IEEE802.16 Multi-class Capacity including AMC scheme and QoS Differentiation for Initial and Bandwidth request ranging

T. Peyre
LIA/CERI, University of Avignon
339, chemin des Meinajaries
Agroparc
BP 1228 84911 AVIGNON
Cedex 9-FRANCE
thierry.peyre@univ-avignon.fr

K. Ibrahimi
LIA/CERI, University of Avignon
339, chemin des Meinajaries
Agroparc
BP 1228 84911 AVIGNON
Cedex 9-FRANCE
LIMIARF, Mohammed V Agdal
Rabat University
4 Av. Ibn Battouta B.P. 1014
MAROC
khalil.ibrahimi@univ-avignon.fr

R. El-Azouzi
LIA/CERI, University of Avignon
339, chemin des Meinajaries
Agroparc
BP 1228 84911 AVIGNON
Cedex 9-FRANCE
rachid.elazouzi@univ-avignon.fr

ABSTRACT

In this paper we study the capacity of the OFDMA-based IEEE802.16 WiMAX network in the presence of two types of traffic, streaming (Real-Time) and elastic (Non-Real-Time) including Adaptive Modulation and Coding (AMC). Many studies in the literature assumed that packets or calls arrive to the system according to poisson process for the sake of analytical simplicity. However, it has been recently proved that the exponential distribution is inappropriate [2]. Based on the generalized traffic processes developed [2], we study the media access control (MAC) layer of WiMAX and develop a resource allocation that maintain the bit rate of real time connections independently of the user position in the cell. Using Markovian analysis we evaluate the impact of our resource allocation on the non-real time connection (NRT) as expected delay and throughput.

Keywords

IEEE802.16e, QoS, wireless communication, Discrete time markov chain

1. INTRODUCTION

Both researchers and industrial actors agree on the massive potential of IEEE802.16 networks as a major communication technology that offers wireless broadband access. Its ability to manage a large spectrum of QoS requirements enables it to carry all metropolitan communication services. As of performance, the IEEE society claims an average throughput up to 30 Mbps on a 15 km coverage radius and a mobility speed up to 80 miles/h. With all these capabilities, it is expected to offer what 3G, DSL, and HFC cable technologies can collectively offer. In addition, the standard allows to link the others commonly used wireless technologies such as 3G, IEEE802.11 and HSDPA. Interested readers are kindly referred to Reference [12] which provides a comprehensive analysis of the standard strengths and weaknesses as compared to other wireless technologies. IEEE802.16e defines five QoS classes [6]: i. Unsolicited Grant Service (UGS) for constant-bit-rate traffic, delay-and-jitter-sensitive applications such as Voice over IP, ii. real-time Polling Service (rtPS), also specified for streaming applications but with higher priority on all other classes, iii. Extended rtPS (ErtPS) adds a bound on the jitter, iv. non-rtPS (nrtPS) for elastic applications and v. the traditional Best-Effort (BE).

In this paper, we consider a single IEEE802.16 cell partitioned in r regions. Each region uses a different modulation and coding technique as described in [5]. We consider two classes of traffic: real-time (RT), corresponding to UGS or rtPS classes, and non-real-time (NRT), corresponding to nrtPS and BE. In addition, we consider the case when the codes dedicated to the ranging requests are distributed between the classes [2].

Many studies in the literature assumed that packets or calls arrive to the system according to poisson process for the sake of analytical simplicity. Moreover, its assume that the arrival processes of all types of connections are independents. In WiMAX, since all contending mobiles share a finite number of CDMA codes, these arrival processes should be dependents (see [3] and [2]). In [2], we developed a MAC access model of ranging request based on different class priorities, using the backoff parameters differentiation and the codes available for these classes. We considered the case of code-based classification which uses a ranging code partitioning between the different classes with a dedicated code subrange for real-time traffic and an other code subrange shared by both classes. This previous study allows us to characterize the arrival process of ranging request for both

ValueTools 2008, October 21-23, 2008, Athens, GREECE. Copyright ©2008 ICST ISBN 978-963-9799-31-8. .

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

real time and non-real connections.

Our aim in this work is to propose a new capacity model which integrates the new multi-class IEEE802.16e MAC access proposed in [2]. Based on the generalized traffic process of arrival ranging request, we develop a new resource allocation algorithm for real-time connections based on the user position in the cell as well as its radio condition as given by the signal-to-Noise Ratio (SNR). In this scheme, all RT connections in the system have the same rate per bit and, the duration of RT call does not depend on the number of connections in the system. Whereas, the duration of a NRT call depends on the dynamic assignment and the user position in the cell. Using Discrete Time Markov Chain (DTMC) introducing the Multi-class prioritization scheme and the classified batch arrival process defined in [2], we propose a probabilistic model that takes features of interest.

In the literature, the performance evaluation of the MAC access of IEEE802.16 networks has been mostly done via simulation; not much analytical work has been produced. The capacity of the OFDMA-CDMA ranging subsystem in IEEE802.16 has been studied in few papers. In [10], the authors analyzed the performance of random access protocols which use ranging subchannel in OFDMA-CDMA environment, in terms of mean delay time (MDT) and first exit time (FET). In [7], the idea is to control adaptively the size of each ranging code for IR, PR, and BR ranging in order to implement efficient random access. In [11], the authors evaluate the capacity of a ranging subchannel in terms of the ranging code error probability versus the number of active users who attempt ranging. Recently, several works addressing QoS in general and call admission control (CAC) in particular have been produced. For instance, in [8], an admission control scheme is proposed. It ensures highest priority to UGS flows while maximizing overall bandwidth by means of bandwidth borrowing. In [9], QoS is treated on the basis of classical intserv and diffserv paradigms as well as their mapping to IEEE802.16 MAC layer.

The remainder of this paper is outlined as follows. In section 2, we review the mechanisms for supporting QoS at the IEEE 802.16 MAC layer. In section 3, we develop a markov based analytical model according with our CAC algorithm and arrival process of calls developed in [14]. We evaluate the performance of our new scheme in Section 5. We conclude in Section 6.

2. MAC OVERVIEW IN IEEE802.16E

The MAC layer in IEEE802.16e is based on the concept of scalable Orthogonal Frequency Division Multiple Access (OFDMA), itself based on OFDM [6]; the scalability is realized by adjusting the FFT size which translates into 10.94 KHz spacing of sub-carrier frequency. As for the physical layer, it implements, in addition to Adaptive Modulation and Coding (AMC), an enhancement of Adaptive Antenna System (AAS), Hybrid Automatic Repeat reQuest (HARQ) and CQICH, fast channel feedback that informs the base station about the state of the propagation channel. Eventually, the physical layer specifies a ranging subchannel and a set of pseudo-noise codes which adds a CDMA feature to OFDMA [11].

2.1 Classes of service

IEEE802.16e defines five QoS classes [6]: i. Unsolicited Grant Service (UGS) for constant-bit-rate traffic, delay-and-jitter-sensitive applications such as Voice over IP, ii. real-time Polling Service (rtPS), also specified for streaming applications but with higher priority on all other classes, iii. Extended rtPS (ErtPS) adds a bound on the jitter, iv. non-rtPS (nrtPS) for elastic applications and v. the traditional Best-Effort (BE). The implementation of these QoS classes takes place at the MAC layer via a classifier and a scheduler. It operates at the flow level, defined by a service flow ID, a connection ID pair, uplink or downlink direction and a set of QoS metrics.

- Unsolicited Grant Service (UGS). In this case, the mobile sends through the ranging request its needed resources. The request indicates the bandwidth required and its periodicity. If more resources are required later, the request will be transmitted through the data frame suffix.

- Real time Polling Service (rtPS). Here the mobile request indicates the minimum and maximum bandwidth for its service. Moreover, its needs will be periodically updated through an additional allocated channel.

- Non real time Polling Service (nrtPS). The mobile asks for a versatile resource through a single request. This also allows to request for a minimum bandwidth.

- Best Effort (BE). The mobiles perform a ranging requests for each needs, with no warranty of QoS.

2.2 Connection process

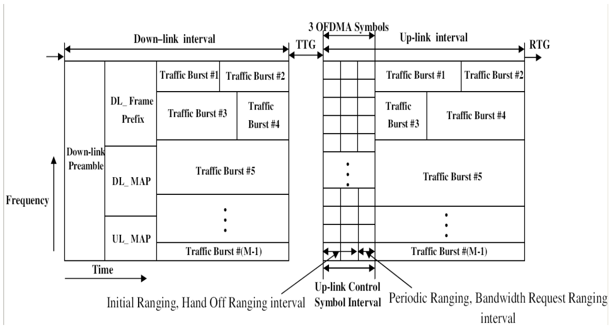

Figure 1: IEEE802.16e MAC Frame

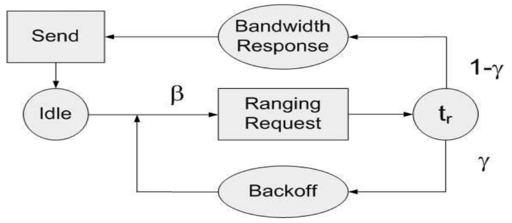

Figure 2: Single node IEEE802.16e backoff process

As explained in [11], the MAC frames are composed of two main TDMA subframes, one for the downlink and one

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

for the uplink (see Figure 1). This ranging interval can manage a large number of contending connections based on the CDMA technique. This makes it possible to share the channel resources between all contending nodes and to minimize the collision probability. In order to ask for a data transmission, a node chooses one of the available CDMA codes and transmits its coded request through the bandwidth request ranging interval. These requests follow a backoff process in case of collision on the selected code. Figure 2 sketches the backoff process in the IEEE802.16e. A collision occurs if two or more nodes choose the same code in the same ranging interval.

Note that the base station has to manage CDMA coding and decoding, resource allocation and flow scheduling from one TDMA frame to the next. And so, the incoming connection request waits for some MAC frames before receiving any response. In fact, the mobile waits for its bandwidth response until a timeout threshold. The IEEE802.16e standard defines a so-called t_r parameter as the maximum number of MAC frames that a contending node can wait before considering that its request has been lost on the wireless channel or in the BS request queue.

2.3 Data transmission and AMC

The IEEE802.16 Physical layer uses an OFDMA sub-carrier allocation policy for the data transmission. The uplink and downlink sub-frames divide the time and frequency space into sub-carriers. As described in [15], the minimum frequency-time unit of sub-channelization is one slot, and a frame is constructed by a number of slots. Different sub-carriers are allocated to a mobile transmission as function of the resource requested by the mobile. Moreover, a sub-channel can be used periodically by different mobiles due to theirs classes of traffic

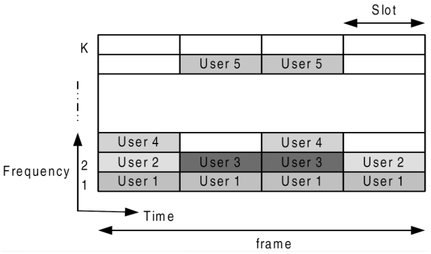

Figure 3: OFDMA sub-carrier allocation

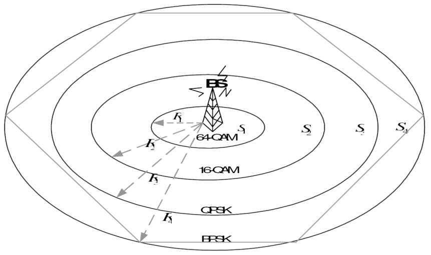

Figure 4: IEEE802.16e AMC regions

As sketched in the figure 3, once a mobile is granted to transmit by a bandwidth response in the DL-MAP, base station assigns one or more subcarrier and hence defines the sub-channel that the mobile will be able to use for its data transmission.

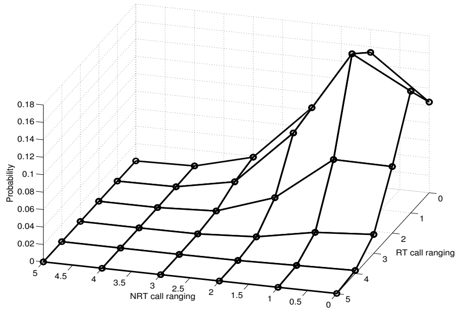

Figure 5: RT and NRT arrivals distribution

Table 1: IEEE802.16e AMC settings

modulation	Coding rate	Receiver SNR (dB)	Surface %
BPSK	1/2	6.4	39.4
QPSK	1/2	9.4	20.75
	3/4	11.2	28.0
16 QAM	1/2	16.4	4.07
	3/4	18.2	5.14
64 QAM	2/3	22.7	0.9
	3/4	24.4	1.74

An IEEE802.16e cell is organized as presented on the figure 4. The table indicates the modulations and codings used in a IEEE802.16e cell as function of the user SNR. The SNR requirement for a BLER less than 10^{-6} depends on the modulation type as specified in the standard [4, Table I]. The number of subcarrier allocated to a mobile directly depends on the available modulation, the type of traffic and the requested bandwidth.

3. SYSTEM MODEL AND RESOURCE ALLOCATION IN WIMAX

3.1 System description

In this paper we consider a single IEEE802.16e cell. Nodes are uniformly distributed on the whole cell and we assume that there is no inter and intra mobility. The cell is decomposed into several regions. The cell population is dispatched between the regions according with the region area coverage. Each region is characterized by the modulation used for data transmission. Due to the AMC scheme described in the previous section, the mobiles use a modulation chosen as function of the receiver SNR.

We assume that the mobiles manage two classes of traffic: Real Time (RT), corresponding to UGS or rtPS classes, and Non Real Time (NRT), corresponding to nrtPS and BE. Note that the IEEE802.16e standard defines a connection based transmission technique. Thus for each new desired transmission the mobile attempts a ranging and, consequently, a single mobile may load the system with several

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

calls.

Concerning the ranging requests, the mobiles attempt by following the connection scheme described in the MAC overview section. But here, we introduce an enhanced ranging scheme originally proposed in [2]: We use a code-based classification where the bandwidth request code range is partitioned between the different classes. From the total code range, we distinguish a dedicated code subrange for real-time traffic and an other code subrange shared by both classes. In [3] concludes that the arrival of RT and NRT connections are dependant and can be approximated by the multiplication of two gaussian distributions. For instance, the figures 5 shows the RT and NRT distribution obtained for 50 users, 4 dedicated codes for RT traffic and 4 others codes shared by RT and NRT traffic. The other communication parameters follow the IEEE802.16e standard [5].

3.2 Connection Admission Control

In this subsection, we define the CAC algorithm used in our IEEE802.16e system. First, we consider a discrete time model wherein a MAC frame consists in a system slot.

The RT traffic are characterized by the same bit rate. Thus, they receive a number of sub-carriers as function of their modulation efficiency (bit per symbol) and hence, as function of their respective region. Note that the system accepts RT calls until its capacity overflows. In addition, our system has the particularity to receive several bandwidth requests during a single slot. In the load traffic, the system can not afford to accept all these requests. Hence, the CAC algorithm should accept some of these calls by prioritizing the ones from the inner region. These calls are those who use the fewer sub-carriers. We assume that the duration of a RT call is not affected by the allocated bandwidth.

Conversely, the NRT traffic have no bandwidth requirement. All the NRT traffic received the same number of sub-carriers. The downlink interval of the IEEE802.16e MAC frame (see figure 1) gathers a large number of sub-carriers. Since the NRT calls tolerate throughput reduction, they will use the sub-carriers left by the RT traffic on the basis of Processor Sharing (PS) [13]. Thus, the bit rate of a NRT call depends on its region (i.e modulation). Note that our CAC algorithm defines a minimum number of sub-carriers which are dedicated to the NRT traffic. Here we want to observe the impact of this singular algorithm property on the NRT traffic throughput, NRT expected delay as well as on the RT blocking probability. Note finally that the NRT traffic remains in the system during a number of slots as function of the consumed resource: the more sub-carriers the NRT call have, the faster the call leave the system.

3.3 Cell decomposition and throughput

We assume that the n mobiles are homogenously distributed. Hence, the number of mobiles n_i contained by a region is function of the region area. The table 1 above provides the coverage ratio of the AMC-based region. Let r_i be the radius of the region i with $r_0 = 0$, and R the cell radius. The populations n_i of regions $i = 1, ..., r$ are given by:

$$n_i = n\frac{r_i^2 - r_{i-1}^2}{R}$$

Note that the bit rate of the RT call is fixed by the ranging request information and we assume that all the RT calls in the system ask for the same bit rate independently of the region. Let R^{RT} be the throughput required by each RT call. The NRT bit rate is function of the available bandwidth in the system and the modulation used in the concerned region. Let R_i^{NRT} and the throughput reached by a NRT call in a region i.

For the NRT throughput calculation, let K be the number of data sub-carriers assigned to each sub-channel, let B be the baud rate (symbol/sec.) and let E_i be the efficiency of the modulation (bits/symbol). Now, consider $BLER_i$ as to the BLock Error Rate for the modulation used in the region i. We denote by L_i^k as the number of sub-carriers allocated for a call of class k in the region i. The physical bit rate R_i^k for a class RT/NRT traffic in region i is given by :

$$R_i^{NRT} = L_i^{NRT} \times K \times B \times E_i \times (1 - BLER_i) \qquad (1)$$

In addition, we can easily determine the number of sub-carriers required by a RT call in the region i :

$$L_i^{RT} = \frac{R^{RT}}{K \times B \times E_i \times (1 - BLER_i)} \qquad (2)$$

3.4 System arrivals and departures

Here, we define the arrivals and departures of calls that occur in the system. In order to determine the capacity of the system, and the performances of our CAC algorithm, we first have to study the request ranging and service processes.

RT and NRT requests income in the system from r regions. Let Z^k be the random variable of the number of incoming connections for class k, $k = \{RT, NRT\}$. The previous work [3] showed that the arrival of RT and NRT ranging requests are dependent and can be approximated by the multiplication of two gaussian distributions. The maximum number of RT (resp. NRT) ranging in function of the code range N_{RT} (resp. N_{NRT}) available for the these sorts of traffic. Remark also that the total number of arrivals can not exceed the total number of code. In our partitioned code ranging scheme: $Z^{RT} + Z^{NRT} \leq N_{RT}$. We denote Z as the arrival process: $Z = (RT_1, ..., RT_r, NRT_1, ..., NRT_r)$ where RT_i (resp. NRT_i) corresponds to the random number of RT (resp. NRT) calls that income in the region i. $P_z(a_1, ..., a_r, b_1, ..., b_r)$

$$= P\big(RT_1 = a_1, ..., RT_r = a_r, NRT_1 = b_1, ..., NRT_r = b_r\big)$$

$$= \frac{\Pi_{i=1}^{r} \begin{pmatrix} n_i \\ a_i \end{pmatrix} \begin{pmatrix} n_i - a_i \\ b_i \end{pmatrix}}{\begin{pmatrix} n \\ a \end{pmatrix} \begin{pmatrix} n - a \\ b \end{pmatrix}} P(Z^{RT} = a, Z^{NRT} = b)$$

where $a = \sum_{i=1}^{r} a_i$ and $b = \sum_{i=1}^{r} b_i$.

Now we determine the call departure as function of the classes k and regions i. First, the RT call duration is independent of the consumed resource. The resource is used for a exponentially distributed time with mean $1/\mu^{RT}$. This mean value does not change with the time and the system load. Conversely, a NRT call duration depends on the available resources shared by all the NRT calls. The service is exponentially distributed with mean μ_i^{NRT}. Please note that this mean change dynamically with the system load. At each slot, the mean number of services evolves with the resources available at the beginning of the slot (MAC frame). But this mean values is independent of the departures occurring in the current slot. Indeed, The IEEE802.16e MAC protocol informs the bandwidth available for the downlink via the

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

DL-MAP. This information is broadcasted at the very beginning of each MAC frame. So, whether a NRT call would want to use the bandwidth freed by the lately call terminations, it will be informs of that only during the next MAC frame DL-MAP. Thus, we consider that the available bandwidth for the NRT call does not change during the frame duration. So the mean time of a NRT call from region i is function of the mean file size $E(Pay)$, and the throughput of the NRT call R_i^{NRT} as follow:

$$\mu_i^{NRT} = \frac{R_i^{NRT}}{E(Pay)}$$

For the mean number of services calculation, we denote by n_i^k the number of calls in the system for the class k and the region i. Let T be the MAC frame duration. Thus, the mean number of call served during a slot is given by:

$$\lambda_i^{RT} = n_i^{RT}\mu^{RT}T \quad \text{and} \quad \lambda_i^{NRT} = n_i^{NRT}\mu_i^{NRT}T \qquad (3)$$

Now, to compute the service distribution, we assume that the system departures follow a discrete Poisson distribution. Let S_i^k be the number of services achieved by the system for calls of class k ($k = RT, NRT$) from the region i. The service distribution is given by:

$$P(S_i^k = x) = \begin{cases} \frac{(n_i^{RT}\mu^{RT}T)^x}{x!}e^{-n_i^{RT}\mu^{RT}T}, & \text{if } k = RT \\ \\ \frac{(n_i^{NRT}\mu_i^{NRT}T)^x}{x!}e^{-n_i^{NRT}\mu_i^{NRT}T}, & \text{if } k = NRT. \end{cases}$$

Finally, we compute the departure distribution of calls in the system. Let D_i^k be the number of calls that leave the system during a slot. We define the departure distribution as follow:

$$P_d(x) = P(D_i^k = x) = \frac{P(S_i^k = x)}{\sum_{j=0}^{n_i^k} P(S_i^k = j)}$$

3.5 System transitions

The system manages both classes of traffic over the r regions. Thus, we represent the system state as a vector $\overrightarrow{n}$. It is composed by the ongoing calls in the system. Let n_i^k be the number of remaining calls for the class k, $k = \{RT, NRT\}$ and region i ($i = 1, ..., r$). For the study needs, we define others vectors $\overrightarrow{n}^{RT}$ and $\overrightarrow{n}^{NRT}$, as respectively the vector of the number of calls for the RT and NRT calls in the system. So, the raw vector is defined as follow:

$$\begin{aligned} \overrightarrow{n} &= (\overrightarrow{n}^{RT}, \overrightarrow{n}^{NRT}) \\ \overrightarrow{n} &= (n_1^{RT}, ..., n_r^{RT}, n_1^{NRT}, ..., n_r^{NRT}) \qquad \overrightarrow{n} \in \mathbb{N}^{2r} \end{aligned}$$

Now, Let L be the total system bandwidth. We consider that a minimal portion of bandwidth can be allocated to all NRT calls noted by L_{min}^{NRT}, and RT calls have the bandwidth left:

$$L^{RT} = L - L_{min}^{NRT}$$

RT calls are assigned a given number of sub-channels per region L_i^{RT}, from the relation (2) among L^{RT}. Thus, the NRT calls use the bandwidth share denoted L^{NRT}. But, since the NRT calls tolerate throughput reduction, they will use the left over capacity on the basis of Processor Sharing (PS) [13]. Thus, due to the number of RT calls in the system, the NRT calls share among them the bandwidth portion L^{NRT} as follows:

$$\begin{aligned} L^{NRT}(\overrightarrow{n}^{RT}) &= L - \sum_{i=1}^{r} n_i^{RT}L_i^{RT} \\ \sum_{i=1}^{r} n_i^{RT}L_i^{RT} &\leq L^{RT} \\ L_i^{NRT}(\overrightarrow{n}^{RT}, \overrightarrow{n}_i^{NRT}) &= \frac{L^{NRT}(\overrightarrow{n}^{RT})}{\sum_{i=1}^{r} n_i^{NRT}} \end{aligned}$$

The state space of the system is obtained by computing all the states where the RT calls do not exceed their bandwidth capacity:

$$E = \{\overrightarrow{n} \in \mathbb{N}^{2r} | \sum_{i=1}^{r} n_i^{RT}L_i^{RT} \leq L^{RT}\} \qquad (4)$$

Then we introduce the vector $\overrightarrow{n}'$ that represents the system state at the next slot. This state is the results of all the arrivals and departures occurred in the state $\overrightarrow{n}$. The vector $\overrightarrow{n}'$ also belongs to the space E and is composed as follow:

$$\overrightarrow{n}' = (n_1'^{RT}, ..., n_r'^{RT}, n_1'^{NRT}, ..., n_r'^{NRT}) \qquad \text{with} \qquad \overrightarrow{n}' \in E$$

The transition probability computation is based on the possible transitions between $\overrightarrow{n}$ and$\overrightarrow{n}'$: $P(\overrightarrow{n}, \overrightarrow{n}')$. Let $x = (x_1^{RT}, ..., x_r^{RT}, x_1^{NRT}, ..., x_r^{NRT})$ be the vector representing the evolution between the state $\overrightarrow{n}$ and $\overrightarrow{n}'$. It is composed by the difference between the arrivals and the departures for the calls of class k in region i. We define the transitions as follows:

$$P(\overrightarrow{n}, \overrightarrow{n}') = P(\overrightarrow{n}' = \overrightarrow{n} + x)$$

Note that the system can evolve from the departure of all the ongoing calls of class k in region i to the arrival of calls from the entire population of the region i, $x_i^k \in [-n_i^k, n_i]$.

The transition probability calculations need to consider the possible evolutions for each region and traffic. But observe that the possible evolution for the RT calls depends on the resources available in L^{RT}. Indeed, the system can afford a limited number of RT calls defined in (4). Due to our CAC algorithm, the system first accepts the RT calls from the best modulations. By contrast, The NRT calls are accepted without limit and independently of the RT call occupancy. Here we need to introduce the particular case where the RT arrivals reach the border of the system capacity. This limit is characterized by the singular region i^* ($i^* = 1, \ldots, r$) wherein one or more request drop appended. Thus, the further region will not be able to accept any other RT. The states which are on the border of the space E. So let i^* ($i^* = 1, \ldots, r$), be the first region where at least one request is blocked by the base station. The values of i^* is given by:

$$\begin{aligned} i^* &= \min\Big(i | L^{RT} - \sum_{j=1}^{r}(n_j^{RT} - d_j^{RT})L_j^{RT} - \sum_{j=1}^{i} a_j L_j^{RT} < 0\Big) \\ i &= 1, ..., r. \end{aligned} \qquad (5)$$

In the sequel, we define below the transition behavior for the general and bordered case:

1. The RT arrivals in the slot never reach the capacity limit. Thus, $\overrightarrow{n}'$ is directly the difference between the arrivals and departures occurred in the regions without

losses. For this scenario, the transition probability is given by:

$$P(\overrightarrow{n}, \overrightarrow{n}') = \sum_{\substack{a_1=l_1\\ \vdots \\ a_r=l_r}}^{c_1,\dots,c_r} \sum_{\substack{b_1=m_1\\ \vdots \\ b_r=m_r}}^{e_1,\dots,e_r} P_z\big(a_1,...,a_r,b_1,...,b_r\big)D$$

where

$$D = \prod_{i=1}^{r} P_d\big(a_i - x_i^{RT}\big)P_d\big(b_i - x_i^{NRT}\big)$$

$$l_i = \max(0, x_i^{RT}) \qquad m_i = \max(0, x_i^{NRT})$$

$$c_i = n_i^{RT} + x_i^{RT} \qquad \text{and} \qquad e_i = n_i^{NRT} + x_i^{NRT}$$

2. The RT arrivals and departures define a border region i^* $(1 \leq i^* \leq r)$. In the sequel, all the incoming calls from the region $i = 1,...,i^*-1$ are accepted. But the ones from the region i^* are accepted only until the limit $x_{i^*}^{RT}$. Obviously, the calls incoming in the regions $j = i^*+1,...,r$ are not accepted. Thus, the transition probability is obtained by computing the cases where the system is not be able to serve enough calls in order to accept all the incoming attempts. In this case, the transition probability compute with the possible arrivals and departures that satisfy the region acceptance behavior previously described. For this, let δ be the Dirac function define as follow:

$$\delta(X) = \begin{cases} 1, & \text{if X is true} \\ 0, & \text{if X is false.} \end{cases}$$

$$P(\overrightarrow{n}, \overrightarrow{n}') = \sum_{\substack{a_1=l_1\\ \vdots \\ a_r=l_r}}^{c_1,\dots,c_r} \sum_{\substack{b_1=m_1\\ \vdots \\ b_r=m_r}}^{e_1,\dots,e_r} \sum_{\substack{d_1=0\\ \vdots \\ d_r=0}}^{n_1^{RT},\dots,n_r^{RT}} A \times D$$

where

$$\begin{aligned} A &= P_z\big(a_1,...,a_r,b_1,...,b_r\big) \\ D &= \delta_1\delta_2\delta_3 \prod_{i=1}^{r} P_d\big(d_i\big)P_d\big(b_i - x_i^{NRT}\big) \\ \delta_1 &= \prod_{j=1}^{i^*-1} \delta\big(n_j^{RT} + a_j - d_j = n_j'^{RT}\big) \\ \delta_2 &= \delta\big(-x_{i^*}^{RT} \leq d_{i^*} < a_{i^*}^{RT} - x_{i^*}^{RT}\big) \\ \delta_3 &= \prod_{j=i^*+1}^{r} \delta\big(d_j = x_j^{RT}\big) \end{aligned}$$

$$l_i = \max(0, x_i^{RT}) \qquad m_i = \max(0, x_i^{NRT})$$

$$c_i = n_i^{RT} + x_i^{RT} \qquad \text{and} \qquad e_i = n_i^{NRT} + x_i^{NRT}$$

Based on all these observations, we can compute the transition matrix P:

$$P = \Big(P(\overrightarrow{n}, \overrightarrow{n}')\Big), \text{ for } (\overrightarrow{n}, \overrightarrow{n}') \in E \times E$$

Now, we determine the steady-state probability vector $\overrightarrow{\Pi}$ and the solution of the steady-state distribution obtained by solving the set of linearly independent equations:

$$\overrightarrow{\Pi} = \{\pi(\overrightarrow{n}) | \overrightarrow{n} \in E\} \qquad \text{with} \qquad \begin{cases} \overrightarrow{\Pi}P = \overrightarrow{\Pi} \\ \sum_{\overrightarrow{n} \in E} \pi(\overrightarrow{n}) = 1 \end{cases}$$

4. PERFORMANCE MEASURES

4.1 Average throughput for NRT calls

Because the RT throughput is fixed by the user, we only focus here on the NRT call average throughput. We computed the physical bit rate reached by the NRT traffic in the relation (1). Here we calculate the average physical bit rate at the steady state. Let R_{NRT}^{tot} be the NRT throughput of the entire cell. The average throughput is given by:

$$E(R_{NRT}^{tot}) = \sum_{\overrightarrow{n} \in E} \pi(\overrightarrow{n}) \sum_{i=1}^{r} n_i^{NRT} R_i^{NRT}$$

4.2 Blocking probability for RT calls

The blocking probability consists in the probability that a RT call is blocked in the region j. Let P_B^j the probability that a RT call incoming from the region j is blocked. To obtain it, we compute the different state evolutions of the system due to all possible request arrivals and call departures. Eventually, we compute the probability $p_{i^*}^j$ that a RT call incoming in the region j $(a_j \geq 1)$ is blocked by the border of the system capacity. Let $\overline{a}_j$ be the maximum number of call that a border region i^* can accept. So, the blocking probability is given by:

$$P_B^j = \sum_{\overrightarrow{n} \in E} \pi(\overrightarrow{n}) \sum_{\substack{a_1=0\\ \vdots \\ a_r=0}}^{c_1,\dots,c_r} \sum_{\substack{b_1=0\\ \vdots \\ b_r=0}}^{e_1,\dots,e_r} \sum_{\substack{d_1=0\\ \vdots \\ d_r=0}}^{n_1^{RT},\dots,n_r^{RT}} A \times D \times p_{i^*}^j \qquad (6)$$

where $c_i = \min(n_i, N_{RT})$, $e_i = \min(n_i, N_{NRT})$ and

$$\begin{aligned} A &= P_z\big(a_1,...,a_r,b_1,...,b_r\big) \\ D &= \prod_{i=1}^{r} P_d\big(d_i\big)P_d\big(b_i - x_i^{NRT}\big) \\ p_{i^*}^j &= \begin{cases} 0, & \text{if} \quad j < i^*; \\ 1, & \text{if} \quad j > i^*; \\ 1 - \frac{\overline{a}_j}{a_j}, & \text{if } j = i^*. \end{cases} \end{aligned}$$

where

$$\begin{aligned} \overline{a}_j = \max\Big(a_i | L^{RT} &- \sum_{k=1}^{r} (n_k^{RT} - d_k^{RT}) L_k^{RT} \\ &- \sum_{l=1}^{i^*-1} a_l L_l^{RT} - a_i L_i^{RT} > 0\Big) \end{aligned}$$

$$i = 1, \dots, r$$

4.3 Mean transfer time for NRT calls

Since the NRT calls are not blocked by the system, they share fairly the available resource among them. Besides, the NRT calls remain in the system during a random period. We compute the mean transfer time by dividing the mean number of NRT call in a region i, $E(NRT_i)$, with the mean

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

NRT-request incoming-rate in the same region, Λ_i^{NRT}. The mean transfer time for the NRT traffic in the region i is given by the Little's law:

$$T_i^{NRT} = \frac{E(NRT_i)}{\Lambda_i^{NRT}} = \frac{\sum_{\overrightarrow{n}\in E}\pi(\overrightarrow{n})n_i^{NRT}}{\sum_{j=0}^{\min(n_i,N_{NRT})} jP(NRT_i=j)}$$

$$P(NRT_i=j) = \sum_{\substack{a_1=0\\ \vdots \\ a_r=0}}^{c_1,\dots,c_r} \sum_{\substack{b_1=0\\ \vdots \\ b_r=0}}^{e_1,\dots,e_r} P_z\left(a_1,...,a_r,b_1,...,b_r\right)$$

For $l = 1,\dots,r$. $c_l = \min(n_l, N_{RT})$ and $e_l = \min(n_l, N_{NRT})$. $b_i = j$ and $e_i = j$.

5. NUMERICAL RESULTS

Considering an OFDMA system (cell) with FFT size 1024 sub-carriers and cell must be decomposed into two regions with two AMC schemes; 64-QAM 3/4 (E_2=3 bits/symbol) and QPSK 1/2 (E_1=1 bits/symbol) respectively. Let $BLER = 0$, $L = 5$, $L_1^{RT} = 1$, $L_2^{RT} = 3$, $K = 48$, $B = 2666$ sym/sec, $E(Pay) = 500000$ bits [1] and $N_{RT} = 2$. The bit rate of RT calls is $R^{RT} = 384$ kbps, T=0.001 and $\mu_{RT} = 1/120$.

Impact of RT call duration on system performances

In order to investigate of the influence of RT call duration speed on NRT calls in our CAC. We plot in the figure 6, the average throughput NRT calls as function of minimum NRT bandwidth, for two cases of RT call duration values, i.e $1/\mu_{RT} = 1$ and $1/\mu_{RT} = 100$. We see in this figure that the average throughput is the same in the both cases. The same manner, we shows in the figures 7 and 8 that the RT call duration variation also have not a particular influence on the system performances. Thus, the RT call duration variation not affects our proposed CAC algorithm.

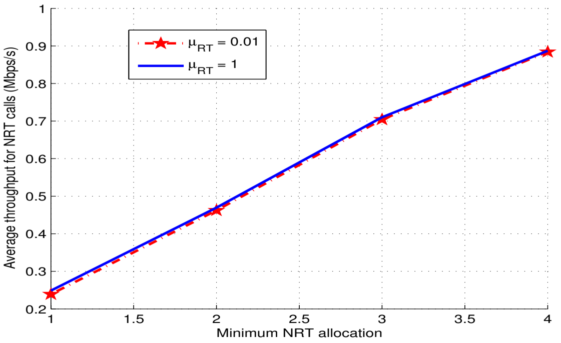

Figure 6: Average NRT throughput versus minimum NRT bandwidth threshold for different RT call durations

Impact of code partitioning

The figure 9 presents the average NRT throughput as function of the bandwidth allocated to the NRT calls for different code partitioning profiles. First, we observe that the average Non Real Time throughput increases linearly with the minimum bandwidth share allocated for the NRT calls (L_m^2in).

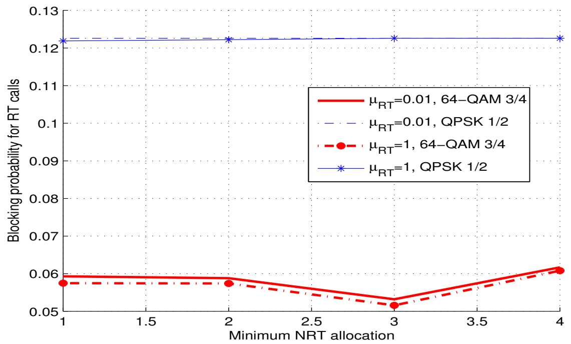

Figure 7: Blocking probability versus minimum NRT bandwidth threshold for different RT call durations

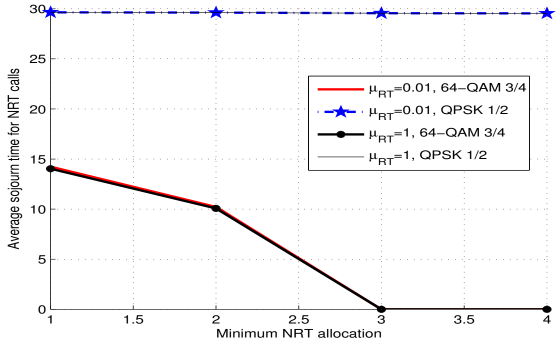

Figure 8: Average sojourn time versus minimum NRT bandwidth threshold for different RT call durations

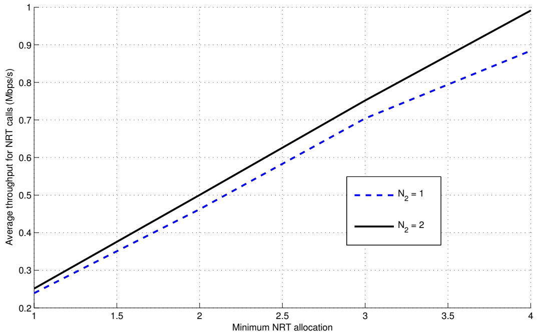

Figure 9: Average NRT throughput versus minimum NRT bandwidth threshold for different code partitioning profiles

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

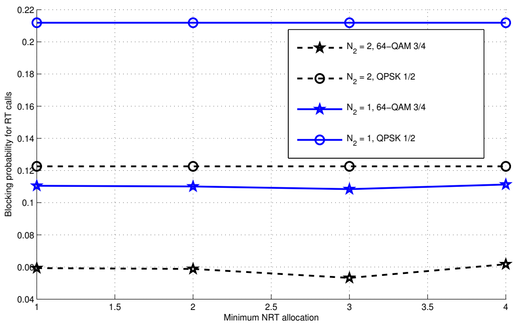

Figure 10: Blocking probability versus minimum NRT bandwidth threshold for different code partitioning profiles

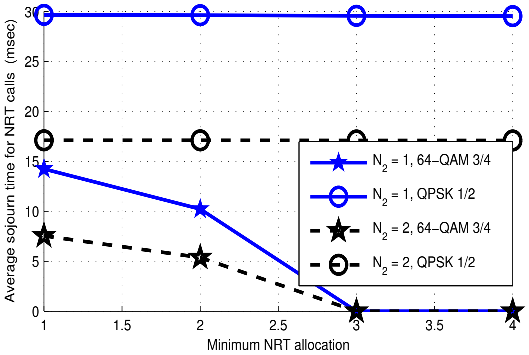

Figure 11: Average sojourn time versus minimum NRT bandwidth threshold for different code partitioning profiles

Second, by using the code partition scheme, we decrease by almost 9% the NRT call throughput. In fact, by lowering the number of codes available for the NRT requests, we reduce the RT request collision and hence they increase the RT occupancy of the system. Observe that this minor impact of the code partition scheme have in fact awful effects on the system performances.

The next figure shows in 11 the mean transfer time for the NRT calls in both regions for different code partitioning profiles. As expected, since the NRT bandwidth threshold increases, the average sojourn time for both regions drastically decreases. Moreover, we easily determine a threshold value where the gain offer by upper values is negligible. In addition, we observe that the use of the code partition scheme leads to the collapse of the sojourn time performance. As explained previously, by reducing the number of codes available for the NRT requests, we decrease the NRT average throughput. Thus, the average sojourn time largely increase in both regions.

Finally, the figure 10 represents the blocking probability in the inner and border regions for different code partitioning profiles. Here, we first observe that the blocking probability for the inner region is sensitive to the NRT bandwidth threshold. However, we also observe that the blocking probability of the border region remains very high and rises slightly with the minimum NRT call bandwidth. This is the consequence of the CAC policy where the system first seeks to accept all the calls that income in the inner region before to accept any call from the border region. This observation leads us to develop a more flexible CAC algorithm that we describe in the future works part of the conclusion. In addition, the figure 10 shows also the devious effect of the code partition scheme. Indeed, if the RT requests have a dedicated code subrange, the mean number of RT requests increases. Therefore, numerous calls will be blocked by the system.

Based on these results, a service provider is able to determine its proper NRT bandwidth thresholds according with the Quality of Service that he might want to introduce for the NRT call customers. Note that this proposition is particularly profitable for the file transfers from the IEEE802.16e nRTPS class.

6. CONCLUSION

In this paper we completed a performance evaluation for the IEEE802.16e standard. We focused on the effects of the standardized AMC [5] and code partitioning [2] schemes .

Since no Connection Admission Control algorithm specifically designed for the IEEE802.16e standard, we desire to define different approaches that set a tradeoff between the Real Time flow prioritization and the Non Real Time flow throughput. Here, we proposed a CAC algorithm where a minimum share of the total bandwidth is allocated to the NRT calls. In addition, the system seeks to first accept the calls incoming in the inner region, i.e. the calls that require fewer resources. The RT calls are characterized by the same physical bit rate needs, and asks for resource as function of the modulation efficiency of their region. The NRT calls share the available bandwidth left by the RT calls and remain in the system according with these resource consumptions.

Our study is based on a Discrete Time Markov Chain (DTMC). The paper provides a complete set of general closed-form relations for the average NRT throughput and sojourn time as well as the RT blocking probability in each region. From this work, we observe the effect of our proposed CAC algorithm in the AMC and code partitioning environment.

The results show that the existence of a minimum bandwidth share for the Non Real Time calls greatly improves the NRT call performances. The CAC algorithm largely increases the NRT throughput and hence drastically decreases the sojourn time. In addition, we observe that the NRT bandwidth allocation have a major impact on the RT blocking probability. However, our study allows the service providers to find a tradeoff as function of their customer needs.

Our future work is now motivated by the observation that the blocking probability remains very high in the border region. Our CAC algorithm strongly prioritize the acceptance of the calls incoming in the inner regions. Thus, the calls from the border regions are often rejected. We would like to propose a more flexible CAC algorithm where an incoming call from a region i is accepted with a probability α_i. Then we will be able to define several α_i distributions which leads to a better performances tradeoff.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421

Please note that the interested reader will refer to [2] where we provide additional figures and analysis. In addition, we also propose new CAC algorithm and tuning proposal for the IEEE802.16e standard.

7. REFERENCES

[1] T. Chahed C. Tarhini. On capacity of ofdma-based ieee802.16 wimax including adaptative modulation and coding (amc) and inter-cell interference. *Local and Metropolitan Area Networks*, June 2007.

[2] T. Peyre R. El-Azouzi and T. Chahed. http://lia.univ-avignon.fr/fich_art/rrpeyre07.pdf.

[3] T. Peyre R. El-Azouzi and T. Chahed. Qos differentiation for initial and bandwidth request ranging in ieee802.16. *Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2008.

[4] IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air interface for fixed broadband wireless access systems, April 2002.

[5] WiMax Forum. Ieee802.16e/d12, air interface for fixed and mobile broadband wireless access systems. IEEE Standard for local and Metropolitan Area Networks, February 2005.

[6] T. Madejski K. Perycz P. Putzolu G. Nair, J.Chou and J. Sydir. Ieee802.16 medium access control and service provisionning. *Intel Technology Journal*, 8, August 2004.

[7] E. S. Hwang-C. H. Cho H. H. Seo, B. H. Ryu and W. Lee N. A study of code partioning scheme of efficient random access in ofdma-cdma ranging subsystem. In *JCCI 2004*, page 262, April 2004.

[8] W. Li H. Wang and D.P. Agrawal. Dynamic admission control and qos for 802.16 wireless man. In *Wireless Telecommunications Symposium*, pages 60 – 66, April 2005.

[9] H. Wang J. Chen, W. Jiao. A service flow management strategy for ieee802.16 broadband wireless access systems in tdd mode. In *IEEE International Conference on Communications, ICC 2005*, volume 5, pages 3422 – 3426, May.

[10] B. H. Ryu-C. H. Cho J. J. Won, H. H. Seo and H. W. Lee. Perfomance analysis of random access protocol in ofdma-cdma. In *KICS Fall Conference*, July 2003.

[11] K. Kim J. You and K. Kim. Capacity evaluation of the ofdma-cdma ranging subsystem in ieee802.16-2004. *Wireless and Mobile Computing, Networking and Communications (WiMob)*.

[12] Dongyan Jia Liangshan Ma. The competition and cooperation of wimax, wlan and 3g", mobile technology, applications and systems. *2nd International Conference on Mobile technology, Applications and systemsIEEE Infocom, Miami*.

[13] F. Delcoigne S. Oueslati-Boulahia N. Benameur, S. Ben Fredj and J.W.Roberts. Integrated admission control for streaming and elastic traffic. *QofIS 2001, Coimbra*, Sept. 2001.

[14] R. El-Azouzi T. Peyre. Performance analysis of single cell ieee802.16e wireless man. *Local Computer Network (LCN)*, Oct. 2007.

[15] H. Yaghoobi. Scalable ofdma physical layer in ieee802.16 wirelessman. *Intel Technology Journal*, pages 201–212, August 2004.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4421
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4421