AMC-aware QoS proposal for OFDMA-based IEEE802.16 WiMAX systems

Chadi Tarhini, Tijani Chahed
GET-Institut National des Télécommunications/UMR CNRS 5157
9 rue C. Fourier - 91011 Evry CEDEX - France
{chadi.tarhini, tijani.chahed}@int-evry.fr

Abstract—OFDMA-based IEEE802.16 implements Adaptive Modulation and Coding (AMC) which results in a different bit rate for each user depending on its position in the cell as well as the radio condition it experiences; users away from the base station experience lower throughput. We hence propose in this paper a new AMC solution that compensates the degradation in modulation by a higher number of sub-carriers so as to maintain the bit rate of streaming flows at a constant level throughout the whole coverage area. This proposal is modeled analytically in a dynamic user configuration, where users of both types, streaming and elastic, remain in the system for a finite duration. And this in multi-cell environment and for different frequency reuse schemes.

Index Terms—QoS, capacity, OFDMA, IEEE802.16, WiMAX, AMC, Reuse partitioning

I. INTRODUCTION

WiMAX systems are based on versions d and e of the IEEE802.16 standard which defines a physical (PHY) and medium access control (MAC) layers for broadband wireless access systems operating at frequencies below 11 GHz. The first of these standards, published in 2004, addresses fixed services while the second, published in 2005, is intended for mobile services.

The IEEE 802.16e specifications define three different PHY layers: single-carrier (SC) transmission, Orthogonal Frequency-Division Multiplexing (OFDM), and OFDM Multiple Access (OFDMA). The multiple access technique used in the first two of these PHY specifications is pure TDMA, while the third technique uses both the time and frequency dimensions for resource allocation. From these three PHY technologies, OFDMA has been selected by WiMAX Forum [1] as the basic technology for portable and mobile applications.

In wireless communication systems, random fluctuations prevent the continuous use of highly bandwidth-efficient modulation and therefore Adaptive Modulation and Coding (AMC) has become a standard approach in recently developed wireless standards, including WiMAX. The idea behind AMC is to dynamically adapt the modulation and coding scheme to the channel conditions so as to achieve the highest spectral efficiency at all times [2]. Adaptive modulation changes the coding scheme and/or modulation method depending on channel state information - choosing it in such a way that it squeezes the most out of what the channel can transmit. In OFDMA, modulation and/or coding can be chosen differently for each sub-carrier, and it can also change with time. Indeed, in the IEEE802.16 standard, coherent modulation schemes are used starting from low efficiency modulations (BPSK with coding rate 1/2) to very high efficiency ones (64-QAM with coding rate 3/4) depending on the Signal-to-Noise Ratio (SNR). This results in lower throughput for users further away from the base station and/or with worse radio conditions [10]. This may be acceptable for elastic traffic, such as data, but not for streaming, such as voice. For IEEE802.16 to be a QoS-capable multimedia system, it should guarantee streaming applications a constant bit rate independent of the user position in the cell as well as radio conditions. To do so, we propose in this work a new AMC mechanism wherein the degradation in modulation for streaming flows is countered by use of a higher number of sub-carriers.

Moreover, in a multi-cell OFDMA system, adjacent cells using sub-carriers of exactly the same frequency and time can cause interference to one another unless inter-cell interference mitigation techniques are applied. This interference takes the form of collisions, the number of which increases as mobile stations get closer to the edge of the cell. To combat this, a reuse partitioning [3] scheme can be applied and controls the reuse factor in different parts of the cell.

II. MODELING PHYSICAL LAYER

A. OFDMA sub-carrier allocations

OFDMA is a multiple access technique which divides the total Fast Fourier Transform (FFT) space into a number of sub-channels (set of sub-carriers that are assigned for data exchange) whereas the time resource is divided into time slots and a frame is constructed by a number of slots. Let N denote the size of FFT and N the total number of data sub-carriers after reserving the pilot and guard sub-carriers which we divide into L groups, each with $K = \frac{N}{L}$ data sub-carriers.

In OFDMA-based WiMAX system, resource allocation is done in the time-frequency domain: a call may share a sub-channel with other users. The user device could choose sub-channels based on geographical location with the potential of eliminating the impact of deep fades.

B. AMC and cell decomposition

In this work, and without loss of generality, we study AMC in the presence of path loss only which we characterize by a certain value $\xi_i$. High efficiency modulation is used for users where $\xi_i \ll \xi$, corresponding to a large SNR. This results
in the division of the cell into r regions, $i = 1...r$, which we may consider to be concentric circles of radius $R_i$ for simplicity, but might be of different topology if we take into account other phenomena, such as fast-fading. In each region, users have the same modulation scheme and experience thus a corresponding bit rate which decreases as users get further from the base station.

To calculate the area covered by each modulation scheme, we must know the maximal distance between Base Station (BS) and users using a corresponding modulation. This distance is determined using the maximal SNR a user should receive without data loss. In [5], switching points between modulation/coding schemes are proposed depending on received SNR and allow us to calculate the maximal distance a user should have. The path loss for the free space model is given by [8]:

$$PL[dB] = -10\log[G_E G_R (\frac{\lambda}{4\pi R})^2]$$
$$= -10\log[G_E] - 10\log[G_R] + 20\log(\frac{\lambda}{4\pi})$$

(1)

where $G_E$ is the emitter antenna gain, $G_R$ is the receiver antenna gain, $R$ is the distance between the emitter and the receiver and $\lambda$ is the wavelength. This path loss is also equal to:

$$PL[dB] = P_E[dBm] - SNR[dB] - N[dBm]$$

(2)

where $P_E$ is the emitted power and $N$ is the thermal noise which is equal to:

$$N[dBm] = 10\log(\tau T W)$$

(3)

$\tau = 1.38 \cdot 10^{-23} \text{watt/K} - H$ is the Boltzmann constant, $T$ is the temperature in Kelvin ($T = 290$) and $W$ is the transmission bandwidth in Hz.

Using the above equations, we can calculate the relationship between the distance and the SNR as follows:

$$R_i = \frac{\lambda \cdot 10^{\frac{P_E[dBm] + 10\log[G_E]}{10} + 10\log[G_R] + SNR[dB] - N[dBm]}}{4\pi}$$

(4)

The area of each region $S_i$ is given by:

$$S_i = \pi \cdot (R_i^2 - R_{i-1}^2)$$

where $R_0 = 0$.

For the sake of illustration, let us consider the following example based on the licensed band for WiMAX to outdoor use in France which starts at a frequency of 3.4GHz and which has system bandwidth equal to 20MHz. At this bandwidth, the thermal noise is equal to $-100.97dBm$. According to the maximum allowed Effective Isotropic Radiated Power (EIRP) of 1W, where the emitters are assumed to have an emission power $P_E$ of 1W for users. We consider without loss of generality the case of antennas in BS and user equipment without gain, i.e., $G_E = G_R = 1$. In Figure 1, we represent the distance assigned to SNR for switching points. The proportion of each surface area per PHY assumption is determined and shown in Table I.

![Fig. 1. Received SNR function of the distance](image)

### Table I

<table>
<thead>
<tr>
<th>Modulation</th>
<th>Coding rate</th>
<th>Receiver SNR(dB)</th>
<th>Surface [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK</td>
<td>1/2</td>
<td>6.4</td>
<td>39.4</td>
</tr>
<tr>
<td>QPSK</td>
<td>1/2</td>
<td>9.4</td>
<td>20.75</td>
</tr>
<tr>
<td>16 QAM</td>
<td>1/2</td>
<td>16.4</td>
<td>7.57</td>
</tr>
<tr>
<td>64 QAM</td>
<td>2/3</td>
<td>22.7</td>
<td>3.91</td>
</tr>
<tr>
<td></td>
<td>3/4</td>
<td>24.4</td>
<td>1.74</td>
</tr>
</tbody>
</table>

C. Throughput of the cell

The instantaneous physical bit rate $R_i^{s,e}$ of streaming or elastic users in region $S_i$ is given by:

$$R_i^{s,e} = \frac{L_i^{s,e} \times K \times C \times \log_2(M)}{T_s \times S_i} \times (1 - BLER)$$

$$= L_i^{s,e} \times K \times B \times E_s \times (1 - BLER)$$

(5)

where $L_i^{s,e}$ is the number of sub-channels to be assigned to streaming/elastic users in region $S_i$, $K$ is the number of data sub-carriers assigned to each sub-channel, $C$ is the coding rate of the $M$-ary modulation, $T_s$ is the OFDMA symbol duration given by:

$$T_s = T_b + T_g$$

with $T_b$ the useful symbol period (in units of microseconds) given by $\frac{N}{W \times T_0}$ and $T_g$ the guard period equal to $G \times T_b$. $W$ is the bandwidth (MHz), $n$ is the sampling factor, $G$ is the ratio of cyclic prefix (CP) to useful time, $S_i$ is the sector coefficient, $B$ is the baud rate (symbols/sec), $E_s$ is the efficiency of the modulation (bits/symbol) in each region $S_i$ and BLER is the perceived Block Error Rate$^1$.

Although streaming flows will have a different number of sub-channels $L_i^{s,e}$ per region, this amounts to one streaming class only with rate $R^s$. As of (TCP-based) elastic calls, since they tolerate reduction in their throughput, we shall not do anything: they will simply share the left over capacity. On

---

$^1$Note that for each SINR value, we can determine a couple of values $(E, BLER)$ and these values are determined by link level curves $E = f(SINR)$ and $BLER = g(SINR)$
the basis of Processor Sharing (PS) [9]. In doing so, we are actually giving priority to streaming calls over data ones. The former are subject to admission control as we cannot accept more users than the (dedicated) resources, sub-carriers in this case, assigned to them.

The number of sub-channels \( L^e_i \) allocated to an elastic call in region \( S_i \) is thus given by:

\[
L^e_i = \left\lfloor \frac{L - \sum_{i=1}^{r} L_i n^e_i}{n^e_i} \right\rfloor
\]  

(6)

\( \lfloor x \rfloor \) indicates the largest integer that is less than or equal to \( x \). \( n^e_i \) and \( n^s_i \) are the number of streaming and elastic flows in region \( S_i \).

In total, we obtain one (high priority) class for streaming applications with constant bit rate \( R^s \) and \( r \) best-effort classes for elastic ones corresponding to the \( r \) regions of our system.

III. MODELING AT THE PHYSICAL LAYER

We assume that streaming calls arrive to region \( S_i \) according to a Poisson process with intensity \( \lambda^s_i \) and use \( l^s_i \) sub-carriers for an exponentially distributed time with mean \( 1/\mu^s \) independent of the share of the resources they get. We also assume that elastic flows arrive to the system according to a Poisson process with intensity \( \lambda^e_i \) and assume for tractability that their service rate is exponentially distributed with mean \( \mu^e_i = \frac{R^e_i}{E[Z]} \) where \( E[Z] \) is the mean file size \( \sum \).

We now model our QoS-capable system as a Continuous Time Markov Chain (CTMC) by taking into account the proposed priorities for the integration of streaming and elastic flows as well as the way they share resources.

The state is characterized by the following row vector:

\[
\begin{pmatrix}
    n^s_1, n^s_2, \ldots, n^s_r, n^e_1, n^e_2, \ldots, n^e_r
\end{pmatrix}
\]

where \( n^s_i \) and \( n^e_i \) for \( i = 1 \ldots r \), represent the number of streaming and elastic calls in region \( S_i \), respectively.

The state space of the system is given by:

\[
\mathcal{S} := \left\{ \begin{pmatrix} n^s & n^e \end{pmatrix} \in \mathbb{N}^{2r} \mid \sum_{i=1}^{r} \left( l^s_i n^s_i + l^e_i n^e_i \right) \leq N \right\}
\]

(7)

where \( l^s_i \) and \( l^e_i \) denote the number of sub-carriers allocated to streaming and elastic calls in region \( S_i \) respectively and \( N \) is the maximum number of sub-carriers in the cell.

A. Analysis

We now determine the steady-state probability vector \( \bar{b} \pi = \{ \pi(\overline{w}) \}_{\overline{w} \in \mathcal{S}} \). Note that the corresponding system is non-homogeneous as the departure rate of elastic calls depends on the overall number of calls in the system whereas streaming calls do not.

The solution of the steady-state distribution is obtained by solving the set of linearly independent equations given by:

\[
\begin{align*}
    \pi \cdot Q &= 0 \\
    \sum_{\overline{w} \in \mathcal{S}} \pi(\overline{w}) &= 1
\end{align*}
\]

To construct the transition matrix \( Q \), we must consider all possible transitions between neighboring states. Let \( q(\overline{w} \rightarrow \overline{w}'') \) denote the transition probability from state \( \overline{w} \) to neighboring states \( \overline{w}'' \). Note that when we accept a new call in region \( S_i \), \( 1 \leq i \leq r \) the state is noted by \( \overline{n}^{s,e}_i \) and when a call terminates the service the next state is \( \overline{n}^{s,e}_i \).

We thus have the following transition rates:

\[
\begin{align*}
    q(\overline{w} \rightarrow \overline{n}^{s,e}_i) &= \lambda^s_i \\
    q(\overline{w} \rightarrow \overline{n}^{s,e}_i) &= n^e_i \mu^e_i \\
    q(\overline{w} \rightarrow \overline{n}^{s,e}_i) &= \lambda^e_i \\
    q(\overline{w} \rightarrow \overline{n}^{s,e}_i) &= n^e_i \mu^e_i
\end{align*}
\]

(9)

and the values \( q(\overline{w} \rightarrow \overline{w}'') \) must be obtained as the sum of all terms in each line in matrix \( Q \) is equal to zero for \( 1 \leq i \leq r \).

B. Performance measures

Based on the steady-state probabilities of our CTMC, we now determine the performance measures relative to our model.

The call blocking probabilities of both types of flows, \( B_i^{s,e} \) and \( B_i^{s,e} \) in region \( S_i \), are obtained by summing up the steady state probability of saturation states:

\[
B_i^{s,e} := \sum_{\overline{w} \in \mathcal{S}} \pi(\overline{w})
\]

(10)

where \( \mathcal{S}_i \) is the subset of states in \( \mathcal{S} \) for which any new call of class \( i \) is blocked when arriving to the system due to lack of resources. Formally,

\[
\mathcal{S}_i := \left\{ \overline{w} \in \mathcal{S} \mid \sum_{k=1}^{r} \left( l^s_k n^s_k + l^e_k n^e_k \right) > N \right\}
\]

The mean transfer time of class \( i \) elastic flows can be calculated from the Little’s formula by:

\[
\bar{T}_i = \frac{n^e_i}{\lambda^e_i (1 - B_i^{s,e})}
\]

(11)

where \( \bar{n}^e_i \) is the mean number of elastic flows in region \( S_i \).

IV. REUSE PARTITIONING IN OFDMA WIMAX

In wireless communications systems, users who are close to the base station (BS) experience a higher signal quality compared to those who are located near the cell edge. Users with high signal quality can tolerate a higher level of interference, which makes a denser reuse of channels possible. Reuse partitioning has been proposed in [6] as a means to increase the total system capacity.

A. Basic concept

Under the reuse partitioning allocation policy, the coverage area of a single cell is divided into several regions, each region having its own reuse factor. Considering two regions only, this amounts to the fact that users in the inner region have access to the total frequency band \( L \) whereas those in the outer region are restricted to only \( L' \) sub-channels in the target cell and \( L'' \) sub-channels in the neighboring cell, with \( L' < L \) and \( L'' < L \) the least overlapping. In doing so, users at the edge shall experience a lower interference figure in terms of collisions, as quantified next.
B. Collisions

There are several approaches as how to distribute available sub-carriers to sub-channels. The simplest one consists of picking randomly sub-carriers from available ones such that any available sub-carrier has the same probability to get allocated to an arriving user. We refer to this method as random allocation of sub-carriers which does not require any coordination between cells. In the case of two cells, the following lemma establishes the mean number of collisions when the number of occupied sub-carriers is \((L \times K_1)\) and \((L \times K_2)\) in cell 1 and cell 2 respectively.

**Lemma 1:** The mean number of collisions in \(L\) sub-channels is given by:

\[
E[C|K_1, K_2] = L \cdot E'[C|K_1, K_2] \tag{12}
\]

where \(E'[C|K_1, K_2]\) is the mean number of collisions in one group of sub-carriers and is in turn given by:

\[
E'[C|K_1, K_2] = \sum_{c=c_{\text{min}}}^{c_{\text{max}}} c \cdot Pr(c|K_1, K_2) \tag{13}
\]

\(Pr(c|K_1, K_2)\) being the probability of having \(c\) collisions in this group and is equal to:

\[
\frac{\sum_{n_1c=0}^{K_1-c} \left( \begin{array}{c} N/L \cdot c \\ c \\ n_1c \end{array} \right) \left( \begin{array}{c} N/L-c-n_1c \\ K_2-c \\ n_1c \end{array} \right)}{\sum_{c=c_{\text{min}}}^{c_{\text{max}}} \sum_{n_1c=0}^{K_1-c} \left( \begin{array}{c} N/L \\ c \\ n_1c \end{array} \right) \left( \begin{array}{c} N/L-c-n_1c \\ K_2-c \\ n_1c \end{array} \right)} \tag{14}
\]

\(c_{\text{min}}\) is the number of colliding sub-carriers. It belongs to \(\{c_{\text{min}} : c_{\text{max}}\}\) with:

\[
c_{\text{min}} = \max(0, K_1 + K_2 - N/L)
\]

and

\[
c_{\text{max}} = \min(K_1, K_2, N/L)
\]

and \(n_1c\) is the number of non-colliding sub-carriers in cell 1; \(n_1c \in \{0 : K_1-c\}\).

The proof is contained in Reference [10].

**Lemma 2:** The mean number of collisions in the system is equal to \([11]\):

\[
E[C] = \sum_{K_1, K_2} \left( \prod_{i=1}^{2} \Pi_i(K_i) \right) E[C|K_1, K_2] \tag{15}
\]

where \(E[C|K_1, K_2]\) is given from Equation \((12)\) and \(\Pi_i(K_i) = \sum_{\vec{n}_1 \in \Theta} \pi(\vec{n}_1) Pr(K_i|\vec{n}_1)\) is the probability of having \(LK_i\) sub-carriers in cell \(i\).

V. NUMERICAL RESULTS

We now present some numerical results considering an OFDMA system with an FFT size of 2048 sub-carriers and 2 regions with AMC respectively 16-QAM 3/4 \((E_1 = 3\) bit/symbol) and QPSK 1/2 \((E_2 = 1\) bit/symbol). These efficiency parameters correspond to the transmission modes with convolutionally-coded modulation. Moreover, let \(L = 6\) sub-channels; \(l_1\) = 1 sub-channel; \(l_2\) = 3 sub-channels; \(\lambda_l = 0.01\) users/sec; \(\lambda_r = 0.01\) users/sec; \(\rho^* = 0.02\); \(B_L E_R = 0\); mean file size \(E[Z] = 500Kb\). These values are deliberately small to ease numerical burden; a scaling up is straight-forward.

Based on Equation \((5)\) and using a baud value \(B = 2666\) symbol/sec and \(K = 48\), \(R_1 = R_2 = 128Kbps\).

Figure 2 shows the blocking probability for streaming flows with and without QoS as a function of an increasing arrival rate of streaming flows.

We observe that without QoS, the blocking rate in the inner region is the same as in the outer region. This is due to the fact that in this case, streaming flows are assigned the same number of sub-carriers in both regions. With our QoS proposal however, the blocking rate increases in both region and is higher in the outer one, as distant users are given more resources. In terms of throughput, recall that our strategy is completely seamless for both types of users which get the same bit rate.

The distinction between inner and outer region is different for data users who indeed belong to two different classes in terms of throughput, one per each region, but have an equal blocking probability, both when QoS is enabled or not. Consequently, the mean transfer times are different in each region, as shown in Figure 3, and get larger when QoS is enabled.

We next turn to the case where reuse partitioning is enabled \((L' = 3\) sub-channels).

Figure 4 shows the blocking probability for streaming flows for the two regions, inner and outer, when both QoS and reuse partitioning are enabled.

We observe that with respect to Figure 2, the blocking probability remains the same in the inner region but gets (unacceptably) higher in the outer one.

Figure 5 shows the mean transfer time in the two regions when again QoS and reuse partitioning are both enabled.

We observe that with respect to Figure 3, both values get higher, with a bias in the outer region.

These increases in the blocking probabilities and mean transfer times are nevertheless counter-balanced by an increase in the total cell throughput, as show in Figure 6 for the case of two cells. This means that the lower number of sub-carriers in the edge of the cells is compensated for by a lower number...
of collisions on those sub-carriers. And hence it is worth the implementation.

VI. CONCLUSION

We proposed in this paper a new QoS paradigm that makes OFDMA-based IEEE802.16 capable of offering streaming applications a sustainable rate throughout the whole coverage area while offering the leftover capacity to elastic ones. We specifically proposed to compensate for the degraded performance of streaming flows that are away from the base station and/or experiencing bad radio conditions, due to path loss and fading, by assigning them a higher number of sub-carriers.

Using reuse partitioning improves the overall system throughput albeit higher individual blocking rates and transfer times.

In our next step, we shall consider the integration of a mobility pattern in WiMAX, in the framework of IEEE802.16e. In this case, our present QoS proposal is even more essential as mobile users need to be guaranteed a sustainable performance as they move around the system.

REFERENCES