The CAC Model and QoS Management in Wireless Multiservice Network

Sonia Ben Rejeb, Zièd Choukair, Sami Tabbane
Unité de recherche Médiatron_Sup'Com
Route de Raoued, km 3.5, 2083 Cité el Ghazala - Tunis

***Abstract*—Call admission control (CAC) is a key for ensuring the quality of service (QoS) in wireless multiservice network. With the advances in wireless communication technology and the growing interest in deploying multimedia services in wireless networks, the issue of providing an efficient CAC has come to the fore. A suitable CAC for the multimedia service networks is expected to make efficient use of the scarce wireless resource while supporting different services with different QoS metrics. In this paper, we propose a CAC model with four priority levels supporting five classes of service (e.g. UGS, ErtPS, rtPS, nrtPS and BE) in a wireless multiservices network: IEEE 802.16e. The idea is to design an efficient CAC that deals with new and handovers (HO) calls. The proposed model manages to establish priority between new end handover calls. It is assumed that the system operates under a reservation channel scheme and a queuing strategy in order to maintain the HO priority. Our model must maintain a balance between two conflicting requirements: maximise the resource utilisation and minimize the forced handover call dropping rate. In order to maintain the maximum resource utilisation, the maximum number of calls should be admitted into a network which may result in unacceptably high HO call dropping rates due to insufficient resources for HO calls. It is very important to propose a CAC model with reserves the minimum amount of necessary resources to maintain an acceptable HO call dropping and provide high resource utilisation.**

Keywords-Wireless multiservice network : IEEE 802.16e, CAC, resource allocation, mobility, QoS.

I. INTRODUCTION

With the constant improvement of wireless technology and the explosive growth of wireless communication market, the demand for newer multimedia applications is increasing rapidly. The convergence of wireless technology and multimedia applications presents network operators with enormous opportunities as well as great challenges. QoS provisioning and mobility management are two key challenging issues that must be addressed in wireless multiservices networks.

QoS provisioning in wireless networks supporting multimedia applications have to meet the expectations of users while maintaining reasonably high utilisation of radio resources.

The QoS provisioning problem is more challenging than in fixed networks for two main reasons. First, the link bandwidth resource is limited in a wireless environment.

Second the changing environment in wireless networks due to the user's mobility and interference results in varying bandwidth. Thus, how to allocate and how to use the limited wireless resources efficiently are to be studied.

On the other hand, the mobility management is another important issue to be addressed when studying the wireless multiservices networks. A mobile user will be able to freely move across the networks while maintaining its current communications. An interrupted communication is a very frustrating phenomenon that may happen to a user. Thus, an efficient CAC protocol must manage to avoid the forced termination of an ongoing call.

The goal of our paper is to study the CAC and the resource allocation in a QoS framework within wireless multiservices network: IEEE 802.16e.

In this paper, the next section is dedicated to the state of the art and the following will present the definition of the problematic. Section IV will develop the call admission control model and QoS improvement. Section V will present the model implementation and simulation results analysis. We close this paper by a conclusion and the perspectives of this work.

II. STATE OF THE ART

Several researches are focused on defining optimal control admission mechanisms in wireless networks.

In [1], the authors propose a control admission policy in wireless networks based on the guard channel. Indeed, in order to give HO calls a higher priority than new calls, it must be exclusively a set of channels, codes, or a bandwidth proportion. If C is the total number of available channels in the cell, it will be divided into two parts: CA is the part used to serve the new call, and CH the part used to serve the HO calls (with C=CA+ CH). A new call is admitted if the total number of calls (including calls HO from other cells) is below the threshold CA. HO call is served if the total number of calls in a cell is less than the total capacity C.

In [2], the authors propose CAC architecture based on the same principle (ie. guard channels) but in the case of a multiservice network that supports different types of traffic

with different QoS requirements. The architecture consists of three services: premium (conversational services), assured (streaming services) and BE (interactive and background services). In order to simplify the calculation, the authors have limited their study on the two first classes but the results can be extended to a larger number of services classes.

In [3] the authors treat the case of a wireless radio network that supports two services: voice and data, and propose a control admission mechanism at four levels of priority in the following order: New data connection <New call voice <HO data call <HO call voice. HO call is served if there is an available channel. If not, the request is in the queue. In each cell, there are two files for HO calls of voice and data. HO voice call on hold is removed from the file if the mobile crosses the area HO without a new channel, or if the communication ends before crossing the area HO. Data connections have not real time constraints, they are more tolerant to delays unlike voice calls. Therefore, it is assumed that there is not an area HO for data users, there is only a border between cells. So, if a HO data call is not served in the current cell, it will be transferred to the file devoted for data calls in the target cell instead it will be deleted. Indeed, the deletion of a HO data call can induce a high penalty due to the connection resetting and the retransmission of a large amount of data.

III. PROBLEMATIC

The 802.16 MAC protocol [4][5][6] is oriented to connection. Security Sublayer (SS) shall establish a connection to the Base Station (BS) to transmit data. The sending process data is presented in Figure 1. It begins with a connection establishment phase which includes the QoS negotiation parameters and control admission. If sufficient resources are available, the connection is accepted. The SS shall send a request for bandwidth allocation. This step requires an effective scheduling of flows in both parties BS and SS. Blocks dotted traces show the parts defined by the standard, they are left to constructors in order to be implemented.

Several researches focused on defining optimal mechanisms for those parts not covered by the standard. In literature, the algorithms proposed to allocate bandwidth to different classes of service using a hierarchical architecture. Indeed, the bandwidth assigned to connections beginning with the higher priority class (e.g. UGS) and then moving on to the lower priority classes. Then, the connections of each class are arranged according to different scheduling mechanisms. For example in [7], rtPS connections are provided by EDF discipline, connections nrtPS follow Weighted Fair Queuing (WFQ) discipline and BE connections are provided by the First in First Out (FIFO) discipline.

In following, we will introduce our model for improving the QoS in 802.16e networks.

We propose a CAC mechanism in this network that can satisfy both of the following purposes:

- The QoS satisfaction constraints of five services classes: UGS, ErtPS, rtPS, nrtPS and BE.
- And call management HO to be higher priority than new calls.

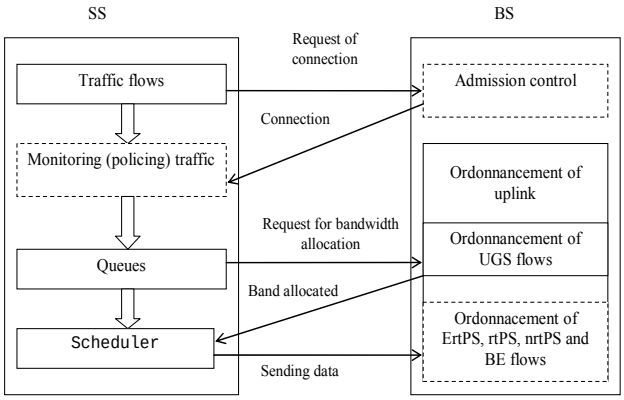

Figure 1. Data sending process and problematic

IV. CAC PROPOSED MODEL AND QOS IMPROVEMENT

A. *CAC proposed model*

In our CAC model (see Figure 2), we are interested in a single cell in a homogeneous network consisting of several cells. Our cell has a constant amount of bandwidth B divided in two proportions, first for the up link and the second for the down link. The cell serves heterogeneous users who require different services (e.g. voice, data, video, etc.) in the form of new (N) or handover (H) calls coming from the adjacent cells.

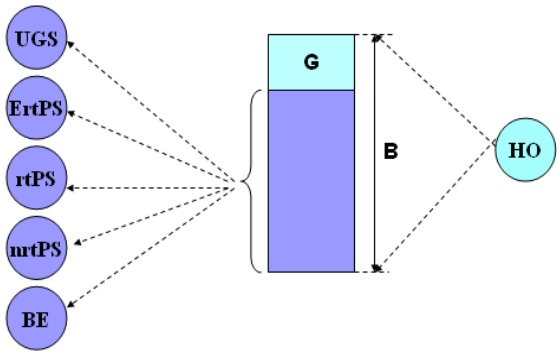

Figure 2. CAC proposed model within wireless multiservice network

According to the requirements of each service class in terms of bandwidth and delay, it is intuitive to consider the following priority system:

$$P(UGS)>P(ErtPS)>P(rtPS)>P(nrtPS)>P(BE) \quad (1)$$

With P(x) denotes the priority of the x class.

Thus, the bandwidth is allocated to connections with highest priority class before the lowest priority class.

The architecture (1) includes 5 priorities, and if we want to take into account HO calls, the architecture (2) includes 10 priorities as follows:

P(UGS,H)>P(ErtPS,H)>P(rtPS,H)>P(nrtPS,H)>P(BE,H)>
P(UGS,N)>P(ErtPS,N)>P(rtPS,N)>P(nrtPS,N)>P(BE,N) (2)

However, this architecture presents an enormous complexity in terms of call management because each defining priority queue is served according to a scheduling algorithm suitable for the service class involved.

In our model and to reduce this complexity we propose to group the service classes according to their time requirements i.e in real-time RT connections and non-real time NRT connections. The Architecture (3) will include the following four priorities:

P (RT, H)> P (NRT, H)> P (RT, N)> P (NRT, N) (3)

With:

- P(RT, H): priority assigned to real-time HO calls (includes HO calls belonging to UGS, rtPS and ErtPS).
- P(NRC, H) : priority assigned to non-real time HO calls (includes HO calls belonging to nrtPS and BE).
- P(RT, N): priority assigned to the new real-time connections (includes new connections belonging to UGS, rtPS and ErtPS).
- P(NRT, N): priority assigned to the new non-real time connections (includes new connections belonging to nrtPS and BE).

Since HO calls have higher priority than new connections, in our model, a bandwidth proportion will be restricted to HO calls.
Real-time calls are characterized by a maximum tolerance to delay. Upon arrival of each real time call, we propose to calculate a deadline after which if the call is not served, it will be rejected.

As for non-real time calls, they will be served according to their order of arrival.

B. Bandwidth reservation: assumptions and parameters

Before detailing the bandwidth reservation algorithm, we present the following assumptions and parameters:

- B: total bandwidth available to satisfy the uplink calls at the BS.
- G: guard band reserved exclusively to serve the HO call.
- U: bandwidth occupied at time t.
- b: bandwidth requested by a new connection.
- b_{HO} : bandwidth requested by a HO call.
- N_a: total number of new calls admitted.
- N_r: total number of new calls rejected.
- H_a: total number of HO calls admitted.
- H_r: total number of HO calls rejected.

A HO call requesting bandwidth b_{HO} is served if: $U + b_{HO} \leq B$. If not, it will be put in a queue.

If the call is a real time (e.g.UGS, rtPS or ErtPS), it must respect the constraint of time. If this constraint is not respected, the call is rejected.

A new connection requiring a bandwidth b is permitted if:

$U+b \leq B-G$. If not, it will be rejected.

The implementation of our model is detailed by the following algorithm:

- Call generation: we take a poissoniens arrival calls.

- Drawing a random value indicating the call type: N: new, or H: HO.

- If N, drawing a second random value indicating whether the call is real time or non-real time: N_{rt} or N_{nrt}
  . If N_{rt}, treatment of new real-time call
  . If N_{nrt}, treatment of new non-real time call
- If H, drawing a second random value indicating whether the call HO is real time or non-real time: H_{rt} or H_{nrt}
  . If H_{rt}, treatment of a HO real-time call
  . If H_{nrt}, treatment of a HO non-real time call

- At the end of the simulation we calculate:
- The blocking probability of new calls: P_{bn}
- The dropping probability of HO calls: P_{ch}
- And the waiting time average of real-time connections.

With: $$P_{bn}=N_r/ (N_a+N_r) \quad (4)$$

$$P_{ch}=H_r/ (H_r+H_a) \quad (5)$$

The treatment of each call type is described by the following procedures:

1) A real-time new call treatment

A real time new call is characterized by (see Figure 3):

- b: bandwidth requested
- d_{max}: maximum delay tolerance
- d: time that the call spent in the queue until the instant t.

The treatment of this type of call is as follows:

✓ Call arrival:

- If $U+b \leq B-G$; the call is admitted: $N_a=N_a+1$ and $U=U+b$ (requested bandwidth reservation)
- If not, d=d+frame duration and as long as $d \leq d_{max}$,
  . If $U+b \leq B-G$; the call is admitted: $N_a=N_a+1$ and $U=U+b$
  . If not d=d+frame duration
- If $d>d_{max}$, $N_r=N_r+1$

This procedure is presented by the following diagram:

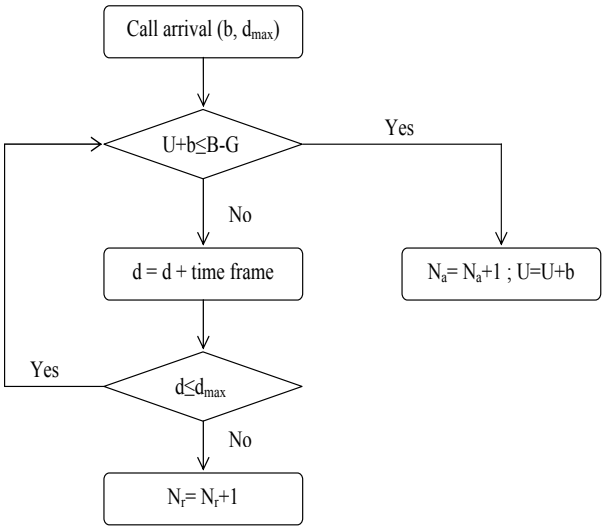

Figure 3. A new real-time call treatment

2) A non-real time new call treatment

A non-real time new call is characterized by (see Figure 4):

- b: bandwidth requested
- b_{min}: reserved minimum rate
- d: time that the call spent in the queue until the instant t

Other parameter is d_{max}: maximum delay tolerance during which the call will wait in the queue. It is necessary to be specified in order not to remain indefinitely waiting in the queue.

The treatment of this type of call is as follows:

✓ Call arrival:

- If U+b≤B-G; the call is admitted: N_a=N_a+1 and U=U+b
- If not, if U+b_{min}≤B-G; the call is admitted: N_a=N_a+1 and U=U+b_{min}
- If not, d=d+frame duration and as long as d≤ d_{max},
  . If U+b≤B-G; the call is admitted: N_a=N_a+1 and U = U+b
  . If not, if U+b_{min}≤B-G; the call is admitted: N_a=N_a+1 and U=U+b_{min}
  . If not d=d+duration of frame
- If d>d_{max}, N_r=N_r+1

This procedure is presented by the following diagram:

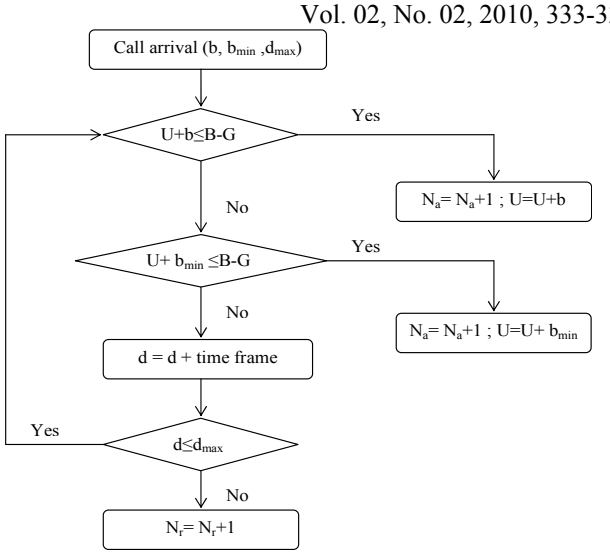

Figure 4. A not-real time new call treatment

3) Real-time HO call Treatment

A real-time HO call is characterized by (see Figure 5):

- b_{HO} : bandwidth required
- d_{max}: maximum delay tolerance
- d: time that the call spent in the queue until the instant t

The treatment of this type of call is as follows:

✓ Call arrival:

- If U+b_{HO}≤B, the call is admitted: H_a=H_a+1 and U=U+b_{HO}
- If not, d=d+frame duration and as long as d≤d_{max},
  . If U+ b_{HO}≤B, the call is admitted: H_a=H_a+1 and U=U+b_{HO}
  . If not d = d+duration of frame
- If d>d_{max}, H_r =H_r+1

This procedure is presented by the following diagram:

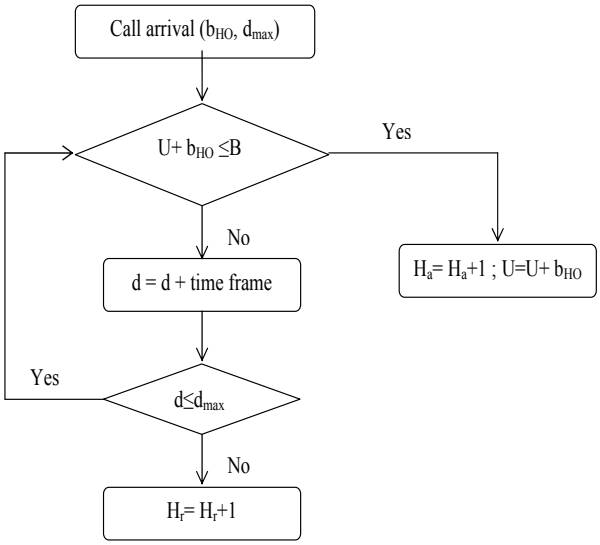

Figure 5. Real-time HO call treatment

4) Non-real time HO call treatment

A not real-time HO call is characterized by (see Figure 6):

- b_{HO} : maximum bandwidth permissible
- $b_{HO,min}$: reserved minimum rate
- d: time that the call spent in the queue until the instant t

Other parameter is d_{max}: maximum delay tolerance during which the call will wait in the queue. It is necessary to be specified in order not to remain indefinitely waiting in the queue.

The treatment of this type of call is as follows:

✓ Call arrival:

- If $U+b_{HO} \leq B$, the call is admitted: $H_a=H_a+1$ and $U=U+b_{HO}$
- If not, if $U+b_{HO,min} \leq B$, the call is admitted: $H_a=H_a+1$and $U=U+b_{HO,min}$
- If not, d=d+frame duration and as long as $d \leq d_{max}$,
  . If $U+b_{HO} \leq B$, the call is admitted: $H_a=H_a+1$ and $U=U+ b_{HO}$
  . If not, if $U+b_{HO,min} \leq B$, the call is accepted: $H_a=H_a+1$ and $U=U+ b_{HO,min}$
  . if not d=d+frame duration
- If $d>d_{max}$, $H_r=H_r+1$

This procedure is presented by the following diagram:

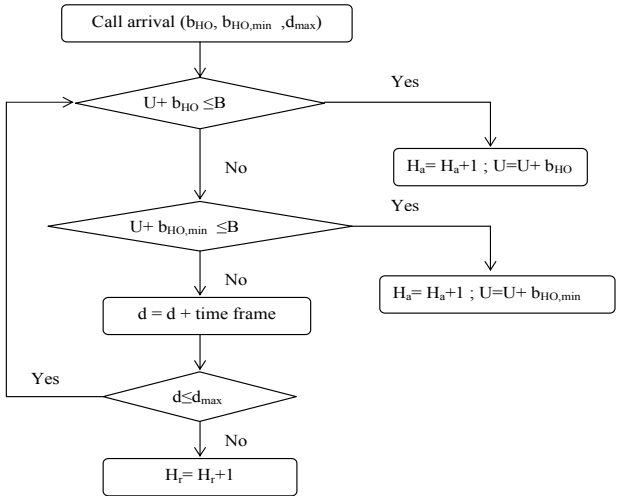

Figure 6. A non-real time HO call treatment

V. CAC MODEL IMPLEMENTATION AND SIMULATION RESULTS ANALYSIS

A. Simulation envirenment

In developing our algorithm, we used Matlab, version 7.0.1. This software is chosen because it has a basis of functions already implanted and necessary for modelling different traffic model laws and modelling queues. Indeed, functions such as: poissrnd or exprnd were very useful for the generation of traffic models based on random processes (poissoniens, exponential, etc.).

Other functions like plot were also very useful for the results representation in curves form.

B. Simulation parameters

In the simulation design, we adopted the following assumptions and parameters [4][5][6]:

- There is a single cell in a homogeneous network consisting of several cells.
- Bandwidth available in the BS:10Mbps divided equally between the uplink and the downlink.
- Frame duration: 5ms.
- The arrival process of new and HO calls are poissonniens with respective parameters λ and μ with $\lambda = \mu * 5$.
- The observation duration is equal to 200frames.
- We consider two types of traffic:

✓ Video traffic using MPEG-4 coding with :

- An average flow of 180Kbps.
- The inter-arrival packets duration is constant and equal to 40ms.
- The session duration is equal to 5s.
- The packet average size is equal to 900 bytes.
- The maximum waiting time for a new connection is equal to 40ms.

✓ FTP traffic: the flow varies between 16 and 256Kbps. The reserved minimum flow is equal to 5Kbps.

C. Presentation and results analysis

1) The guard band Influence on the new calls blocking probability

We have simulated the behaviour of the BS for the arrival rates of new calls λ. These rates vary from 25 to 45calls/s (with arrival rate of HO calls: $\mu=\lambda/5$) in both cases: G=0 and G=B/10 with B=5Mbps. The calculation of new calls blocking probability gave the two following curves (see Figure 7):

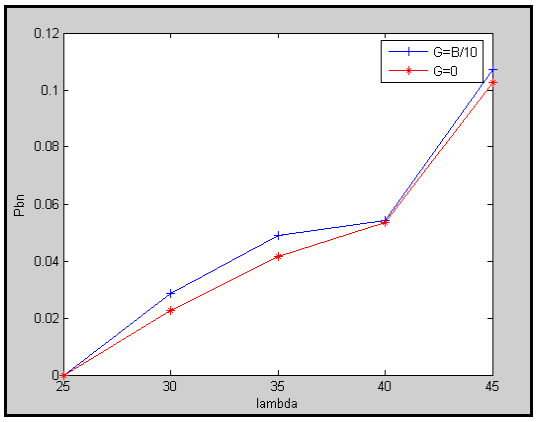

Figure 7. Influence of the guard band on the new call blocking probability

According to Figure 7, we see that the increasing of λ (and therefore μ) leads the increase of new connexions blocking probability. For λ varies between 25 and 35calls/s, we have an acceptable blocking rate (between 0 and 4%), beyond these values, the blockage rate becomes too high, it exceeds 10% when λ exceeds 45calls/s. Comparing these two cases where G=0 and G=B/10,

The first case where G=0 offers the lowest blocking rate, this is well justified since in this case, the new and HO calls are treated similarly: as long as the bandwidth is available, they are admitted. In the second case where G=B/10=500kbps, a part of the bandwidth is reserved exclusively for HO calls therefore the new calls have less chance to be admitted which causes the increased of calls blocking probability.

2) Influence of the guard band on the HO calls dropping probability

In order to study the influence of the guard band on the HO calls dropping probability, we have simulated the BS behaviour with the same parameters used in the precedent paragraph. Simulation results are presented in Figure 8.

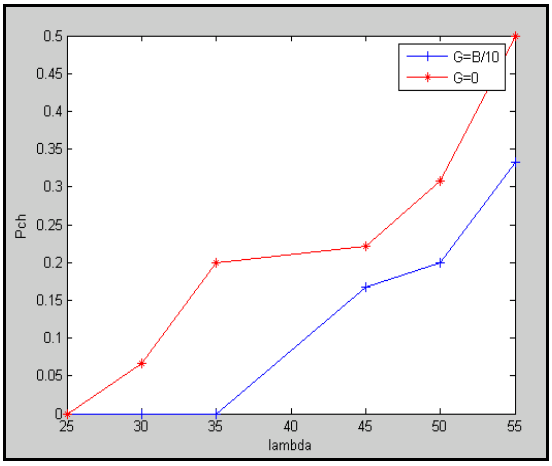

Figure 8. Influence of the guard band on the HO calls dropping probability

We note that the admission control policy with guard band offer a HO calls dropping rate much lower than without guard band that varies from 2 to 20times for an arrival rate λ varying between 30 and 55calls/s. Indeed, till 35calls/s, all HO calls are admitted but over 45calls/s, Pch becomes too high and exceeds 10% which is very inconvenient for the user because the dropping call is less desired than its blocking from the beginning, we may think to act on the parameter G in order to keep this probability below a certain threshold for a very high calls arrival rate.

3) Influence of the guard band on the average waiting time of new real-time connections

In this section, we propose to calculate the average waiting time for new real-time calls in the queue before being served. We adopted the following assumptions: the maximum waiting time of a real time call in the queue is 40ms, after this period if it is not served, it will be rejected. In the case of HO calls, this time is 10ms. By varying λ between 25 and 55calls/s, we obtained the following results (see Figure 9).

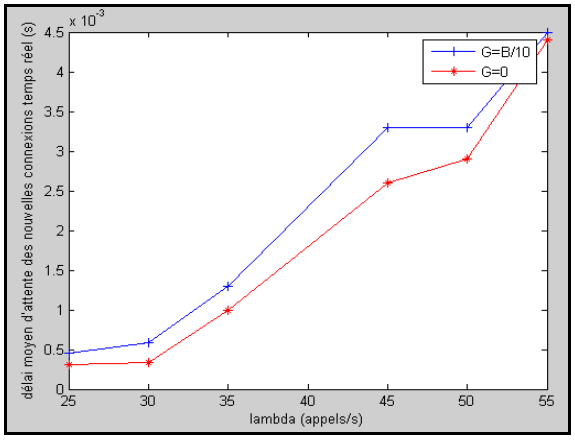

Figure 9. Influence of the guard band on the average waiting time of new real-time connections

We note from Figure 9 that the average waiting time of real time new connections increases with the arrival rate of new calls and the guard band G. Indeed with G=0, the new calls are more likely to be used since they can benefit from the available bandwidth, then they will wait at least in the queue. In the case where G≠0, ie when a proportion of bandwidth is reserved for HO calls, there may have a bandwidth that is available but not accessible by new calls since it is only for HO calls, which leads to waiting in the queue and the call may be rejected if it exceeds the maximum of the period that has been affected.

4) Influence of the guard band on the average waiting time of real-time HO calls

In this section, we adopt the same assumptions as the previous paragraph and we intend to study the effect of guard band on the average waiting time of real time HO calls. The two cases where G=0 and G=B/10 gave the two curves in Figure 10.

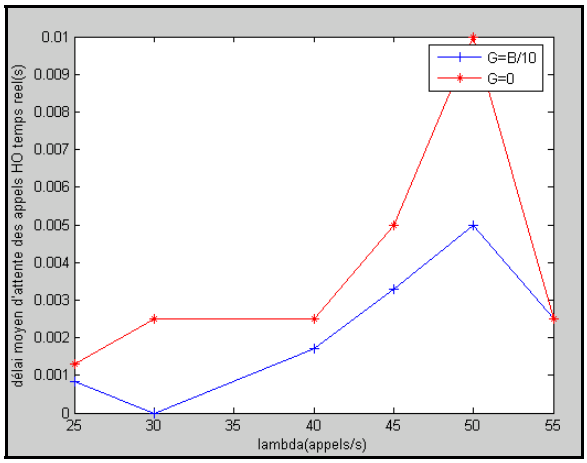

Figure 10: Influence of the guard band on the average waiting time of real-time HO calls

Figure 10 shows that the admission control policy with guard band provides higher performance than that without guard bands in the point of view of the real time HO call management. Indeed, these calls are very demanding in terms of time; therefore, upon a request for migration from one cell to another, it must be ensured that the call waiting time, before being served, it remains below a certain threshold. According to the presented values in this figure, our approach saves us about half the time compared to a policy without guard band.

VI. CONCLUSION AND PERSPECTIVES

In this paper, we presented our approach based on admission control mechanism that takes into account the HO management and different service classes defined by the 802.16e standard. This approach is based on the guard band principle which reserves a bandwidth to meet and give more priority to HO calls.

Compared with the admission control policy without guard band, the simulation results show that this policy provides better performance regarding the management of HO calls. Indeed, it offers a lower dropping probability and a lower average waiting time for real time HO calls. However this policy has a higher rate of blocking new calls.

In our future work, we will try to adjust dynamically the guard band G depending on network load and new call blocking and HO call rates.

This proposed approach can be applied for the fourth generation networks. Its applicability to networks having different technologies can also be envisaged, if complemented with the necessary handover functionalities. Thus, it is worthwhile to study the vertical handover in networks having different technologies.

REFERENCES

[1] Y.Zhang, D.Liu, "An adaptive algorithm for call admission control in wireless networks", Department of Electrical and Computer Engineering University of Illinois, Chicago.

[2] Y.Wei, C.Lin, F.Ren, R.Raad, E.Dutkiewicz, "Dynamic handoff scheme in differentiated QoS wireless multimedia networks", Department of Science and Technology, Tsinghua University, China Motorola Australian Research Centre, Australia, Computer Communications 27 (2004), pp. 1001–1011.

[3] R. Naja, "Mobility managment and resources allocation in wireless multiservice network, ENST, 2003.

[4] M. Castrucci "A framework for resource control in WiMax network", IEEE/NGMAST 2007.

[5] K. Gakar "How many traffic classes do we need in WiMax ", IEEE/WCNC 2007 proceeding, pp. 3706-3711.

[6] M.C.Wood, "An analysis of the design and implementation of QoS over IEEE 802.16", 2006.

[7] K. Wongthavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems", International Journal of Communication Systems, vol. 16, pp. 81–96, 2003.

[8] T.Tsai, C.Jiang, C.Wang, "CAC and packet scheduling using token bucket for IEEE 802.16 networks", National Chengchi University, Taipei, Taiwan, ROC, May 2006.

[9] Wimax Forum, "Mobile WiMAX–Part I: A Technical Overview and Performance Evaluation", August 2006.