP) requires firstly the description of the Confidentiality-Integrity mechanisms of WPA2. Thus, the 802.1 X/EAP authentication mechanisms will be described in then WPA2 entity.

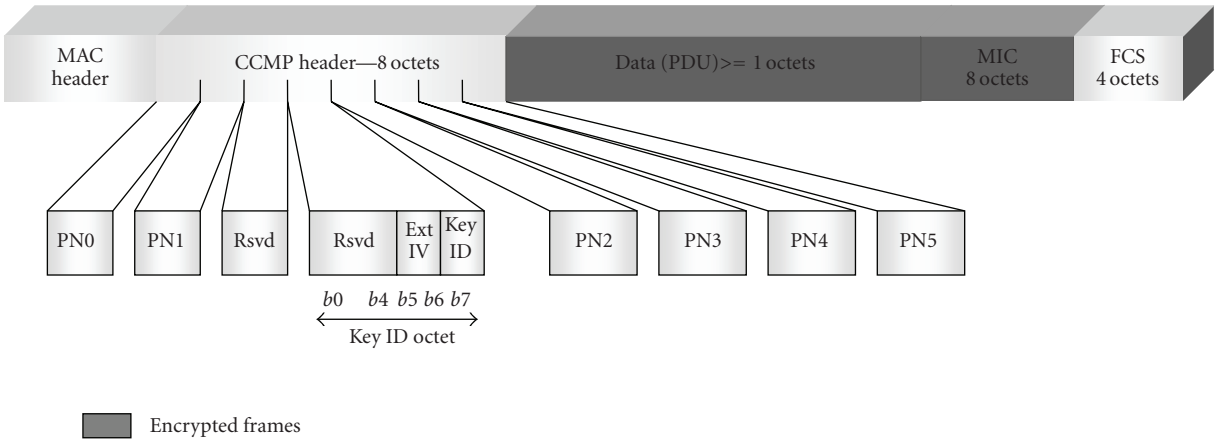


Figure 4: WPA2 MPDU Format.

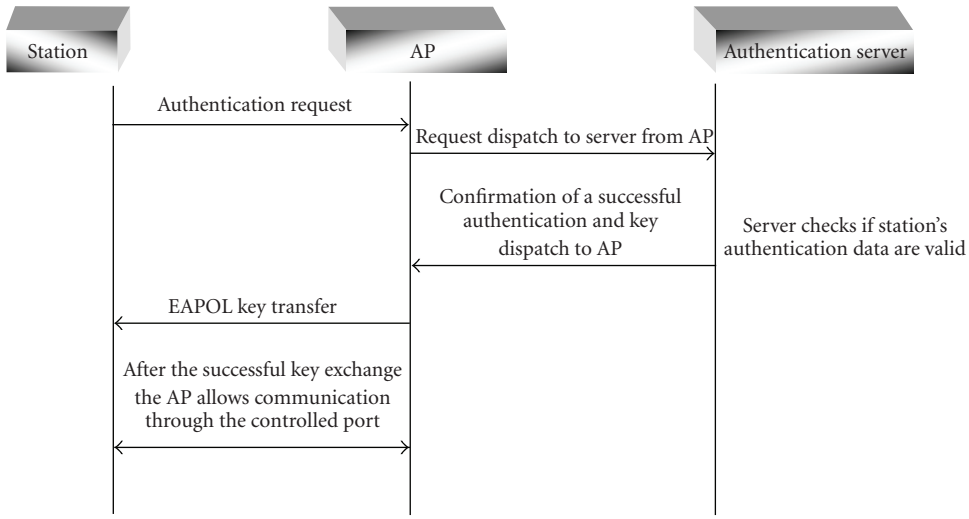


Figure 5: WPA2 Authentication procedure.

2.3. WPA2. The WPA2 name was given for the IEEE 802.11i from the WiFi Alliance. It was designed to provide stronger security with new mechanisms and hardware devices without the WEP bindings. Attention was given so that WPA devices could be associated with WPA2 access points. The security in 802.11i defines the Robust Security Network Association (RSNA), which is the indicator of the modern secured wireless communication implementation in WiFi, and separates security into two important modes: the pre-RSNA with WEP and WPA and to RSNA with WPA2 as described in this section.

2.3.1. Confidentiality. WPA2 uses the Counter-Mode/Cipher Block Chaining (CBC)-MAC Protocol (CCMP) for confidentiality as well as integrity. For data confidentiality CCMP uses AES in counter mode with 128 bit key and 128 bit block size. The encrypted parts of the MPDU are the payload and the MIC field (see Figure 4).

2.3.2. Integrity. CCM-MAC operations expand the original MPDU size by 16 octets—8 octets for the CCMP Header field and 8 octets for the MIC field. CCM requires a fresh temporal key for every session and a unique nonce value for each frame, protected by a given temporal key. For this purpose, a 48-bit packet number is used. CCM does not use the WEP ICV anymore. Leaving aside the integrity protection of the MPDU, CCM protects some Additional Authentication Data (AAD). The AAD is constructed from the MPDU header and it includes subfields from MAC frame control, addresses from source and destination fields, Sequence Control (SC), QoS control field, and therefore provides enhanced integrity protection.

2.3.3. Authentication. For authentication WPA2 provides the strong 802.1X method, which transmits key information between authenticator and supplicant. IEEE 802.1X has three main entities: The Supplicant (WS), the Authenticator (AP) and the Authentication server. The authenticator does not do the authentication; the Authentication server does this task through the authenticator. Between the supplicant and the authenticator the 802.1X protocol is implemented; between the authenticator and the authentication server the protocol

is not defined. Nevertheless, RADIUS is typically used. The EAP method used (de facto the EAP-TLS is used [9]) by IEEE 802.1X will support mutual authentication, as the station needs assurance that the AP is a legitimate AP.

The initial traffic for authentication (see Figure 5) takes place between the supplicant and the authentication server through the uncontrolled port. Once the authentication server authenticates the supplicant, it informs the authenticator for the successful authentication and it passes keying material to the authenticator. Key material exchange between the supplicant and the authenticator is implemented with the Extensible Authentication Protocol over LANs (EAPOL). If all exchanges are successful the Authenticator allows traffic through the controlled port.

2.3.4. Key Derivation and Management. Due to the fact that the 802.11i has more than one confidentiality protocols, the AP uses a ciphersuite to notify for all the data-confidentiality protocols allowed to be used (e.g., CCMP or TKIP). The client then chooses the parameters and it sends the choices back to the AP. The chosen parameters must match the available options from the list; if not, the AP will deny the association by sending a proper message. Right after the cipher suite is chosen, the key exchange is taking place. A key hierarchy is implemented to create keys for the EAPOL handshaking and the WPA2 security mechanisms. There are two key hierarchies in the 802.11i standard.

(i) Pairwise Key Hierarchy for Unicast Traffic Protection. The first key of the hierarchy is the 256 bit Pairwise Master Key (PMK). The PMK derivation depends on the authentication method used. If the 802.1X method is used, the PMK is derived from server and the first 256 bits of the Authentication, Authorization, and Accounting (AAA) key. If a preshared key is used, the password is used to create the PMK. The Pairwise key hierarchy generates the Pairwise Transient Key (PTK) from PMK. Some of the parameters are the source and the transmit address, plus, nonce from the client and the authenticator. From PTK three keys are derived. (i) The 128 bit EAPOL Key Confirmation Key (KCK), which is used for data origin authenticity in the authentication procedure that follows with HMAC-MD5, or SHA-1 algorithm. (ii) The 128 bit EAPOL Key Encryption Key (KEK), which provides traffic key confidentiality during authentication handshaking with RC4, or AES with Key Wrap. (iii) The 256 bit for TKIP or the 128 bit Temporal Key (TK) for AES-CCMP; it is used for WPA2 confidentiality.

(ii) Group Key Hierarchy for Multicast and Broadcast Traffic Protection. The first key created is the Group Master Key (GMK), which is a random number, which AP can periodically reinitialize it. The key which is derived from GMK is created with a pseudorandom function with parameters from GMK, the authenticator MAC address and a nonce from the authenticator, called Group Temporal Key (GTK). Its length is 256 bit with TKIP, and 128 bit for CCMP. The temporal key derived from GTK is 256 bit with TKIP, and 128 bit for CCMP and it is used for confidentiality.

Two are the EAPOL-key exchanges in the 802.11i standard: the 4-way and the group handshake.

The supplicant and the authenticator use this handshake to confirm the existence of the PMK, verify the selection of the cipher suite, and derive a fresh Pairwise Transient Key (PTK) for the following data session [10]. The 4-way handshake is comprised of 4 messages between the supplicant and the authenticator [11] (see Figure 6).

(i) *Message* 1. The authenticator sends a nonce (ANonce) to supplicant.

(ii) *Message* 2. The Supplicant creates its own nonce (SNonce) and sends it to authenticator. With ANonce and SNonce available, the supplicant calculates the PTK. The supplicant also sends the security parameters that it used during association, and the message is authenticated and verified with KCK from authenticator.

(iii) *Message* 3. The authenticator sends the GTK encrypted with KEK and the security parameters that sent out with its beacons. The message then is authenticated with KCK from supplicant to verify that the information sent from authenticator is valid.

(iv) *Message* 4. With this message, PTKs are ready to be used from WPA2 confidentiality protocol.

With the Group key handshake, a 4-way handshake precedes this procedure and includes the GTK conveyance in Message 3. The group key handshake updates the GTK.

(i) *Message* 1. The authenticator sends to the supplicant the GTK encrypted using the KEK and the message is subject to an authentication check.

(ii) *Message* 2. With this message, the group temporal keys (GTKs) are ready to be used from the WPA2 confidentiality protocol.

When clients roam between access points the result is a decrease in system performance as the load to authentication server is increased. A convenient way of the WPA2 to effectively resolve this issue is the key caching. With key caching the client station and the access point retain the security association when the client station roams to another access point. When a client returns to an access point, it sends the key name in the association request from AP. The client can send more than one key name in the association request. If the access point sends a success in the association response, then the client and access point proceed directly to the 4-way handshake.

After the thorough analysis of the WPA2, it must be stressed that many modern hardware devices use AES-CCMP in the WPA security, besides the TKIP option, combined with shared-key authentication, instead of the 802.1X authentication that WPA2 uses. This case resembles

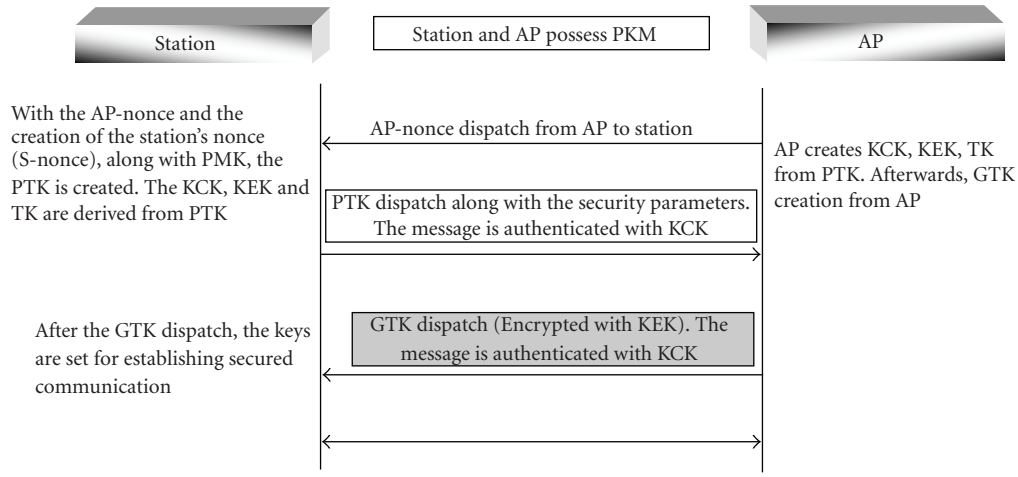


Figure 6: EAPOL key material exchange.

with WPA2 security and it should be referred as such, for the following two reasons.

(1) Although WPA is a part of the 802.11i standard, it is designed to allow legacy equipment with WEP security to upgrade their firmware.

(2) The AES-CCMP implementation in the 802.11i standard defines the Robust Security Network Association (RSNA), and indicates the modern secured wireless communication implementation in WiFi.

3. WiMAX Security Mechanisms

Security in 802.16/e was thoroughly designed as an important part of the standard architecture due to the additional possible weaknesses that wireless communication endures, especially now where the specific network deployment is to cover much larger areas. The security protocol is applied in the privacy sublayer which is positioned at the bottom of the MAC layer, and it provides mechanisms to ensure confidentiality, integrity and client authentication with the implementation of a Key Management Protocol (PKM). PKM provides also secure key distribution between BS and SS. The security information set (keys and cryptographic suites) between BS and SS is defined with the implementation of the Security Association (SA). The information included in a SA varies according to the suite it is used. The SA maintains the security state relevant to a connection [12]. SA is identified using a 16-bit SA identifier (SAID). There are three SA types.

(i) *Primary SA.* Each SS entering the network establishes an exclusive Primary SA with its BS. SS's SAID will be equal to the basic Connection ID (CID). The task of the Primary SA is to map the Secondary Management Connection.

(ii) *Static SA.* Static SAs are provisioned from the BS and they are created during the initialization process of a SS. For the basic unicast service a Static SA is created. If a SS has subscribed to additional services, additional SAs are created respectively. Static SAs can be shared by multiple SSs (multicasting).

(iii) *Dynamic SA.* A Dynamic SA is created and terminated on the fly, in response to the initiation and termination of specific service flows. Like Static SAs, Dynamic SAs can be shared by multiple SSs.

Primary and Basic Management connections are not mapped to a SA, while all transport connections are mapped to an existing SA. The BS ensures that each SS has access only to authorized SAs. Key synchronization between SS and BS is regulated from PKM.

3.1. Security Mechanisms in 802.16. The PKM protocol is used by the SS for authentication, traffic key material derivation by the BS, periodic reauthorization, and key refresh.

3.1.1. Authentication. The SS authentication is controlled from the Authorization Finite State Machine (FSM) (see Figure 13). The state machine consists of six stages (Start, Authorize wait, Authorized, Reauthorize Wait, Authorize Reject Wait and Silent), and eight distinct events (Communication Established, Timeout, Authorization Grace Timeout, Reauthorize, Authorization Reply, Authorization Invalid, Permanent Authorization Reject, Authorization Reject). In the authentication procedure the BS handles the following tasks.

(i) Authenticates the identity of a SS,

(ii) Assigns to the authenticated SS the SAIDs and the properties of Primary, and Static SAs key information,

(iii) Provides to the authenticated SS the shared secret, a 160-bit Authorization Key (AK) to initiate the following key management process.

The authorization process (see Figure 7) begins with the Authentication Information message from SS to BS.

The message contains the X.509 certificate which is bound with SS's MAC address. The certificate is issued by the manufacturer or an external authority for the SS. The X.509 authentication service is part of the X.500 series of recommendations that define a directory service. The directory is, in effect, a server or distributed set of servers that maintain a database of information about users. The core of X.509 is the public key cryptography and the digital signatures, and since the standard does not dictate a specific algorithm, RSA (asymmetric cryptography) is recommended [13]. The scheme is complete with the existence of a Certificate Authority (CA). CA issues certificates and binds each entity with a private-public key pair [14]. It is imperative that both parties entrust the CA. In 802.16 authentication, the issuer is the manufacturer or another trusted entity.

The X.509 v.3 for the 802.16 standard contains the following information:

(i) version of the X.509 certificate,

(ii) the unique Certificate serial number which the CA issues,

(iii) certificate signature. Public Key Cryptography Standard (PKCS) #1 with RSA cipher and SHA-1 hashing algorithm,

(iv) certificate (CA) issuer,

(v) certificate validity period,

(vi) certificate subject, which indentifies the entity whose public key is certified,

(vii) subject's public key, which provides the certificate holder's public key, identifies how the public key is used, and it is restricted to RSA encryption. The key size is at least 1024 bit and 2048 bit maximum,

(viii) the certificate issuer unique ID; Optional field to allow reuse of issuer name over time,

(ix) the certificate subject unique ID; Optional field to allow reuse of subject name over time,

(x) certificate extensions,

(xi) signature algorithm (PKCS#1),

(xii) signature value which is the digital signature of the Abstract Syntax Notation 1 Distinguished Encoding Rules (ASN.1 DER) encoding of the rest of the certificate.

The first message that SS sends is informative and it provides a mechanism for the BS to obtain information for the certificate of the SS. However, the BS may choose to ignore it. In the second message (Authorization Request) that is sent right after the first one, the SS requests authorization. The message includes (i) the X.509 certificate, (ii) the list of the cryptographic suite identifiers, each implementing a pair of packet data encryption and authentication algorithms that SS supports, (iii) the SS's Basic CID, which is the first static CID that BS assigns to SS during initial ranging. As mentioned earlier, the primary SAID is equal to the Basic CID.

When the BS receives the message, it authorizes the SS via the X.509 certificate, it checks for basic unicast services and other possible additional services the SS has subscribed for, and finally, it determines the cryptographic suite from the SS's list of the second message. Then, with a random or pseudo-random function, the BS generates the AK and encrypts it with the SS's public key. The encrypted AK is sent from the BS in an Authorization Reply message along with:

(i) A 4-bit key sequence number that distinguishes successive generations of AKs.

(ii) The SAIDs of the single primary and static SAs the SS is authorized to obtain key material for. The authorization reply does not identify any Dynamic SAs.

When the SS receives the message, it decrypts the AK with its private key, reads the defined cipher suite and the SAIDs, and then proceeds to key exchange with the BS. The AK remains active until it expires according to the predefined lifetime set by the BS. The SS periodically refreshes the AK by issuing authorization requests. The BS is able to support two active AKs simultaneously for each SS. Those keys must have overlapping times. Additionally, BS is always ready to send an AK to a SS upon request. The AK transition period begins when the BS receives an authorization message from a SS and the BS has a single active AK for that SS. Right after the BS receives the message, it activates the second AK which has a sequence number increased by one from the older AK, and it sends it to the SS. The lifetime of the second AK is the remaining lifetime of the older AK, plus the predefined AK lifetime. The lifetime ranges from one day to 70 days, with a default value of 7 days. If the SS does not reauthorize itself before the expiration of the current AK key, the BS does not create the sequentially next AK and considers the SS unauthorized.

3.1.2. Key Derivation and Management. With the AK delivered to SS, a key derivation will proceed to create the necessary traffic key material to implement the security mechanisms. From AK three keys will be derived.

(i) The Key Encryption Key (KEK). KEK is responsible for the encryption of the Temporal Encryption Key (TEK), that BS sends to each SS. TEKs are used for the MPDU encryption to ensure confidentiality.

(ii) The Downlink Hash function-based Message Authentication Code (HMAC_KEY_D). For the BS, the HMAC_KEY_D is used to calculate the HMAC digest for some of the management messages that it sends to SS, while for the SS it is used to verify the HMAC-Digest from the aforementioned received messages.

(iii) The Uplink Hash function-based Authentication Code (HMAC_KEY_U). For the SS, the HMAC_KEY_U is used to calculate the HMAC-Digest for some management messages that it sends to the BS, while the BS uses it to verify the HMAC-Digest of the management messages sent from the SS.

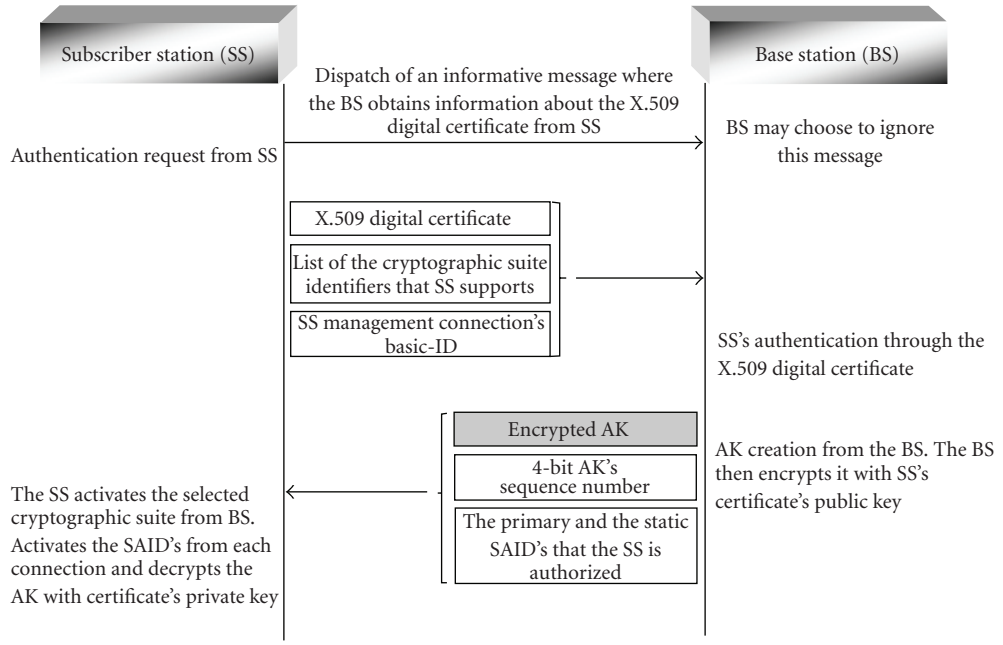


Figure 7: 802.16 Authentication Process.

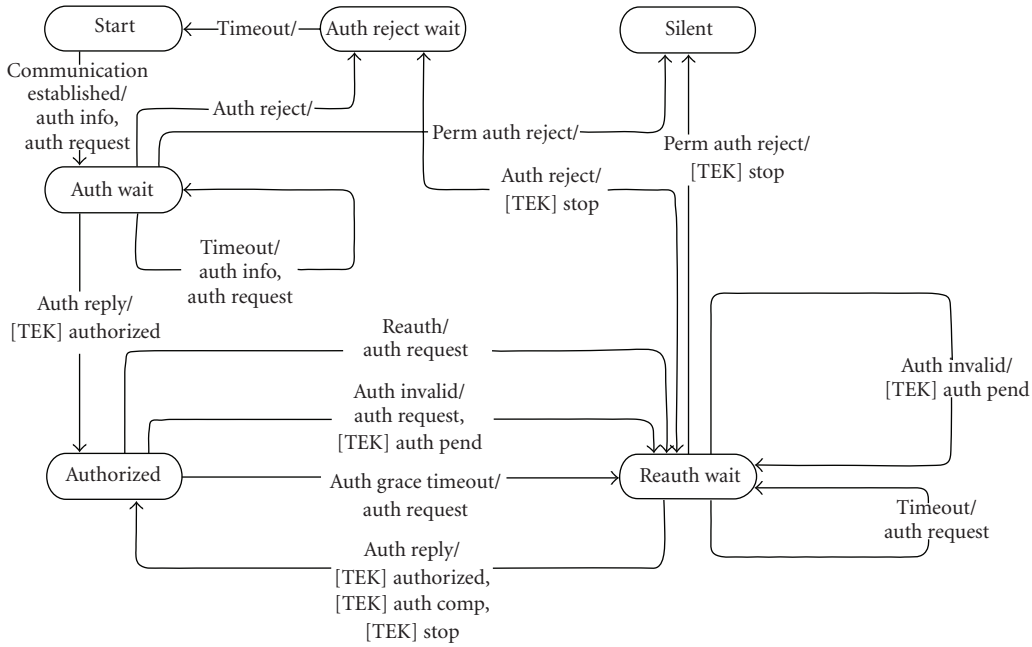


Figure 8: Authorization Finite State Machine Flow Diagram.

The BS is responsible to keep the keying information for every SS that joins the network. After key derivation the SS starts a separate TEK state machine for each of the SAIDs (the single primary and any static SA that the BS has assigned to SS). The TEK state machine (see Figure 8) consists of six stages (Start, Operational Wait, Operational Reauthorize Wait, Operational, Rekey Wait, Rekey, Reauthorize Wait), and nine events (Stop, Authorized, Authorization Pending, Authorization Complete, TEK Invalid, Timeout, TEK Refresh Timeout, Key Reply, Key Reject). Its task is to manage key material associated with the respective SAID. Each TEK state machine operates with a key request scheduling algorithm to refresh key material for their respective SAID. The BS always keeps two sets of active TEKs along with their respective 64-bit IV for each SAID. For TEK and IV generation, the BS uses a random or a pseudorandom function. The lifetime for each TEK is between 30 minutes to 7 days, with the default value set to 12 hours. The two TEKs

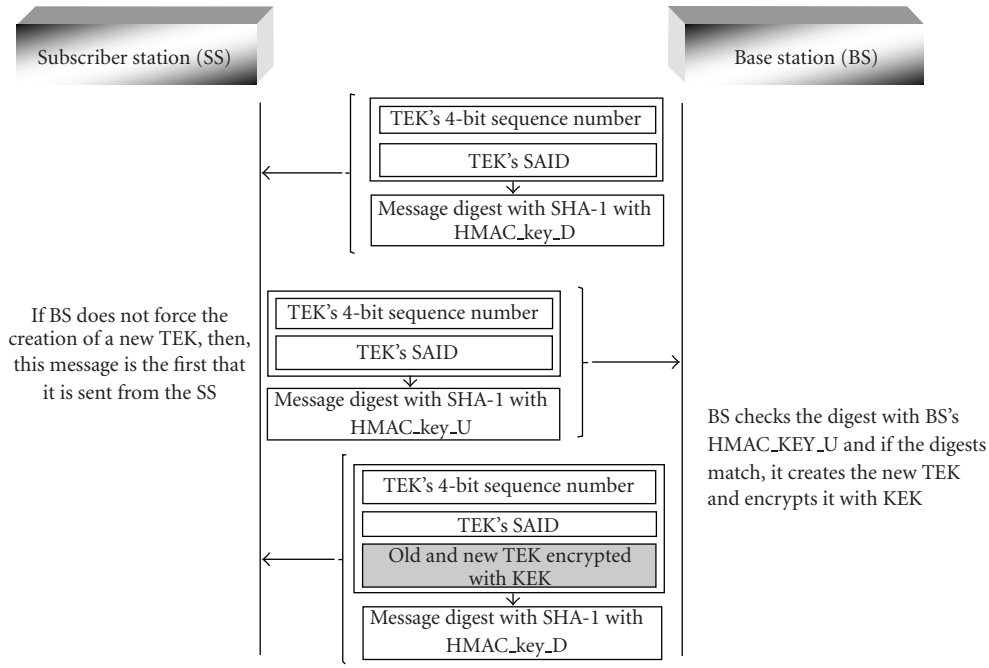


Figure 9: SA-TEK 3-way handshake.

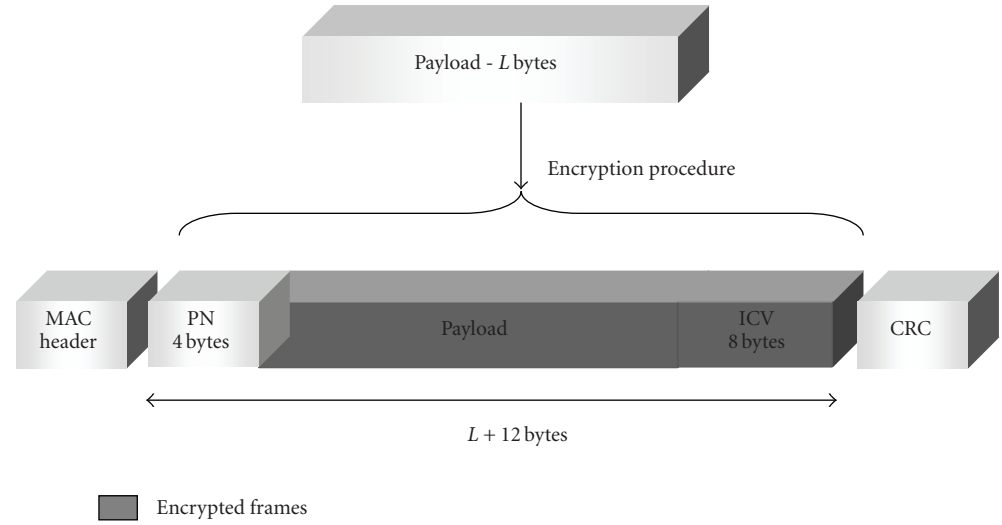


Figure 10: MAC 802.16 encryption frames.

have overlapping lifetimes, just like the AK keys, and the sequence number of the newer is the older number plus one. Each new TEK becomes active halfway through the lifetime of its successor. For each SAID, the BS uses the older of the two active TEKs for encryption of the downlink traffic, while for the uplink traffic uses the older or the newer.

The PKM protocol for the TEK refresh procedure uses the SA-TEK 3-way handshake (see Figure 9) [12].

(*i*) *Message* 1. This message is optional, and BS uses it only when it wants to force a re-key of an SA, or create a new one. In this message the BS sends the key sequence number, its SAID and the digest of this message with the HMAC_KEY_D.

(*ii*) *Message* 2. If BS does not force re-keying, message 2 is the first message that the SS sends to re-key each SA. In this message, the SS sends to BS the key sequence number, its SAID, and the digest of the message with the HMAC_KEY_U.

(*iii*) *Message* 3. The BS receives the second message, verifies the digest with HMAC_KEY_U and if is successful, it sends back the key sequence number, the SAID, the old and the new TEK with their parameters, along with the message digest. The BS encrypts the old and the new TEK with KEK and sends it to SS.

For the Mesh Mode, each node after authorization starts for each of its neighbors a separate TEK state machine

for each of the SAIDs identified during the authentication procedure. The node has the task to maintain the two active TEKs for each SAID between itself and all the other nodes that it initiated the TEK exchange with. The TEK state machine is responsible to maintain keying material. The neighbor replies to the Key Request message with a Key Reply message. The message contains the BS's active TEK for a specific SAID and it is encrypted with the node's public key.

3.1.3. Confidentiality. Confidentiality includes data and TEK Encryption.

Data Encryption. In data encryption the encrypted frames are the MPDU payload along with the 64-bit ICV of the payload (see Figure 10). The ICV is added right after the PDU. At the front, a 32-bit Packet Number (PN) is appended. For the sake of uniqueness, there are separate ranges of values for the uplink and the downlink [15]. According to the TEK length, two encryption methods are implemented.

(i) DES in Cipher Block Chaining (CBC) mode, when TEK is 64 bit. DES in CBC mode uses a 56-bit key with a 64-bit block encryption along with the 64-bit IV. The function actually expects the 64-bit TEK key, but only the 56 bits are used [13]. With the DES-CBC mode, each encrypted ciphertext block is XORed with the next plaintext block to be encrypted, and therefore, it makes the blocks dependent on all the previous blocks. Consequently, in order to find the plaintext of a particular block, the ciphertext, the key, and the ciphertext of the previous block must be known. The first encrypted block has no previous ciphertext, and so the plaintext is XORed with the IV. This mode of operation improves security from the regular DES.

(ii) AES in CCM mode when TEK is 128-bit. The AES in CCM mode uses a 128-bit key and 128-bit block size. The key-PN combination will not be used more than once. The reason is that two sent packets encoded with the same key-PN combination eliminate the security guarantees of the CCM mode. For this reason, and only in the AES-CCM mode, when more than half of the available numbers of the 32-bit PN have been exhausted, the SS schedules a new Key Request, to obtain new key material and avoid this incident.

TEK Encryption. The TEK encryption is again dependent on its key-size. If the size of the TEK is 64-bit, the 112-bit 3-DES is used. The keying material of 3-DES consists of two distinct DES keys. The 64 most significant bits of the KEK are used in the encryption. If the TEK size is 128-bit, the 128-bit AES in ECB mode will be used with a 128-bit KEK. Another encryption method for the 128-bit TEK is the RSA encryption with the SS's public key.

3.1.4. Integrity. For data traffic integrity, ICV is calculated from two modes:

(i) CBC mode. The downlink CBC IV is initialized as the XOR of the IV included in the TEK's SAID, and the content of the PHY synchronization field of the latest DL_MAP. The uplink CBC IV is initialized as the XOR of the IV included in the TEK's SAID, and the content of the PHY synchronization field of the DL_MAP that is in effect when the UL_MAP is created.

(ii) CCM mode**.** The CCM provides data integrity and data origin authentication for some data outside the payload. The ICV is computed from the ESP header, the Payload, and the ESP trailer fields, which is significantly smaller than the CCM-imposed limit. The ESP payload is composed from the IV, the encrypted payload and the Authentication data as it is defined in the RFC 4309 ("Using Advanced Encryption Standard CCM Mode with IPsec Encapsulating Security Payload").

For the management messages integrity, two 160-bit keys (HMAC_KEY_D, HMAC_KEY_U) are used to create the HMAC digest for integrity protection and authentication, by implementing the Secure Hash Algorithm (SHA-1). The digest is calculated over the entire MAC management message, except from the HMAC digests and the HMAC tupple attributes. The HMAC Sequence number in the HMAC tupple is the AK sequence number from which the HMAC_KEY has been derived.

3.2. Security Mechanisms in 802.16e. Although IEEE 802.16-2004 has a strong security protocol, the introduction of the 802.16e corrigendum with its mobility services has enhanced and corrected weaknesses appearing in the 802.16 standard. Due to mobility features introduced with 802.16e, the SS becomes a Mobile Station (MS) as well.

3.2.1. Authentication. With the 802.16e standard, the PKM protocol besides the unilateral authentication of the SS, it can implement mutual authentication for BS and SS. Two methods are used for authentication (see Figure 11): The known X.509 digital certificate with RSA public key encryption as described in the 802.16 authentication, and the EAP method. EAP is a generic authentication protocol and thereby it has to use a particular credential for authentication selected by the operator. Two are the credential types: The X.509 digital certificate of EAP-TLS, and a Subscriber Identity Module for EAP-SIM. The EAP methods are not part of the protocol, but they must fulfill some mandatory criteria (Generation of Symmetric Keying Material, Key strength, Mutual Authentication Support, Share State Equivalence, Resistance to Dictionary attacks, Protection of Man in the Middle attacks) as defined in RFC 4017.

The new feature in 802.16e is the implementation of two Privacy Key Management protocols PKM v.1, and PKM v.2. The difference between the two versions is that PKM v.2 implements more enhanced security features than PKM v.1 does. For both versions, the Authorization Finite State Machine (FSM) remains as described in 802.16 standard.

3.2.2. PKM v.1 Authentication. Authentication with PKM v.1 is the same as described in the 802.16 standard, and it is

unilateral (only SS is authenticated). The procedure uses X.509 v.3 digital certificates with RSA public key encryption for authorization and the following SAID allocation for the single primary, and any static SAs the SS is subscribed for, along with the AK derivation. For the SS's X.509 certificate, the Certificate Issuer Unique ID and the Certificate Subject Unique ID fields are omitted. The EAP in PKM v.1 is optional and applicable only if specifically required. As noted in "Authentication, Authorization, and Accounting (AAA) Key Management Requirements (RFC4017)": *EAP selects one end-to-end authentication mechanism. The mechanisms defined in [RFC3748] only support unilateral authentication, and they do not support mutual authentication or key derivation. As a result, these mechanisms do not fulfil the security requirements for many deployment scenarios, including Wireless LAN authentication [RFC4017]. To ensure adequate security and interoperability, EAP applications need to specify mandatory-to-implement algorithms.* IEEE 802.16e does not specify a mandatory-to-implement EAP method, nor does it specify the required security properties of EAP methods are to be used. The specification as it stands permits implementations to use the EAP MD5-Challenge, which does not generate keys and is vulnerable to dictionary attacks [16].

3.2.3. PKM v.2 Authentication. In PKMv2, RSA and EAP can be used in different deployments such as RSA, RSA-EAP, EAP and EAP-EAP. With two authentication schemes, there are two sources possible for keying material derivation. The RSA based authentication initially creates the pre-Primary AK (pre-PAK), and the EAP creates the Master Session Key (MSK), both for key derivation and management.

The enhancement in the protocol is the mutual authentication between BS and SS. With mutual authentication, the BS presents its own certificate to each SS joins the network. This certificate presents the following.

(i) Country Name (Country of operation)

(ii) Organization Name (Name of infrastructure operator)

(iii) Organizational Unit Name (Wireless MAN)

(iv) Common Name (Serial number)

(v) Common Name (The operator defined BS ID).

Like in PKM v.1, the Certificate Issuer Unique ID and the Certificate Subject Unique ID of the SS's X.509 certificate fields are omitted.

Mutual authentication is performed in two schemes. In the first only the mutual authentication is used, while in the second, mutual authentication is followed by EAP authentication. In the latter case, the mutual authentication is implemented only for initial network entry, while EAP is implemented in the re-entry authentication.

The authorization process (see Figure 12) begins again like in 802.16 with the Authentication Information message from SS to BS. Right after, the SS sends the Authorization Request message consisted of: (i) the SS's X.509 certificate, (ii) the list of the cryptographic suite identifiers, each implementing a pair of packet data encryption and authentication

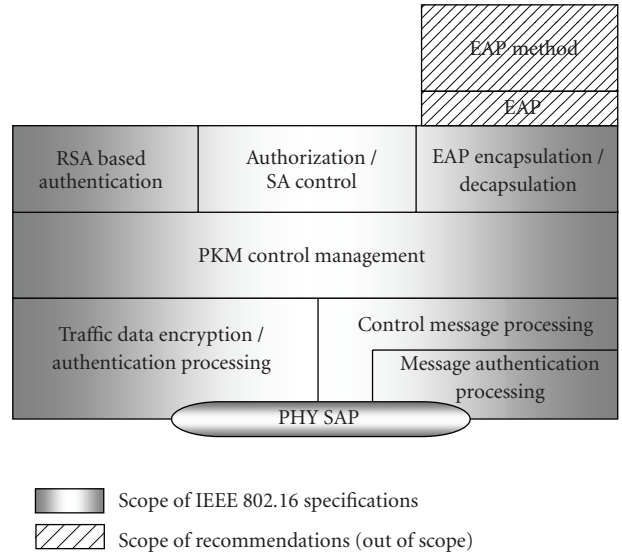


Figure 11: 802.16e security sublayer.

algorithms that SS supports, (iii) the SS's Basic CID, which is the first static CID that the BS assigns to the SS during initial ranging, (iv) A 64-bit random number generated in the SS (SSNonce).

Again, when the BS receives the message, it validates the SS's identity with the X.509 certificate, it checks for basic unicast services and possibly additional statically services the SS is subscribed for, and finally, it determines the cryptographic suite from SS's list from the second message. Then, the BS generates the pre-PAK and encrypts it with the SS's public key. The encrypted pre-PAK is sent from BS in an Authorization Reply message along with the following.

(i) The BS's certificate.

(ii) A 4-bit key PAK sequence number that distinguishes successive generations of AKs.

(iii) The lifetime of PAK

(iv) The SAIDs of the single primary and static SAs the SS is authorized to obtain key material for.

(v) The 64-bit SSNonce.

(vi) A 64-bit random number (BSNonce) generated in the BS to ensure along with SS's nonce the liveness of the message for replay attacks prevention.

(vii) An RSA signature for every attribute in the authorization reply message to ensure message integrity.

When the SS receives the message; it decrypts the pre-PAK with its private key, reads the defined cipher suite and the SAIDs, and proceeds to key exchange with BS.

3.2.4. PKM v.2 Key Derivation and Management. In 802.16 with PKM, the AK derived from BS right after the Authorization Request from SS; the same is implemented with PKM v.1. In PKM v.2 the different authentication schemes (RSA, RSA-EAP, EAP, EAP-EAP) use different key material to

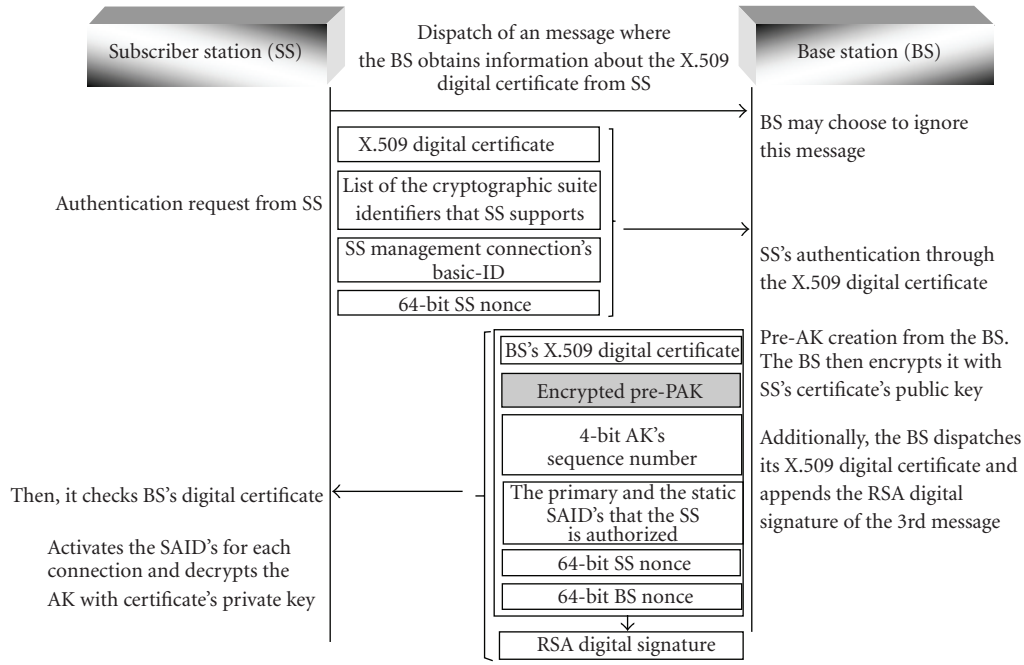


Figure 12: 802.16e Authentication Process with PKM v.2.

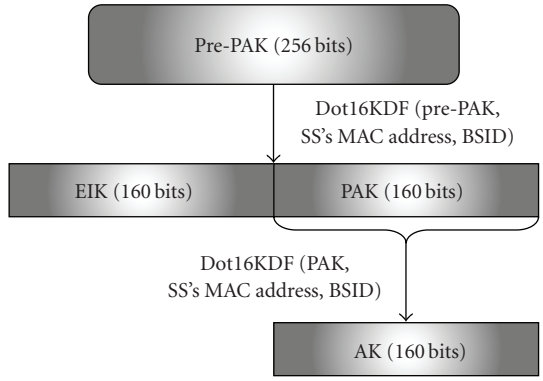


Figure 13: AK derivation with RSA authentication.

construct the 160-bit AK. All the key derivations though, are based on the Dot16KF algorithm, a CTR mode construction that can be used for the creation of an arbitrary amount of keying material from source keying material. If RSA authentication is used, the initial key material is the 256-bit pre-PAK sent from BS to SS. If EAP is used, the key transferred to 802.16e layer is the 512-bit Master Session Key (MSK), which is known to the AAA server, the Authenticator, and the SS. For every authentication scheme, the AK will derive with the following way.

(*i*) *RSA Authentication Only.* From pre-PAK, the SS's MAC address and the BSID, two 160-bit keys are generated. The PAK and the EAP Integrity Key (EIK). With the two new keys along with SS's MAC address and the BSID, the AK is derived (see Figure 13).

(*ii*) *EAP Authentication Only.* From MSK, the 160-bit Pairwise Master Key (PMK) is derived, and optionally the EIK with a MSK truncation to 320 bits. From PKM, the SS's MAC address and the BSID, the AK is derived. During authentication the BS will provide to SS the respective 4-bit PMK sequence number, as it happens with PAK and RSA. The SS caches the PMK upon successful authentication, as the Authenticator does upon its receipt via the AAA protocol. When a new PMK is cached for an SS, the authenticator deletes the old PMK which was used for the specific SS (see Figure 14).

(*i*) *RSA-EAP Authentication.* With the RSA encryption as it was described before, the PAK and the EIK are derived. From EAP in a similar way as before, the PMK is generated. From PAK XORed with PMK, the SS's MAC address and the BSID, the AK is finally created (see Figure 15).

(*ii*) *EAP-EAP Authentication.* From the first EAP authentication, two keys are generated; the PMK-1 and the EIK. From

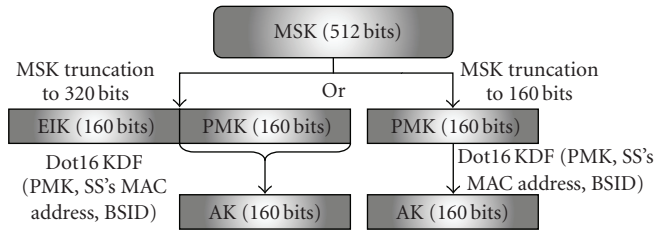


Figure 14: AK derivation with EAP authentication.

the second EAP authentication, only the second PMK-2 is created. With PMK-1 XORed with PMK-2, the SS's MAC address and the BSID the AK is derived (see Figure 16).

Like in 802.16, the SS periodically refresh its AK by reissuing Authorization Requests to the BS, and both SS and BS hold simultaneously two active AK's with overlapping times. The only enhancement in PKM v.2 for the AK is the introduction of a 64-bit ID for each AK (AKID). The AKID is created from AK, AK sequence number, the SS's MAC address and the BSID.

After the AK generation as described in 802.16, three keys are created. One of the three keys is the 128-bit KEK for TEK encryption during the SA-TEK 3-way handshake. The other two keys, the downlink message authentication key and the uplink authentication key will derive according to the used MAC mode. With PKM v.2 two MACs can be implemented. The known from 802.16 HMAC and the new Cipher based MAC (CMAC). In the latter case, the calculated hash value is derived from the CMAC algorithm with AES. The value is calculated over a field that contains: (i) the 64-bit AKID, (ii) the 32-bit CMAC packet number counter, (iii) the 16-bit connection ID, (iv) a 16-bit zero padding for the header alignment with the AES block size, and (v) the entire MAC management message. With CMAC the downlink authentication key CMAC_KEY_D is used to authenticate management messages in the downlink direction, while the respective CMAC_KEY_U is used to authenticate management messages in the uplink direction. Therefore, from AK and the implemented MAC, two options are available.

(i) AK with HMAC**:** In this case the derived keys are: the 128-bit KEK, the 160-bit HMAC_KEY_U and the 160-bit HMAC_KEY_D,

(ii) AK with CMAC**:** In this case the derived keys are the 128-bit KEK, the 128-bit CMAC_KEY_U. and the 128-bit CMAC_KEY_D.

It must be stressed that if only EAP authentication is used, the EIK will be used instead of the AK to generate the aforementioned keys.

The TEK state machine remains the same as described in 802.16 managing key material associated with the respective SAID, but due to the supported multicast features TEK consists of an additional state (Multicast and Broadcast Rekey Interim Wait), and two more events (Group- KEK Updated and GTEK Updated) to the rest described in 802.16. The difference is that the PKM v.2 implements an enhanced

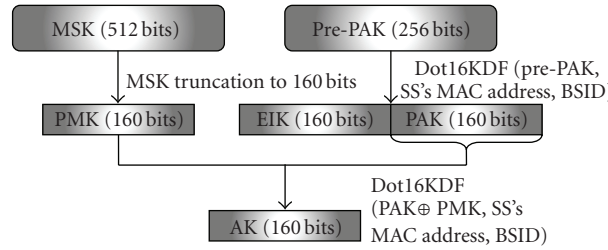


Figure 15: AK derivation with RSA-EAP authentication.

SA-TEK 3-way handshake, which operates in the following way (see Figure 17).

(*i*) *Message* 1. During the initial network entry or a reauthorization, the BS sends a SA-TEK challenge, which includes a random number (BS-Nonce), to the SS with HMAC/CMAC protection. If the BS does not receive a SA-TEK Request message within a certain period of time, it resends the SA-TEK challenge. If again for a certain number of times the BS does not receive a SA-TEK Request, it starts another full authentication procedure or it drops the SS.

(*ii*) *Message* 2. The SS sends the SA-TEK request along with the random number from the SA-TEK challenge, protected with the HMAC/CMAC. In case where the SS does not receive a SA-TEK Response from the BS, it transmits the message again for a specific number of times. If again receives no Response, it fully initiates the authentication procedure.

(*iii*) *Message* 3. When the BS receives the SA-TEK Request from the SS, it performs a number of checks before sending the SA-TEK Response message: (i) confirms that the AKID corresponds to the current AK. If it does not correspond, the BS ignores the message; (ii) verifies the HMAC/CMAC. If it is invalid, the BS ignores the message; (iii) verifies that the BSNonce received from SS with the SA-TEK Request matches with the sent random number in the first message. This process adds freshness to the messages and therefore prevents replay attacks. If the number is different, the BS ignores the message; (iv) checks the SS's security parameters, and if they do not match it reports it to the higher layers. If the validation is successful the BS sends the SA-TEK Response message protected with HMAC/CMAC. For unicast SAs, the BS for each SAID sends the TEK, the TEK's lifetime, the TEK's sequence number, and the 64-bit CBC IV, encrypted

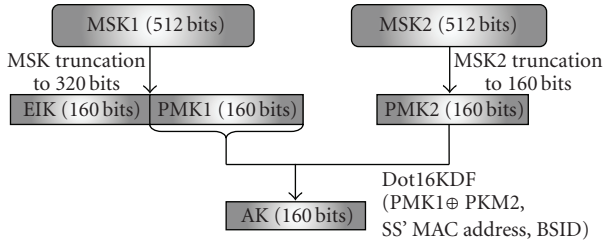


Figure 16: AK derivation with EAP-EAP authentication.

with KEK. In case of group or multicast SAs, the BS for a specific GSAID sends the GTEK, the GKEK, the GTEK remaining lifetime, the GTEK's sequence number and the CBC IV, encrypted with KEK.

When the SS receives the SA-TEK Response message it verifies the HMAC/CMAC digest. If it is valid the SS installs the TEK and its parameters, otherwise, the SS ignores the message.

3.2.5. Multicasting Key Derivation. In multicasting, the key derivation starts with the random generation of the 128-bit Group KEK (GKEK) from the BS and the 64-bit GKEK ID. The key encrypted with KEK is transmitted to SS. There is one GKEK per Group Security Association (GSA) and it is used to encrypt the Group TEK (GTEK) sent in multicast messages to the SSs join the group. GTEK is used to encrypt multicast data packets and it is randomly generated from the BS. GKEK generates the CMAC_KEY_GD for the authentication of multicast messages. The GSA contains keying material and it is used to secure multicast groups. It is defined separately from SAs because they offer lower security, since each of the members joining the group share the keying material and consecutively can forge traffic as if it came from any other member of the group.

3.2.6. Confidentiality with PKM v.2. The length of the TEK an the KEK keys must be either 64 or 128 bits. If the SA implements a cipher suite with a block size of 128 bits, the TEK and the KEK are 128-bit long. Otherwise the length is 64 bits.

Data Encryption. In data encryption, the encrypted frames are the MPDU payload along with the 64-bit Ciphertext Message Authentication Code (see Figure 18). The Ciphertext MAC is added right after the PDU, while at the front, the 32-bit Packet Number (PN) is appended. Again, for the PN there are separate ranges of values for the uplink and the downlink. According to the TEK length, three encryption methods are implemented.

(i) DES in Cipher Block Chaining (CBC) mode using a 56-bit key with 64-bit block encryption along with the 64-bit IV,

(ii) AES in CCM mode with 128-bit key and 128-bit block size,

(iii) AES in CBC mode with 128-bit TEK key and 128-bit block size.

TEK Encryption. The KEK is used for the encryption of the TEK. If it is to encrypt a 128-bit TEK, the 128-bit of the KEK are used directly, otherwise, if TEK is 64-bit long the KEK splits in two 64-bit DES keys. The TEK encryption methods are

(i) 3-DES for 64-bit TEK encryption

(ii) AES in ECB mode for 128-bit TEK encryption

(iii) RSA with SS's public key for 128-bit TEK encryption

(iv) AES Key Wrap for 128-bit TEK encryption. The AES Key Wrap is designed to encrypt key data, and the algorithm accepts both the ciphertext and the ICV, as it is defined in the RFC 3394 ("Advanced Encryption Standard Key Wrap Algorithm").

Group KEK Encryption. The GKEK is encrypted with KEK and the encryption methods are the aforementioned methods used for the TEK.

3.2.7. Integrity with PKM v.2. For the MPDU payload integrity, the ICV can be derived from three modes.

(i) DES-CBC mode. The downlink CBC IV now is initialized as the XOR of the IV included in the TEK's SAID, and the content of the PHY synchronization field of the current frame number. The uplink CBC IV is initialized as the XOR of the IV included in the TEK's SAID, and the content of the PHY synchronization field of the Frame Number of the frame where the relevant UL_MAP was transmitted.

(ii) AES-CCM mode. The integrity procedure of the AES-CCM is the same as it was described for the 801.16 and the PKM protocol.

(iii) AES-CBC mode. The CBC IV created with the XOR of: (i) the CBC IV parameter included in the TEK keying information, (ii) the 128-bit concatenation of the 48-bit MPDU header, (iii) the PHY synchronization value of the MPA that the data transmission occurs, (iv) the 48-bit MAC address and the Zero hit counter.

For management message integrity protection and authentication two MAC modes are implemented.

(i) The HMAC digest with the Secure Hash Algorithm (SHA-1). In PKM v.2 the short-HMAC calculation include the HMAC packet number concatenated after the MAC management message. The HMAC packet number is the AK sequence number.

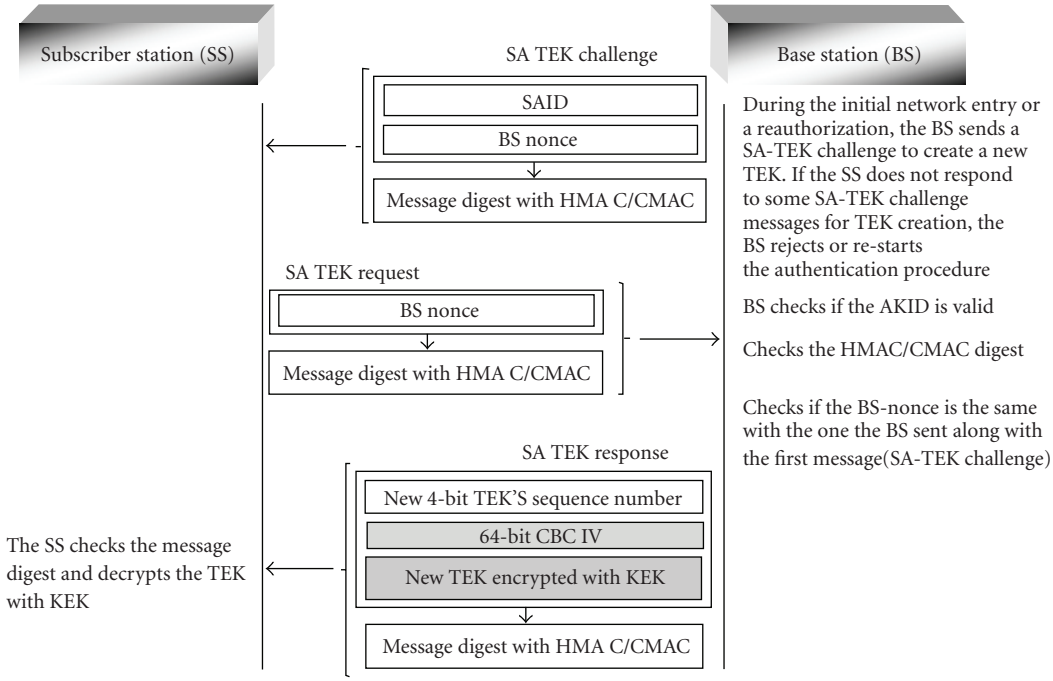


Figure 17: SA-TEK 3-way handshake with PKM v.2.

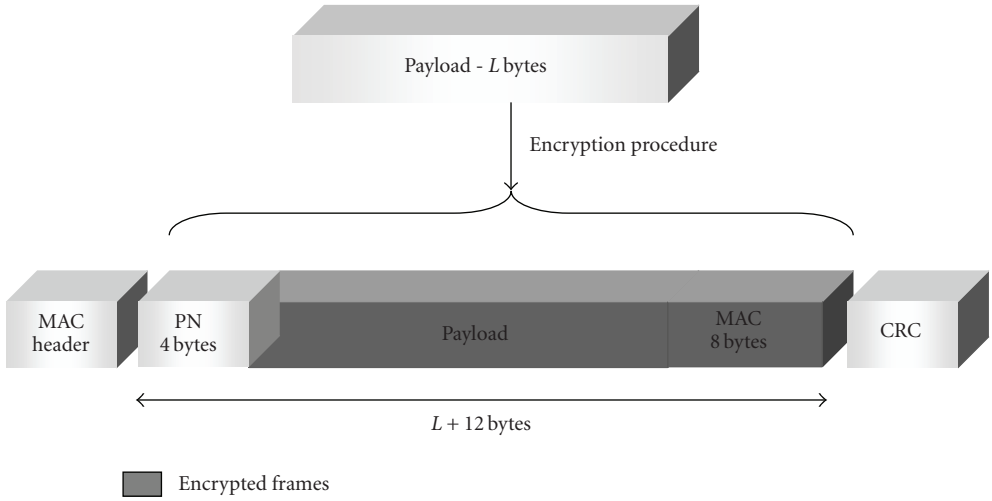


Figure 18: MAC 802.16e encryption frames.

(ii) The CMAC value is implemented as it was described earlier in the PKM v.2 Key derivation and management entity.

4. WiFi-WiMAX Security Comparison

In this section we present a summary of the security mechanisms for authentication, key derivation and management, confidentiality, and integrity procedures applied in WiFi and WiMAX networks.

From the security description in sections WiFi and WiMAX, and with the aid of the following Table 1, it is easy to conclude that WiMAX security is much stronger than it is in WiFi. One of the reasons of course is the large areas that WiMAX covers, and therefore, such conditions demand secure operational conditions of the network, which requires strong security mechanisms.

On the other hand WiFi undoubtedly covers small areas comparing to WiMAX but many WiFi network deployments in companies, industries, agencies and in many cases domestic users, handle valuable confidential information that cannot be compromised. In this case, WiFi security is demanded to be as strong in performance as it happens with the WiMAX mechanisms. Having said that, it is apparent that WEP and WPA security, with RC4 encryption and shared-key authentication, is not adequate to provide guaranteed confidentiality, integrity and secure user-authentication.

Table 1: WiFi and WiMAX security comparison.

(a)

IEEE Protocol		WiFi		
		WEP	WPA	WPA2
Authentication	Method	Open System Authentication Shared Key Authentication	802.1X authentication Shared Key Authentication	802.11X authentication with (RADIUS) server. The EAP method used by IEEE 802.1X will support mutual authentication, as the STA needs assurance that the AP is legitimate.
Key Derivation and Management	Key Management and short description	The keys from traffic encryption are consisted of the concatenation of the 40 bit shared key and the 24 bit IV for a 64 bit key. Most of the vendors use a 104 bit shared key concatenated with the 24 bit IV to create a 124 bit key	TKIP. The 48-bit IV field is used as MPDU TKIP Sequence Counter (TSC). TKIP uses key mixing consisted of the Temporal Key (TK), the Transmit Address (TA), and the TSC for the WEP seed. The WEP seed produced from the aforementioned parameters operates just like the WEP IV. Therefore, assures that every data packet is sent with its own unique encryption key	*Pairwise key hierarchy* for unicast traffic protection. The first key is the 256 bit PMK. PMK derivation depends on the authentication method. If 802.1X is used, the PMK derives from server and the first 256 bits AAA key. If pre-shared key is used, the password is used to create the PMK. PMK generates the PTK from PMK. From PTK three keys are derived. (I) The 128 bit EAPOL KCK, for data origin authenticity in the authentication procedure. (II) The 128 bit EAPOL KEK. (III) The 256 bit for TKIP or 128 bit for AES-CCMP Temporal Key (TK) for WPA2 traffic confidentiality. *Group key hierarchy* for multicast and broadcast traffic protection. The first key created is the GMK. The key GTK. Its length is 256 bit with TKIP, and 128 bit for CCMP. The TK derived from GTK is 256 bit with TKIP, and 128 bit for CCMP and it is used for confidentiality
Confidentiality	Traffic Key Encryption Algorithm	None	None	TK encryption: (I) RC4 with 128-bit KEK. (II) With AES Key Wrap with 128 bit KEK.
	Cipher Algorithms for traffic Data and Key size	RC4 with 64 bit key (WEP-40) RC4 with 128 bit key (WEP-104)	RC4 with 256-bit key.	AES-CCM with 128 bit TK
	Encrypted Frames	MPDU + ICV	MPDU + MIC + ICV	MPDU + MIC
Integrity	Integrity Algorithm	32 bit ICV with CRC-32	(i) 64 bit Michael MIC. (ii) 32 bit ICV	(i) 64 bit CCM MIC for traffic messages (ii-a) HMAC-MD5 with KCK, (ii-b) HMAC-SHA1 with 128 bit KCK for EAPOL 4-way handshake.
	Protected Frames	MPDU	[Michael MIC]: MSDU Sender and Destination Address (SA, DA), the MSDU Priority, and the MSDU payload [ICV]: MPDU	[MIC]: MPDU+ Additional Authentication Data (AAD). The AAD is comprised of the MPDU header, subfields from MAC frame control, addresses from source and destination fields, Sequence Control (SC), QoS control field. [HMAC]: EAPOL 4-way handshake messages

(b)

	IEEE Protocol	WiMAX	
		802.16	802.16e
Authentication	Method	Privacy Key Management Protocol (PKM). Only SS authentication with X.509 version 3 and RSA public-key cryptography	2 PKM versions. V.1 is the 802.16 PKM, and V.2 is more enhanced with mutual authentication option (BS presents its certificate to SS). Two authentication schemes can be used separately or combined: RSA, EAP, RSA-EAP, EAP after EAP authentication. For RSA, client authentication with X.509 v.3 certificates. EAP uses credentials: X509 certificate for EAP-TLS, or Subscriber Identity Module for EAP-SIM.
Key Derivation and Management	Key Management and short description	After Certificate approval, BS sends authorization reply with Authorization Key (AK) encrypted with client's Public Key, and the Security Association set Identity (SAID). From AK derives KEK, HMAC_KEY_U, HMAC_KEY_D, (U for uplink and D for downlink). The last two keys used for the HMAC digest for management messages. For every SAID, a TEK state machine is responsible for key material usage. TEK sends periodically messages for key content refresh. TEK key material is used for uplink and downlink encryption. BS maintains 2 sets of active AKs and TEKs, old and new for each SAID. There is a 4-bit AK sequence number increased by one for each new AK. Additionally a 32-bit packet number (PN). Both prevent replay attacks	AK in PKM v.2 operates as in PKM. In PKM v.2, there two key material primary sources. For RSA, BS' initial key material is the 256-bit pre-PAK (primary authorization key). Pre-PAK gives 160 bit PAK and 160 bit EIK (EAP Integrity Key). PAK+EIK+SS MAC address + BSID generate AK. For EAP only, the initial key is the 512-bit Master Session Key (MSK) and generates the 160 bit Pairwise Master Key (PMK) and optionally the 160 bit EIK with MSK truncation to 320 bits. From PMK+SS' MAC address + BSID AK derives. For RSA-EAP, PAK and EIK derive from RSA and PMK from EAP. AK is generated from PAK XOR PMK+ SS' MAC address + BSID. For EAP after EAP, PMK1 and EIK derive and from 2nd EAP PMK2 derives. PMK1 XOR PMK2+SS' MAC address and BSID, the AK derives. From AK 3 keys derive: One is the 128-bit KEK and the other two are: (I) The 160 bit HMAC_KEY_U and HMAC_KEY_D, if HMAC is used, and (II) The 128 bit CMAC_KEY_U and CMAC_KEY_D, if CMAC is used. If EAP only is used, the three aforementioned keys will derive from EIK. All key derivations are based on the Dot16KF algorithm
Confidentiality	Traffic Key Encryption Algorithm	(i)112 bit 3-DES with 64 bit KEK, if TEK is 64 bits. (ii) AES in ECB mode with 128 bit KEK, if TEK is 128 bits. (iii) RSA encryption with SS's public key if TEK is 128 bits.	(i) 112 bit 3-DES with 64 bit KEK, if TEK is 64 bits. (ii) AES in ECB mode with 128 bit KEK, if TEK is 128 bits. (iii) RSA with SS's public key if TEK is 128 bits. (iv) AES Key Wrap with 128-bit KEK for 128-bit TEK encryption.
	Cipher Algorithms for traffic Data and Key size	(i) DES- CBC with 56 bit TEK and 64 bit block encryption along with 64 bit IV. (ii) AES in CCM mode with 128 bit TEK.	(i) DES in CBC mode. (ii) AES in CCM mode. (iii) AES in CBC mode with 128 bit TEK.
	Encrypted Frames	MPDU + ICV	MPDU + MAC (Message Authentication Code)
Integrity	Integrity Algorithm	(i) DES-CBC mode for 64 bit ICV. (ii) AES-CCM mode for 64 bit ICV. (iii) SHA-1 for HMAC.	(i) DES-CBC mode for 64 bit MAC. (ii) AES-CCM mode for 64 bit MAC. (iii) AES-CBC mode for 64 bit MAC. (iv) SHA-1 for HMAC Digest. (v) AES-CMAC value.
	Protected Frames	[ICV]: MPDU + additional packet information. [HMAC]: Management messages.	[MAC]: MPDU = additional packet information. [HMAC]: Management messages. [CMAC]: Management messages + additional information.

On the other hand, the Robust Security Network Association (RSNA) with the 802.11i and the WPA2 does provide a secure wireless network operation, and it is the only security mechanism in WiFi that operates with AES encryption, CCMP integrity mechanisms, key derivation and management with EAPOL, and secured user-authentication with the 802.1X protocol, that resembles with the strong mechanisms that WiMAX uses.

5. Threat Model for WiFi and WiMAX Networks

Wireless networks face potentially more threats due to the lack of physical infrastructure. Some of the consequences of these attacks include the loss of proprietary information, legal and recovery costs, and the loss of network service. Network security attacks are typically divided into passive and active attacks [17].

In passive attacks an unauthorized entity monitors the traffic, but does not modify its content. Passive attacks are divided in two categories.

(1) eavesdropping, where the adversary monitors the transmissions between a station/SS and an AP/BS,

(2) traffic analysis where the adversary listens into the transmission in order to obtain information from the transmitted packet-flow.

In active attacks, the adversary proceeds to actions in order to achieve his malicious intentions, using sometimes information obtained from earlier passive attacks. Active attacks can be divided in four categories.

(1) Masquerading (Spoofing). This type of attack is actually a man-in-the-middle attack, where an adversary places himself between two parties and manipulates the communication between them. There are two types of spoofing: AP/BS, and MAC address spoofing. In the first, the adversary pretends to be a legitimate AP/BS and tricks users to join the rogue AP/BS network and therefore gains access to information, possible valuable for malicious purposes. With MAC address spoofing, where the MAC address is used to authenticate a station/SS, an adversary can replicate the address of a user.

(2) Replay attacks. With this attack an adversary reuses valid transmitted packets that he has intercepted, without modifying the message during retransmission.

(3) Message modification attacks, where the adversary tampers the content of legitimate messages.

(4) Denial-of-Service (DoS), where the adversary prevents the normal network operation with various ways in PHY and in MAC layer. In PHY layer the attack methods are: (i) jamming, where a device emits electromagnetic energy on the network's frequencies. The energy makes the frequencies unusable by the network, causing a denial of service. (ii) Scrambling, which is similar to jamming but it is applied for short intervals of time and targeted to specific frames or parts of frames, usually control or management messages, in order to disturb the normal network operation [15]. In MAC layer the attack is implemented with the transmission of messages, aiming to decrease the network efficiency.

5.1. WiFi Threat Analysis. The operation of WiFi for almost a decade has revealed various serious security weaknesses like cryptographic vulnerabilities, network exploitations and denial of service attacks, which easily can compromise the wireless network security.

5.1.1. Passive Attacks. The passive attacks in WiFi networks can provide valuable information to adversaries. With eavesdropping, it is possible to gain information about the parties' identity and the time they communicate. With traffic analysis it is possible to analyze traffic patterns and determine the content of communication, as short bursts of activity could mean instant messaging and steady streaming could reveal video conferencing. Additionally monitoring and traffic analysis is the first step to proceed and break cryptographic keys and thereby compromise the network confidentiality and the authentication procedures. Passive attacks, due to the characteristics of the wireless network, are applicable to all WiFi schemes, namely WEP/WPA/WPA2, since all packet traffic can be sniffed and stored.

5.1.2. Active Attacks

Key Cracking. As mentioned earlier, traffic analysis is the first step to cryptographic keys cracking. Indeed, the IV portion of the RC4 key is not encrypted, which allows an eavesdropper by analyzing a relatively small amount of network traffic to recover the key having the IV value known with the advantage of the small 24-bit IV key space, and a weakness in the way WEP implements the RC4 algorithm. Thus, if two messages have the same IV, and the plaintext of either message is known, it is relatively easy for an adversary to determine the plaintext of the second message [8]. Additionally, many messages contain common protocol headers or other easily guessable contents, and therefore, it is possible to identify the original plaintext contents with minimal effort. Even traffic with sequentially increasing IV values is susceptible to attack. There are 16.777.216 million possible IV values; on a busy WLAN, the entire IV space may be exhausted in a few hours. When the IV is chosen randomly, which represents the best possible generic IV selection algorithm, by the birthday paradox two IVs already have a 50% chance of colliding after about 2^{12} frames [18, 19]. As analyzed before, the use of stream ciphers is dangerous and therefore WEP and WAP face a serious threat. With the implementation of AES-CCM with 128-bit key in WPA2, the traffic data confidentiality is well secured. Shared key authentication in WEP and WPA can be breached quite easily. One way is a man in the middle attack where an adversary eavesdrops, captures and views the clear-text challenge value and the encrypted response.

Then he can analyze with off-line brute-force or dictionary attacks the clear-text and the encrypted challenge and thus determine the WEP key stream. Moreover, authentication attack can be achieved by injecting properly encrypted WEP messages without the key [18]. Another problem with shared key authentication is that all devices have to use the same WEP key because WEP does not support key management as WPA and WPA2 when they use 802.1X authentication with EAPOL 4-way handshake. Therefore, if the key is compromised, it needs immediately to be replaced from all stations.

Masquerading: Spoofing. Another way to surpass authentication is the MAC address spoofing [10]. Even if the 48-bit address is large enough to prevent brute force guessing, methods for MAC address filtering and the fact that the address is broadcasted freely in the wireless network, makes it easy for an adversary to obtain it by sniffing the victim's communication. With various programs available to change the MAC address in a PC network adapter within minutes, even if the value in the hardware is encoded and cannot be changed, the firmware value can be altered [20]. Moreover, due to the fact that the AP is not authenticated to the station, an adversary can masquerade a legitimate AP and spoof a station to join the malicious network. The 802.1X supports mutual authentication and therefore the station is secured that the AP is legitimate. On the other hand, 802.1X with EAP-TLS prevents an adversary from forging, modifying, and replaying authentication packets, provided mutual authentication is used. Nevertheless, during the 4-way handshake a session hijacking is possible after the 3rd message sent from AP for successful EAP. At this point, the adversary sends a disassociation management frame to the station-victim to get disassociated, while the 802.1X state machine of the authenticator still remains in the authenticated state. The consequence of this is the network access gaining from the adversary using the MAC address of the authenticated supplicant [21]. Besides that, 802.1X authentication is a very strong authentication mechanism and undoubtedly is preferred in WLANs.

Replay Attacks. WEP does not provide protection against replay attacks because it does not include features such as an incrementing counter, nonce, timestamps that could detect replayed messages immediately. In WPA/WPA2 the 48-bit unique number for each packet is sufficient to prevent replay attacks.

Message Modification. Except from the confidentiality breaching of the implemented algorithms, the integrity algorithm, the CRC-32 can be tampered with bit flipping attacks, since an adversary knows which CRC-32-bit will have to change when message bits are altered even if the CRC-32 ICV is encrypted, because a property of stream ciphers, such as WEP's RC4, is that bit flipping survives the encryption process, as the same bits flip whether or not encryption is used [22]. Michael MIC on the other hand prevents an adversary from inserting modified messages. Even if the adversary intercepts a packet and forwards it to the victim-station later with a valid encrypted MIC, the station will check that the PN is out-of-order and the packet will be discarded. With CCM the integrity of the message is much more secured because besides the payload, CCM authenticates Additional Authentication Data (AAD) as MAC frame control, Sequence Control (SC), addresses from source and destination fields, making thus the message modification impossible, even in the fields sent clear in the air. Additionally message authentication in EAPOL 4-way handshake provides a secure way to key distribution. Although 802.11i protects data frames, it does not offer integrity protection to control or management messages. An attacker can exploit the fact that management frames are not authenticated, and thereby, he can use such messages to destabilize the normal network operation. A message modification threat concerning all WiFi schemes is the IP redirection attack. In this attack the AP acts a router with internet connectivity, which is usually the case, and the adversary all it has to do is to sniff an encrypted packet off the air [18], modify it by giving it a new IP destination, and redirect it to an address belongs to him. Later on, the AP will decrypt and send the packet to the new malicious destination, where the adversary can read the packet in the clear.

DoS Attacks. DoS attacks in WiFi can cause serious implications in the network efficiency. In the PHY layer, jamming can affect the network operation not only intentionally by an adversary, but from other WLANs transmitting in the same frequency, which is something possible since channels in the ISM band are very few. In the MAC layer, the availability can be suspended with flooding attack, where the adversary takes advantage of the CSMA/CA mechanism by constantly transmitting many short-length packets in a fast rate. The effect of this effort is that each station within the network range assumes that the medium is busy and, therefore, each station listens to the medium and waits patiently for its turn to transmit for as long the adversary uses this attack. The implementation of this attack can be achieved easily [23] by placing a wireless network interface card into a test mode where it continuously transmits a test pattern. Another DoS threat is the De-authentication attack, where the adversary, as a legitimate AP, uses the deauthentication message to all stations ordering them to quit the network. The attack is successful since the AP address has been found, which is easy as it is transmitted in the clear, and the adversary has only to listen to the medium and obtain it [24]. With the address available, the adversary transmits the de-authentication message as a legitimate AP. Consequently, every station gets misled and stops communication with the network, having again to repeat the authentication procedure. Another threat is packet removal by an adversary and thus prevention from reaching its destination. This can be done if the adversary interferes in the reception process by causing CRC errors so that the receiver drops the packet. Additionally, if the adversary uses a bidirectional antenna, he can delete the packet on the receiver's side, and simultaneously using another antenna to receive the packet

for himself if he wants so [10]. The aforementioned DoS attacks can be implemented in every WiFi scheme.

5.2. WiMAX Threat Analysis. The security in IEEE 802.16-2004 and 802.16e standards is one of the most important issues in the protocol architecture. The implementation of strong and efficient mechanisms makes the WiMAX security very efficient. Nevertheless, in this short period of their existence, various weaknesses have emerged. Some of the possible threats are similar to the ones that WiFi faced; this observation stresses on the importance of the WiFi threat analysis and the prevention measures that can be taken for WiMAX. Of course, threats in WiFi did not appear right after the introduction of the standards; it took a long period of efforts and computing time from hackers, Government Agencies, Universities, and Research Institutions to reveal the security vulnerabilities issues. This is very important because WiMAX is new and not sufficiently operated to reveal the actual weaknesses it might face, making thus the threat analysis evaluation based on WiFi attacks and estimated vulnerabilities from the new mechanisms of the standard.

5.2.1. Passive Attacks. As mentioned earlier, passive attacks are achievable in a wireless network during packet transmission. Eavesdropping and traffic analysis threats can be used to determine the behavior of an entity about the transmitting times. Moreover, due to the fact that management messages are sent in the clear, they can provide valuable information about the location of the SS at a certain period of time [15]. Additionally monitoring and traffic analysis is necessary to proceed with cryptographic keys cracking to compromise the confidentiality and authentication mechanisms.

5.2.2. Active Attacks

Key Cracking. Cryptographic immunity in WiMAX is based on the fact that the AK remains secret between the BS and the SS. If this is not the case, security is breached. Therefore, the AK generation mechanism and the AK generation material are two important issues. The AK creation according to the standard is assumed to be random with the usage of a uniform probability distribution; if this is the case, it must be explicitly defined. Another important matter is the key material used for the AK generation. The standard defines the BS responsible for the AK creation. The potential problem is if the random number generator appears specific bias to expose the AK. The same issue appears with TEK generation, as the standard fails to specify that the TEK is created using a uniform probability distribution and a cryptographic-quality random number generator [12]. TEK's lifetime is important if the usage period is approaching its maximum value (7 days) and the DES-CBC cipher is implemented. In 1998, the Electronic Frontier Foundation [13, 25] broke a DES encryption in less than three days period, using a DES cracker-machine with a structure costing less than 250.000$. It is obvious that after a decade where computation efficiency is enormous and the hardware costs are constantly decreased, the DES cipher should be considered weak. DES uses a 64-bit block size. One theorem [12] describes that a CBC mode using a block cipher with an n-bit block cipher loses its security after operating on $2^{n/2}$ blocks with the same encryption key. Therefore, with $n = 64$, the maximum safely protected 64-bit blocks are 2^{32}. With an average throughput of 10 Mbps the 2^{32} blocks are produced within 7.6 hours approximately and thereby if TEK's lifetime is at the default value, namely 12 hours, the security can be compromised. Furthermore, the CBC mode requires a random IV to ensure security but the standard uses a predictable IV [12]. On the other hand, AES with key size of 128 bits, and the consideration of the current and the projected technology, makes brute-force attacks impractical [13]; thereby, the usage of AES-CCM and additionally the AES-CBC for the 802.16e, makes data traffic secured. Nevertheless, AES-CCM faces a potential threat when the key-PN combination is used more than once; the reason is that two packets encoded with the same key-PN combination eliminate the security guarantees of the CCM mode. To prevent this, the new key request as described in the standard, demands renewal when more than half of the available numbers of the 32-bit PN have been exhausted. Finally, TEK encryption is well secured with all encryption schemes. Considering though energy consumption, the RSA encryption of TEK with SS's public key and the calculating cost, makes this scheme useful only if for some reason the KEKs cannot be usable for a period of time.

Masquerading: Spoofing. In case of unilateral and not mutual authentication, a rogue BS can masquerade a legitimate BS and spoof a number of SSs by using the BS's address, stolen over the air by intercepting management messages. Nevertheless, since the adversary has to transmit during the legitimate transmission, the procedure is more difficult due to the time division model [15]. Moreover, the signal of the rogue BS must be stronger from that of the legitimate BS. If this is done, the adversary waits until a time slot is allocated to the legitimate BS and commences the attack. As in WiFi, the threat of MAC address spoofing is viable. As it is defined in the standard, each SS has a 48-bit MAC address burned into the firmware and it is used as verification element during authentication procedure from the BS. Currently all 802.16 based network equipment is in the form of standalone units, where MAC address modifications require changes at the firmware level which is difficult unless aid if provided from the manufacturer [20]. Unfortunately this will change since one of the WiMAX Forum members, Intel, announced that it plans to sell IEEE 802.16 compliant chipsets inside laptops [26]. If this is to be implemented, spoofing a MAC address will be easy for WiMAX as it is for WiFi.

Replay Attacks. The PKM v.1 authentication protocol is susceptible to replay attacks since the first and the second message from the SS, and the third message from the BS, do not provide any freshness with nonce or time-stamping, nor implement any message authentication scheme. If the adversary replays any of the three messages the receiver, either the BS or the SS, cannot determine who really the

sender is. Despite the fact that replay authentication messages attacks cannot expose the strongly encrypted AK, it can lead though to a severe result. The reason is that if BS has a timeout value to reject authorization requests (Auth-REQ) from the same SS within a certain period of time, the rightful request from the victim SS will be ignored and thereby leads to Denial of Service (DoS). In case where the BS accepts the requests, a new AK generation will take place continuously leading to exhaustion of the BS's capabilities [27]. In PKM v.2 RSA authentication, the BSNonce along with the SSNonce from the second message ensure freshness against replay attacks on the third message. Nevertheless, a replay attack on the second message just as described before in PKM v.1 is possible since the BS cannot realize that the SSNonce is not fresh. A replay attack can appear in both PKM SA-TEK 3-way handshake versions. In PKM v.1, a request message sent from a SS at an earlier time can be constantly replayed by an adversary, forcing the BS to reply with new TEK key material, exhausting thus the BS's capabilities. Nevertheless, message replay attack cannot succeed anytime. The threat is successful only if the used for the replay attack intercepted message had the same AK during the actual time of the attack. That is, each message is authenticated with an HMAC digest; if the HMAC_KEY_U used for the digest during the message creation, derived from a different AK than the current, the digest would not match and the message would be discarded, leading thus to a failed replay attack. Unfortunately, AK's lifetime ranges between 1 to 70 days with default value the 7 days, making thus the attack very possible for a long period of time. In PKM v.2 the replay attack cannot succeed because of the BSNonce in the SA-TEK challenge message. Since the fact that the BS sends SA-TEK challenges with different nonce, the adversary cannot succeed if he replays the SA-TEK request message, because the BSNonce in the replayed message is not longer valid and thereby, the message will be discarded from the BS. The data traffic is also secure from replay attacks, since each packet has a 32-bit number (PN) preventing from repeated packet numbers.

Message Modification. Authentication and integrity protection in each MPDU payload with DES-CBC, AES-CCM, and additionally AES-CBC for PKM v.2 makes message modification a failed attack. Moreover, management message authentication with HMAC and CMAC is secured to modification. Another weak point appears in the third message sent by the BS in PKM v.1 authentication procedure where message integrity mechanism does not exist. A man in the middle attack is possible to intercept and modify the third message, causing a serious DoS attack. Since that the message does not have any integrity mechanism the adversary can modify the encrypted AK and send it to the victim SS. The SS will decrypt a different AK from the initial legitimate key generated from BS. The usage of the wrong AK key from the SS will lead to the creation of non-legitimate KEK, HMAC_KEY_D, HMAC_KEY_U keys, and consecutively to the decryption from the SS of the TEK sent from the BS with a wrong KEK. As a consequence, the communication between SS and BS will be impossible, since all management messages sent from SS will have different HMAC digests and they will be discarded from BS and vice versa, and moreover, the data traffic encryption-decryption procedure with TEK will lead to the impossible revelation of the plaintext. The problem is fixed in PKM v.2 since the BS uses RSA signature to ensure the integrity of the message and thereby any modification on the encrypted AK will be known to the SS, since the signature comparison from the BS and the signature of the modified message from SS will be different, and therefore the message will be discarded. Leaving aside the secure message authentication implemented in WiMAX, replay and message injection attacks face another difficulty—the timing and the synchronization to inject a message. The adversary has to find an open slot in the schedule and get prepared for his transmission. Even if the adversary knows the propagation delay as a part of the initialization procedure, when he has to inject the message from a BS, he does not know how much propagation delay will meet. Moreover, the adversary has to surpass the stateful characteristic of the WiMAX MAC layer. MAC accepts messages only at certain times, and thereby, it will not respond to messages exceeding this period of time [20]. Therefore, the aforementioned difficulties make replay and message injection a very difficult task to do.

DoS Attacks. WiMAX like every wireless network is susceptible to jamming and scrambling. Nevertheless jamming can be detected quite easily and cannot affect the network severely. Scrambling as mentioned, targets selective control or management messages in order to destabilize the normal network operation, especially when they are time sensitive messages such as channel measurement report requests or responses, which are not delay tolerant. Moreover slots of data traffic can be scrambled, forcing the victim-users to retransmit. Scrambling though needs to surpass important technical difficulties to be successful. The reason is that the adversary must interpret control information and send noise during specific intervals [15]. As shown in WiFi, a deauthentication attack leads to serious DoS. In WiMAX the corresponding message is the Reset Command (RES-CMD) message, where the SS upon receiving this message begins complete reset. An exploitation of this message by an adversary is not possible since the specific management message is authenticated, and thus, a serious DoS attack is prevented. Nevertheless, through the authorization state machine and the Auth Invalid message, a similar DoS attack is possible. The Auth Invalid message can be exploited by an adversary for the flowing reasons.

(i) It is not authenticated and thus can be easily created.

(ii) The message will be accepted from the SS at anytime.

(iii) The message does not utilize the PKM Identifier serial number, and therefore the SS will not discard it as a message with an unmatched Identifier field.

Thereby, if the adversary attacks with this message, it causes a SS transition from the Authorized state to the Reauth Wait state. When the Reauth Wait timer expires, a Reauth Request is sent by the SS, requesting another chance to rejoin the

Table 2: WiFi and WiMAX threat analysis comparative overview.

(a)

IEEE Protocol		WiFi		
		WEP	WPA	WPA2
Passive attacks	Eavesdropping	Cannot be avoided. (i) Traffic patterns can determine the content of communication (Video conferencing, Instant messaging) (ii) Station's and AP's MAC address interception	Cannot be avoided. (i) Traffic patterns can determine the content of communication (Video conferencing, Instant messaging) (ii) Station's and AP's MAC address interception	Cannot be avoided. (i) Traffic patterns can determine the content of communication (Video conferencing, Instant messaging) (ii) Station's and AP's MAC address interception
	Traffic analysis	Cannot be avoided	Cannot be avoided	Cannot be avoided
Active attacks	Key cracking	RC4 key cracking very possible	RC4 key cracking very possible	AES provides safety—No key cracking possible
	User-Authentication Breaching	(i) Shared key authentication weak due to RC4 (Brute force, dictionary attacks) (ii) Firmware change leads to authentication breaching	(i) Shared key authentication weak due to RC4 (ii) Firmware change leads to authentication breaching (iii) 802.1X very secure	(i) Firmware change leads to authentication breaching (ii) 802.1X very secure
	Masquerading (Spoofing)	(i) Station masquerading (ii) AP masquerading	(i) Station masquerading (ii) AP masquerading (When 802.1X is not used)	802.1X authentication very strong but session hijacking is possible after the 3rd message from the AP for successful EAP
	Replay attacks	Yes, no mechanism to prevent replay attacks	48-bit TKIP sequence counter (TSC) to prevent replay attacks	48-bit packet counter to prevent replay attacks
	Message modification attacks	CRC-32 weak to prevent such attacks	(i) CRC-32 weak to prevent such attacks (ii) MIC prevents such attacks on MSDU	CCMP provides safety in modification attacks
	DoS attacks (PHY layer)	Jamming	Jamming	Jamming
	DoS attacks (MAC layer)	(i) Network block with CSMA/CA exploitation (ii) De-authentication attack (iii) Deliberate CRC errors	(i) Network block with CSMA/CA exploitation (ii) De-authentication attack (iii) Deliberate CRC errors	(i) Network operation blocking with CSMA/CA exploitation (ii) De-authentication attack

(b)

IEEE Protocol		WiMAX	
		802.16	802.16e
Passive attacks	Eavesdropping	Cannot be avoided. (i) Information disclosure of the SS's location at certain period of times due to the fact that management messages are sent in the clear (ii) SS's and BS's MAC address interception	Cannot be avoided. (i) Information disclosure of the SS's location at certain period of times due to the fact that management messages are sent in the clear (ii) SS's and BS's MAC address interception
	Traffic analysis	Cannot be avoided	Cannot be avoided

(b) Continued.

	IEEE Protocol	WiMAX	
Active attacks	Key cracking	(i) With DES-CBC there is possibility of cracking if TEK (ii) With AES-CCM, threat if PN-key combination is used more than once (iii) TEK encryption well secured	(i) With DES-CBC there is possibility of cracking (ii) With AES-CCM, threat if PN-key combination is used more than once (iii) With AES-CBC, no key cracking possible (iv) TEK encryption well secured
	User-Authentication Breaching	If network equipment stop being standalone units, as it is the case now, and instead 802.16 compliant chipsets take their place inside laptops, as it was announced from WiMAX forum members, the change of Firmware can lead to authentication breaching	If network equipment stop being standalone units, as it is the case now, and instead 802.16 compliant chipsets take their place inside laptops, as it was announced from WiMAX forum members, the change of Firmware can lead to authentication breaching
	Masquerading (Spoofing)	(i) SS's MAC address spoofing (ii) Lack of mutual authentication could lead to BS's spoofing	(i) SS's MAC address spoofing (ii) Lack of mutual authentication with PKM v.1 could lead to BS's spoofing
	Replay attacks	(i) In PKM authentication, replay attack on the 2nd and 3rd message (ii) In SA-TEK 3-way handshake replay attack possible if AK hasn't changed	(i) In PKM v.1 authentication, replay attack on the 2nd and 3rd message (ii) In PKM v.1 SA-TEK 3-way handshake replay attack possible if AK hasn't changed (iii) In PKM v.2 authentication, replay attack on the 2nd message
	Message modification attacks	(i) Message modification of the 3rd message in PKM of the encrypted AK (ii) For data traffic integrity, DES-CBC and AES-CCM mode ensure safety on message modification attacks (iii) The HMAC protected Management messages are safe on modification attacks	(i) For data traffic integrity, DES-CBC, AES-CCM and AES-CBC mode ensure safety on message modification attacks (ii) The HMAC and CMAC protected Management messages are safe on modification attacks
	DoS attacks (PHY layer)	(i) Jamming (ii) Scrambling (on control and management messages)	(i) Jamming (ii) Scrambling (on control and management messages)
	DoS attacks (MAC layer)	(i) Message modification of the 3rd message in PKM (ii) Replay attacks on 2nd message in PKM authentication (iii) Replay attack in SA-TEK 3-way handshake, if AK hasn't changed (iv) DoS attacks with Reset Command (RES-CMD) management message (v) DoS attacks with Ranging Response (RNG_RSP) set to value 2 [Abort]	(i) Message modification of the 3rd message in PKM v.1 (ii) Replay attacks on 2nd message in PKM v.1 and v.2 authentication (iii) Replay attack in PKM v.1 SA-TEK 3-way handshake, if AK hasn't changed (iv) DoS attacks with Reset Command (RES-CMD) management message (v) DoS attacks with Ranging Response (RNG_RSP) set to value 2 [Abort]

network. The period of the Reauth Wait timer is measured in seconds and if additionally an Auth Reject message is sent at this point, it will lead the SS to the Silent state where it ceases subscriber traffic, responding only to BS's management messages [20]. The usage of the Auth Reject message is achievable since that it is not authenticated as well. The Ranging Request (RNG-REQ) message is the very first message sent by an SS seeking to join a network where the SS requests transmission timing, power, frequency and burst profile information. RNG-REQ is also sent periodically for SS's adjustments. Moreover, the BS can use this message when it demands uplink and downlink channel changing, power transmission modifications and finally, termination of all transmissions and MAC re-initialization of a SS. It

is obvious that if this message could be exploited by an adversary, it would cause a serious DoS attack. Unfortunately, this message is not encrypted, authenticated and it is stateless, making it thus a candidate for DoS attack. Thereby, an adversary can spoof a specific SS by sending an RNG-RSP message, with the ranging status field set to value 2, which means "abort" [20]. The SS's address can be easily obtained by sniffing the channel IDs it uses.

5.3. WiFi-WiMAX Threat Analysis Overview. In this entity with the aid of the following table (see Table 2) we present a summary of the possible threats that WiFi and WiMAX could face during the network operation.

In WiFi, the establishment of the Robust Security Network Association (RSNA) with the 802.11i founds the implementation of a really secure wireless network operation. The pre-RSNA period with WEP and WPA, and the implementation of RC4 encryption in the information confidentiality (privacy) and the user authentication operation, is not secure and easily can be breached. Additionally, the CRC32 checksum cannot guarantee the information integrity of the MPDU's. Moreover, the often key renewal is not an easy task because it requires a key method delivery which is out of the pre-RSNA WiFi operation. On the other hand, the RSNA period forms a secure operation of WiFi. The usage of the AES-CCMP encryption scheme in the confidentiality (privacy) of the information makes key cracking impossible, The CCMP implementation guarantees the integrity of the MPDU along with some Additional Authentication Data (AAD), and the 802.1X authentication provides secure key management and user authentication procedure. Nevertheless, due to the nature of the protocol architecture, the RSNA appears the same weaknesses like WEP and WPA, with two important DoS attacks:

(i) transmission prevention with the fast and constant transmission of short packets, taking advantage of the CSMA/CA algorithm operation,

(ii) De-Authentication attack which uses the ability of the MAC address forging with a simple firmware change.

As mentioned before, WiMAX implements much more enhanced security mechanisms to prevent any possible threats. Leaving aside the specific cryptographic suites that WiMAX uses, the protocol architecture can be characterized with two important features: (a) MAC has a connection-oriented architecture, assigning each slot to a certain connection, each one belonging to various services, like network management and data transport, all of which implement its own security parameters, (b) the stateful characteristic of the WiMAX MAC layer where MAC accepts messages only at certain times, rejecting thereby messages exceeding a defined period of time.

The aforementioned characteristics prevent many Denial of Service attacks, as described in the threat analysis section, make any connection exploitation and message injection extremely difficult. In addition to the sophisticated MAC operation, the WiMAX implemented security mechanisms enhance even more the network security. It is apparent from the detailed description of the WiMAX security mechanisms that user-authentication becomes secure with the X.509 certificates and the RSA asymmetric encryption, especially with PKM v.2 where mutual authentication is needed. Nevertheless, the 802.16 PKM authentication, as shown before, appears some flaws that could lead to some DoS attacks. The confidentiality and the integrity with WiMAX are well secured, although the TEK lifetime could be an issue when DES-CBC is used for data traffic encryption. Even if some management messages implement integrity mechanisms with HMAC of CMAC digests, and thus provide protection on modification attacks, the lack of the implementation to all management messages as shown could lead to serious DoS attacks. As a conclusion it can be stressed that WiMAX implements strong security mechanisms, much more enhanced from WiFi, especially with the 802.160e standard which is used for full mobility characteristics.

In the case of mobility though, an important issue should be determined that concerns the hand-over procedure of a mobile station. The hand-over mechanism is not defined in the 802.16e protocol and it is extremely important to be the fast, secure at the key exchange and the probable authentication procedure, and finally, seamless in real-time applications during the mobile station transfer from one Base Station to another.

6. Guidelines for Secure WiFi and WiMAX Networks

From the WiFi and WiMAX threat analysis, we concluded that WiMAX implements stronger security mechanisms and succeeds to block most of the threats in a wireless network. Nevertheless some weaknesses still exist in WiMAX as well; in the following, we will try to identify the recommendations for WiFi and WiMAX, on how specific mechanisms should be used, how specific security options shall be set and if new security mechanisms, additional to the ones available with WiFi and WiMAX, are needed in order for the network will operate more securely and robustly.

Passive attacks in any wireless network are unavoidable since all messages are transmitted freely in the air. If the network is to ensure the confidentiality of the data traffic by implementing strong encryption schemes as it is recommended later we could minimize the risks of passive attacks.

6.1. Guidelines for WiFi Networks

6.1.1. WEP Security. Threat analysis showed how insufficient is WEP security. The possibilities to enhance security are limited, and if WEP is the only available solution the only thing that can be done to enhance security is the constant key renewal is short periods of time (i.e., each day).

6.1.2. WPA Security. The usage of RC4 encryption faces the same important security issue as described in WEP, even if TKIP uses a different key for each MPDU encryption. Therefore, confidentiality and user shared-key authentication could be compromised as well. The only thing that can

be done, as well as in WEP, is the often key renewal in short periods of times.

In case where WPA can implement the AES-CCMP encryption-integrity security scheme, it is important to be the selected choice in order to provide secure confidentiality and integrity of the transmitted information.

With MIC (Michael) and the TSC operation, WPA succeeds to protect the integrity of MSDUs and the replay attack threat.

User authentication is well secured if the 802.1X authentication is to be used.

6.1.3. WPA2. As noted before, the implementation of the 802.11i protocol in WPA2 defines the Robust Security Network Association era where WiFi networks can be considered very safe. The confidentiality is totally guaranteed with AES encryption, while integrity is likewise secured with the CCMP implementation of the AES-CCMP scheme, where besides the MPDU, some additional authentication data (AAD) are protected as well. As mentioned with WPA, the 802.1X authentication ensures secured authentication procedure.

Nevertheless, as described in threat analysis, 802.1X can face a serious threat that could lead to a user-authentication breaching, and to a DoS attack with the transmission of a De_Auth message (Deauthentication attack). This attack appears in each WiFi security scheme and the reason is the lack of authentication in the De_Auth message.

Therefore in order to prevent this threat, a modification in the WPA and the WPA security operation can be implemented when the 802.1X authentication is used. With 801.X and the EAPOL operation, both parties-Station and AP, possess the 128 bit EAPOL Key Confirmation Key (KCK). This key is used for data origin authenticity and it can be used in the De_Auth message authentication in order to determine that the message not only left from the AP with the specific MAC address that could be changed as shown before, but it must have a legitimate digest produced with the KCK key from the authentic AP, and only the Station can confirm it.

6.2. Guidelines for WiMAX Networks

6.2.1. General Guidelines. WiMAX has already shown some *cryptographic vulnerabilities*; some of them can be fixed if the following issues and specific cipher suites are followed.

(*i*) *Random Number Generation.* A random AK and TEK generation with the usage of a *uniform probability distribution* without any bias is needed. Such a generator must be explicitly defined by the implementation [12]. Additionally, the random number could be a concatenation of two random numbers created from the BS and the SS respectively. This would prevent any possible bias if the random generation is done only by the BS.

(*ii*) *The Lifetime of Keys (AK, TEK).* Since it is understood that short-time key generations will affect the network operation by keeping the BS busy more often with key renewals, the AK can be left at its default value (7 days) and below since the strong encryption (RSA—public key) is used and it cannot reveal the AK easily. Similarly TEK's lifetime should be set not more than its default value 12 hours. This is an acceptable lifetime to ensure that TEK's immunity to key-cracking is guaranteed. It should be noted that increasing the lifetime of keys, may have some (relatively small) positive impact on performance, it does though increase significantly the exposure to key attacks.

The WiMAX forum defines two system profiles; one based on the 802.16-2004 revision of the IEEE 802.16 standard and the other based on the 802.16e amendment. The first targets the requirements of the fixed and nomadic market, and is the first to be commercially available. The 802.16e version has been designed with portable and mobile access in mind, but it will also support fixed and nomadic access. Thereby, since the cryptographic suites for two system profiles are different, we will also differentiate the security planning guidelines.

6.2.2. Guidelines for the 802.16-2004 *Profile.* The following security mechanisms should be selected for the 802.16-2004 profile in order to ensure strong authentication, confidentiality and integrity.

(*i*) *Data Traffic Confidentiality and Authenticity.* the AES-CCM mode should be implemented with 128-bit TEKs, ensuring a strong encryption mechanism. Additionally CCM provides extra data origin authentication for some data outside the payload. If DES-CBC mode is to be implemented, though, it is important to generate an IV randomly with a uniform probability distribution for each packet to ensure secured encryption.

(*ii*) *TEK Confidentiality.* Either 3DES or preferably AES-ECB will provide strong security. RSA public key encryption is not recommended due to large computational costs. It can be implemented though if for some reason the KEK production or the usage is problematic.

(*iii*) *Integrity*. HMAC with SHA-1 is the only applicable management message integrity mechanism, but ensures message authenticity.

The following modifications could enhance the security offered by the 802.16-2004 profile.

(*i*) *Signature on the Third Message.* during authentication for integrity protection with the SS's RSA public key and SHA-1 or MD-5 hash algorithm for message modification prevention. Additionally, time-stamping in the second and the third message is required for replay attack protection. Nonce is not recommended as showed since that the SSNonce in the second message does not prevent a continuous replay attack. Even if the computational cost for the signatures and the time-stamping is increased, it is a onetime procedure for the whole session and it is imperative to be implemented to ensure secure authentication.

(*ii*) *Mutual Authentication.* solution prevents masquerading attacks. Therefore, the BS shall present its certificate within the third message as in RSA PKM v.2.

(*iii*) *Time-Stamping in SA-TEK 3-Way Handshake.* in a similar way with the authentication procedure, a time-stamping should be added in the messages to prevent replay attacks. With this feature, the SA-TEK 3-way handshake will be secured.

(*iv*) *Authenticated Management Messages.* In order to prevent DoS attacks, which cause obstruction in the normal operation of the management messages, all management messages should be authenticated.

6.2.3. Guidelines for the 802.16e Profile. The second system profile, the 802.16e includes all the security schemes that are implemented in the 802.16-2004 standard profile. Therefore, all the security enhancements discussed in the previous section should also be considered with the 802.16e profile in the case where PKM v.1 is to be used.

The 802.16e has stronger and more efficient security mechanisms and thereby the PKM v.2 protocol should be used wherever possible. In this case the security planning guidelines are the following.

(*i*) *RSA along with EAP.* authentication provides strong security with *mutual authentication.* The EAP scheme is not defined within the standard but the EAP-TLS or EAP-SIM should be implemented. It is recommended that even if the authentication procedure demands extra computational cost and time, it must be used because it ensures safe authentication.

(*ii*) *Data Traffic Confidentiality.* The AES-CCM or the AES-CBC mode with 128-bit TEK provides strong encryption. Additionally, CCM or CBC provides secure data integrity.

(*iii*) *TEK Confidentiality.* The AES Key Wrap is preferable because it is specifically designed to encrypt key data, and the algorithm accepts both the ciphertext and the ICV. If it cannot be implemented, either 3DES or preferably AES-ECB mode will provide secured TEKs.

(*iv*) *Message Authentication.* The hash AES-CMAC value is the strongest integrity mechanism because except the management message, it is calculated over additional fields like the 64-bit AKID, the 32-bit CMAC PN counter, and the 16-bit connection ID. Thereby it is the preferable solution for secure message authentication. Of course HMAC can be selected if AES-CMAC cannot be implemented.

Additional modifications in PKM v.2 are suggested in the following areas.

(i) Although RSA in PKM v.2 implements nonce for the second and the third message, as described in the section on WiMAX threat analysis, the second message remains exposed to replay attacks. Time-stamping must be used instead of nonce in order to ensure replay attack protection. In additionally, RSA signatures in authentication messages should be added to prevent message modifications.

(ii) All management messages should be authenticated.

Also, it is clear that the standard misses to define as secure seamless hand-off mechanism. In the following we describe such a mechanism which if implemented will enhance the security of mobility processes.

7. Open Issues and Conclusions

The first target of this work is to analyze and compare the WiFi and WiMAX wireless network security. An important conclusion from this comparison is the highly sophisticated design of the WiMAX networks. An important reason is the operational characteristics of the WiMAX networks, covering large areas and serving many more users than a WiFi network does. Nevertheless, the protection of the information cannot be relevant to the aforementioned characteristics and every security mechanism should guarantee it. Therefore, having WiMAX security as a pattern, it can be said that WPA2 implements similar strong security characteristics and it is the only secure solution in a WiFi network.

The second target of this work is the threat analysis of WiFi and WiMAX. The conclusions from this analysis present similar results as above. In WiFi an important number of threats can create serious problems, where in WiMAX most of these threats are prevented. The reason is the enhanced security mechanisms of WiMAX, along with the operational characteristics of MAC layer. Of course, some threats are still exist, especially in 802.16-2004 standard. In addition to the already defined possible threats, in this work we indicated a weak point in the 802.16 authentication procedure with the message modification attack in the third message sent from the BS and we propose the implementation of the 802.16e authentication mechanism in the guidelines to fix it.

The highest level of security is met in the 802.16e standard, where most of the 802.16-2004 standard security issues are fixed, and simultaneously, supports the mobility feature which is very important in the contemporary way of life. Nevertheless, it leaves two important matters open as far as security is concerned. The first is the implementation of the EAP mechanism. As noted, all EAP applications need to specify mandatory-to-implement algorithms to ensure security and mutual authentication. The second issue is the mechanism to ensure soft HO. Even if WiMAX Forum [7] expects that the initial products will support only simple mobility with hard HOs, which are less complex than soft HOs, but they have a high latency and increased energy consumption. The 802.16e will finally implement full mobility, mobile VoIP, and real-time applications. Security issues remain open for this implementation as pre-authentication procedure is out of the scope of the standard. Nevertheless, a seamless, fast and secure way of key management and transfer during pre-authentication with the aim to avoid a

full repeated authentication procedure, ensuring a smooth transcend from the serving BS to the target BS, remains an open matter.

The demand for wireless broadband access is growing fast and the success is highly dependent on the security it is provided. The implementation of the security guidelines for WiFi and WiMAX networks as described before will prevent any possible threats, enhance and fix indicated flaws, and form a safe environment where wireless communication shall be embraced from users.

Acknowledgments

The author acknowledges that this article reflects personal opinion and it does not in any way represent the opinion of ENISA or any other person or an ENISA body in any way whatsoever.

References

[1] L.M.S.C. of the IEEE Computer Society, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band," ANSI/IEEE Standard 802.11-1999™.

[2] L.M.S.C. of the IEEE Computer Society, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE Standard 802.11b™-1999.

[3] L.M.S.C. of the IEEE Computer Society, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," Amendment 6: Medium Access Control (MAC) Security Enhancements. IEEE Standard 802.11g™-2003.

[4] L.M.S.C. of the IEEE Computer Society, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," Amendment 6: Medium Access Control (MAC) Security Enhancements. IEEE Standard 802.11i™-2004.

[5] L.M.S.C. of the IEEE Computer Society, "Air Interface for Fixed Broadband Access Systems," IEEE Standard 802.16™-2004.

[6] L.M.S.C. of the IEEE Computer Society, "Air Interface for Fixed Broadband Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," IEEE Standard 802.16e™-2005 and IEEE Standard 802.16™-2004/Cor1-2005.

[7] WiMAX Forum, "Fixed, nomadic, portable and mobile applications for 802.16-2004 and 802.16e WiMAX networks," November 2005.

[8] S. Fluhrer, I. Martin, and A. Shamir, "Weaknesses in the key scheduling algorithm of RC4," in *Proceedings of the 8th Annual Workshop on Selected Areas in Cryptography*, Toronto, Canada, August 2001.

[9] B. Aboba and D. Simon, "PPP EAP TLS authentication protocol," *RFC 2716*, October 1999.

[10] C. He and J. C. Mitchell, "Security analysis and improvements for IEEE 802.11i," in *Proceedings of the 12th Annual Network and Distributed System Security Symposium (NDSS '05)*, pp. 90–110, February 2005.

[11] D. Halasz, "IEEE 802.11i and wireless security," August 2004, http://www.embedded.com/.

[12] D. Johnston and J. Walker, "Overview of IEEE 802.16 security," *IEEE Security and Privacy*, vol. 2, no. 3, pp. 40–48, 2004.

[13] W. Stallings, *Cryptography and Network Security*, Pearson Education, 4th edition, 2006.

[14] C. Adams and S. Lloyd, *Understanding PKI*, Addison-Wesley, Reading, Mass, USA, 2nd edition, 2003.

[15] M. Barbeau, "WiMax/802.16 threat analysis," in *Proceedings of the 1st ACM International Workshop on Quality of Service and Security in Wireless and Mobile Networks (Q2SWinet '05)*, pp. 8–15, Montreal, Canada, October 2005.

[16] B. Aboba, "EAP-only security review on 802.16," IETF Liaison to IEEE 802.

[17] T. Karagiannis and L. Owens, "Recommendations of the National Institute of Standards and Technology, Wireless Network Security—802.11, Bluetooth and Handheld Devices," NIST Special Publication 800-48, November 2002.

[18] N. Borisov, I. Goldberg, and D. Wagner, "Intercepting mobile communications: the insecurity of 802.11," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking (MOBICOM '01)*, pp. 180–188, Rome, Italy, July 2001.

[19] A. Stubblefield, J. Ionannidis, and A. D. Rubin, "Using the Fluhrer, Mantin, and Shamir attack to break WEP," in *Proceedings of ISOC Symposium on Network and Distributed System Security*, February 2002.

[20] D. D. Boom, *Denial of service vulnerabilities in IEEE 802.16 wireless networks*, M.S. thesis, Naval Postgraduate School, Monterey, Calif, USA, September 2004.

[21] A. Mishra and W. Arbaugh, *An Initial Analysis of the IEEE 802.1X Standard*, Department of Computer Science, University of Maryland, 2002.

[22] K. Scarfone, L. Owens, B. Eydt, and S. Frankel, "Establishing Wireless Robust Security Networks to IEEE 802.11i," NIST Special Publications 800-97. February 2007.

[23] C. Wullems, K. Tham, J. Smith, and M. Looi, "A trivial denial of service attack on IEEE 802.11 direct sequence spread spectrum wireless LANs," in *Proceedings of Wireless Telecommunications Symposium (WTS '04)*, pp. 129–136, Pomana, Calif, USA, May 2004.

[24] J. Bellardo and S. Savage, "802.11 Denial-of-Service Attacks: Real Vulnerabilities and Practice Solutions," Department of Computer Science and Engineering, University of California at San Diego.

[25] Electronic Frontier Foundation, *Cracking DES: Secrets of Encryption Research, Wiretap Politics and Chip Design*, O'Reilly, Sebastopol, Calif, USA, 1998.

[26] The Register, Intel: WiMAX in notebooks by 2006, September 2004, http://www.theregister.co.uk/2004/07/02/intel_wimax/.

[27] S. Xu and C.-T. Huang, "Attacks on PKM protocols of IEEE 802.16 and its later versions," in *Proceedings of the 3rd International Symposium on Wireless Communication Systems (ISWCS '06)*, pp. 185–189, Valencia, Spain, September 2006.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 413987, 12 pages
doi:10.1155/2010/413987

Research Article

Investigation of Cooperation Technologies in Heterogeneous Wireless Networks

Zhuo Sun and Wenbo Wang

Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts & Telecommunications, P.O. Box 93, Beijing 100876, China

Correspondence should be addressed to Zhuo Sun, zhuosun@bupt.edu.cn

Received 29 September 2009; Accepted 2 February 2010

Academic Editor: Rashid Saeed

Heterogeneous wireless networks based on varieties of radio access technologies (RATs) and standards will coexist in the future. In order to exploit this potential multiaccess gain, it is required that different RATs are managed in a cooperative fashion. This paper proposes two advanced functional architecture supporting the functionalities of interworking between WiMAX and 3GPP networks as a specific case: Radio Control Server- (RCS-) and Access Point- (AP-) based centralized architectures. The key technologies supporting the interworking are then investigated, including proposing the Generic Link Layer (GLL) and researching the multiradio resource management (MRRM) mechanisms. This paper elaborates on these topics, and the corresponding solutions are proposed with preliminary results.

1. Introduction

In the near future, multitude of wireless communication network based on a variety of radio access technologies (RATs) and standards will emerge and coexist. The availability of multiple access alternatives offers the capability of increasing the overall transmission capacity, providing better service quality and reducing the deployment costs for wireless access. In order to exploit this potential multiaccess gain, it is required that different RATs are managed in a co-operative fashion. In the design of such a co-operative network, the main challenge will be bridging between different networks technologies and hiding the network complexity and difference from both application developers and subscribers and provide the user seamless and QoS guaranteed services. The trend will also bring about a revolution in almost all fields of wireless communications, such as network architecture, protocol model, radio resource management, and user terminal.

There are always plenty of prior researches on the cooperation of heterogeneous RATs, including a number of IST FP projects [1]. However, in the view of this paper, two technologies play an important and foundational role in efficient cooperation between different radio technologies, including: *Generic Link Layer (GLL) and Multiradio Resource Management (MRRM)*.

The generic link layer and multiradio resource management are firstly discussed in Ambient Networks Project [2]. The GLL may be identified as a toolbox of link layer functions, which is designed with the capabilities of universal link layer data processing and reconfiguration to enable different radio access networks to cooperate on the link layer. GLL not only can offer the lossless and efficient solution for intersystem handover, but also make the possibility of multiradio transmission (or reception) diversity and multiradio multi hop. Multiradio Transmission Diversity (MRTD), implies the sequential or parallel use of multiple RAs for the transmission of a traffic flow. Multiradio Multihop (MRMH) implies link layer support for multiple RAs along each wireless connection over a multi-hop communication route. Moreover, in the heterogeneous relay network, in order to provide the better end-to-end QoS guarantee a unified expression or evaluation of QoS capability through a transmission link is needed. QoS Mapping is used to translate QoS guarantee provided by the next hop into their effects on the previous hop (sender).

MRRM is a control-plane functionality designed to manage all the available radio access resources in a coordinated manner, such as load balance, radio access selection, and mobility management. (In other papers, the MRRM item may be replaced by Joint Radio Resource Management (JRRM) and Common Radio Resource Management sometimes.) The aim of introducing MRRM is to efficiently use the radio resources in a multiaccess network, it is important to provide optimum radio resource management functionalities between the different RATs in the RAN.

In the following, these aforementioned issues will be elaborated, respectively. This paper is organized as follows. Firstly, in Section 2 we propose advanced interworking networks architecture by taking WiMAX and 3GPP long-term evolution networks as a specific case. The GLL adopted in the protocol architecture is introduced, and the investigation of several novel concepts of GLL is presented in Section 3. Section 4 discusses the key functionality and mechanisms of MRRM, especially for load balance and RA selection. Section 5 concludes this paper.

2. Interworking Architecture Based on Multiradio Access

It is important to note that having a well-defined interworking architecture, which is a very challenging task to researchers, will accelerate the creation of enriched services through the co-operation of networks. In this paper, we focus in particular on an interesting use case: the integration of mobile WiMAX within 3GPP LTE networks. This integration is facilitated by the evolved packet network architecture, which has recently been standardized by 3GPP in the context of Release 8 specifications [3].

After the introduction of the IP transport in R4 and R5, 3GPP TSG RAN group studied the UTRAN architecture evolution items [4, 5] to improve the radio performance and transport layer utilization; this work continues in release 8 [3]. In [4], several UTRAN architecture enhancement proposals are presented based on: separation of control and user plane, redefinition of UTRAN nodes functionalities, and separation of functional entities for cell, multicell, and user related functions. Another aspect in the scope of this work is the functionality increase of nobe B, which moves parts of the RNC functionalities to an evolved node B(iNodeB) including cell specific radio resource management, soft HO management and radio processing (MAC, RLC, and PDCP), and user data handling.

In [6], it provides some approaches to co-operation between multiple RATs in a multiradio environment which are investigated in the work package "multiradio access" (MRA) of the Wireless World Initiative—Ambient Networks project. A multiradio access (MRA) interworking architecture is also proposed in [6], and different levels of co-operation have been studied based on two concepts: generic link layer and multiradio resource management in order to exploit the potential multiaccess gain.

Herein we adopt these ideas and clues [3–7] and propose two architectures of WiMAX and LTE interworking with necessary logical nodes and interfaces. The two architectures are designed on the basis of different levels of interworking, and each of them can combine several RATs within a single RAN and allow a flexible deployment of network nodes and the interconnecting transport network. They not only combine common functions of different RATs but also are build on a Generic Link Layer (GLL) [6] which generalizes some common link layer functions for different RATs, such as queuing of data packets, higher layer header compression, segmentation and retransmission functionality and an enhanced Radio Resource Control (RRC) layer which adds the Multiradio resource management (MRRM) [6] functionalities.

2.1. Radio Control Server- (RCS-) Based Centralized Architecture. Figure 1 shows a proposal of WiMAX and LTE interworking architecture that consists of the following logical nodes and networks.

(i) User Terminal (UT): this logical node consists of all functionalities necessary for an end user to access either WiMAX or LTE network.

(ii) Relay Node (RN): it consists of forwarding functionality in order to extend the network's coverage area and simplify the network planning.

(iii) Base station (BS): it is a pure WiMAX Access Point (AP).

(iv) Radio Control Server (RCS): the one in WiMAX network controls the BSs with associated UTs and the one in LTE controls Node Bs with associated UTs.

(v) Multiradio Control Server (MRCS): this node is defined to control and coordinate some RCSs for interworking.

(vi) Bearer Gateway (BG): this node acting as Access Router (AR), assigns IP address, and so forth, and consists of GLL and WiMAX and LTE RATs specific user plane functions.

In this proposed architecture, WiMAX and LTE RANs co-operate in a loose mode based on the RCSs and MRCSs. The evolutional RAN architecture of 3 G as aforementioned is adopted with an evolved node B and separation of user and control plane. The new introduced RCSs and MRCSs will play an important role in the cooperation of the two different RATs. Actually, MRCS and BG are two different logical nodes, but they can be located in the same communication entity. MRCS is used to complete the functionalities in the control plane, while BG domains the user plane.

The radio interface protocol stack in the control plane is described as shown in Figure 2. Note that UT not only can directly communicate with BS/Node B, but also can communicate via RN. The GLL is defined above (or within) the L2 and below Radio Resource Control (RRC) layer. It needs to notice that the GLL entity in BS/node B is optional in the loose cooperation scenario. In RRC layer, the MRRM controls the radio connection and management of radio resource for different RATs and different hops, by cooperation between different MRRM entities in MRCS,

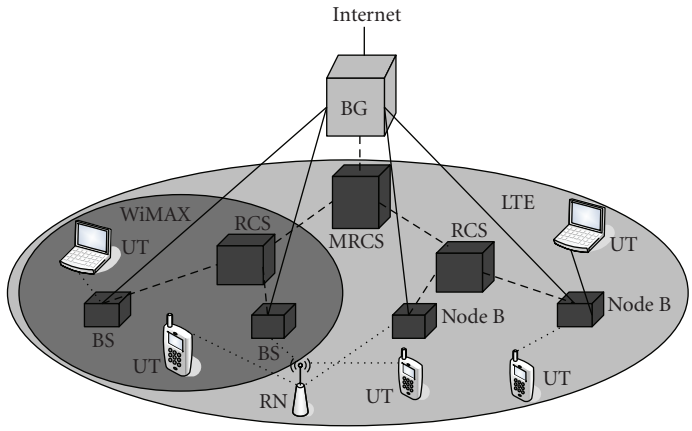


Figure 1: RCS based centralized architecture.

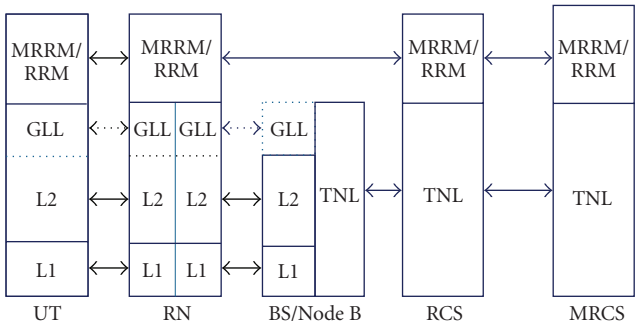


Figure 2: Interface protocol architecture in the control plane.

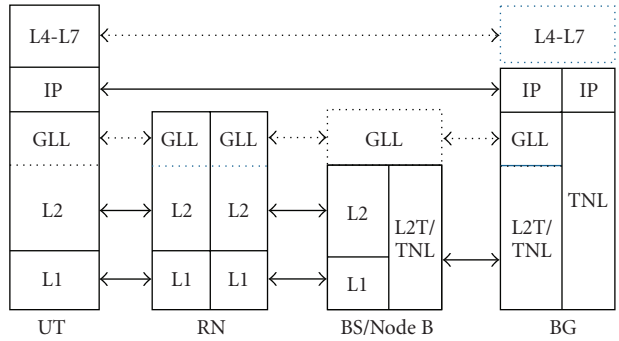


Figure 3: Interface protocol architecture in the user plane.

RCS, RN, and UT. TNL which means Transport Network Layer is used to carry the radio interface protocols between infrastructure nodes.

The radio interface protocol in the user plane is described as Figure 3. The interface between different communication entities in the user plane is the same as that in the control plane. However, RCS and MRCS are not concerned here as user and control planes are separate. But Bearer Gateway (BG) is needed to convey different data formats, respectively, from LTE or WiMAX to the IP core network and vice versa. Therefore GLL should be involved in this node. IP packets are transmitted between the BG and the Node B/BS via some layer two tunnels (L2Ts) based on some specific tunnelling protocol.

2.2. Access Point- (AP-) Based Centralized Architecture. Figure 4 shows another proposal of WiMAX and LTE interworking architecture that consists of the following logical nodes.

(i) User Terminal (UT): this logical node consists of all functionalities necessary for a terminal user to access either WiMAX or LTE at least.

(ii) Relay Node (RN): it consists of retransmission in order to extend coverage area.

(iii) Radio Access Technology Access Point (RAT AP): it is a combined WiMAX and LTE Access Point in one node with GLL features.

(iv) Radio Control Server (RCS): this is a general controller of RAT AP which performs both RRM and MRRM functions.

(v) Access Router (AR): the Access Router assigns IP address and carries out routing functions which depends on route parameters, and so forth. It may include or not include GLL because all RAT APs provide an identical format of data (all IP packets).

In this proposed architecture, WiMAX and LTE RANs co-operate in a tight mode based on the RCSs and ARs. The evolutional RAN architecture of 3G as aforementioned is adopted with an evolved node B and separation of user and control plane. RAT-AP supports both WiMAX and LTE access technologies, and Access Router (AR) is independent of any RATs and needed for routing functionalities. Therefore GLL should not be involved in AR and RCS.

3. Generic Link Layer

Generic link layer as an additional communication layer that provides universal link layer data processing for multiple radio access technologies may be identified as a toolbox

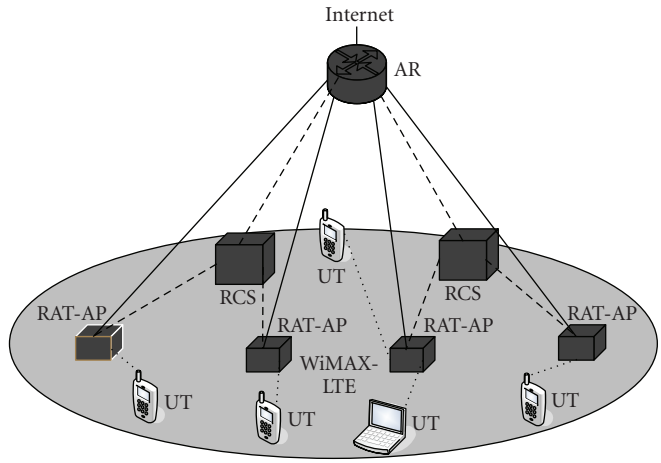


Figure 4: AP-based centralized architecture.

of link layer functions that can be readily adapted to the characteristics of both legacy and new (as yet unforeseen) radio access technologies. Figures 2 and 3 depict reference protocol model of generic Link Layer in heterogeneous networks. In these figures, both the RAN and terminal have installed the GLL logical architecture to support the efficient cooperation between different radio access technologies.

One of the important functions of introducing GLL is to enable lossless and efficient intersystem handover. Considering an intersystem handover process without GLL, a mobile terminal dynamically selects one of the two available radio access networks. During the lifetime of a session, an intersystem handover from RAN A to RAN B is executed in the case of the movement of the terminal or a change of the radio link quality, the radio link in RAN A is torn down and a new radio link is established in RAN B. In consequence, all buffers in the old link layer of RAN A are flushed and all data stored for transmission is discarded. Consequently, such an intersystem handover can lead to a significant amount of packet losses. The motivation for a generic link layer is to overcome this problem by making radio access networks cooperate on the link layer. If the radio link layers are compatible, the old radio link layer state can be handed over to the new radio link layer that continues the transmission in a seamless way, where the generic link layer is used for both radio links in the different radio access networks with different configurations.

More specifically, GLL should have the following functions [6]: (1) provides a unified interface to the upper layers, acting as a multi-RAT convergence layer, hiding the heterogeneity of the underlying multi-RAT environment, (2) controls and maybe complement the RLC/MAC functionalities supported by the multiple RATs in order to maximize the application layer performance while utilizing the radio resources allocated by the MRRM, (3) provides a modular architecture that readily caters for the integration and co-operation of different types of legacy and future RATs, (4) provides support for novel concepts such as dynamic scheduling of user packets across multiple RATs selected by the MRMM and other forms of multiradio macro diversity, (5) provides link layer context information to the higher layers for supporting efficient inter-RAT mobility management.

The proposed GLL facilitates two novel applications. The first one, named Multiradio Transmission Diversity, implies the sequential or parallel use of multiple RAs for the transmission of a traffic flow. The second one, termed Multiradio Multi-Hop networking, implies link layer support for multiple RAs along each wireless connection over a multi-hop communication route.

3.1. Multiradio Transmission Diversity (MRTD). Multiradio transmission diversity (MRTD) is defined as a well-defined split of a data-flow (on IP or MAC PDU level) between two communicating entities over more than one RAT. The transmitting entity may select one or more RATs among the available ones to achieve the gain of multiradio diversity. Different MRTD schemes are possible. When referring to the scheme of selecting the multiple RAs at any given time for transmission of user data, MRTD is classified as well two schemes: switched (sequential) and parallel MRTD [8].

For switched MRTD, user's data, equivalent in size to the payload of MAC PDUs, is transmitted via only one RA PHY layer at any given time. Successive MAC PDUs may be transmitted via different RA physical layers. The paper [9] studies packet scheduling algorithms in order to exploit multiradio transmission diversity in multiradio access networks, where the packet scheduling process is viewed as a combination of user scheduling and radio access allocation. In [10], the authors address the problem of multiuser scheduling with multiradio access selection, it shows that performance gains are possible and come from multiuser diversity as well as multiradio diversity while both the best user and the best radio access were selected. Parallel MRTD is implemented by simultaneously transmitting the copies of same data over multiple RAs, in other word different RAs are allowed to serve the same entity, so as to increase the robustness. At the reception, the received packets from different radio accesses can be combined based on some

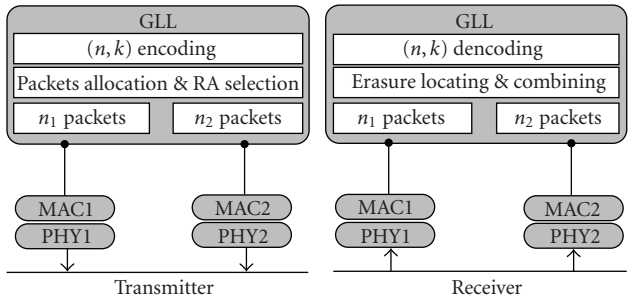


Figure 5: MRTD based on Packet level FEC.

strategies to achieve the gain of MRTD. The researching of parallel MRTD schemes is still scare up to now.

Both switched and parallel MRTD can provide considerable performance gain, but there are also some constraints for them. For switched MRTD, it supposes every selected radio access can provide enough bandwidth or data rate serving for the user, but which is not always possible. For parallel MRTD, the transmission efficiency is decreased due to that the reduplicate packets need to be transmitted simultaneously. Therefore, in this paper, a novel MRTD scheme based on packet level FEC (MRTD-PFEC) is proposed, which both considers the constraints of maximum data rate of each RA for the user on the one hand and integrates with packet level FEC to achieve better transmission efficiency than parallel MRTD scheme. The brief idea of the scheme is described as follow: the source packets (original information packets) will be firstly coded at generic link layer for the enhanced capability of error correction, then the coded packets are allocated over different RAs according to the specified selection algorithm, in order to minimizing the probability of irrecoverable loss at receiver side. At the reception, the source packets can be recovered based on combined decoding procedure at GLL.

We firstly assume that the sender with multi-mode has more than one ($l > 1$) available radio accesses (RAs) for simultaneous transmission, and these l radio access networks (RANs) are interworking in a cooperated fashion. Moreover, the terminal is designed with the functionality of MRRM and GLL. For simplicity, the number of available RAs is assumed as two ($l = 2$) in the following analysis, but it can be extended to the case with $l > 2$.

Figure 5 give a modal structure of the proposed MRTD scheme. The implementation procedure of MRTD-based packet level FEC scheme consists of four steps: *packet level encoding, channel measurement, packets allocation, and receiving and decoding* sequentially. As the most important step, the packet allocation process will be elaborated in the following.

3.1.1. Packet Level Encoding. At the sender, the data from upper layer (e.g., IP) are segmented into packet with fixed length L (bits) at GLL. The GLL packets are sequentially buffered and the continuous k packets are coded into n packets by using the (n,k) packet level forward error correction.

Different from bit level correction strategies, packet level correction operates on sequences of packets and deals with straight packet losses, while bit level correction operates on sequences of bits and deals with unpredictable bit error. For packet level FEC, one of advantages is that the decoder can know where the errors are by use of a Cyclic Redundancy Check (CRC), while the CRC field exists in each packet. These known error locations are called erasures, with which the decoder can correct more errors than that without the information of error locations. A (n,k) block erasure code takes k source packets and produces n encoded packets in such a way that any subset of k encoded packets (and their identity) allows the reconstruction of the source packets in the decoder and can recover from up to $n-k$ losses in a group of n encoded blocks.

When using the *Vandermonde code* [11] as the erasure code, the coding process can be represented as

$$y_{(n)} = G_{(n\times k)} \times x_{(k)}. \tag{1}$$

where $x = x_0 \cdots x_{k-1}$ are the source data, G is an $n \times k$ encoding matrix with rank k and consists in using coefficients of the form

$$g_{ij} = x_i^{j-1}. \tag{2}$$

It should be pointed out that the redundancy level $n - k/n$ is determined by the requirement of tolerant error rate for the service.

3.1.2. Channel Measurement. We assume that the instantaneous channel state of one RA link between sender and receiver is available sender through specific feedback and measurement mechanism, which is beyond the scope of this paper and would not be detailed here. Then, the average channel signal-to-noise ratio (SNR) can be calculated as

$$\gamma = \alpha\gamma_t + (1-\alpha)\overline{\gamma}, \tag{3}$$

where γ_t is the instantaneous channel, SNR, $\overline{\gamma}$ is the average channel SNR before the time t, and α is a constant. The average channel SNR will be used in the following step.

3.1.3. Packets Allocation. The goal of packets allocation is adaptive to the capability and reliability of the available transmission channels (i.e., RAs) in order to exploit the maximum gain of MRTD. Herein, we give an allocation strategy with the goal of minimizing the probability of irrecoverable loss at receiver.

In Section 2, we mention that a (n,k) block erasure code takes k source packets and produces n encoded packets in such a way that any subset of k encoded packets (and their identity) allows the reconstruction of the source packets in the decoder and can recover from up to $n-k$ losses in a group of n encoded blocks. Therefore, the probability of irrecoverable loss equals the probability of more than $n-k$ lost packets out of n packets sent via the two RAs.

We divide the n packets into two groups with the length of n_1 and n_2, respectively, and the process of separation satisfies the following condition:

$$n = n_1 + n_2. \tag{4}$$

Assuming the transmission via different RAs is independent, according to the separation, the probability of irrecoverable loss at the receiver can be expressed as

$$C(k, n_1, n_2) = \sum_{j=n-k}^{n} \sum_{i=0}^{j} P_1(i, n_1) P_2(j-i, n_2), \tag{5}$$

Subjected to (5), $n_1 L/T_1 < B_1, n_2 L/T_2 < B_2$, where $P_1(i, n_1)$ represents the fact that there are i packets lost out of n_1 packets sent via RA 1, $P_2(j-i, n_2)$ represents the fact that there are $j-i$ packets lost out of n_2 packets sent via RA 2. $C(k, n_1, n_2)$ is the probability that total more than $n-k$ packets are lost out of a total $n_1 + n_2$ packets sent by both senders. T_l is the total time duration of sending n_l packets via RA l, and B_l is the constraint of maximum data rates ($l = 1, 2$).

The goal of the allocation algorithm is to select the optimized value of n_1, n_2 to minimize the probability of irrecoverable loss:

$$(n_1, n_2) = \arg \min_{n_1, n_2} C(k, n_1, n_2). \tag{6}$$

The process of searching for n_1, n_2 is fast since only n comparisons are required for the senders.

3.1.4. Receiving and Decoding. At the receiver, both the two parts of received packets from RA 1 and 2 are collected. The packets detected by CRC with error are discarded firstly. If there are more than k packets without error, recovery of original data is possible by solving the linear system

$$\underline{y}' = G' \underline{x} \Longrightarrow \underline{x} = G'^{-1} \underline{y}', \tag{7}$$

where $\underline{x}$ is the source data, $\underline{y}'$ is a subset of k components of $\underline{y}$ available at the receiver, and matrix G' is the subset of rows from G corresponding to the components of $\underline{y}'$.

Otherwise, the retransmitting strategy will be triggered to retransmit some of the error packets until more than k packets are received without error. The retransmitting process is also beyond the scope of this paper and will not be described in detail.

The simulation works have been carried out to investigate the performance of our proposed MRTD-FEC scheme based on the last section. Three types of MRTD scheme are compared together.

Switched MRTD. The packet at GLL is sent via the selected RA, where the maximum throughput RA selection strategy proposed in [9] is used.

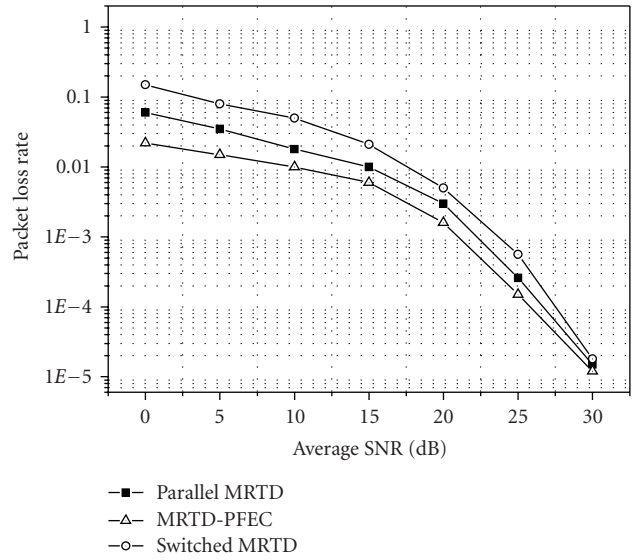


Figure 6: Packet loss rate versus average SNR.

Parallel MRTD. The packet at GLL is duplicated and sent via both the RAs. And the proposed *MRTD-PFEC* isin the paper.

To assess the performance of the proposed scheme clearly, Figure 6 shows the packet loss rate versus different average signal noise ratio. (Herein, we adopt the same average SNR for the user on the two RAs because the instantaneous SNR on different RAs are different at a time.) From the figure, it can be seen that the least packet loss rate happens in the MRTD-PFEC strategy, especially in the cases of lower average SNR. When the SNR is increasing and above certain value, all of the MRTD schemes have almost the close performance. That can be explained when the channel state was favorable with high robustness, the diversity and error correction strategies will be needed rarely and not contribute the correction of packet losses sufficiently.

Figure 7 depicts the average expected goodput versus different average SNR. We can observe when the channel state is in a bad condition that the MRTD-PFEC can provide the best expected goodput among the three strategies. When the SNR increases and the channel condition changes better, the switched MRTD outperforms MRTD-PFEC and parallel MRTD strategies since the needs of error correction and diversity are reduced but the pain of increased overhead introduced by MRTD-PFEC outstands. When combining with the results of packet loss rate, we can conclude that the MRTD-PFEC performs well especially when the channel states of the available RAs are in a bad condition as well.

3.2. Multiradio Multihop. From the multiradio access perspective, the scenarios that need to be targeted are quite different from the ones that have been traditionally associated to ad hoc (multi-hop) networks. Multi-hop communications are thought to be an extension of the current wireless communications paradigm, characterized by having, in most of the cases, a single hop between the end user and the point

of attachment to the network. In contrast with this, multihop extensions appear as an appropriate way of extending coverage in a quick and efficient manner, so as to serve punctual increases of traffic demand. This can be achieved either by having dedicated relaying nodes, usually deployed by the operators, or working at unlicensed bands or even by letting end users to become forwarding nodes.

In WINNER project [12], the same concept of heterogeneous relay node is proposed. A heterogeneous relay node is a network element that is wirelessly connected to another relay node or a BS by means of a given radio access technology, and it serves another relay node or a UT using a different radio access technology. Figure 8 illustrates the scenario with a heterogeneous relay node, in which a subscriber can connect to both RAN1 BS and RAN2 BS through the relay node.

There are a number of interest issues and potential solutions with regards to the realization of MRMH networks.

(1) Multi-Hop ARQ as a unified error recovery protocol spanning over the complete multi-hop route may be described in terms of a two-stage error recovery process with respect to different radio access technologies.

(2) A special issue that needs to be addressed is that of different Layer 2 segmentation sizes per hop in cases where different RATs are used along the multihop route. This causes a problem that no common sequence numbering scheme can be used along the route.

(3) The capacity of a multi-hop route is typically determined by the bottleneck hop or "weakest link." Therefore, it is not realistic to have more data in flight on the multi-hop route than being required for utilizing the bottleneck capacity (or some anticipated variations thereof). A further advantage of a common multi-hop ARQ layer is that a bottleneck node can use a flow control mechanism in order to avoid extensive data buffering. This reduces the amount of data that needs to be recovered in cases where the route changes. To facilitate the prioritization of certain types of packets (e.g., ARQ signaling), a priority-based queuing discipline is required.

(4) MRMH can be combined with MRTD. Henceforth two-route selection mechanisms can be identified: one addresses the problem within the route (i.e., at the relay nodes) and another addresses it from the edge-nodes of the network (i.e., infrastructure nodes or user terminals).

3.3. QoS Mapping. Providing a seamless and adptive QoS in a heterogeneous network is a key issue. The research work of QoS has been mainly in the context of individual system, and much less process has been in addressing the issue of QoS guarantee in the heterogeneous networks. In [13], the author proposes a QoS framework integrating a three-plane network infrastructure and a unified terminal cross-layer adaptation platform for heterogeneous environment.

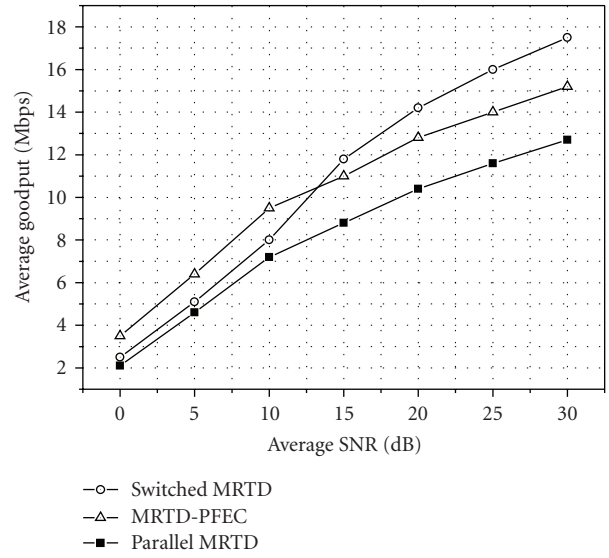


Figure 7: Expected goodput versus average SNR.

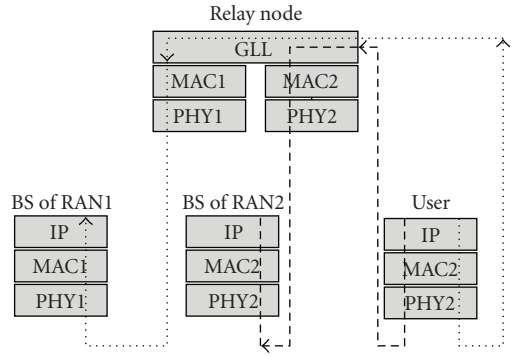


Figure 8: Heterogeneous relay.

However, there are no research results considering the end-to-end QoS guarantee over multiradio multi-hop link, so here we give a possible solution based on QoS mapping mechanism.

QoS mapping is usually referred for cross-layer of the protocol stack [14], herein which is needed to translate QoS guarantee provided by the next hop into their effects on the previous hop (sender), in order to give a unified evaluation of QoS capability of the end-to-end link. We can illustrate mapping process with some preliminary results related to segmentation and reassembly.

Considering a specific multiradio multi-hop scenario showed by Figure 9, there is a relay node connecting RAT-1 BS and RAT-2 UT. When the downlink RAT-1 MAC PDUs (denoted as RAT-1 PDU) pass through the relay node, each of them will be processed through the General Link Layer in relay node. In GLL, RAT-1PDUs will be segmented and reassembled in several RAT-2 PDUs, then the overall packet losses and delay are determined not only by RAT-1 link but also by RAT-2 link. In the following, the packet loss probability in the second hop with consideration of

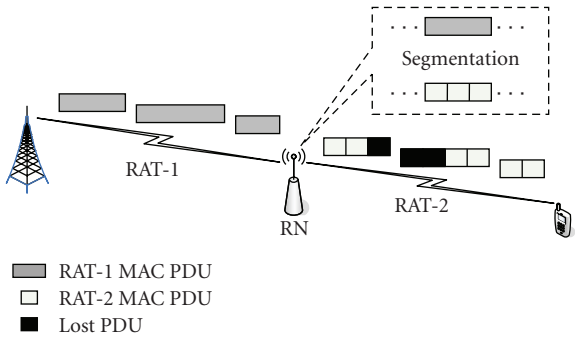


Figure 9: Segmentation and reassembly in MRMH link.

segmentation and reassembly are derived and mapped to the first link or sender (BS).

For simplicity, assuming a RAT-1 PDU can be divided into $N(N > 1)$ RAT-2 PDUs, which are labeled from 0 to $N - 1$. The loss probability of the ith RAT-2 PDU is independent of others, defined as p_i. Then, we can obtain the probability of successful transmission, which contains transmission of N RAT-2 PDUs and indicates the successful transmission probability of the corresponding RAT-1 PDU in the second hop

$$P_W = 1 - \prod_{i=0}^{N-1} 1 - p_i. \tag{8}$$

Because of the related fading characters of wireless channel, the loss of each PDU is relative with the previous PDU, a Markovian model is adopted where the probability of a packet loss depends only on whether the previous packet was also lost. Let M_i represent the event that the ith RAT-2 PDU is lost, then we have

$$P[M_i \mid M_{i-1}] = \alpha p, \tag{9}$$

$$P\left[M_i \mid \overline{M_{i-1}}\right] = p, \tag{10}$$

where $\alpha > 1$ and $0 < \alpha p < 1$, and α represents the relativity of the channel conditions in the time intervals for the transmission of two continuous RAT-2 PDUs. The larger α is, the more similar the channel conditions are.

Then, the probability of the ith RAT-2 PDU delivering correctly can be calculated as

$$\begin{aligned} P\left[\overline{M_i}\right] &= P\left[\overline{M_i} \mid M_{i-1}\right] P[M_{i-1}] + P\left[\overline{M_i} \mid \overline{M_{i-1}}\right] P\left[\overline{M_{i-1}}\right] \\ &= (1 - \alpha p) P[M_{i-1}] + (1 - p) P\left[\overline{M_{i-1}}\right]. \end{aligned} \tag{11}$$

The steady-state probability that a RAT-2 PDU is delivered correctly can be derived from (9), denoted by β,

$$\beta = \frac{1 - \alpha p}{1 + p - \alpha p}. \tag{12}$$

Finally, according to (6), the overall packet loss probability can be obtained

$$P_W = 1 - \beta^N. \tag{13}$$

That is the capability of packet losses provided by the second hop, which is actually the effect on the previous hop. The investigation of delay mapping can be analyzed in a similar way. Based on the result of QoS mapping, the unified expression of QoS capability through a multiradio multi-hop link is achieved, which can be used in resource allocation and scheduling for specific service to provide a better QoS guarantee, which is beyond the scope of this paper.

4. Multiradio Resource Management

To use the radio resources efficently in a multiaccess network, it is important to provide optimum radio resource management functionalities between the different RATs in the RAN. MRRM can operate at system, session, and flow level. At the system level, MRRM performs, for example, spectrum, load, and congestion control across two or more RAs. At the session level, MRRM coordinates decisions on different associated flows, where MRRM operations can be triggered either by system level operations or directly by session/flow level events, for example, session arrivals, or MRRM works through the establishment and maintenance of different RA.

The MRRM concept is divided into two logical parts on the basis of already existing intrinsic RRM functions. (1) RA coordination functions: the scope of these generic functions spans over the available RAs and typically includes functions such as dynamic RA addition and removal, inter-MRRM communication, RA selection, inter-RA handover, congestion control, load sharing, adaptation of the allocated resources in a coordinated manner across several available RAs, and so forth. (2) Network-complementing RRM functions: these technology-specific functions are particularly designed for one or more RAT(s). However, these functions do not replace the existing RRM functions of RAT(s) but rather complement them. These functions may provide missing, or complement inadequate RRM functions of an underlying RAT, for example, providing admission control, congestion control, intra-RAT handover. They are responsible for the RAT-specific interaction of the RA coordination functions and act as an adaptation function towards the network-intrinsic RRM functions. Hence, they appropriately translate format/terminology or commands into supporting effective interaction.

A nonexhaustive list of the most important RRM issues in multiradio access networks will be presented as follows.

4.1. Radio Access Network (RAN) Selection. Future devices can incorporate more than one access method to enjoy the seamless and variable services. The technological solutions should be transparent to the end user and one automatic

means of evaluating the optimum choice to satisfy a set of services. Therefore, one of the principle research challenges involved in heterogeneous networks is the network selection to determine the appropriate radio accesses from those available RAs for the users. A perfect RA selection scheme should not only benefit from being able to access his/her subscribed services anywhere and anytime with high QoS and less cost, but also can improve overall efficiency of spectrum utilization.

At present, many researches aiming at this issue have been done, and they have put forward some fundamental algorithms for the heterogeneous systems. In the traditional methods such as [15], only the radio signal strength (RSS) threshold and hysteretic values are considered and processed in a fuzzy logic-based algorithm. However, in such a multiradio access environment, the traditional algorithm is not sufficient to make a handoff decision, since they do not take the current context or user's preference into account. When considering more handoff decision factors, a number of two-dimension cost functions such as [16] are developed. In one dimension, the function reflects the types of services requested by the user; while in the second dimension, it represents the cost of the network according to specific parameters. However, this method is not flexible for variable scenario, and the considered factor is not enough to describe the requirements in the RA selection process.

Herein we propose an optimized cost function-based RA selection algorithm. The purpose of the RAN selection algorithm is to optimize a predefined cost function including minimizing the consumed resources and/or "minimal price" for the session, guarantee the required QoS, and increase the overall spectrum efficiency. The algorithm is flexible for many scenarios by through parameters weight regulation. The implementation of the algorithm can be divided into three stages (depicted as Figure 10): Trigger and information collection, parameters processing, and RA selection.

In the first stage, the selection process will be triggered by several conditions, such as a new service generated, user profiles changed, or a new available access point detected. Next, some parameters used in the RAN selection procedure are collected. These parameters consist of radio propagation conditions, load situation in each RAN, required QoS level by the application, achievable level of QoS per RAN, consumed resources the corresponding charge per RAN, and so forth. In this scheme these parameters can be divided into two parts.

In the second stage, it is to calculate the weights of each parameter in the predefined cost function. The weight factors reflect the dominances of the particular requirements with respect to the user. AHP [17] as a mathematical-based technology to analyze complex problems and assist in finding the best solution by synthesizing all deciding factors is adopt to derive the weights of QoS parameters on the basis of user's preference and service application. Then we should normalize these parameters. Because these parameters have different characters, the normalization of the data is performed through two methods: larger the better, or smaller the better.

Larger the better:

$$x_i^*(j) = \frac{x_i(j) - l_j}{u_j - l_j}. \tag{14}$$

Smaller the better

$$x_i^*(j) = \frac{u_j - x_i(j)}{u_j - l_j}, \tag{15}$$

where $u_j = \max\{x_1(j), x_2(j), \ldots, x_n(j)\}$, $l_j = \min\{x_1(j), x_2(j), \ldots, x_n(j)\}$.

In the last stage, based on the prepared parameters and information, the cost function can be calculated for each user-network pair. The cost function for ith user on kth radio access is predefined as

$$\begin{aligned} F(i,k) = {} & W_{\text{se}} \times \text{SE} + W_c \times \text{Cost} \\ & + W_\alpha \times \alpha + W_\beta \times \beta + W_\gamma \times \gamma, \end{aligned} \tag{16}$$

where SE is the spectrum efficiency and Cost represents the cost of a specific network per data unit. α and β are the required bit rate and BER of specific service respectively. γ is a required Grade Of Service (GOS) of a specific network. The spectrum efficiency can be configured out by this expression

$$\text{SE} = \frac{\text{ErlangsPerCell} \times \text{Bitrate} \times \text{Activity Factor}}{\text{System Bandwidth} \times \text{Cell Area}}, \tag{17}$$

where Activity Factor is the weight attributed to different service. Based on the results, the K_i network with the maximum value of the cost function will be selected for ith user to access

$$K_i = \arg\max_k F(i,k). \tag{18}$$

Figures 11 and 12 are the simulation result. In the simulation, we compare the performance of the Resource Utilization and Percentage Of Satisfied Users between using AHP selection and Random selection algorithm. Resource Utilization can be defined as the ratio of used bandwidth and the total system bandwidth. Percentage Of Satisfied Users can be defined as the ratio of the user number which get the service which they want and the total user number. Through these two figures, it is very clear that the system performance get improvement.

4.2. Load Balancing. Balancing the load between multiple systems allows for a better utilization of the radio resource as a whole and an improvement of the systems' capacity. Many intelligent algorithms have been proposed to balance the load between different radio technologies, but few researches address the theoretical analysis for the load balance strategies. Reference [18] analyzes multiple bearer services allocation onto different subsystems in multiaccess wireless systems.

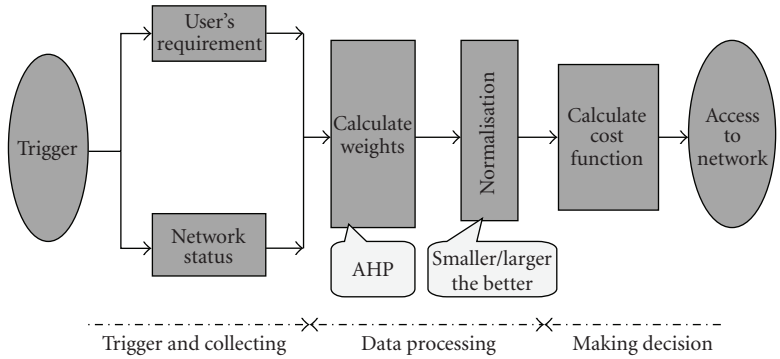


Figure 10: Cost function-based RA selection scheme.

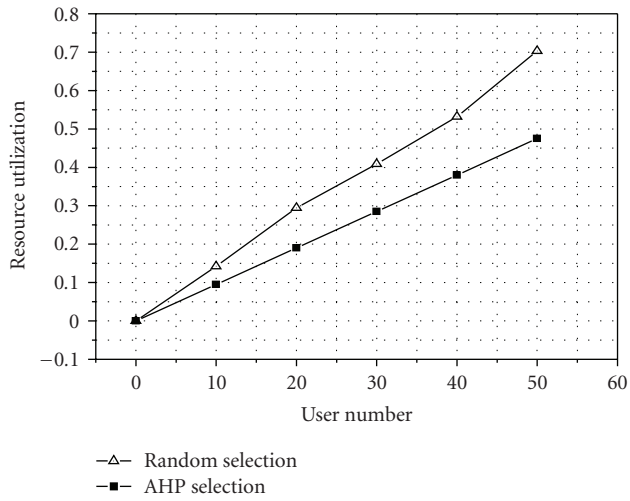


Figure 11: Resource utilization.

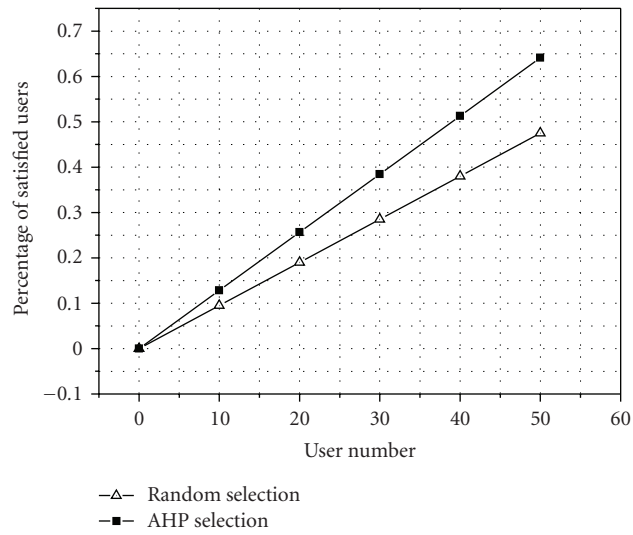


Figure 12: Percentage of satisfied users.

Considering subsystem's multi-service capacities and capacity constraints, near-optimum subsystem service allocations that maximize combined multi-service capacity are derived from simple optimization procedures. However, this work cannot be applied to give a theoretical evaluation for certain load balance strategies. In order to solve this problem, we put forward a theoretical framework, which can be used to evaluate the performance of dynamic load-balancing strategies.

In our analysis, for simplicity a scenario with two kinds of RATs overlapped is considered, we also suppose that two networks have the same capacity C, and each service utilizes the single unit of resource in TDMA. Based on certain load balance strategy, the user or service of one overloaded network or cell can be transferred to the light-load network or cell. We also assume that call requests arrive according to a Poisson process and call arrival rates in RAN 1 and RAN 2 are λ_1 and λ_2, respectively, and service times in both networks are exponentially distributed with parameter μ. By applying the multidimensional Markov chain to model the load state of both the two networks, the blocking probability between the two inter-working networks can be derived in a simple way.

Assuming that $P(0,0)$ is the idle-state probability and $s(i_1;i_2)$ are the states which the two networks experienced, so the probabilities of all the states are derived and satisfied

$$\begin{aligned} &P[s(i_1 \le c;\ i_2 \le c)] \\ &\quad + P[s(i_1 > c;\ 0 \le i_2 \le 2c - i_1) \\ &\qquad \cap\ s(0 \le i_1 \le 2c - i_2;\ i_2 > c)] \\ &\quad + P[s(i_1 \le c;\ i_2 \ge 2c - i_1 + 1) \\ &\qquad \cap\ s(i_1 \ge 2c - i_2 + 1;\ i_2 \le c)] = 1. \end{aligned} \tag{19}$$

The expression of each element of the formulation will depend on the certain load balance strategy. When a "simple borrowing" load balance scheme [19] is employed, the probabilities of all the states are given as

$$
\begin{aligned}
&P[s(i_1 \le c;\ i_2 \le c)] \\
&\quad = P(0,0)\sum_{i_1=0}^{c}\sum_{i_2}^{c}\frac{T_1^{i_1}T_2^{i_2}}{i_1!i_2!}, \\
&P[s(i_1 > c;\ 0 \le i_2 \le 2c - i_1),\ s(0 \le i_1 \le 2c - i_2;\ i_2 > c)] \\
&\quad = P(0,0)\sum_{i_1=c+1}^{2c}\sum_{i_2=0}^{2c-i_1}\frac{T_1^{i_1}T_2^{i_2} + T_2^{i_1}T_1^{i_2}}{i_1!i_2!}, \\
&P[s(i_1 > c;\ i_2 > c)] \\
&\quad = P(0,0)\frac{c^c}{c!}\sum_{i_1=c+1}^{\infty}\left(\frac{T_1}{c}\right)^{i_1} \times \frac{c^c}{c!}\sum_{i_2=c+1}^{\infty}\left(\frac{T_2}{c}\right)^{i_2}, \\
&P[s(i_1 \le c;\ i_2 \ge 2c - i_1 + 1),\ s(i_1 \ge 2c - i_2 + 1;\ i_2 \le c)] \\
&\quad = P(0,0)\sum_{i_2=0}^{c}\sum_{i_1=2c-i_2+1}^{\infty}\frac{(2c-i_2)^{2c-i_2}}{(2c-i_2)!} \\
&\qquad \times\left[\left(\frac{T_1}{2c-i_2}\right)^{i_1}\frac{T_2^{i_2}}{i_2!} + \left(\frac{T_2}{2c-i_2}\right)^{i_1}\frac{T_1^{i_2}}{i_2!}\right],
\end{aligned}
\tag{20}
$$

where $T_1 = \lambda_1/\mu$ and $T_2 = \lambda_2/\mu$ are the *traffic intensities* of networks 1 and 2, respectively, so $P(0,0)$ can be calculated from the above relation.

The call blocking probability of network i ($i = 1$), denoted by Pb_i, is given as

$$
\begin{aligned}
Pb_i = P(0,0)\Bigg\{&\frac{c^c}{c!}\sum_{i_1=c}^{\infty}\left(\frac{T_1}{c}\right)^{i_1} \times \frac{c^c}{c!}\sum_{i_2=c}^{\infty}\left(\frac{T_2}{c}\right)^{i_2} \\
&+ \sum_{i_2=0}^{c}\sum_{i_1=2c-i_2}^{\infty}\frac{(2c-i_2)^{2c-i_2}}{(2c-i_2)!}\left(\frac{T_1}{2c-i_2}\right)^{i_1}\frac{T_2^{i_2}}{i_2!} \\
&- \frac{T_1^c}{c!}\frac{T_2^c}{c!}\Bigg\}.
\end{aligned}
\tag{21}
$$

The call blocking probability of network 2 can be calculated similarly.

In contrast to interworking, the probability of one network without interworking can also be calculated as

$$
Pb_s = P(0)\frac{c^c}{c!}\sum_{i_1=c}^{\infty}\left(\frac{T_1}{c}\right)^{i_1}, \tag{22}
$$

where $P(0)$ can be derived from the following relation:

$$
P(0)\left[\sum_{i_1=0}^{c-1}\frac{T_1^{i_1}}{i_1!} + \frac{c^c}{c!}\sum_{i_1=c}^{\infty}\left(\frac{T_1}{c}\right)^{i_1}\right] = 1. \tag{23}
$$

When the two networks have the same capacity 12 ($C = 12$), Figures 13 and 14 show the blocking probability of RAN 1 in both interworking and non-interworking case, with the constant traffic intensity of RAN 2 ($T_2 = 8$, $T_2 = 10$). It may be observed that the blocking probability is eased in the interworking case, and the profit is more evident when the traffic became more heavy.

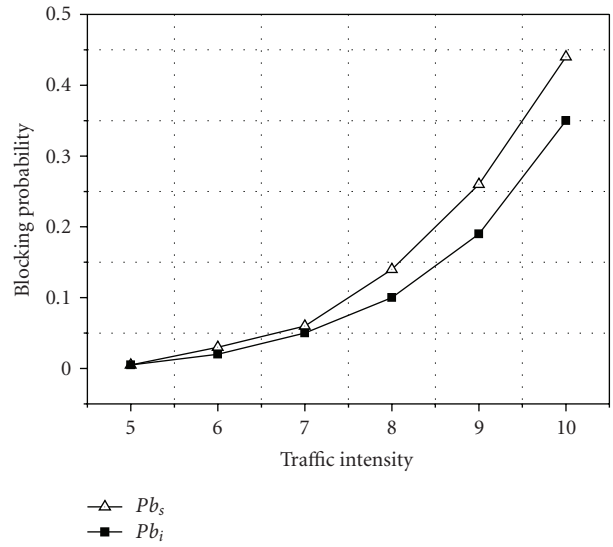


Figure 13: Blocking probability versus traffic intensity, for $C = 12$, $T_2 = 8$.

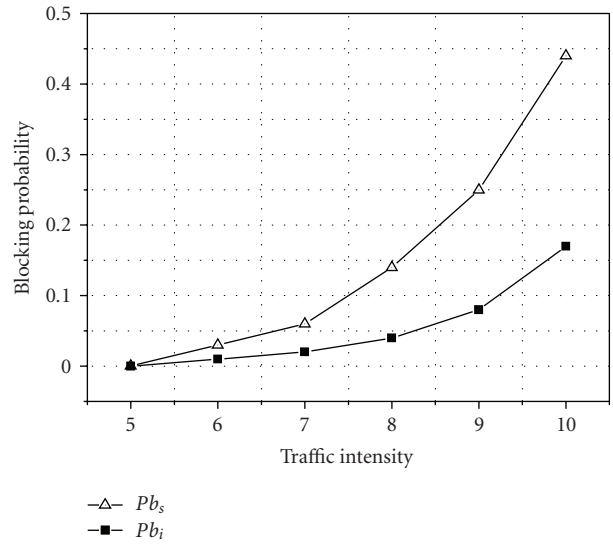


Figure 14: Call blocking probability versus traffic intensity, for $C = 12$, $T_2 = 10$.

5. Conclusion

Cooperation mechanisms between different radio access technologies in the heterogeneous network environments is one of the hot issues in the following years, which may cover most of the foundational fields of wireless communications, such as link layer protocol design, radio resource management and power saving and QoS guarantee. This paper fistly proposes two interworking network architectures to make different RATs cooperate, which makes subscribers access anywhere with the best techniques, the interworking between WiMAX and 3GPP LTE networks is taken as the specific case. Then this paper elaborates several important issues including GLL, MRRM, in order to allow efficient

cooperation between different radio access technologies. The potential state-of-the-art challenges are presented for these corresponding topics. Moreover, some solutions and mechanisms are proposed with numeric results.

References

[1] Information Society Technologies, http://cordis.europa.eu/ist/.

[2] WWI Ambient Networks, Deliverable: MRA Architecture (D2.2, Version 1.0), February 2005.

[3] 3GPP TS 23.002, "Network Archetecture (Release 8)," December 2008.

[4] 3GPP TR 25.897, "Feasibility on the Evolution of UTRAN Architecture," Release6, v0.3.1, August 2003.

[5] 3GPP TR 25.882, "3GPP System Architecture Evolution: Report on Technical Options and Conclusions," Release7, v1.0.0, March 2006.

[6] J. Sachs, L. Muñoz, R. Aguero, et al., "Future wireless communication based on multi-radio access," in *Proceedings of the 11th Meeting of the Wireless World Research Forum (WWRF '04)*, Oslo, Norway, June 2004.

[7] WiMAX Forum, "WiMAX Forum Network Architeture (stage 3)," Rel.1, January 2008.

[8] K. Dimou, R. Agero, et al., "Generic link layer: a solution for multi-radio transmission diversity in communication networks beyond 3G," in *Proceedings of the 54th IEEE Vehicular Technology Conference (VTC '05)*, 2005.

[9] G. P. Koudouridis, H. R. Karimi, and K. Dimou, "Switched multi-radio transmission diversity in future access networks," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '05)*, 2005.

[10] R. Veronesi, "Multiuser scheduling with multi radio access selection," in *Proceedings of the 2nd International Symposium on Wireless Communication Systems*, pp. 455–459, 2005.

[11] L. Rizzo, "Effective erasure codes for reliable computer communication protocols," *ACM Computer Communication Review*, vol. 27, no. 2, pp. 24–36, 1997.

[12] IST WINNER project, http://www.ist-winner.org/.

[13] X. Gao, G. Wu, and T. Miki, "End-to-end QoS provisioning in mobile heterogeneous networks," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 24–34, 2004.

[14] L. A. DaSilva, "QoS mapping along the protocol stack: discussion and preliminary results," in *Proceedings of the IEEE International Conference on Communications (ICC '00)*, vol. 2, pp. 713–717, June 2000.

[15] N. D. Tripathi, J. H. Reed, and H. F. Vanlandingham, "Adaptive handoff algorithms for cellular overlay systems using fuzzy logic," in *Proceedings of the IEEE 49th Vehicular Technology Conference (VTC '99)*, vol. 2, pp. 1413–1418, Houston, Tex, USA, May 1999.

[16] H. S. Park, S. H. Yoon, T. H. Kim, J. S. Park, M. S. Do, and J. Y. Lee, "Vertical handoff procedure and algorithm between IEEE802.11 WLAN and CDMA cellular network," in *Proceedings of the 7th International Conference on Mobile Communications (CDMA '03)*, Lecture Notes in Computer Science, pp. 103–112, Seoul, South Korea, November 2003.

[17] T. L. Saaty, *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, RWS, Pittsburgh, Pa, USA, 2000.

[18] A. Furuskär, "Allocation of multiple services in multiaccess wireless systems," in *Proceedings of the IEEE Mobile and Wireless Communication Networks (MWCN '02)*, pp. 261–265, September 2002.

[19] T. J. Kahwa and N. D. Georganas, "A hybrid channel assignment scheme in large-scale, cellular-structured mobile communication systems," *IEEE Transactions on Communications*, vol. 26, no. 4, pp. 432–438, 1978.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 628657, 9 pages
doi:10.1155/2010/628657

Research Article

A Multistandard Frequency Offset Synchronization Scheme for 802.11n, 802.16d, LTE, and DVB-T/H Systems

Javier González-Bayón,[1] Carlos Carreras,[1] and Ove Edfors[2]

[1] *Departamento de Ingeniería Electrónica, E.T.S.I. Telecomunicación, Universidad Politécnica Madrid, 28040 Madrid, Spain*
[2] *Department of Electrical and Information Technology, Lund University, 22100 Lund, Sweden*

Correspondence should be addressed to Javier González-Bayón, javier@die.upm.es

Received 21 October 2009; Accepted 24 December 2009

Academic Editor: Francisco Falcone

Carrier frequency offset (CFO) synchronization is a crucial issue in the implementation of orthogonal frequency division multiplexing (OFDM) systems. Since current technology tends to implement different standards in the same wireless device, a common frequency synchronization structure is desirable. Knowledge of the physical frame and performance and cost system requirements are needed to choose the most suitable scheme. This paper analyzes the performance and FPGA resource requirements of several data-aided (DA) and decision-directed (DD) schemes for four wireless standards: 802.11n, 802.16d, LTE, and DVB-T/H. Performance results of the different methods are shown as BER plots and their resource requirements are evaluated in terms of the number of computations and operators that are needed for each scheme. As a result, a common architecture for the four standards is proposed. It improves the overall performance of the best of the schemes when the four standards are considered while reducing the required resources by 50%.

1. Introduction

OFDM has been the focus of a wide variety of studies in wireless communication systems because of its high transmission capability and its robustness to the effects of frequency-selective multipath channels. Several existing and upcoming standards, among them are WiFi 802.11n [1], WiMAX 802.16d [2], LTE [3], and DVB-T/H [4, 5], are based on the OFDM concept. It is expected that several of them will coexist and, in many cases, operate concurrently on the same wireless terminal. This opens up for receiver/transmitter algorithm design where the basic algorithm structure is shared between the different OFDM-based standards, allowing for both efficient implementations and efficient use of resources on a common baseband processing platform. Several approaches to multistandard solutions can be found in the literature [6–9], but none of them deals with the synchronization problem in detail.

It is well known that OFDM systems are more sensitive to an offset in the carrier frequency than single carrier schemes at the same bit rate. This CFO causes loss of orthogonality of the multiplexed signals creating intercarrier interference (ICI) and introducing a constant increment in the phase of the samples.

Frequency synchronization is often performed in two phases: acquisition and tracking. At the start of the sequence the acquisition stage is used to perform a first estimation of the CFO of the signal [10–14]. In a circuit-switched system the acquisition phase can be fairly long since it only represents a small percentage of the total transmitted sequence. Some systems like LTE, DVB, and cellular systems are circuitswitched. In packet-switched systems, as 802.16d and 802.11n, the acquisition phase is more important since the transmission sequences are short. The most common approach in such systems is to use a preamble for acquisition. As it will be shown, the acquisition stage is a well-defined task that can be easily adapted to all standards being considered. Therefore, the paper focuses specially on the tracking stage.

After acquisition, the problem of tracking has to be solved. Since acquisition is never performed perfectly and conditions are not static in a real system, there still remains a residual CFO that needs to be corrected. The tracking stage

can be non-data-aided [15], when no extra information is included in the transmitted data (as in DD methods) or data aided [12, 16], when periodically transmitted training symbols and/or known pilot subcarriers are used.

In this paper, different frequency synchronization schemes are evaluated for the addressed standards with an explicit aim to reuse as much as possible the algorithm structure when switching between standards because of the limited resources available in the target architecture. Therefore, algorithm and architectural design are approached together from the beginning of the design flow. In this study, FPGAs have been selected as target architecture for these systems because of their support for reconfigurability, parallelism, and increased performance over software-based (e.g., DSP) solutions.

The main contributions of this paper are as following:

(1) detailed performance analysis of CFO synchronization schemes (mainstream and alternative) for four current wireless communications standards,

(2) comparative evaluation of their computational requirements,

(3) proposal of feasible architectures for multistandard devices.

The paper is structured as follows. The OFDM signal and the different standard frames are introduced in Sections 2 and 3. The acquisition and the different tracking schemes are presented in Sections 4 and 5. BER results for the different standards are given in Section 6. Implementation issues are considered in Section 7. Finally, Section 8 concludes the paper.

2. The OFDM Signal

The baseband scheme of a digitally implemented OFDM transmission system with CFO correction enabled is provided in Figure 1. Considering an OFDM system, the data source emits symbols (d_i) which belong to a BPSK, QPSK, 16-QAM, or 64-QAM constellation and are assumed to be equiprobable and statistically independent. The sequence d_i is serial to parallel converted into blocks of N symbols ($d_{k,l}$ denotes the k_{th} symbol of l_{th} block where $k = 0, \ldots, N-1$, and $l = -\infty, \ldots, +\infty$). These blocks are generated with period $T_s = T + T_g$ (T: useful period, T_g: guard interval). After the inverse FFT (IFFT) is applied to each block with period T_s, a cyclic prefix (CP) is inserted by prefixing the resulting N samples ($s'_{n,l}, k = 0, \ldots, N-1$) with a replica of the last N_g samples. Thus, each block is made of $N_s = N + N_g$ samples called an "OFDM symbol".

Since the carrier frequency difference between the transmitter and the receiver Δf can be modeled as a time-variant phase offset, $e^{j2\pi\Delta f t}$, the received OFDM signal can be represented as

$$r(t) = e^{j2\pi\Delta f t} s(t) * h(t,\tau) + w(t), \tag{1}$$

where $w(t)$ is the additive white Gaussian noise (AWGN), $s(t)$ is the transmitted baseband OFDM signal, $h(t,\tau)$ is the channel impulse response with τ being the delay spread, and "$*$" denotes linear convolution.

Assuming that $r(t)$ is sampled at the transmit interval T with perfect timing, the samples blocked for the lth FFT are

$$r_{n,l} = r\Big[\big(n + N_g + lN_s\big)T\Big], \quad 0 \le k < N, \ -\infty < l < +\infty. \tag{2}$$

The resulting samples from the FFT obtained in (2) are [17]

$$c_{k,l} = e^{j\pi((N-1)/N)\varepsilon} e^{j2\pi((lN_s+N_g)/N)\varepsilon} \frac{\sin(\pi\varepsilon)}{N\sin(\pi\varepsilon/N)} H_{k,l} d_{k,l} + \mathrm{ICI}_{k,l} + W_{k,l}, \quad 0 \le k < N, \ -\infty < l < +\infty, \tag{3}$$

where $\varepsilon = \Delta f T$ is the CFO normalized with respect to the subcarrier spacing. Likewise, $H_{k,l}$ is the channel coefficient on the k_{th} subcarrier with the assumption that the channel is stationary during at least one symbol, $\mathrm{ICI}_{k,l}$ is the intercarrier interference noise due to loss of orthogonality and, $W_{k,l}$ is a zero-mean stationary complex process. The first term is the data value $d_{k,l}$ modified by the channel transfer function, experiencing an amplitude reduction and phase shift due to the frequency offset.

3. The Standard Frames

The IEEE 802.11n standard is the latest in the 802.11 family. It adds extra functionality and provides better spectral efficiency. High data rates are achieved through space division multiplexing and multiple-input-multiple-output (MIMO) antenna configurations, though this paper will focus on the single input and output (SISO) antenna case. This standard defines a physical layer that can use 64 or 128 subcarriers with local oscillator frequencies of 2.4 GHz or 5 GHz. Also, it can operate in three modes: legacy, high throughput, and mixed. This paper focuses on the mixed and legacy modes where the preamble is composed of repeated patterns in the time domain called short training field (STF) and long training field (LTF) and other signal field preambles, as illustrated in Figure 2. The correlation properties of STF and LTF allow CFO estimation in the acquisition stage. Also, 802.11n allocates a number of boosted pilot subcarriers (4 or 6) in the data symbols for channel estimation and synchronization purposes.

The IEEE 802.16d standard (also known as fixed WiMAX) defines a physical layer that uses 256 subcarriers which are modulated with BPSK, QPSK, 16-QAM, or 64-QAM constellations. The transmission according to IEEE 802.16 is done in bursts, similarly to 802.11n. The WiMAX OFDM preamble is defined differently for uplink and downlink communications [2]. In both cases, the time domain signal of the preamble has a repeated pattern. The long preamble, used for downlink, consists of two symbols: a 4×64 pattern symbol, where a 64-sample pattern is repeated 4 times, and a 2×128 pattern symbol with two repetitions of a 128-sample pattern. The uplink uses a short preamble with just a 2×128 pattern symbol. This work will focus on the uplink frame. Eight boosted subcarriers are allocated for pilot signals and a number of the highest and lowest

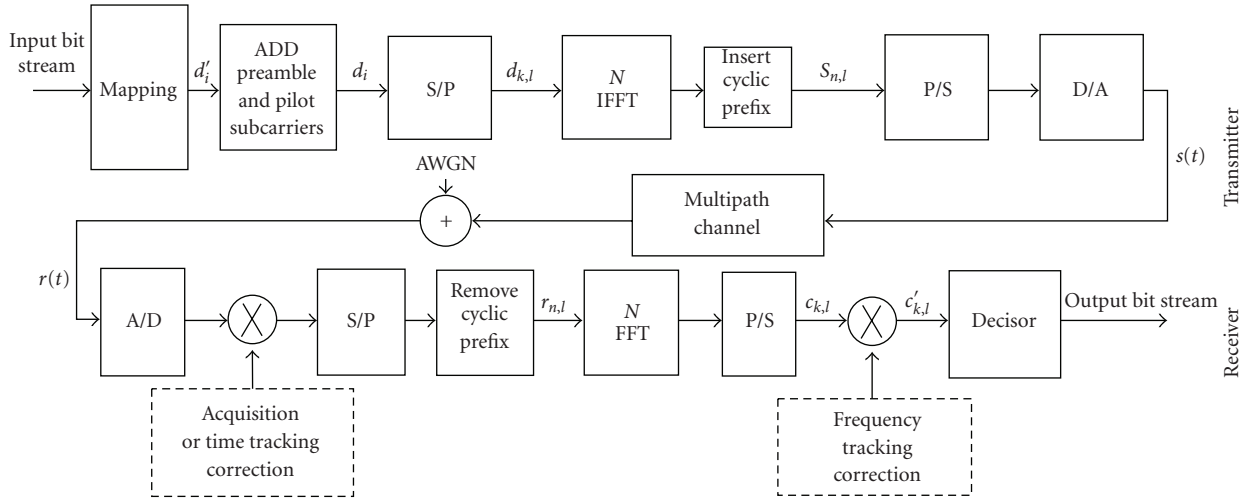


Figure 1: OFDM block diagram.

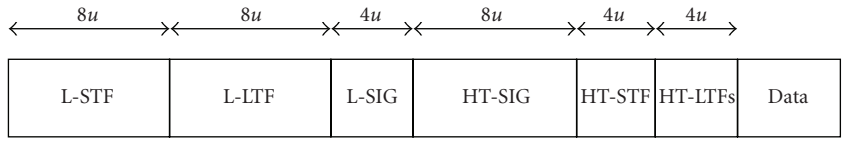


Figure 2: Preamble for 802.11n mixed mode.

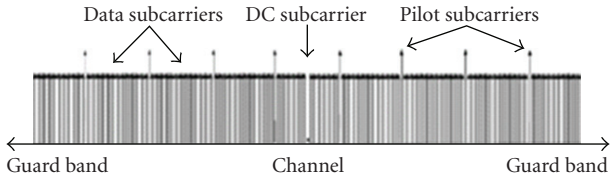


Figure 3: Frequency domain for 802.16d.

frequency subcarriers are null. The shape of the WiMAX OFDM signal in the frequency domain is shown in Figure 3.

LTE is a project belonging to the Third Generation Partnership Project (3GPP) to improve the Universal Mobile Telecommunications Systems (UMTS) and to cope with future communications requirements. LTE uses OFDM in the downlink which results in high spectral efficiency. It is also designed to be flexible in the channel allocation. In contrast to packet-oriented networks, LTE does not include a preamble to facilitate timing and frequency synchronization. Instead, pilot subcarriers are embedded in the frame as shown in Figure 4. In the normal mode, pilot subcarriers are transmitted every six subcarriers during the first and fifth OFDM symbols of each slot. This paper deals exclusively with the Frequency Division Duplex (FDD) mode defined in the standard.

Systems using DVB standards focus on digital television and data services. Even though the DVB-T standard is prepared for mobile reception, there are some factors that have to be considered when the end device is running under limited power constraints. This was the major motivation to develop a new broadcast standard aimed for handheld devices. This standard is denoted as DVB-H. It contains two major additions to the DVB-T standard, namely, time slicing and a new mode of operation called 4K. However, the physical frame has the same structure as in DVB-T. Therefore, similar synchronization schemes can be performed for both standards. DVB-H specifies three possible OFDM modes (2K, 4K, and 8K). As with LTE, DVB-T/H does not include a preamble for timing and frequency synchronization purposes. It defines dedicated synchronization subcarriers embedded into the OFDM data stream: continual (periodicity in the time domain) and scattered pilot subcarriers (periodicity in the frequency domain). Both continual and scattered pilots are transmitted at a boosted power level and their position can be observed in Figure 5.

In order to choose a suitable frequency synchronization scheme, special attention must be paid to the reference OFDM symbols and pilot subcarriers. In 802.11n and 802.16d there is a preamble amended at the beginning of the frame, whereas in LTE and DVB-T/H there is no preamble. Therefore, correlation properties introduced by the CP should be used in the acquisition stage for these two standards. Continual pilot subcarriers are defined in 802.11n, 802.16d, and DVB-T/H but LTE only includes pilot subcarriers at some specific OFDM symbols. Thus, data-aided tracking performance would perform better in 802.11n, 802.16d, and DVB-T/H than in LTE if the pilots are used for tracking purposes. From these observations, it seems that using a decision-directed algorithm in the

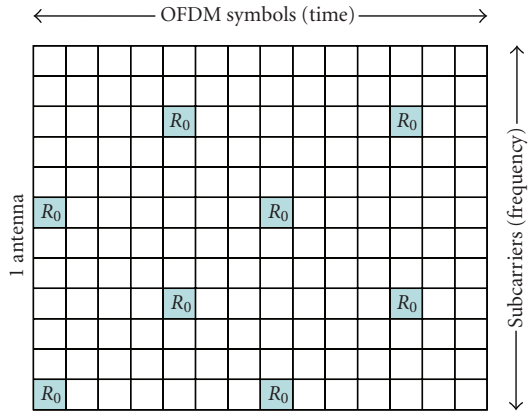


Figure 4: Reference pilots in LTE.

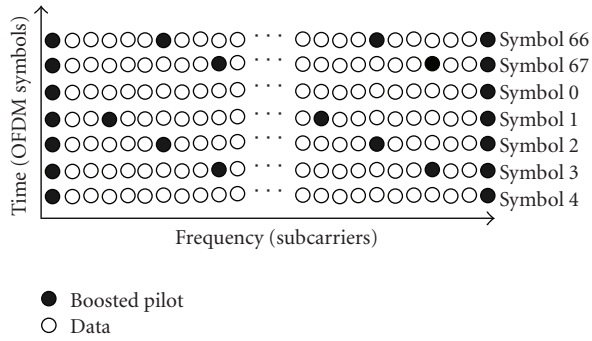


Figure 5: DVB-T/H pilot structure.

tracking stage would lead to a more homogeneous approach in a multistandard system.

4. CFO Acquisition Schemes

Most of the solutions for acquisition use the aid of pilot symbols, which are assumed to be known at the receiver. An alternative technique is to use the redundant information included in the CP [10]. Furthermore, CFO acquisition can be divided in two steps as explained in [11, 12]. In the first step, the fractional part of the CFO is estimated and corrected, allowing for the integer part of the CFO to be estimated and corrected in the second step.

The 802.11n and 802.16d standards include a preamble at the beginning of the frame. This preamble has an OFDM symbol with a repeated pattern in the time domain. The Moose algorithm [13] can be used to perform the fractional acquisition stage by using this symbol. Let there be L complex samples in each half of the training symbol, and let the correlation parts be

$$P = \sum_{m=0}^{L-1} \left(r_m^* r_{m+L}\right). \tag{4}$$

Considering the LTF symbol, for example, where the first half is identical to the second one (in time order), except for a phase shift caused by the carrier frequency offset, then the normalized frequency offset estimate is

$$\hat{\phi} = \text{angle}(P). \tag{5}$$

Subcarrier spacing for 802.11n is 312.5 KHz. Assuming a 25 ppm local oscillator and a carrier frequency of 2.4 GHz, the signal can experience a CFO of less than ± 0.6 times the subcarrier spacing. Thus, the integer estimation of the CFO can be avoided. Similar calculations and conclusions can be obtained for 802.16d.

LTE does not include a preamble in its frame, so a blind method should be used to accomplish CFO acquisition. Subcarrier spacing in LTE systems is 15 KHz; thus, normalized CFO can be higher than one. According to [12], first the fractional part of the CFO can be estimated by using the CP allocated in the OFDM symbol as shown in (4) and (5), where r_m and r_{m+L} are now the cyclic prefix and its copy, and $L = N$. After that, integer estimation can be performed in the frequency domain by using a modification of the algorithm described in [12]:

$$x_k = cp_{l,k} \cdot p_{l,k}, \tag{6}$$

$$\hat{n}_I = \arg\max_{m \in I} \left| \sum_{k \in cp+m} x_k \right|, \tag{7}$$

where $cp_{l,k}$ are the received pilot subcarriers inserted in the lth OFDM symbol, $p_{l,k}$ are the known values of the pilot subcarriers, and I is determined from $[-n_{\max}, n_{\max}]$. Due to the LTE pilot subcarrier structure, $n_{\max} = 5$. By using the known values of the pilot subcarriers in (7), the integer part of the CFO can be calculated using only the first OFDM symbol ($l = 1$).

The DVB-T/H frame does not include a preamble and it also has pilot subcarriers in the first OFDM symbol likewise LTE, so a similar approach to LTE acquisition can be used. The main difference between integer estimation in LTE and DVB-T/H is the length of the cyclic prefix and the number of pilot subcarriers that can vary depending on the transmission mode, thus increasing or decreasing the CFO estimation performance and its computational complexity.

It can be concluded that the same algorithm (4) and (5) can be applied in the four standards for fractional CFO acquisition by using the CP or the available preamble, whereas a similar method (6) and (7) can be used for integer acquisition in LTE and DVB-T/H where it is needed. Since algorithm reuse can be accomplished easily in the acquisition stage, the rest of the paper will focus on the tracking stage.

5. CFO Tracking Schemes

After acquisition, there still remains a little variation in the residual CFO. If that variation is not tracked and corrected, constellation points will fall in a different quadrant after a number of OFDM symbols, thus significantly degrading the system performance. For example, a residual CFO = 0.02 introduces a subcarrier rotation of 22° after three OFDM

symbols for a DVB 2K mode with CP = 64 and QPSK constellation. Thus, accuracy and speed of convergence are important when implementing the CFO tracking closed loop. Although, this residual CFO also introduces ICI, it can be considered negligible in most cases, depending on the conditions and specifications. Therefore, the tracking effort should be aimed at correcting CFO rotation.

It should be mentioned that for DVB-T/H, channel estimation and equalization could be performed during all the data transmission by using the continual pilot subcarriers. This equalization would also correct partially the residual CFO rotation. However, even for this standard a residual CFO tracking scheme is highly recommended [12]. The CFO tracking scheme will be more critical for packet-switched systems, as 802.16d and 802.11n, where channel estimation is performed only at the beginning of the frame by using the preamble.

The so-called decision-directed methods (non-data-aided methods) compare the received data subcarriers with sliced versions (as fed from the demapper) to give a larger number of estimates. The Decision-Directed Time-Frequency Loop (DD-TFL) proposed in [15] for CFO tracking in the 802.11g standard is based on two feedback loops in the time and the frequency domain and it uses all the data subcarriers to perform the estimations. Adaptations of this scheme for the 802.16d standard are found in [16] where the Decision-Directed Frequency Loop (DD-FL) and Data-Aided Frequency Loop (DA-FL) schemes are presented. DD-FL avoids the use of the time loop and uses less number of subcarriers per symbol to perform the tracking stage. By using DA-FL, the pilot subcarriers inserted in the data stream are used instead of the data subcarriers to perform the CFO estimations. DA-FL and DD-FL aim at reducing the CFO tracking computational complexity with almost no performance penalty. Other CFO tracking methods can be found in the literature as the classical scheme presented in [12]. However this DA tracking scheme requires pilot subcarriers in two consecutive OFDM symbols and this condition is not met by LTE. Therefore, this method is not considered in this work.

DA-FL, DD-FL, and DD-TFL can be adapted to other standard frames. The 802.11n, 802.16d, and DVB-T/H frames include pilot subcarriers in every OFDM symbol, whereas LTE includes pilot subcarriers in some specific symbols. Therefore, DA-FL performance is expected to worsen for this standard.

The DA-FL scheme [16] uses pilot subcarriers inserted in the OFDM data symbols. Its structure is represented in Figure 6.

The sequence $c_{k,l}$ after the FFT at the receiver is modified at every subcarrier as

$$c'_{k,l} = c_{k,l} e^{-j\Psi_{k,l}}, \quad 0 \le k \le N. \tag{8}$$

The corrected data symbols $c'_{k,l}$ may then be demapped to a bit stream. In the phase error detector (PED), the subcarrier pilots, $p_{k,l}$, are used for extracting the error increment $E_{k,l}$ according to one of the algorithms proposed in [18]. In particular, the algorithm selected here to extract the error increment computes

$$\begin{aligned} e^{I}_{k,l} &= \operatorname{imag}(p_{k,l}) - \operatorname{imag}\left(p'_{k,l}\right), \\ e^{QI}_{k,l} &= \operatorname{real}(p_{k,l}) - \operatorname{real}\left(p'_{k,l}\right), \\ E_{k,l} &= e^{Q}_{k,l} \operatorname{sgn}(\operatorname{real}(p_{k,l})) - e^{I}_{k,l} \operatorname{sgn}(\operatorname{imag}(p_{k,l})), \end{aligned} \tag{9}$$

where $p'_{k,l}$ are the known values of the pilot subcarriers and sgn() is the sign function. After error extraction, the error increment $E_{k,l}$ is attenuated and enters the filter directly. Then, the estimated phase error $\Psi_{k,l}$ is applied to the post-FFT data symbol $c_{k,l}$. Therefore, CFO correction is updated as many times as pilot subcarriers are inserted in the OFDM symbol. Since this scheme performs correction in the frequency domain, it corrects the phase rotation and not the ICI introduced by the CFO. An important point to remark is that by using algorithm described in (9, 10, 11) no complex multiplications are needed. This is an important improvement over classical tracking schemes as in [12].

The structure of the DD-FL scheme [16] is represented in Figure 7. This scheme also uses the error extraction algorithm described by (9). These equations are adapted to a decision-directed scheme by substituting the pilot subcarriers data ($p_{k,l}$) by the data samples, and the known value of the pilot subcarriers ($p'_{k,l}$) by the samples at the output of the decisor.

The DD-TFL scheme [15] is composed of two tracking loops as it can be observed in Figure 8. The frequency loop uses the information provided by the output of the decisor to build the tracking system. In the time loop, the error $E_{k,l}$ estimated by the decision-directed phase error detector (DD-PED) is fed to the time branch and is averaged before entering the filter. As a result, the pre-FFT sample $r_{n,l}$ is rotated as

$$r'_{k,l} = r_{k,l} e^{-j(n+Ng+lNs)\Psi_l}, \quad 0 \le n \le N. \tag{10}$$

This time branch is able to correct the ICI introduced by the residual CFO; thus, a better performance is expected when compared to DD-FL.

These tracking schemes can be used on the four standards. The two DD methods can use all or some of the available data subcarriers to perform the tracking. In this work, all data subcarriers are used for 802.11n, eight data subcarriers are used for 802.16d, every 6th subcarrier is used for LTE, and every 38th subcarrier is used for DVB-H/D. By choosing these values, simulations provide meaningful results and simulation times are not prohibitive. The DA method uses all the pilot subcarriers available in the frame.

6. BER Results

BER results for the complete synchronization system are obtained for each standard. A Rayleigh channel consisting of two paths is considered. The channel is perfectly estimated at the receiver and it is corrected using zero-forcing equalization. There is no coding of the QPSK signal, so performance

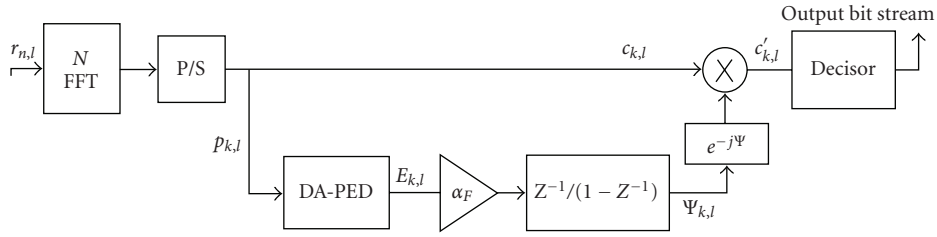


Figure 6: DA-FL scheme.

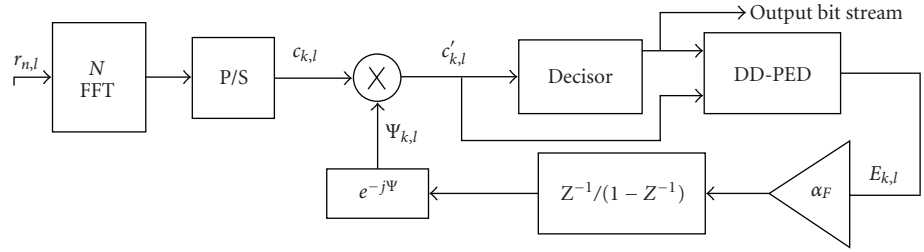


Figure 7: DD-FL scheme.

of the different schemes is shown through raw BER values. It is assumed that timing synchronization is perfectly achieved. The BER values are calculated by averaging the error bits throughout 10000 frames.

The 802.11n frame is simulated considering a system with a 64-point FFT and four pilots per symbol. The CP is composed of 16 samples. Each frame is composed of 100 OFDM symbols and the normalized CFO introduced in the system is 0.6. Similar length frame and normalized CFO are used for 802.16d. This standard requires a 256-point FFT and $N_g = 32$ is used. In the case of LTE, the frame is composed of 140 OFDM symbols with an FFT size of 512, $N_g = 64$ and CFO = 2.7. Finally, in the case of DVB-H, the frame is composed of 40 OFDM symbols with a 4048-point FFT, $N_g = 128$, and a normalized frequency offset of 2.7. Table 1 summarizes the chosen parameters for the different standards. First of all, some previous simulations were performed to find the appropriate attenuation (α_T, α_F) of the filters of the loops. Table 2 collects the values finally selected. Once the optimum attenuation values for the different schemes and standards were found, the BER results were obtained for a system where both CFO acquisition and tracking were enabled. Acquisition was performed for each standard as explained in Section 4, whereas three different tracking schemes (DA-FL, DD-FL, and DD-TFL) were evaluated for each standard.

Figure 9 shows the BER results for 802.11n. DA-FL obtains the best response and, for low noise values, DD-FL and DD-TFL approximate to the offset free case as well. This is because DD schemes rely on hits in the decisor block to work correctly. Hence, when noise decreases and less errors occur at the decisor, DD performance increases.

Figure 10 displays the results for 802.16d. The DA-FL scheme improves the BER obtained by the DD schemes. In a similar way to 802.11n, the DD schemes approximate to DA-FL performance when the noise decreases. It is possible to improve DD performance in this case by increasing the number of data subcarriers used in the tracking estimation. However, this would also increase the computational requirements.

Figure 11 shows the plot for LTE. It can be observed that DA-FL performance is unacceptable, while DD schemes obtain BER values close to the offset free case. This is because there are no pilots inserted in every OFDM symbol, so tracking convergence is not fast enough for DA-FL. Thus, this standard encourages the use of DD methods. As it was

Table 1: Parameters for the different standards.

	802.11n	802.16d	LTE	DVB-T/H
N_{FFT}	64	256	512	4048
N_g	16	32	64	128
Ts (us)	4	72	83	448
CFO	0.6	0.6	2.7	2.7
Pilot subcarriers per OFDM symbol	4	8	50	89
Data subcarriers in DD schemes	48	8	50	89
Frame length (OFDM symbols)	100	100	140	40

Table 2: Optimal loop parameters.

	DA-FL	DD-FL	DD-TFL	
802.11n	$\alpha_F = 7 \times 10^{-2}$	$\alpha_T = 5 \times 10^{-5}$	$\alpha_F = 5 \times 10^{-5}$	$\alpha_T = 10^{-3}$
802.16d	$\alpha_F = 2 \times 10^{-4}$	$\alpha_T = 10^{-4}$	$\alpha_F = 10^{-4}$	$\alpha_T = 10^{-5}$
LTE	$\alpha_F = 10^{-4}$	$\alpha_T = 10^{-5}$	$\alpha_F = 10^{-5}$	$\alpha_T = 10^{-3}$
DVB-T/H	$\alpha_F = 10^{-2}$	$\alpha_T = 10^{-2}$	$\alpha_F = 10^{-2}$	$\alpha_T = 10^{-3}$

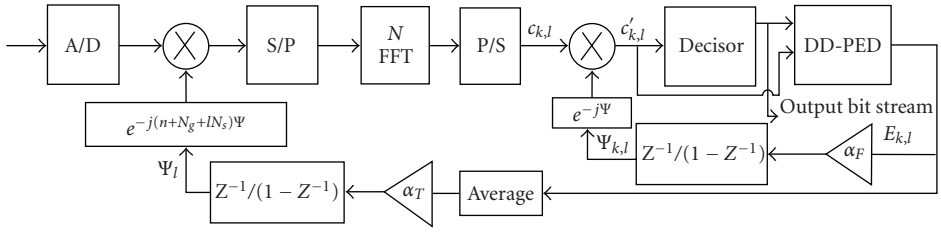


Figure 8: DD-TFL scheme.

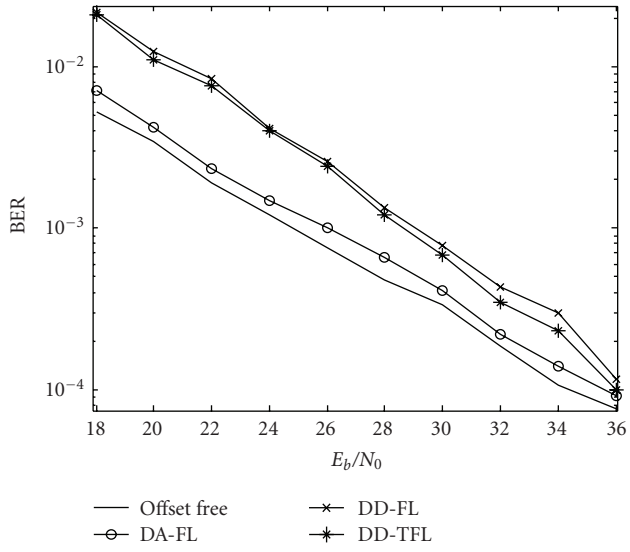


Figure 9: BER values for 802.11n.

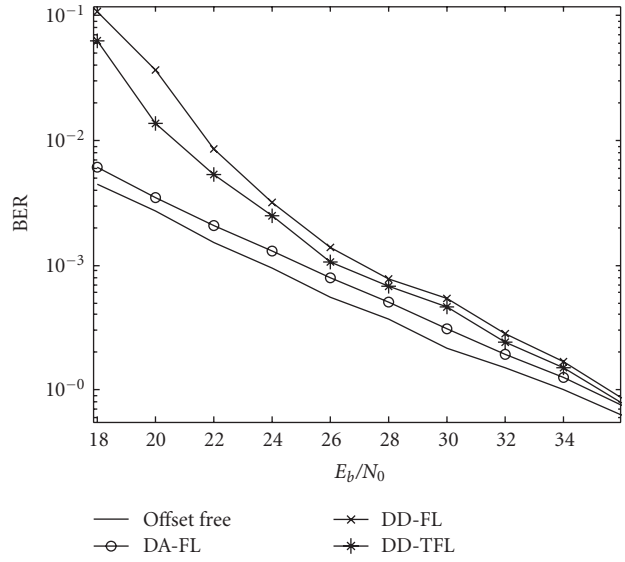


Figure 10: BER values for 802.16d.

expected, DD-TFL behaves better than DD-FL although the difference is small.

Figure 12 displays the results for DVB-T/H. The DA-FL scheme clearly outperforms the DD schemes. That is due to the "small" number of data subcarriers used for CFO tracking. It is possible to improve the DD performance, similarly to 802.16d and LTE by increasing the number of data subcarriers and the computational complexity.

Therefore, from the previous performance results it can be concluded that DD-TFL is the best option for a common implementation for the three standards since it improves slightly the DD-FL performance and DA-TL has an unacceptable performance for LTE.

7. Implementation Issues

The BER performance of the different schemes has been shown in Section 6. However, there still remains an important issue that needs to be considered for implementation purposes: their computational complexity. This is a key issue when determining the number of hardware resources needed for portable, battery-powered systems. Computations are described in terms of real multiplications (M), additions (S), and multiplications by a constant (MC). A complex multiplication is implemented using 3 M and 5 S. CFO correction is implemented through a complex multiplication. On the other hand, the required FPGA resources are described in terms of embedded multipliers and adders (EM/A).

Table 3: Number of operations and resources.

		DA-FL	DD-FL	DD-TFL
Ops	802.11n	36/75/14	36/108/36	72/180/12
	802.16d	9/19/2	9/21/4	18/40/4
	LTE	11/20/2	11/23/4	21/42/4
	DVB-T/H	22/37/1	22/37/1	44/74/1
Res	802.11n	6/9	6/9	12/17
	802.16d	6/9	6/9	12/17
	LTE	6/9	6/9	12/17
	DVB-T/H	6/9	6/9	12/17

Table 3 describes the three synchronization schemes for each standard according to their (M/S/MC) computations, as millions of operations per second, and their required (EM/A) resources. The computations per second are calculated taking into account the operations performed by each method, including the algorithm, the filter, and the correction, and considering the bit rates defined in the standards. The required resources are obtained by scheduling the operations involved assuming that they are performed iteratively subcarrier by subcarrier. No other sharing of resources has been considered in the architecture.

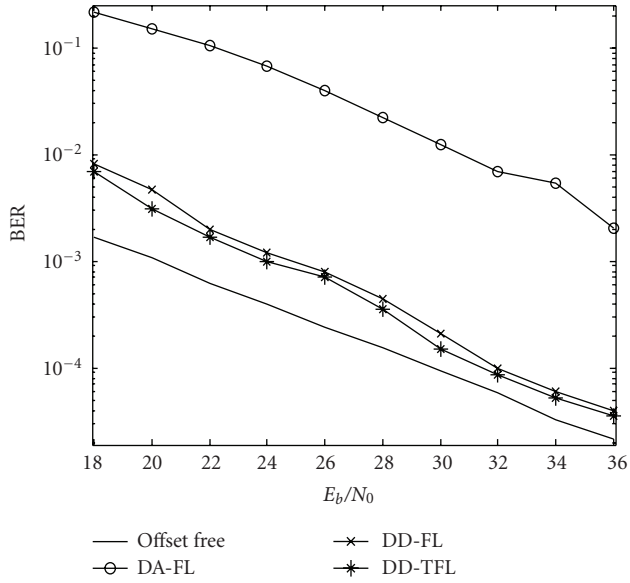


Figure 11: BER values for LTE.

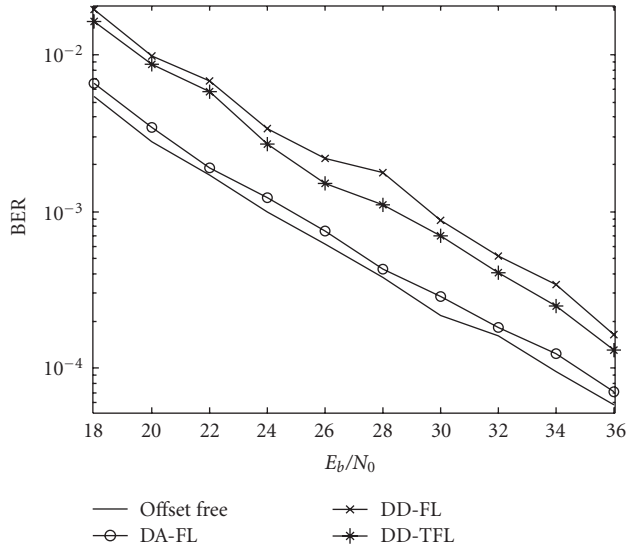


Figure 12: BER values for DVB-T/H.

It can be observed that DA-FL and DD-FL need less than a half of the number of operations required by DD-TFL. Therefore, DD-TFL not only would require more resources, but also would consume more power. In this framework, a new analysis of the results obtained in Section 6 reveals that the advantage of DD-TFL over DD-FL can be considered negligible. It is also important to note that DD-FL and DD-TFL will increase or reduce their computations (and also their performance) depending on the actual number of data subcarriers in the OFDM symbol. Therefore, when considering computational requirements in addition to performance, it turns out that the best alternative is DD-FL.

Nevertheless, an even better solution can be found by looking at the structure of the three tracking schemes. Since DA-FL and DD-FL use the same estimation algorithm, both schemes can be implemented using the same resources and work for the four different frames (DA-FL for 802.11n, 802.16d, and DVB-T/H, and DD-FL for LTE). To accomplish that, only two memories with the number and position of the pilot or data subcarriers involved in the tracking are needed to switch between DA-FL and DD-FL. This solution also offers more possibilities to reuse the EMs available in the FPGA.

Table 4: Features of the three solutions.

	DD-TFL	DD-FL	DA-FL & DD-FL
EM (% total)	18%	9%	9%
RE (% time)	84%	2% to 59%	2-3%
dB losses	1.6 to 3	1.8 to 3.6	0.5 to 1.8

Table 4 summarizes the three possible multistandard solutions considering that the target device is a Virtex 4 xc4vlx60 which contains 66 EM. For each solution, it includes the percentage of EMs used in the FPGA, the resource utilization (RE) described as a percentage of the total time, and the range of signal losses in dB for a target BER = 10^{-4}. In the case of resource utilization, the percentages are obtained from the ratio of the subcarriers that are being used to calculate the CFO estimates with respect to the total number of subcarriers available in each OFDM symbol. These percentages somehow describe the possibilities of further resource reuse. Some values in the table are given as ranges that include the results for the four standards being evaluated. For example, the DD-FL solution allows a 2% resource utilization for DVB-T/H, 3% for 802.16d, 10% for LTE, and 59% for 802.11n.

8. Conclusions

In this work, a comparison of different frequency synchronization schemes for four wireless communications standards (802.11n, 802.16d, LTE, and DVB-T/H) has been presented, aimed at a multistandard FPGA implementation. Focus is on the tracking stage, as acquisition is performed using the same algorithm for 802.11n, 802.16d, LTE, and DVB-T/H. In the case of 802.11n and 802.16d, only fractional CFO acquisition is performed over the preamble.

Despite the frame differences between the standards, three different methods to accomplish CFO tracking have been evaluated. DA-FL performs well for 802.11n, 802.16d, and DVB-H/T. However, DA-FL performance for LTE is unacceptable due to the fact that no pilot subcarriers are inserted at each OFDM symbol. DD-TFL is the scheme with best performance for the four standards but, after analyzing the computational requirements and the possibilities of resource reuse, DD-FL appears as a more balanced solution. Furthermore, a solution that combines DA-FL for 802.11n, 802.16d and DVB-T/H standards and DD-FL for LTE by including a small additional memory to switch between standards has been proposed, showing overall better performance than DD-TFL and requiring only half of its resources.

Acknowledgment

The work presented in this paper has been supported in part by the Spanish Ministry of Science and Innovation under projects no. TEC2006-13067-C03-03 and no. TEC2009-14219-C03-02 and by the European Commission under the FP7-ICT project MULTI-BASE (216541).

References

[1] "IEEE draft standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific equirements—part 11: wireless lAN medium access control (MAC) and physical layer (PHY) specifications amendment: enhancements for higher throughput," June 2009.

[2] "IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems," IEEE 802.16, 2004.

[3] A. B. Ericsson, "Long term evolution (LTE): an introduction," White paper, October 2007.

[4] ETSI EN 300 744, "Digital video broadcasting (DVB): frame structure, channel coding and modulation for digital terrestrial television (DVB-T)," Tech. Rep., ETSI, 2004.

[5] DVB-H-Transmission Systems for Handheld Terminals- EN 302 204 v1.1.1, http://www.dvb-h.org/.

[6] C. Garuda and M. Ismail, "A multi-standard OFDM-MIMO transceiver for WLAN applications," in *Proceedings of the 48th IEEE International Midwest Symposium on Circuits and Systems*, pp. 1613–1616, Cincinnati, Ohio, USA, August 2005.

[7] B. Mennenga, J. Guo, and G. Fettweis, "A component based reconfigurable baseband architecture," in *Proceedings of the 16th IST Mobile and Wireless Communication Summit*, pp. 1–5, Budapest, Hungary, July 2007.

[8] R. Barrak, A. Ghazel, and F. Ghannouchi, "Optimized multistandard rf subsampling receiver architecture," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2901–2909, 2009.

[9] F. Gallazi, G. Torlli, P. Malcovati, and V. Ferragina, "A digital multistandard reconfigurable FIR filter for wireless applications," in *Proceedings of the 14th IEEE International Conference on Electronics, Circuits and Systems*, pp. 808–811, Marrakech, Morocco, December 2007.

[10] J.-J. van de Beek, M. Sandell, and P. O. Borjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1800–1805, 1997.

[11] T. Schmidl and D. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Transactions on Communications*, vol. 45, no. 12, pp. 1613–1621, 1997.

[12] M. Speth, S. Fechtel, G. Fock, and H. Meyr, "Optimum receiver design for OFDM-based broadband transmission—part II: a case study," *IEEE Transactions on Communications*, vol. 49, no. 4, pp. 571–578, 2001.

[13] P. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Transactions on Communication*, vol. 42, pp. 2901–2914, 1994.

[14] G. Santella, "A frequency and symbol synchronization system for OFDM signals: architecture and simulation results," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 1, pp. 254–275, 2000.

[15] L. Kuang, Z. Ni, J. Lu, and J. Zheng, "A time-frequency decision-feedback loop for carrier frequency offset tracking in OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 367–373, 2005.

[16] J. González-Bayón, C. Carreras, and A. Fernández-Herrero, "Comparative evaluation of carrier frequency offset tracking schemes for WiMAX OFDM systems," in *Proceedings of the IEEE Symposium on Signal Processing and Information Technology (ISSPIT '07)*, Cairo, Egypt, December 2007.

[17] M. Speth, S. Fechtel, G. Fock, and H. Meyr, "Optimum receiver design for wireless broadband systems using OFDM—part I," *IEEE Transactions on Communictions*, vol. 47, pp. 1668–1677, 1999.

[18] S. Moridi and H. Sri, "Analysis of four decision-feedback carrier recovery loops in the presence of intersymbol interference," *IEEE Transactions on Communictions*, vol. 33, pp. 543–550, 1985.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 279807, 12 pages
doi:10.1155/2010/279807

Research Article

Capacity Evaluation for IEEE 802.16e Mobile WiMAX

Chakchai So-In, Raj Jain, and Abdel-Karim Tamimi

Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA

Correspondence should be addressed to Chakchai So-In, cs5@cse.wustl.edu

Received 21 September 2009; Accepted 2 December 2009

Academic Editor: Rashid Saeed

We present a simple analytical method for capacity evaluation of IEEE 802.16e Mobile WiMAX networks. Various overheads that impact the capacity are explained and methods to reduce these overheads are also presented. The advantage of a simple model is that the effect of each decision and sensitivity to various parameters can be seen easily. We illustrate the model by estimating the capacity for three sample applications—Mobile TV, VoIP, and data. The analysis process helps explain various features of IEEE 802.16e Mobile WiMAX. It is shown that proper use of overhead reducing mechanisms and proper scheduling can make an order of magnitude difference in performance. This capacity evaluation method can also be used for validation of simulation models.

1. Introduction

IEEE 802.16e Mobile WiMAX is the standard [1] for broadband (high-speed) wireless access (BWA) in a metropolitan area. Many carriers all over the world have been deploying Mobile WiMAX infrastructure and equipment. For interoperability testing, several WiMAX profiles have been developed by WiMAX Forum.

The key concern of these providers is how many users they can support for various types of applications in a given environment or what value should be used for various parameters. This often requires detailed simulations and can be time consuming. In addition, studying sensitivity of the results to various input values requires multiple runs of the simulation further increasing the cost and complexity of the analysis. Therefore, in this paper we present a simple analytical method of estimating the number of users on a Mobile WiMAX system. This model has been developed for and used extensively in WiMAX Forum [2].

There are four goals of this paper. First, we want to present a simple way to compute the number of users supported for various applications. The input parameters can be easily changed allowing service providers and users to see the effect of parameter change and to study the sensitivity to various parameters. Second, we explain all the factors that affect the performance. In particular, there are several overheads. Unless steps are taken to avoid these, the performance results can be very misleading. Note that the standard specifies these overhead reduction methods; however, they are not often modeled. Third, proper scheduling can make an order of magnitude difference in the capacity since it can change the number of bursts and the associated overheads significantly. Fourth, the method can also be used to validate simulation models that can handle more sophisticated configurations.

This paper is organized as follows. In Section 2, we present an overview of Mobile WiMAX physical layer (PHY). Understanding this is important for performance modeling. In Section 3, Mobile WiMAX system and configuration parameters are discussed. The key input to any capacity planning and evaluation exercise is the workload. We present three sample workloads consisting of Mobile TV, VoIP, and data applications in Section 4. Our analysis is general and can be used for any other application workload. Section 5 explains both upper and lower layer overheads and ways to reduce those overheads. The number of users supported for the three workloads is finally presented in Section 6. It is shown that with proper scheduling, capacity can be improved significantly. Both error-free perfect channel and imperfect channel results are also presented. Finally, the conclusions are drawn in Section 7.

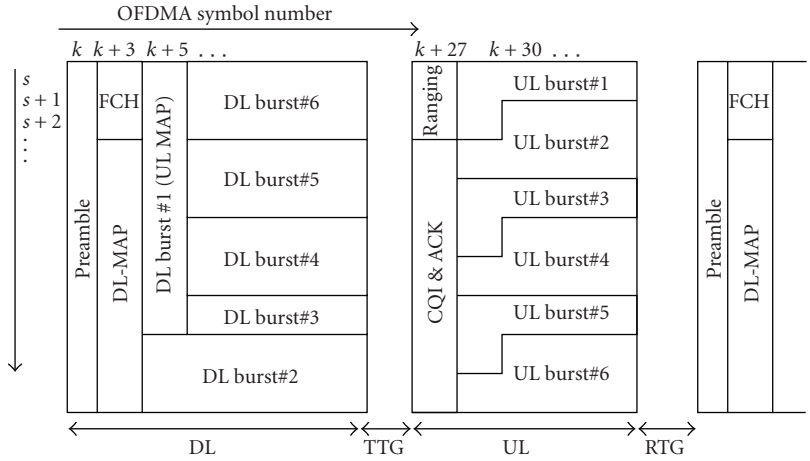


Figure 1: A Sample OFDMA frame structure.

2. Overview of Mobile WIMAX PHY

One of the key developments of the last decade in the field of wireless broadband is the practical adoption and cost effective implementation of an Orthogonal Frequency Division Multiple Access (OFDMA). Today, almost all upcoming broadband access technologies including Mobile WiMAX and its competitors use OFDMA. For performance modeling of Mobile WiMAX, it is important to understand OFDMA. Therefore, we provide a very brief explanation that helps us introduce the terms that are used later in our analysis. For further details, we refer the reader to one of several good books and survey on Mobile WiMAX [3–7].

Unlike WiFi and many cellular technologies which use fixed width channels, Mobile WiMAX allows almost any available spectrum width to be used. Allowed channel bandwidths vary from 1.25 MHz to 28 MHz. The channel is divided into many equally spaced subcarriers. For example, a 10 MHz channel is divided into 1024 subcarriers some of which are used for data transmission while others are reserved for monitoring the quality of the channel (pilot subcarriers), for providing safety zone (guard subcarriers) between the channels, or for using as a reference frequency (DC subcarrier).

The data and pilot subcarriers are modulated using one of several available MCSs (Modulation and Coding Schemes). Quadrature Phase Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM) are examples of modulation methods. Coding refers to the Forward Error Correction (FEC) bits. Thus, QAM-64 1/3 indicates an MCS with 6-bit (64 combinations) QAM modulated symbols and the error correction bits take up 2/3 of the bits leaving only 1/3 for data.

In traditional cellular networks, the downlink—Base Station (BS) to Mobile Station (MS)—and uplink (MS to BS) use different frequencies. This is called Frequency Division Duplexing (FDD). Mobile WiMAX allows not only FDD but also Time Division Duplexing (TDD) in which the downlink (DL) and uplink (UL) share the same frequency but alternate in time. The transmission consists of frames as shown in Figure 1. The DL subframe and UL subframe are separated by a TTG (Transmit to Transmit Gap) and RTG (Receive to Transmit Gap). The frames are shown in two dimensions with frequency along the vertical axis and time along the horizontal axis.

In OFDMA, each MS is allocated only a subset of the subcarriers. The available subcarriers are grouped into a few subchannels and the MS is allocated one or more subchannels for a specified number of symbols. The mapping process from logical subchannel to multiple physical subcarriers is called a permutation. Basically, there are two types of permutations: distributed and adjacent. The distributed subcarrier permutation is suitable for mobile users while adjacent permutation is for fixed (stationary) users. Of these, Partially Used Subchannelization (PUSC) is the most common used in a mobile wireless environment [3]. Others include Fully Used Subchannelization (FUSC) and Adaptive Modulation and Coding (band-AMC). In PUSC, subcarriers forming a subchannel are selected randomly from all available subcarriers. Thus, the subcarriers forming a subchannel may not be adjacent in frequency.

Users are allocated a variable number of *slots* in the downlink and uplink. The exact definition of slots depends upon the subchannelization method and on the direction of transmission (DL or UL). Figures 2 and 3 show slot formation for PUSC. In uplink (Figure 2), a slot consists of 6 *tiles* where each tile consists of 4 subcarriers over 3 symbol times. Of the 12 subcarrier-symbol combinations in a tile, 4 are used for pilot and 8 are used for data. The slot, therefore, consists of 24 subcarriers over 3 symbol times. The 24 subcarriers form a subchannel. Therefore, at 10 MHz, 1024 subcarriers form 35 UL subchannels. The slot formation in downlink is different and is shown in Figure 3. In the downlink, a slot consists of 2 clusters where each cluster consists of 14 subcarriers over 2 symbol times. Thus, a slot consists of 28 subcarriers over two symbol times. The group of 28 subcarriers is called a subchannel resulting in 30 DL subchannels from 1024 subcarriers at 10 MHz.

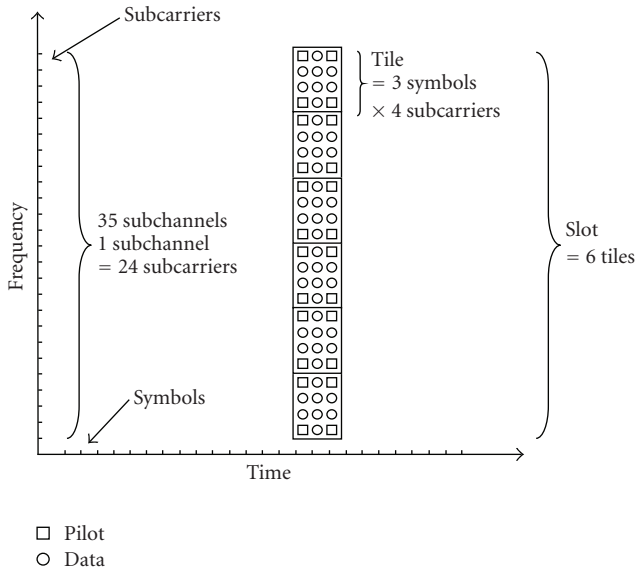


Figure 2: Symbols, tiles, and slots in uplink PUSC.

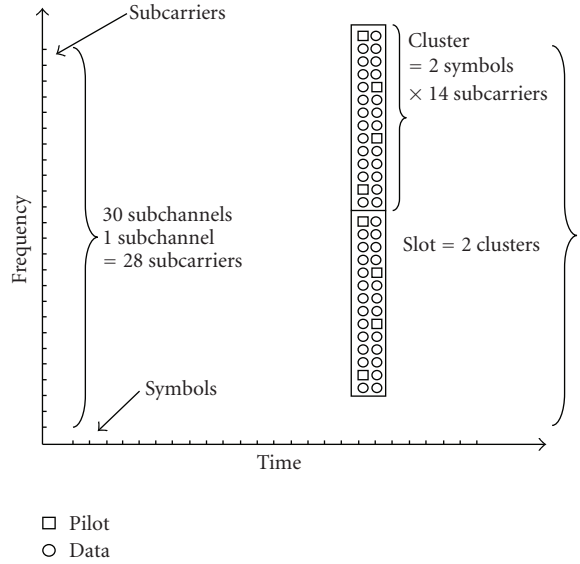


Figure 3: Symbols, clusters, and slots in downlink PUSC.

The Mobile WiMAX DL subframe, as shown in Figure 1, starts with one symbol-column of preamble. Other than preamble, all other transmissions use slots as discussed above. The first field in DL subframe after the preamble is a 24-bit Frame Control Header (FCH). For high reliability, FCH is transmitted with the most robust MCS (QPSK 1/2) and is repeated 4 times. Next field is DL-MAP which specifies the burst profile of all user bursts in the DL subframe. DL-MAP has a fixed part which is always transmitted and a variable part which depends upon the number of bursts in DL subframe. This is followed by UL-MAP which specifies the burst profile for all bursts in the UL subframe. It also consists of a fixed part and a variable part. Both DL MAP and UL MAP are transmitted using QPSK 1/2 MCS.

3. Mobile WiMAX Configuration Parameters and Characteristics

The key parameters of Mobile WiMAX PHY are summarized in Tables 1 through 3.

Table 1 lists the OFDMA parameters for various channel widths. Note that the product of subcarrier spacing and FFT size is equal to the product of channel bandwidth and sampling factor. For example, for a 10 MHz channel, 10.93 kHz×1024 = 10 MHz×28/25. This table shows that at 10 MHz the OFDMA symbol time is 102.8 microseconds and so there are 48.6 symbols in a 5 millisecond frame. Of these, 1.6 symbols are used for TTG and RTG leaving 47 symbols. If n of these are used for DL, then $47 - n$ are available for uplink. Since DL slots occupy 2 symbols and UL slots occupy 3 symbols, it is best to divide these 47 symbols such that $47 - n$ is a multiple of 3 and n is of the form $2k + 1$. For a DL : UL ratio of 2 : 1, these considerations would result in a DL subframe of 29 symbols and UL subframe of 18 symbols. In this case, the DL subframe will consist of a total of 14×30 or 420 slots. The UL subframe will consist of 6×35 or 210 slots.

Table 2 lists the number of data, pilot, and guard subcarriers for various channel widths. A PUSC subchannelization is assumed, which is the most common subchannelization [3].

Table 3 lists the number of bytes per slot for various MCS values. For each MCS, the number of bytes is equal to [number bits per symbols × Coding Rate × 48 data subcarriers and symbols per slot/8 bits]. Note that for UL, the maximum MCS level is QAM-16 2/3 [2].

This analysis method can be used for any allowed channel width, any frame duration, or any subchannelization. We assume a 10 MHz Mobile WiMAX TDD system with 5-millisecond frame duration, PUSC subchannelization mode, and a DL : UL ratio of 2 : 1. These are the default values recommended by Mobile WiMAX forum system evaluation methodology and are also common values used in practice. The number of DL and UL slots for this configuration can be computed as shown in Table 4.

4. Traffic Models and Workload Characteristics

The key input to any capacity planning exercise is the workload. In particular, all statements about number of subscribers supported assume a certain workload for the subscriber. The main problem is that workload varies widely with types of users, types of applications, and time of the day. One advantage of the simple analytical approach presented in this paper is that the workload can be easily changed and the effect of various parameters can be seen almost instantaneously. With simulation models, every change would require several hours of simulation reruns. In this section, we present three sample workloads consisting of Mobile TV, VoIP, and data applications. We use these workloads to demonstrate various steps in capacity estimation.

The VoIP workload is symmetric in that the DL data rate is equal to the UL data rate. It consists of very small packets that are generated periodically. The packet size and

Table 1: OFDMA parameters for Mobile WiMAX [3, 8, 9].

Parameters	Values						
System bandwidth (MHz)	1.25	5	10	20	3.5	7	8.75
Sampling factor	28/25				8/7		
Sampling frequency (F_s, MHz)	1.4	5.6	11.2	22.4	4	8	10
Sample time ($1/F_s$, nsec)	714	178	89	44	250	125	100
FFT size (N_{FFT})	128	512	1,024	2,048	512	1,024	1,024
Subcarrier spacing (Δf, kHz)	10.93				7.81		9.76
Useful symbol time ($T_b = 1/\Delta f$, μs)	91.4				128		102.4
Guard time ($T_g = T_b/8$, μs)	11.4				16		12.8
OFDMA symbol time ($T_s = T_b + T_g$, μs)	102.8				144		115.2

Table 2: Number of subcarriers in PUSC [8].

Parameters	Values				
(a) DL					
System bandwidth (MHz)	1.25	2.5	5	10	20
FFT size	128	N/A	512	1,024	2,084
number of guard subcarriers	43	N/A	91	183	367
number of used subcarriers	85	N/A	421	841	1,681
number of pilot subcarriers	12	N/A	60	120	240
number of data subcarriers	72	N/A	360	720	140
(b) UL					
System bandwidth (MHz)	1.25	2.5	5	10	20
FFT size	128	N/A	512	1,024	2,084
number of guard subcarriers	31	N/A	103	183	367
number of used subcarriers	97	N/A	409	841	1,681

Table 3: Slot capacity for various MCSs.

MCS	Bits per symbol	Coding Rate	DL bytesper slot	UL bytesper slot
QPSK 1/8	2	0.125	1.5	1.5
QPSK 1/4	2	0.25	3.0	3.0
QPSK 1/2	2	0.50	6.0	6.0
QPSK 3/4	2	0.75	9.0	9.0
QAM-16 1/2	4	0.50	12.0	12.0
QAM-16 2/3	4	0.67	16.0	16.0
QAM-16 3/4	4	0.75	18.0	16.0
QAM-64 1/2	6	0.60	18.0	16.0
QAM-64 2/3	6	0.67	24.0	16.0
QAM-64 3/4	6	0.75	27.0	N/A
QAM-64 5/6	6	0.83	30.0	N/A

the period depend upon the vocoder used. G723.1 Annex A is used in our analysis and results in a data rate of 5.3 kbps, 20 bytes voice packet every 30 millisecond. Note that other vocoder parameters can be also used and they are listed in Table 5.

The Mobile TV workload depends upon the quality and size of the display. In our analysis, a sample measurement on a small screen Mobile TV device produced an average packet size of 984 bytes every 30 millisecond resulting in an average data rate of 350.4 kbps [11, 12]. Note that Mobile TV workload is highly asymmetric with almost all of the traffic going downlink. Table 6 also shows other types of Mobile TV workload.

For data workload, we selected the Hypertext Transfer Protocol (HTTP) workload recommended by the 3rd Generation Partnership Project (3GPP) [13]. The parameters of HTTP workload are summarized in Table 7.

The characteristics of the three workloads are summarized in Table 8. In this table, we also include higher level headers, that is, IP, UDP, and TCP, with a header compression mechanism. Detailed explanation of PHS (Payload Header Suppression) and ROHC (Robust Header Compression) is presented in the next section. Given ROHC, the data rate with higher level headers ($R_{\text{with Header}}$) is calculated by

$$R_{\text{with Header}} = R \times \frac{(\text{MSDU} + \text{Header})}{\text{MSDU}}. \quad (1)$$

Here, MSDU is the MAC SDU size and R is the application data rate. Given the R, number of bytes per frame per user can be derived from $R_{\text{with Header}} \times$ frame_duration. For example, for Mobile TV, with 983.5 bytes of MAC SDU size and 350 kbps of application data rate, with ROHC type 1, MAC SDU size with header is 983.5 + 1 bytes and as a result, the data rate with header is 350.4 kbps and results in 216 bytes per frame.

Table 4: Mobile WiMAX system configurations.

Configurations	Downlink	Uplink
DL and UL symbols excluding preamble	28	18
Ranging, CQI, and ACK (symbols columns)	N/A	3
number of symbol columns per Cluster/Tile	2	3
number of subcarriers per Cluster/Tile	14	4
Symbols × Subcarriers per Cluster/Tile	28	12
Symbols × Data Subcarriers per Cluster/Tile	24	8
number of pilots per Cluster/Tile	4	4
number of Clusters/number Tiles per Slot	2	6
Subcarriers × Symbols per Slot	56	72
Data Subcarriers × Symbols per Slot	48	48
Data Subcarriers × Symbols per DL and UL Subframe	23,520	12,600
Number of Slots	**420**	**175**

Table 5: Vocoder parameters [10].

Vocoder	AMR	G.729A	G.711	G.723.1	
				A	B
Source bit rate (kbps)	4.5 to 12.2	8	64	5.3	6.3
Frame duration (millisecond)	20	10	10	30	30
Payload (bytes) (Active, Inactive)	(33, 7)	(20, 0)	(20, 0)	(20, 0)	(20, 0)

5. Overhead Analysis

In this section, we consider both upper and lower layer overheads in detail.

5.1. Upper Layer Overhead. Table 7 which lists the characteristics of our Mobile TV, VoIP, and data workloads includes the type of transport layer used: either Real Time Transport Protocol (RTP) or TCP. This affects the upper layer protocol overhead. RTP over UDP over IP (12 + 8 + 20) or TCP over IP (20 + 20), can result in a per packet header overhead of 40 bytes. This is significant and can severely reduce the capacity of any wireless system.

There are two ways to reduce upper layer overheads and to improve the number of supported users. These are Payload Header Suppression (PHS) and Robust Header Compression (ROHC). PHS is a Mobile WiMAX feature. It allows the sender not to send fixed portions of the headers and can reduce the 40-byte header overhead down to 3 bytes. ROHC, specified by the Internet Engineering Task Force (IETF), is another higher layer compression scheme. It can reduce the higher layer overhead to 1 to 3 bytes. In our analysis, we used ROHC-RTP packet type 0 with R-0 mode. In this mode, all RTP sequence numbers functions are known to the decompressor. This results in a net higher layer overhead of just 1 byte [5, 14, 15].

For small packet size workloads, such as VoIP, header suppression and compression can make a significant impact on the capacity. We have seen several published studies that use uncompressed headers resulting in significantly reduced performance which would not be the case in practice.

PHS or ROHC can significantly improve the capacity and should be used in any capacity planning or estimation.

Note that one option with VoIP traffic is that of silence suppression which if implemented can increase the VoIP capacity by the inverse of fraction of time the user is active (not silent). As a result in this analysis, given a silence suppression option, a number of supported users are twice as much as that without this option.

5.2. Lower Layer Overhead. In this section, we analyze the overheads at MAC and PHY layers. Basically, there is a 6-byte MAC header and optionally several 2-byte subheaders. The PHY overhead can be divided into DL overhead and UL overhead. Each of these three overheads is discussed next.

5.2.1. MAC Overhead. At MAC layer, the smallest unit is MAC protocol data unit (MPDU). As shown in Figure 4, each PDU has at least 6-bytes of MAC header and a variable length payload consisting of a number of optional subheaders, data, and an optional 4-byte Cyclic Redundancy Check (CRC). The optional subheaders include fragmentation, packing, mesh, and general subheaders. Each of these is 2 bytes long.

In addition to generic MAC PDUs, there are bandwidth request PDUs. These are 6 bytes in length. Bandwidth requests can also be piggybacked on data PDUs as a 2-byte subheader. Note that in this analysis, we do not consider the effect of polling and/or other bandwidth request mechanisms.

Table 6: Mobile TV workload parameters [12].

Applications	Format	Data rate	Notes
Mobile phone video	H.264 ASP	176 kbps	176 × 144, 20 frame per second
Smartphone video	H.264 ASP	324 kbps	320 × 240, 24 frame per second
IPTV video	H.264 Baseline	850 kbps	480 × 480, 30 frame per second
Sample video trace [11]	MPEG2	350 kbps	Average Packet Size = 984 bytes

Table 7: Web workload characteristics.

Parameters	Values
Main page size (bytes)	10,710
Embedded object size (bytes)	7,758
Number of embedded objects	5.64
Reading time (second)	30
Parsing time (second)	0.13
Request size (bytes)	350
Big packet size (bytes)	1,422
Small packet size (bytes)	498
% of big packets	76
% of small packets	24

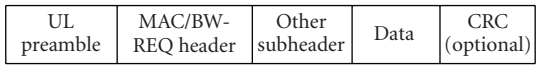


Figure 4: UL burst preamble and MAC PDU (MPDU).

Consider fragmentation and packing subheaders. As shown in Table 9, the user bytes per frame in downlink are 219, 3.5, and 9.1 bytes for Mobile TV, VoIP, and Web, respectively. In each frame, a 2-byte fragmentation subheader is needed for all types of traffic. Packing is not used for the simple scheduler used here.

However, in the enhanced scheduler, given a variation of deadline, packing multiple SDU is possible. Table 9 also shows an example when deadline is put into consideration. In this analysis, the deadlines of Mobile TV, VoIP, and Web traffic are set to 10, 60, and 250 millisecond. As a result, 437.9, 42.0, and 454.9 bytes are allocated per user. These configuration results in one 2-byte fragmentation overhead for Mobile TV and Web traffic but two 2-byte packing overheads with no fragmentation for VoIP. Table 9 also shows the detailed explanation of fragmentation and packing overheads in downlink. Note that the calculation for uplink is very similar.

5.2.2. Downlink Overhead. In DL subframe, the overhead consists of preamble, FCH, DL-MAP, and UL-MAP. The MAP entries can result in a significant amount of overhead since they are repeated 4 times. WiMAX Forum recommends using compressed MAP [3], which reduces the DL-MAP entry overhead to 11 bytes including 4 bytes for CRC [1]. The fixed UL-MAP is 6 bytes long with an optional 4-byte CRC. With a repetition code of 4 and QPSK, both fixed DL-MAP and UL-MAP take up 16 slots.

The variable part of DL-MAP consists of one entry per bursts and requires 60 bits per entry. Similarly, the variable part of UL-MAP consists of one entry per bursts and requires 52 bits per entry. These are all repeated 4 times and use only QPSK MCS. It should be pointed out that the repetition consists of repeating slots (and not bytes). Thus, both DL and UL MAPs entries also take up 16 slots each per burst.

Equation (2) show the details of UL and DL MAPs overhead computation:

$$\begin{aligned} UL\_MAP(\text{bytes}) &= \frac{48 + 52 \times \#UL\_\text{users}}{8}, \\ DL\_MAP(\text{bytes}) &= \frac{88 + 60 \times \#DL\_\text{users}}{8}, \\ DL\_MAP(\text{slots}) &= \left\lceil \frac{UL\_MAP}{S_i} \right\rceil \times r, \\ UL\_MAP(\text{slots}) &= \left\lceil \frac{DL\_MAP}{S_i} \right\rceil \times r. \end{aligned} \tag{2}$$

Here, r is the repetition factor and S_i is the slot size (bytes) given ith modulation and coding scheme. Note that basically QPSK1/2 is used for the computation of UL and DL MAPs.

5.2.3. Uplink Overhead. The UL subframe also has fixed and variable parts (see Figure 1). Ranging and contention are in the fixed portion. Their size is defined by the network administrator. These regions are allocated not in units of slots but in units of *transmission_opportunities*. For example, in CDMA initial ranging, one opportunity is 6 subchannels and 2 symbol times.

The other fixed portion is Channel Quality Indication (CQI) and ACKnowledgements (ACKs). These regions are also defined by the network administrator. Obviously, more fixed portions are allocated; less number of slots is available for the user workloads. In our analysis, we allocated three OFDMA symbol columns for all fixed regions.

Each UL burst begins with a UL preamble. Typically, one OFDMA symbol is used for short preamble and two for long preamble. In this analysis, we do not consider one short symbol (a fraction of one slot); however, users can add an appropriate size of this symbol to the analysis.

6. Pitfalls

Many Mobile WiMAX analyses ignore the overheads described in Section 5, namely, UL-MAP, DL-MAP, and MAC overheads. In this section, we show that these overheads have a significant impact on the number of users

Table 8: Summary of workload characteristics.

Parameters	Mobile TV	VoIP	Data (Web)
Types of transport layer	RTP	RTP	TCP
Average packet size (bytes)	983.5	20.0	1,200.2
Average data rate (kbps) w/o headers	350.0	5.3	14.5
UL : DL traffic ratio	0	1	0.006
Silence suppression (VoIP only)	N/A	Yes	N/A
Fraction of time user is active		0.5	
ROHC packet type	1	1	TCP
Overhead with ROHC (bytes)	1	1	8
Payload Header Suppression (PHS)	No	No	No
MAC SDU size with header	984.5	21.0	1,208.2
Data rate (kbps) after headers	350.4	5.6	14.6
Bytes/frame per user (DL)	219.0	3.5	9.1
Bytes/frame per user (UL)	0.0	3.5	0.1

Table 9: Fragmentation and packing subheaders.

Parameters	Mobile TV	VoIP	Data (Web)
Average packet size with higher level header (bytes)	984.5	21.0	1,208.2
Simple scheduler			
Bytes/5 millisecond frame per user	219.0	3.5	9.1
Number of fragmentation subheaders	1	1	1
Number of packing subheaders	0	0	0
Enhanced scheduler			
Deadline (millisecond)	10	60	250
Bytes/5 millisecond frame per user	437.9	42.0	454.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0

Table 10: Example of capacity evaluation using a simple scheduler.

Parameters	Mobile TV	VoIP	Data (Web)
MAC SDU size with header (bytes)	984.5	21.0	1,208.2
Data rate (kbps) with upper layer headers	350.4	5.6	14.6
(a) DL			
Bytes/5 millisecond frame per user (DL)	219.0	3.5	9.1
Number of fragmentation subheaders	1	1	1
Number of packing subheaders	0	0	0
DL data slots per user with MAC header + packing and fragmentation subheaders	38	2	3
Total slots per user (Data + DL-MAP IE + UL-MAP IE)	46	18	19
Number of users (DL)	9	35	33
(b) UL			
Bytes/5 millisecond frame per user (UL)	0.0	3.5	0.1
number of fragmentation subheaders	0	1	1
number of packing subheaders	0	0	0
UL data slots per user with MAC header + packing and fragmentation subheaders	0	2	2
Number of users (UL)	8	87	87
Number of users (min of UL and DL)	9	35	33
Number of users with silence suppression	**9**	**70**	**33**

supported. Since some of these overheads depend upon the number of users, the scheduler needs to be aware of this additional need while admitting and scheduling the users [4, 17]. We present two case studies. The first one assumes an error-free channel while the second extends the results to a case in which different users have different error rates due to channel conditions.

6.1. Case Study 1: Error-Free Channel. Given the user workload characteristics and the overheads discussed so far, it is straightforward to compute the system capacity for any given workload. Using the slot capacity indicated in Table 3, for various MCSs, we can compute the number of users supported.

One way to compute the number of users is simply to divide the channel capacity by the bytes required by the user payload and overhead [4]. This is shown in Table 10. The table assumes QPSK 1/2 MCS for all users. This can be repeated for other MCSs. The final results are as shown in Figure 5. The number of users supported varies from 2 to 82 depending upon the workload and the MCS.

The number of users depends upon the available capacity which depends on the MAP overhead, which in turn is determined by the number of users. To avoid this recursion, we use (3) to (5) that give a very good approximation for the number of supported users using a ceiling function:

$$\begin{aligned}\#DL\_\text{slots} = &\left\lceil \frac{DL\_MAP + CRC + \#DL\_\text{users} \times DIE}{S_i} \right\rceil \times r \\ &+ \left\lceil \frac{UL\_MAP + CRC + \#UL\_\text{users} \times UIE}{S_i} \right\rceil \times r \\ &+ \#DL_{\text{users}} \times \left\lceil \frac{D}{S_k} \right\rceil,\end{aligned} \tag{3}$$

$$\#UL\_\text{slots} = \#UL\_\text{users} \times \left\lceil \frac{D}{S_k} \right\rceil, \tag{4}$$

$$D = B + MAC_{\text{header}} + \text{Subheaders}. \tag{5}$$

Here, D is the data size (per frame) including overheads, B is the bytes per frame, and MAC_{header} is 6 bytes. Subheaders are fragmentation and packing subheaders, 2 bytes each if present. *DIE* and *UIE* are the sizes of downlink and uplink map information elements (IEs). Note that *DL_MAP* and *UL_MAP* are fixed MAP parts and also in terms of bytes. Again, r is the repetition factor and S_i is the slot size (bytes) given ith modulation and coding scheme. number *DL_slots* is the total number of DL slots without preamble and number *UL_slots* are the total number of UL slots without ranging, ACK, and CQICH.

For example, consider VoIP with QPSK 1/2 (slot size = 6 bytes) and repetition of four. Equation (3) results 35 users in the downlink. The derivation is as follows:

$$\begin{aligned}\#DL\_\text{slots} = 420 &= \left\lceil \frac{11 + 4 + \#DL\_\text{users} \times 60/8}{6} \right\rceil \times 4 \\ &+ \left\lceil \frac{6 + 4 + \#UL\_\text{users} \times 52/8}{6} \right\rceil \\ &\times 4 + \#DL_{\text{users}} \times \left\lceil \frac{11.5}{6} \right\rceil.\end{aligned} \tag{6}$$

For uplink, from (4)and (5), the number of UL users is 87:

$$\#UL\_\text{slots} = 175 = \#UL_{\text{users}} \times \left\lceil \frac{3.5 + 6 + 2}{6} \right\rceil. \tag{7}$$

Finally, after calculating the number of supported users for both DL and UL, the total number of supported users is the minimum of those two numbers. In this example, the total number of supported users is 35, (minimum of 35 and 87). In this case, the downlink is the bottleneck mostly due to the large overhead. Together with silence suppression, the absolute number of supported users can be up to $2 \times 35 = 70$ users. Figure 5 shows the number of supported users for various MCSs.

The main problem with the analysis presented above is that it assumes that every user is scheduled in every frame. Since there is a significant per burst overhead, this type of allocation will result in too much overhead and too little capacity. Also, since every packet (SDU) is fragmented, a 2-byte fragmentation subheader is added to each MAC PDU.

What we discussed above is a common pitfall. The analysis assumes a dumb scheduler. A smarter scheduler will try to aggregate payloads for each user and thus minimizing the number of bursts. We call this the enhanced scheduler. It works as follows. Given n users with any particular workload, we divide the users in k groups of n/k users each. The first group is scheduled in the first frame; the second group is scheduled in the second frame, and so on. The cycle is repeated every k frames. Of course, k should be selected to match the delay requirements of the workload.

For example, with VoIP users, a VoIP packet is generated every 30 millisecond, but assuming 60 millisecond is an acceptable delay, we can schedule a VoIP user every 12th Mobile WiMAX frame (recall that each Mobile WiMAX frame is 5 millisecond) and send two VoIP packets in one frame as compared to the previous scheduler which would send 1/6th of the VoIP packet in every frame and thereby aggravating the problem of small payloads. Two 2-byte packing headers have to be added in the MAC payload along with the two SDUs.

Table 11 shows the capacity analysis for the three workloads with QPSK 1/2 MCS and the enhanced scheduler. The results for other MCSs can be similarly computed. These results are plotted in Figure 6. Note that the number of users supported has gone up significantly. Compared to Figure 5, there is a capacity improvement by a factor of 1 to 20 depending upon the workload and the MCS.

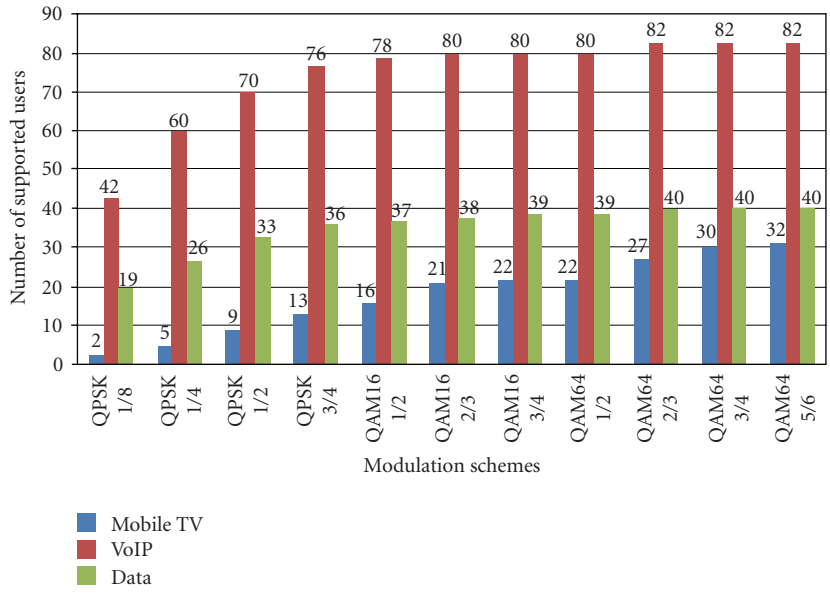


Figure 5: Number of users supported in a lossless channel (Simple scheduler).

Table 11: Example of capacity evaluation using an enhanced scheduler.

Parameters	Mobile TV	VoIP	Data (Web)
MAC SDU size with header (bytes)	984.5	21.0	1,208.2
Data rate (kbps) with upper layer headers	350.4	2.8	14.6
Deadline (millisecond)	10	60	250
(a) DL			
Bytes/5 millisecond frame per user (DL)	437.9	42.0	454.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0
DL data slots per user with MAC header + packing and fragmentation subheaders	75	9	78
Total slots per user (Data + DL-MAP IE + UL-MAP IE)	83	25	94
Number of users (DL)	10	269	233
(b) UL			
Bytes/5 millisecond frame per user (UL)	0.0	42.0	2.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0
UL data slots per user with MAC header + packing and fragmentation subheaders	0	9	2
Number of users (UL)	8	228	4350
Net number of users (min of UL and DL)	10	228	233
Number of users with silence suppression	**10**	**456**	**233**

Proper scheduling can change the capacity by an order of magnitude. Making less frequent but bigger allocations can reduce the overhead significantly.

The number of supported users for this scheduler is derived from the same equations that were used with the simple scheduler. However, the enhanced scheduler allocates as large size as possible given the deadlines. For example, for Mobile TV with a 10-millisecond deadline, instead of 219 bytes, the scheduler allocates 437.9 bytes within a single frame and for VoIP with 60-millisecond deadline, instead of 3.5 bytes per frame, it allocates 42 bytes and that results in 2 packing overheads instead of 1 fragmentation overhead.

In Table 11, the number of supported users for VoIP is 228. This number is based on the fact that 42 bytes are allocated for each user every 60 millisecond:

$$\left\lceil \frac{\#\text{slots\_subframe}}{\#\text{slots\_aggregated\_users}} \right\rceil \times \frac{\text{deadline}}{5 \text{ millisecond}} \quad (8)$$

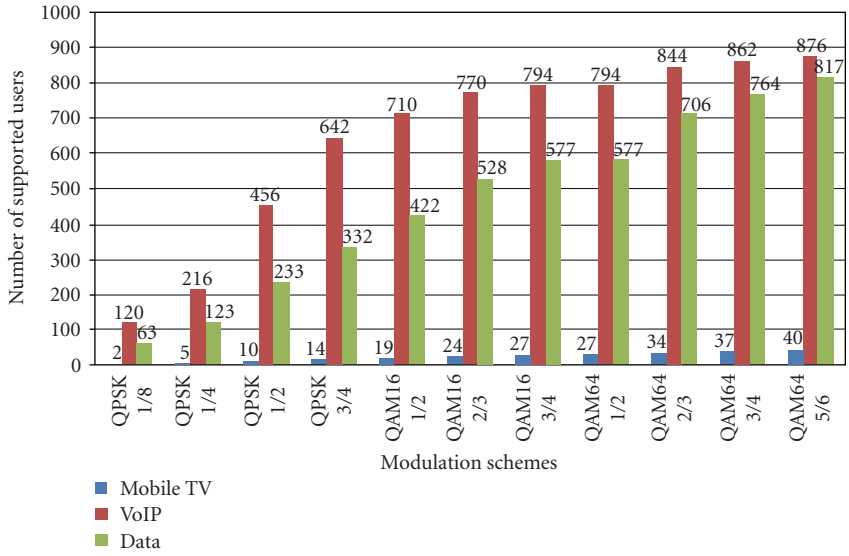


Figure 6: Number of users supported in a lossless channel (Enhanced Scheduler).

Table 12: Simulation parameters [16].

Parameters	Value
Channel model	ITU Veh-B (6 taps) 120 km/hr
Channel bandwidth	10 MHz
Frequency band	2.35 GHz
Forward Error Correction	Convolution Turbo Coding
Bit Error Rate threshold	10^{-5}
MS receiver noise figure	6.5 dB
BS antenna transmit power	35 dBm
BS receiver noise figure	4.5 dB
Path loss PL(distance)	$37 \times \log 10(\text{distance}) + 20 \times \log 10(\text{frequency}) + 43.58$
Shadowing	Log normal with $\sigma = 10$
number of sectors per cell	3
Frequency reuse	1/3

Table 13: Percent MCS for 1×1 and 2×2 antennas [16].

Average MCS	1 Antenna		2 Antenna	
	%DL	%UL	%DL	%UL
FADE	4.75	1.92	3.03	1.21
QPSK 1/8	7.06	3.54	4.06	1.68
QPSK 1/4	16.34	12.46	14.64	8.65
QPSK 1/2	15.30	20.01	13.15	14.05
QPSK 3/4	12.14	21.23	10.28	15.3
QAM-16 1/2	20.99	34.33	16.12	29.97
QAM-16 2/3	0.00	0.00	0.00	0.00
QAM-16 3/4	9.31	5.91	14.18	22.86
QAM-64 1/2	0.00	0.00	0.00	0.00
QAM-64 2/3	14.11	0.59	24.53	6.27

With the configuration in Table 11, the number of supported users is $\lceil 175/9 \rceil \times 60/5 = 228$ users. With silence suppression, the absolute number of supported users is $2 \times 228 = 456$. Note that the number of DL users is computed using (3), (4), and (5), and then (9) can be applied. The calculations for Mobile TV and Data are similar to that for VoIP.

The per-user overheads impact the downlink capacity more than the uplink capacity. The downlink subframe has DL-MAP and UL-MAP entries for all DL and UL bursts and these entries can take up a significant part of the capacity and so minimizing the number of bursts increases the capacity.

Note that there is a limit to aggregation of payloads and minimization of bursts. First, the delay requirements for the payload should be met and so a burst may have to be scheduled even if the payload size is small. In these cases, multiuser bursts in which the payload for multiple users is aggregated in one DL burst with the same MCS can help reduce the number of bursts. This is allowed by the IEEE 802.16e standards and applies only to the downlink bursts.

Table 14: Number of supported users in a lossy channel.

Workload	1 Antenna		2 Antenna	
	Simple scheduler	Enhanced scheduler	Simple scheduler	Enhanced scheduler
Mobile TV	14	16	17	20
VoIP	76	672	78	720
Data	36	369	37	438

The second consideration is that the payload cannot be aggregated beyond the frame size. For example, with QPSK 1/2, a Mobile TV application will generate enough load to fill the entire DL subframe every 10 millisecond or every 2 frames. This is much smaller than the required delay of 30 millisecond between the frames.

6.2. Case Study 2: Imperfect Channel. In Section 6.1, we saw that the aggregation has more impact on performance with higher MCSs (which allow higher capacity and hence more aggregation). However, it is not always possible to use these higher MCSs. The MCS is limited by the quality of the channel. As a result, we present a capacity analysis assuming a mix of channels with varying quality resulting in different levels of MCS for different users.

Table 12 lists the channel parameters used in a simulation by Leiba et al. [16]. They showed that under these conditions, the number of users in a cell which were able to achieve any particular MCS was as listed in Table 13. Two cases are listed: single antenna systems and two antenna systems.

Average bytes per slot in each direction can be calculated by summing the product (percentage users with an MCS $\times$ number of bytes per slot for that MCS). For 1 antenna systems this gives 10.19 bytes for the downlink and 8.86 bytes for the uplink. For 2 antenna systems, we get 12.59 bytes for the downlink and 11.73 bytes for the uplink.

Table 14 shows the number of users supported for both simple and enhanced schedulers. The results show that the enhanced scheduler still increases the number of users by an order of magnitude, especially for VoIP and data users.

7. Conclusions

In this paper, we explained how to compute the capacity of a Mobile WiMAX system and account for various overheads. We illustrated the methodology using three sample workloads consisting of Mobile TV, VoIP, and data users. Analysis such as the one presented in this paper can be easily programmed in a simple program or a spread sheet and effect of various parameters can be analyzed instantaneously. This can be used to study the sensitivity to various parameters so that parameters that have significant impact can be analyzed in detail by simulation. This analysis can also be used to validate simulations.

However, there are a few assumptions in the analysis such as the effect of bandwidth request mechanism, two-dimensional downlink mapping, and the imprecise calculation of slot-based versus bytes-based. Moreover, we do not consider (H)ARQ [18]. In addition, the number of supported users is calculated with the assumption that there is only one traffic type. Finally, fixed UL-MAP is always in the DL subframe though there is no UL traffic such as Mobile TV [4].

We showed that proper accounting of overheads is important in capacity estimation. A number of methods are available to reduce these overheads and these should be used in all deployments. In particular, robust header compression or payload header suppression and compressed MAPs are examples of methods for reducing the overhead.

Proper scheduling of user payloads can change the capacity by an order of magnitude. The users should be scheduled so that their numbers of bursts are minimized while still meeting their delay constraint. This reduces the overhead significantly particularly for small packet traffic such as VoIP.

We also showed that our analysis can be used for loss-free channel as well as for noisy channels with loss.

Acknowledgment

This work was sponsored in part by a grant from Application Working Group of WiMAX Forum. "WiMAX," "Mobile WiMAX," "Fixed WiMAX," "WiMAX Forum," "WiMAX Certified," "WiMAX Forum Certified," the WiMAX Forum logo and the WiMAX Forum Certified logo are trademarks of the WiMAX Forum.

References

[1] IEEE P802.16Rev2/D2, "DRAFT Standard for Local and metropolitan area networks," Part 16: Air Interface for Broadband Wireless Access Systems, p. 2094, December 2007.

[2] C. So-In, R. Jain, and A.-K. Tamimi, "AWG Analytical Model for Application Capacity Planning over WiMAX V0.8," WiMAX Forum, Application Working Group (AWG) Contribution, September 2009, http://cse.wustl.edu/~jain/papers/capmodel.xls.

[3] WiMAX Forum, "WiMAX System Evaluation Methodology V2.1," p. 230, July 2008, http://www.wimaxforum.org/resources/documents/technical.

[4] 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, 2009.

[5] C. Eklund, R.-B. Marks, S. Ponnuswamy, K.-L. Stanwood, and N.-V. Waes, *WirelessMAN Inside the IEEE 802.16 Standard for Wireless Metropolitan Networks*, IEEE Standards Information Network/IEEE Press, Piscataway, NJ, USA, 2006.

[6] G. Jeffrey, J. Andrews, A. Arunabha-Ghosh, and R. Muhamed, *Fundamentals of WiMAX Understanding Broadband Wireless Networking*, Prentice-Hall PTR, Upper Saddle River, NJ, USA, 2007.

[7] L. Nuaymi, *WiMAX: Technology for Broadband Wireless Access*, John Wiley & Sons, New York, NY, USA, 2007.

[8] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 wirelessMAN," *Intel Technology Journal*, vol. 8, no. 3, pp. 202–212, 2004.

[9] R. Jain, C. So-In, and A.-K. Tamimi, "System-level modeling of IEEE 802.16E mobile WiMAX networks: key issues," *IEEE Wireless Communications*, vol. 15, no. 5, pp. 73–79, 2008.

[10] R. Srinivasan, T. Papathanassiou, and S. Timiri, "Mobile WiMAX VoIP capacity system level simulations," Application Working Group, WiMAX Forum, Beaverton, Ore, USA, 2007.

[11] A.-K. Tamimi, R. Jain, and C. So-In, "SAM: simplified seasonal ARIMA model for wireless broadband access enabled mobile devices," in *Proceedings of IEEE International Symposium on Multimedia (ISM '08)*, pp. 178–183, Berkeley, Calif, USA, December 2008.

[12] D. Ozdemir and F. Retnasothie, "WiMAX capacity estimation for triple play services including mobile TV, VoIP and Internet," Application Working Group, WiMAX Forum, Beaverton, Ore, USA, 2007.

[13] 3rd Generation Partnership Project, "HTTP and FTP traffic model for 1xEV-DV simulations," *3GPP2-C50-EVAL-2001022-0xx*, 2001.

[14] G. Pelletier, K. Sandlund, L.-E. Jonsson, and M. West, "RObust Header Compression (ROHC): A Profile for TCP/IP (ROHC-TCP)," *RFC 4996*, January 2006.

[15] L.-E. Jonsson, G. Pelletier, and K. Sandlund, "Framework and four profiles: RTP, UDP, ESP and uncompressed," *RFC 3095*, July 2001.

[16] Y. Leiba, Y. Segal, Z. Hadad, and I. Kitroser, "Coverage/capacity simulations for OFDMA PHY in with ITU-T channel model," type C802.16d-03/78, IEEE, November 2004.

[17] C. So-In, R. Jain, and A.-K. Tamimi, "A deficit round robin with fragmentation scheduler for IEEE 802.16e mobile WiMAX," in *Proceedings of IEEE Sarnoff Symposium (SARNOFF '09)*, pp. 1–7, Princeton, NJ, USA, March-April 2009.

[18] A. Sayenko, O. Alanen, and T. Hamalainen, "ARQ aware scheduling for the IEEE 802.16 base station," in *Proceedings of IEEE International Conference on Communications (ICC '08)*, pp. 2667–2673, Beijing, China, May 2008.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 806279, 9 pages
doi:10.1155/2010/806279

Research Article

Effective Scheme of Channel Tracking and Estimation for Mobile WiMAX DL-PUSC System

Phuong Thi Thu Pham[1] and Tomohisa Wada[1, 2]

[1] *Information Engineering Department, Graduate School of Engineering and Science, University of The Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan*
[2] *Magna Design Net, Inc., L1831-1 Oroku, Naha-city, Okinawa 901-0155, Japan*

Correspondence should be addressed to Phuong Thi Thu Pham, thuphuong@lsi.ie.u-ryukyu.ac.jp

Received 28 September 2009; Accepted 30 November 2009

Academic Editor: Rashid Saeed

This paper introduces an effective joint scheme of channel estimation and tracking for downlink partial usage of subchannel (DL-PUSC) mode of mobile WiMAX system. Based on the pilot pattern of this particular system, some channel estimation methods including conventional interpolations and a more favorable least-squares line fitting (LSLF) technique are comparatively studied. Besides, channel estimation performance can be remarkably improved by taking advantage of channel tracking derived from the preamble symbol. System performances in terms of packet error rate (PER) and user link throughput are investigated in various channels adopted from the well-known ITU models for mobile environments. Simulation results show a significant performance enhancement when the proposed joint scheme is utilized, at least 5 dB, compared to only commonly used channel estimation approaches.

1. Introduction

Wireless metropolitan area network (Wireless MAN) or worldwide interoperability for microwave access (WiMAX), which is defined in IEEE Std. 802.16d/e [1, 2], is a new technology that provides wireless access in fixed and mobile environments. Some modes in this system utilize orthogonal frequency division multiple access (OFDMA) technique as a modulation method. This technique is adopted from the powerful orthogonal frequency division multiplexing (OFDM) which effectively mitigates the impairment of the time-variant frequency selective fading channel [3, 4]. A typical OFDM system is shown in Figure 1.

DL-PUSC, as specified in [1], is one of the multiple access modes for downlink direction which is popularly used for performance analysis. This scheme divides OFDM symbol into subchannels and assigns them to users/subscribers. Each subchannel is further partitioned into groups of 14 consecutive subcarriers called clusters. Clusters of one user are not continuously connected but are pseudorandomly permuted over OFDM symbol among users so that data of different users are treated equally over the effect of fading channels. Hence, channel estimation and equalization for recovering the original signal of each user must be performed from cluster to cluster.

Pilot-based channel estimation is widely used in OFDM transmission system. By scattering known data called pilots into OFDM symbol at the transmitter, calculating channel values at pilot positions and then interpolating the whole channel values for data subcarriers at the receiver, transmitted information can be recovered. There are many techniques reported for channel estimation; some conventional methods like linear and cubic spline [5–8] interpolations are commonly used due to their low complexity for practical implementation, yet low efficiency. Other methods like transform-domain processing [9, 10] perform better but require higher computation for executing DFT/IDFT. Theoretical optimum method like minimum mean square error estimator (MMSEE) [11–15] gives best performance but is too complicate for practical realization.

Preamble is a special OFDM symbol transmitted at the beginning of transmission frame. Since it contains lots of

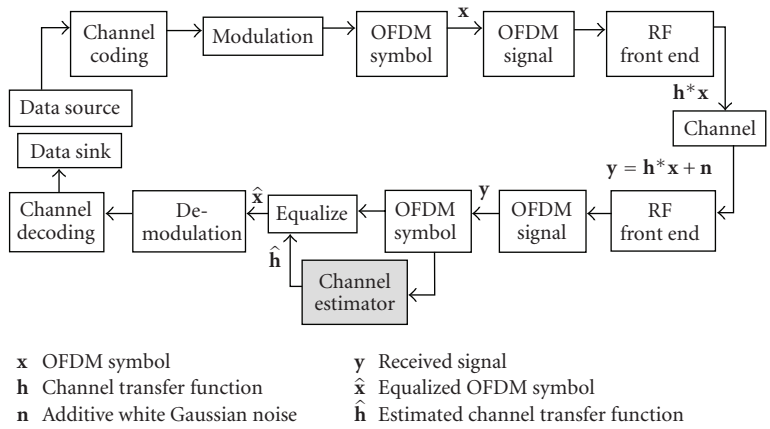


Figure 1: An overview of OFDM system.

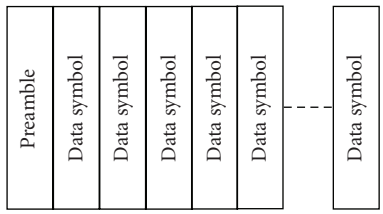


Figure 2: Basic transmission frame.

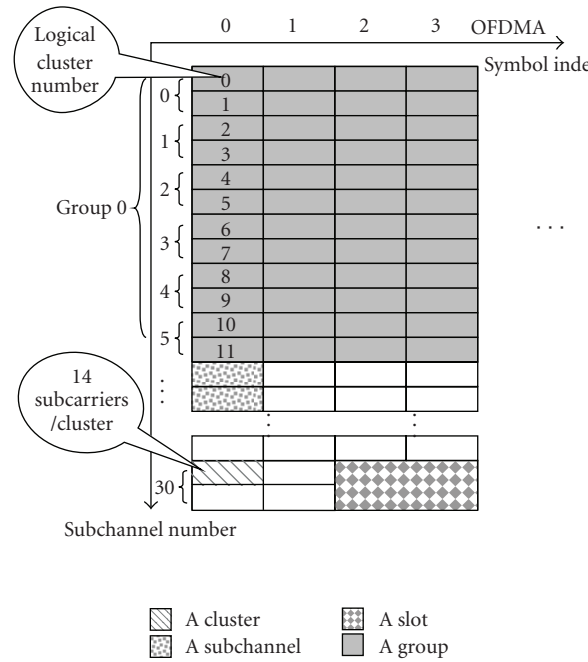


Figure 3: Basic elements in DL-PUSC mode.

pilots usually evenly distributed, channel estimation task for preamble symbol is quite easy and accurate. Moreover, as the radio channel is often slowly faded, by some tracking algorithm, the estimated channel can be exploited to enhance the performance of estimation for the subsequent data symbols.

In this paper, a new scheme of channel estimation for DL-PUSC system is proposed. This approach uses LSLF technique combined with channel tracking to form a joint channel estimation scheme. Comparisons between this new method and other conventional approaches such as linear and cubic spline interpolations by simulating a typical 1024-point FFT system profile in different ITU mobile channel models are given to show that a significant improvement in system performance can be achieved.

The following parts of this paper are organized as follows. Section 2 addresses the transmission structure and the signal model. Channel estimation methods such as the commonly used linear and cubic spline interpolations and the new method using LSLF technique are highlighted in Section 3. Joint scheme with tracking algorithm is introduced in Section 4. Simulation setup, results, and discussion are given in Section 5. Finally, Section 6 summarizes and concludes the paper.

2. System Description

2.1. Transmission Structure. A basic transmission structure is shown in Figure 2 in which a preamble symbol starts the frame and data symbols are transmitted right after. Preamble symbol is designed according to different profiles of transmission, that is, in 1024-point FFT mode, there are 284 boosted BPSK pilots, each every three subcarriers, starting from subcarrier index 86 (indexing starts from 0). Subcarriers at other positions are set to 0. Pilot values are generated by a particular Pseudo-Noise code related to IDCell and segment parameters [1].

In DL-PUSC mode, an OFDM symbol is divided into subchannels; each of those is associated to a specific user. Subchannel is further partitioned into clusters; each of which contains a group of 14 consecutive subcarriers. When transmitted, clusters of different users will be permuted among themselves; therefore, they are scattered over the OFDM symbol. The data symbol structure is shown in Figure 3.

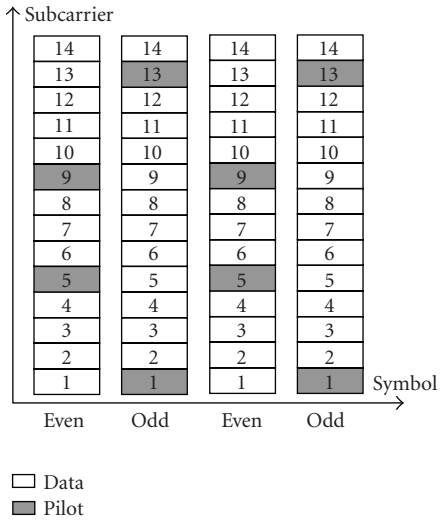


Figure 4: Pilot pattern of data symbol in a cluster.

Pilot pattern of a cluster in a DL-PUSC frame of data symbols is shown in Figure 4. Pilots are allocated at subcarrier $\{5, 9\}$ for even symbol and at subcarrier $\{1, 13\}$ for odd symbol.

2.2. Signal Model. Assume that transmitted frame has M OFDM data symbols in which $\mathbf{x}_m = (x_{0,m}, x_{1,m}, x_{2,m}, \ldots, x_{N-1,m})^T$ $0 < m < M-1$ is the symbol at time m and N is the number of subcarriers in OFDM symbol.

At the receiver, if intersymbol interference is negligible, received signal could be derived as

$$\mathbf{y}_m = \mathbf{A}\mathbf{h}_m + \mathbf{w}_m, \tag{1}$$

where $\mathbf{A}$ is the $N \times N$ diagonal matrix whose values are $\mathbf{x}_m$. The channel frequency response $\mathbf{h}_m = \mathbf{F}\mathbf{g}_m(t)$ is the DFT of the time-varying multipath fading channel impulse response $\mathbf{g}_m(t)$ of which a discrete-time version can be obtained as in [9, 10]. $\mathbf{F}$ is the $N \times L$ matrix whose entries are $f_{n,l} = (1/\sqrt{N}e^{-j2\pi(nl/N)})$ $0 \le n \le N-1$, $0 \le l \le L-1$ where L is the number of channel impulse response taps, and $\mathbf{w}_m$ is the additive white Gaussian noise.

In order to recover $\mathbf{x}_m$ from $\mathbf{y}_m$, the channel $\mathbf{h}_m$ has to be estimated by exploiting the pilots which are located at predefined positions in OFDM symbols. The least-square values of channel frequency response at for pilots are obtained by

$$h^{\mathrm{LS}}_{n_p,m} = \frac{y_{n_p,m}}{\mathrm{x}_{n_p,m}}, \tag{2}$$

where n_p denotes the pilot position (in this particular case, $n_p = \{1, 5, 9, 13\}$, $p = 0, \ldots, 3$).

The goal is to estimate all the channel values $\mathbf{h}^{\mathrm{EST}}_m$ at all data subcarriers from the values of $\{h^{\mathrm{LS}}_{n_p,m}\}$ so that $\mathbf{h}^{\mathrm{EST}}_m$ should be as much similar as possible to $\mathbf{h}_m$.

Here, we have two kinds of pilot pattern depending on whether they belong to preamble symbol or data symbol. It is obvious that the density of pilot subcarriers in preamble symbol is higher than that in data symbol. Thus, in order to estimate the whole channel to have a reference for channel tracking, conventional method like linear interpolation can be utilized to get a good tradeoff between complexity and performance. On the other hand, for data symbol, in this particular case of DL-PUSC, the interpolation task must be performed from cluster to cluster, and because each cluster contains only 14 consecutive subcarriers, the channel on each cluster can be approximated as a "line"; this fact inspires the idea of using LSLF technique to estimate the partial channel. Therefore, a comparative study is carried out to demonstrate the superiority of this approach to other commonly used methods such as linear [5, 6] and cubic spline [7, 8] interpolations.

Due to the pilot arrangement in data symbols, it is necessary to perform a two-dimension (2D) interpolation scheme or two 1D estimations in cascade. As the number of pilots in time axis is more than that in frequency axis, it is more convenient to estimate first in time and then in frequency.

3. Channel Estimation

3.1. Conventional Methods

3.1.1. Linear Interpolation. Figure 5 shows an example of the linear interpolation technique in time direction for an even number of OFDM symbols in which pilots of even symbols are located at the 5th and 9th locations while those of odd symbols are resided in the 1st and 13th places. Consider

$$h^{\mathrm{EST}}_{\{1,13\},m} = \begin{cases} h^{\mathrm{LS}}_{\{1,13\},1}, & m = 0, \\ \dfrac{h^{\mathrm{LS}}_{\{1,13\},m-1} + h^{\mathrm{LS}}_{\{1,13\},m+1}}{2} & m = 2, 4, \ldots, M-2, \end{cases}$$

$$h^{\mathrm{EST}}_{\{5,9\},m} = \begin{cases} \dfrac{h^{\mathrm{LS}}_{\{5,9\},m-1} + h^{\mathrm{LS}}_{\{5,9\},m+1}}{2}, & m = 1, 3, \ldots, M-3, \\ h^{\mathrm{LS}}_{\{5,9\},M-2}, & m = M-1. \end{cases} \tag{3}$$

Then, interpolation in frequency direction can be evaluated as

$$h^{\mathrm{EST}}_{k,m} = k\Delta + h^{\mathrm{LS/EST}}_{n_p,m}, \quad k = 1, \ldots, 4,$$
$$\Delta = \frac{h^{\mathrm{LS/EST}}_{n_{p+1},m} - h^{\mathrm{LS/EST}}_{n_p,m}}{4}, \quad p = 0, 1, 2, \tag{4}$$

where k denotes the position of channel value inside the interval of two adjacent pilots $(h^{\mathrm{LS/EST}}_{n_p,m}, h^{\mathrm{LS/EST}}_{n_{p+1},m})$.

3.1.2. Cubic Spline Interpolation. With this method, because we do not have enough pilots in frequency direction, interpolation in time has to be performed first. Moreover, it also requires having at least 8 OFDM symbols to have

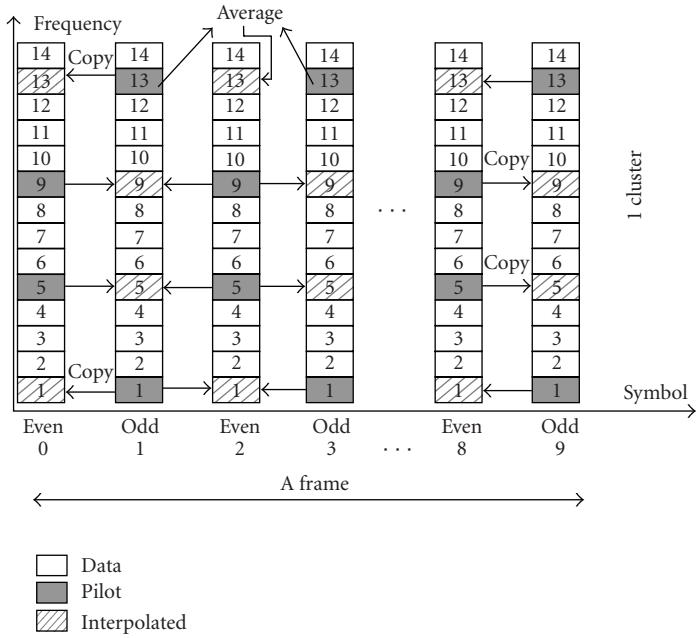


Figure 5: Linear interpolation in time.

enough pilot points. Parameters of a third-ordered equation are calculated as indicated in [8] and [16]. Then, the channel values at data positions within the appropriate interval are calculated. After interpolation in time, there are enough points to carry out this task again in frequency direction.

3.2. Channel Estimation Using LSLF. Since a cluster size is small, channel transfer function of that cluster can be considered a "line". Therefore, the least-squares line that fits all the pilots is thought to be a better approximation of the ideal channel.

3.2.1. Interpolation in Time. Suppose that an even number of OFDM symbols appears in frame and is indexed from 0 to $M-1$. In time direction, for a cluster, there are four sub-streams according to subcarrier indexes $\{1,5,9,13\}$ containing channel values at pilots. Stream 1 and stream 13 have pilots at odd locations $\{1,3,5,\ldots,M-1\}$ while stream 5 and stream 9 have pilots at even locations $\{0,2,4,\ldots,M-2\}$. Define the channel values at pilot positions vector $\mathbf{p}=\{p_k\}$ and pilot position vector $\mathbf{l}=\{l_k\}$, $(k=1,2,\ldots,M/2)$ in which $\mathbf{l}=(0,2,\ldots,M-2)$ for stream $\{5,9\}$ and $\mathbf{l}=(1,3,\ldots,M-1)$ for stream $\{1,13\}$; LSLF technique will find the pair of coefficients $\omega=\begin{pmatrix}a\\b\end{pmatrix}$ to form the line containing the set of points: $\mathbf{y}=\{y_k\}$; $y_k=al_k+b$ so that the least-squares error

$$s=\sum_{k=1}^{M/2}(p_k-y_k)^2=\sum_{k=1}^{M/2}(p_k-al_k-b)^2 \tag{5}$$

is minimized. That means to find a pair of coefficients $\{a,b\}$ so that they minimize s and so vanish the partial derivatives $(\partial s/\partial a)$ and $(\partial s/\partial b)$. Therefore, from [17], problem turns into solving this system of equations

$$\begin{aligned}&\frac{\partial s}{\partial a}=0,\\&\frac{\partial s}{\partial b}=0\end{aligned} \tag{6}$$

$$\Longleftrightarrow\begin{cases}a=\dfrac{(M/2)\sum_{k=1}^{M/2}l_kp_k-\sum_{k=1}^{M/2}l_k\sum_{k=1}^{M/2}p_k}{(M/2)\sum_{k=1}^{M/2}l_k^2-\left(\sum_{k=1}^{M/2}l_k\right)^2},\\[2ex] b=\dfrac{\sum_{k=1}^{M/2}l_k^2\sum_{k=1}^{M/2}p_k-\sum_{k=1}^{M/2}l_k\sum_{k=1}^{M/2}l_kp_k}{(M/2)\sum_{k=1}^{M/2}l_k^2-\left(\sum_{k=1}^{M/2}l_k\right)^2}.\end{cases} \tag{7}$$

Channel values at all locations including data and pilots in stream 1, 5, 9, and 13 will be calculated by applying

$$h_{k'}^{\text{dataEST}}=al_{k'}+b,\quad k'=0,1,\ldots,M-1. \tag{8}$$

It is important to derive the maximum number of OFDM symbols M so that the fitting by using a "line" is reasonable. By the fact that the fading channel will change in time with a coherent time T_c as mentioned deeper in next section, it is clear to see that the limit should be $MT_{\text{symbol}}<T_c$. So a rough limit range for M can be $4\le M<T_c/T_{\text{symbol}}$ and for convenient M should be even number.

3.2.2. Interpolation in Frequency. The same routine as in time axis can be used to interpolate in frequency axis. Now, there is a block of M clusters; each cluster contains 14 subcarriers in which 4 locations were estimated values from the previous task. One note is that all clusters now have "pilots" at the same indexes; hence, the complexity is less.

Table 1: Profile parameters.

Bandwidth	8.75 MHz	FFT size	1024
Sampling factor n	8/7	Number of used subcarriers N used	840
Sampling frequency	10 MHz	Frame structure	1 preamble symbol + 48 data symbol
Subcarrier space	9.765625 KHz	Modulation mode	QPSK 16-QAM 64-QAM
Useful symbol time Tb	102.4 μs	CP ratio G	1/8
Guard interval Tg	12.8 μs	Channel coding	CC(171,133) rate 1/2
OFDM symbol time Ts	115.2 μs	Carrier frequency	2.3 GHz
Number of user	3	System mode	DL-PUSC

Table 2: Profiles of channels used in simulation.

Model 1 Ped.B	Path power (dB)	−3.9	−4.8	−8.8	−11.9	−11.7	−27.8
	Path delay (μs)	0	0.2	0.8	1.2	2.3	3.7
Model 2 Veh.A	Path power (dB)	−3.1	−4.1	−12.1	−13.1	−18.1	−23.1
	Path delay (μs)	0	0.31	0.71	1.09	1.73	2.51

Again, it is crucial to judge the assumption that channel transfer function in a cluster can be viewed as line. The fact that the frequency range of 14 subcarriers or one cluster should be less than the coherent bandwidth B_c of the fading channel [18] gives $B_c \approx 1/5\sigma_\tau$ in which σ_τ denotes the root mean squared delay spread of the multipath fading channel. Another factor is the bandwidth of the designed system; the performance of this method would degrade when system bandwidth is significantly broader than B_c so that a small portion as cluster could also be frequency selective.

4. Joint Scheme with Channel Tracking

For preamble symbol, the least-square channel values at pilot positions are

$$h_{n_p}^{\text{preLS}} = \frac{y_{n_p}^{\text{pre}}}{x_{n_p}^{\text{pre}}}. \tag{9}$$

In the case of DL-PUSC mode, pilots in preamble symbol are evenly spaced scattered, one every three subcarriers. The whole channel values can be linearly interpolated (except the two zero-guard interval regions) as

$$h_{n_p+i}^{\text{preEST}} = \left(\frac{L-i}{L}\right) h_{n_p}^{\text{preLS}} + \left(\frac{i}{L}\right) h_{n_{p+1}}^{\text{preLS}}, \tag{10}$$

where $i = 1, 2$; $L = 3$; $n_p = 86 + 3p$; $p = 0, \ldots, 282$.

In fact, the channel does not stay the same over time but slowly changes; this is due to the relative movement of all the components influencing the transmission. The most impact factor is the relative speed between mobile station and base station that causes a Doppler frequency shift f_D. The coherent time $T_c \approx 1/f_D$ over which the channel can be viewed as unchanged is in the order of several milliseconds to hundreds of milliseconds. Hence, it is considered "slow" when comparing to an OFDM symbol time slot. As a result, after estimating the whole channel from the preamble symbol, performance of channel estimation for successive data symbols can be enhanced by a tracking algorithm [9].

Suppose that $h_{n,m}^{\text{dataEST}}$ is the estimated channel value at subcarrier n of data symbol m; it can be recalculated so that some useful information from the preamble symbol can be involved to reduce the distance between it and the real channel, and thus enhance the estimation performance:

$$\hat{h}_{n,m}^{\text{data}} = \left(\frac{M-m-1}{M}\right) h_n^{\text{preEST}} + \left(\frac{m+1}{M}\right) h_{n,m}^{\text{dataEST}}, \tag{11}$$

where $n = 0, \ldots, N-1$; $m = 0, \ldots, M-1$.

This tracking algorithm means that the nearer the data symbol is located to the preamble symbol; the more influent it gets from the estimated channel given by the preamble and vice versa.

5. Simulation Results

5.1. Simulation Setup. The typical 1024-point FFT profile, whose parameters are given in Table 1, is chosen for simulation. The number of users requesting service is assumed to be 3.

Channel models are taken from ITU models for mobile environments [19], and carrier frequency is set to be 2.3 GHz:

(i) Model 1: ITU_Pedestrian B (Ped.B), speed 6 km/h, and fading frequency $f_D \approx 12.78$ Hz,

(ii) Model 2: ITU_Vehicular A (Veh.A), speed 30 km/h with $f_D \approx 63.89$ Hz, and speed 120 km/h with $f_D \approx 255.56$ Hz.

These channel models are time-variant frequency selective channels in Non Line-Of-Sight mobile conditions. Their specific parameters are given in Table 2.

System performance is demonstrated as packet error rate (PER) versus signal-to-noise ratio (SNR) and user link throughput defined as

$$T = D(1 - \text{PER}) \tag{12}$$

where D is the peak data rate given by

$$D = N_s N_b R_c / T_s \tag{13}$$

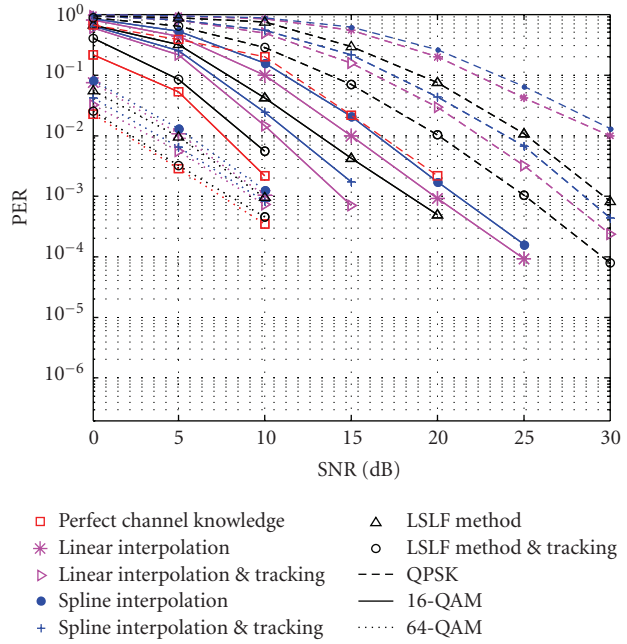


Figure 6: PERs in Ped.B 6 km/h.

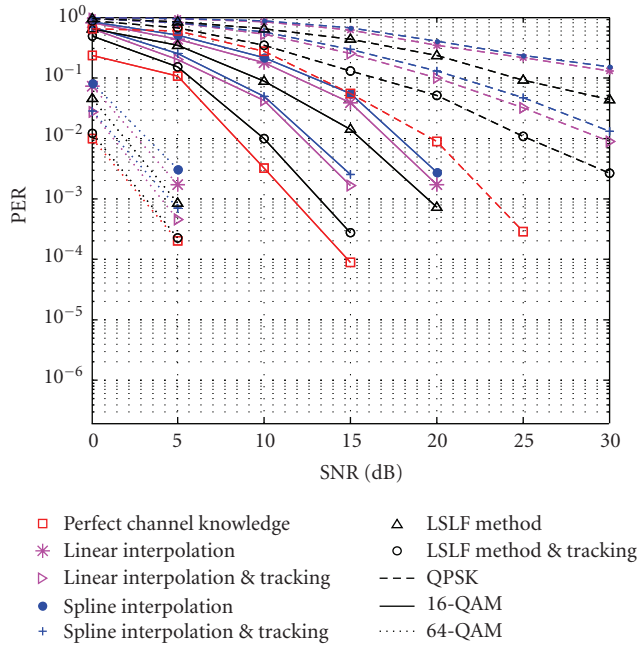


Figure 7: PERs in Veh.A 30 km/h.

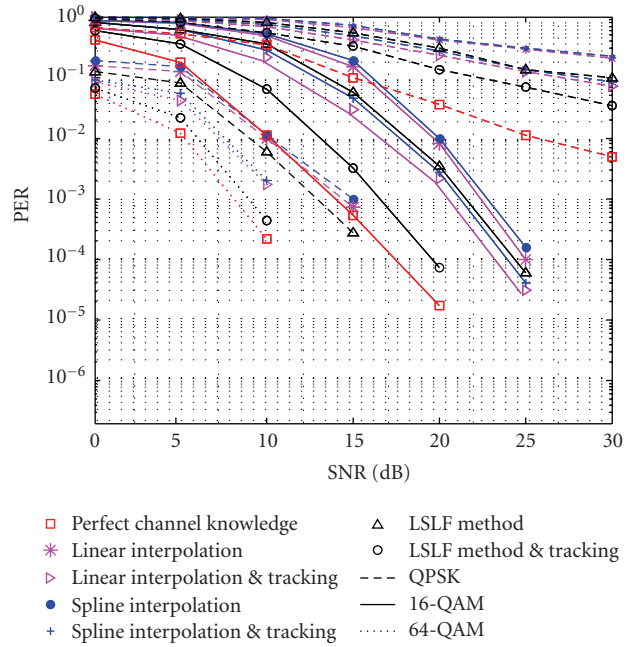


Figure 8: PERs in Veh.A 120 km/h.

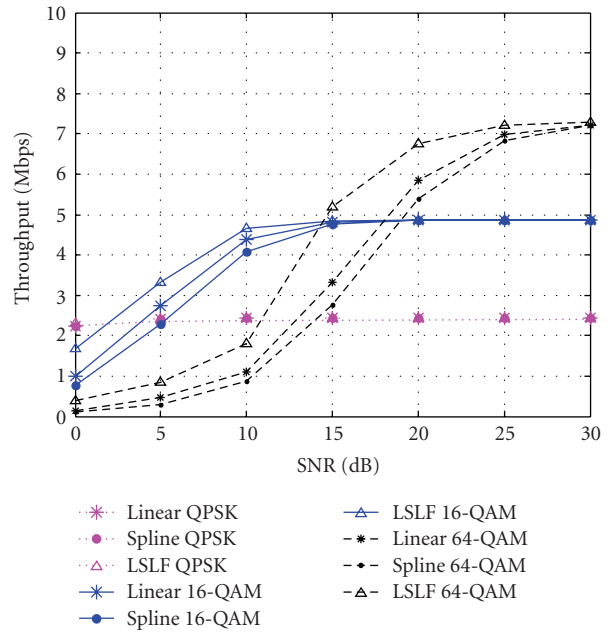


Figure 9: Throughput in Ped.B 6 km/h without using channel tracking.

in which N_s, N_b, R_c, and T_s denote the number of subcarriers assigned to a user, number of data bits in a subcarrier, channel coding rate, and the OFDM symbol time, respectively.

5.2. Simulation Results and Discussion. Simulation results are shown in Figure 6 to Figure 14. The very first notice is that in all channel conditions the LSLF approach always outperforms the other two conventional methods. The improvement varies depending on which modulation mode is used. It is also very clear to see that when channel tracking comes into play, regardless of modulation schemes and channel conditions, the performance is remarkably boosted. The joint scheme of LSLF channel estimation and channel tracking appears to be the best, very robust, and not only highly surpassing other schemes but also able to reach very near to the perfect channel knowledge case.

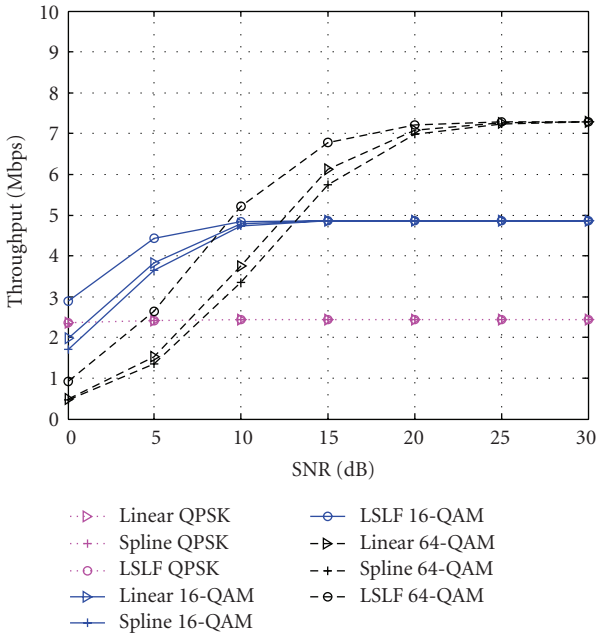


Figure 10: Throughput in Ped.B 6 km/h using channel tracking.

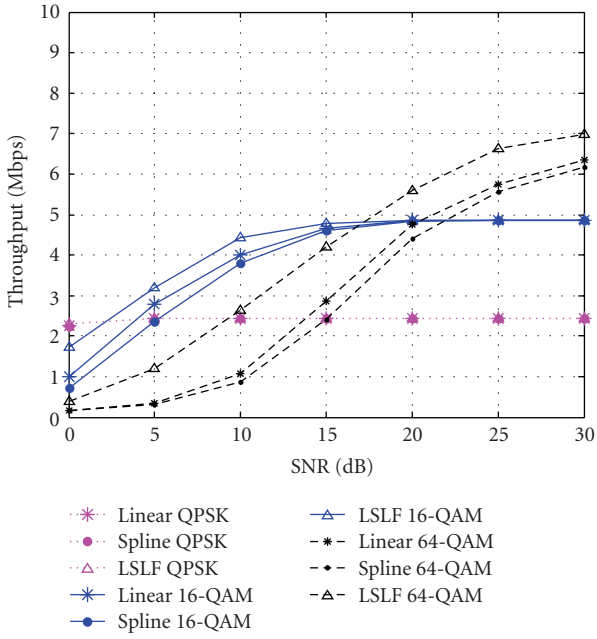


Figure 11: Throughput in Veh.A 30 km/h without using channel tracking.

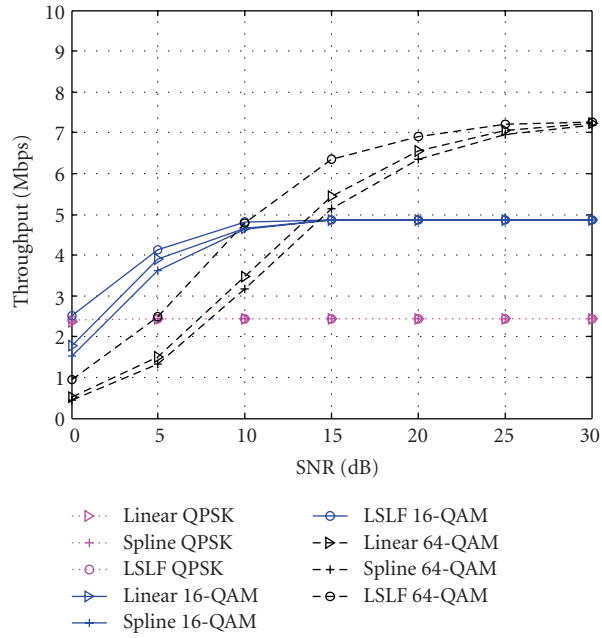


Figure 12: Throughput in Veh.A 30 km/h using channel tracking.

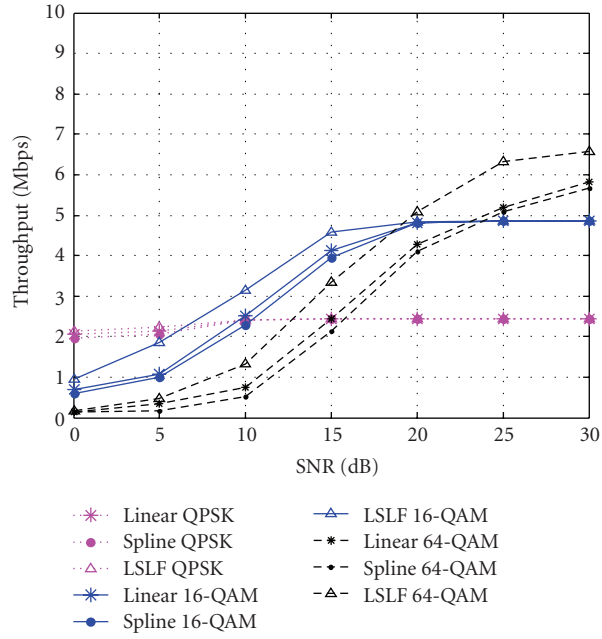


Figure 13: Throughput in Veh.A 120 km/h without using channel tracking.

Figures 6, 7, and 8 show PERs in different channel models. Ped.B channel has quite long delay spread, causing severely frequency-selective faded channel and limiting the performance of channel estimation, particularly in frequency axis. However, due to the slow moving speed, the channel does not change rapidly, giving some favor to estimation in time. One can notice that the effect given by channel tracking in this channel is not as strong as that in Veh.A channels. On the other hand, Veh.A channels have smaller delay spread but higher moving speed, meaning that the channel within a cluster is flat but it changes faster. The coherent time of this channel in case of speed 120 km/h is in order of several milliseconds which can degrade the system performance since it might go below the frame time. However, the LSLF still works properly and with channel tracking; at least 5-dB enhancement can be achieved.

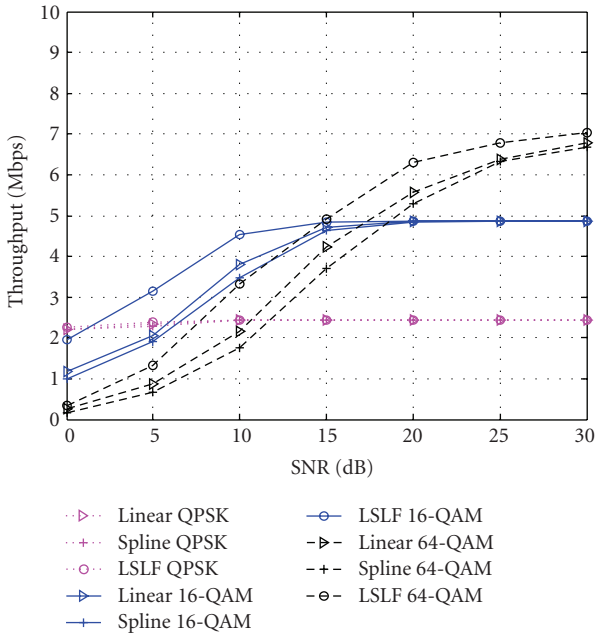


Figure 14: Throughput in Veh.A 120 km/h using channel tracking.

Figures 9, 10, 11, 12, 13, and 14 show the user link throughputs in various channel conditions without and with channel tracking. Obviously, channel tracking give a noticeable improvement and the joint scheme of LSLF channel estimation with tracking significantly increases the link performance.

Another notice is that the performance improvement also depends on modulation modes. The higher modulation modes always suffer higher error in channel estimation, leading to more degradation compared to the perfect channel knowledge case whereas in lower modulation mode, for example, QPSK, the joint scheme is able to reach the ideal case.

Last but not least, it is worth to examine roughly the complexity of LSLF method for practical implementation. From equations (7), it is obvious that the LSLF method needs more computation than linear and cubic spline interpolation but it does not require any complicated process or special design structure. There is no complex operation since the in-phase and quadrature components can be treated separately whereas there are also some terms in (7) that can be reused. Therefore, the superior performance gain obtained by the joint scheme with channel tracking makes this method very promising for realization.

6. Conclusions

This paper has studied a joint channel tracking and estimating scheme which is highly suitable for OFDMA DL-PUSC mode of mobile WiMAX system. System simulation with various standardized channel models for mobile environments showed impressive improvements in both PER and user link throughput. Low complexity and high performance give this joint scheme a high potential for practical implementation.

References

[1] "IEEE Standard for Local and Metropolitan area networks Part 16," The Institute of Electrical and Electronics Engineering, Inc. Std. IEEE 802.16e, 2005.

[2] "IEEE Standard for Local and Metropolitan area networks Part 16," The Institute of Electrical and Electronics Engineering, Inc. Std.IEEE 802.16d, 2004.

[3] G. Parsaee and A. Yarali, "OFDMA for the 4th generation cellular networks," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE '04)*, vol. 4, pp. 2325–2330, 2004.

[4] O. Edfors, M. Sandell, J.-J. Van de Beek, D. Landström, and F. Sjöberg, *An Introduction to Orthogonal Frequency Division Multiplexing*, Luleå Tekniska Universitet, Luleå, Sweden, 1996.

[5] Y. Shen and E. F. Martinez, "WiMAX channel estimation: algorithms and implementations," Tech. Rep. AN3429, Freescale Semiconductor Inc., Brooklyn, NY, USA, 2007.

[6] Y. Shen and E. F. Martinez, "Channel estimation in OFDM systems," Tech. Rep. AN3059, Freescale Semiconductor Inc., Brooklyn, NY, USA, 2006.

[7] M. Henkel, C. Schilling, and W. Schroer, "Comparison of channel estimation methods for pilot aided OFDM systems," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '07)*, pp. 1435–1439, Dublin, Ireland, April 2007.

[8] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "A study of channel estimation in OFDM systems," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '02)*, vol. 56, no. 2, pp. 894–898, 2002.

[9] X. Dong, X. Xie, and X. Chen, "Joint channel estimation for WiMAX by preamble and uneven pilot," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '07)*, pp. 1104–1107, September 2007.

[10] Y. Zhao and A. Huang, "A novel channel estimation method for OFDM mobile communication systems based on pilot signals and transform-domain processing," in *Proceedings of the 47th IEEE Vehicular Technology Conference (VTC '97)*, vol. 3, pp. 2089–2093, May 1997.

[11] J.-J. van de Beek, O. Edfors, M. Sandell, S. Wilson, and P. Borjesson, "On channel estimation in OFDM systems," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '95)*, vol. 2, pp. 815–819, Chicago, Ill, USA, July 1995.

[12] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "Channel estimation techniques based on pilot arrangement in OFDM systems," *IEEE Transactions on Broadcasting*, vol. 48, no. 3, pp. 223–229, 2002.

[13] T. Yücek, M. K. Özdemir, H. Arslan, and F. E. Retnasothie, "A comparative study of initial downlink channel estimation algorithms for mobile WiMAX," in *Proceedings of the IEEE Mobile WiMAX Symposium*, pp. 32–37, March 2007.

[14] M. Morelli and U. Mengali, "A comparison of pilot-aided channel estimation methods for OFDM systems," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3065–3073, 2001.

[15] M. Sandell and O. Edfors, "A comparative study of pilot-based channel estimators for wireless OFDM," Tech. Rep., Signal Processing Division, Luleå University of Technology, Luleå, Sweden, September 1996.

[16] E. Weisstein, *Cubic Spline*, The MathWorld Book, Wolfram Math World.

[17] E. Weisstein, *Least Squares Fitting*, The MathWorld Book, Wolfram Math World.

[18] B. Sklar, "Rayleigh fading channels in mobile digital communication systems part I: characterization," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 136–146, 1997.

[19] Recommendation ITU-R M.1225, "Guidelines for evaluation of radio transmission technologies for IMT-2000," 1997.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 916569, 15 pages
doi:10.1155/2010/916569

Research Article

Paging and Location Management in IEEE 802.16j Multihop Relay Network

Kuan-Po Lin and Hung-Yu Wei

Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan

Correspondence should be addressed to Hung-Yu Wei, hywei@cc.ee.ntu.edu.tw

Received 29 September 2009; Accepted 15 December 2009

Academic Editor: Rashid Saeed

IEEE 802.16j is an emerging wireless broadband networking standard that integrates infrastructure base stations with multihop relay technology. Based on the idle mode operation in IEEE 802.16j, we propose a novel location management and paging scheme. It integrates the paging area-based and the timer-based location update mechanism. In paging area-based scheme, an idle mode mobile station updates when it moves to a new paging area. In timer-based scheme, an idle mode MS updates when the location update timer expires. In this work, we formulate the mathematical model to evaluate the performance of the proposed paging scheme. A new random walk mobility model that is suitable for modeling in multihop relay network is created. Optimization of location update timer is also investigated.

1. Introduction

IEEE 802.16 standard [1] (or WIMAX) is an emerging broadband wireless access system to provide users with high-speed multimedia services. The IEEE 802.16e standard provides mobility support for WiMAX system. Mobile Stations (MSs) are usually powered by battery. Paging mechanism and MS idle mode operation are defined to save power in mobile IEEE 802.16e system. Recently, the IEEE 802.16j Multihop Relay (MR) standard is proposed to support for multihop relay communications with Relay station (RS) [2–4]. IEEE 802.16j standard provides better network coverage and enhance system throughput performance. In 802.16j network, the base station is called Multihop Relay BS (MR-BS). Relay Stations (RSs) relay signaling and data messages between the MR-BS and the MS.

In WiMAX system, MS enters idle mode to save power when there is no data to transmit or to receive. Whenever an incoming data message arrives, the network applies paging mechanism to wake up the dormant MS. During idle mode operation, MS still needs to update its location occasionally so that network only needs to perform broadcast paging in selected cells when a data message arrives. Tradeoff between signaling cost and location precision of idle mode MS is the main design issue in paging and location update protocol design.

Conventional cellular network paging and location management design could be categorized as follows (1) Location-based paging area schemes [5]: users update when they move across the border between different paging areas. Paging area might be overlapping or nonoverlapping. (2) Time-based schemes [6]: users update periodically when the update timer expires. (3) Distance-based schemes [7–9]: users update when moving a fixed distance away from the last updating location. (4) Movement-based schemes: users update based on the number of passing stations. (5) Velocity-based schemes: users update based on the velocity. (6) Profile-based schemes [10]: users update according to their behaviors. Some schemes apply an integrated approach to reduce the signaling cost [11]. Paging for microcell/macrocell overlay is also studied [12]. Pipeline paging technique could be applied to reduce the paging delay [13].

In this paper, we propose a novel paging and location update algorithm that integrates timer-based scheme and location-based paging area scheme for IEEE 802.16j system. For performance evaluation, we investigate a random walk mobility model that is suitable to evaluate the mobility issue in multihop relay cellular network like 802.16j, as base

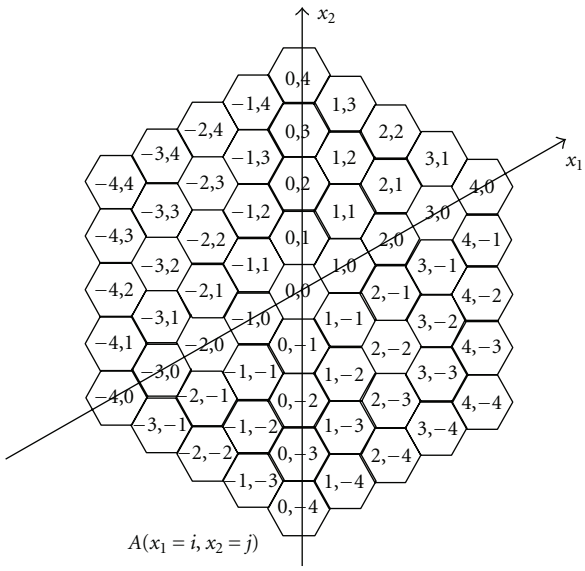


Figure 1: Absolute Geographical Location Model: $A(x1;\ x2)$.

stations and relay stations operate differently but coexist in this type of network. The mobility model is described and validated in Section 2. The paging scheme design is presented in Section 3. In Section 4, we evaluate the system performance analytically. The optimization of location update period is presented in Section 5. Performance results are presented in Section 6. Finally, we conclude the paper in Section 7.

2. Mobility Model

Random walk model is widely used for modelling mobility in cellular networks [6, 11, 14, 15]. Markov chain formulation is used to compute the probability that MS movement. Labelling and grouping cells based on geometric symmetry reduces the complexity of the model. Akyildiz et al. proposed a random walk model for MS mobility in cellular networks [14]. In this model, MSs move in the hexagonal cell. The probability that MS moves to an adjacent hexagonal cell is a system parameter. When the MS moves to an adjacent cell, it has the uniform probability to move to one of the 6 adjacent hexagonal cells. The cellular random walk model is no longer applicable in multihop relay network as some cells are base stations and some are relay stations.

In the proposed model, the probability of MS movement from arbitrary cell i to arbitrary cell j could be computed while computational complexity is limited. The goals of the proposed random walk mobility model for multihop relay networks are to (1) uniquely identify the relay station cells and (2) simplify the mathematical model based on the symmetric property.

An MR-BS (multihop relay base station) or an RS (relay station) is located in the center of a hexagonal cell. Random walk mobility model is applied to characterize the movement of mobile stations (MSs). The Absolute Geographical Location is applied to uniquely identify the hexagonal cells. The Relative Moving Distance is applied to reduce the complexity of the random walk mathematical model. Rules of mapping between Absolute Geographical Location and Relative Moving Distance will also be described in this section.

2.1. Absolute Geographical Location Model. The Absolute Geographical Location is used to uniquely identify the geographical location of each hexagonal cell. Unlike the random mobility model described in [14], hexagonal cells have to be uniquely labelled to distinguish MR-BS and RS. As shown in Figure 1, we apply oblique coordinates with axis x_1 and x_2 to label the hexagonal cells. Each cell is uniquely identified as $A(x_1 = i, x_2 = j)$. The origin of the oblique coordinate is $A(0, 0)$, where MR-BS is usually located.

2.2. Relative Moving Distance Model. As described previously, the Absolute Geographical Location $A(i, j)$ indicates the geographical cell location. Due to the symmetric property of random walk mobility model, the probability of an MS moving from cell $A(i, j)$ to new cell $A(m, n)$ is the same as moving from cell $A(0, 0)$ to $A(m - i, n - j)$. Thus, in terms of moving probability between cells, we can model that the moving probability by considering the probability of an MS moves from the origin $R(0, 0)$ to $R(u, k)$ in the Relative Moving Distance model. The MS moving probability $P_{R(u,k)}$ in the Relative Moving Distance model is the same as $P_{A(i,j)\to A(m,n)}$ and $P_{A(0,0)\to A(m-i,n-j)}$ in the Absolute Geographical Location model.

The Relative Moving Distance model is consisted of n_r tiers of hexagonal cells. A 5-tier Relative Moving Distance model is shown in Figure 2. In the boundary of the wireless network, an MS may enter an outer cell and does not come back to the network. In the Markov Chain models, those outer cells will be modeled as absorbing states. The outer cells are the fifth tier of the network, which is denoted as *out*, as shown in Figure 2.

There are three axes (v_1, v_2, and v_3) across the origin $R(0, 0)$, and the network is divided into six regions. A hexagonal cell is labelled as $R(u, k)$. The (u, k) tuple is labelled based on the oblique coordinate system with axes $v_1 = (0, 1)$ and $v_2 = (1, 0)$. Note that, for the cells in the same tier, the sum of u and k is the same and is equal to the tier number n_r. Cells in Relative Moving Distance model are symmetric. This model provides mobility information for Absolute Geographical Location. An MS movement from (i, j) to (m, n) in the Absolute Geographical Location model will be transformed to an MS movement from $(0, 0)$ to $(|m - i|, |n - j|)$ in the Relative Moving Distance model. Notice that the u, k in Relative Moving Distance model are all nonnegative integer; hence, absolute value operation is taken during the transformation.

2.3. Simplified Moving Distance Model. Since the six regions in the Relative Moving Distance model shown in Figure 2 are symmetric in terms of MS moving probability, we could further simplify the moving distance model. Figure 3 illustrates the Simplified Moving Distance model, which is

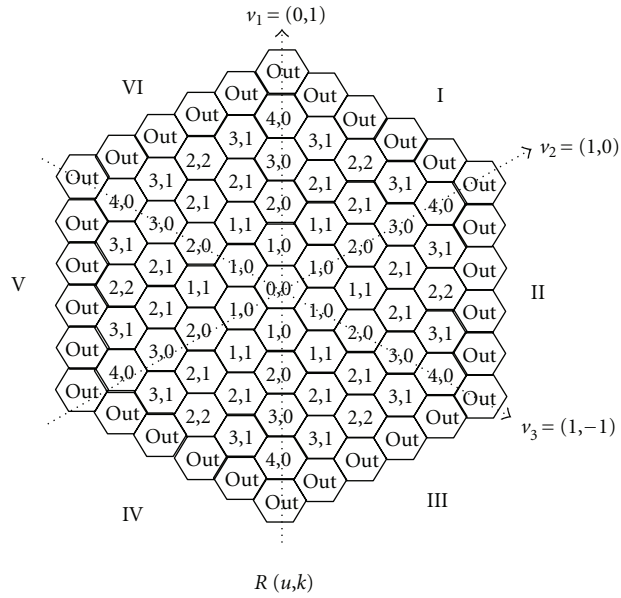


Figure 2: Relative Moving Distance Model: $R(u,k)$.

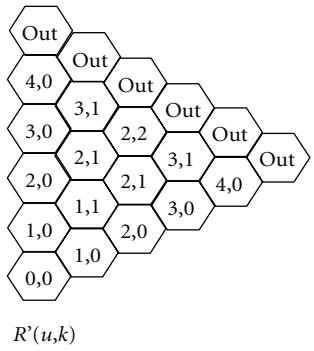


Figure 3: Simplified Moving Distance Model:$R'(u,k)$.

actually the Region I of the original Relative Moving Distance model. The cell in the Simplified Moving Distance model is denoted as $R'(u,k)$, where u,k are non-negative integers and $u \geq k$.

2.4. Rules of Mapping. We will describe a set of mapping rules that transforms the relative moving distance to the absolute geographical location. Because of the Markov property, the future MS movement depends only on the current location state. In the Relative Moving Distance model, a mobile station always starts from $R(0,0)$ as we proposed this *relative* mobility model for movement from the current location of the MS. The coordinate space is considered to be shifted so that the origin of the coordinate space is centered at the current MS location.

We observe the geometric property of the hexagonal topology to create 3 mapping rules to simplify the model. We classify the 6 regions in Figure 2 based on the geometric properties. Region I and IV will apply Mapping Rule I. In Regions I and IV, we find that $(m-i)(n-j) \geq 0$ is always true. Regions II and V will apply Mapping Rule II. In Regions II and V, we find that $(m-i)(n-j) < 0$ and $|m-i| \geq |n-j|$ is always true. Regions II and VI will apply Mapping Rule III. In Regions II and VI, we find that $(m-i)(n-j) < 0$ and $|m-i| < |n-j|$ is always true. Based on the geometric property, these 3 classifications of mapping rules will be discussed in Theorems 1, 2, and 3, respectively.

Moving from $A(i,j)$ to $A(m,n)$ in a given time interval is transformed to moving between $R(0,0)$ and $R(u,k)$ in the same time interval. If a user starts at $A(i,j)$ and locates in $A(m,n)$ after i unit time, the probability is equal to that of moving from $R(0,0)$ to $R(u,k)$ after i unit time. We define $P^i_{R(u,k)}$ as the probability that an MS moves from $R(0,0)$ to $R(u,k)$ after i unit time:

$$P^i_{A(i,j)\to A(m,n)} = P^i_{A(0,0)\to A(m-i,\ n-j)} = P^i_{R(u,k)}. \tag{1}$$

In the Relative Moving Distance model, three axes divide the network into six regions. As the Relative Moving Distance model applies an MS-centric view that considers relative movement from the starting location, the MS movement is always starting from $R(0,0)$. The MS movement in the original Absolute Geographical Location from $A(i,j)$ to $A(m,n)$ is equivalent to the transformed MS movement from $R(0,0)$ to $R(m-i,n-j)$. The movement to $R(m-i,n-j)$ could be classified based on values of $m-i$ and $n-j$. The classification of the mapping rules also corresponds to the mobile movement in the six regions shown in Figure 2.

All cells in Regions I and IV have the property $(m-i)(n-j) \geq 0$. The relative movement vector $(m-i,n-j)$ can be denoted as a linear composition of two axes $v_1 = (0,1)$ and $v_2 = (1,0)$ with integer coefficients a and b :

$$a \cdot v_1 + b \cdot v_2 = a \cdot (0,1) + b \cdot (1,0) = (m-i, n-j). \tag{2}$$

In the Simplified Moving Distance model $R'(u,k)$, u and k are non-negative integers. We solve the above equation and derive the non-negative solution by taking absolute values $a = |n-j|$, $b = |m-i|$. Since $u \geq k$ in the Simplified Moving Distance model, as shown in Figure 4, u is the larger one among a and b while the smaller one is k.

For example, as shown in Figure 4, $A(1,3)$ can be decomposed as the linear combination of v_1 and v_2. Notice that the moving probability from $A(0,0)$ to $A(1,3)$ is the same as the moving probability from $R'(0,0)$ to $R'(3,1)$:

$$(1,3) = a \cdot v_1 + b \cdot v_2 = a \cdot (0,1) + b \cdot (1,0) \Longrightarrow a = 3, b = 1. \tag{3}$$

From observation, the Mapping Rule I maps the absolute geographical location to the relative moving distance model in Regions I and IV, as shown in Figure 2. Notice that the relative moving values $(m-i)$ and $(n-j)$ are both positive values (in Region I) or both negative values (in Region IV).

Theorem 1 (Mapping Rule I). *While* $(m-i)(n-j) \geq 0$,

$$P^i_{A(i,j)\to A(m,n)} = \begin{cases} P^i_{R'(|m-i|,|n-j|)}, & \text{if } |m-i| \geq |n-j|, \\ P^i_{R'(|n-j|,|m-i|)}, & \text{if } |m-i| < |n-j|. \end{cases} \tag{4}$$

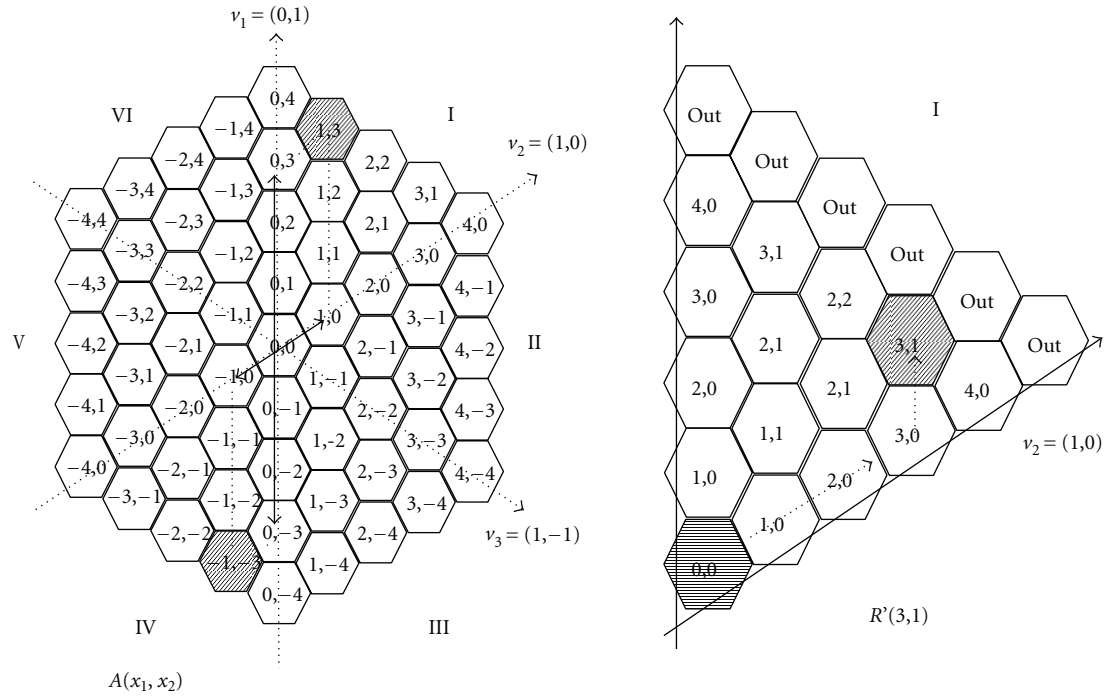


Figure 4: Mapping example.

If an MS moves to cells in Region II or V in Figure 2, the following two properties hold: $(m-i)(n-j) < 0$ and $|m-i| \geq |n-j|$. The relative movement vector $(m-i, n-j)$ can be denoted as a linear combination of $v_2 = (1,0)$ and $v_3 = (1,-1)$. We can get $a = |m+n-i-j|$, $b = |n-j|$ by solving the equation:

$$a \cdot v_2 + b \cdot v_3 = a \cdot (1,0) + b \cdot (1,-1) = (m-i, n-j). \tag{5}$$

Theorem 2 (Mapping Rule II). *While* $(m-i)(n-j) < 0$ *and* $|m-i| \geq |n-j|$,

$$P^i_{A(i,j)\to A(m,n)} = \begin{cases} P^i_{R'(|n-j|,|m+n-i-j|)} & \text{if } |n-j| \geq |m+n-i-j|, \\ P^i_{R'(|m+n-i-j|,|n-j|)} & \text{if } |n-j| < |m+n-i-j|. \end{cases} \tag{6}$$

If an MS moves to Region III or VI, the following two properties hold: $(m-i)(n-j) < 0$ and $|m-i| < |n-j|$. The relative movement vector $(m-i, n-j)$ can be denoted as a linear combination of $-v_1 = (0,-1)$, and $v_3 = (1,-1)$. We can get $a = |m-i|$, and $b = |m+n-i-j|$ by solving the equation:

$$a \cdot v_3 + b \cdot (-v_1) = a \cdot (1,-1) + b \cdot (0,-1) = (m-i, n-j). \tag{7}$$

Theorem 3 (Mapping Rule III). *While* $(m-i)(n-j) < 0$ *and* $|m-i| < |n-j|$,

$$P^i_{A(i,j)\to A(m,n)} = \begin{cases} P^i_{R'(|m-i|,|m+n-i-j|)} & \text{if } |m-i| \geq |m+n-i-j|, \\ P^i_{R'(|m+n-i-j|,|m-i|)} & \text{if } |m-i| < |m+n-i-j|. \end{cases} \tag{8}$$

An example of mapping movement to Region III is shown in Figure 5. The left part of the figure is the Absolute Geographical Location. An MS moves from $A(-1,2)$ to $A(1,-2)$. The right part of figure is the equivalent Relative Moving Distance. Considering the starting point $A(1,-2)$ as the center of the map, the destination $A(-1,2)$ is in Region III. Applying Theorem 3 and setting $(i,j) = (-1,2)$ and $(m,n) = (1,-2)$, we can obtain $P^i_{A(-1,2)\to A(1,-2)} = P^i_{R'(2,2)}$.

2.5. Calculation of User Movement Probability. The user movement is modelled by the random walk mobility model. As described previously, the computation of MS movement could be simplified by the transformation and mapping to the Simplified Moving Distance Model $R'(u,k)$. The mobile network model has n_r tiers of cells. The value of n_r must be large enough so that the probability of users moving outside is small. Depending on the requirements of

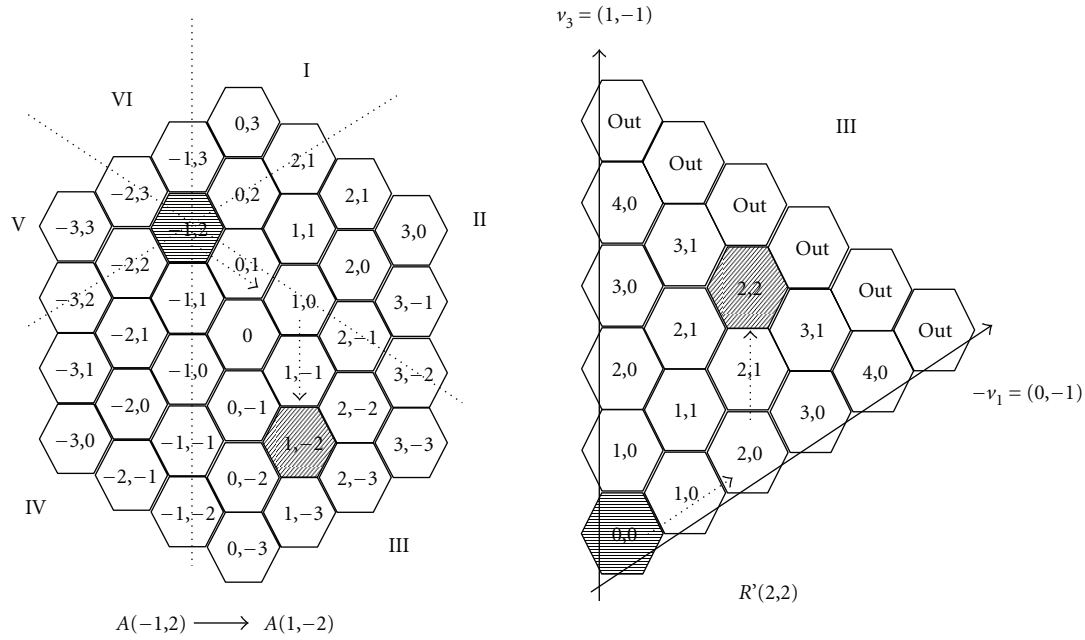


Figure 5: Mapping example.

modelling various mobility protocols, the value of n_r should be selected accordingly.

To further simplify the notation, we map each cell $R'(u,k)$ in Simplified Moving Distance Model to a new state S_x, as shown in Figure 6. The states are relabelled from inner cells toward outer cells. For example, the origin $R'(0,0)$ is denoted as S_1. Likewise, the $R'(1,0)$ is denoted as S_2, and so forth. As we observed, the relabeling based on geometrical symmetry could be used to simplify the following mobility model formulation. A discrete-time Markov Chain model, as shown in Figure 7, is created to compute the MS movement probability. We denote the probability that an MS stays in the same cell in the next time slot as p. The probability that an MS moves to a neighboring cell in the next time slot is thus $1-p$, which is denoted as q. In the random walk model, the MS has probability p to stay in the same cell and $q/6$ to move to another adjacent cell (notice that there are 6 neighboring cells). By observation of the geometric properties of the hexagonal topology, the random walk mobility could be formulated as the Markov Chain shown in Figure 6.

We define the matrix O_i to represent the probability that an MS is in state S_x after i unit time slots. The size of an n_r-tier network is denoted as $S(n_r)$. Hence, the size of O_i is 1 by $S(n_r)$:

$$\begin{aligned} O_i &= \left(P^i_{S_1} \;\; P^i_{S_2} \;\; P^i_{S_3} \;\; P^i_{S_4} \;\; \ldots\right)_{1\times S(n_r)} \\ &= \left(P^i_{R(0,0)} \;\; P^i_{R(1,0)} \;\; P^i_{R(2,0)} \;\; P^i_{R(1,1)} \;\; \ldots\right)_{1\times S(n_r)}. \end{aligned} \tag{9}$$

In the relative moving model, the initial location of an MS is at the origin at time 0. The initial state O_0 is described

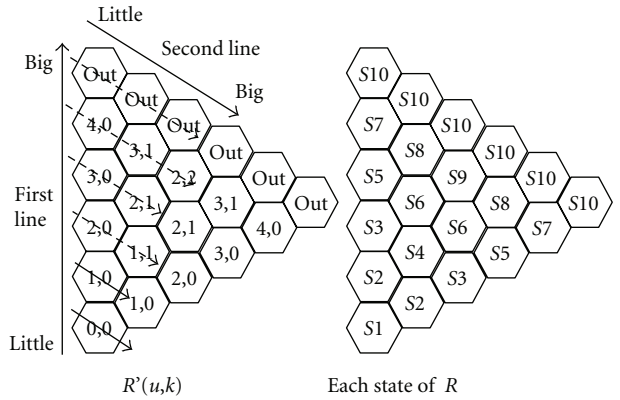


Figure 6: Relabelling the Markov Chain states S_x.

as follows:

$$O_0 = \left(1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; \ldots\right)_{1\times S(n_r)} \tag{10}$$

The probabilistic transition matrix in the Markov Chain model is denoted as Ts. It is an $S(n_r)$ by $S(n_r)$ matrix. As shown in Figure 2, the number of tiers in the hexagonal topology is symmetric. We derive the value of $S(n_r)$ based on observing the geometric property of the hexagonal network topology:

$$S(n_r) = \begin{cases} \dfrac{n_r^2 + 2n_r + 5}{4}, & \text{if } n_r \text{ is odd,} \\ \dfrac{n_r^2 + 2n_r + 4}{4}, & \text{if } n_r \text{ is even.} \end{cases} \tag{11}$$

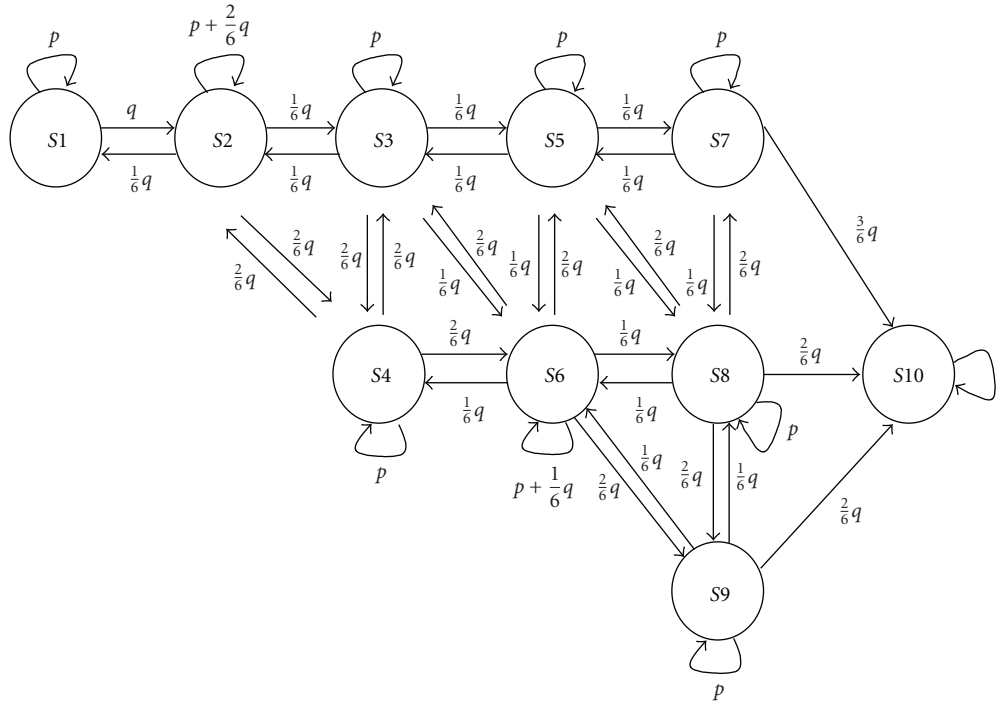


Figure 7: Markov Chain model.

By observing the mobility symmetry in Figure 6, the Markov Chain state transition diagram is drawn in Figure 7. Now, we will write down the state transition probability of the Markov Chain model of Figure 7 in matrix form. An element in Ts is the probability of moving from one state to another state during one unit time in the Markov Chain model:

$$Ts = \begin{pmatrix} p & q & 0 & 0 & 0 & \dots & 0 \\ \frac{q}{6} & p+\frac{q}{3} & \frac{q}{6} & \frac{q}{3} & 0 & \dots & 0 \\ 0 & \frac{q}{6} & p & \frac{q}{3} & \frac{q}{6} & \dots & 0 \\ 0 & \frac{q}{3} & \frac{q}{3} & p & 0 & \dots & 0 \\ 0 & 0 & \frac{q}{6} & 0 & p & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{S(n_r)\times S(n_r)}, \quad (12)$$

$$O_{i+1} = O_i Ts.$$

Based on the state diagram shown in Figure 7, the elements of Ts can be obtained. From the definition, the Markovian state probability in time slot i could be computed by iteratively multiply the current state probability with transition matrix. We can then calculate O_i with Ts iteratively:

$$O_1 = O_0 Ts = \begin{pmatrix} p & q & 0 & 0 & 0 & \dots \end{pmatrix}_{1\times S(n_r)},$$

$$O_2 = O_1 Ts = \begin{pmatrix} p^2+\frac{1}{6}q^2 & pq+q\left(p+\frac{1}{3}q\right) & \frac{1}{6}q^2 & \frac{1}{3}q^2 & 0 \end{pmatrix}_{1\times S(n_r)},$$

$$\vdots$$

$$O_i = O_0 (Ts)^i. \quad (13)$$

The movement probability could be computed with (13). It multiplies Ts by i times. To reduce the computational complicity, we can diagonalize the matrix Ts and derive matrix D and V. D is the diagonal matrix of eigenvalues. V consists of the eigenvectors of Ts. We can obtain the state probability quicker by applying (16):

$$Ts = VDV^{-1}, \quad (14)$$

$$(Ts)^i = VD^iV^{-1}, \quad (15)$$

$$O_i = O_0 VD^iV^{-1}. \quad (16)$$

2.6. Validation of the Mobility Model. Similar to the previous work [14], we validate the mathematical model by simulation. The network tier n_r is 3, and two mobility scenarios $p = 0.8$ or $p = 0.9$ are simulated. The movement probability values after 100 time slots are shown in Table 1. Math 1 method is the result of $O_0 T_s^{100}$ computation based on (13).

Math 2 method is the results of diagonalized computation based on(16)

The initial state probability matrix O_0 is 1 for the center cell and 0 for other cells:

$$O_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}_{1\times S(3)}. \tag{17}$$

The transition matrix Ts is

$$Ts = \begin{pmatrix} p & q & 0 & 0 & 0 \\ \frac{q}{6} & p+\frac{q}{3} & \frac{q}{6} & \frac{q}{3} & 0 \\ 0 & \frac{q}{6} & p & \frac{q}{3} & \frac{q}{2} \\ 0 & \frac{q}{3} & \frac{q}{3} & p & \frac{q}{3} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{S(3)\times S(3)}. \tag{18}$$

The diagonal matrix D is

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1-0.188q & 0 & 0 & 0 \\ 0 & 0 & 1-0.795q & 0 & 0 \\ 0 & 0 & 0 & 1-1.270q & 0 \\ 0 & 0 & 0 & 0 & 1-1.412q \end{pmatrix}_{S(3)\times S(3).}. \tag{19}$$

The transpose of matrix O_i is shown as the following:

$$O_i^T = \begin{pmatrix} 0.140 * (1-0.188q)^i + 0.362(1-0.795q)^i + 0.496(1-1.270q)^i \\ 0.678 * (1-0.188q)^i + 0.275(1-0.795q)^i - 0.954(1-1.270q)^i \\ 0.304 * (1-0.188q)^i - 0.897(1-0.795q)^i + 0.592(1-1.270q)^i \\ 0.408 * (1-0.188q)^i - 0.537(1-0.795q)^i + 0.129(1-1.270q)^i \\ 1-1.532 * (1-0.188q)^i + 0.795(1-0.795q)^i - 0.263(1-1.270q)^i \end{pmatrix}_{S(3)\times 1.} \tag{20}$$

Table 1: The simulation and the math calculation.

$p = 0.8$	$R(0,0)$	$R(1,0)$	$R(2,0)$	$R(1,1)$	outside
Simulation	0.003045	0.014737	0.006624	0.008770	0.966824
Math 1	0.003021	0.014718	0.006617	0.008759	0.966883
Error1	0.78%	0.25%	0.11%	0.11%	0.01%
Math 2	0.003041	0.014660	0.006580	0.008827	0.966891
Error2	0.13%	0.15%	0.63%	0.89%	0.01%
$p = 0.9$	$R(0,0)$	$R(1,0)$	$R(2,0)$	$R(1,1)$	outside
Simulation	0.021103	0.101835	0.045706	0.060456	0.770900
Math 1	0.021009	0.102053	0.045636	0.060503	0.770799
Error1	0.45%	0.21%	0.15%	0.08%	0.01%
Math 2	0.021161	0.101646	0.045363	0.061028	0.770801
Error2	0.27%	0.19%	0.75%	0.95%	0.01%

We implement the Monte Carlo simulation in C++ to model the random walk mobility model in the hexagonal topology. Each MS has probability p to stay in the same cell and probability $(1-p)/6$ to move to any adjacent hexagonal cell. Totally 1000000 simulation runs are conducted. The uniformly random walk mobility simulation results are compared with the Markov Chain analysis results. As shown in Table 1, the differences between the mathematical models and simulation results are always less than 1%. In addition, we observe that the diagonalized method effectively reduces the computation time.

Table 2: Paging and Idle Mode Related Signaling Messages.

Message name	Message description
DREG-REQ	SS De-registration message
DREG-CMD	De/Re-register Command
MOB_PAG-ADV	BS broadcast paging message
RNG-REQ	Ranging Request
RNG-RSP	Ranging Response

3. IEEE 802.16j Multihop Paging

3.1. IEEE 802.16j Idle Mode. Idle mode operation reduces control signaling cost and MS energy consumption. An MS in idle mode periodically listens to the downlink broadcasting paging messages without registering to a specific BS. RSs relay all paging messages between MS and MR-BS. In this paper, we consider nontransparent mode operation in 802.16j system. Idle mode and paging operations are illustrated in Figure 8.

3.1.1. Entering Idle Mode. Before entering idle mode, an MS sends Deregistration message (DREG-REQ) to the MR-BS. Then the MR-BS replies De/Reregister Command message (DREG-CMD) to MS. These two signaling messages are used to synchronize the paging listening time. For an MS serving by the relay stations, the access RS will relay all deregistration messages and paging messages between the MR-BS and the

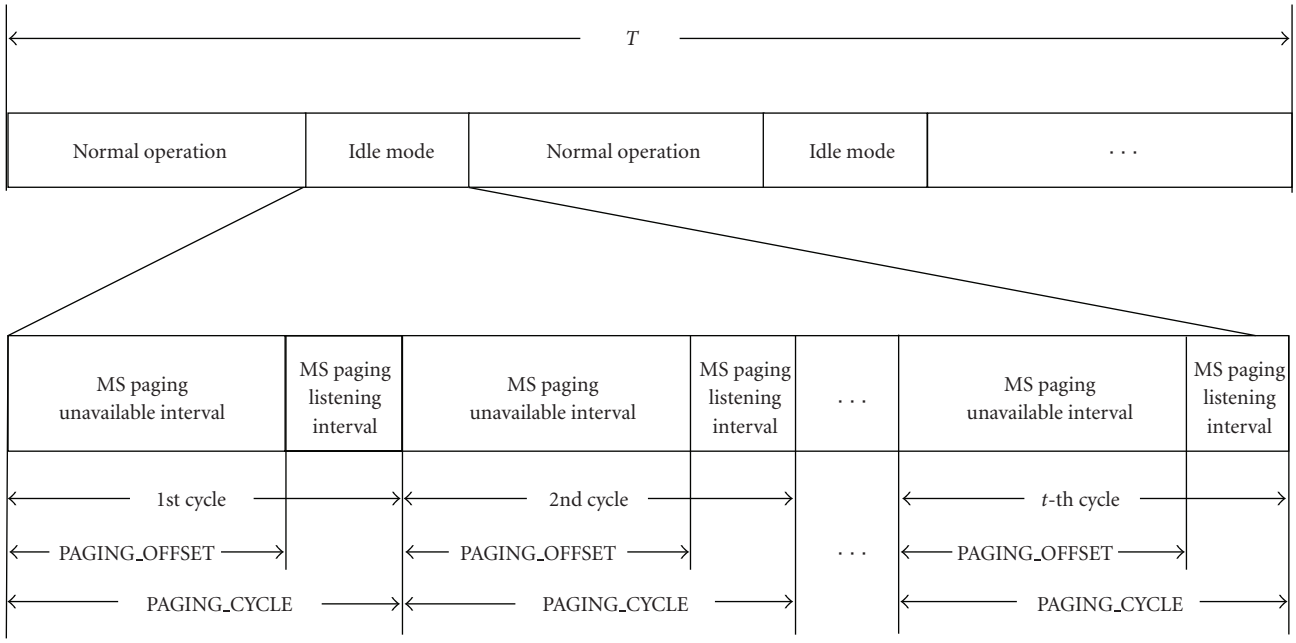


Figure 8: Active mode and idle mode operation.

MS. Notice that the control signaling cost is multiplied by the number of relay hops in this scenario.

3.1.2. Idle Mode Operation. As shown in Figure 9, there are two types of time intervals in idle mode operation: MS Paging Unavailable Interval and MS Paging Listening Interval. During MS Paging Unavailable Interval, an MS turns down radio interface to save power. In MS Paging Listening Interval, an MS listens to the downlink broadcast of paging advertisement messages (MOB_PAG-ADV). The listening interval has a period of *PAGING_CYCLE*. The *PAGING_OFFSET* parameter is used to separate MSs in different paging groups. An MS is synchronized to the periodic listening intervals based on the *PAGING_CYCLE* and *PAGING_OFFSET* given in a MOB_PAG-ADV message.

3.1.3. Termination of Idle Mode. At the end of MS listening interval, an MS must decide whether to leave idle mode or not. If an MS would like to transmit data, it must leave idle mode and enter active mode for normal operation. When an MS decides to terminate the idle mode, it will start the network reentry process by first sending Ranging Request (RNG-REQ) message to MR-BS. Then MR-BS will reply with Ranging Response (RNG-RSP) message to the MS. Then the MS can send the location update message and start the normal active mode operation. Relay stations will forward signaling messages, such as RNG-REQ and RNG-RSP, between MS and MR-BS when needed.

The paging operation is initiated when the system wants to find an MS. For example, a new data packet is arrived and is to be delivered to the MS. The network will check the paging information database that records the associated paging group of the to-be-paged MS. All MR-BS and access relay stations in the paging group will send broadcast paging message MOB_PAG-ADV with the MS's MAC address. Once the MS receives the broadcast paging message, it will terminate the idle mode and go back to normal mode. The MOB_PAG-ADV broadcasting is initiated from the MR-BS and is forwarded through relay stations.

3.2. Paging Methods. In the network topology, MR-BS and RS are assumed to be located at the center of hexagonal cells. A cell is consisted of 1 MR-BS and 6 RSs as shown in Figure 10. Packets are either directly transmitted from MR-BS to MS, if an MS is located in the central cell, or forwarded through two-hop-relay transmission. When the network is going to page an MS, the paging message is forward from the MR-BS to the six RSs. Then the MR-BS and the 6 RSs will broadcast paging messages to MS (i.e., 7 transmissions are needed). Thus, the total signaling cost in one paging event is

$$N_{P_1} = N_{P_1}(\text{Relay}) + N_{P_1}(\text{Broadcast}) = 6 + 7 = 13. \quad (21)$$

Our paging scheme includes both paging area-based update mechanism and timer-based update mechanism. Several cells are grouped into one paging area. An MS roams between different paging areas and sends an update when it moves across the border. If a message arrives, the network only broadcasts the messages in one paging area to find the user. For example, the paging areas can be allocated as shown in Figure 11. There are totally 14 paging areas shown in this figure. In this example, one cell includes one base station and six relay stations, as shown in Figure 10. Notice that a hexagon that labelled with number has a base station, and other neighboring hexagons without number have relay stations, as shown in Figure 11.

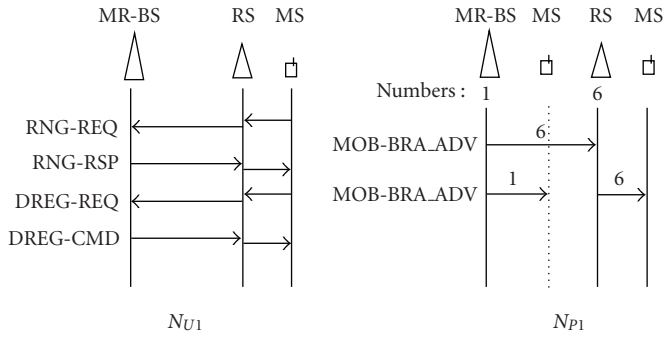


Figure 9: Signaling flow and signaling cost N_{U_1}, N_{P_1}.

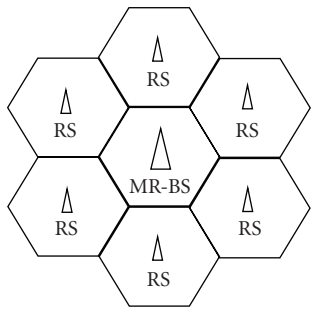


Figure 10: IEEE 802.16j multihop cellular structure: base stations and relay stations.

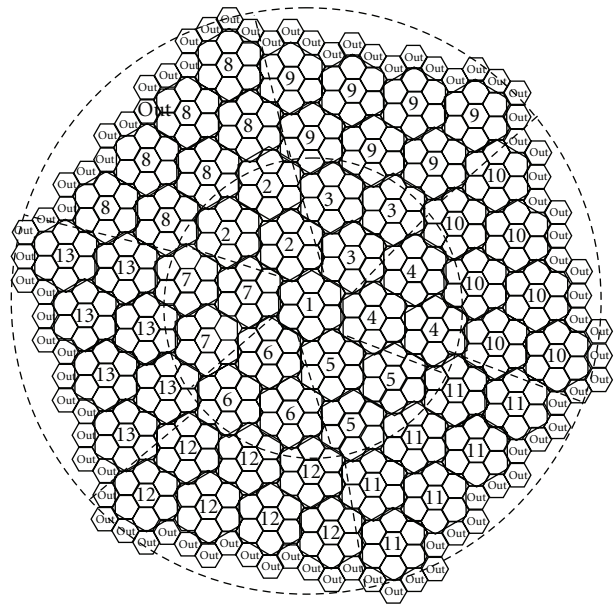


Figure 11: Example of paging area topology with 13 paging area $a_j, j = 1, \ldots, 13$. Base stations are located in hexagons with labelled numbers (Paging Area ID). Relay stations are located in hexagons without number.

Before an MS enters idle mode, the serving base station exchanges DREG-REQ and DREG-CMD messages with it. The last serving cell will be denoted as paging areas 1 as shown in Figure 11. In idle mode, an MS still needs to listen to paging-related information periodically. During every MS Listening interval, the MS listens to broadcast messages, which contains paging-related information. From this information, if the MS detects that it moves to a different paging area, it must notify the network about the paging area change. We call this update the Paging Area Notification (PA Notification). Hence, when a data message arrives, the network knows the right paging area to find the idle mode MS. When an MS moves to a new paging area, the MS will always first send update to the RS and then forwarded the signaling message to BS. In PA Notification, there are totally N_{U_1} signaling cost, which is defined by the number of signaling message transmitted weighed by the number of hops to be forwarded. In the 2-hop multihop cellular structure, as shown in Figure 10, the PA Notification signaling cost is

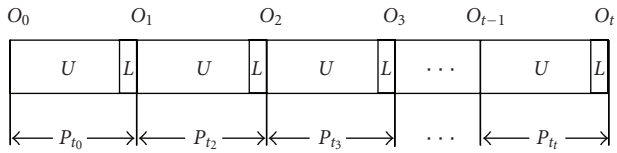
$$N_{U_1} = \text{messages} \times \text{relay} = 4 \times 2 = 8. \tag{22}$$

In the proposed paging scheme, the paging area topology is MS-centric. When an MS updates the exact cell location to the network, the system recomputes the paging area, and the current cell becomes the centralized cell in the paging area, which is labelled with 1 as shown in Figure 11. Paging areas will only be reset in two circumstances: (1) data message arrival (and system create paging message to locate the MS) or (2) timer-based update (timer expires after t).

The first case occurs after the data messages arrive, and the network starts the broadcast paging procedure. All cells in the paging area, where the MS located, will send broadcast paging messages. The second case is timer-based location update. If no message arrives after t time slots, the MS must update its location to avoid losing track of its location. After timer expires, the MS goes into active mode, updates its location, and resets the paging area (set the current cell as paging area 1) before it enters idle mode again.

4. Paging Performance Analysis

Signaling cost in wireless network paging design is critical. In this section, we will investigate the signaling cost in the proposed paging scheme. The 802.16j paging cycle strucutre



L: MS paging listening interval
U: MS paging unavailable interval
P_{t_l} : Probability that arrival in the cycle
O_i: MS movement probability

Figure 12: IEEE 802.16j paging cycle.

is shown in Figure 12. MS Paging unavailable interval and MS Paging Listening Interval appear alternatively. In MS Paging unavailable interval, the MS enters idle mode and does not receive packets from the network. In MS Paging Listening Interval, the MS listens to the paging channel to find whether paging messages are sent. The process could be modeled as discrete events including MS movement and paging arrival occur at the MS Paging Listening Interval. For performance evaluation, we compute the probabilities of the MS movement events and paging arrival events accordingly.

4.1. Interrupted Versus Uninterrupted Idle Periods. We denote the overall time duration as T. During this time, we could further categorize the time period into two types: interrupted idle period and uninterrupted idle period.

4.1.1. Interrupted Idle Period. A paging message arrives and terminates an interrupted idle period. We calculate the N_i, the number of interrupted Idle periods during the total duration T, and N_u, the number of uninterrupted idle periods during T. The paging message arrival follows Poisson random process with rate λ. Hence, the expected number of paging message arrival during time T is λT. The number of interrupted idle periods is

$$N_i \cong \lambda T. \tag{23}$$

4.1.2. Uninterrupted Idle Period. No paging messages arrive during an uninterrupted idle period. An uninterrupted idle period is terminated due to the timer-based forcing update. The mobile-centric location area is reset after timer-based paging area update period t. One additional cycle for active mode operation for the location area reset is needed. Thus, the length of an uninterrupted idle period is $t + 1$ cycles. We denote the time duration from entering idle mode to the paging arrival time as t_p. The expected value of t_p is denoted as $\overline{t_p}$.

During total duration T, the expected interrupted time periods is $\lambda T\overline{t_p}$ cycles. So the number of uninterrupted idle periods is the remaining uninterrupted time during T divided by the duration of an uninterrupted idle period. The expected number of uninterrupted idle periods is

$$N_u = \frac{E\left[T - \lambda T\left(\overline{t_p} + 1\right)\right]}{E[t+1]}. \tag{24}$$

In an interrupted idle period, the signaling messages include paging and location updating. In an uninterrupted idle period, the signaling messages only include location updating at the end of the period.

4.2. Broadcast Paging. Broadcast paging event only occurs during an interrupted period. If the call arrives between $i - 1$ and i cycle, the system broadcasts a paging message to the paging area where the MS locates. We can derive the probability of the MS in a paging area from the probability computation in Section 2.

The total paging signaling cost of one MS at cycle i is "the probability of the MS in paging area a_j" multiplied by "the signaling cost in paging area a_j." We have calculated the paging signaling cost in one multihop cell, N_{P_1}, in (21). Thus, the total paging signaling cost is (the probability of the MS in paging area a_j) $\times N_{P_1} \times$ (the number of multihop cells in paging area a_j).

Based on the mobility model described in Section 2, we can readily compute the probability of an MS in an paging area after time t_p. For example, the paging area a_1 shown in the center of Figure 11 has 1 multihop relay cell, which includes 1 BS hexagonal cell marked with 1 and 6 RS hexagonal cells surrounding the BS. The probability of an MS is located with the paging area a_1 after time t_p is $O_{t_p}[1, 1, 0, 0, \dots]'_{S(n)\times 1}$. The mobility matrix that corresponds to paging area $a_1[1, 1, 0, 0, \dots]'_{S(n)\times 1}$ is denoted as Sp_1. Similarly, Sp_j is the matrix corresponding to paging area a_j. Notice that Sp_j only depends on the paging area topology and is independent of t_p. Considering the whole wireless networks, we have Sp :

$$Sp = \sum_{\forall i} Sp_i. \tag{25}$$

According to the random walk mobility model, the MS location state probability is O_{t_p}. For each paging event, the signaling cost is $N_{P_1} O_{t_p} Sp$. The cost of paging signaling during total time duration T is

$$\text{Paging\_signaling} = N_i N_{P_1} O_{t_p} Sp. \tag{26}$$

Similar to [9], we will compute t_p. The Poisson arrival is

$$P\left(n_p, \Delta t\right) = \frac{e^{-\lambda\Delta t}(\lambda\Delta t)^{n_p}}{n_p!}. \tag{27}$$

The number of arrived paging message is denoted as n_p. If $n_p = 0, P(n_p = 0, \Delta t) = e^{-\lambda\Delta t}$, it implies that no message arrives. If $n_p \neq 0, P(n_p \neq 0, \Delta t) = 1 - e^{-\lambda\Delta t}$, it implies that at least one message arrives. The probability that paging message arrival time t_p falls between $i - 1$ and i, as shown in Figure 12, is

$$\begin{aligned} P_{t_i} &= \prod_{j=0}^{i-2} P\left(n_p = 0,\ j \le t_p < j+1\right) P\left(n_p \neq 0,\ i-1 \le t_p < i\right) \\ &= e^{-\lambda(i-1)}\left(1 - e^{-\lambda}\right). \end{aligned} \tag{28}$$

Then, when $i-1 \le t_i < i$, we calculate $\overline{t_i}$, the expected value of a message arrival time that falls between $i-1$ and i [16]:

$$\overline{t_i} = \frac{\int_{i-1}^{i} \lambda x e^{-\lambda x} dx}{\int_{i-1}^{i} \lambda e^{-\lambda x} dx} = i + \frac{-e^{\lambda}}{e^{\lambda}-1} + \frac{1}{\lambda}. \tag{29}$$

The average value $\overline{t_p}$ is

$$\begin{aligned} \overline{t_p} &= \sum_{i=1}^{t} [P_{t_i} \overline{t_i}] = \left(1-e^{-\lambda}\right) \sum_{i=1}^{t} i e^{-\lambda(i-1)} \\ &\quad + \left(1-e^{-\lambda t}\right)\left(-\frac{e^{\lambda}}{e^{\lambda}-1} + \frac{1}{\lambda}\right) \\ &= \frac{1}{\lambda}\left(1-e^{-\lambda t}\right) - t e^{-\lambda t}. \end{aligned} \tag{30}$$

Thus, from (26), the overall paging signaling cost is

$$\text{Paging\_signaling} = N_i N_{P_1} \frac{\sum_{i=1}^{t} O_i Sp P_{t_i}}{\sum_{i=1}^{t} P_{t_i}}. \tag{31}$$

4.3. Paging Area Notification (PA Notification). If the MS moves across the border between two different paging areas, the MS must notify the network about the PA change. The MS update signaling cost of each PA Notification event is denoted as N_{U_1}. The corresponding PA notification probability between cycle i and $i+1$ is the summation of the probability across the paging area border, according to the previously described random walk mobility model and the paging area topology. There are totally $N_{U_1} = 4 \times 2 = 8$ signaling message transmissions when an MS updates.

Similar to the Sp_i formulation, the mobility matrix for PA notification event, in which an MS moves away from paging area a_i, is denoted as Su_i. Similarly, when we consider the whole network, we have Su as follows:

$$Su = \sum_{\forall i} Su_i. \tag{32}$$

4.3.1. Uninterrupted Idle Period. The update signaling during time i to $i+1$ is $N_{U_1} O_i Su$. In an uninterrupted idle period, there are totally t MS Paging listening intervals, since an uninterrupted idle period is terminated by the timer-based update after time t. In each MS Paging listening interval, the MS checks if PA changes. The expected PA Notification signaling cost in one uninterrupted idle period is:

$$N_{U_1}\left(\sum_{i=0}^{t-1} O_i Su + \frac{1}{2}\right). \tag{33}$$

During the total duration T, the number of uninterrupted idle periods N_u is

$$N_u = \frac{\sum_{i=1}^{t} P_{t_i}(T - \lambda T(\overline{t_i}+1))}{\sum_{i=1}^{t} P_{t_i}(t+1)} = \frac{\sum_{i=1}^{t} P_{t_i} T - \lambda T\left(\overline{t_p} + P_{t_i}\right)}{\sum_{i=1}^{t} P_{t_i}(t+1)}. \tag{34}$$

The total update singling cost in all uninterrupted idle periods will be

$$\text{Update\_signaling\_un} = N_u N_{U_1}\left(\sum_{i=0}^{t-1} O_i Su + \frac{1}{2}\right). \tag{35}$$

4.3.2. Interrupted Idle Period. In an interrupted idle period, there are totally $t_p - 1$ cycles, since an interrupted idle period is terminated by message arrival at time t_p. The expected PA Notification signaling cost in an interrupted idle period is

$$N_{U_1}\left(\frac{\sum_{i=0}^{t-1} O_i Su P_{t_i}}{\sum_{i=0}^{t-1} P_{t_i}} + \frac{1}{2}\right). \tag{36}$$

During the total time duration T, the number of interrupted idle period is N_i. The total update singling cost in all interrupted idle periods will be

$$\text{Update\_signaling\_in} = N_i N_{U_1}\left(\frac{\sum_{i=0}^{t-1}(O_i Su + (1/2)) P_{t_i}}{\sum_{i=0}^{t-1} P_{t_i}} + \frac{1}{2}\right). \tag{37}$$

4.4. Timer-Based Paging Area Update. Timer-based paging area update (Timer-Based PA Update) occurs when the update timer t expires. The system recomputes the MS-centric paging area, as shown in Figure 11. In addition, the same MS-centric paging area recomputation occurs when an MS goes into active mode, which happens after a data message arrives. During T, the expected data message arrival is λT. As the signaling message flow is the same in the timer-based PA update and the paging due to data arrival, we will lump together the signaling cost into one term in this subsection.

The number of total PA update, which includes both Timer-Based PA Update and PA update due to data arrival, is $(\sum_{i=1}^{t} P_{t_i} T - \lambda T(\overline{t_p} + P_{t_i}))/\sum_{i=1}^{t} P_{t_i}(t+1)$. For each PA update, the signaling cost is denoted as N_A :

$$N_A = \text{messages} \times \text{relay} = 4 \times 2 = 8. \tag{38}$$

Notice that the N_A timer-based PA update signaling messages are the same as the PA notification signaling messages N_{U_1}, since similar signaling message flow is applied.

So the total timer-based PA update signaling cost is

$$\text{Timer\_signaling} = \lambda T N_A + \frac{\sum_{i=1}^{t} P_{t_i} T - \lambda T\left(\overline{t_p} + P_{t_i}\right)}{\sum_{i=1}^{t} P_{t_i}(t+1)} N_A. \tag{39}$$

From (23), (31), (34), (35), (37), and (39) the total signaling cost is

$$\begin{aligned} S\_\text{total} &= \lambda T N_{P_1} \frac{\sum_{i=1}^{t} O_i Sp P_{t_i}}{\sum_{i=1}^{t} P_{t_i}} + \lambda T N_{U_1} \left(\frac{\sum_{i=0}^{t-1} O_i Su P_{t_i}}{\sum_{i=0}^{t-1} P_{t_i}} + \frac{1}{2} \right) \\ &+ \frac{\sum_{i=1}^{t} P_{t_i} T - \lambda T \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_{U_1} \left(\sum_{i=0}^{t-1} O_i Su + \frac{1}{2} \right) \\ &+ \lambda T N_A + \frac{\sum_{i=1}^{t} P_{t_i} T - \lambda T \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_A. \end{aligned} \tag{40}$$

5. Optimized Timer-Based Location Update t^*

In the previous section, we derive the signaling cost given parameters p, λ, T, and t. In this section, we will optimize the timer-based update period t to minimize the overall signaling cost. The total time duration T, which is just an observation time period, does not affect the optimization results. We will normalize the formulation by defining $S_0 = S\_\text{total}/T$. After normalization of (40), we have

$$\begin{aligned} S_0 &= \lambda N_{P_1} \frac{\sum_{i=1}^{t} O_i Sp P_{t_i}}{\sum_{i=1}^{t} P_{t_i}} + \lambda N_{U_1} \left(\frac{\sum_{i=0}^{t-1} O_i Su P_{t_i}}{\sum_{i=0}^{t-1} P_{t_i}} + \frac{1}{2} \right) \\ &+ \frac{\sum_{i=1}^{t} P_{t_i} - \lambda \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_{U_1} \left(\sum_{i=0}^{t-1} O_i Su + \frac{1}{2} \right) \\ &+ \lambda N_A + \frac{\sum_{i=1}^{t} P_{t_i} - \lambda \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_A. \end{aligned} \tag{41}$$

The Ts matrix is an $S(n_r)$ by $S(n_r)$ matrix. After diagonizing the matrix, the matrix O_i is composed of eiganvalues e_1 to $e_{S(n_r)}$ and some constant values. To simplify the S_0 notation, we define E_u^i and E_p^i as follows:

$$\begin{aligned} E_u^i &= O_i Su = \sum_{k=1}^{S(n_r)} u_k e_k^i, \\ E_p^i &= O_i Sp = \sum_{k=1}^{S(n_r)} p_k e_k^i. \end{aligned} \tag{42}$$

Notice that the parameters u_k and p_k are constants, for all $k \in [1, S(n_r)]$. Then, the normalized signaling cost is

$$\begin{aligned} S_0 &= \lambda N_{P_1} \frac{\sum_{i=1}^{t} E_p^i P_{t_i}}{\sum_{i=1}^{t} P_{t_i}} + \lambda N_{U_1} \frac{\sum_{i=0}^{t-1} E_u^i P_{t_i}}{\sum_{i=0}^{t-1} P_{t_i}} + \frac{\lambda N_{U_1}}{2} \\ &+ \frac{\sum_{i=1}^{t} P_{t_i} - \lambda \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_{U_1} \left(\sum_{i=0}^{t-1} E_u^i + \frac{1}{2} \right) \\ &+ \lambda N_A + \frac{\sum_{i=1}^{t} P_{t_i} - \lambda \left(\overline{t_p} + P_{t_i} \right)}{\sum_{i=1}^{t} P_{t_i} (t+1)} N_A. \end{aligned} \tag{43}$$

After substituting (28) and (30) for P_{t_i} and t_p and some computation, we could obtain

$$\begin{aligned} S_0 &= \lambda N_{P_1} \sum_{k=1}^{S(n_r)} \frac{p_k e_k \left(1 - e^{\lambda} \right) \left(e_k^t - e^{\lambda t} \right)}{(1 - e^{\lambda t})(e_k - e^{\lambda})} \\ &+ \lambda N_{U_1} \sum_{k=1}^{S(n_r)} \frac{u_k \left(1 - e^{\lambda} \right) \left(e_k^t - e^{\lambda t} \right)}{(1 - e^{\lambda t})(e_k - e^{\lambda})} + \frac{\lambda N_{U_1}}{2} \\ &+ \frac{\lambda \left(1 - e^{\lambda t} + t \right)}{(e^{\lambda t} - 1)(t+1)} N_{U_1} \left(\sum_{k=1}^{S(n_r)} u_k \frac{\left(e_k^t - 1 \right)}{e_k - 1} + \frac{1}{2} \right) \\ &+ \lambda N_A + \frac{\lambda \left(1 - e^{\lambda t} + t \right)}{(e^{\lambda t} - 1)(t+1)} N_A. \end{aligned} \tag{44}$$

To find the optimized t, we take the first-order derivatives:

$$\begin{aligned} \frac{dS_0}{dt} = \sum_{k=1}^{S(n_r)} \Bigg\{ & \lambda (N_{P_1} e_k p_k + N_{U_1} u_k) \\ & \times \frac{\left(1 - e^{\lambda} \right)}{\left(e^{\lambda t} - 1 \right)^2} \frac{\left(\lambda e_k^t e^{\lambda t} - \lambda e^{\lambda t} + \mathfrak{A} \right)}{(e_k - e^{\lambda})} \\ & + \frac{N_{U_1} \lambda u_k}{(t+1)(e_k - 1)(e^{\lambda t} - 1)} \\ & \times \Bigg[\left(1 - \lambda e^{\lambda t} \right) \left(e_k^t - 1 \right) + \left(t - e^{\lambda t} + 1 \right) \\ & \times \left(\log(e_k) e_k^t - \frac{\lambda \left(e_k^t - 1 \right) e^{\lambda t}}{(e^{\lambda t} - 1)} - \frac{\left(e_k^t - 1 \right)}{(t+1)} \right) \Bigg] \\ & - \left(\frac{N_{U_1}}{2} + N_A \right) \lambda e^{\lambda t} \frac{1 + \lambda t + \lambda t^2 - e^{\lambda t}}{\left(e^{\lambda t} - 1 \right)^2 (t+1)^2} \Bigg\} = 0, \end{aligned} \tag{45}$$

where $\mathfrak{A}$ denotes $e_k^t \log(e_k) - e^{\lambda t} e_k^t \log(e_k)$.

By solving $dS_0/dt = 0$, we will get the optimal paging area update timer t^*.

6. Performance Evaluation

The PA Notification signaling cost decreases as t increases because, in our paging area topology, the size of paging area near the center is smaller than the size of paging area away from the center. As expected, the timer-based PA update signaling cost decreases as t increases. As t increases, the low PA update frequency reduces the signaling cost; however, the location tracking of MS becomes coarser. The broadcast paging signaling cost depends on the data message arrival rate λ. In addition, if an MS goes to *outside* state of the paging area, mostly due to infrequent paging area update, the network needs to broadcast the whole network to locate the MS. Tradeoffs between frequency of paging area update and the broadcasting cost could be observed in the figures.

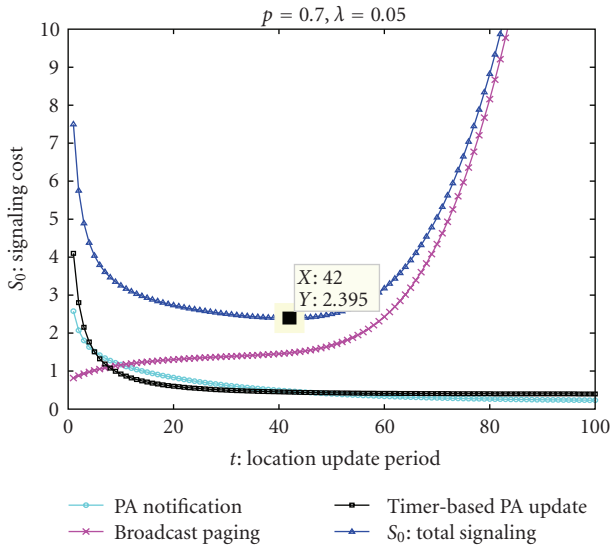


Figure 13: Signaling cost: high mobility and low message arrival rate.

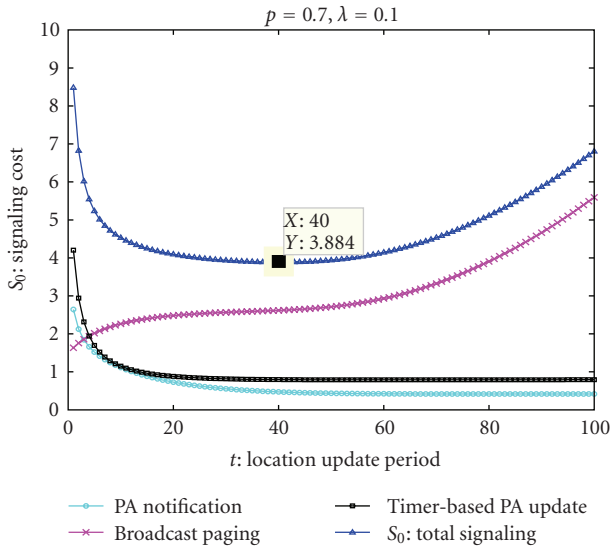


Figure 14: Signaling cost: high mobility and high message arrival rate.

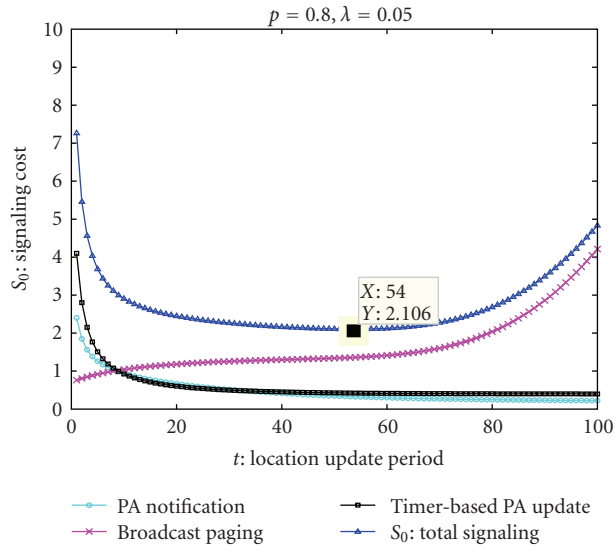


Figure 15: Signaling cost: low mobility and low message arrival rate.

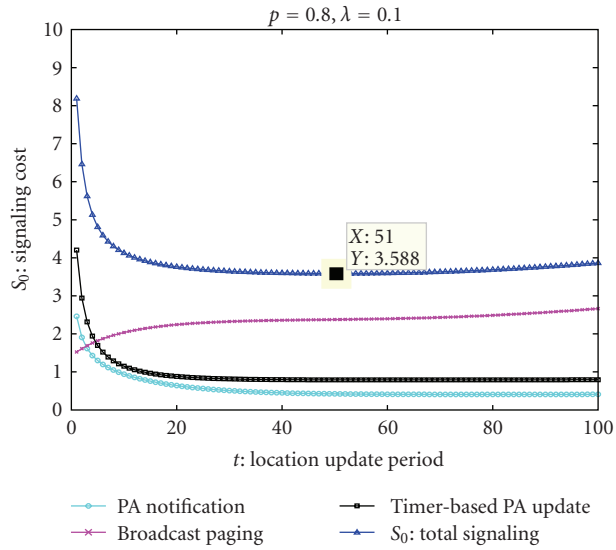


Figure 16: Signaling cost: low mobility and high message arrival rate.

6.1. Finding Optimized Location Update Timer. In Figures 13, 14, 15, and 16, we illustrate the signaling cost of the proposed IEEE 802.16j paging scheme in different mobility scenarios and paging arrival scenarios. In each figure, the three signaling cost components, PA Notification, Broadcast Paging, and Timer-Based PA Update, are shown, respectively. The optimal value of the total signaling cost S_0 is also labelled.

Figures 13 and 14 show the performance differences between a high message arrival rate (λ) scenario and a low message arrival rate scenario. In the three signaling cost components, the broadcast paging cost changes the most. With small λ, the signaling cost grows more steeply as t increases. The reason is that the broadcast paging signaling cost becomes large when the MS location is updated infrequently. When an MS receives a message more frequently, it goes into active mode more frequently. When an MS goes into active mode and then reenters the idle mode, the paging area is updated. Consequently, the MS less likely goes to outside area.

Comparing Figures 13 and 15, the mobility parameter p differs. Notice that a high p indicates the low mobility scenario since p defines the probability that an MS stays in the same cell during unit time. When MS mobility is high, the optimal t^* is smaller to keep the needed precision of location tracking.

In Figure 16, as the MS mobility is low and the data message arrival rate is high, the probability that an MS

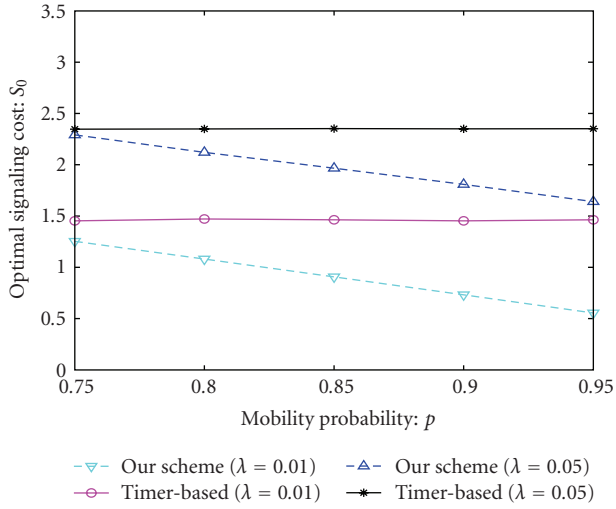


Figure 17: Comparison to timer-based scheme.

stays in the central region of the paging area is high. The probability of an MS that moves out of the paging areas is low; hence, the cases of network-wide broadcasting to find MS rarely occur. The signaling cost of Broadcast Paging component is relative flat, compared with the other three figures. Hence, the overall signaling cost is related flat when t is large.

6.2. Comparing to Pure Timer-Based Scheme. In addition to update the paging area topology when an MS does not update its location for time t, an MS notifies the network when an MS moves across the border of paging areas in the proposed paging scheme. On the contrary, a pure timer-based paging algorithm might only update the location of an MS only when the t timer expires. In Figure 17, we compare the proposed scheme and the pure timer-based scheme. The proposed scheme has a lower signaling cost than the pure timer-based scheme as shown in the figure.

7. Conclusion

In this work, we investigated the paging and location management scheme in the IEEE 802.16j multihop relay networks. The paging scheme is compatible with the idle mode operation in the IEEE 802.16j standard and integrates with the paging area design and timer-based location update mechanism scheme. We propose a generalized random walk mobility model that is suitable for investigating user mobility in multihop cellular relay system, for example, IEEE 802.16j. The analytical mobility model is shown to match the simulation results. We applied this random walk mobility model to analyze the proposed paging scheme. The proposed scheme performs well compared to naive timer-based scheme. In addition, the proposed paging area update optimization has been shown to minimize the signaling cost effectively. In the future, we plan to further invest age advanced paging and location update algorithms to further enhance the signaling cost and paging delay. Moreover, nonrandom-walk mobility model for IEEE 802.16j is an interesting future work item to study. Advanced paging and location update scheme over generalized user mobility model will play a critical role in optimization the IEEE 802.16j relay network.

Notations

$A(x1, x2)$:	Absolute Geographical Location
Ts:	Transition matrix
p:	Probability of MS stay in the same cell at next cycle
T:	Total time of MS operation
q:	Probability of MS stay in a different cell at next cycle
$R(u, k)$:	Relative Moving Distance
u, k:	Index of a Relative Moving Distance cell
n_r:	Number of tiers in Relative Moving Distance topology
S_i:	The number of states
$P^t_{R(u,k)}$:	Probability at state $R(u, k)$
$S(n)$:	Number of states at the nth tier
O_i:	Each state's probability at time i
i:	After i time slots of the timer-based location update
t:	The timer for timer-based location update
D:	Diagonal matrix of Ts
V:	Eigenvector matrix of Ts
Sp:	Total paging signaling cost matrix
Sp_i:	Paging signaling cost matrix in paging area i
Su:	Total update signaling cost matrix
Su_i:	Update signaling cost matrix in paging area i
S:	Total signaling cost
$p_1 \dots p_{65}$:	Coefficients of eigenvalues in matrix Sp
$e_1 \dots e_{65}$:	Eigenvalue of matrix Ts
$u_1 \dots u_{65}$:	Coefficients of eigenvalues in matrix Su
N_i:	Number of interrupted idle period
N_u:	Number of uninterrupted idle period
N_{U_1}:	Signaling cost in 1 MS update operation
N_{P_1}:	Signaling cost in 1 broadcast paging operation
N_A:	Signaling cost in 1 paging area update operation
t_p:	Message arrival time
$\overline{t_p}$:	Average message arrival time
n_p:	Message arrival number
S_{total}:	Total signaling cost during T
S_0:	Normalized signaling cost in one time slot
t^*:	The optimal value of update timer t.

Acknowledgment

This work was partly supported by the Industrial Technology Research Institute (ITRI).

References

[1] "IEEE Standard for Local and metropolitan area networks—part 16: Air Interface for Broadband Wireless Access Systems," IEEE Std 802.16-2009, May 2009.

[2] "IEEE Standard for Local and metropolitan area networks—part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 1: Multihop Relay Specification," IEEE Std 802.16j-2009, May 2009.
[3] S. W. Peters and R. W. Heath Jr., "The future of WiMAX: multihop relaying with IEEE 802.16j," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 104–111, 2009.
[4] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 115–121, 2009.
[5] S.-R. Yang, Y.-C. Lin, and Y.-B. Lin, "Performance of mobile telecommunications network with overlapping location area configuration," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1285–1292, 2008.
[6] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: to update or not to update?" *Wireless Networks*, vol. 1, no. 2, pp. 175–185, 1995.
[7] J. S. M. Ho and I. F. Akyildiz, "Mobile user location update and paging under delay constraints," *Wireless Networks*, vol. 1, no. 4, pp. 413–425, 1995.
[8] Y.-H. Zhu and V. C. M. Leung, "Derivation of moving distance distribution to enhance sequential paging in distance-based mobility management for PCS networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 11, pp. 3029–3033, 2006.
[9] C. K. Ng and H. W. Chan, "Enhanced distance-based location management of mobile communication systems using a cell coordinates approach," *IEEE Transactions on Mobile Computing*, vol. 4, no. 1, pp. 41–55, 2005.
[10] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th Annual International Conference on Mobile Computing and Networking (MobiCom '07)*, pp. 123–134, Montreal, Canada, September 2007.
[11] Z. Liu and T. D. Bui, "Dynamical mobile terminal location registration in wireless PCS networks," *IEEE Transactions on Mobile Computing*, vol. 4, no. 6, pp. 630–639, 2005.
[12] X. Wu, B. Mukherjee, and B. Bhargava, "A crossing-tier location update/paging scheme in hierarchical cellular networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 839–848, 2006.
[13] Y. Xiao, H. Chen, and M. Guizani, "Performance evaluation of pipeline paging under paging delay constraint for wireless systems," *IEEE Transactions on Mobile Computing*, vol. 5, no. 1, pp. 64–76, 2006.
[14] I. F. Akyildiz, Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "A new random walk model for PCS networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 7, pp. 1254–1260, 2000.
[15] I. F. Akyildiz and W. Wang, "A dynamic location management scheme for next-generation multitier PCS systems," *IEEE Transactions on Wireless Communications*, vol. 1, no. 1, pp. 178–189, 2002.
[16] Y. Zhang and M. Fujise, "Energy management in the IEEE 802.16e MAC," *IEEE Communications Letters*, vol. 10, no. 4, pp. 311–313, 2006.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 602973, 7 pages
doi:10.1155/2010/602973

Research Article

Seamless Video Session Handoff between WLANs

Claudio de Castro Monteiro,[1] Paulo Roberto de Lira Gondim,[2] and Vinícius de Miranda Rios[3]

[1] *Computation Department, Federal Institute of Education, Science and Technology of Tocantins IFTO, Palmas 77.021-090, Brazil*
[2] *Electrical Engineering Department, Faculty of Technology, University of Brasilia UnB, Brasília 70.910-900, Brazil*
[3] *Informatics Department, University of Tocantins UNITINS, Tocantins, Brazil*

Correspondence should be addressed to Claudio de Castro Monteiro, ccm.monteiro@ieee.org

Received 1 October 2009; Accepted 16 December 2009

Academic Editor: Francisco Falcone

Handoff in a distributed IEEE 802.11 Wireless LAN network is a source of significant amount of problems on the video transmission environment. The visual quality of video streaming applications is lowered when stations are in handoff status. In this paper, we introduce an architecture of a session proxy (SP), which tries to preserve the quality of the streaming video upon each handoff between access points. We have evaluated thresholds of RSSI and Loss Frame Rate (LFR) for deciding the moment when the handoff process shall begin. Our solution performance was evaluated in a testbed implementation for MPEG-4 video on demand with one video server (VLS) and two FreeBSD-based access points supporting Mobile IP, DHCP Server and IAPP approach.

1. Introduction

Nowadays, the most used pattern for WLANs by the market is IEEE 802.11 [1] and its extensions such as 802.11a [2] 802.11b [3], 802.11g [4], 802.11e [5], 802.11n [6], among others.

Several studies have been conducted with the intention of analyzing the advantages and disadvantages of the use of wireless networks [7–10], also approaching possible methods or mechanisms to avoid or reduce the problems inherent in their use.

The success story of 802.11 Wireless LAN can be attributed due to its high bit rate, easy installation, and low price. The 802.11 MAC protocol originally has objective to work at the home or office environment, but nowadays the IEEE is extending the protocol towards mobile environments, with direct application for data delivery to distant access, integrating branches of the 3G technology [1]. However, currently, seamless session continuity is still out of reach, especially for video streaming applications. The first step to achieve session continuity during handoffs in WLAN was made by the IEEE 802.11f Inter Access Point Protocol (IAPP) [11], that recommend this as a good practice. In order to limit the packets loss due to the network disconnection of a wireless client during handoff, this standard recommend, the transfer of the "context" from the previous access point to the next. This technique can work very well for nonreal time applications and transport protocols such as web browsing using TCP. We will show in this paper that this is not the case for video streaming real-time applications, especially for streaming video on demand.

The focus of our work is to preserve real-time video streaming session during handoff process in WLANs. For this, we analyzed fading of the wireless signal. However, in case of a handoff between two WLAN access points, a sudden loss of packet occurs and the mobile node will not be able to preserve the visual quality. That is the reason we chose to analyze the Rate Frame Loss (RFL) also, trying to identify the moment of handoff process start.

We can find studies about this problem. Some use techniques of crosslayer to adapt the video quality when WLAN is congested [12]. In our network, the unique source of packet loss is handoff related.

Other approaches adopt frames network retransmission, changing the ARQ mechanism, using information from link layer to adapt the frames' retransmission [13].

Here we assume that our Session Proxy (SP) and Access Points (APs) have buffer enough to store packets to overcome the delay variation and frame loss rate caused by the handoff. Thus SP always has enough data to send to AP and the AP to the mobile node (MN).

In this paper we propose a solution based on a Session Proxy, located in the mobile operator network. We assume Access Points (AP) architecture with IP router, Mobile IP, and 802.11-IAPP functionalities. The SP is RTSP session aware and tries to preserve the quality streaming video, during handoff process in WLAN. We evaluated the performance of our solution and it has been compared to the standard IAPP approach. In the next sections we are going to comment about related works in the literature, our network solution architecture, experiment methodology, and testbed used and show the results. We finish with a conclusion.

2. Related Works

The problem caused on video quality by the handoff in WLAN has been discussed by some works. However, this problem has been divided in two parts: the part that study adaptation forms WLANs for video streams traffic; and part which studies forms to keep adaptation when a handoff process occurs. We analyze works that follow these two scenarios.

In [14] is proposed a novel mechanism of RTS classification based on stations transmission rate. This work aims to control the multitransmission rate anomaly in 802.11 networks, improving video streaming quality to receivers.

A proposal of novel adaptive algorithm that improves the efficiency of datagram streaming over IEEE 802.11 networks is presented in [15]. It uses the signal quality information to adapt the transmission and therefore improves the network utilization. This work estimates thresholds based on SNR and packets loss rate to adapt stream application.

A proposal of a handoff study in Mobile IP networks and Mobile IP Protocol Extensions for Handoff Latency Minimization was showed in [16], indicating that native Mobile IP has high handoff latency and that its proposed to improve in 15% the performance of handoff latency.

In [17] was proposed a proxy-based multimedia scheme for control Real-Time Streaming Protocol (RTSP) to support fast signaling at home network. The testbed implementation showed that the proposed scheme improves the performance compared with RTSP in terms of latency time, but not resolve the RTSP session continuity problem. The proposal reduces latency time but the loss rate is big enough for RTSP session not to continue.

A proposal of an Ethernet Soft Switch architecture to solve the problem of frame loss during handoff process at video streaming transmission is present in [18]. In this work, on-demand video streams were transmitted to mobile node while it moves between access points. In these experiments, there were limited resources and mobile node had enough cache for receive the frames in the access points. The base of the proposal is to establish different retransmissions methods for I, B, and P frames, to keep the received video stream quality.

A discussion about how WLAN roaming habilities are affected by new standards is present in [19]. The standards considered were IEEE 802.11i, IEEE 802.11e, and new IEEE 802.11r. This last one was developed to address issues faced by real-time applications that implement the service's security and quality enhancements. The performance evaluation of 802.11r prototype and the 802.11i baseline mechanisms shows a voice application using 802.11r to achieve significantly shorter transition time and reduced packet loss during AP-AP transition and can therefore realize a noticeable improvement in voice quality, but nothing is noticed about video streaming transmission.

In [20], is proposed a low-latency Mobile IP handoff scheme that can reduce the handoff infrastructure's latency mode in wireless LANs to less than 100 milliseconds. The proposal tries to resolve the mobility intra-WLAN measuring multiple AP's signal strength working in infrastructure mode. It accelerates the detection of link-layer handoff by replaying cached foreign agent advertisements. The proposal is transparent to the Mobile IP software installed on mobile and wired nodes. The authors show how efficient the proposal is, with a mechanism of bandwidth guarantee in 802.11e-based standard wireless LAN. This implementation does not predict mobile node's handoff, leaving this work under responsibility of IAPP mechanism. It proposes an acceleration handoff's detection.

In work developed in [21], one analytical modeling of handoff latency for FMIPv6 and HMIPv6, using WLANs as access networks, was present. This model considers factors of both link and network layer that influences the Mobile IP handoff delay. The results show an improving performance in the MIPv6, which help in the handoff process. However, the solution forces clients to have support MIPv6.

In [22] is proposed a framework for multimedia delivery and adaptation in mobile environments. This work introduces the concept of Personal Address (PA), which is a network address associated to the user instead of a network interface. The proposed framework works at the network layer and it moves the PA among networks and devices to deliver media in a seamless and transparent way. The authors claim that location's transparency sponsored by PA allows the user to receive multimedia data independent of the IP network. However, the solution presented uses Mobile IP and do not show the impact generated in the transmission multimedia session continuity, caused by implementing the entities managed by PAs.

All related works studied try to resolve problems in video streams quality in 802.11 networks. Some tries to test technologies with Mobile IP, others to implement IEEE 802.11f and its recommendations, and others yet to bring new concepts with "personal address." However, this problems increase when there is one video stream transmission during handoff process. Usually, video stream sessions have a synchronization time that does not support the handoff latency between two access points. The studies found in the literature handle problems with enlace retransmission

```
receive_socket(socket, RTSP_request);
registered_session(session_ID, RTSP, IP_MAC, 0);
open_socket(socket1, IP_server);
send_socket(socket1, RTSP_request);
receive_socket(socket1, RTSP_response);
send_socket(socket, RTSP_response);

while(Session_ID < > 0)
{
        receive_socket(socket, RTSP_packets);
        receive_socket(socket2, status, MAC_AP);
        if(status==1)
        {
                FrameID=frame_ID;
                start_cache(Session_ID);
        }
        if(status < >1)
        {
                send_socket(socket1, RTSP_packets);
        }
        sendcache_socket(socket1, RTSP_packets);
}
```

Algorithm 1

techniques, with the separation frames types and delivering only the necessary or usually with application Mobile IP and IAPP's technologies. Then, our solution is based on set that meets Mobile IP, IAPP, AP router based, and the Session Proxy (SP). The proposal tries to resolve the session continuity problem after handoff, ensuring the transmitted video's quality on receiver (PSNR).

3. Proposal

In our proposal, we suggest the insertion, in the architecture of wireless operator, of two components: a session proxy (SP), and an 802.11 access point FreeBSD-based with IP router, DHCP server, and IAPP functionality. In Figure 1, these components and its links can be seen.

3.1. Functionality. The main idea is to use the SP to ensure the session's continuity even after long periods of link's discontinuity, using for this, the prediction of handoff of the MN, through the thresholds defined after extensive experiments detailed at session B and displayed in Table 1.

Thus, the MN authenticates is associated with AP1 and receives an IP address dynamically through DHCP server. The MN requests an open session's RTSP with the video server. This request will be received by SP, registered with the structure shown in Table 2 and then forward to the video server, according with Algorithm 1.

The video server then opens an RTSP session with the SP, which will begin to receive the frames, transferring them to AP1, which deliver it to MN. This process will continue up until the AP1 that identifies the mobile node is coming at handoff zone (where RSSI and LFR are at BETA level), starting the frame cache then indicating to SP for start frame cache also. At this point, AP1 cache frames intended to mobile node and SP cache frames intended to AP1, using the data structure shown in Table 2.

Table 1: Thresholds for prediction of handoff.

ALFA	RSSI > 40	LFR < 10%	PSNR > 35
BETA	40 ≥ RSSI > 30	LFR < 20%	29 > PSNR > 26
GAMA	30 ≥ RSSI	LFR ≥ 20%	PSNR < 18

Table 2: Session registration cache structure.

Session ID	Service ID	IP association	Frame ID

When the MN reaches the GAMA level, the AP1 records in the session registration cache the identifier of the last frame received by the MN and continues with the video server session open, receiving frames, inserting in the cache and transmitting to AP1, which will also be doing caching of frames received. Record done, AP1 finishes the association with the MN and informs the SP that the mobile is not in its association's list. This fact informs to AP1 that must start transmission of frames in its cache since the last frame that was received by the MN should be sent to AP2 via IAPP.

3.2. Handoff Decision. To achieve these thresholds, we performed 200 video stream transfers in the MPEG-4 format, for each of the three scenarios below, that was obtained with the average results of the values expressed in Table 1 and in Figures 2 and 3.

Thus, to predict the handoff of the mobile node, the APs uses the Algorithm 2 to determine the signal levels of the link mobile, starting so the cache of frames.

After the frames start being cached by AP1 and SP, the MN starts the GAMA level, which will have its RTSP session open with the SP discontinued and their frames will be saved in their caches. Therefore, if the MN is back to BETA level, associated with either AP1 or AP2, it will receive the video from the next frame after the last received, generating a guarantee of delivery the entire video's contents.

4. Testbed Scenario

To validate our proposal, we set up a scenery's piece illustrated in Figure 1. We use a set of software and hardware that generate the desired scenario's implementation.

In our testbed, we use three computers with VLS [23] doing RTSP video stream, one computer doing the SP functions, two access points FreeBSD-based with Mobile IP KAME [24], and IAPP implementations.

The links *video servers* → *SP*, *SP* → *AP*, and *AP* → *AP* at 100 Mbps and links *AP* → *MN* at 54 Mpbs.

Each station can establish AP connection if and only if its transmission rate is equal or higher than 2 Mbps, according to selection RTS mechanism proposed for [14].

The APs were configured in channels 1 and 11, respectively, to avoid adjacent channel interference.

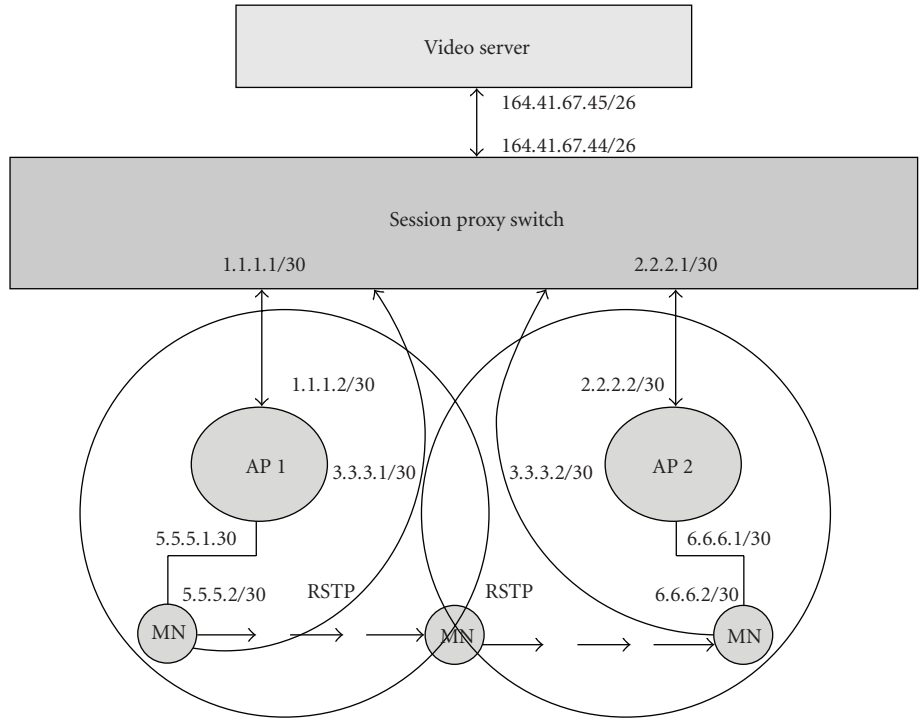


Figure 1: Proposed elements in our testbed scenery.

```
l_loss=icmp_request(IP_MN);
l_rssi=rssi_verify(MAC_MN);
if (l_rssi <= 40 and l_rssi > 30)
{
        if(l_loss < 20)
        {
                send_SP(1, MAC_AP);
                start_cache(sessao_ID, ID_Frame);
        }
}
else if ( l_rssi <= 30)
   {
        if (l_loss >= 20 )
        {
                send_SP(2,ID_Frame);
                handoff(MAC);
                start_cache(sessao_ID,ID_Frame);
                send_IAPP(MAC_AP, ID_Frame+1);
        }
}
...
```

Algorithm 2

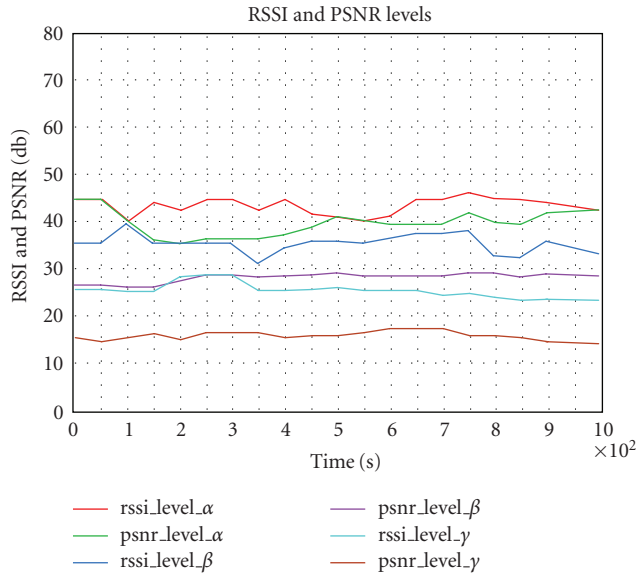


Figure 2: Thresholds RSSI and PSNR levels.

A reduced and expert version of FreeBSD operating system was developed [25] and embedded on IDE flash card. Each AP has three network interfaces: two IEEE 802.3 at 100 Mbps and one at IEEE 802.11g at 54 Mbps with Atheros chipset.

For tests, we use a video file with 16.6 minutes, at MPEG-4 format. This video was stored at video server and streamed for VLS to SP at 30 fps. The video was streamed 200 times at scenarios shown in Table 3. Use the UNIX *ifconfig* command in AP reducing RSSI levels during the time transmission in order to simulate the changes in proposed levels (MN movement). The results are the average of these 200 transmissions.

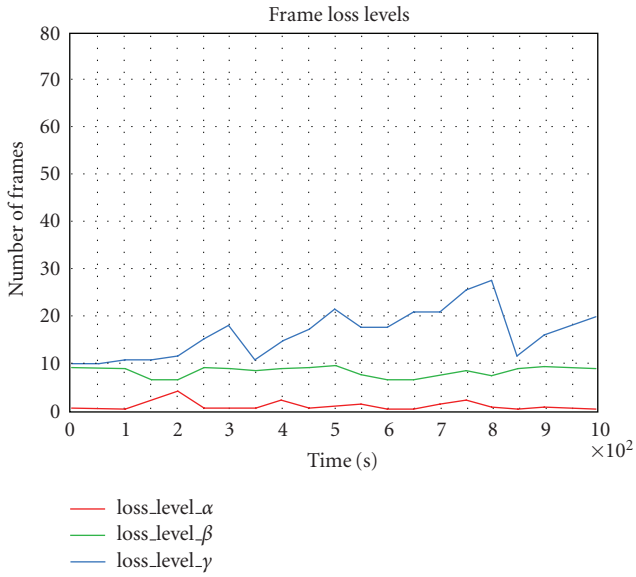


Figure 3: Thresholds packets LOSS levels.

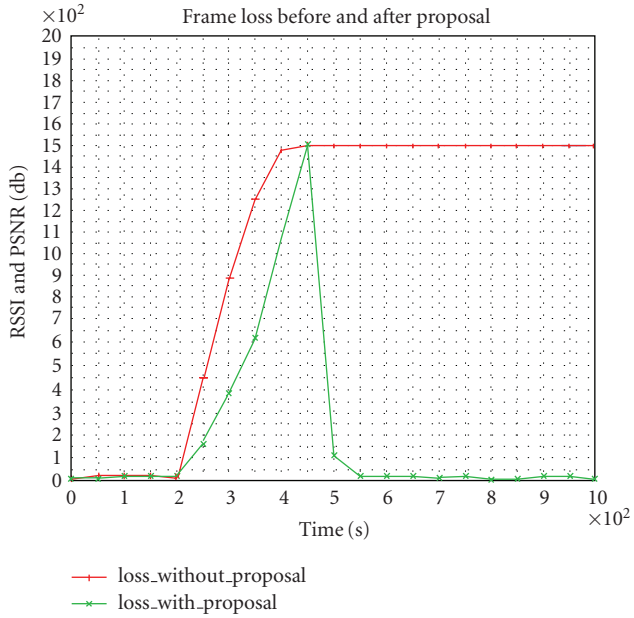


Figure 4: Average of number of loss frames during transmissions without and with proposal.

5. Obtained Results

After the experiments, notice that the farther from the AP is MN, in other words, approaching the limits of his cell, the MN has reduced its level of RSSI. In the configured environment with Mobile IP and IAPP, the level of the MN's RSSI reaches zero at the physical handoff, recovering their intensity once MN is associated to the new AP.

The time between the link-off of the old AP and link-on in the new AP, taking into account its authentication, combined with the time taken by the DHCP server to provide an IP address to the MN and the time of negotiation between

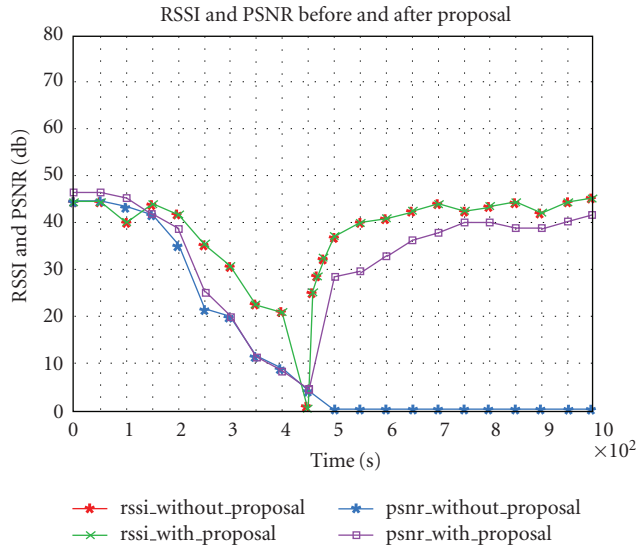


Figure 5: Average of PSNR and RSSI during transmissions without and with proposal.

Table 3: Scenarios for obtaining of thresholds.

Scenario 1	An AP1 at channel 10 An AP2 at channel 09 (adjacent channel interference) An station without movement at 2 m of AP1
Scenario 2	An AP1 at channel 10 An AP2 at channel 09 (adjacent channel interference) An station without movement at 10 m of AP1
Scenario 3	An AP1 at channel 10 An AP2 at channel 09 (adjacent channel interference) An station without movement at 25 m of AP1

the HA and FA was in our experiment, about 10 seconds, enough time to RTSP started session with the server to be closed by an absolute inability of the protocol to resequence the frames lost (in the case 30 fps × 10 s = 300 frames).

Thus, without the application of SP, proposed in this work, the level packets loss generated by the handoff between APs reaches 1500 frames in the interval of 50 seconds, showing a total connection loss. After the handoff done, the RTSP session is lost and the frames' level lost does not recover anymore, remaining in 1500, as shown in Figure 4.

The visual impact on the quality of received video is large. Considering that the PSNR measured every 50 seconds of transmission can be seen in Figure 5; after the handoff, the PSNR values remain at zero until the end of transmission, considering the permanent loss session's RTSP.

While analyzing Figures 4 and 5, we can see that with the implementation of our proposal the frame loss is not prevented during the handoff, but we signal to the SP and the APs, to the cache of frames transmitted to the MN, delivering the same to it, as soon as the association with the other

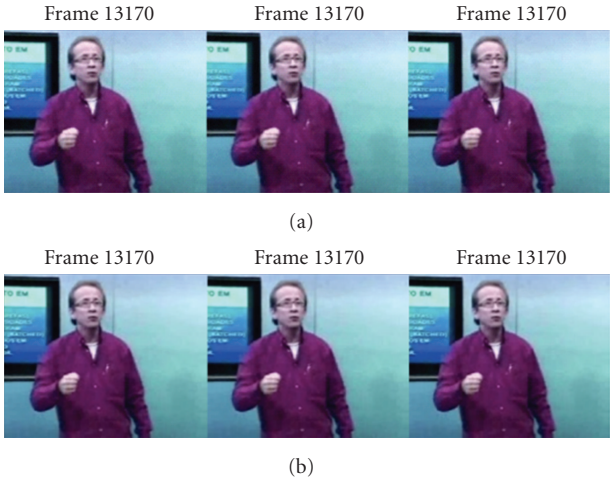


Figure 6: Video sequence after and before handoff without proposal.

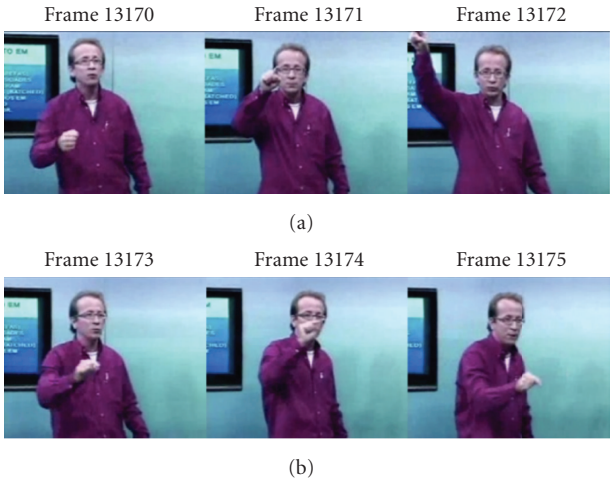


Figure 7: Video sequence after and before handoff with proposal.

AP is complete and that the RSSI level is sufficient (BETA level). This allows the packets lost recovery, reducing the frame loss' rate after the handoff, so MN receives the frames that was not received during the connection discontinuation. This increases the average PSNR of the video forwarded, monitored in the transmission each 50 seconds.

In Figure 6 is shown an example of sequence frames received before and after handoff, between 400 and 600 seconds. MN receives the frame 13170 at 439. After this time, MN entering in GAMA level ends the AP that does not send next frames. Without our proposal implementation, notice that PSNR measured at Figure 5 remains in zero after 450 seconds, due to high-loss frame.

Moreover, Figure 7 shows other sequence frames, with our proposal implementation. Note that different frames continue being received, increasing the PSNR values. The same way, at 400 seconds, the MN receives frame 13170. After this time, our SP mechanism works the cache frames. Then, after 490 seconds, the MN reaches the acceptable BETA RSSI level (after handoff) and receives the frames 13171 and all others from the video.

We can verify that the advantage of our proposal is preserve the continuity session, ensuring that user in MN receive all video's content.

6. Conclusions

After experiments being conducted, we concluded that during handoff between access points, the use of IAPP and Mobile IP is not sufficient to solve continuity frames problems, as a result of long time passed during handoff, generating high packet loss. The SP's idea brought higher implementation flexibility, considering that it acts in networks level, receiving physical level information to decide the moment that comes before physical handoff.

Our proposal offers a good solution for IPTV scenarios, with delivery video-on-demand and live transmissions (without interaction) at last miles, where users can move it between APs forming BSSs (typically airports, bus stations, shoppings, university campus, etc).

As future works, we can quote the application of our proposal in ubiquitous environment, considering two access networks: UMTS and WLAN. We want to show that SP implementation works well with heterogeneous networks too, taking that implements the level sensitivity at mobile node.

References

[1] IEEE 802.11r—Fast Roaming/Fast BSS Transition.
[2] IEEE802.11a-std, 1999, http://www.ieee802.org/11/.
[3] IEEE802.11b-std, 1999, http://www.ieee802.org/11/.
[4] IEEE802.11g-std, 1999, http://www.ieee802.org/11/.
[5] IEEE802.11e-std, 1999, http://www.ieee802.org/11/.
[6] Y. Xioa, "IEEE 802.11N: enhancements for higher throughput in wireless Lans," in *Proceedings of the IEEE Wireless Communications & Networking Conference (WCNC '05)*, Dracena, Brazil, December 2005.
[7] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *Proceedings of the 22nd IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM '03)*, vol. 2, pp. 836–843, San Franciso, Calif, USA, April 2003.
[8] M. Fonseca, E. Jamhour, C. Mendes, and A. Munaretto, "Extensão do mecanismo RTS/CTS para otimização de desempenho em redes sem fio," in *Proceedings of the 25th Simpósio Brasileiro de Telecomunicações*, 2007.
[9] G. Bianchi, L. Fsatta, and M. Oliveri, "Perfomance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LAN's," in *Proceedings of the 7th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '96)*, pp. 392–396, Taipei, Taiwan, October 1996.
[10] Y. Xiao, "IEEE 802.11n: enhancements for higher throughput in wireless Lans," *IEEE Wireless Communications*, vol. 12, no. 6, pp. 82–91, 2005.
[11] IEEE802.11f-2003. IEEE Trial-Use Recommended Pratice for Multi-vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution System Supporting IEEE802.11 Operation.

[12] G. Convertino, D. Melpignano, E. Piccinelli, F. Rovati, and F. Sigona, "Wireless adaptative video streaming by real-time channel estimation and video transcoding," in *Proceedings of the International Conference on Consumer Electronics (ICCE '05)*, pp. 179–180, Singapore, December 2005.

[13] P. Bucciol, G. Davini, E. Masala, E. Filippi, and J. C. De Martins, "Cross layer perceptual ARQ for H.264 video streaming over 802.11 wireless networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 5, pp. 3027–3031, Dallas, Tex, USA, 2004.

[14] C. de Castro Monteiro and P. R. Gondim, "Improving video quality in 802.11 networks," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, Rio de Janiero, Brazil, April 2009.

[15] AF Conceição and F. Kon, "Desenvolvimento de aplicações adaptativas para redes IEEE 802.11," in *Proceedings of the 24th Simpósio Brasileiro de Redes de Computadores (SBRC '06)*, Prague, Czech Republic, March 2006.

[16] R. Malekian, "The study of handover in mobile IP networks," in *Proceedings of the 3rd International Conference on Broadband Communications (BROADCOM '08)*, Pretoria, Gauteng, South Africa, November 2008.

[17] J.-M. Lee, M.-J. Yu, S.-G. Choi, and B.-S. Seo, "Proxy-based multimedia signaling scheme using RTSP for seamless service mobility in home network," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 481–486, 2008.

[18] T. Van Leeuwen, I. Moerman, and P. Demeester, "Preserving streaming video quality in mobile wireless LAN networks," in *Proceedings of the 63rd IEEE Vehicular Technology Conference (VTC '06)*, vol. 2, pp. 971–975, Melbourne, Australia, May 2006.

[19] S. Bangolae, C. Bell, and E. Qi, "Performance study of fast BSS transition using IEEE 802.11r," in *Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC '06)*, vol. 2006, pp. 737–742, Vancouver, Canada, July 2006.

[20] S. Sharma, N. Zhu, and T.-C. Chiueh, "Low-latency mobile IP handoff for infrastructure-mode wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 4, pp. 643–652, 2004.

[21] J. Xie, I. Howitt, and I. Shibeika, "IEEE 802.11-based mobile IP fast handoff latency analysis," in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 6055–6060, Glasgow, Scotland, June 2007.

[22] R. Bolla, S. Mangialardi, R. Rapuzzi, and M. Repetto, "Streaming multimedia contents to nomadic users in ubiquitous computing environments," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, Rio de Janiero, Brazil, April 2009.

[23] VideoLan Software Suite, http://www.videolan.org/.

[24] L. Stewart, M. Banh, and G. Armitage, "Implementing an IPv6 and Mobile Ipv6 testbed using FreeBSD 4.9 and KAME," CAIA Technical Report, 2004.

[25] C. de Castro Monteiro, http://www.bacuri.org/.

Hindawi Publishing Corporation
Journal of Computer Systems, Networks, and Communications
Volume 2010, Article ID 793807, 16 pages
doi:10.1155/2010/793807

Research Article

Multimode Flex-Interleaver Core for Baseband Processor Platform

Rizwan Asghar and Dake Liu

Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden

Correspondence should be addressed to Rizwan Asghar, rizwan@isy.liu.se

Received 25 August 2009; Accepted 12 October 2009

Academic Editor: Rashid Saeed

This paper presents a flexible interleaver architecture supporting multiple standards like WLAN, WiMAX, HSPA+, 3GPP-LTE, and DVB. Algorithmic level optimizations like 2D transformation and realization of recursive computation are applied, which appear to be the key to reach to an efficient hardware multiplexing among different interleaver implementations. The presented hardware enables the mapping of vital types of interleavers including multiple block interleavers and convolutional interleaver onto a single architecture. By exploiting the hardware reuse methodology the silicon cost is reduced, and it consumes 0.126 mm^2 area in total in 65 nm CMOS process for a fully reconfigurable architecture. It can operate at a frequency of 166 MHz, providing a maximum throughput up to 664 Mbps for a multistream system and 166 Mbps for single stream communication systems, respectively. One of the vital requirements for multimode operation is the fast switching between different standards, which is supported by this hardware with minimal cycle cost overheads. Maximum flexibility and fast switchability among multiple standards during run time makes the proposed architecture a right choice for the radio baseband processing platform.

1. Introduction

Growth of high-performance wireless communication systems has been drastically increased over the last few years. Due to rapid advancements and changes in radio communication systems, there is always a need of flexible and general purpose solutions for processing the data. The solution not only requires adopting the variances within a particular standard but also needs to cover a range of standards to enable a true multimode environment. The symbol processing is usually done in baseband processors. A fully flexible and programmable baseband processor [1–3] provides a platform for true multimode communication. To handle the fast transition between different standards, such type of platform is needed in both mobile devices and especially in base stations. Other than symbol processing, one of the challenging area is the provision of flexible subsystems for forward error correction (FEC). FEC subsystems can further be divided in two categories, channel coding/decoding and interleaving/deinterleaving. Among these categories, interleavers and deinterleavers appeared to be more silicon consuming due to the silicon cost of the permutation tables used in conventional approaches. For multistandard support devices the silicon cost of the permutation tables can grow much higher, resulting in an unefficient solution. Therefore, the hardware reuse among different interleaver modules to support multimode processing platform is of significance. This paper presents a flexible and low-cost hardware interleaver architecture which covers a range of interleavers adopted in different communication standards like HSPA Evolution (HSPA+) [4], 3GPP-LTE [5], WiMAX; IEEE 802.16e [6], WLAN; IEEE 802.11a/b/g [7], IEEE 802.11n [8], and DVB-T/H [9].

Interleaving plays a vital role in improving the performance of FEC in terms of bit error rate. The primary function of the interleaver is to improve the distance properties of the coding schemes and to disperse the sequence of bits in a bit stream so as to minimize the effect of burst errors introduced in transmission [10, 11]. The main categories of interleavers are block interleavers and convolutional interleavers. In block interleavers the data are written row wise in a memory configured as a row-column matrix and then read column-wise after applying certain intra-row and inter-row permutations. They are usually specified in the

form of a row-column matrix with row and/or column permutations given in tabular form, however; they can also be specified by a modulo function having more complex functions involved to define the permutation patterns. On the other hand, convolutional interleavers use multiple first-in-first-out (FIFO) cells with different width and depth. They are defined mainly by two parameters, the depth of memory cells and number of branches.

Looking at the range of interleavers used in different standards (Table 1) it seems difficult to converge to a single architecture; however, the fact that multimode coverage does not require multiple interleavers to work at the same time provides opportunities to use hardware multiplexing. The multimode functionality is then achieved by fast switching between standards. This research is to merge the functionality of different types of interleavers into a single architecture to demonstrate a way to reuse the hardware for a variety of interleavers having different structural properties. The method in general is the so-called hardware multiplexing technique well presented in [12]. It starts at analyzing and profiling multiple implementation flows, identifying opportunities of hardware multiplexing, and eventually fine tuning the microarchitecture, using minimal hardware, and maximal reuse of multifunctions.

This paper is organized as follows. Section 2 presents the previous work done for the interleaver algorithm implementations. The challenges involved to cover the wide range of standards are mentioned in Section 3. It also presents a shared data flow and hardware cost associated with different implementations. Section 4 provides the detailed explanation of the unified interleaver architecture and its subblocks. A brief explanation of the algorithmic transformations and optimizations used for efficient mapping onto single architecture is given in Section 5 with selected example cases. The usage of the proposed architecture while integrating into baseband system is explained in Section 6. Section 7 provided the VLSI implementation results and comparison to others followed by a conclusion in Section 8.

2. Previous Work

A variety of interleaver implementations having different structural properties have been addressed in literature. The main area of focus has been low cost and throughput. Most of the work covers a single or a couple of interleaver implementations which is not sufficient for a true multimode operation. The design of interleaver architecture for turbo code internal interleaver has been addressed in [13–17]. Some of these designs targeted very low-cost solutions. A recent work in [18] provides a good unified design for different standards; however, it covers only the turbo code interleavers and does not meet the complete baseband processing requirements demanding an all-in-one solution. The work in [19–22] covers the DVB-related interleaver implementations. Literature [23–27] focuses on more than one interleaver implementations with reconfigurability for multiple variants of wireless LAN and DVB. High-throughput interleaver architectures for emerging wireless communications based on MIMO-OFDM techniques have been addressed in [25, 27]. These techniques require multiple-stream processing in parallel, thus requiring parallel addresses generation and memory architecture as shown in Figure 1.

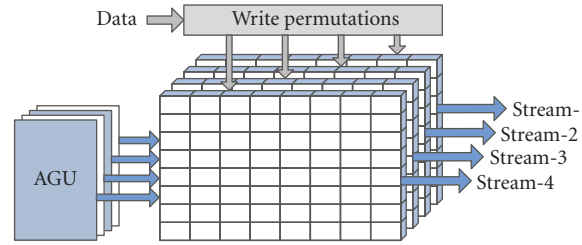


Figure 1: 3D view of interleaver configuration for a multistream communication system.

Some commercial solutions [28–30] from major FPGA vendors are also available for general purpose use. The available literature reveals that they do not compute the row or column permutations on the fly; instead they take row or column permutation tables in the form of a configuration file as input and use them to generate the final interleaved address. In this way, the complexity for on-the-fly computation of permutation patterns is avoided. This approach needs extra memory to store the permutation patterns. As these implementations are targeted for FPGA use only, they also enjoy the availability of dual port block RAM, which is not a good choice for chip implementations.

3. Shared Data Flow and Algorithm Analysis

The motivation of the research is to explore an all-in-one reconfigurable architecture which can help to meet fast time-to-market requirements from industry and customers. A summary of targeted interleaver implementations which are being widely used is provided in Table 1. The broadness of the interleaving algorithms gives rise to many challenges when considering a true multimode interleaver implementation. The main challenges are as follows:

(i) on the fly computation of permutation patterns,

(ii) wide range of interleaving block sizes,

(iii) wide range of algorithms,

(iv) fast switching between different standards,

(v) sufficient throughput for high-speed communications,

(vi) maximum standard coverage,

(vii) acceptable silicon cost and power consumption.

Exploring the similarities between different interleaving algorithms a shared data flow in general is shown in Figure 2. This data flow is shared by different interleaver types summarized in Table 1. Many of the interleaver algorithms, for example, [4, 6–9] need some preprocessing before starting actual interleaving process. Therefore the whole data flow has been divided into two phases named as precomputation phase as shown in Figure 2(a) and the execution phase as shown in Figure 2(b). There are many

Table 1: List of algorithms and permutations in different interleaver implementations and the cost comparison.

Standard	Interleaver type	Algorithm/permutation methodology	HW cost	
			Addr. Gen. @65 nm (μm^2)	Data memory @6 soft bits (kbits)
HSPA+	BTC	Multistep computation including intra-row permutation computation $S(j) = (v \times S(j-1))\%p$; $r(i) = T(q(i))$; $U(i,j) = S((j \times r(i))\%(p-1))$; $q\text{mod}(i) = r(i)\%(p-1)$; $RA(i,j) = \{RA(i,j-1) + q\text{mod}(i)\}\%(p-1)$; $I_{i,j} = \{C \times r(i)\} + U(i,j)$	12816	59.92
	1st, 2nd, and HS-DSCH int.	Standard block interleaving with given column permutations. $\pi(k) = \left(P\left\lfloor \frac{k}{R} \right\rfloor + C \times (k\%R)\right)\%K_\pi$	2288	29.96
LTE	QPP for BTC	$I_{(x)} = (f_1 \,.\, x + f_2 \,.\, x^2)\%N$	3744	72.0
	Sub-Blk. int.	Standard block interleaving with given column permutations.	2080	36.0
WiMAX	Channel interleaver	Two step permutation $M_k = \left(\frac{N}{d}\right) \times (k\%d) + \left\lfloor \frac{k}{d} \right\rfloor$; $J_k = s \times \left\lfloor \frac{M_k}{s} \right\rfloor + \left(\left(M_k + N - \left\lfloor d \times \frac{M_k}{N} \right\rfloor\right)\%s\right)$	8944	9.0
	Blk. int. b/w RS & CC	Standard block interleaver without any permutations	2080	19.92
	CTC interleaver	$I_{(x\%4=0)} = (P_0 \cdot x + 1)\%N$; $I_{(x\%4=1)} = \left(P_0 \cdot x + 1 + \frac{N}{2} + P1\right)\%N$; $I_{(x\%4=2)} = (P_0 \cdot x + 1 + P1)\%N$; $I_{(x\%4=3)} = \left(P_0 \cdot x + 1 + \frac{N}{2} + P3\right)\%N$	7280	56.25
WLAN	Channel interleaver	Two step permutation $M_k = \left(\frac{N}{d}\right) \times (k\%d) + \left\lfloor \frac{k}{d} \right\rfloor$; $J_k = s \times \left\lfloor \frac{M_k}{s} \right\rfloor + \left(\left(M_k + N - \left\lfloor d \times \frac{M_k}{N} \right\rfloor\right)\%s\right)$	8944	1.68
802.11n	Ch. Interleaver with frequency rotation	Two step permutation as above, with extra frequency interleaving, that is, $R_k = \left[J_k - \left\{\left(((i_{ss}-1) \times 2)\%3 + 3\left\lfloor \frac{i_{ss}-1}{3} \right\rfloor\right) \times N_{ROT} \times N_{BPSC}\right\}\right]\%N$	11563	24.54
DVB-H	Outer conv. interleaver	Permutation defined by depth of first FIFO branch (M) and number of total braches.	12272	8.76
	Inner bit interleaver	Six parallel interleavers with different cyclic shift $H_e(w) = (w + \Delta)\%126$; where $\Delta = 0, 63, 105, 42, 21$ and 84	3120	0.738
	Inner symbol interleaver	$y_{H(q)} = x_q$ for even symbols; $y_q = x_{H(q)}$ for odd symbols; where $H(q) = (i\%2) \times 2^{N_r-1} + \sum_{j=0}^{N_r-2} R_i(j) \times 2^j$;	3536	35.4
General purpose use	Row or/and Col. Perm. Given	Standard block interleaver with or without row or/and column permutation.	3952	24.0
Total cost	$\sum$ (all)	Independent implementations	$\sim$ 82619	$\sim$ 378.0
This work	Reconfigurable Solution	HW Multiplexed Design	27757	72.0

minor differences in both the phases when we consider different types of interleavers; however, one of the main differences might be due to the type of interleaver, that is, block interleaver or convolutional interleaver. Other than the differences in address calculation for the two categories, a major difference is the memory access mechanism. In case of block interleaver the memory read and write is explicit but a convolutional interleaver needs to write and

Table 2: Architecture exploration for different standards.

Standard	Interleaver type	Block size	Adders/ comparator	Multiplier	HW LUT	Configurable LUT/registers	Memory size (SB: soft bits)
HSPA+	Prime interleaver for BTC	5114	7	1	20×5b 440×7b 52×14b	20×8b 256×8b	$2 \times 5114 \times$ SB
	1st, 2nd, and HS-DSCH interleaving	5114	2	1	15×3b 32×5b	—	$5114 \times$ SB
3GPP-LTE	QPP interleaver for BTC	6144	5	—	188×19b	2×13b	$2 \times 6144 \times$ SB
	Sub-Block interleaver	6144	2	1	32×5b	—	$6144 \times$ SB
WiMAX (802.16e)	Channel interleaver	1536	5	1	15×4b	2×2b 1×11b	$1536 \times$ SB
	Block interleaver b/w RS and CC	2550	2	1	—	—	2550×8b
	CTC interleaver	2400	4	—	32×27b	1×12b	$4 \times 2400 \times$ SB
WLAN (802.11 a/b/g)	Channel interleaver	288	5	1	15×4b	2×2b 1×9b	$288 \times$ SB
802.11n Enhanced WLAN	Channel interleaver with frequency rotation	2592	9	1	30×4b 24×9b	2×2b 2×10b	$4 \times 648 \times$ SB
DVB ETSI EN 300-744	Outer convolutional interleaver	1122	4	1	—	11×11b	357×8b 765×8b
	Inner bit interleaver	126	8	—	—	21×1b 126×1b	$2 \times 126 \times 1$b $2 \times 126 \times 2$b
	Inner symbol interleaver	6048	1	—	30×1b	—	6048×6b
General purpose use	Row or/and Column permutation given as a table	4096	2	1	—	256×8b 64×6b	$4096 \times$ SB

read at the same time. This demands a dual port memory; however, it has been dealt by dividing the memories and introducing a delay in the read path. To get the general idea of cost saving by using hardware multiplexed architecture with shared data flow, each of the algorithms is implemented separately after applying appropriate algorithmic transformations. Comparing the hardware cost for different implementations as given in Table 1, the proposed hardware multiplexed architecture based on shared data flow provides 3 times lower silicon cost for address generation and about 5 times lower silicon cost for data memory in shared mode. Going through all the interleaver implementations given in Table 1, different hardware requirements for computing elements and memory are summarized in Table 2. Looking at the modulo computation requirements, the use of adder appears to be the common computing element for all kinds of implementations. Further observation reveals that adder is mostly followed by a selection logic. Therefore, a common computing cell named *acc_sel* as shown in Figure 3 is used to cover all the cases. Table 2 shows that the computational part of the reconfigurable implementation can be restricted to have 8 additions, 1 multiplication, and a comparator.

The memory requirements for different implementations are also very wide, due to different sizes, width, memory banks and ports. The memory organization and address computation is explained in detail in the next section.

4. Multimode Interleaver Architecture

The study from algorithm analysis provides the basis to multiplex the hardware intensive components and combine the functionality of multiple types of interleavers. The architecture for the multimode interleaver is given in Figure 4. The hardware partitioning is done in such a way that all computation intensive components are included in the address generation block. The other partitioned blocks are register file, control-FSM, and memory organization block. These blocks are briefly described in the following subsections.

4.1. Address Generation (ADG) Block. Address generation is the main concern for any kind of interleaving. Unified address generation is achieved by multiplexing the computation intensive blocks mentioned in Table 2. The address generation hardware is shown in detail in Figure 4. It is surrounded by other blocks like control FSM, register file, and some lookup tables. It utilizes 8 *acc_sel* units with a multiplier and a comparator. The reconfigurability is mainly achieved through changing the behavior of *acc_sel*

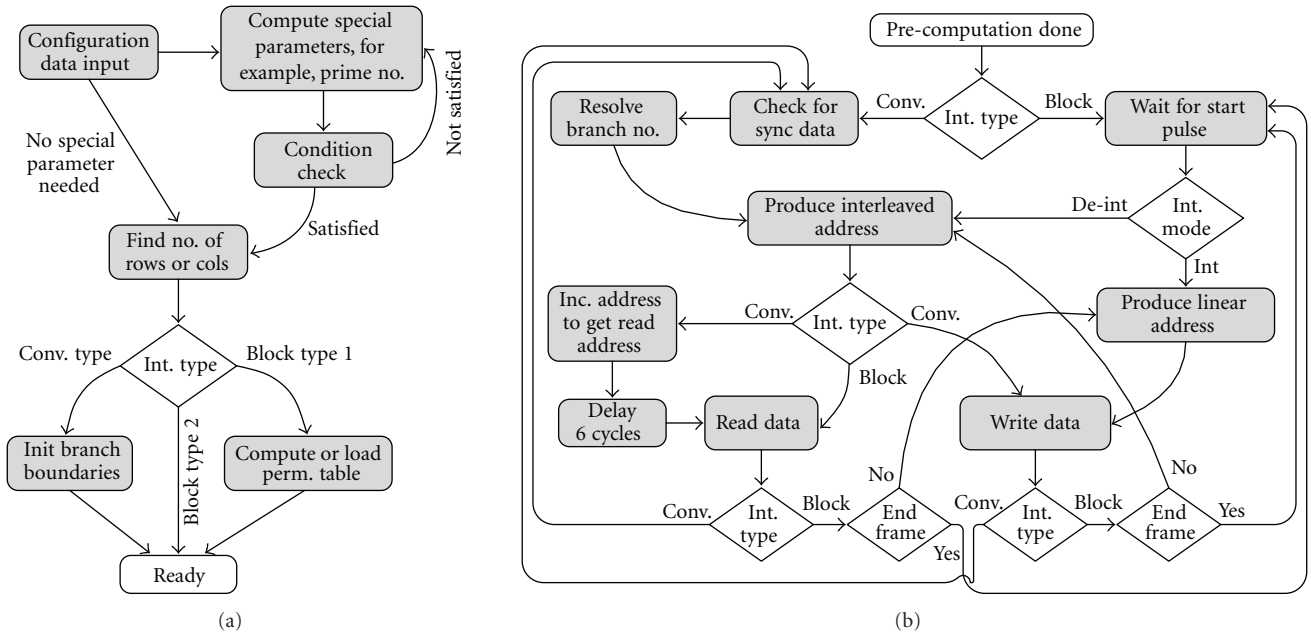


Figure 2: Data flow graph for (a) precomputation phase (b) execution phase.

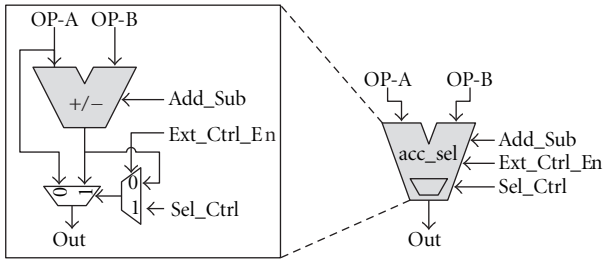


Figure 3: An accumulation and selection cell (acc_sel).

and appropriate multiplexer selection. The control signals *Add_Sub*, *Ext_Ctrl_En* and, *Sel_Ctrl* are used to define the behavior of *acc_sel* block. Using these signals in an appropriate way this block can be configured as an adder, a subtractor, a modulo operation with MSB of output as select line, or just a bypass. All the combinations are fully utilized and make it a very useful common computing element. The address generation block takes the configuration vector and configures itself with the help of a decoder block and part of the LUT. The configuration vector is 32 bit wide, which defines block size, interleaver depth, interleaving modes, and modulation schemes.

The ADG block generates the interleaved address based on all the permutations involved for implementing a block interleaver, whereas it generates memory read and write addresses concurrently while implementing a convolutional interleaver. The role of ADG block to be used as an interleaver or deinterleaver is mainly controlled by the controller after employing an addressing combination (permuted or sequential addressing) for writes and reads from the memory.

4.2. Control FSM. Two modes of operation for the hardware are defined as precomputation mode and execution mode. In order to handle the sequence of operations in the two modes a multistate control-FSM is used. The flow graph of the control-FSM is shown in Figure 5. During precomputation phase, the FSM may perform two main functions: (1) computation of necessary parameters required for interleaver address computation and (2) initialization of registers to become ready for execution phase. Other than IDLE state, 5 states (S1∼S4, S8) are assigned for precomputation. The common parameter to be computed in the precomputation phase is number of rows or columns; however, some specific parameters like prime number p; and intra-row permutation sequence $S(j)$ in WCDMA turbo code interleaver are also computed during this phase. For the interleaver functions which do not require precomputation, the initialization steps for precomputation are bypassed, and the control FSM directly jumps to the execution phase. The extra cycle cost associated with the precomputation has been investigated for the current implementation and the results are presented in a later section. In the execution phase, the control-FSM helps in sequencing the loading of data frames into memory or reading data frames from memory. In total 4 states (S5∼S7, S9) are assigned for execution phase. S9 is used for convolutional interleaver case only, whereas states S5∼S7 are reused for all types of interleavers. During the execution phase the control-FSM keeps track of block size also by employing row and column counter, thus providing the block synchronization required for each type of interleaver implementation.

4.3. Register File. The requirement of temporary storage of parameters arises with many types of interleaver implementations. Register requirements from different implementations are listed in Table 2. Some special usage configuration is also required for different cases; for example, WCDMA turbo

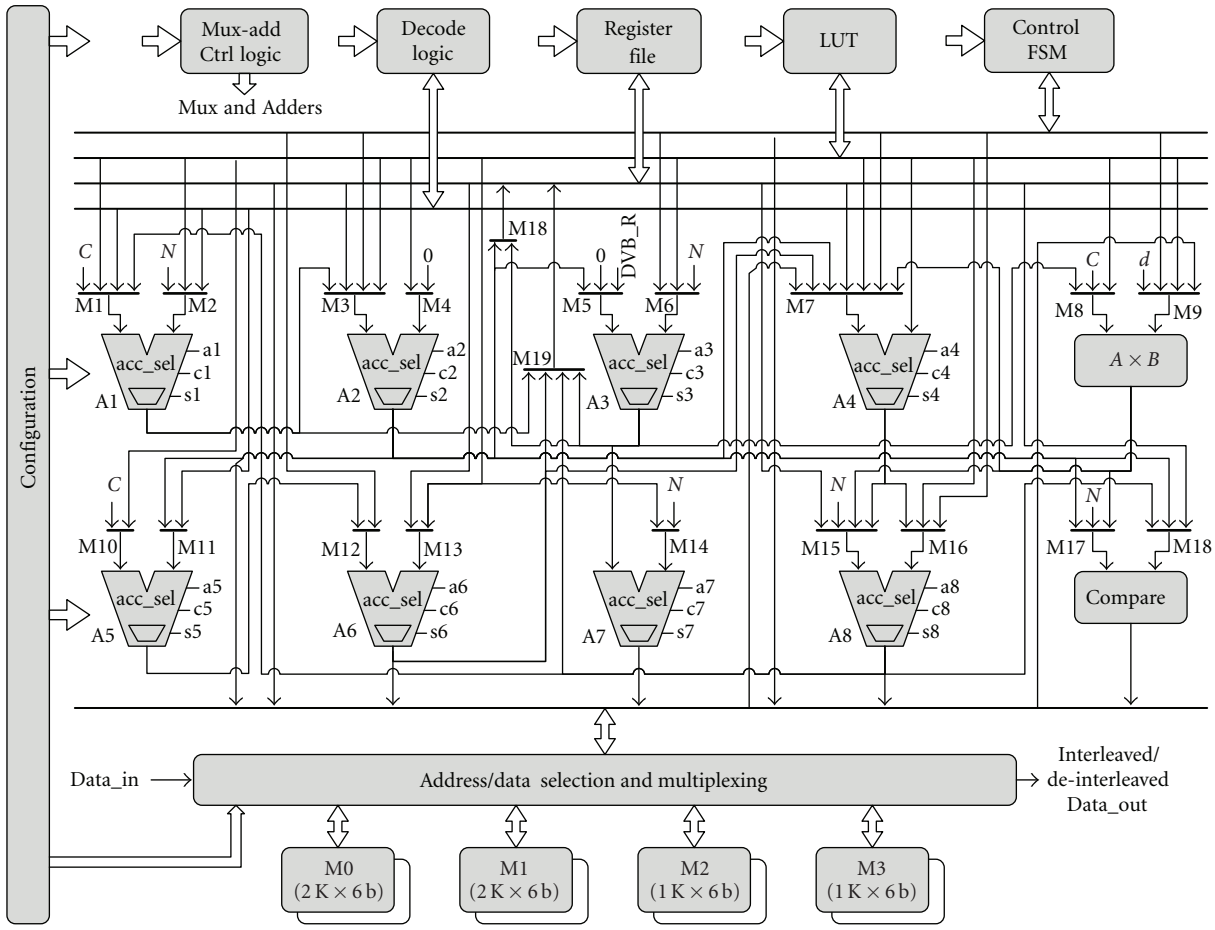


Figure 4: Address generation schematic in detail.

code interleaver needs 20 registers to form a circular buffer, convolutional interleaver in DVB requires 11 registers to be used as a general purpose register file, and the bit interleaver in DVB requires a long chain of single bit registers. Due to small size and special configuration requirements, a general purpose register file is not feasible here, and a fully customized register file is used. The width of registers is not the same and it is optimized as per requirement from different implementations. The registers can also be connected to form a chain, thus the single bit buffer for a bit interleaver is managed by circulating the shifted output inside register file. The two data input ports of the register file are fed through multiplexers M18 and M19 as shown in Figure 4.

4.4. Memory Organization. Memory requirements for different types of interleaver implementations are very much different as listed in Table 2. Also, soft bit processing in the decoder implies different requirements of bit width for different conditions and decoding architectures. The maximum width requirement is 6 bits for symbol interleaving and 8 bits for part of the memory in WCDMA. Multistream transmission requires multiple banks of memories in parallel. The size of the memory is taken as $2 \times 6144 \times SB$, which is due to large block size requirements for 3GPP-LTE, 3GPP-WCDMA, and DVB.

Memory partitioning is mainly motivated by the high-throughput requirements from the multistream system, for example, 802.11n. It requires four memory banks in parallel which appears to be a good choice to meet other requirements as well. Parallel memory banks can also be used in series to form a big memory. Partial parallelism can also be used where larger memory width is needed. Another worth full benefit of using multiple memory banks is avoiding the use of dual port RAM, which is not silicon efficient. Thus all the memories in the design are single port memories. The interleaved addresses for block and convolution interleavers computed by address generation block are combined according to the configuration requirement to make the final memory address. Figure 6 shows the memory organization with address selection logic. Particularly for convolutional interleaving, a small delay line with depth of 6 in the path of read addresses and control signals is used to avoid the data write and read for the same memory in a single clock cycle.

5. Algorithm Transformation for Efficient Mapping

The main objective is to use single architecture for interleaver implementation with maximum hardware sharing among

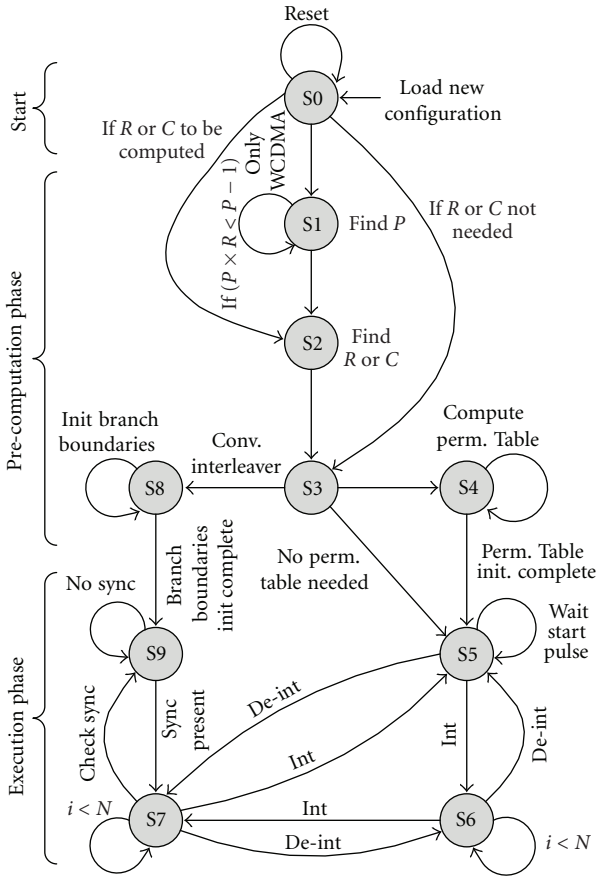


Figure 5: FSM state graph.

different algorithms. The versatility of interleaving algorithms makes it an in-efficient implementation when original algorithms are directly mapped to same architecture. On the other hand some transformations based on modular algebra can be applied on the original algorithms to make them hardware efficient. Same algorithmic transformations can be used to reach to an efficient hardware multiplexing among different standards. The following subsections present some transformation examples for selected algorithms which are very much versatile in the implementation point of view. These subsections cover channel interleaving for WiMAX and WLAN including 802.11n with frequency rotation, turbo code block interleaving for LTE, WiMAX, and HSPA Evolution, and convolutional interleaving used in DVB.

5.1. Channel Interleaving in WiMAX and WLAN. The channel interleaving in 802.11a/b/g (WLAN) and 802.16e (WiMAX) is of the same type. The interleaver function defined by a set of two equations for two steps of permutations, provides spatial interleaving, whereas the newly evolved standard 802.11n [8] based on MIMO-OFDM employs frequency interleaving in addition to spatial interleaving. Most of literature available [31–36] covers the performance and evaluation of WLAN interleaver design for a high-speed communication system; however, some recent work [23–27] focuses on interleaver architecture design including some complexity reduction techniques along with feasibility to gain higher throughput. The 2D realization of interleaver functions is exploited to enable efficient hardware implementation. The two steps of permutations for index k for interleaver data are expressed by the following equations:

$$M_k = \left(\frac{N}{d}\right) \times (k\%d) + \left\lfloor \frac{k}{d} \right\rfloor, \tag{1}$$

$$J_k = s \times \left\lfloor \frac{M_k}{s} \right\rfloor + \left(\left(M_k + N - \left\lfloor d \times \frac{M_k}{N} \right\rfloor \right) \% s \right). \tag{2}$$

Here N is the block size corresponding to number of coded bits per allocated subchannels and the parameter s is defined as $s = \max\{1, N_{\mathrm{BPSC}}/2\}$ where N_{BPSC} is the number of coded bits per subcarrier, (i.e., 1, 2, 4 or 6 for BPSK, QPSK, 16-QAM, or 64-QAM, resp.). The operator % is the modulo function computing the remainder and the operator $\lfloor x \rfloor$ is the floor function, that is, rounding x towards zero. The range of n and k is defined as $0, 1, 2, \ldots (N-1)$. The direct implementation of the above mentioned equations is very much hardware in-efficient and also the mapping onto the proposed unified interleaver architecture is not possible. Therefore, realization of two 1D equations into 2D space and computation of interleaved address in recursive way is adopted to reduce the hardware complexity as explained in the following subsections.

5.1.1. BPSK-QPSK. As N_{BPSC} is 1 and 2 for BPSK and QPSK, respectively; thus $s = 1$ for both cases and (2) simplifies to the following form:

$$J_k = \left(\frac{N}{d}\right) \times (k\%d) + \left\lfloor \frac{k}{d} \right\rfloor. \tag{3}$$

Considering the interleaver as a block interleaver, the parameter d is usually considered as total number of columns N_{COL}, and parameter N/d is taken as total number of rows N_{ROW}, but the column and row definition are swapped hereafter. The parameter d is taken as total number of rows and parameter N/d is taken as total number of columns. The functionality still remains the same, with the benefit that it ends up with the recursive expression for all the modulation schemes. According to new definitions, the term $(k\%d)$ provides the behavior of row counter and the term $\lfloor k/d \rfloor$ provides the behavior of column counter. Thus introducing two new variables i and j as two dimensions, such that j increments when i expires, the ranges for i and j are mentioned as follows:

$$i = 0, 1, \ldots (d-1), \qquad j = 0, 1, \ldots \left(\frac{N}{d} - 1\right), \tag{4}$$

which satisfies against k when $i = (k\%d)$ and $j = \lfloor k/d \rfloor$. Defining total number of columns as $C = N/d$, (3) can be written as

$$J_{i,j} = C \times i + j. \tag{5}$$

The recursive form after handling the exception against $i = 0$ can be written as

$$J_{i,j} = \begin{cases} j, & \text{if } (i = 0), \\ J_{(i-1),j} + C, & \text{otherwise.} \end{cases} \tag{6}$$

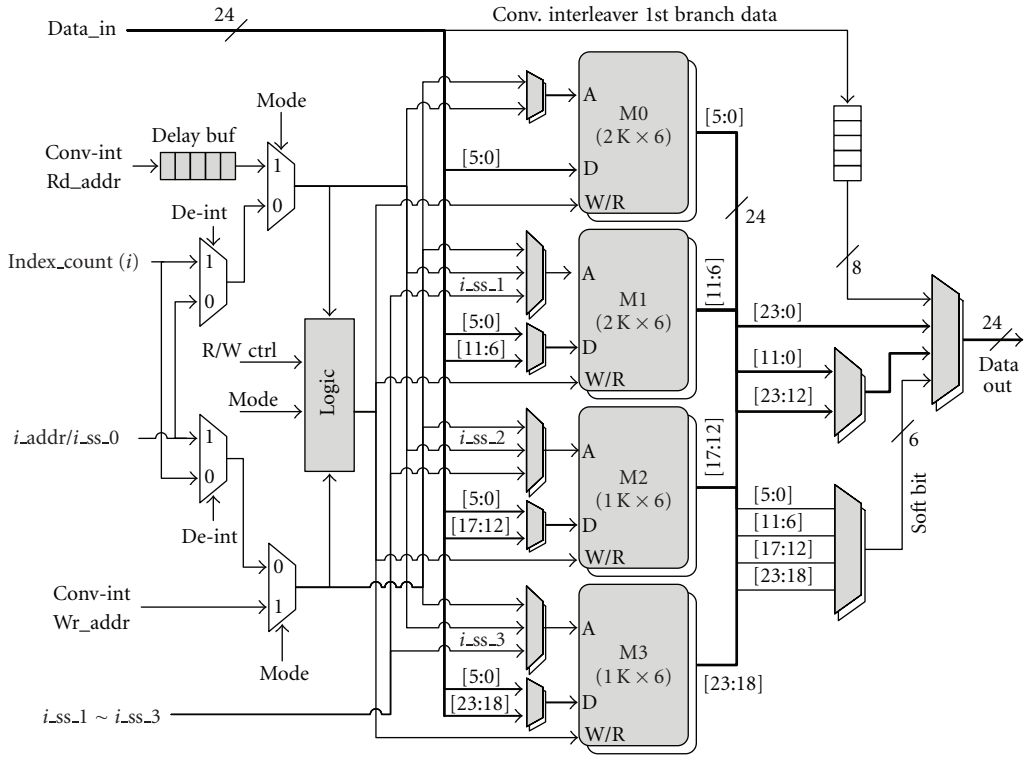


Figure 6: Memory address selection and data handle.

Defining row counter i as $i = R_c$ and column counter j as $j = C_c$, the hardware for (6) is shown in Figure 7(a). The case of BPSK and QPSK do not carry any specific inter-row or inter-column permutation pattern; thus it ends up with relatively simple hardware, but it provides the basis for analysis for 16-QAM and 64-QAM cases, which are more complicated.

5.1.2. 16-QAM. 16-QAM scheme has 4 code bits per subcarrier; thus parameter s is 2 and (2) becomes

$$J_k = 2 \times \left\lfloor \frac{M_k}{2} \right\rfloor + \left(\left(M_k + N + \left\lfloor \frac{d \times M_k}{N} \right\rfloor \right) \%2 \right). \tag{7}$$

Like BPSK/QPSK case, algebraic only steps cannot be used here to proceed due to the presence of floor and modulo functions. Instead, all the possible block sizes for 16-QAM are analyzed to restructure the above equation. The following structure appears to be equivalent to (7) and at the same time resembles the structure of (3); thus it suits well for hardware multiplexing:

$$J_k = \left(\frac{N}{d} \right) \times (k\%d) + \left\lfloor \frac{k}{d} \right\rfloor + r_k^2. \tag{8}$$

The extra term r_k^2 is defined by the following expression:

$$\begin{aligned} r_k^2 = {} & [(1 - (k\%2)) - (k\%2)] \left\{ 1 - \left(\left\lfloor \frac{k}{d} \right\rfloor \%2 \right) \right\} \\ & + [((k\%2) - 1) + (k\%2)] \left\{ \left\lfloor \frac{k}{d} \right\rfloor \%2 \right\}. \end{aligned} \tag{9}$$

This term appears due to the reason that the interleaver for 16-QAM carries specific permutation patterns, making the structure more complicated. Considering the 2-dimensions i and j having range as mentioned in (4), the behavior of the term $k\%2$ is the same as that of $i\%2$, when i is the row counter. Thus (8) can be written in 2D representation as follows:

$$J_{i,j} = \begin{cases} j, & \text{if } (i = 0), \\ J_{(i-1),j} + C + r_{i,j}^2, & \text{otherwise}, \end{cases} \tag{10}$$

where

$$\begin{aligned} r_{i,j}^2 = {} & [(1 - (i\%2)) - (i\%2)] \{1 - (j\%2)\} \\ & + [(i\%2) + (1 - (i\%2))] \{j\%2\}. \end{aligned} \tag{11}$$

The term can further be simplified to a smaller expression but it is easy to realize the hardware from its current form. The modulo terms can be implemented by using the LSB of row counter R_c and column counter C_c, and the required sequence can be generated with the help of an XOR gate and an adder as shown in Figure 7(b).

5.1.3. 64-QAM. The parameter s is 3 for 64-QAM; thus (2) becomes

$$J_k = 3 \times \left\lfloor \frac{M_k}{3} \right\rfloor + \left(\left(M_k + N + \left\lfloor \frac{d \times M_k}{N} \right\rfloor \right) \%3 \right). \tag{12}$$

The presence of modulo function $x\%3$ makes it much harder to reach some valid mathematical expression algebraically. Different structures for all possible block sizes for

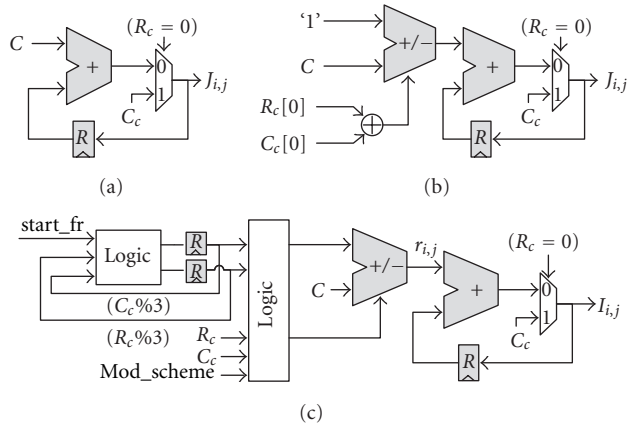


Figure 7: Interleaver address generation for (a) BPSK-QPSK, (b) 16-QAM, and (c) combined for all modulation schemes.

64-QAM are analyzed and the structure similar to (6) and (10) and equivalent to (12) is given as follows:

$$J_{i,j} = \begin{cases} j, & \text{if } (i = 0), \\ J_{(i-1),j} + C + r^3_{i,j}, & \text{otherwise,} \end{cases} \tag{13}$$

where i and j represent two dimensions and their range is given by (4). Defining $i' = (i\%3)$ and $j' = (j\%3)$, $r^3_{i,j}$ is given as

$$\begin{aligned} r^3_{i,j} = {} & \left((1 - j') + \frac{j'(j'-1)}{2}\right)\left\{2\left((1 - i') + \frac{i'(i'-1)}{2}\right) - \left(i' - \frac{i'(i'-1)}{2}\right)\right\} \\ & + (j' - j'(j'-1))\{2(i' - i'(i'-1)) - ((1 - i') + i'(i'-1))\} \\ & + \left(\frac{j'(j'-1)}{2}\right)\left\{2\left(\frac{i'(i'-1)}{2}\right) - \left(1 - \frac{i'(i'-1)}{2}\right)\right\}. \end{aligned} \tag{14}$$

The term $r^3_{i,j}$ provides the inter-row and inter-column permutation for $s = 3$ against row counter i and column counter j. The expression for $r^3_{i,j}$ looks very long and complicated, but eventually, it gives a hardware efficient solution as the terms inside braces are easier to generate through a very small lookup table. The generic form for (6), (10), and (13) to compute the interleaved address $I_{i,j}$ can be written as

$$I_{i,j} = \begin{cases} j, & \text{if } (i = 0), \\ I_{(i-1),j} + C + r^s_{i,j}, & \text{otherwise.} \end{cases} \tag{15}$$

Here parameter s distinguishes different modulation schemes. For BPSK/QPSK $r^1_{i,j} = 0$, and for 16-QAM and 64-QAM, $r^2_{i,j}$ and $r^3_{i,j}$ are given by (9) and (14), respectively. The hardware realization supporting all modulation schemes is shown in Figure 7(c). It appears to be a much optimized implementation as it involves only two additions, some registers, and a very small lookup table.

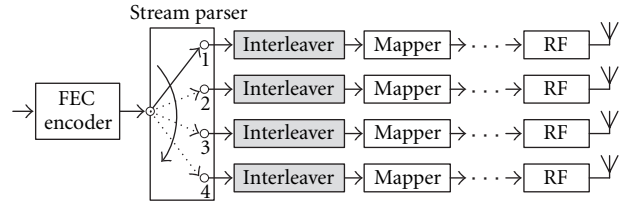


Figure 8: Use of interleaver in multiple spatial streams (802.11n).

5.2. Frequency Interleaving in 802.11n. The transmission in 802.11n can be distributed among four spatial streams as shown in Figure 8. The interleaving requires frequency rotation in case more than one spatial streams are being transmitted. The frequency rotation is applied to the output of the second permutation J_k. The expression for frequency rotation for spatial stream i_{ss} is given as follows:

$$R_k = \left[J_k - \left\{\left(((i_{\text{ss}} - 1) \times 2)\%3 + 3\left\lfloor \frac{i_{\text{ss}} - 1}{3} \right\rfloor\right) \times N_{\text{ROT}} \times N_{\text{BPSC}}\right\}\right]\%N. \tag{16}$$

Here N_{ROT} is the parameter which defines different frequency rotations for 20 MHz and 40 MHz case in 802.11n. The frequency rotation also depends on the index of the spatial stream i_{ss}, thus each spatial stream faces different frequency rotations. Defining the rotation term as J_{ROT}, that is,

$$J_{\text{ROT}} = \left\{\left(((i_{\text{ss}} - 1) \times 2)\%3 + 3\left\lfloor \frac{i_{\text{ss}} - 1}{3} \right\rfloor\right) \times N_{\text{ROT}} \times N_{\text{BPSC}}\right\}, \tag{17}$$

we have

$$R_k = (J_k - J_{\text{ROT}})\%N. \tag{18}$$

The range for the term $(J_k - J_{\text{ROT}})$ is not bounded and it can have value greater than $2N$; thus direct implementation cannot be low cost. Analyzing the two terms $[J_k\%N]$ and $(-J_{\text{ROT}})\%N$ separately, it is observed that the second term provides the starting point for computing the rotation R_k. As the rotation is fixed for a specific spatial stream, thus the starting value $r_{ks} = (-J_{\text{ROT}})\%N$ also holds for all run time computations. Equation (18) in combination with (10) can be written as

$$J^{i_{\text{ss}}}_{i,j} \equiv R_k = \left(J_{i,j} + r_{ks}\right)\%N. \tag{19}$$

Here $J^{i_{\text{ss}}}_{i,j}$ is the joint address after applying both, spatial interleaving and frequency interleaving against row index i, column index j and spatial stream index i_{ss}. A lookup can be used for the starting values for r_{ks} against different spatial

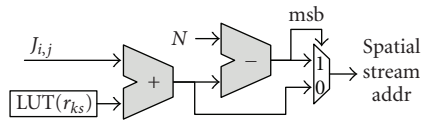


Figure 9: HW for frequency rotation in 802.11n.

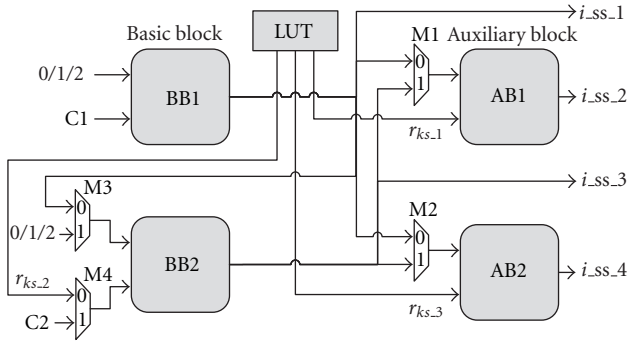


Figure 10: HW for quad stream interleaver.

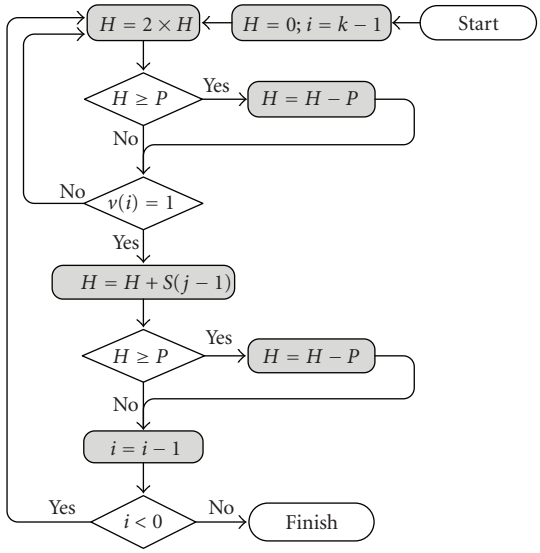


Figure 11: Flow graph for interleaved modulo multiplication algorithm.

streams. The r_{ks} values for all the cases follow the condition, that is, $(r_{ks} < N)$ which depicts that the term $(J_k + r_{ks})$ cannot be larger than $2N$. Therefore, the frequency rotation can be computed with a very small hardware as shown in Figure 9.

5.3. Multistream Interleaver Support in 802.11n. The spatial interleaver address generation block shown in Figure 7(c) is denoted as Basic Block (BB) and the frequency rotation block as shown in Figure 9 is denoted as Auxiliary Block (AB). Both these blocks combine to form a complete address generation circuit for one spatial stream. In order to provide support for four streams in parallel, one may consider replicating the two blocks four times. However, an optimized solution would be to use 2 basic blocks and 2 auxiliary blocks, still providing support for 4 spatial streams. The hardware block diagram to generate the interleaver addresses for multiple streams in parallel is shown in Figure 10. This hardware supports quick configuration changes thus providing full support to any multitasking environment. If some new combination of modulation schemes is needed to be implemented, which is not supported already, the interfacing processor can do task scheduling for different types of modulation schemes.

5.4. Turbo Code Interleaver for HSPA+. The channel coding block in HAPA+ including WCDMA uses turbo coding [37] for forward error correction. 3GPP standard [4] proposes the algorithm for block interleaving in turbo encoding/decoding as mentioned below. Here N is the block size, R is the row size, and C is the column size in bits.

(i) Find appropriate number of rows "R", prime number "p", and primitive root "v" for particular block size as given in the standard.

(ii) Col Size:

$$\begin{aligned} &C = p - 1, \quad \text{if } (N \leq R \times (p-1)), \\ &C = p, \quad \text{if } (R \times (p-1) < N \leq (R \times p)), \\ &C = p + 1, \quad \text{if } (R \times p < N). \end{aligned} \tag{20}$$

(iii) Construct intra-row permutation sequence S(j) by

$$S(j) = [v \times S(j-1)]\%p; \quad j = 1, 2, \ldots p-2. \tag{21}$$

(iv) Determine the least prime integer sequence $q(i)$ for $i = 1, 2, \ldots R-1$, by taking $q(0) = 1$, such that $g.c.d(q(i), p-1) = 1$, $q(i) > 6$ and $q(i) > q(i-1)$.

(v) Apply inter-row permutations to $q(i)$ to find $r(i)$:

$$r(i) = T(q(i)). \tag{22}$$

(vi) Perform the intra-row permutations Ui, j, for $i = 0, 1, \ldots R-1$ and $j = 0, 1, \ldots p-2$.

If $(C = p)$: $Ui, j = S[(j \times r(i)) \bmod (p-1)]$ and $Ui, (p-1) = 0$.

If $(C = p+1)$: $Ui, j = S[(j \times r(i)) \bmod (p-1)]$, and $Ui, (p-1) = 0$, $Ui, p = p$, and if $(N = R \times C)$ then exchange $U(R-1, 0)$ with $U(R-1, p)$.

If $(C = p-1)$: $Ui, j = S[(j \times r(i)) \bmod (p-1)] - 1$.

(vii) Perform the inter-row permutations.

(viii) Read the address columns wise.

The presence of complex functions like modulo computation, intra-row and inter-row permutations, multiplications, finding least prime integers, and computing greatest common divisor makes it in-efficient while implementing it in its original form. Further, to get one interleaving address in each cycle, some preprocessing is also required where parameters like total number of rows or columns, least prime number sequence $q(i)$, inter-row permutation patterns $T(i)$, intra-row permutations $S(j)$, prime number

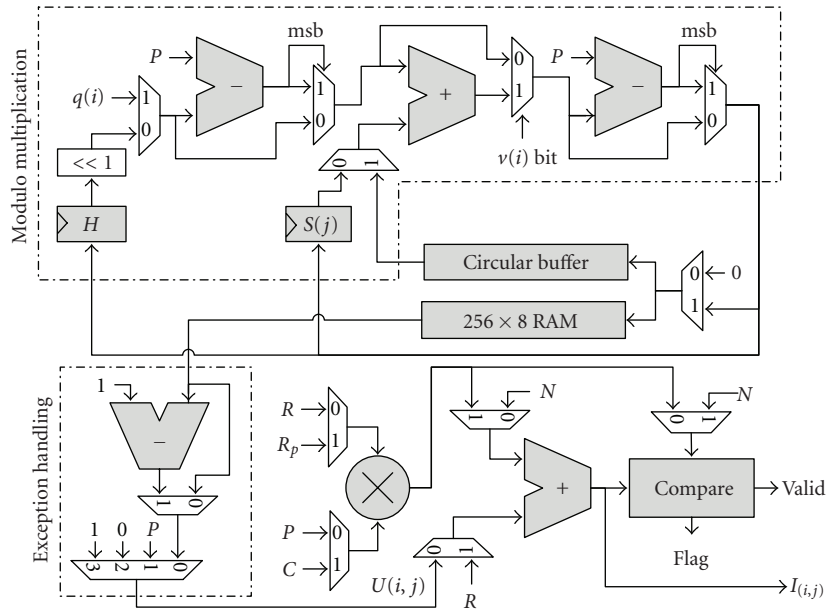


Figure 12: WCDMA turbo code interleaver hardware.

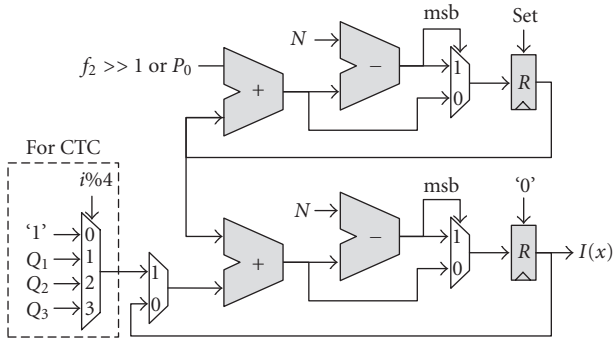


Figure 13: Simplified HW for 3GPP-LTE and CTC interleaver.

p, and associated integer v are computed. Some of these parameters can be computed using lookup tables while the others need some close loop or recursive computations. The simplifications considered in the implementation are discussed in the following paragraphs.

One of the main hurdles to generate on-the-fly interleaved address is the computation of intra-row permutation sequence $S(j)$. Before applying the intra-row permutations, the term $(j \times r(i)\%(p-1))$ is computed which produces random values due to $r(i)$ and modulo function. These random values appear as index to compute $S(j)$, due to which it may require many clock cycles to be computed on-the-fly. To resolve it, some precomputations are made and results are stored in a memory. These precomputations involve the computation of a modulo function which requires a divider for direct implementation. To avoid the use of divider, indirect computation of modulo function is done by using Interleaved Modulo Multiplication Algorithm [38]. It computes the modulo function in an iterative way requiring more than one clock cycles. Looking at maximum value of v, which is 5 bits, a maximum of 5 iterations are needed to compute one modulo multiplication. The algorithm to compute the Interleaved Modulo Multiplications is shown in Figure 11 and the hardware required is shown in Figure 12. This hardware produces the data for memory while in precomputation phase; however, same hardware is utilized to generate the address for the memory, while in execution phase. The usage of memory depends on the parameter p and it will be filled upto $(p-2)$ locations.

Finding $q\text{mod}(i) = q(i)\%(p-1)$ instead of direct computation of least prime number sequence $q(i)$ gives the benefit of computing the RAM address recursively and avoiding computation of the modulo function. This idea was introduced in [13] and later on it has been used in [14, 16, 17]. The computation of $q(i)\%(p-1)$ can be managed by a subtractor and a look up table, provided that all the values of $q(i)$ placed in the look up table satisfy the condition $q(i) < 2(p-1)$. The similarities between different sequences for $q(i)\%(p-1)$ for all possible p values are very helpful to improve the efficiency of the lookup table. The parameters p and v are stored in combined fashion in a lookup table of size 52×14b. The lookup table is addressed via a counter. Against each value of p, the condition $(p \times R \geq N - R)$ is checked using a comparator to find the appropriate value for p and v. Once p is found, the total number of columns C can have only three values, that is, $p-1$, p, or $p+1$. Hence C is found in at most three clock cycles by checking the condition $(R \times C \geq N)$. The recursive function used to compute the RAM address with the help of parameter $q\text{mod}(i)$ is given by

$$RA(i,j) = \{RA(i,j-1) + q\text{mod}(i)\}\%(p-1). \quad (23)$$

The data from RAM are denoted as $U(i,j)$ after passing through some exception handling logic. Parameter $U(i,j)$ provides the intra-row permutation pattern for a particular row. The final interleaved address $I_{i,j}$ can be found

by combining the inter-row permutation with intra-row permutation as follows:

$$I_{i,j} = \{C \times r(i)\} + U(i,j). \tag{24}$$

The complete hardware for interleaver address generation for Turbo Code interleaver is shown in Figure 12. It can be mapped to the proposed unified interleaver architecture quite efficiently.

5.5. Turbo Code Interleaving in 3GPP-LTE and WiMAX. The newly evolved standard, 3GPP LTE [5], involves interleaving in the channel coding and rate matching section. The interleaving in rate matching is called subblock interleaving and is based on simple block interleaving scheme. The channel coding in LTE involves Turbo Code with an internal interleaver. The type of interleaver here is different and it is based on quadratic permutation polynomial (QPP), which provides very compact representation. The turbo interleaver in LTE is specified by the following quadratic permutation polynomial:

$$I_{(x)} = (f_1 \cdot x + f_2 \cdot x^2)\%N. \tag{25}$$

Here $x = 0, 1, 2, \ldots (N-1)$, with N as block size. This polynomial provides deterministic interleaver behavior for different block sizes and appropriate values of f_1 and f_2. Direct implementation of the permutation polynomial given in (25) is hardware in-efficient due to multiplications, modulo function, and bit growth problem. To simplify the hardware, (25) can be rewritten for recursive computation as

$$I_{(x+1)} = (I_{(x)} + g_{(x)})\%N, \tag{26}$$

where $g_{(x)} = (f_1 + f_2 + 2 \cdot f_2 \cdot x)\%N$. This can also be computed recursively as

$$g_{(x+1)} = (g_{(x)} + 2 \cdot f_2)\%N. \tag{27}$$

The two recursive terms mentioned in (26) and (27) are easy to implement in hardware (Figure 13) with the help of a LUT to provide the starting values for $g_{(x)}$ and f_2.

WiMAX standard [6] uses convolutional turbo coding (CTC) also termed duo-binary turbo coding. They can offer many advantages like performance, over classical single-binary turbo codes [39]. Parameters to define the interleaver function as described in [6] are designated as P_0, P_1, P_2, and P_3. Two steps of interleaving are described as follows.

Step 1. Let the incoming sequence be

$$u_0 = [(A_0, B_0), (A_1, B_1), (A_2, B_2), \ldots (A_{N-1}, B_{N-1})]; \tag{28}$$

for $x = 0 \cdots N-1$, *if* $(i\%2) = 1$, *then* $(A_i, B_i) = (B_i, A_i)$.

The new sequence is

$$u_1 = [(A_0, B_0), (B_1, A_1), (A_3, B_3), \ldots (B_{N-1}, A_{N-1})]. \tag{29}$$

Step 2. The function $I_{(x)}$ provides the address of the couple from the sequence u_1 that will be mapped onto address x of the interleaved sequence. $I_{(x)}$ is defined by the set of four expressions with a switch selection as follows:

for $x = 0 \cdots N-1$
switch $(x\%4)$.
case 0: $I_{(x\%4=0)} = (P_0 \cdot x + 1)\%N$.
case 1: $I_{(x\%4=1)} = (P_0 \cdot x + 1 + N/2 + P_1)\%N$.
case 2: $I_{(x\%4=2)} = (P_0 \cdot x + 1 + P_2)\%N$.
case 3: $I_{(x\%4=3)} = (P_0 \cdot x + 1 + N/2 + P_3)\%N$.

Combining the four equations provided in step-2, the interleaver function $I_{(x)}$ becomes

$$I_{(x)} = (\beta_x + Q_x)\%N, \tag{30}$$

where β_x can be computed using recursion, that is, $\beta_{(x+1)} = (\beta_x + P_0)\%N$ by taking $\beta_0 = 0 \cdot Q_x$ is given by

$$Q_x = \begin{cases} 1, & \text{if } (j\%4 = 0), \\ 1 + \dfrac{N}{2} + P_1, & \text{if } (j\%4 = 1), \\ 1 + P_2, & \text{if } (j\%4 = 2), \\ 1 + \dfrac{N}{2} + P_3, & \text{if } (j\%4 = 3). \end{cases} \tag{31}$$

As range of β_x and Q_x is less than N, thus I_x can be computed by using addition and subtraction with compare and select logic as shown in Figure 13.

5.6. Convolutional Interleaving in DVB. The convolutional interleaver used in DVB is based on the Forney [40] and Ramsey type III approach [41]. The convolutional interleaver being part of outer coding resides in between RS encoding and convolutional encoding. The convolutional interleaver for DVB consists of 12 branches as shown in Figure 14. Each branch j is composed of first-in-first-out (FIFO) shift registers with depth $j \times M$, where $M = 17$ for DVB. The

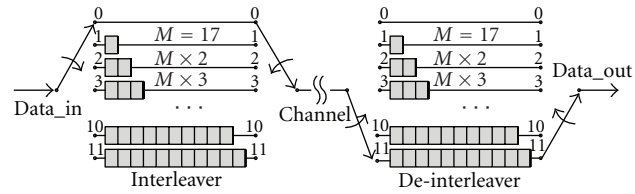


Figure 14: Convolutional interleaver and deinterleaver in DVB.

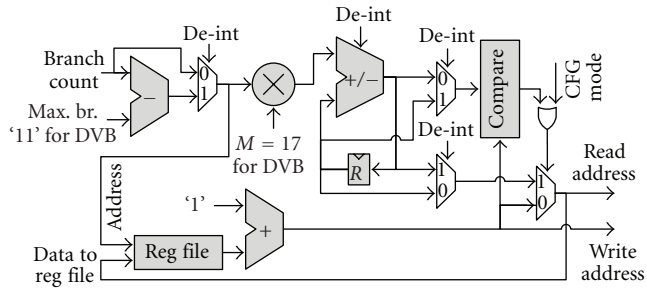


Figure 15: HW for RAM read/write address generation for convolutional interleaver.

Table 3: Precomputation cycle cost for different standards.

Standard	Worst case precomputation cycle cost
802.11 a/b/g—WLAN Channel interleaver	20
802.16e—WiMAX Channel interleaver	98
3GPP—WCDMA Block turbo code (Depends on Block size "N")	15 for ($N = 40$) 23 for ($N = 41$) 802 for ($N = 5040$) 563 for ($N = 5114$)
ETSI EN 300-744—DVB Inner symbol interleaver	15
802.11n—Extended WLAN	38
General purpose use	Depends on external HW, that is, loading the permutations
All others	Less than 3

Table 4: Summary of implementation results.

Parameter	Value
Target technology	65 nm
Memory configuration	$2048 \times 6b \times 4$; $1024 \times 6b \times 4$
Total memory	72 Kbit
Memory area	97972 μm^2
Memory power consumption	10.5 mW
Logic area	28436 μm^2
Total area	0.126 mm^2
Clock rate	166 MHz
Throughput (Max)	664 Mbps
Total power consumption	11.7 mW

packet of 204 bytes consisting of one sync byte (0×47 or $0\times B8$) is entered into the interleaver in a periodic way. For synchronization purpose the sync bytes are always routed to *branch*-0 of interleaver.

Convolutional interleaving is best suited for real time applications with some added benefits of half the latency and less memory utilization as compared to block interleaving. Recently, convolutional interleavers have been analyzed to work with Turbo codes [42–44], with improved performance, which make them more versatile; thus general and reconfigurable convolutional interleaver architecture integrated with block interleaver functionality can be of significance.

Implementation of convolutional interleavers using first-in-first-out (FIFO) register cells is silicon inefficient. To achieve a silicon efficient solution, RAM-based implementation is adopted. The memory partitioning is made in such a way that by applying appropriate read/write addresses in a cyclic way, it exhibits the branch behavior as required by a convolutional interleaver. RAM write and read addresses are generated by the hardware shown in Figure 15. The hardware components used here are almost the same as used by interleaver implementation for other standards, thus providing the basis for multiplexing the hardware blocks for reuse. To keep track of next write address for each branch, 11 registers are needed, which provides the idea of using cyclic pointers instead of using FIFO shift registers. For each branch the corresponding write address is provided by the concerned pointer register and next write address (which is also called current read address) is computed by using an addition and a comparison with the branch boundaries. Other reference implementations have used branch boundary tables directly, but to keep the design general, the branch boundaries are computed on-the-fly using an adder and a multiplier in connection with a branch counter.

For implementing a convolutional deinterleaver, the same hardware is used by implementing the branch counter in reverse order (decrementing by 1). In this way, same branch boundaries are used, and the only difference is that the sync byte in the data is now synchronized with the largest branch size as shown in Figure 14. Keeping the same branch boundaries for the deinterleaver, the width of the pointer register becomes fixed. This gives an additional benefit that the width of pointer register may be optimized efficiently.

6. Integration into Baseband System

The multimode interleaver architecture can perform interleaving or deinterleaving for various communication systems. It is targeted to be used as an accelerator core with a programmable baseband processor. The usage of the multimode interleaver core completely depends on the capability of the baseband processor. For lower throughput requirements only a single core can be utilized with baseband processor and the operations are performed sequentially. However, as a matter of fact, usual system level implementations require interleaver at multiple stages. Number of stages can be up to three, for example, WCDMA (turbo code interleaving, 1st interleaving, and 2nd interleaving). A fully parallel implementation can be realized by using three instances of the proposed multimode interleaver core, but in order to optimize the hardware cost a wise usage would be to use two instances hooked up with the main bus of the processor as shown in Figure 16. In this way the interleaving stages can be categorized as channel interleaving and coding/decoding interleaving. Further optimizations can be made in the two cores to fit in the particular requirements, for example, one interleaver core dedicated for coding/decoding and the second core dedicated for channel interleaving. By doing so the reduction of silicon cost associated with address generation is not significant, however, memory sizes can be optimized as per the targeted implementations, which can reduce the silicon cost significantly. For current implementation of multimode interleaver, the input memory used for any kind of decoding is considered to be the part of baseband processor data memory. In this way the extra memory inside interleaver core can be avoided which might be redundant in many cases. However, the integration of input memory in the main decoding operation is facilitated by the interleaver core by providing the address for input memory. In this way the interleaved/deinterleaved data can be fed to the decoder block in synchronized manner.

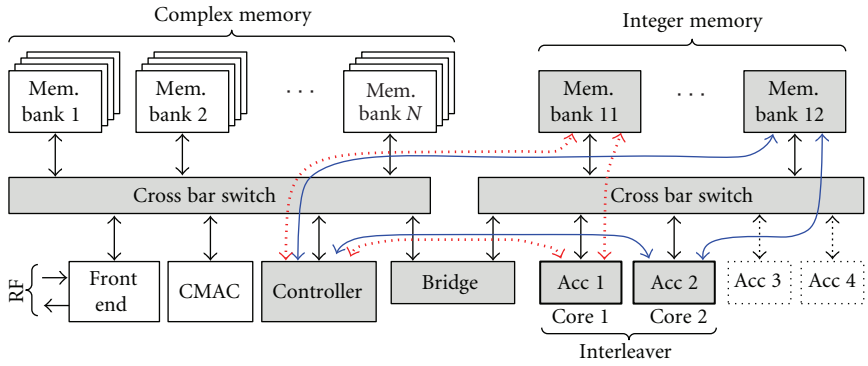


Figure 16: Integration of interleaver core with baseband processor.

Table 5: HW comparison with other implementations.

Implementation	Standard coverage	Technology	Operating frequency	Power	Memory size	Total core size
Xilinx [28] Virtex-5	General purpose (commercial use)	FPGA	262/360 MHz Speed Grade -1/-3	—	18 Kbits	210 LUTs + Memory
Altera [29] FLEX-10KE	General purpose (commercial use)	FPGA	120 MHz	—	16 Kbits	392 LEs + Memory
Lattice [30] ispXPGA	General purpose (commercial use)	FPGA	132 MHz	—	36 Kbits	284 LUTs + Memory
Shin and Park [13]	WCDMA turbo code; cdma2000	0.25 μm	—	—	35 Kbits	2.678 mm^2
Asghar et al. [18]	WCDMA, LTE, WiMAX and DVB-SH Turbo Code Interleaver Only	65 nm	200 MHz	10.04 mW	30 Kbits	0.084 mm^2
Chang and Ding [23]	WiMAX, WLAN, DVB	0.18 μm	100 MHz	—	12 Kbits	0.60 mm^2
Chang [24]	WiMAX, WLAN, DVB	0.18 μm	150 MHz	—	12 Kbits	0.484 mm^2
Wu et al. [25]	WiMAX, WLAN, 802.11n	0.18 μm	200 MHz	—	32 Kbits	0.72 mm^2
Asghar and Liu [26]	WiMAX, WLAN, DVB	0.12 μm	140 MHz	3.5 mW	12 Kbits	0.18 mm^2
Asghar and Liu [27]	WiMAX, WLAN, 802.11n	65 nm	225 MHz	4 mW	15.6 Kbits	0.035 mm^2
Horvath et al. [20]	DVB bit and symbol interleaver	0.6 μm	36.57 MHz	300 mW	48 Kbits	69 mm^2
Chang [21]	DVB bit and symbol interleaver	0.35 μm	—	—	52.2 Kbits	2.9 mm^2
This work	All range including WLAN, WiMAX, DVB, HSPA+, LTE, 802.11n and General purpose implementation	65 nm	166 MHz	11.7 mW	72 Kbits	0.126 mm^2

Although the main focus is to support the targeted standards, however, programmability of the processor may target some different types of interleaver implementation which is not directly supported by this core. To make it still usable, support for some indirect implementation of any block interleaver with or without having row or column permutations is also provided. In this case the interleaver core is configured to implement a general interleaver with external permutation patterns. The permutation patterns are computed inside baseband processor using its programmability feature and loaded in a couple of the interleaver memories during precomputation phase. Excluding these memories, a restriction on maximum block size (i.e., 4096) will be imposed in this case. This type of approach is adopted by all commercially available interleaver implementations like Xilinx [28], Altera [29], and Lattice Semiconductor [30]. The computation of interleaver permutations on processor side and loading them into memory can impose more computation and time overheads on the processor side. Another drawback is that it does not support fast switching between different interleaver

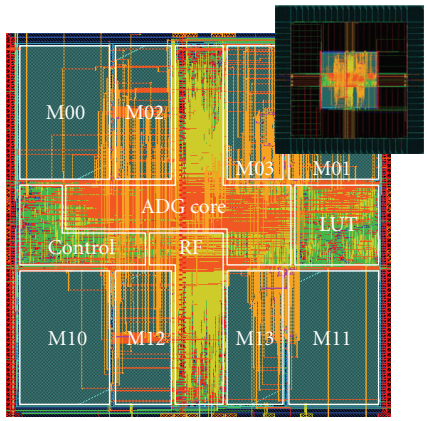


Figure 17: Layout of proposed multimode interleaver.

implementations. A real multimode processor may require fast transition from one standard to another; therefore, it is not a perfect choice for a real multimode environment. However, it is supported by the proposed multimode interleaver core for the completeness of the design.

7. Implementation Results

The reconfigurable hardware interleaver design shown in Figure 4 provides the complete solution for multimode radio baseband processing. The wide range of standard support is the key benefit associated with it. The RTL code for the reconfigurable interleaver design was written in Verilog HDL and the correctness of the design was verified by testing for maximum possible cases. Targeting the use of interleaver core with a multimode baseband processor, one of the important parameters to be investigated is precomputation cycle cost. A lower precomputation cycle cost is beneficial for fast switching between different standards. Table 3 shows the worst case cycle cost during precomputation for different interleavers. It is observed that the cycle cost in WCDMA is higher for some block sizes, but still it works fine, as it is less than the frame size and it can be easily hidden behind the first SISO decoding by the turbo decoder. The worst case precomputation cycle cost for other interleaver implementations is not very high. Therefore, the design supports fast switching among different standards and hence it is very much suitable for a multimode environment.

The multimode interleaver design was implemented in 65 nm standard CMOS technology and it consumes 0.126 mm^2 area. The chip layout is shown in Figure 17 and the summary of the implementation results is provided in Table 4. The design can run at a frequency of 166 MHz and consumes 11.7 mW power in total. Therefore, having 4-bit parallel processing for four spatial streams (e.g., 802.11n) maximum throughput can reach up to 664 Mbps. However, this throughput is limited to 166 Mbps for single stream communication systems. Table 5 provides the comparison of the proposed design to others in terms of standard coverage, silicon cost, and power consumption. The reference implementations have lower standard coverage as compared to the proposed design. Though more silicon is needed for more standard coverage, our solution still provides a good tradeoff with an acceptable silicon cost and power consumption.

8. Conclusion

This paper presents a flexible and reconfigurable interleaver architecture for multimode communication environment. The presented architecture supports a number of standards including WLAN, WiMAX, HSPA+, 3GPP-LTE, and DVB, thus providing coverage for maximum range. To meet the design challenges, the algorithmic level simplifications like 2D transformation of interleaver functions and recursive computation for different implementations are used. The major focus has been to compute the permutation patterns on-the-fly with flexibility. Architecture level results have shown that the design provides a good tradeoff in term of silicon cost and reconfigurability when comparing with other reference designs with less standard coverage. As compared to individual implementations for different standards, the proposed unified address generation offers a reduction of silicon by a factor of three. Finally, the basic requirement of a multimode processor platform, that is, fast switching between different standards has been met with minimal precomputation cycle cost. It enables the processor to use the interleaver core for one standard at some time and use it for another standard in the next time slot by just changing the configuration vector and small preprocessing overheads.

References

[1] A. Nilsson, E. Tell, and D. Liu, "An 11mm^2, 70 mW fully-programmable baseband processor for mobile WiMAX and DVB-T/H in 0.12 μm CMOS," *IEEE Journal of Solid State Circuits*, vol. 44, pp. 90–97, 2009.

[2] E. Tell, A. Nilsson, and D. Liu, "A low area and low power programmable baseband processor architecture," in *Proceedings of the 5th International Workshop on System-On-Chip for Real-Time Applications (IWSOC '05)*, pp. 347–351, Banff, Canada, July 2005.

[3] J. Glossner, D. Iancu, J. Lu, E. Hokenek, and M. Moudgill, "A software-defined communications baseband design," *IEEE Communications Magazine*, vol. 41, no. 1, pp. 120–128, 2003.

[4] 3GPP, "Technical specification group radio access network; multiplexing and channel coding (FDD)," Technical Specification 25.212 V8.4.0, December 2008.

[5] 3GPP-LTE, "Technical specification group radio access network; E-UTRA; multiplexing and channel coding, release 8," Technical Specification 3GPP TS 36.212 v8.0.0, 2007–2009.

[6] IEEE 802.16e-2005, "IEEE standard for local and metropolitan area networks—part 16: air interface for fixed broadband wireless access systems—amendment 2," 2005.

[7] IEEE 802.11-2007, "Standard for local and metropolitan area networks—part 11: WLAN medium access control (MAC) and physical layer (PHY) specs," rev. of IEEE Std. 802.11-1999.

[8] IEEE P802.11n/D2.0, "Draft standard for enhanced WLAN for higher throughput," February 2007.

[9] ETSI EN 300-744 V1.5.1, "Digital video broadcasting (DVB); framing structure, channel coding and modulation for digital terrestrial television," November 2004.

[10] S. Lin and D. J. Costello Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1983.

[11] B. Sklar, *Digital Communications: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2nd edition, 2001.

[12] D. Liu, *Embedded DSP Processor Design, Application Specific Instruction Set Processors*, Morgan Kaufmann, San Mateo, Calif, USA, 2008.

[13] M.-C. Shin and I.-C. Park, "Processor-based turbo interleaver for multiple third-generation wireless standards," *IEEE Communications Letters*, vol. 7, no. 5, pp. 210–212, 2003.

[14] R. Asghar and D. Liu, "Very low cost configurable hardware interleaver for 3G turbo decoding," in *Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA '08)*, pp. 1–5, Damascus, Syria, April 2008.

[15] P. Ampadu and K. Kornegay, "An efficient hardware interleaver for 3G turbo decoding," in *Proceedings of IEEE Radio and Wireless Conference (RAWCON '03)*, pp. 199–120, August 2003.

[16] Z. Wang and Q. Li, "Very low-complexity hardware interleaver for turbo decoding," *IEEE Transactions on Circuits and Systems II*, vol. 54, no. 7, pp. 636–640, 2007.

[17] R. Asghar and D. Liu, "Dual standard re-configurable hardware interleaver for turbo decoding," in *Proceedings of the 3rd International Symposium on Wireless Pervasive Computing (ISWPC '08)*, pp. 768–772, Santorini, Greece, May 2008.

[18] R. Asghar, D. Wu, J. Eilert, and D. Liu, "Memory conflict analysis and implementation of a re-configurable interleaver architecture supporting unified parallel turbo decoding," *Journal of Signal Processing Systems*. In press.

[19] J. B. Kim, Y. J. Lim, and M. H. Lee, "A low complexity FEC design for DAB," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '01)*, vol. 4, pp. 522–525, Sydney, Australia, May 2001.

[20] L. Horvath, I. B. Dhaou, H. Tenhunen, and J. Isoaho, "A novel, high-speed, reconfigurable demapper-symbol deinterleaver architecture for DVB-T," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '99)*, vol. 4, pp. 382–385, Orlando, Fla, USA, May-June 1999.

[21] Y. -N. Chang, "Design of an efficient memory-based DVB-T channel decoder," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '05)*, vol. 5, pp. 5019–5022, Kaohsiung, Taiwan, May 2005.

[22] H. Afshari and M. Kamarei, "A novel symbol interleaver address generation architecture for DVB-T modulator," in *Proceedings of the International Symposium on Communications and Information Technologies (ISCIT '06)*, pp. 989–993, Bangkok, Thailand, October 2006.

[23] Y.-N. Chang and Y.-C. Ding, "A low-cost dual-mode deinterleaver design," in *Proceedings of IEEE International Conference on Consumer Electronics (ICCE '07)*, pp. 1–2, Las Vegas, Nev, USA, January 2007.

[24] Y. N. Chang, "A low-cost dual mode de-interleaver design," *IEEE Transaction on Consumer Electronics*, vol. 54, no. 2, pp. 326–332, 2008.

[25] Y.-W. Wu, P. Ting, and H.-P. Ma, "A high speed interleaver for emerging wireless communications," in *Proceedings of the International Conference on Wireless Networks, Communications and Mobile Computing*, vol. 2, pp. 1192–1197, Maui, Hawaii, USA, June 2005.

[26] R. Asghar and D. Liu, "Low complexity multi mode interleaver core for WiMAX with support for convolutional interleaving," *International Journal of Electronics, Communications and Computer Engineering*, vol. 1, no. 1, pp. 20–29, 2009.

[27] R. Asghar and D. Liu, "Low complexity hardware interleaver for MIMO-OFDM based wireless LAN," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '09)*, pp. 1747–1750, Taipei, Taiwan, May 2009.

[28] Xilinx Inc., "Interleaver/De-Interleaver," Product Specification, v5.1, DS250, March 2008.

[29] Altera Inc., "Symbol Interleaver/De-Interleaver Core," Mega Core Function User's Guide, ver. 1.3.0, June 2002.

[30] Lattice Semiconductor Inc., "Interleaver/De-Interleaver IP Core," ispLever Core User's Guide, ipug_61_02.5, August 2008.

[31] X.-F. Wang, Y. R. Shayan, and M. Zeng, "On the code and interleaver design of broadband OFDM Systems," *IEEE Communications Letters*, vol. 8, no. 11, pp. 653–655, 2004.

[32] R. Van Nee, V. K. Jones, G. Awater, A. Van Zelst, J. Gardner, and G. Steele, "The 802.11n MIMO-OFDM standard for wireless LAN and beyond," *Wireless Personal Communications*, vol. 37, no. 3-4, pp. 445–453, 2006.

[33] H. Niu, X. Ouyang, and C. Ngo, "Interleaver design for MIMO-OFDM based wireless LAN," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '06)*, vol. 4, pp. 1825–1829, Las Vegas, Nev, USA, 2006.

[34] J. Baltersee, G. Fock, and H. Meyr, "Achievable rate of MIMO channels with data-aided channel estimation and perfect interleaving," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 12, pp. 2358–2368, 2001.

[35] S. Ramseier, "Shuffling bits in time and frequency—an optimum interleaver for OFDM," in *Proceedings of the IEEE International Conference on Communications (ICC '03)*, vol. 5, pp. 3418–3422, May 2003.

[36] V. D. Nguyen and H.-P. Kuchenbecker, "Block interleaving for soft decision viterbi decoding in OFDM systems," in *Proceedings of the 54th IEEE Vehicular Technology Conference (VTC '01)*, vol. 1, pp. 470–474, 2001.

[37] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and encoding: turbo-codes," in *Proceedings of IEEE International Conference on Communications (ICC '93)*, vol. 2, pp. 1064–1070, Geneva, Switzerland, May 1993.

[38] G. R. Blakley, "A computer algorithm for calculating the product A*B mod M," *IEEE Transactions on Computers*, vol. 32, no. 5, pp. 497–500, 1983.

[39] J.-H. Kim and I.-C. Park, "Duo-binary circular turbo decoder based on border metric encoding for WiMAX," in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC '08)*, pp. 109–110, Seoul, Korea, March 2008.

[40] G. D. Forney, "Burst- correcting codes for the classic bursty channel," *IEEE Transactions on Communications*, vol. 19, no. 5, part 2, pp. 772–781, 1971.

[41] J. L. Ramsey, "Realization of optimum interleavers," *IEEE Transactions on Information Theory*, vol. 16, no. 3, pp. 338–345, 1970.

[42] S. Vafi and T. Wysocki, "Weight distribution of turbo codes with convolutional interleavers," *IET Communications*, vol. 1, no. 1, pp. 71–78, 2007.

[43] E. K. Hall and S. G. Wilson, "Stream-oriented turbo codes," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1813–1831, 2001.

[44] S. Vafi and T. Wysocki, "On the performance of turbo codes with convolutional interleavers," in *Proceedings of Asia-Pacific Conference on Communications*, pp. 222–226, Perth, Wash, USA, October 2005.