Proposal of a Novel Bandwidth Management Framework for IEEE 802.16 Based on Aggregation

Mohamad El Masri, Slim Abdellatif, Guy Juanole

Proposal of a novel bandwidth management framework for IEEE 802.16 based on aggregation

Mohamad El Masri
LAAS-CNRS
Université de Toulouse
Toulouse, France
Email:masri@laas.fr

Slim Abdellatif
LAAS-CNRS
Université de Toulouse
Toulouse, France
Email:slim@laas.fr

Guy Juanole
LAAS-CNRS
Université de Toulouse
Toulouse, France
Email:juanole@laas.fr

***Abstract*—In an environment where IEEE 802.16 access networks are used as a backbone for a heterogeneous wireless access network, the question of Quality of Service brings along a questioning on several other aspects: scalability, guarantees, scheduling among others. We give in this paper a general solution answering those questions. This solution is based on an aggregated bandwidth management in the backbone. This aggregation along with specific bandwidth request and scheduling policies are herein specified and discussed.**

I. Introduction

The IEEE 802.16 standard [1], [2] (on which is based the Worldwide Interoperability for Microwave Access - WiMAX) defines the physical layer and Medium Access Control (MAC) layer for Broadband Wireless Access systems supporting multimedia services. The WiMAX Forum is dedicated to certifying the interoperability of WiMAX products. WiMAX's connection oriented features, its contention-less access schemes and its mechanisms providing services of different degrees dedicated to multimedia applications makes WiMAX *The ultimate solution for Quality of Service* [3]. A classical point-to-multipoint (PMP) WiMAX network is made out of a centralized Base Station (BS) that is responsible for organizing its downlink (DL) access to the medium and that of the Subscriber Stations (SS) uplink (UL) access. Access to the medium for data transfer is done in a connection oriented contention-less polling fashion and the multiplexing of UL and DL accesses is done using TDMA mechanisms. Each connection, identified by its CID, will have (depending on its class of service) several ways to individually request its bandwidth needs. In early WiMAX forum documents such as [4], one of the suggested business case scenarios was the use of the IEEE 802.16 access as a WiFi Hot Spot Backhaul. However, along with the Quality of Service comes the question of scalability and of guarantees. It is clear that in such a scenario, leaving to the WiMAX BS the responsibility of managing the bandwidth provision of each connection individually becomes a serious scalability issue. This is also the case in other scenarios presented in the business case analysis [4]: for example serving high speed internet access in rural areas where DSL services are not available.

In this paper, we present and discuss a bandwidth management framework for WiMAX based on an aggregated management of the bandwidth provisioning. The framework we propose is driven by two main ideas: the first is to adapt the bandwidth management procedures in WiMAX in order to achieve a Latency-Rate server (LR server) behavior [5]. The second idea is to make the bandwidth request-grant policies more flexible and simpler by mean of aggregation. This work is in progress, the framework is undergoing performance evaluation. The paper is organized as follows: we first present the different classes of service defined by the WiMAX standard [1] as well as the related bandwidth management mechanisms, we then give an overview of related literature. The third section details our framework proposal. The benefits of applying this framework are then discussed and a case study is presented. The last section concludes the paper and gives an overview of future work.

II. Bandwidth provision in WiMAX

A. Detailing the services

WiMAX specifies four different scheduling services to which are mapped all of the uplink connections. To each of those services is attached a set of rules specifying the request-grant policy that must be followed regarding an uplink connection. The services are:

- The Unsolicited Grant Service (UGS): supports flows generating a periodically fixed amount of Data to be sent.
- The Real-Time Polling Service (rtPS): supports dynamic flows that are real time sensitive.
- The non-Real-Time Polling Service (nrtPS): supports flows which are delay and jitter tolerant.
- The Best Effort service (BE): is a classical best effort service.

We detail in the following the request-grant policies attached to each of the services.

B. Bandwidth requests: the WiMAX way

Bandwidth requests are done in WiMAX on a per connection basis. several ways are available to allow a connection to request bandwidth or to imply needs. Upon establishing a connection, a specification of the flow using the connection is communicated to the BS. This specification can be considered as the first long term bandwidth request made by the connection. Other methods to request bandwidth (or to imply the need

to be polled) while the connection is alive are: unicast request opportunities, contention request opportunities, piggybacking and the Poll Me bit. We detail hereafter how each service uses the request methods.

a) UGS: A UGS connection is periodically granted air time without having to specifically request it. The amount of the grant is fixed upon set up of the connection, based on the Maximum Sustained traffic of the flow. A UGS connection is not allowed to use any contention based request period. If a UGS connection transmit depth queue is exceeded (due to a lost UL-MAP or due to clock mismatch) it sets in outbound packets the SI bit (Slip Indicator bit) informing the BS of the situation. The PM bit (Poll Me bit) in outbound UGS packets can be used to request polls for other non-UGS connections.

b) rtPS: A rtPS connection is provided with periodic unicast request opportunities (a unicast opportunity is a period where only the destined connection can express its needs). Those opportunities will be used by the connection to express its needs depending on its queue situation. A rtPS connection is not allowed to use contention request opportunities (which are periods where several connections may express their needs in a CSMA/CA fashion contention based access).

c) nrtPS: A nrtPS connection is provided with regular unicast request opportunities (the standard specifies an interval on the order of one second or less). A nrtPS connection can also use contention request opportunities.

d) BE: A BE connection may be granted unicast request opportunities by the BS. It may also use contention request opportunities in order to express its needs.

C. The grants

IEEE 802.16 SS medium access for data transmission (Uplink data transmission) is done in a contention-less, polling based fashion. Within the BS, a scheduling algorithm, that is not specified in the standard, is supposed to build the UL-MAP (map of the transmission opportunities granted for the uplink direction). The UL-MAP is supposed to be built based on the information the BS has about the active connections and on the requests received from the different connections (as specified earlier). Two grant modes were defined at different stages of standardization by the IEEE 802.16 workgroup: GPC (Grant Per Connection) and GPSS (Grant Per Subscriber Station). In GPC mode, the connections are individually granted data transmission opportunities by the BS; the UL-MAP specifies the time range each connection in each SS should use. In GPSS mode, transmission opportunity is granted to the SS; the SS would have the responsibility to schedule each of its connections in the granted time. Note that the GPC mode is obsolete, latest IEEE 802.16 standard [1] does not mention it.

D. Overview of the literature

The IEEE 802.16 standard chose not to give any specifications of the different QoS mechanisms to be used but only gave a general framework. Recent work concentrated on these aspects. Architectures focusing on the different uplink scheduling algorithms to be used for the different classes of service were given in [6], [7]. A QoS architecture was defined in [8] which instantiated the different blocs of a QoS architecture (admission control, classifiers, schedulers, traffic shaper and different queuing mechanisms). However few concentrated on the scalability issues. The first step in this direction was pulling the GPC mechanism out of the standard which was a serious flaw in terms of scalability. Other works [9] concentrated on the optimal duration of the contention period in order to reduce collision probabilities or on adapting Uplink/Downlink ratio to traffic shape [10]. Our work focuses on the bandwidth management mechanisms. We propose a redesign of the request mechanisms based on an aggregated management. Our design will offer rate guarantees and latency bounds to sensitive flows by adopting a Latency Rate [5] server behavior. The design will also provide a flexible and simplified bandwidth management procedure.

III. A NOVEL BANDWIDTH MANAGEMENT FRAMEWORK

A. Latency Rate servers

In [5], Stiliadis and Varma define a general model for traffic scheduling algorithms called Latency-Rate Servers (LR servers). We give here a quick overview of the benefits of the model, details and proofs are available in [5].
The LR server model gives conditions on ρ_i, $A_i(\tau,t)$ and $W_i(\tau,t)$ during a session i busy period whose starting time is τ [5]. ρ_i denotes the allocated rate to a session i, $A_i(\tau,t)$ is the arrivals from session i during the interval $(\tau,t]$. $W_i(\tau,t)$ is the service received by session i in the interval $(\tau,t]$. When those conditions apply during a busy period of session i, an $LR(\rho_i,\Theta_i)$ server will be able to offer guarantees on the rate ρ_i, bounds on the delays constrained by Θ_i and properties on the queue size necessary for the session. The guarantees on the rate and on the delays are the properties of interest in our case. The service provided by a WiMAX SS/Network can be made to act as an LR server towards a flow. Proof of this is not given here due to lack of space. What follows describes the necessary procedures leading to such a behavior.

B. WiMAX, an LR server

Let i be a given flow (which in our case consists in all time sensitive flows of an SS: UGS, rtPS and nrtPS). Let τ be the beginning of a backlog period of i. During this period, i's transmission queue is by definition never empty. The proposed system is based on a specific form of an **aggregated bandwidth request** built at time t_{rk} during the backlogged period:

$$BW\_Rq(t_{rk}) = \\ min\{Q_i(t_{rk}), max\{0, g_i(t_{rk}-\tau) - W_i(t_{rk}-\tau)\}\}$$

where $Q_i(t_{rk})$ is the amount of enqueued traffic for flow i, g_i is the long term rate we wish to guarantee to flow i. If all requests of this type are positively granted to flow i (property that must be guaranteed using a mandatory admission controller) then **the service offered to flow i respects an $LR(\rho_i,\Theta_i)$ model** (with $\rho_i = g_i$ and $\Theta_i = 2F\_T + UL\_F\_T$

F_T being the WiMAX frame time, UL_F_T being the uplink frame time). The consequences of which are guaranteed rates and delay bounds. The bounds on the delays can be reduced by applying anticipation, i.e. in the aggregated bandwidth request: instead of considering the amount of service due at the time of request ($g_i(t_{rk} - \tau)$) taking a later time into account. We call this **provisioning policies**.
The following section will detail each of the aspects of the framework and analyze it.

C. Specifying the architecture

1) An Admission Control Algorithm: The main aim of the admission controller is to make sure the provisioning as proposed is possible. Thus, in order for the behavior of WiMAX to be an LR server, admission control is mandatory. The different policies for the provisioning are detailed in III-C3. The admission control algorithm must adapt to the adopted provisioning policy in order to guarantee that all the requests made by the SSs respecting the allocated rate would be positively granted. The admission control algorithm must be applied to flows using UGS and rtPS services. It may also be applied to nrtPS flows if guarantees are to be given to such flows. It is useless to apply admission control to BE flows since no guarantees are to be given to such flows (this is obviously the definition of a best effort service), since the medium access in WiMAX is controlled and contention-less, new Best Effort flows will not affect other flows' performance. It may however be interesting, in order to avoid BE flows starvation, to reserve a part of the total bandwidth to such flows.

2) Aggregating the requests:

a) Specifying the aggregation: In the context we presented in introduction, an idea emerges: the provisioning is made in an aggregated-decentralized fashion (the grants being made in the GPSS mode) and the scheduling of the connections is made by the SS, the complexity of the request procedure proposed by WiMAX becomes avoidable (not useless but avoidable). Aggregating the requests will allow the SS a better flexibility in its bandwidth management and will allow (in the context where the number of connections per SS is high) to reduce the overhead caused by the multiple requests. The flexibility will come along with a reduced complexity of the procedure. The complex request methods (unicast polls, multicast polls, PM bit, piggybacking) will make place for a single, less complex and more flexible bandwidth management procedure. The newly built aggregated request must be sent in a contention-less fashion (as a specific UGS connection for example), this is necessary in order to reduce the loss probability of the request. The request must also be sent at the end of the uplink slot allocated to the SS (this way the SS has updated information about the state of its queues when sending the request). This will allow to reduce the contention period at the beginning of each frame to the strict minimum (reducing it to the initial ranging part). It is obvious that this does not respect the IEEE 802.16 standard [1]. We however think of the aggregation as a way to make the whole bandwidth management easier to apprehend and of an added flexibility. It is however necessary to note that the request aggregation is not a sine qua non condition for the architecture to work.

b) Format of the request: The single aggregated request per SS will contain (in addition to the usual headers) two fields representing the aggregated uplink needs of all the SS's connections. The first field is called the "contracted bytes"; the second is called the "additional bytes". The exact content of each field will differ following the provisioning policy adopted (this will be explained in III-C3) however, conceptually, and as the name of each field indicates, the first field will contain information on the needed bandwidth that is within the contracted allocated rate (as per the provisioning policy) and the second will contain the needs of the station that do not go within the contracted allocated rate. The first field will thus contain, as per the stated condition in III-B, the bandwidth that must be allocated to the SS in order for the system to work properly.

3) Provisioning policies: We define three provisioning policies which are mainly based on the degree of anticipation in a request. The degree of anticipation in the request-provisioning policy will have a direct effect on the bounded delays that the LR server will be able to offer. We consider the policies from the aggregated requests' point of view: meaning that the policies will differ with regards to what qualifies to be considered as falling within the "Contracted Bytes" (or CB in the following) part of the aggregated request, or within the "Additional Bytes" (AB in the following) part of it.
In addition to the terminology defined in III-A (mainly ρ, Θ and W), we define tx_k as the beginning of the transmission time allocated to the SS in k^{th} frame , tx_{k+1} is the possible beginning of transmission time of the SS in the following frame. We also define tr_k as the time of transmission of the SS's aggregate request in the current frame (which will usually be at the end of the uplink transmission time allocated to the SS as specified by the UL-MAP), tr_{k+1} is the possible time of aggregate request transmission by the SS in the following frame. Q_{UGS} is the amount of UGS bytes enqueued within the SS at the moment the request packet is being built, the same applies to Q_{rtPS}, Q_{nrtPS} and Q_{BE}. ε_{UGS} is the slip amount for the UGS queue: it is the value indicating the amount of needed bandwidth (considering the state of the queue) that goes beyond the contracted request: $\varepsilon_{UGS} = max(Q_{UGS} - R_{UGS}, 0)$, R_{UGS} being the total amount of UGS bytes falling within the contracted request. The same applies to ε_{rtPS} and ε_{nrtPS}.
A general formula of the anticipation policies would be:

$$CB = (\rho(t) - W(tr_k)) + \varepsilon_{UGS}$$
$$\text{with } \rho(t) = \sum \rho_{UGS} + \sum \rho_{rtPS} + \sum \rho_{nrtPS}$$
$$\text{and } AB = \varepsilon_{rtPS} + \varepsilon_{nrtPS} + Q_{BE}$$

The value of t defines the anticipation degree. We instantiate t to three interesting values giving three different policies:

- **No anticipation** In which the request will only consider the amount of service due at the time of build of the request: $t = tr_k$

- **Simple anticipation** In which the request will consider the amount of service due at the following request time: $t = tr_{k+1}$
- **Far anticipation** In which the request will consider the amount of service due one further transmission time: $t = tx_{k+2}$

The "No anticipation" policy and the "Simple anticipation" policy are illustrated in figure 1.

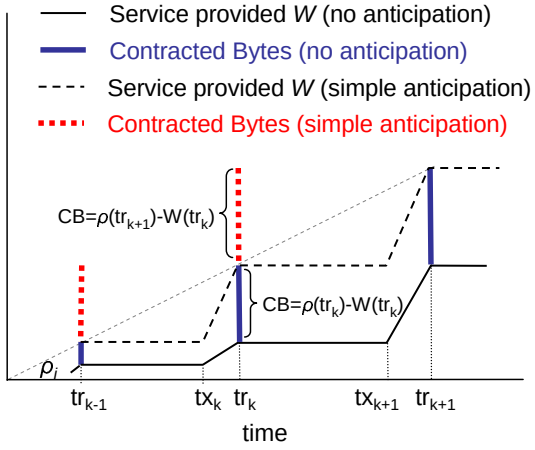

Fig. 1. No anticipation and simple anticipation policies

a) Flexibility in bandwidth management: The property of aggregation of bandwidth management (i.e. by aggregating on the one hand the requests and on the other hand the grants) of the designed bandwidth management implies additional flexibility. This flexibility is the main reason why the proposed policies with anticipation do not necessarily mean a bandwidth waste. The bandwidth allocated to the SS based on the request it made can thus be used to serve different flows including BE flows.

b) Attached admission control: As said earlier, the admission control algorithm is closely connected to the provision policy chosen by the network administrator. Depending on the chosen policy, the guarantees given to the flows will differ and so will the admission condition. The admission control algorithm will be activated only for UGS, rtPS and nrtPS flows. By limiting the air time dedicated to the flows with guarantees (UGS, rtPS and nrtPS) to only a specified percentage α (80 % for example) , the admission controller will be able to avoid starvation of BE flows without submitting them to a strict control.

The admission controller will base its decision of admission of a new flow on available air time and the total contracted bytes accorded to each station in the chosen provisioning policy. When a new flow submits an admission request, the following condition should be true in order for the new flow to be accepted:

$$\sum_{stations} (normalized\_\rho(F\_T) + SSTG) < \alpha * UL\_F\_T$$

with $normalized\_\rho$ being the necessary time for transmission of the requested rate at the station's transmission profile and SSTG being the Subscriber Station Transition Gap. We thus have $normalized\_\rho(F\_T) = \frac{\rho_i * F\_T}{StationTxRate}$.

IV. Analysis and Discussion

In this section, we analyze the different properties of the system as a whole: the designed system will inherit the properties of an LR server which will first be exposed. We will then present the different qualitative properties of the designed system and give a case study illustrating the different properties.

A. Latency-rate server properties

With the properties of an LR sever insured by the WiMAX bandwidth management system as described earlier, several properties are directly inherited: mainly guarantees on the delays. As described in section III-A and as can be clearly seen in figure 1we can give a worst case scenario bound for delays for each of the provision scenarios, the worst case scenario being a peak arrival within contract right after the request has been sent: in a no anticipation policy a service conforming to the contracted rate starts no later than $3F\_T$. This delay falls to one F_T with the simple and far anticipation policies.

B. Other properties of the designed system

1) Number of requests: The designed mechanisms allow reducing the overhead caused by the number of sent requests per SS. An SS in a standard WiMAX bandwidth management will have to send several requests per Super Frame, this is reduced to one request per Super Frame per SS in the designed architecture.

2) Increased determinism: In our proposal the single aggregated request of the SS is sent in a contention-less fashion, unlike standard WiMAX architecture where some bandwidth requests can be sent in contention zones. This property of the new design increases the determinism of the WiMAX bandwidth management.

3) Flexibility and simplicity: The aggregation of the requests and of the grants will allow a better flexibility of the bandwidth management. It will also render the whole request mechanisms simpler. The SS will have a better knowledge of the state of its queues, this will allow it to adapt the request to its needs and to organize the scheduling accordingly.

C. A case study

We propose, in order to show the different properties brought by the designed system of bandwidth management a case study. We place ourselves in the context presented earlier: a WiMAX network serving WiFi hotspots. We present the scenario of our case study and analyze some positive aspects of the proposal.

1) The scenario: Figure 2 describes the case study scenario we are about to analyze. A WiMAX BS acts as a part of a backbone element serving WiFi hotspots. The WiFi hotspots have in fact both a 802.11 and a 802.16 interface, acting thus as a WiFi access point and a WiMAX SS. Each WiFi access point serves several stations with different types of service

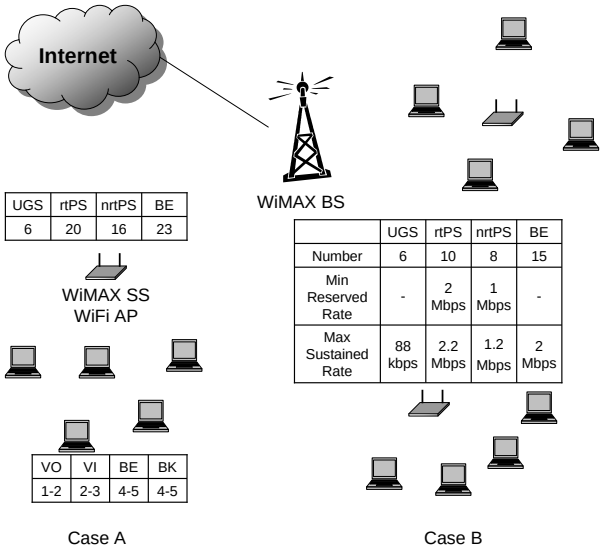

Fig. 2. Case study scenario: in case A we give a view of the number of flows per 802.11e access category and the total number of connections going through the WiMAX SS, Case B gives an overview of the specifications of the different flows

(applications range from time sensitive voice applications to highly loaded ftp applications). Each application requesting access to the network is considered, at the WiMAX level, as a connection belonging to a service type respecting the specification of the application. We thus consider that each SS-Access Point will have at any time of the network's lifetime several WiMAX connections belonging to each type of service. Details on each specific case can be seen in figure 2.

2) The benefits: We can analyze the benefits we get from the design in this scenario.

- In case A, we analyze the number of bandwidth requests sent out by SS. If we consider as an example the number of applications given in figure (2,A) we can clearly see the benefits of our architecture in terms of reduced overhead. In a classical WiMAX architecture, several bandwidth requests will have to be sent in order for the connections to express their needs (say about 20 requests per frame in our case). This number is reduced in our scheme to only 1 aggregated request per Frame. The complexity of the procedure is also reduced, for example, in standard WiMAX nrtPS connections will sometimes have to ask for a poll in order to be able to send a bandwidth request. This 3 way procedure and alike complex bandwidth request procedures are now reduced. The architecture will also allow reducing the length of the contention period present at the beginning of each frame.
- WiMAX makes use of contention based bandwidth request periods in order to allow nrtPS and BE connections to express their needs in a highly loaded scenario as may be case A. This will cause possible collisions. Bandwidth requests in our case are sent in a contention-less fashion. There are no risks of collision between the requests and thus the determinism of the request procedure is increased.
- Case B will help the reader understand the benefits of the aggregation in our proposal (i.e. including the LR server behavior and the anticipation policies). When having flows as those shown in figure (2,B) with a considerable variability in the rates, anticipating the needs as explained in paragraph III-C3 will allow reducing the delay bounds for time sensitive flows. One may think this anticipation may cause a waste of bandwidth (when the anticipated needs do not come true). In our case, the bandwidth management being flexible, the SS will be able to serve the high number of BE flows instead of wasting the allocated bandwidth.

V. Conclusions

We propose in this paper a bandwidth management scheme for IEEE 802.16 WiMAX. This scheme is based on an aggregated bandwidth management and an *LR* server like behavior. The architecture gives the system interesting properties among which are delay bounds and rate guarantees for time sensitive flows. The bandwidth management is simplified and made flexible. A case study has been proposed in order to highlight the different properties of the designed system. The system's implementation is ongoing, future work will include the implementation of the system and its evaluation by means of simulation.

References

[1] 802.16, *IEEE Standard for Local and Metropolitan area Networks Part16 : Air Interface for fixed broadband Wireless Access Systems*, 2004.

[2] C. Eklund, R. Marks, K. Stanwood, and S. Wang, "Ieee standard 802.16: a technical overview of the wirelessman air interface for broadband wireless access," *Communications Magazine, IEEE*, vol. 40, no. 6, pp. 98–107, Jun 2002.

[3] M. Marchese and M. Mongelli, "Optimal bandwidth provision at wimax mac service access point on uplink direction," *Communications, 2007. ICC '07. IEEE International Conference on*, pp. 80–85, 24-28 June 2007.

[4] W. Forum, *Business Case Models for Fixed Broadband Wireless Access based on WiMAX Technology and the 802.16 Standard*, 2004.

[5] D. Stiliadis and A. Varma, "Latency-rate servers: a general model for analysis of traffic scheduling algorithms," *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, pp. 611–624, 1998.

[6] Q. Liu, X. Wang, and G. Giannakis, "A cross-layer scheduling algorithm with qos support in wireless networks," *Vehicular Technology, IEEE Transactions on*, vol. 55, no. 3, pp. 839–847, May 2006.

[7] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in ieee 802.16 broadband wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 6, pp. 668–679, June 2006.

[8] J. Delicado, L. Orozco-Barbosa, F. Delicado, and P. Cuenca, "A qos-aware protocol architecture for wimax," *Electrical and Computer Engineering, 2006. CCECE '06. Canadian Conference on*, pp. 1779–1782, May 2006.

[9] S.-M. Oh and J.-H. Kim, "The analysis of the optimal contention period for broadband wireless access network," *Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on*, pp. 215–219, March 2005.

[10] C.-H. Chiang, W. Liao, and T. Liu, "Adaptive downlink/uplink bandwidth allocation in ieee 802.16 (wimax) wireless networks: A cross-layer approach," *Global Telecommunications Conference, 2007. GLOBECOM '07. IEEE*, pp. 4775–4779, Nov. 2007.