This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

Optimal Bandwidth Provision at WiMAX MAC Service Access Point on Uplink Direction

Mario Marchese, *Senior Member*, *IEEE*, Maurizio Mongelli, *Member*, *IEEE*
DIST - Department of Communication, Computer and System Sciences
University of Genoa
Via Opera Pia 13, 16145, Genoa (Italy)
{ Mario.Marchese, Maurizio.Mongelli }@unige.it

***Abstract*—In this paper, the IEEE 802.16 protocol is investigated with respect to the bandwidth provision problem arising at the *Medium Access Control* (MAC) layer. The aim is to optimally tune the resource allocation to match QoS requirements. Traffic flows are originated at network layers overlying the 802.16 protocol stack. This leads to the investigation of a novel control algorithm, suited to optimal bandwidth allocation and *Call Admission Control* in the presence of statistically heterogeneous flows. Specific implementation details are provided to match the application of the control algorithm using the regular 802.16 request-grant protocol. Simulation results validate the proposed approach.**

Keywords-802.16 MAC layer, Cross-layer Quality of Service, Dynamic Bandwidth Allocation, Neural Feedback Control.

I. Introduction

The IEEE 802.16 standard [1], also known as *Worldwide Interoperability for Microwave Access* (WiMAX), is the emerging technology for broadband wireless access. It is considered as the ultimate solution for *Quality of Service* (QoS) delivery in wireless infrastructures. Several economical studies foresee the application of WiMAX infrastructures in different market segments and business areas in the next few years [2].

The WiMAX architecture specifies the composition of a *Wireless Metropolitan Area Network* (W-MAN), which consists of a *Base station* (BS) and some *Subscriber Stations* (SS). The 802.16 standard allows implementing QoS guarantees in both uplink (from SS to BS) and downlink (from BS to SS) directions. A TDMA (*Time Division Multiple Access*) is used to access the uplink, while a plain TDM (*Time Division Multiplexing*) is implemented for downlink.

WiMAX standard does not specify QoS implementations details. Classification, resource allocation and scheduling of packets are left to the choice of implementers. The QoS-based services defined at the network layer (IP, ATM) are mapped into the WiMAX MAC core through a proper interface or SAP (*Service Access Point*) [1]. The SAP introduces specific cross-layer considerations [3-7] (also known as *QoS Mapping*) that are currently hot topics of research related to open standardization activities in other environments [8, 9]. Actually, the main problem is the relation between upper layers, where groups of flows characterized by different performance requirements may be stored in differentiated buffers, and the MAC layer, where the number of available queues is necessarily lower for implementation reasons.

The problem is then providing each single MAC queue with the necessary bandwidth to satisfy the requirements of all the traffic flows conveyed in it. In this perspective, a novel control algorithm able to capture the "bandwidth need" of the different flows conveyed within the WiMAX MAC core is investigated.

The remainder of the paper is organized as follows. The characteristics of the WiMAX MAC layer are summarized in the next section. Section III contains the state of the art on WiMAX QoS. An introduction to the bandwidth allocation problem, used for QoS mapping through layers, is detailed in section IV. The formalization of the novel functional optimization approach provided by this paper is reported in section V while the related bandwidth allocation solution is provided in section VI. The performance evaluation, obtained by simulation analysis, is the object of section VII. The conclusions and the directions for future research are contained in section VIII.

II. The WiMAX Medium Access Control Layer

WiMAX environment consists of a central radio *Base Station* (BS) and a number of *Subscriber Stations* (SS). A SS typically covers a single residential or business building. BS is connected to public networks via cable fiber. The BS transmits a TDM signal, where the time slots are allocated serially for single SS. Uplink sharing is ruled through TDMA. Both time division duplexing, where the uplink and downlink share a channel but do not transmit simultaneously, and frequency-division duplexing, in which the uplink and downlink operate on separate channels, sometimes simultaneously, are allowed [10].

Even if WiMAX physical layer is currently a hot topic of research involving *spatial multiplexing*, *hybrid ARQ*, *interference cancellation* and *power allocation* [11], no further detail is given about it because the focus is on the WiMAX MAC features.

A. The MAC Protocol

The MAC layer is composed of three sublayers. from bottom to top: the *Security Sublayer* (PS), the *MAC Common Part Sublayer* (CPS), and the *Service Specific Convergence Sublayer* (CS) The former deals with security and network access authentication procedures. CPS carries out the key MAC functions. It is connection oriented. The CS sublayer provides the interface to the upper layers, decides the MAC service class for the specific connection and initializes the resource allocation requests of the CPS.

A MAC connection is identified by a 16-bit *Connection Identifier* (CID). The *MAC Protocol Data Unit* (M-PDU) is the data unit exchanged between the MAC layers of BS and SS. The CS sublayer receives external network *Service Data Units* (SDUs) through the *CS-Service Access Point* (CS-SAP), and associates them to the proper MAC service flow and CID. In practice, the IP packets are conveyed to CPS queues after CS filtering. MAC CPS receives the data and encloses it in the MAC PDU to send it to the destination. MAC PDU consists of a fixed length header, a variable length payload and an optional *Cycle Redundancy Check* (CRC). Two types of headers are standardized: *generic headers* (GH), to send MAC management messages and CS data and bandwidth request headers. Additionally, MAC PDU may contain different types of subheaders: the Grant Management subheader, used by a SS to request bandwidth to the BS, topical to implement the control scheme proposed in this paper, and the packing and fragmentation subheaders, related to packing (multiple SDUs into a single MAC PDU) and fragmentation functionalities.

B. The MAC Services

Each packet traversing the MAC interface in the uplink direction is mapped to a *Scheduling Service* (SCS), which is associated with a set of rules imposed by the BS "*responsible for allocating the uplink capacity and the request grant protocol between the SS and the BS*" [10]. A set of QoS parameters is also associated to a SCS. During the connection set up phase, the SCS is chosen and activated if sufficient resources are available. A unique CID is assigned to all activated connections of a given SCS.

Four SCS are defined. The *Unsolicited Grant Service* (UGS) is designed for CBR-like real time services, such as *Voice over IP* (VoIP) without silence suppression, ATM CBR and SDH E1/T1 over ATM. The BS schedules a fixed size data grant periodically, without an explicit request from any SS. The *Real time Polling Service* (rtPS) is dedicated to real time bursty traffic, dynamic in nature, such as VoIP (with silence suppression) and real time streaming audio-video [10]. The *Non real time Polling Service* (nrtPS) is related to non-real time bursty traffic with some QoS guarantees (e.g., the aggregation of FTP or Web connections [10]). The *Best Effort* (BE) service is designed to support regular Internet BE traffic.

III. WIMAX QOS: STATE OF THE ART AND QOS MAPPING

While extensive signaling and bandwidth request mechanisms are provided in the standard, details of scheduling and reservation management are not standardized, thus allowing vendors to differentiate their equipments. In other words, 802.16 standards do not suggest how to schedule the packets to meet QoS requirements but fixes the protocol features that can help it. [13] proposes architecture for dynamic bandwidth allocation. It implements traffic policing and exploits WRR (*Weighted Round Robin*) and priority scheduling algorithms for downlink and uplink service differentiation, respectively. In [14], an uplink scheduling differentiation mechanism is proposed, based on the GPC mode. A QoS scheduling architecture is also studied in [12, 15] to provide QoS guarantees to WiMAX applications. [5, 16] extend the QoS features of the standard through traffic policing and *Call Admission Control* (CAC).

The mentioned works propose solutions for QoS management by heuristically matching the QoS mapping operations between MAC and upper layer. This paper, on the other hand, goes deep into QoS mapping optimization by considering the peculiar characteristics of the SS requests in dependence of the bandwidth need resulting from the aggregation (*traffic grooming* in the [5] terminology) of SS connections. Taking the architectures in [5, 12, 16] as a reference, a limited number of MAC queues conveying traffic characterized by a large set of QoS requirements is used.

IV. THE UPLINK BANDWIDTH MANAGEMENT PROBLEM

As outlined above, the data flow traversing the 802.16 SS in uplink direction may be modeled as a cascade of buffers implemented, respectively, at IP layer and at MAC layer [5, 12, 16]. In shorts, the data packet are stored in IP queues and, after filtering by the CS-SAP, are sent to MAC CPS buffers through the MAC-SAP (*MAC – Service Access Point*) interface. Fig. 1 shows a possible example where the IP traffic, composed of video and voice packets (differentiated into two IP queues), is conveyed to one single MAC buffer. The upper layers (IP, in this case) are unaware of the local implementation of the QoS management within the MAC queues. Each single queue at MAC layer is representative of a specific SCS (*Scheduling Service*) and can benefit of a proper request-grant protocol (UGS, rtPS, nrtPS, BE) for bandwidth management, as specified in section III. The idea is that the MAC layer receives specific service requirements (called QoS_i^*, in the reminder of the paper) from the IP layer differentiated for traffic classes (the index *i* identifies an IP traffic class) and must assign to the MAC CPS queue server (the SCS) sufficient bandwidth to guarantee all the required service requirements. QoS_i^* levels flow from IP to MAC through the CS-SAP interface, which is the access point of IP to the services offered by MAC. The CS is then responsible for service requests towards the MAC CPS through the MAC-SAP (see the Annex C of the standard [1] for details). Actually, the bandwidth allocation performed at the MAC CPS buffers is hided to upper layers because it is a sublayer concern. In this view, the aim of the paper is to develop a control methodology to be implemented at the MAC-SAP to support resource allocation.

As outlined in [5-9], the general concept of SAP leads to specific cross-layer management issues: the QoS paradigm applied at the upper layers must be "mapped" into the 802.16 MAC queues so that applications are unaware of the QoS protocols changes. More specifically, the MAC-SAP leads to the following problems: **1)** encapsulation change: QoS must be assured within the WiMAX system, despite the change of encapsulation format and the adoption of specific *fragmentation* and *packing* procedures locally implemented at the MAC level; **2)** traffic aggregation.

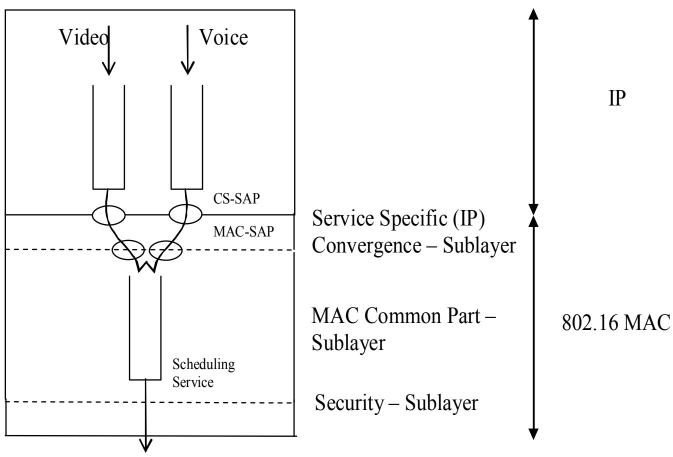

Figure 1. Data flow model for the *Subscriber Station* (SS) uplink [5, 16].

In more detail concerning the latter, the traffic coming from the IP layer needs to be aggregated within the SS MAC queues, thus generating heterogeneous trunks from the QoS requirement viewpoint. For example, a given SCS (say, rtPS) queue may see the aggregation of different real time traffic categories (such as VoIP and video) within a single MAC buffer. The mentioned traffic categories surely have different QoS requirements. In the specific WiMAX architecture the aggregation process may take place: **1)** when generating a connection before it enters one of the SS uplink queues (some applications may be multiplexed together in a single connection), **2)** when connections are multiplexed within SS uplink queues [12, 16] or **3)** when flows coming from different SSes at BS are mixed in the BS uplink queues [12, 16].

As a consequence, the concept of *Equivalent Bandwidth* (EqB) (usually defined as the *minimum bandwidth allocation necessary to guarantee a specific QoS to a traffic flow*), is generalized, since what is needed here is the minimum bandwidth provision that satisfies all the QoS levels required by the different classes aggregated in the same uplink trunk. Many studies confirm the efficiency of aggregating homogeneous traffic, but the performance of non-homogeneous trunks (from the statistical behaviour and QoS requirement viewpoints) is still an open issue [18].

V. THE MATHEMATICAL MODEL

To match optimal bandwidth provision in the presence of the "generalized" concept of EqB, a proper optimization framework capturing the concept of cross-layer QoS mapping is developed. From now on, the QoS is expressed in terms of IP metrics: *Packet Loss Probability* (PLP), *Average Delay* (AD), and *Delay Jitter* (DJ) of IP packets [21].

It is important to define the *QoS observation horizon T* [22], as the time interval during which the QoS levels actually achieved for a specific flow are monitored. For example, $T = 1\,\text{m}$ for PLP<1% means that during each period of 1 minute, the averaged PLP must be lower than 1%.

Just one MAC buffer within an SS is considered (as in Fig. 1) for the sake of simplicity. The structure may be simply duplicated for each MAC queue.

Let $\alpha_i(t)$ be the *stochastic input rate process* coming from the buffer of an IP service class $i = 1,...,N$ (for instance, the voice service in Fig. 1), at a give time instant t, and entering the WiMAX system through the MAC-SAP. Up to N IP service classes may be aggregated together within the MAC buffer. Let $\boldsymbol{\alpha}$ be the aggregate vector of all input rate processes α_i, $i = 1,...,N$, namely, $\boldsymbol{\alpha} = col\{\alpha_1,...,\alpha_N\}$. The overall input rate process of the MAC buffer, denoted by α_{MAC}, at a given time instant t, is obtained as $\alpha_{MAC}(t) = \sum_{i=1}^{N} \alpha_i(t)$. The MAC buffer serves the queued traffic according to one of the mentioned SCS (i.e., UGS, rtPS, nrtPS, BE).

A sequence of $\alpha -$ *observation time horizons* (different from the QoS observation horizon T defined above), where α_{MAC} is monitored is defined. A new bandwidth request may be performed by the SS to the BS at the end of each $\alpha -$ observation time horizon. The specific protocol used for bandwidth requests will be detailed later.

Let $\hat{t}$ be the duration of the $\alpha -$ observation time horizon. Let θ be the service rate of the mentioned MAC buffer. New service rate reallocations are performed for $t = k\hat{t}$, $k = 1, 2, ...$. Let:

$$\boldsymbol{I}(k\hat{t}) = col\{\alpha_{MAC}((k-\Xi)\hat{t}),...,\alpha_{MAC}((k-1)\hat{t}))\} \quad (1)$$

be an aggregate vector that maintains a finite horizon memory (of depth Ξ) over the values assumed by α_{MAC} during the time interval $[(k-\Xi)\hat{t},(k-1)\hat{t}]$. Note that $\hat{t}$ denotes also the reallocation period.

Let $J_{k\hat{t}}(\theta(k\hat{t})) = \underset{\boldsymbol{\alpha}}{E}\left\{\text{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\theta(k\hat{t}),\boldsymbol{\alpha}\right]\right\}$ be the functional cost by considering the bandwidth reallocation performed at time $k\hat{t}$ and a QoS observation horizon T (beginning at time $k\hat{t}$) to monitor QoS parameters. The quantity $\text{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\theta(k\hat{t}),\boldsymbol{\alpha}\right]$ is defined in (2):

$$\text{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\theta(k\hat{t}),\boldsymbol{\alpha}\right] = \sum_{i=1}^{N}\left(\text{QoS}_i^{[k\hat{t},k\hat{t}+T]}\left[\theta(k\hat{t}),\boldsymbol{\alpha}\right] - \text{QoS}_i^*\right)^2 \quad (2).$$

The function $\text{QoS}_i^{[k\hat{t},k\hat{t}+T]}\left[\theta(k\hat{t})\right]$ represents the QoS of the IP service class i actually measured within the MAC queue according to the bandwidth reallocation $\theta(k\hat{t})$ and to the current realization of the stochastic processes $\boldsymbol{\alpha}$ in the time period $\left[k\hat{t}, k\hat{t}+T\right]$. QoS_i^* is the desired QoS performance level for the service class i, which is transmitted from the IP to the MAC layer through the MAC-SAP interface. In practice, the MAC layer offers a service to the IP layer fixed by an agreement expressed in terms of objective performance metrics, i.e., PLP, AD, and DJ.

Let $f(\boldsymbol{I}(k\hat{t}))$ be a reallocation law, which provides the service rate reallocation $\theta(k\hat{t})$ of the buffer as a function of the current information vector:

$$\theta(k\hat{t}) = f(\boldsymbol{I}(k\hat{t})) \tag{3}$$

The bandwidth provision problem for the uplink MAC buffer can now be stated.

WiMAX Functional Resource Allocation Problem – ***WFRAP***: it finds the optimal bandwidth reallocation function $f^*(\cdot)$, such that the cost:

$$J_{k\hat{t}}(\theta(k\hat{t})) = \underset{\boldsymbol{\alpha}}{E}\left\{\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}\left[f(\boldsymbol{I}(k\hat{t})),\boldsymbol{\alpha}\right]\right\} \tag{4}$$

is minimized.

VI. The Control Algorithm

A. The Extended Ritz method

In order to approximate the optimal resource allocation law $f^*(\cdot)$, a modification of the *Extended Ritz* method [23] is applied. The Extended Ritz method approximates the solution of a functional optimization problem by fixing the structure of the decision functions. Among the possible form choices of the decision functions, this paper uses a *feedforward neural network* (NN) (with a single scalar output). It is denoted by $\bar{f}(\boldsymbol{I},\boldsymbol{w})$, being $\boldsymbol{I}$ the input of the NN and $\boldsymbol{w}$ the NN weights to be optimized. The scalar output of the NN, denoted by $\bar{\theta}$, is obtained as:

$$\bar{\theta} = \bar{f}(\boldsymbol{I},\boldsymbol{w});\ \bar{\theta}\in[0.0,1.0] \tag{5}$$

$\bar{\theta}\in[0.0,1.0]$ since sigmoid functions are chosen for the NN output layer. The service rate is constrained to a given domain, i.e., $\theta\in[0,MaxBw]$, where $MaxBw$ is the maximum available bandwidth for the MAC buffer under study. In order to guarantee the fulfillment of the constraint, a *normalization operator* $n[\cdot]$ is applied to the output of the neural network.

$$\theta(k\hat{t}) = n\left[\bar{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w})\right];\quad n(x) = MaxBw\cdot x; \tag{6}$$

The composition $n\left[\bar{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w})\right]$ of the neural approximation $\bar{f}(\boldsymbol{I},\boldsymbol{w})$ and of the normalization operator $n[\cdot]$ is identified as $\hat{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w})$ and is called *neural bandwidth allocation function* (NBAF). It follows that a cost function is obtained by substituting the structure of the NBAF into the cost in (4), which now depends on the parameter vector $\boldsymbol{w}$. It leads to the mathematical programming problem defined below.

Problem ***WFRAP$_w$***: it finds the optimal parameter vector $\boldsymbol{w}^*$ such that the cost:

$$\underset{\boldsymbol{\alpha}}{E}\left\{\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\hat{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w}),\boldsymbol{\alpha}\right]\right\} \tag{7}$$

is minimized. In this way, the functional optimization problem ***WFRAP*** has been reduced to an unconstrained nonlinear programming one.

B. The training algorithm

To solve ***WFRAP$_w$***, a *stochastic approximating* gradient-based algorithm of the form:

$$\boldsymbol{w}^{h+1} = \boldsymbol{w}^h - \zeta_h\,\nabla_{\boldsymbol{w}}\,\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\hat{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w}^h),\boldsymbol{\alpha}^h\right],\ h=0,1,\ldots \tag{8}$$

is applied, where the index h denotes both the steps of the iterative procedure and the generation of the h-th realization of the stochastic processes $\boldsymbol{\alpha}$. The components of the gradient $\nabla_{\boldsymbol{w}}\,\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}\left[\hat{f}(\boldsymbol{I}(k\hat{t}),\boldsymbol{w}^h),\boldsymbol{\alpha}^h\right]$ can be obtained by applying the regular *backpropagation equations* used to train neural networks. The backpropagation procedure must be initialized by means of the quantities $\dfrac{\partial\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}}{\partial\bar{\theta}},\ i=1,\ldots,N$ (i.e., the gradient $\nabla_{\bar{\theta}}\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}(\theta,\boldsymbol{\alpha}^h)$). Unfortunately, as outlined above, such quantities cannot be obtained analytically, because no closed form is available for the functional cost. The gradient $\nabla_{\bar{\theta}}\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}(\theta,\boldsymbol{\alpha}^h)$ is then estimated by means of *Infinitesimal Perturbation Analysis* (IPA) as in [23].

C. Control algorithm and grant request protocol

The NBAF (introduced previously for just one MAC queue) may be used to tune the bandwidth of all MAC queues (i.e., all the SCSes). The *Grant Management Subheader* (GMSH) is exploited to this aim. The GMSH is a lightweight way to attach a request of uplink bandwidth, without the need of transmitting a complete MAC PDU. A possible use of NBAF by using the features of the WiMAX MAC protocol is reported in the following for the grant services.

If the CID in the GH (Generic Header) indicates that a channel is using UGS, only two bits of GMSH are used by the standard. The *slip indicator* (SI) bit is used by the SS to inform the BS that the rate of arrival of the data to be sent is faster than the granted uplink rate. It acts as a request to the BS to make additional uplink grants. A portion of the 14-bit left unused by the standard for UGS might be used to transfer information about the current state of the UGS uplink buffer to BS. The information would be expressed in terms of the mentioned information vector $\boldsymbol{I}(\cdot)$ defined in (1). The NBAF will be then located in the BS and is used to infer the next bandwidth grant for the UGS of a given SS.

In the case of any other SCS (rtPS, nrtPS), of main interest for the control scheme presented, the GMSH uses a slightly different format to piggyback bandwidth grant to BS. The piggyback request is composed by a 16-bit number that explicitly represents the number of uplink bytes being requested by SS for the specific buffer. In this case, the NBAF should be locally implemented within the SS and directly computes the next request to be sent to the BS.

VII. PERFORMANCE EVALUATION

To test the proposed control methodology, a C++ simulator has been develop for the IP and WiMAX MAC queues, having in mind the aggregation architecture shown in Fig. 1. A single rtPS MAC buffer is considered. A heterogeneous trunk of VoIP and video traffics is considered. Both VoIP and video sources are injected together in the WiMAX core (as in the case shown in Fig. 1).

VoIP sources are modeled as an exponentially modulated on-off process, with mean on and off times (as in the ITU P.59 recommendation) equal to 1.008 s and 1.587 s, respectively. When in the active state, they are 16.0 kbps flows over RTP/UDP/IP. The VoIP packet size is 80 bytes. The VoIP QoS targets are PLP=0.01 (1%) and AD=30 ms. The arrival frequency λ_{VoIP} is exponentially distributed with average 3 calls per minute, being the rtPS buffer supposed to be the aggregation point of all voice applications of the SS. The average call duration, μ_{VoIP}, log-normal distributed, is 10 minutes.

As far as the video service is concerned, real traces taken from [24] have been used. Data are H.263 encoded and have an *average bit rate* ($\overline{B}_{Video}$) of 260 kbps and a *peak bit rate* (B^{p}_{Video}) ranging from 1.3 to 1.5 Mbps, depending on the specific trace. Each video trace lasts about 1 hour. The video QoS targets are PLP=0.01 (1%) and AD=20 ms. The QoS observation horizon T lasts 5 minutes.

The *Packet CS* encapsulation format ([1], pag. 20) of VoIP and video packets (IP SDU) is implemented without header suppression. The MAC overhead corresponds to 48-bit, due to the GH (Generic Header) without CRC. The MAC payload is the IP SDU. Fragmentation of video SDUs and packing of SDU data ([1], pag. 125) are applied to generate M-PDU payloads of 1400 bytes each. It requires the addition of the 24-bit packing subheader. Thus, an M-PDU payload is composed of a 1000-bytes video SDU fragment and 5 VoIP SDUs. Such payload size is true on average, since a small amount of M-PDUs contain the last fragment of video SDU (ranging from 100 to 800 bytes). As outlined before, packing and fragmentation, together with statistical heterogeneity of the α_{MAC} process, make the analytical derivation of the QoS "seen" by a specific flow an impracticable task.

Two different NBAFs are used, the first one (denoted by ${}^{PLP}\hat{f}(\boldsymbol{I}(\cdot), {}^{PLP}\boldsymbol{w}^{*})$) is trained for the solution of problem $\boldsymbol{WFRAP_{w}}$ where the QoS_{i}^{*} targets in (2) are the PLPs defined for VoIP and video, respectively (i = VoIP, Video). The second one (denoted by ${}^{AD}\hat{f}(\boldsymbol{I}(\cdot), {}^{AD}\boldsymbol{w}^{*})$) is trained with respect to the AD constraints.

The $\alpha-$ observation horizon $\hat{t}$ is set to 30 seconds and the depth Ξ of the time horizon of the information vector $\boldsymbol{I}(\cdot)$ is 5 for both NBAFs. Each NBAF is implemented by a feedforward neural network with 20 hyperbolic tangent neural units in the hidden layer and with a sigmoid output layer. The optimal parameters' vector $\boldsymbol{w}^{*}$ characterizing each NBAF is obtained off line by means of the stochastic gradient technique described in subsection VI.B. The simulation scenario for NBAF training implies the generation of traffic samples coming from the aggregation of the VoIP flows together with a video trace and according to the chosen calls statistics ($\lambda_{VoIP}, \mu_{VoIP}$). As to further details on the training procedure (whose simulation time in this case took around 7.6 hours with an AMD Athlon @1.8 GHz), the reader is refered to [23].

After training, the NBAFs performance is tested over a simulation horizon of about 7 hours. Since the most stringent QoS requirement is not known a priori, bandwidth reallocations are driven in real time by both the trained NBAFs according to:

$$\theta(k\hat{t}) = Max\left\{ {}^{PLP}\hat{f}(\boldsymbol{I}(k\hat{t}), {}^{PLP}\boldsymbol{w}^{*}),\ {}^{AD}\hat{f}(\boldsymbol{I}(k\hat{t}), {}^{AD}\boldsymbol{w}^{*}) \right\},\ k = 1,2,... \quad (10)$$

A different video trace is used in each repetition of the simulation scenario.

Fig. 2 shows the allocations obtained by (10) during the system evolution. Figs 3 and 4 show the PLP and the AD of video measured at the MAC buffer, respectively, in dependence of different video traces. Each point represents the performance metric averaged over the last QoS observation horizon T (5 minutes). The tags "Jurassic", "MrBean" and "Silence" mean the adoption of "Jurassic Park", "Silence of the lambs" and "Mr. Bean" traces, respectively. The straight line in Figs 3 and 4 denotes the QoS video targets (PLP=10^{-3}, AD=20 ms). The changes in the service rate are due to the time varying number of active VoIP calls whose variation is highlighted in Fig. 5 with respect to the allocations for the "Jurassic" case. The quick response to traffic changes and the maintenance of the QoS are outstanding. Only some small spikes of performance degradation arise (having an overall duration around 7% of the total simulation horizon, see Figs 3 and 4 when the number of active VoIP calls suddenly increases (Fig. 5). Similar results may be obtained for the other video traces used during training.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, the bandwidth allocation problem has been investigated in relation to the QoS support in a WiMAX environment. A novel control mechanism has been developed to this aim, in the presence of heterogeneous traffic trunks.

Directions for future research may rely on a deep investigation of the WiMAX system performance considering the entire QoS architecture, e.g., in dependence of different traffic categories and considering the downlink component of the system, too.

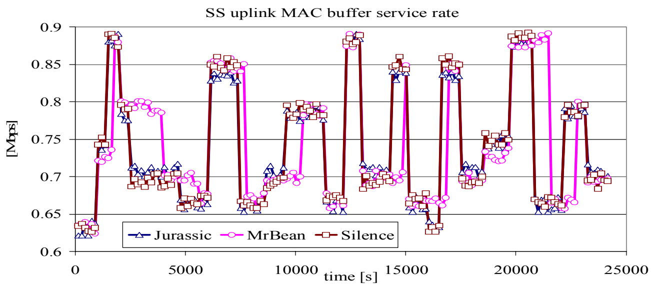

Figure 2. SS uplink MAC buffer service rate.

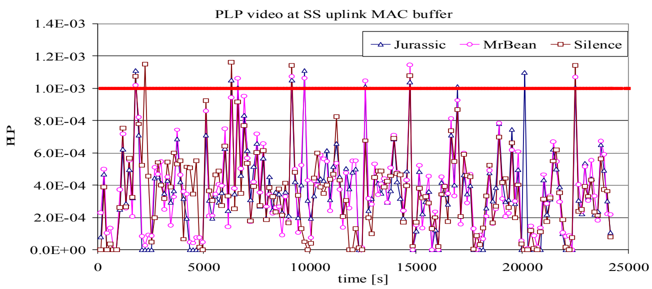

Figure 3. PLP video at SS uplink MAC buffer.

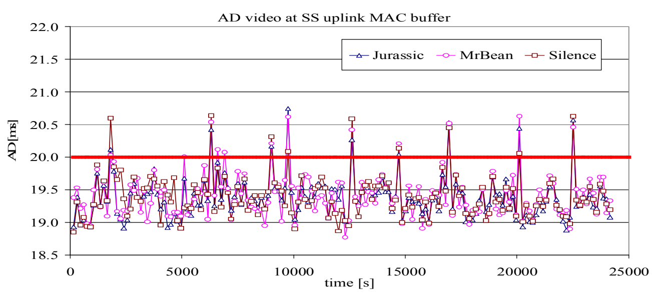

Figure 4. AD video at SS uplink MAC buffer.

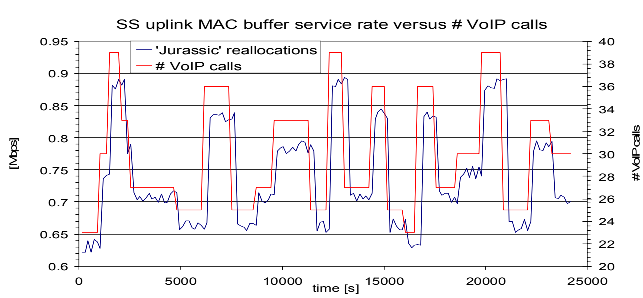

Figure 5. Service rate versus traffic variations.

References

[1] IEEE. Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Oct. 2004.

[2] WiMAX Forum. Business Case Models for Fixed Broadband Wireless Access based on WiMAX Technology and the 802.16 Standard, WiMAX Forum Technical Report, no. 10, Oct. 2004.

[3] "Cross-Layer Protocol Engineering for Wireless Mobile Networks: Part 1," *IEEE Comm. Mag.*, vol. 43, no. 12, Dec. 2005.

[4] ETSI. Satellite Earth Stations and Systems (SES). Broadband Satellite Multimedia (BSM) Services and Architectures: DiffServ Mappings and Negotiations. Draft ETSI TS 102 464 V 0.0.1, Jan. 2006.

[5] J. Chen, W. Jiao, Q. Guo, "Providing Integrated Qos Control for IEEE 802.16 Broadband Wireless Access," *62nd IEEE Vehicular Technology Conference*, Dallas, TX, Sept. 25-28, 2005.

[6] V. Kawadia, P. R. Kumar, "A Cautionary Perspective on Cross Layer Design," *IEEE Wireless Comm.*, vol. 12, no. 1, Feb. 2005, pp. 3-11.

[7] S. Shakkottai, T. S. Rappaport, P. C. Karlsson, "Cross-layer Design for Wireless Networks," *IEEE Comm. Mag.*, vol. 41, no. 10, Oct 2003 pp. 74-80.

[8] ETSI. Satellite Earth Stations and Systems (SES). Broadband Satellite Multimedia (BSM) Services and Architectures: QoS Functional Architecture. Draft ETSI TS 102 462 V0.4.2, Jan. 2006.

[9] M. Marchese, M. Mongelli, "On-line Bandwidth Control for Quality of Service Mapping over Satellite Independent Service Access Points," *Computer Networks*, vol. 50, no. 12, Aug. 2006, pp. 1885-2126.

[10] C. Eklund, R. B. Marks, K. L. Stanwood, S. Wang, "IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access," *IEEE Comm. Mag.*, vol. 40, no. 6, Jun 2002, pp. 98-107.

[11] A. Ghosh, D. R. Wolter, J. G. Andrews, R. Chen, "Broadband Wireless Access with WiMAX/802.16: Current Performance Benchmarks and Future potential," *IEEE Comm. Mag.*, vol. 43, no. 2, Feb. 2005, pp. 129-136.

[12] S. Maheshwari, *An Efficient QoS Scheduling Architecture of IEEE 802.16 Wireless MANs*, MTech Dissertation, Indian Institute of Information Tehcnology, 2005.

[13] G. Chu, D. Wang, S. Mei, "A QoS Architecture for the MAC protocol of IEEE 802.16 BWA system," *IEEE Internat. Conf. on Commun. 2002 (ICC 2002)*, New York, April-May 2002, pp. 435-439.

[14] M. Hawa, D. W. Petr, "Quality of Service Scheduling in Cable and Broadband Wireless Access Systems," *10th IEEE Internat. Workshop on Quality of Service 2002*, Miami, FL, May 2002, pp. 247-255.

[15] J. Chen, W. Jiao, H. Wang, "A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode," *IEEE Internat. Conf. on Commun. 2005 (ICC 2005)*, Seoul, May 16-20, 2005, pp. 3422-3426.

[16] K. Wongthavarawat, A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *Int. J. of Commun. Sys.*, vol. 16, no. 1, Feb. 2003, pp. 81-96.

[17] J. Schmitt, "Translation of Specification units between IP and ATM Quality of Service Declarations," *Internat. J. Commun. Sys.*, vol. 16, no. 4, 2003, pp. 291-310.

[18] C.-F. Su, G. de Veciana, "Statistical Multiplexing and Mix-Dependent Alternative Routing in Multiservice VP Networks," *IEEE/ACM Trans. on Networking*, vol. 8, no. 1, Feb. 2000, pp. 99-108.

[19] R. Guérin, H. Ahmadi, M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, September 1991 pp. 968-981.

[20] E. W. Knightly, N. B. Shroff, "Admission Control for Statistical QoS: Theory and Practice," *IEEE Network*, vol. 13, no. 2, March/April 1999, pp. 20-29.

[21] ITU-T. Series Y: Global Information Infrastructure and Internet Protocol Aspects. Network Performance Objectives for IP-based services. Rec. Y.1541, May 2002.

[22] D. Eun, N. Shroff, "A Measurement-Analytic Approach for Qos Estimation in a Network Based on the Dominant Time Scale," *IEEE/ACM Trans. on Networking*, vol. 11, no. 2, Apr. 2003, pp. 222-235.

[23] M. Baglietto, F. Davoli, M. Marchese, M. Mongelli, "Neural Approximation of Open-loop-Feedback Rate Control in Satellite Networks," *IEEE Trans. Neural Networks*, vol. 16, no 5, Sept. 2005, pp. 1195- 1211.

[24] http://www-tkn.ee.tu-berlin.de/research/trace/trace.html.