Luis ORTIZ[1], Victor RANGEL[1], Javier GOMEZ[1], Raul AQUINO[2], Miguel LOPEZ-GUERRERO[3]

National Autonomous University of Mexico (1), University of Colima (2), Metropolitan Autonomous University (3)

Performance Evaluation of VoIP Traffic over the IEEE 802.16e Protocol with Different Modulation and Coding Schemes

Abstract. *Supporting as many Voice-over-IP (VoIP) users as possible in a Broadband Wireless Access network using limited radio resources is a critical issue. However, performance of VoIP services is affected by several parameters defined in the IEEE 802.16e protocol. In this paper we developed a theoretical model and an algorithm to evaluate the performance of some of the most important VoIP codecs. Simulation results validate the theoretical model to achieve the maximum number of VoIP streams for different configurations of an IEEE 802.16e system.*

Streszczenie. *W artykule przedstawiono model oraz algorytm działania, służące do oceny jakości pracy niektórych z najważniejszych codec'ów dla bramek VoIP przy obsłudze dużej ilości użytkowników. Badania symulacyjne potwierdziły skuteczność działania modelu, w celu osiągnięcia maksymalnej ilości strumieni VoIP przy różnych konfiguracjach systemu IEEE 802.16e.* ***(Działanie protokołu IEEE 802.16e przy dużej ilości użytkowników VoIP – zagadnienie schematów modulacji i kodowania).***

Keywords: IEEE 802.16e, Mobile WiMAX, VoIP, Performance Evaluation.
Słowa kluczowe: IEEE 802.16e, Mobile WiMAX, VoIP, ocena działania.

Introduction

Supporting as many VoIP users as possible using limited radio resources is a very important issue because VoIP is expected to be widely supported by mobile wireless networks. However, performance of VoIP services in the IEEE 802.16e standard [1] is affected by various factors such as signaling overhead, ranging regions and wasted symbols due to "rectangulation and quantization".

Rectangulation is the process of allocating bandwidth resources in the downlink (DL) channel on a square or rectangular region of the frame structure. Quantization is the process of allocating these resources using the minimum allocation unit, denominated "quantum map". Considering Partial Usage of Sub-Channels (PUSC), a quantum map or a slot in the DL channel is one sub channel consisting of two OFDMA (Orthogonal Frequency Division Multiple Access) symbols. A quantum map in the uplink (UL) channel is one sub channel consisting of three OFDMA symbols. These settings seriously affect VoIP service performance because a VoIP packet allocation generally does not fit well in the DL and UL reservation space and usually multiple quantum maps are required.

Signaling overhead of MAP messages also affects VoIP services, because such overhead increases when the Base Station (BS) schedules small-sized VoIP packets. Some studies have evaluated VoIP performance taking into consideration mapping overhead [2], [3]. However, in these studies neither wasted resources in the DL channel due to quantization and rectangulation were considered nor power calibration (ranging) and contention signaling on the UL channel were taken into account.

A previous work presented an analytical model to evaluate VoIP performance [4]. This study takes into account the quantization and ranging regions in the DL and UL channels, respectively. However, wasted symbols due to rectangulation were not considered. The performance of some speech codecs has been evaluated in [5] and [6]. In [6] wasted symbols were considered, however they were considered as a result of variations of VoIP inter-arrival times and packet sizes. Moreover, in [5-6], the mapping overhead was not considered and the performance optimization of speech codecs based on OFDMA symbols for UL and DL channels was not addressed.

For a detailed description of 1) frame format, 2) initialization and registration process, 3) bandwidth reservation process and 4) Quality of Service classes of the IEEE 802.16e protocol, we refer the reader to our previous work [7], which also includes a simple performance analysis of mobile WiMAX networks for VoIP streams considering only G.711 and G.723 speech codecs.

In this paper, we present a performance optimization of a mobile WiMAX network in order to support the maximum number of VoIP streams, using most common codecs: G.711, G.729, G.728 and G.723.1. We also implemented a simulation model to analyze end-to-end delays and other important issues that could not be captured by the theoretical model. Several modulations and codifications were taken into account in order to evaluate the performance of different VoIP codecs using three packet encapsulation types: without Header Suppression (-HS), with Header Suppression (+HS [8-9]) and compressed Real-time Transport Protocol (+cRTP[10]).

Performance Analysis of VoIP Traffic

In this section, we present the performance analysis of the IEEE 802.16e MAC protocol when VoIP traffic is carried using a 20 MHz channel. The theoretical model we have derived for the performance analysis can also be used to study other applications. However, in this study we evaluate CBR traffic to load the network with short VoIP packets, when the UGS service class is used. From Fig. 1, we can see that the DL sub frame is comprised of a Preamble, a FCH (frame control header), a DL-MAP message, a UL-MAP message and DL bursts. According to the standard [1] the Preamble and the FCH are of constant size, but the DL-MAP and the UL-MAP are of variable size. Here DL bursts are also constants since they are used to transport fixed-size VoIP frames. Therefore in order to know the number of VoIP streams supported in the DL sub frame ($VoIPstreams_{DL}$), we just need to compute the available number of OFDMA symbols at the PHY layer in the DL sub frame (Avl_{smbDL}), subtract the overhead (FCH, DL-MAP and UL-MAP), and compute how many DL VoIP bursts fit in the last symbols, (considering the total wasted symbols, if any). Similarly, we follow the same procedure to compute the number of VoIP streams supported in the UL sub frame ($VoIPstreams_{UL}$). We just need to compute the available number of OFDMA symbols at the PHY layer in the UL sub frame (Avl_{smbUL}), subtract the ranging regions and compute how many UL VoIP bursts fit in the last symbols (in this case there are no wasted symbols).

Finally, the maximum number of VoIP streams supported ($MaxVoIPstreams$) in a 20 MHz channel for the transmission of voice traffic will be the minimum of $VoIPstreams_{DL}$ and $VoIPstreams_{UL}$.

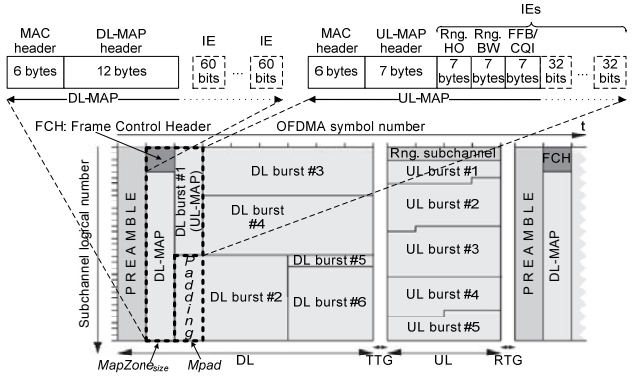

Fig.1. Frame structure for the IEEE 802.16e MAC protocol

Theoretical Model

For the modeling of the IEEE 802.16e protocol, we used the parameters given in Table 1. These parameters include the default values given by the standard [1]. As the grants (used to transmit data traffic) have to be reserved in quantum map units, the available number of OFDMA symbols (Avl_{smbDL}) has to be rounded to multiples of quantum maps. Hence, the available number of OFDMA symbols in the DL sub frame is given by Eq. (1); here we have taken out one OFDMA symbol for the Preamble,

(1) $$Avl_{smbDL} = \left\lfloor \frac{OFDMA_{smbDL} - 1}{Qsmb_{DL}} \right\rfloor * Qsmb_{DL} * Data_{sbcrDL}$$

where: $OFDMA_{smbDL}$ – number of OFDMA symbols used in the DL sub frame; $Qsmb_{DL}$ – quantum symbol for the DL sub frame (in OFDMA symbols); $Data_{sbcrDL}$ – data Subcarriers.

Table 1. MAC and PHY layer parameters for a 20 MHz channel

Parameter	Definition	Default value	
$Frame_d$	Frame duration	5ms	
FCH_{sbch}	FCH sub channels	1	
FCH_{smb}	FCH symbols	2	
FFB_{smb}	Symbols for Fast Feedback Channel Quality Information (FFB/CQI)	6	
FFB_{sbch}	Sub channels for FFB/CQI.	1	
$IErng_{bytes}$	Ranging information element size (bytes)	7	
$MACHdr_{bytes}$	MAC header (bytes)	6	
N	Number of VoIP streams	-	
$OFDMA_{smb}$	OFDMA symbols (see Table 2)	θ[1-4]	
Rng_{smbBW}	Symbols for ranging and BW request	1	
Rng_{smbHO}	Symbols for ranging handoff	2	
Rng_{sbchBW}	Sub channels for ranging and BW request	6	
Rng_{sbchHO}	Sub channels for ranging handoff	6	
		Subframe	
		UL	***DL***
$Data_{sbcr}$	Data Subcarriers	1120	1440
$IEsize_{bits}$	Information element size	32	60
$MapHdr_{bytes}$	Map header (bytes)	7	12
$Qsmb$	Quantum symbol size	3	2
$RepCnt$	Repetition count	1	4
$SbCh$	Sub channels	70	60
$SbCr_{sbch}$	Sub carriers per Subchannel	16	24

According to the standard [1], $OFDMA_{smbDL}$ can be set to different values (denoted by "θ" in Table 2). For each VoIP codec, modulation and codification, we derived an algorithm that chooses the best configuration from Table 2 in order to achieve the maximum number of VoIP streams on each channel.

Table 2. UL and DL sub frame configuration

Configuration	$OFDMA_{smb}$	
	UL	DL
θ0	9	38
θ1	12	35
θ2	15	32
θ3	18	29
θ4	21	26
θ5	24	23

One of the performance problems is the *"signaling overhead of control messages*" (SOCM) consumed in the DL sub frame, such as the FCH, DL-MAP and UL-MAP messages. The map zone size ($MapZone_{size}$) computes the number of OFDMA symbols consumed by SOCM messages as the number of VoIP streams increases. Thus, $MapZone_{size}$ is given by

(2) $$MapZone_{size} = FCH_{size} + Map_{sizeDL} + Map_{sizeUL}$$

where: Map_{sizeDL} - length of the DL-MAP sub frame; Map_{sizeUL} - length of the UL-MAP sub frame as shown in Fig.1; FCH_{size} - length of the frame control header given by

(3) $$FCH_{size} = \left\lceil \frac{FCH_{smb} * FCH_{sbch} * SbCr_{sbchDL}}{Qmap_{DL}} \right\rceil * Qmap_{DL} * RepCnt_{DL}$$

where: FCH_{smb}, FCH_{sbch} and $SbCr_{sbchDL}$ are the number of OFDMA symbols, the number of sub channels and the number of subcarriers per sub channel in the DL direction assigned to the FCH region, respectively; $RepCnt_{DL}$ - DL repetition count; $Qmap_{DL}$ - DL Quantum MAP given by

(4) $$Qmap_{DL} = Qsmb_{DL} * SbCr_{sbchDL}$$

where: $Qsmb_{DL}$ - length of the quantum symbol in the DL sub frame.

The DL-MAP sub frame contains Information Elements (IEs) used by the subscriber stations (SSs) to decode their grants in the DL sub frame. The DL-MAP size (Map_{sizeDL}) depends on the number of VoIP streams (N) allocated in the DL sub frame. Therefore the DL-MAP size is defined as

(5) $$Map_{sizeDL} = \left\lceil \frac{\left(MACHdr_{bytes} + MapHdr_{bytesDL}\right) * 8 + N * IEsize_{bitsDL}}{Qmap_{DL}} \right\rceil \left(\frac{Qmap_{DL}}{RepCnt_{DL}^{-1}} \right); \quad \text{for } N = 0,1,2,3,\ldots$$

where: $MACHdr_{bytes}$ - generic MAC header; $MapHdr_{bytesDL}$ - DL-MAP header; $IEsize_{bitsDL}$ - DL IE size as is shown at the top of Fig. 1.

In (3) and (5), $RepCnt_{DL}$ is used because the BS must ensure that SOCM messages for SS operation are correctly received.

Similarly, we computed the UL-MAP size as:

(6) $$Map_{sizeUL} = \left\lceil \frac{\left(MACHdr_{bytes} + MapHdr_{bytesUL} + 3 * IErng_{bytes}\right) * 8 + N * IEsize_{bitsUL}}{Qmap_{DL}} \right\rceil \left(\frac{Qmap_{DL}}{RepCnt_{UL}^{-1}} \right); \quad \text{for } N = 0,1,2,3,\ldots$$

where: $MapHdr_{bytesUL}$ - UL-MAP header; $IErng_{bytes}$ - Ranging IE size; $IEsize_{bitsUL}$ - UL IE size; $RepCnt_{UL}$ - UL repetition count.

In (6), $IErng_{bytes}$ is used by the Handoff region, the Bandwidth Requests region and the Fast Feedback Channel Quality Information (FFB/CQI) region.

Then, the number of VoIP streams supported in the DL sub frame is defined by:

$$VoIPstreams_{DL} = \frac{CIntArv_{time} * \max(N)}{Frame_d} \quad \text{for } N = 0,1,2,3,\ldots \tag{7}$$
$$\left| \; Avl_{smbDL} - MapZone_{size} - N * SSVoIP_{DL} - Mpad - Wstsmb_{(N)} \geq 0; \right.$$

where: $CIntArv_{time}$ - codec inter-arrival time (see Table 3); $Frame_d$ - Frame duration; $SSVoIP_{DL}$ - DL VoIP stream size; $Mpad$ - padding of map zone wasted by $MapZone_{size}$; $Wstsmb_{(N)}$ - total wasted symbols in the DL sub frame.

Table 3. Characteristics of the considered codecs

Codec	Bit rate	Codec inter-arrival time (ms) $CIntArv_{time}$	VoIP frame size (application layer)
G.711	64 kbps	10	80 bytes
G.723	5.3 kbps	30	20 bytes
G.726	32 kbps	10	40 bytes
G.728	16 kbps	2.5	40 bits (5 bytes)
G.729	8 kbps	10	10 bytes

Both $Mpad$ and $Wstsmb_{(N)}$ are wasted symbols due to rectangulation and quantization. $max(N)$ means the maximum N such that Avl_{smbDL}-$MapZone_{size}$-N*$SSVoIP_{DL}$-$Mpad$-$Wstsmb_{(N)} \geq 0$. In order to compute $SSVoIP_{DL}$ we need to obtain the VoIP frame size at the PHY layer ($VoIPFrame_{PHY}$) and then apply the modulation and codification overhead factor. Thus, the DL VoIP stream size is defined by:

$$SSVoIP_{DL} = \left\lceil \frac{VoIPFrame_{PHY}}{M * cc * Qmap_{DL}} \right\rceil * Qmap_{DL} \tag{8}$$

where: M - number of bits per symbol (2 for QPSK, 4 for 16-QAM and 6 for 64-QAM), cc - convolutional coding rate (1/2, 2/3, 3/4 or 5/6).

For the performance analysis, the most common VoIP codecs were considered (that is, G.711, G.723, G.726, G.728 and G.729). These are described as follows:

1) Codec G.711 [11] was considered in order to load the IEEE 802.16e network and because this codec will be used for quality voice calls. G.711 is the mandatory codec according to the ITU-T H.323 conferencing standard [12], which uses Pulse Code Modulation to produce a data rate of 64 kbps at the application layer. This codec creates and encapsulates an 80-byte VoIP frame every 10 ms.

2) According to the ITU, IETF and the VoIP Forum, G.723.1 (G.723 from now on) [13] is the preferred speech codec for Internet telephony applications. This codec generates a data rate of 5.3 kbps at the application layer, where a 20-byte VoIP frame is generated every 30 ms.

3) Codec G.726 uses Adaptive Differential Pulse Code Modulation (ADPCM) scheme according to the ITU G.726 recommendation [14]. This codec generates a data rate of 32 kbps at the application layer, where a 40-byte VoIP frame is generated every 10 ms.

4) According ITU 6.728 recommendation [15], codec G.728 uses Low-Delay Code Excited Linear Prediction (LD-CELP) and generates a data rate of 16 kbps at the application layer and a 40-bit VoIP frame is generated every 2.5 ms.

5) Codec G.729 uses Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP) speech compression algorithm, approved by ITU [16]. It is mostly used in VoIP applications where bandwidth must be economized. It generates a 10-byte VoIP frame every 10 ms, producing a data rate of 8 kbps.

As an example of how to obtain the $SSVoIP_{DL}$, Fig. 2a illustrates the encapsulation process for a G.711 codec using two different modulations QPSK 1/2 (M=2, cc=1/2) and 64-QAM 3/4 (M=6, cc=3/4).

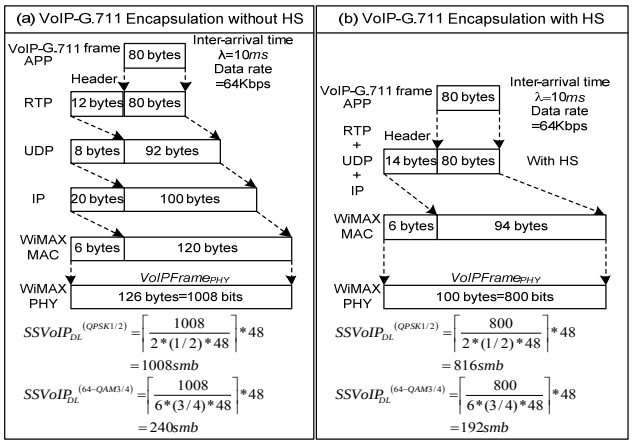

Fig.2. VoIP encapsulation for a G.711 codec, with and without header suppression, (a) and (b), respectively

According to [7] and [8], header suppression (HS) is possible, so we can disregard fixed fields of the RTP, UDP and IP headers. This results in a reduction from 40 bytes to 14 bytes of header as shown in Fig. 2b. This reduction of RTP+UDP+IP headers (VoIP frame overhead), will increase system performance as indicated in the following sections.

As we mentioned, padding symbols are wasted by the $MapZone_{size}$. This is because the map zone has to fill the symbols from the last sub channels to form a rectangular region, as it is shown in Fig. 1. Thus wasted padding symbols by the $MapZone_{size}$ are defined by

$$Mpad = \left(\left\lceil \frac{MapZone_{size}}{Qsmb_{DL} * Data_{sbcrDL}} \right\rceil * Qsmb_{DL} * Data_{sbcrDL} \right) - MapZone_{size} \tag{9}$$

With each grant being allocated at the last sub channels of the DL sub frame, it could be possible that data do not fit well in these sub channels. So, data allocation has to be moved to start at the next OFDMA symbol of the first sub channel. This generates an offset as is shown in Fig. 3.

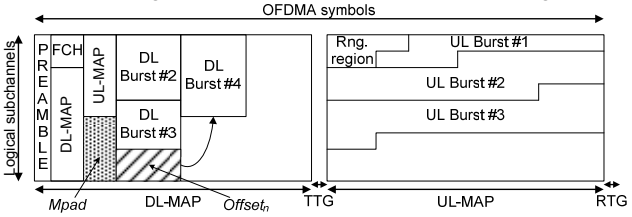

Fig.3. Offset generated by data allocation

Thus, the offset generated by the n-th data allocation is defined as:

$$Offset_n = \left\lceil SbCh_{DL} - Mod\left(\frac{(n-1) * SSVoIP_{DL} + Wstsmb_{(n-1)}}{Qmap_{DL}} \right) \right\rceil * Qmap_{DL}; \quad \text{for } n \geq 1 \tag{10}$$

where: $SbCh_{DL}$ – sub channels on the DL sub frame.

This expression considers the previous VoIP streams already allocated in the DL sub frame ($(n\text{-}1)$*$SSVoIP_{DL}$) and the total wasted symbols due to all previous data allocations in the DL sub frame ($Wstsmb_{(n-1)}$). Therefore, the total

number of wasted symbols in the DL sub frame due to N VoIP-stream allocations is obtained as:

(11) $$Wstsmb_{(N)} = \sum_{n=1}^{N} Goffset_n$$

where: $Goffset_n$ - grant offset wasted. We need to guarantee that $Offset_n$ corresponds to a grant that was tried to be allocated at the last sub channels of the DL sub frame. Hence, the grant offset wasted by the n-th $SSVoIP_{DL}$ allocation in the DL sub frame is defined by:

(12) $$Goffset_n = \begin{cases} Offset_n; & if\ Offset_n < SSVoIP_{DL} \\ 0 & otherwise \end{cases}$$

We also need to know the number of VoIP streams supported in the UL sub frame. Thus, the available number of OFDMA symbols in the UL sub frame (Avl_{smbDL}) is computed as:

(13) $$Avl_{smbUL} = \left[\begin{array}{l} \left(\left(\left\lfloor \frac{OFDMA_{smbUL}}{Qsmb_{UL}} \right\rfloor * Qsmb_{UL} * Data_{sbcrUL} \right) \right. \\ \left. - Rng_{smbHO} * Rng_{sbchHO} * SbCr_{sbchUL} \right) \\ - Rng_{smbBW} * Rng_{sbchBW} * SbCr_{sbchUL} \\ - FFB_{smb} * FFB_{sbch} * SbCr_{sbchUL} \end{array} \right]$$

where: $OFDMA_{smbUL}$ – number of OFDMA symbols used in the UL sub frame; $Qsmb_{UL}$ – quantum symbol for the UL sub frame (in OFDMA symbols); $Data_{sbcrUL}$ – data subcarriers in the UL sub frame; $SbCr_{sbchUL}$ - number of subcarriers per sub channel in the UL sub frame; Rng_{smbHO} and Rng_{sbchHO} are the number of symbols and sub channels used for ranging and handoff, respectively. Rng_{smbBW} and Rng_{sbchBW} are the number of symbols and sub channels used for ranging and bandwidth (BW) request, respectively. FFB_{smb} and FFB_{sbch} are the number of symbols and sub channels used for Fast Feedback Channel Quality Information (FFB/CQI), respectively.

Then, the number of VoIP streams supported in the UL sub frame is obtained as:

(14) $$VoIPstreams_{UL} = \left(\frac{CIntArv_{time}}{Frame_d} \right) \left\lfloor \frac{Avl_{smbUL}}{SSVoIP_{UL}} \right\rfloor$$

where: $SSVoIP_{UL}$ - UL VoIP stream size given by

(15) $$SSVoIP_{UL} = \left\lceil \frac{VoIPFrame_{PHY}}{M * cc * Qmap_{UL}} \right\rceil * Qmap_{UL}$$

where: $Qmap_{UL}$ - UL Quantum MAP.

In (15) we also have applied the minimum reservation unit in the UL channel ($Qmap_{UL}$), which is defined by:

(16) $$Qmap_{UL} = Qsmb_{UL} * SbCr_{sbchUL}$$

where: $Qsmb_{UL}$ - length of the quantum symbol in the UL sub frame.

Finally, the maximum number of VoIP streams supported is defined as *min(VoIPstreams*$_{DL}$, $VoIPstreams_{UL}$).

Simulation Model

In order to validate the theoretical model, we implemented a mobile WiMAX network simulation model based on the OPNET MODELER package v.16. At the top level of the IEEE 802.16e network model there are network components, for example the BS, SSs and servers, as is shown in Fig. 4a. The next hierarchical level, Fig. 4b, defines the functionality of a SS in terms of components such as traffic sources, TCP/UDP, IP, MAC and PHY interfaces. The operation of each component is defined by a Finite State Machine (an example of which is shown in Fig. 4c).

The actions of a component at a particular state are defined in Proto-C code (see Fig. 4d). This approach allows modifications to be applied to the operation of the IEEE 802.16e MAC protocol and different optimizations and enhancements to be tested. The parameters used for the simulation model were the same as the ones used in the theoretical model (shown in Table 1).

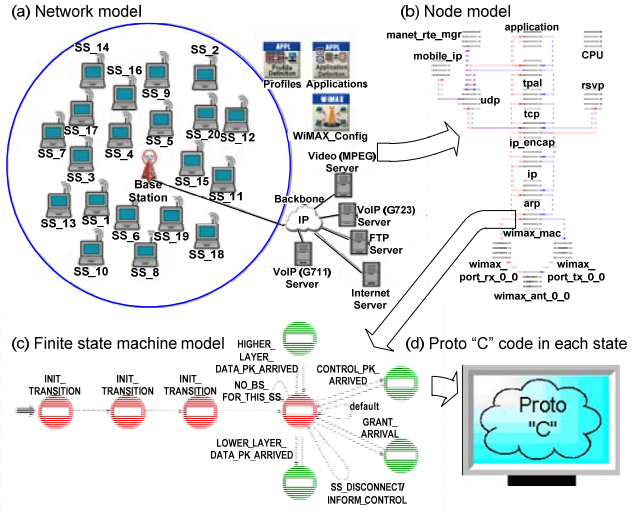

Fig.4. IEEE 802.16e simulation model

Results

The performance analysis of VoIP traffic in a mobile WiMAX network is of great importance for the 4G telecommunications' community. This study determines the maximum number of VoIP phone calls that can be supported so that a WiMAX Mobile network is not overloaded, otherwise the network would result in a lower system performance.

For the performance analysis we modeled a 20 MHz time division duplex (TDD) channel, using the configuration parameters as indicated in Table 1. For the first performance scenario, we evaluated two different codecs, G.711 and G.723. G.711 was chosen because is used for quality voice calls and thus it consumes more resources than the other ones. In contrast, G.723 was chosen because is the preferred speech codec for Internet telephony applications and opposite to G.711, it has a low bandwidth consumption. For each codec we also employed two modulations: QPSK with convolutional coding = 1/2 (QPSK 1/2) and 64-QAM with convolutional coding = 3/4 (64-QAM 3/4).

We considered different frame configurations (see Table 2) in order to evaluate the system throughput and find out the maximum number of supported VoIP streams.

In Fig. 5(a-b) we can see the throughput for the UL direction, here both models (theoretical and simulation) were used, and the results are in good agreement. We also found the same throughput values for the DL direction, thus Fig. 5 also applies for the DL channel. In Fig. 5a, the maximum number of quality phone calls (*VoIPstreams*) supported was of 38 using G.711 codec with QPSK 1/2, without HS (-HS) and θ3. This is the result of having 38 outgoing VoIP streams in the UL sub frame and 38 ingoing VoIP streams in the DL sub frame.

When HS and θ4 frame configuration is considered this number is increased by 42.1%, so the maximum number of quality phone calls grows to 54. This is because considering HS, the VoIP frame overhead is reduced considerably and more VoIP streams can be allocated. By changing the modulation to 64-QAM 3/4 and using a θ3 frame configuration without considering HS, the maximum number of quality phone calls becomes of 144. However, when HS is employed, this number increases to 170 when a θ2 frame configuration is used, resulting in an 18% increase.

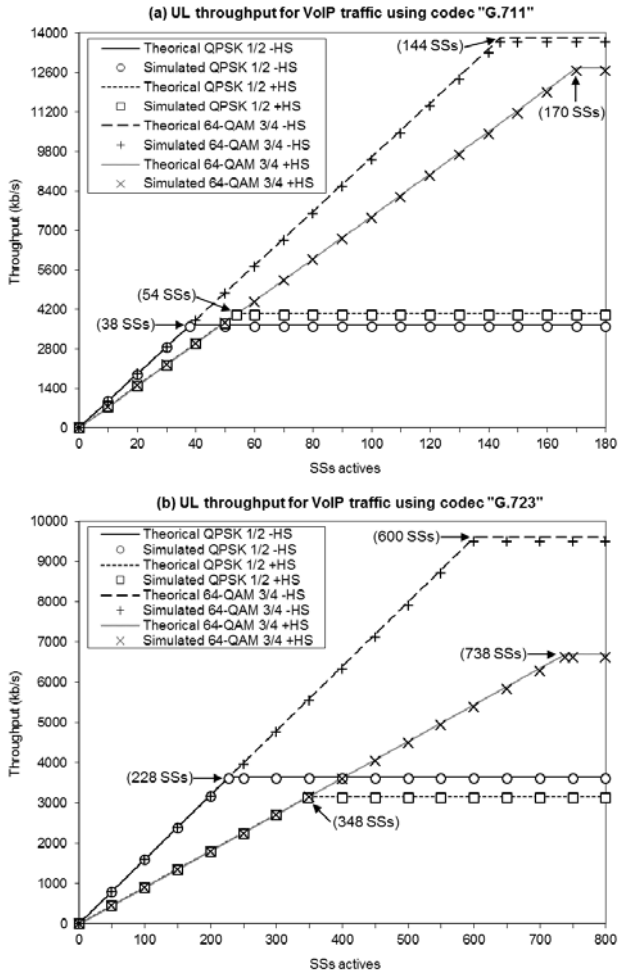

Fig.5. Maximum throughput of VoIP traffic in a 20 MHz channel

Although more VoIP streams are supported, we can see a throughput reduction (8.13%). This is due to the fact that with HS in effect a significant amount of resources are freed and this amount largely exceeds the amount consumed by the VoIP streams. For instance, with -HS the maximum throughput achieved was 13.8 Mbps (= 144 SSs * 96 kbps, where DL-MAP + UL-MAP = 4.1 Msmb/s) compared to 12.8 Mbps (= 170 SSs * 75.2 kbps, where DL-MAP + UL-MAP = 4.9 Msmb/s) when HS was considered. Moreover in this case, more OFDMA symbols were configured to DL sub frame (θ2), to compensate for resources consumed by the signaling overhead of MAP messages.

Fig. 6 shows the allocations of VoIP streams (bursts) in both directions, DL and UL, where the empty space could not be allocated for the transmission of VoIP traffic, since it is not possible to have fragmented VoIP frames when UGS is used and due to the rectangulation and quantization process. However most of this empty space is allocated for the transmission of more VoIP bursts when 64-QAM 3/4 and HS are considered, due to the fact that VoIP burst size is considerably reduced and fits better in the unscheduled symbols.

Similarly, Fig. 5b shows the UL throughput for codec G.723, which also applies to DL direction. We can see that the maximum number of VoIP phone calls is considerably increased from 228 (with -HS, θ4, QPSK 1/2) to 348 (with HS, θ3, QPSK 1/2). Here, there was an increase of 52.6% in VoIP phone calls.

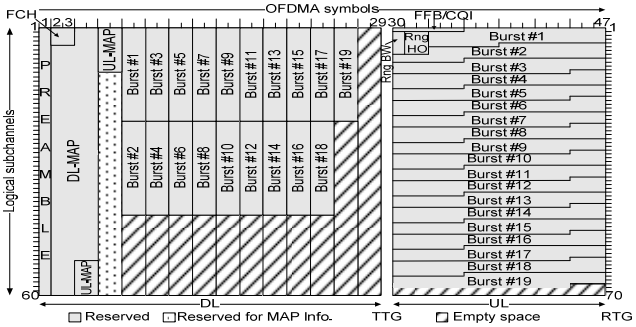

Fig.6. MAP and VoIP burst allocation for codec G.711-QPSK1/2

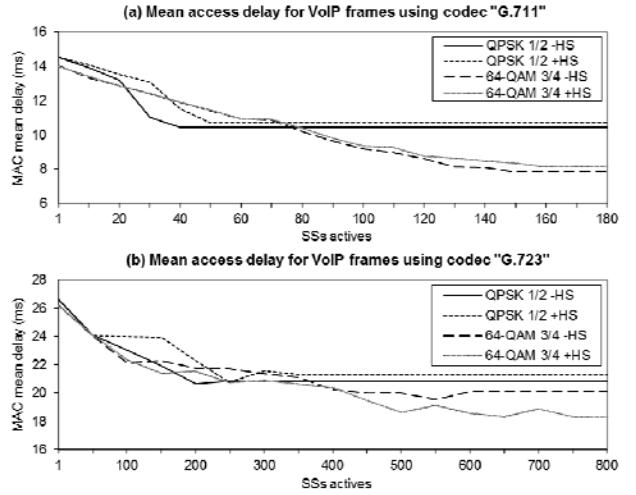

Fig.7. VoIP traffic mean access delay in a 20 MHz channel

However, these VoIP phone calls are performed with a medium quality, since MOS (Mean Opinion Score) = 3.6 for codec G.723, compared to MOS = 4.4 for codec G.711. Once again, we can see a throughput reduction (15.6%) due VoIP frame overhead reduction. By using 64-QAM 3/4, the number of VoIP phone calls can be increased from 600, (with -HS, θ2) to 738 (with HS, θ1). Here the increase was of 23% in VoIP phone calls, but the throughput reduction was also important (43.5%). When a big amount of VoIP phone calls is considered for HS, the VoIP frame overhead reduction is significant. Moreover, when small-sized VoIP frames are considered (for instance G.723, with 64-QAM 3/4) VoIP bursts fit better in the last sub channels of DL sub frame and thus fewer symbols are wasted. This analysis can be directly applied to fixed nodes, where the modulation type can be negotiated with the BS at connection setup, however for mobiles nodes, it is recommended to use QPSK 1/2 for bandwidth estimation and use unscheduled symbols for nrtPS or BE services, since these types of service can support fragmentation.

Fig. 7 shows the mean access delay of VoIP frames in the UL direction. According to "PacketCable™ Audio/Video Codecs Specification" [17], in order to estimate the one-way delay we need to know: 1) Coding delay (consisting of Encoding and Decoding delays), 2) Access delay (consisting of MAC access delay, transmission delay and

propagation delay), and 3) Look-ahead delay. Coding and look-ahead delays are constants and account for 20 ms and 67 ms for codec G.711 and G.723, respectively. In Fig. 7a, for codec G.711 we see that the simulated mean access delay is less than 15 ms, plus coding and look-ahead delays the point to point (PtP) delay remains under 45ms, which is under the maximum PtP delay allowed for VoIP calls, 150ms. For codec G.723, as shown in Fig. 7b, the simulated mean access delay remained under 27ms, this delay becomes less than 94ms when coding and look-ahead delays are considered, which is still below the maximum PtP delay.

Performance optimization for VoIP Traffic

In order to carry out the performance optimization for VoIP traffic, we designed an algorithm which chooses the best θ frame configuration from Table 2 in order to achieve the maximum number of VoIP phone calls. We evaluated the performance of codecs, G.711, G.723, G.726, G.728 and G.729 using different modulations and codings (QPSK 1/2, QPSK 3/4, 16-QAM 1/2, 16-QAM 3/4, 64-QAM 2/3, 64-QAM 3/4 and 64-QAM 5/6). We also evaluated the performance of VoIP traffic considering two repetition counts for FCH and DL-MAP (*RepCnt*=4 and *RepCnt*=1, respectively); moreover HS and cRTP were also considered.

The operating principle of the algorithm is based on the coding repetition count (*CRepCount*), which is defined by:

$$CRepCount = \begin{cases} 1; & if \quad \frac{CIntArv_{time}}{Frame_d} < 1 \\ \frac{CIntArv_{time}}{Frame_d} & otherwise \end{cases} \tag{18}$$

The coding repetition count specifies the number of frames a VoIP stream has to wait in order to be allocated, as it is shown in Fig. 8. Here, we can see a VoIP stream that is allocated every frame (*CRepCount*=1), one that is allocated every two frames (*CRepCount*=2), another one that is allocated every three frames (*CRepCount*=3) and finally, one more that is allocated every six frames (*CRepCount*=6). As we can see, in this case $Frame_n$=$Frame_{n+6}$. We called this a "cycle", because based on the fact that VoIP streams are CBR traffic, then the same VoIP streams have to be allocated every six frames .

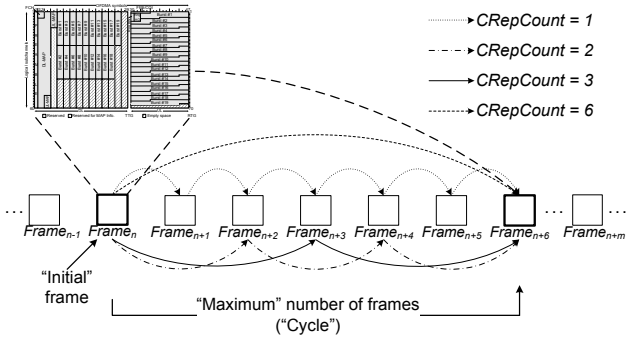

Fig.8. Algorithm operation principle

Therefore, if $Frame_n$ is filled with data allocations (for UGS traffic), $Frame_{n+6}$ will be filled too. This means that only $Frame_{n+1}$ to $Frame_{n+5}$ have available symbols for more data allocations. Although this algorithm can model, different VoIP codecs simultaneously, we modeled one VoIP codec at a time in order to evaluate the performance of each VoIP codec individually.

In Fig. 9 we show the algorithm used to optimize VoIP traffic performance. This algorithm begins by initializing a temporary variable, *maxvoipstr*. This variable stores the maximum number of VoIP streams reached at each iteration. Then in a Round Robin (RR) mode, each of the configuration parameters is selected. First, the DL-MAP repetition count is selected (1 or 4). Then, a VoIP codec, a modulation and codification are selected. Next, a packet encapsulation (-HS, +HS or +cRTP) is selected.

Finally, the frame configuration θ is selected. Once all parameters are selected, the algorithm computes the maximum number of supported VoIP streams (*MaxVoIPstreams*) using the theoretical model described in section II. This value is compared with the maximum number of VoIP streams reached (*maxvoipstr*), and the highest value of these two variables is stored in the *maxvoipstr* temporary variable. This guarantees that at the end of all iterations the maximum number of VoIP streams reached will be stored in *maxvoipstr*. Therefore, Table 4 shows the maximum number of VoIP streams reached with the best θ frame configuration for the different VoIP codecs modeled, using different packet encapsulation types. Here, the FCH and DL-MAP repetition count (DLr) is also indicated.

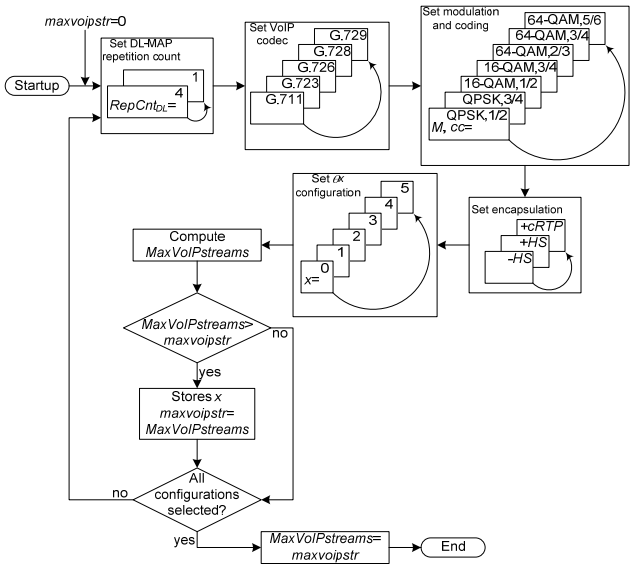

Fig.9. Algorithm for VoIP traffic performance optimization

Discussion and Conclusions

The performance optimization presented in this paper indicates that VoIP streams under different configurations can be supported by the mobile WiMAX protocol. There are, however, performance issues that need to be considered.

The general trend from the results was that the system would comfortably support a number of active SSs making a VoIP phone call, where the maximum system throughput is obtained at the point when all available OFDMA symbol are scheduled. After that point, even a slight increase in the number of VoIP phone calls results in system instability. Performance deterioration is not gradual and the packet access delay increases rapidly after the threshold point if there is no control on the accepted traffic. Results shown in Fig. 5 and 6 were obtained using a call admission control (CAC) scheme at call setup (using the simulation model), that computes the available number of OFDMA symbols in each direction (DL and UL). A new call is accepted if there are enough available symbols to allocate SS_{VoIP} [smb/s] in each direction. In general, it was demonstrated that with the use of header suppression, bandwidth efficiency is considerably increased to a large extent, achieving a much higher figure regarding the maximum sustainable number of VoIP streams. We observed that considering HS and cRTP, VoIP streams fit better in a DL sub frame and fewer symbols are wasted. Therefore more VoIP bursts can be

allocated instead of VoIP frame overhead. In addition, by considering cRTP, the RTP, UDP and IP headers can be reduced to only two bytes where no UDP checksums are sent. Moreover, system performance highly depends of the DL-MAP repetition count.

For the performance optimization, we used the default value $RepCnt$=4, however, having $RepCnt$=1, and combined with cRTP, the number of VoIP-G.723 phone calls that the WiMAX mobile system could support, increases up to 1800 for 64-QAM 5/6. Further research will focus on performance analysis of VoIP with mobile SSs considering also silence suppression that reduces VoIP bandwidth by a 60%.

Table 4. Maximum number or VoIP streams

Codec	Mod & cc	DLr	-HS	+HS	+cRTP
G.711	QPSK1/2	4	38/θ3	54/θ4	64/θ4
G.711	QPSK3/4	4	64/θ4	76/θ4	84/θ4
G.711	16-QAM1/2	4	76/θ4	90/θ3	102/θ3
G.711	16-QAM3/4	4	116/θ3	136/θ3	144/θ3
G.711	64-QAM3/4	4	144/θ3	170/θ2	170/θ2
G.711	64-QAM5/6	4	144/θ3	170/θ2	200/θ2
G.723	QPSK1/2	4	228/θ4	348/θ3	432/θ3
G.723	QPSK3/4	4	306/θ3	432/θ3	510/θ2
G.723	16-QAM1/2	4	408/θ3	510/θ2	600/θ2
G.723	16-QAM3/4	4	510/θ2	600/θ2	738/θ1
G.723	64-QAM3/4	4	600/θ2	738/θ1	738/θ1
G.723	64-QAM5/6	4	600/θ2	738/θ1	924/θ0
G.726	QPSK1/2	4	64/θ4	84/θ4	90/θ3
G.726	QPSK3/4	4	84/θ4	116/θ3	136/θ3
G.726	16-QAM1/2	4	102/θ3	144/θ3	144/θ3
G.726	16-QAM3/4	4	144/θ3	170/θ2	200/θ2
G.726	64-QAM3/4	4	170/θ2	200/θ2	246/θ1
G.726	64-QAM5/6	4	200/θ2	246/θ1	246/θ1
G.728	QPSK1/2	4	22/θ3	36/θ3	50/θ2
G.728	QPSK3/4	4	34/θ3	50/θ2	61/θ1
G.728	16-QAM1/2	4	36/θ3	50/θ2	61/θ1
G.728	16-QAM3/4	4	50/θ2	61/θ1	77/θ0
G.728	64-QAM3/4	4	61/θ1	77/θ0	77/θ0
G.728	64-QAM5/6	4	61/θ1	77/θ0	77/θ0
G.729	QPSK1/2	4	84/θ4	144/θ3	170/θ2
G.729	QPSK3/4	4	116/θ3	170/θ2	200/θ2
G.729	16-QAM1/2	4	144/θ3	200/θ2	246/θ1
G.729	16-QAM3/4	4	170/θ2	246/θ1	246/θ1
G.729	64-QAM3/4	4	200/θ2	246/θ1	308/θ0
G.729	64-QAM5/6	4	246/θ1	308/θ0	308/θ0
G.711	QPSK1/2	1	44/θ4	56/θ4	72/θ5
G.711	QPSK3/4	1	72/θ5	90/θ5	108/θ5
G.711	16-QAM1/2	1	90/θ5	108/θ5	120/θ4
G.711	16-QAM3/4	1	136/θ4	160/θ4	192/θ4
G.711	64-QAM3/4	1	192/θ4	240/θ4	240/θ4
G.711	64-QAM5/6	1	192/θ4	240/θ4	280/θ4
G.723	QPSK1/2	1	270/θ5	408/θ4	576/θ4
G.723	QPSK3/4	1	360/θ4	576/θ4	720/θ4
G.723	16-QAM1/2	1	480/θ4	720/θ4	840/θ4
G.723	16-QAM3/4	1	720/θ4	840/θ4	1236/θ3
G.723	64-QAM3/4	1	840/θ4	1236/θ3	1236/θ3
G.723	64-QAM5/6	1	840/θ4	1236/θ3	1800/θ2
G.726	QPSK1/2	1	72/θ5	108/θ5	108/θ5
G.726	QPSK3/4	1	108/θ5	136/θ4	160/θ4
G.726	16-QAM1/2	1	120/θ4	192/θ4	192/θ4
G.726	16-QAM3/4	1	192/θ4	240/θ4	280/θ4
G.726	64-QAM3/4	1	240/θ4	280/θ4	412/θ3
G.726	64-QAM5/6	1	280/θ4	412/θ3	412/θ3
G.728	QPSK1/2	1	27/θ5	48/θ4	70/θ4
G.728	QPSK3/4	1	40/θ4	70/θ4	103/θ3
G.728	16-QAM1/2	1	48/θ4	70/θ4	103/θ3
G.728	16-QAM3/4	1	70/θ4	103/θ3	150/θ2
G.728	64-QAM3/4	1	103/θ3	150/θ2	150/θ2
G.728	64-QAM5/6	1	103/θ3	150/θ2	150/θ2
G.729	QPSK1/2	1	108/θ5	192/θ4	240/θ4
G.729	QPSK3/4	1	136/θ4	240/θ4	280/θ4
G.729	16-QAM1/2	1	192/θ4	280/θ4	412/θ3
G.729	16-QAM3/4	1	240/θ4	412/θ3	412/θ3
G.729	64-QAM3/4	1	280/θ4	412/θ3	600/θ2
G.729	64-QAM5/6	1	412/θ3	600/θ2	600/θ2

This work was supported by DGAPA, National Autonomous University of Mexico (UNAM) under Grant IN IN108910 and IN106609. By research funds from CONACyT grants 105279 and 105117.

REFERENCES

[1] IEEE 802.16e-2005, IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed BWA Systems, Amendment for PHY and MAC for Combined Fixed and Mobile Operation in Licensed Bands, *IEEE*, Dec. 2005.

[2] JaeWoo S., A Downlink Performance Analysis of VoIP Services over an IEEE 802.16e OFDMA System, *IEEE Communications Letters*, Feb. 2007.

[3] JaeWoo S., Performance Analysis of a Semi-Fixed Mapping Scheme for VoIP Services in Wireless OFDMA Systems, *Fifth International Conference on Wireless Mobile Communications*, Oct. 2009.

[4] Jae-Woo S., Performance Analysis of VoIP Services in the IEEE 802.16e OFDMA System with Inband Signaling, *IEEE Transactions on Vehicular Technology*, May 2008.

[5] Kaarthick B., Yeshwenth V. J., Nagarajan N., Rajeev, Investigating the performance of various vocoders for fair scheduling algorithm in WiMAX, *First Asian Himalayas International Conference*, Nov. 2009.

[6] Howon L., Hyu-Dae K., Dong-Ho C., Extended-rtPS$^+$ Considering Characteristics of VoIP Codecs in Mobile WiMAX, *IEEE International Symposium on PIMRC*, Dec. 2008.

[7] Ortiz L., Rangel V., Gomez J., R. A. Santos R.A., Lopez-Guerrero M., Performance optimization of mobile WiMAX netwoks for VoIP stream, *Proceedings of KTTO´2*011, Bielsko-Biala, Poland, 22-24 June 2011.

[8] Rangel V. Performance Evaluation and Optimisation of the DVB/DAVIC Cable Modem Protocol (Phd. thesis), University of Sheffield, Jun. 2002.

[9] ETSI ES 200 800 v.1.3.1. Digital Video Broadcasting: Interaction Channel for Cable TV Distribution Systems (CATV), *ETSI*, Oct. 2001.

[10] Casner S., Jacobson V., RFC2508 - Compressing IP/UDP/RTP Headers for Low-Speed Serial L, *Cisco Systems*, Feb. 1999.

[11] ITU-T Rec. G.711, Pulse Code Modulation (PCM) of Voice Frequencies, *ITU-T*, Nov. 1988.

[12] ITU-T Rec. H.323, Packet-Based Multimedia Communications Systems, *ITU-T*, Dec. 2009.

[13] ITU-T Rec. G.723.1, Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, *ITU-T*, May. 2006.

[14] ITU-T Rec. G.726, 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM), *ITU-T*, Dec. 1990.

[15] ITU-T Rec. G.728, Coding of Spech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction, *ITU-T*, Sep. 1992.

[16] ITU-T Rec. G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP), *ITU-T*, Jan. 2007.

[17] PacketCable, PacketCable™ Audio/Video Codecs Specification, *CableLabs*, PacketCable Project, Dec. 2001.

***Authors**: MSc. Luis Ortiz, Víctor Rangel, Ph.D., Javier Gómez, Ph.D., National Autonomous University of Mexico, Department of Telecommunications, Av. Universidad s/n Colonia Chamilpa, 143 87 Mexico City, Mexico, E-mail: lortiz@fi-b.unam.mx, victor@fi-b.unam.mx, javierg@fi-b.unam.mx; Raúl Aquino, Ph.D., University of Colima, School of Telematics, Avenida Universidad 333, 280 40 Colima, Mexico, E-mail: aquinor@ucol.mx; Miguel López-Guerrero, Ph.D., Metropolitan Autonomous University, Department of Electrical Engineering, Prolongación Canal de Miramontes 3855, 143 87 Mexico City, Mexico, E-mail: milo@xanum.uam.mx.*