INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2007; **20**:1059–1079
Published online 12 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dac.862

Neural bandwidth allocation function (NBAF) control scheme at WiMAX MAC layer interface

Mario Marchese and Maurizio Mongelli*,†

DIST-Department of Communication, Computer and System Sciences, University of Genoa, Via Opera Pia 13, Genova 16145, Italy

SUMMARY

The paper proposes a bandwidth allocation scheme to be applied at the interface between upper layers (IP, in this paper) and *Medium Access Control* (MAC) layer over IEEE 802.16 protocol stack. The aim is to optimally tune the resource allocation to match objective QoS (*Quality of Service*) requirements. Traffic flows characterized by different performance requirements at the IP layer are conveyed to the IEEE 802.16 MAC layer. This process leads to the need for providing the necessary bandwidth at the MAC layer so that the traffic flow can receive the requested QoS.

The proposed control algorithm is based on real measures processed by a neural network and it is studied within the framework of 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 by using the existing features of 802.16 request–grant protocol acting at MAC layer. The performance evaluation reported in the paper shows the quick reaction of the bandwidth allocation scheme to traffic variations and the advantage provided in the number of accepted calls. Copyright © 2006 John Wiley & Sons, Ltd.

Received 11 November 2005; Revised 26 April 2006; Accepted 20 September 2006

KEY WORDS: 802.16 MAC layer; cross-layer quality of service (QoS); dynamic bandwidth allocation; neural control

1. INTRODUCTION

Recent estimations indicate that more than one million residential people and one hundred thousand institutions are not covered by terrestrial broadband access. The nature itself of wireless systems joined to new technical solutions allows offering an immediate coverage at high

*Correspondence to: Maurizio Mongelli, DIST-Department of Communication, Computer and System Sciences, University of Genoa, Via Opera Pia 13, Genova 16145, Italy.

†E-mail: mopa@dist.unige.it

speed. The challenge is if broadband wireless technology can fill the digital divide at service cost, reliability and quality comparable with terrestrial solutions. Great challenges concerning digital divide are also evident in Africa, Asia and South America and wireless technology have a real chance to play a key role. Also in the U.S., the lack of economical access to wired broadband resources at a significant number of facilities will pose a critical hindrance to business operations. Broadband wireless systems will be essential to enable comprehensive broadband services with the performance required to support the mission-critical applications needed by corporate and markets.

The IEEE 802.16 standard [1], also known as *Worldwide Interoperability for Microwave Access* (WiMAX), is the emerging technology for broadband wireless access [2]. Several economical studies foresee the application of WiMAX infrastructures in different market segments and business areas in the next few years [3]. It is considered as the ultimate solution for *Quality of Service* (QoS) delivery in wireless broadband infrastructures.

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.

Each uplink connection has its specific ID and deserves a specific QoS and the medium access control (MAC) protocol assures the required QoS with a proper resource allocation. If the necessary resources are not available, the incoming connection is rejected.

QoS is expressed in terms of *Packet Loss Probability* (PLP), *Average Delay* (AD) or *Delay Jitter* (DJ) of the packets. WiMAX standard does not specify QoS implementations details. Classification, resource allocation and scheduling are left as the responsibility of vendors.

In the downlink direction, scheduling over TDM is relatively simple because the BS can have information about the state of resources within the uplink subsystem. On the other hand, TDMA uplink scheduling in SS is more complex. SSs are the ingress point in the system for the traffic streams and represent the first point where resource allocation is provided. Uplink scheduling involves both BS and SS. BS is responsible for the resource allocation. The communication between SS and BS is guaranteed through a proper request–grant protocol, which is part of the MAC standard.

Additionally, WiMAX MAC offers QoS functionalities to the upper layers. The QoS-based services defined at the network layer (IP, ATM) are mapped into the WiMAX MAC core [1]. This process should lead to an intelligent use of bandwidth resources in the uplink direction, in dependence of the available service classes but also introduces specific cross-layer considerations [4–7], which are currently hot topics of research that are also related to open standardization activities in other environments [8, 9].

Operations performed at different layers of the protocol stack to match QoS in dependence of the specific layer features (e.g. concerning flow identification, aggregation capabilities, resource allocation) are called *QoS mapping*. 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. As a result, groups of connections with heterogeneous performance requirements need to be conveyed in the same MAC buffer.

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 aim is to properly tune the service rate of WiMAX MAC queues, in dependence of the current state of traffic and QoS requests, so that all the different performance requirements are satisfied, independent of the traffic multiplexing performed at MAC layer. The idea is to obtain a WiMAX resource allocation that is transparent to the layers overlying the 802.16 protocol stack, but satisfies the QoS constrains defined by the upper layers.

The remainder of the paper is organized as follows. The characteristics of the WiMAX MAC layer are summarized in the next two sections. Section 4 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 5. The formalization of the novel functional optimization approach provided by this paper is reported in Section 6 while the related bandwidth allocation solution is provided in Section 7. The performance evaluation, obtained by simulation analysis, is the subject of Section 8. The conclusions and the directions for future research are provided in Section 9.

2. CHANNEL ACCESS AND QOS IN WIRELESS SYSTEMS

In general, wireless communication takes place over a shared medium. This situation leads to collisions when multiple hosts try accessing the channel with packets transmissions. A control access scheme to the shared channel is therefore needed. In this view, two solutions are available: (1) *collision-based* channel access and (2) *collision-free* channel access. Each type of mechanisms provides different schemes of QoS [2].

(1) The principles of the first method are collision avoidance and collision resolution. A typical example is Ethernet CSMA/CD of wired LANs systems. When a collision occurs a retransmission is needed. Topical performance variable is collision probability. It depends on the number of active users and on the traffic load. They both increase the number of collisions and retransmissions. In this situation, it is hard to implement QoS. As a consequence, the scheme typically provides *Best Effort* (BE) service. Improvements may be obtained through overprovisioning or by adding a priority scheme. The latter consists of controlling different backoff windows to prioritize the classes. This solution is implemented in IEEE 802.11e (QoS WLAN) networks.

(2) In the second method, the channel is arbitrated in order to *a priori* avoid collisions. Only one host is allowed to transmit packets at a given time. Usually, a TDMA scheme is applied where the channel access opportunity is divided into frames and each frame is divided into time slots. The number of time slots assigned to a host reflects the bandwidth allocated for the host. Differently from the solution above, collision-free access requires a 'master' managing the global slots assignment. A proper signalling to change the assignment over time is needed, too. Actually, the slots organization can change dynamically during the lifetime of sessions as a function of the traffic load, QoS requirements and channel conditions. Collision-free channel access with TDMA is chosen in WiMAX to provide full access control and to implement tight QoS.

DOI: 10.1002/dac

3. THE WIMAX MEDIUM ACCESS CONTROL LAYER

WiMAX standard gives specification for MAC and physical layers for a Wireless MAN. The most important characteristics of it are summarized in the following by paying special attention to the instruments available for QoS support, which have a topical role in implementing the control algorithm introduced in this paper.

WiMAX environment consists of a central radio BS and a number of SSs. A SS typically covers a single residential or business building. BS is connected to public networks *via* cable fibre. 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 Automatic Repeat reQuest (ARQ), interference cancellation and power allocation [11], no further detail is given about it because the focus is on the WiMAX MAC features.

3.1. 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 (also inherently connectionless services are mapped to a connection) and is designed to support hundreds of users per channel with a variety of services. The CS sublayer provides the interface to the upper layers. It classifies the network service data units within the MAC system. It decides the MAC service class for the specific connection and initializes the resource allocation requests of the CPS. There are two general CS: ATM CS, designed for ATM services, and Packet CS, defined to map IPv4, IPv6, Ethernet and VLAN services, which is considered in this paper.

A MAC connection is identified by a 16-bit *Connection Identifier* (CID). Each SS has a standard 48-bit MAC address, which uniquely identifies the SS. 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 with the proper MAC service flow and CID. In practice, data coming from upper layer are received by the CS through the CS-SAP; differentiated between ATM and Packet; and conveyed to a specific buffer of the CPS layer. The stack may be simplified in this paper because the ATM CS is ignored and only the Packet (IP) service is considered. 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.

DOI: 10.1002/dac

3.2. *The MAC services*

WiMAX environment provides QoS support for both uplink and downlink traffic but this paper concerns only uplink bandwidth allocation. In this framework, 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 with 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. It eliminates the overhead and latency of SS requests. 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]. It offers periodic dedicated request opportunities to meet real-time requirements. Each SS emits explicit bandwidth requests, granted on the real need of a MAC connection. The *Non-real time Polling Service* (nrtPS) is related to non-real time bursty traffic with some QoS guarantees (for example, the aggregation of FTP or Web connections [10]). It is similar to rtPS but allows random access opportunities to send bandwidth requests. The BE service is designed to support regular Internet BE traffic without any guarantee.

3.3. *Grant per subscriber station versus grant per connection*

As mentioned above, each SS makes use of bandwidth request mechanisms to specify uplink bandwidth needs to BS. The requests may be either explicitly requested by special packets or piggybacked within a data packet. Requests can be aggregate or incremental. There are two classes of SS: the *Grant per Connection* (GPC) class, able to accept grants explicitly assigned to a specific connection; the *Grant per Subscriber Station* (GPSS) class, which accepts bandwidth grants assigned to the station. The bandwidth assigned to the connections of a singe GPSS station is aggregated into a single grant. A GPSS station can manage the bandwidth assigned flexibly because it can use the bandwidth requested by a connection (but no longer needed to it due to a status change from the last request) for other connections. GPSS is more scalable than GPC [10]. GPSS stations are only considered in this work (as in [12]).

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

While extensive signalling 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 [1, p. 139]. 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. Due to its recent application, WiMAX literature contains some QoS architecture solutions, but it is still limited concerning cross-layer techniques for QoS mapping between MAC and external (upper) layers.

Chut *et al.* [13] proposed architecture for dynamic bandwidth allocation by using GPSS mode. It implements traffic policing and exploits WRR (*Weighted Round Robin*) and priority

DOI: 10.1002/dac

scheduling algorithms for downlink and uplink service differentiation, respectively. Hawa and Petr [14] proposed an uplink scheduling differentiation mechanism based on the GPC mode. A QoS scheduling architecture is also studied in [12] to provide QoS guarantees to WiMAX applications. It uses a simple control algorithm to compute SS aggregate requests. Another MAC scheduling architecture is proposed and tested in [15, 16]. Chen *et al.* [5] and Wongthavarawat and Ganz [16] extend the QoS features of 802.16 standard by introducing traffic policing within SS and both an uplink hierarchical scheduler and a *Call Admission Control* (CAC) within BS. Service differentiation and QoS (in terms of bandwidth and delay requirements) are provided by using hierarchical scheduling (with per connection granularity in [5]).

The mentioned works propose solutions for QoS management by heuristically matching the QoS mapping operations between MAC and upper layer (e.g. in [16] per mean rate allocation is provided for rtPS connections). 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. More specifically, the aim is to capture the bandwidth needs arising at the uplink level through a measurement-based control and to allocate MAC resources according to them. Actually, if a given SCS can represent either an individual application or a group of applications [5, 16, 17], the resulting uplink traffic buffer serving that SCS must guarantee a wide range of QoS requirements, in dependence of the application categories aggregated in that SCS. In the presence of a large set of applications requiring diverse performance constraints, the bandwidth allocation mechanism should be intelligent enough to capture the bandwidth needs of the different connections when they are conveyed within a single MAC trunk. Taking the architectures in [5, 12, 17] as a reference, a limited number of MAC queues conveying traffic characterized by a large set of QoS requirements is used.

Aggregating in the same queue traffic that is heterogeneous, not only for the statistical features but also for QoS requirements, gives rise to a complex bandwidth allocation problem, which is investigated in the following.

5. THE UPLINK BANDWIDTH MANAGEMENT PROBLEM

As outlined above, the data flow traversing the 802.16 SS in uplink direction may be modelled as a cascade of buffers implemented, respectively, at IP layer and at MAC layer [5, 12, 17]. 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. Figure 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 and can benefit of a proper request–grant protocol (UGS, rtPS, nrtPS, BE) for bandwidth management, as specified in Section 3. 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

DOI: 10.1002/dac

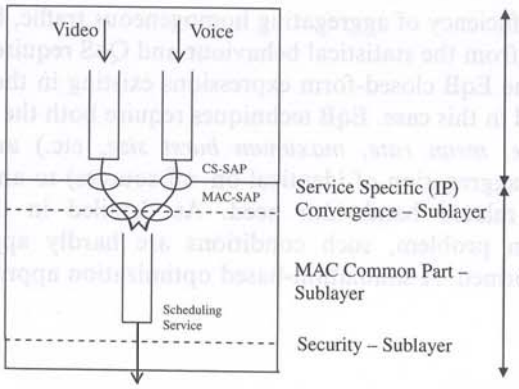

Figure 1. Data flow model for the *Subscriber Station* (SS) uplink [5, 12, 17].

(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:

Encapsulation change. QoS must be assured within the WiMAX system, despite the change of encapsulation format. The MAC overhead has an impact on bandwidth dimensioning [9, 18]. Moreover, the adoption of specific *fragmentation* and *packing* procedures, locally implemented at the MAC level, makes the QoS mapping problem analytically untractable. An example is reported in the performance evaluation section.

Traffic aggregation. The number of available queues for traffic classification and resource allocation at MAC layer may be lower than the one at IP layer. It depends on the technological constraints of the specific SS equipment (see, e.g. the SAP implementation for the satellite case [4, 8]).

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) [16]; (2) when connections are multiplexed within SS uplink queues [12, 17]; or (3) when flows coming from different SSs at BS are mixed in the BS uplink queues [12, 17].

In any case, 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.

DOI: 10.1002/dac

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 [19, 20]. The EqB closed-form expressions existing in the literature (see, e.g. [19–22]) can be hardly applied in this case. EqB techniques require both the adoption of specific traffic descriptors (*peak rate*, *mean rate*, *maximum burst size*, etc.) and the presence of homogeneous flows (e.g. the aggregation of identical on–off sources) to analytically derive the multiplexing gain and the related bandwidth need. As detailed in the following when formulating the optimization problem, such conditions are hardly applicable when QoS mapping operations are performed. A simulation-based optimization approach is the only way to avoid heuristic solutions.

6. THE MATHEMATICAL MODEL

To match the 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. The methodology is studied specifically for WiMAX environment by taking care of grant–request protocol functionalities to support bandwidth requests and CAC between BS and SSs. However, it can be generalized to match QoS mapping between IP and other MAC systems, such as for the satellite access case [8, 9].

Since the IP protocol is widely considered as the business and implementation choice for QoS deployment in the next future, IP is taken as the reference network layer overlying the WiMAX protocol stack. The QoS is therefore expressed in terms of IP metrics: PLP, AD, and DJ of IP packets [23].

It is important to define the QoS observation horizon T [24], as the time interval during which the QoS levels actually achieved for a specific flow are monitored. For example, $T = 1$ min for PLP $< 1\%$ means that during each period of 1 min, the averaged PLP must be lower than 1%.

Just one MAC buffer within an SS is considered (as in Figure 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 $i = 1, \ldots, N$ (for instance, the voice service in Figure 1), at a give time instant t, and entering the WiMAX system. 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, $i = 1, \ldots, N$, namely, $\boldsymbol{\alpha} = \mathrm{col}\{\alpha_1, \ldots, \alpha_N\}$. The overall input rate process of the MAC buffer, denoted by α_{MAC}, at a given time instant t, is obtained as $\alpha_{\mathrm{MAC}}(t) = \sum_{i=1}^{N} \alpha_i(t)$. The MAC buffer serves the queued traffic according to one of the mentioned SCS (i.e. UGS, rtPS, nrtPS, BE).

A sequence of α-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 α-observation time horizon. The specific protocol used for bandwidth requests will be detailed later.

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

$$\mathbf{I}(k\hat{t}) = \mathrm{col}\{\alpha_{\mathrm{MAC}}((k - \Xi)\hat{t}), \ldots, \alpha_{\mathrm{MAC}}((k - 1)\hat{t})\} \tag{1}$$

DOI: 10.1002/dac

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 $\hat{t}$ also denotes the reallocation period.

Let $J_{k\hat{t}}(\theta(k\hat{t}))=E_{\boldsymbol{\alpha}}\{\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}[\theta(k\hat{t}),\boldsymbol{\alpha}]\}$ 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 $\mathrm{QoS}^{[k\hat{t},k\hat{t}+T]}[\theta(k\hat{t}),\boldsymbol{\alpha}]$ is defined in (2).

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

The function $\mathrm{QoS}_i^{[k\hat{t},k\hat{t}+T]}[\theta(k\hat{t})]$ 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 $[k\hat{t},k\hat{t}+T]$. QoS_i^* is the desired QoS performance level for the service class i, which can be transmitted from the IP to the MAC layer through the CS-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 $f(\mathbf{I}(k\hat{t}))$ 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:

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

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:

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

is minimized.

Owing to the choice of the cost function (2), the optimal solution $f^*(\cdot)$ guarantees the alignment of the measured QoS levels $(\mathrm{QoS}_i^{[\cdot]}(\cdot))$ with the QoS target levels (QoS_i^*). The control strategy resulting from the solution of problem WFRAP for successive time instants $k\hat{t}$, $k=0,1,\ldots$ can be considered as an *Open Loop Feedback Control* (OLFC) [25]. It is able to perform on-line dynamic reactions to variable traffic conditions.

Due to the operations performed at the MAC-SAP interface (i.e. encapsulation change and traffic aggregation), no closed-form expressions for the cost functions $\mathrm{QoS}_i^{[\cdot]}(\cdot)$ are obtainable. The finite horizon functional cost (4) can be computed only through a simulation-based approximation. Analogous considerations have been outlined for a similar traffic aggregation problem studied in [19]. In this perspective, the neural approximation investigated below follows the direction of Montecarlo simulation approximation, by exploiting a sequence of estimates of (4) according to the possible stationary configurations of the stochastic processes involved in the problem.

It is also worth noting that even if EqB closed-form expressions for the cost functions $\mathrm{QoS}_i^{[\cdot]}(\cdot)$ were available (for example, with respect to the PLP in the presence of homogeneous on–off sources), their application in real time would require a mapping between the statistical behaviour of the sources and the traffic descriptors used in the EqB function. This operation may constitute a complicated traffic estimation problem. Moreover, the possible continuous online minimization of the EqB function to compute bandwidth reallocations is computationally expansive. The proposed methodology, on the other hand, guarantees instruments to react online to unexpected traffic changes with small computational effort. An example is reported in the performance evaluation section.

The next step is to formulate a technique aimed at solving the defined functional optimization problem.

7. THE CONTROL ALGORITHM

7.1. The extended Ritz method

In order to approximate the optimal resource allocation law $f^*(\cdot)$, a modification of the *Extended Ritz* method [26] is applied. The Extended Ritz method [27] 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 $\bar{f}(\mathbf{I}, \mathbf{w})$, being $\mathbf{I}$ the input of the NN and $\mathbf{w}$ the NN weights to be optimized. The scalar output of the NN, denoted by $\bar{\theta}$, is obtained as

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

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

$$\theta(k\hat{t}) = n[\bar{f}(\mathbf{I}(k\hat{t}), \mathbf{w})], \quad n(x) = \mathrm{Max}\, Bw \cdot x \tag{6}$$

The composition $n[\bar{f}(\mathbf{I}(k\hat{t}), \mathbf{w})]$ of the neural approximation $\bar{f}(\mathbf{I}, \mathbf{w})$ and of the normalization operator $n[\cdot]$ is identified as $\hat{f}(\mathbf{I}(k\hat{t}), \mathbf{w})$ 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 $\mathbf{w}$. It leads to the mathematical programming problem defined below.

Problem $\mathbf{WFRAP_w}$ finds the optimal parameter vector $\mathbf{w}^*$ such that the cost:

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

is minimized.

In this way, the functional optimization problem **WFRAP** has been reduced to an unconstrained non-linear programming one.

7.2. The training algorithm

To solve $\mathbf{WFRAP_w}$, a *stochastic approximating* [28] gradient-based algorithm of the form:

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

is applied, where the index h denotes both the steps of the iterative procedure and the generation of the hth realization of the stochastic processes $\boldsymbol{\alpha}$.

The components of the gradient $\nabla_{\mathbf{w}} \mathrm{QoS}^{[k\hat{t}, k\hat{t}+T]}[\hat{f}(\mathbf{I}(k\hat{t}), \mathbf{w}^h), \boldsymbol{\alpha}^h]$ can be obtained by applying the regular *backpropagation equations* used to train neural networks [29]. The backpropagation procedure must be initialized by means of the quantities $\partial \mathrm{QoS}^{[k\hat{t}, k\hat{t}+T]} / \partial \bar{\theta},\ i = 1, \ldots, N$ (i.e. the gradient $\nabla_{\bar{\theta}}\mathrm{QoS}^{[k\hat{\imath},k\hat{\imath}+T]}(\theta,\alpha^h))$. Unfortunately, as outlined above, such quantities cannot be obtained analytically, because no closed form is available for the functional cost.

The gradient $\nabla_{\bar{\theta}}\mathrm{QoS}^{[k\hat{\imath},k\hat{\imath}+T]}(\theta,\alpha^h)$ is then estimated by means of *Infinitesimal Perturbation Analysis* (IPA) [30]. IPA is a sensitive estimation for *Discrete Event Systems* and allows getting derivative estimators of the buffer performance (loss and delay volumes) as functions of the available resources (service rate and buffer size).

7.3. QoS control in wireless systems

This is the first time the neural control studied in [26] is used to tune bandwidth provision to support QoS. In this view, the paper also constitutes a research improvement in the field of QoS control methodologies for wireless systems.

Traditional PID (*proportional integral derivative*) controllers, despite the related simple approach (almost primitive, in eyes of some control theorists), often yield good performance, as, for instance, in TCP control and WLAN QoS research fields (see, e.g. [31–33]).

More specifically, in [31, 33], a PID controller is applied to allow capacity optimization and the maintenance of QoS differentiation in a WLAN system. However, it is worth noting that since PID controllers performance is guaranteed in steady state, the QoS requirements may not be always assured over specific time periods [24]. Moreover, some specific metrics (such as PLP), are not taken into account explicitly in [31, 33]. The PID parameters may not also depend explicitly on the chosen metrics. As a consequence, the PID parameters must be tuned with care to obtain good performance. In this perspective, more sophisticated techniques (e.g. fuzzy tuning, neural networks, identification methodologies, learning theory) have been studied and applied to achieve optimized performance (see, e.g. [9, 26, 34] and reference therein).

The control paradigm proposed follows this principle and makes use of a neural feedback control to guarantee different QoS metrics over the chosen observation horizon T. In [34], a similar receding-horizon prediction method is applied to drive bandwidth reallocations of MAC queues in a GEO satellite systems. A possible drawback relies on the required on-line computational burden of the technique. As outlined below, the proposed approach is suitable for real-time implementation since it requires a small computational effort after the training has taken effect.

In [9], an optimization approach is proposed to match PLP requirements with small on-line computational effort. A gradient descent is applied on-line to compute bandwidth needs over a satellite SAP interface. Since the algorithm proposed is driven by IPA in real time, it may introduce suboptimal performance, which the NBAF studied here tries avoiding.

7.4. QoS measurement and computational complexity

In the specific architecture, the QoS is expressed in SDUs and implemented within the MAC frame composed of MAC PDUs. As a consequence, QoS depends on the fragmentation and packing procedures applied by CPS sublayer. For instance, a SDU packet is considered lost if a least one PDU containing a portion of that packet is lost.

The NBAF is trained in dependence of the SDUs performance actually measured at MAC layer. In real time, it performs a sequence of reallocations as function of the traffic samples collected in the information vector. No other knowledge on the state of the sources is required.

DOI: 10.1002/dac

The control technique is therefore suited to on-line control of the MAC frame actually implemented at the uplink subsystem.

The training procedure described above can be performed off-line. The related computational burden does not influence the on-line performance of the system. When the system works in real time, the optimized NBAF $\hat{f}(\cdot, \mathbf{w}^*)$ (already computed off-line), is directly applied, thus obtaining necessary resource allocations simply accessing a memory. Moreover, no further conceptual difficulty is involved if one would want to employ the data collected on-line to perform real-time 'adjustment' of the parameters' vector $\mathbf{w}^*$ through (8). The only need to perform on-line training is related to the application of IPA during the system evolution to update the gradient used in (8). However, also the computational complexity for computing IPA-based gradients is low.

7.5. Aggregating versus separating traffic flows

The buffer structure of Figure 1 is taken as reference when developing the NBAF methodology. It means that, for the sake of simplicity, a single buffer is considered at MAC layer, thus generating the need of aggregating traffic. The proposed control methodology is then used to optimally tune the buffer service rate. Optimal statistical multiplexing when a pool of buffers is available [19] is not explicitly considered. As such, it would be interesting to compare the complexity of the NBAF solution to one in which the wireless system had enough queues so that mixing different traffic classes in the same queue is not needed. This subject is left open for future analysis. The interested reader is referred to [19,35], where resource allocation trade-off between aggregating and separating traffic is investigated.

7.6. 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 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 $\mathbf{I}(\cdot)$ 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.

DOI: 10.1002/dac

7.7. Call admission control

The bandwidth allocation presented may simply work together with a CAC. Following [31], the CAC can exploit the presence of the measurement-based control to drive CAC decisions.

Let J^q the number of active connections currently served by the qth uplink SCS ($q =$ UGS, rtPS, nrtPS, BE), whose bandwidth grant is computed by the qth trained NBAF denoted by $\hat{f}_q(\mathrm{I}(\cdot), \mathbf{w}_q^*)$. An additional incoming connection, identified as (J^q+1), for the qth MAC service buffer is admitted if and only if:

$$B_{\mathrm{p}}^{J^q+1} + \hat{f}_q(\mathrm{I}(\cdot), \mathbf{w}_q^*) \leqslant \mathrm{Max}\, Bw_q \tag{9}$$

where $B_p^{J^q+1}$ represents the peak bandwidth of the (J^q+1)th source. The suggested CAC takes into account the bandwidth actually used by the flows. If the CAC depends only on the peak rates of the sources (as, e.g. proposed in [36]), the CAC decisions lead to an underutilization of the available capacity, as shown in [31].

8. PERFORMANCE EVALUATION

To test the proposed control methodology, a C ++ simulator has been developed for the IP and WiMAX MAC queues, having in mind the aggregation architecture shown in Figure 1. The aims of the performance evaluation are:

Allocations optimality. To highlight both QoS preservation and adaptive response to traffic variations for a single rtPS buffer using the NBAF methodology.

CAC. To check the efficiency of the NBAF-driven CAC in presence of rtPS and UGS traffics.

On-line training. To test the on-line performance in the presence of unpredicted traffic variations.

A single SS is considered as in [17]. The intervention of multiple SSs is left for future analysis. The PLP and AD performance metrics are tested here. However, the analysis can also be generalized for other QoS metrics (e.g. DJ). The simulations scenarios are settled having in mind the regular *small and medium-sized enterprize* (SME) access scenario as, e.g. in [3, 16]. Since the study focuses on MAC QoS, ideal channel conditions are assumed (as in [16]), thus packet corruption due to the wireless channel never occurs.

8.1. NBAF bandwidth provision

A single rtPS MAC buffer is considered for now. Following [20], 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 Figure 1).

VoIP sources are modelled as an exponentially modulated on–off process, with mean on and off times (as in the ITU P.59 recommendation) equal to 1.008 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 three 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 [37], is 10 min.

As far as the video service is concerned, real traces taken from [38] have been used. Data are H.263 encoded and have an *average bit rate* ($\bar{B}_{\mathrm{Video}}$) of 260 kbps and a *peak bit rate* ($B_{\mathrm{Video}}^{\mathrm{p}}$) ranging

DOI: 10.1002/dac

from 1.3 to 1.5 Mbps, depending on the specific trace. Each video trace lasts about 1 h. The video QoS targets are PLP $= 0.01$ (1%) and AD $= 20$ ms. The QoS observation horizon T lasts 5 min.

The *Packet CS* encapsulation format [1, p. 20] of VoIP and video packets (IP SDU) is implemented without header suppression. The MAC overhead corresponds to 48-bit, due to the GH without CRC. The MAC payload is the IP SDU. Fragmentation of video SDUs and packing of SDU data [1, p. 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 five 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 ${}^{\mathrm{PLP}}\hat{f}(\mathbf{I}(\cdot),{}^{\mathrm{PLP}}\mathbf{w}^*))$ is trained for the solution of problem $\mathbf{WFRAP_w}$ where the QoS_i^* targets in (2) are the PLPs defined for VoIP and video, respectively ($i =$ VoIP, Video). The second one (denoted by ${}^{\mathrm{AD}}\hat{f}(\mathbf{I}(\cdot),{}^{\mathrm{AD}}\mathbf{w}^*))$ is trained with respect to the AD constraints.

The α-observation horizon $\hat{t}$ is set to 30 s and the depth Ξ of the time horizon of the information vector $\mathbf{I}(\cdot)$ 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 $\mathbf{w}^*$ characterizing each NBAF is obtained off-line by means of the stochastic gradient technique described in Section 7.2.

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 (λ_{VoIP}, μ_{VoIP}). Operatively, at step $h = 0$ of the training phase, a random initialization of the vector $\mathbf{w} = \mathbf{w}^0$ is performed. Then, a sample path is generated according to the chosen statistics over a time horizon $[0, \Xi\hat{t} + \Sigma T]$ (T denotes again the dimension of the QoS observation horizon). At time $\bar{t} = \Xi\hat{t}$, the information vector $\mathbf{I}(\bar{t})$ is collected and the bandwidth reallocation is computed by using the NBAF. Such a reallocation is applied over a time horizon $[\bar{t}, \bar{t} + T]$ and the gradient $\nabla_{\mathbf{w}}\mathrm{QoS}^{[\bar{t},\bar{t}+T]}[\hat{f}(\mathbf{I}(\bar{t}), \mathbf{w}^0), \boldsymbol{\alpha}^0]$ is approximated through the IPA equations [30] relative to the loss performance metric when PLP is the target (i.e. training ${}^{\mathrm{PLP}}\hat{f}(\mathbf{I}(\cdot),{}^{\mathrm{PLP}}\mathbf{w}^*))$ and relative to delay volume metric when AD is the target (i.e. training ${}^{\mathrm{AD}}\hat{f}(\mathbf{I}(\cdot),{}^{\mathrm{AD}}\mathbf{w}^*))$. The vector $\mathbf{w}^1$ is then computed by applying (8). Then (step $h = 2$), time is moved to $\bar{t} = \Xi\hat{t} + T$, the information vector $\mathbf{I}(\bar{t})$ is collected and the same operations, thus obtaining $\mathbf{w}^2$, is repeated. The same procedure is applied until $t = \Xi\hat{t} + \Sigma T$ (step $h = \Sigma$). After that, a new sample path is generated and a new simulation is performed. The same steps are repeated until the cost function (7) exhibits a steady-state minimum. Each Σ' steps ($\Sigma' = \Sigma/20$), the current video trace is changed with another one, randomly chosen out of the first five available in [38]. Up to 2500 independent replications of the training simulation scenario have been used to successfully terminate the training phase. The simulation time of the training phase took around 7.6 h with an AMD Athlon at 1.8 GHz.

After training, the NBAFs performance is tested over a simulation horizon of about 7 h. 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:

$$\theta(k\hat{t}) = \mathrm{Max}\{{}^{\mathrm{PLP}}\hat{f}(\mathbf{I}(k\hat{t}),{}^{\mathrm{PLP}}\mathbf{w}^*),\ {}^{\mathrm{AD}}\hat{f}(\mathbf{I}(k\hat{t}),{}^{\mathrm{AD}}\mathbf{w}^*)\}, \quad k = 1, 2, \ldots \tag{10}$$

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

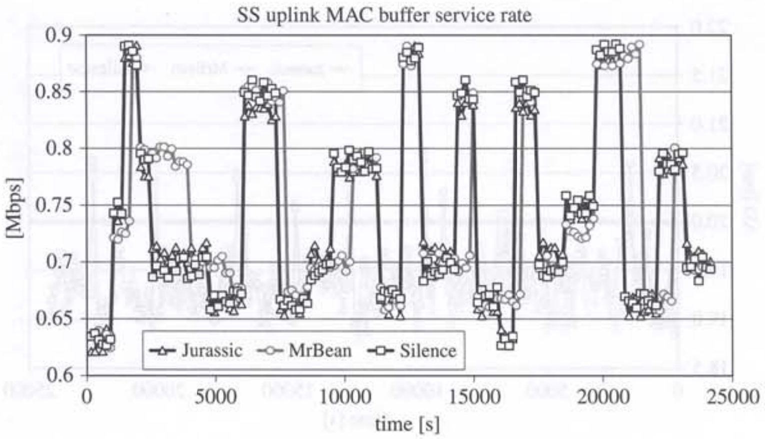

Figure 2. SS uplink MAC buffer service rate.

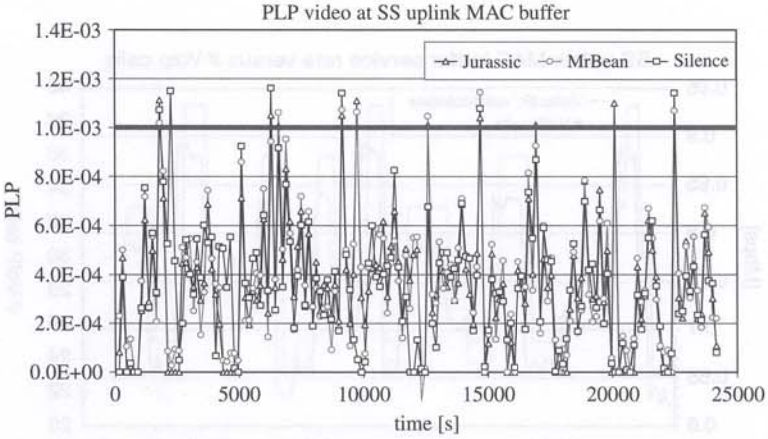

Figure 3. PLP video at SS uplink MAC buffer.

Figure 2 shows the allocations obtained by (10) during the system evolution. Figures 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 min). 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 Figures 3 and 4 denotes the QoS video targets ($\mathrm{PLP} = 10^{-3}$, $\mathrm{AD} = 20$ ms).

The changes in the service rate are due to the time varying number of active VoIP calls whose variation is highlighted in Figure 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 Figures 3 and 4) when the number of active VoIP calls suddenly increases (Figure 5). Similar results may be obtained for the other video traces used during training.

DOI: 10.1002/dac

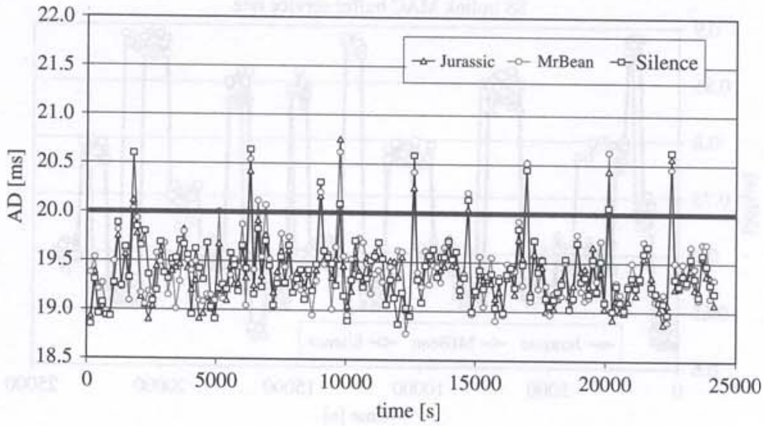

Figure 4. AD video at SS uplink MAC buffer.

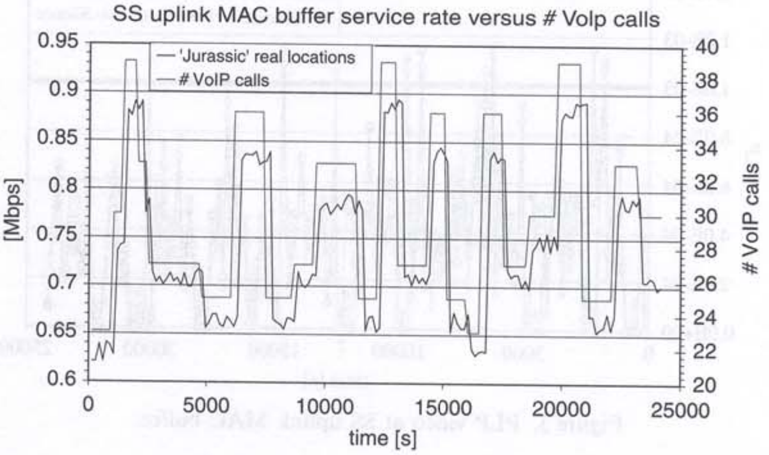

Figure 5. Service rate *versus* traffic variations.

8.2. *NBAF CAC*

This part is devoted to the investigation of the advantage provided in the CAC performance by the NBAF approach in presence of UGS and rtPS traffics. The two SCS are implemented through two MAC service buffers. UGS is characterized by virtual leased lines requiring a bandwidth allocation of 512 kbps for each connection. UGS connection request arrival frequency is 0.1 calls per minute. rtPS SCS conveys the heterogeneous VoIP and video trunk outlined in the previous subsection. rtPS statistics are: $\lambda_{\text{video}} = 1$ call/30 min (exponentially distributed), $\mu_{\text{video}} = 1$ h (exponentially distributed), $\mu_{\text{VoIP}} = 10$ min (log-normally distributed). λ_{VoIP} is changed in successive simulations to test different channel utilizations conditions. The uplink channel capacity (C_{uplink}) is 5 Mbps as in [12]. For the sake of simplicity, a *complete*

DOI: 10.1002/dac

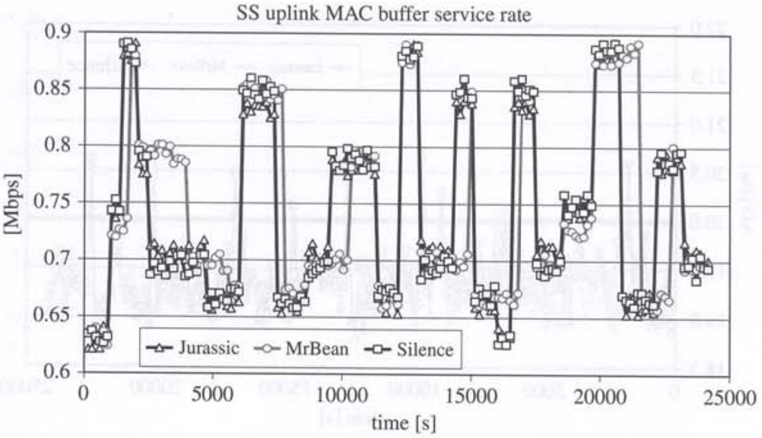

Figure 2. SS uplink MAC buffer service rate.

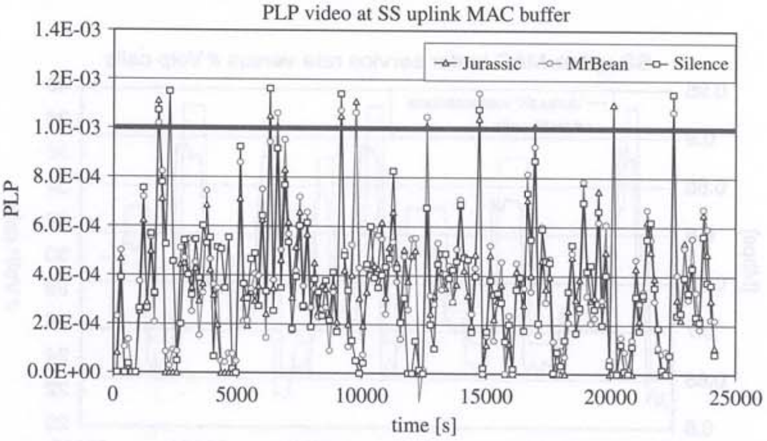

Figure 3. PLP video at SS uplink MAC buffer.

Figure 2 shows the allocations obtained by (10) during the system evolution. Figures 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 min). 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 Figures 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 Figure 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 Figures 3 and 4) when the number of active VoIP calls suddenly increases (Figure 5). Similar results may be obtained for the other video traces used during training.

sharing (CS) approach is applied for UGS and rtPS. It means that no *a priori* bandwidth separations are provided between the classes (i.e. $\text{Max Bw}_q = C_{\text{uplink}},\ q = \text{UGS, rtPS}$). An incoming connection of a given SCS enters the network if the required bandwidth is lower than the residual channel capacity. The approach has the advantage of simplicity but some SCS may starve in presence of other SCS with higher call frequencies. It leads to the application of specific *service separation* (SvS) schemes (i.e. computing $\text{Max Bw}_{\text{SCS}}$ in function of the SCS statistics) [39], to derive guaranteed blocking probability performance for each SCS. The NBAF CAC rule may be integrated within the SvS mode but it is out of the scope of this paper.

Bandwidth provision and CAC for UGS are trivial, since UGS deals (in this case) with simple CBR-like allocations. Bandwidth provision for rtPS is computed by the trained NBAFs using (10). rtPS CAC is driven by following the rule in (9), detailed in Section 7.7, where the current resource utilization computed by the NBAF is obtained by (10).

The NBAF approach is compared with a *peak bandwidth* approach (PeakB) and an *Equivalent Bandwidth* approach (EqB). PeakB does not consider the real utilization of the capacity. It also applies the CAC rule (9) but applying the sum of the peak rates of all the active connections instead of (10). EqB is taken directly from [21] having in mind the performance evaluations of [19, 20] concerning traffic aggregation. No other EqB techniques are applicable in this context because no traffic descriptors are obtainable for the rtPS α_{MAC} process. $m_{[(k-1)\hat{t}_{\text{EqB}},\ k\hat{t}_{\text{EqB}}]}$ and $\sigma_{[(k-1)\hat{t}_{\text{EqB}},\ k\hat{t}_{\text{EqB}}]}$ denote the measured *mean* and *variance* of the input rate process of the rtPS buffer during the EqB α-observation horizon $[(k-1)\hat{t}_{\text{EqB}}, k\hat{t}_{\text{EqB}}]$, whose duration is $\hat{t}_{\text{EqB}}$. The value of $\hat{t}_{\text{EqB}}$ has been accurately dimensioned to optimize EqB performance in terms of trade off between stability of bandwidth allocations and achieved QoS [20]. The result is $\hat{t}_{\text{EqB}} =$ 6 min. A new rtPS buffer service rate reallocation is performed every $t = k\hat{t}_{\text{EqB}}, = 1, 2, \ldots,$ as ruled by (11):

$$\theta_{rtPS}(k\hat{t}_{\text{EqB}}) = m_{[(k-1)\hat{t}_{\text{EqB}}, k\hat{t}_{\text{EqB}}]} + \zeta(\varepsilon) \cdot \sigma_{[(k-1)\hat{t}_{\text{EqB}}, k\hat{t}_{\text{EqB}}]} \tag{11}$$

where $\zeta(\varepsilon) = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)}$, being ε is the most stringent PLP requirement among the connections of the rtPS trunk. It allows guaranteeing all the QoS thresholds in the MAC queue, but it introduces bandwidth waste [19]. As a consequence, it overestimates the necessary resources and leads to poor CAC performance as shown below. The EqB CAC rule consists of using (11) in place of (10) when computing the current utilization of the rtPS buffer in (9).

EqB technique (11) is suited for PLP control. Actually, the delay control may also matched by properly dimensioning the maximum buffer size (as widely done in the literature), but it requires a precise knowledge of buffer lengths, which is not required for the application of the control algorithm proposed. Moreover, even if this approach reveals to be a good heuristic for the maintenance of reasonable upper bound on the *Maximum Transfer Delay*, it often overestimates the bandwidth requirement related to the AD constraint.

Figure 6 shows the *blocking probability* (BlP) of VoIP calls by using the three mentioned strategies as a function of the VoIP call arrival frequency. The width of the confidence interval over the following simulated BlP performance measures is less than 1% for the 95% of the cases. The trend clearly outlines that the NBAF allows improving the available channel capacity, optimally tuning the bandwidth provision. Actually NBAF exploits in real time the off-line training phase where it learned the real sensitivity of QoS metrics in dependence of the system statistics. Comments of the same sort might be applied for video and UGS BlP. Similar results can be obtained by fixing λ_{VoIP} and introducing a variable λ_{video}.

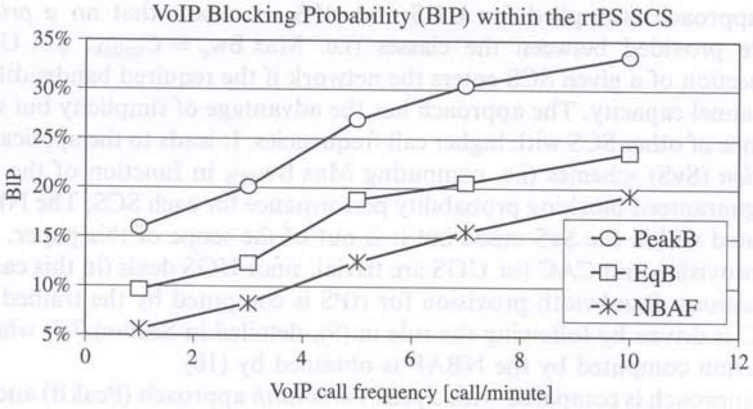

Figure 6. VoIP blocking probability *versus* VoIP call arrival frequency.

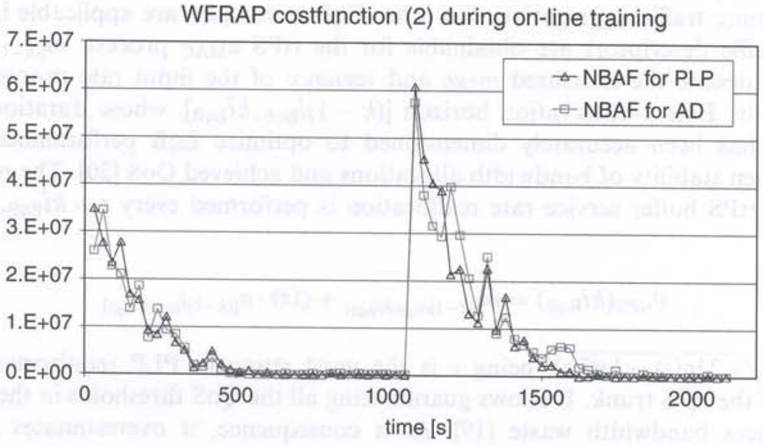

Figure 7. On-line NBAFs training for PLP and AD.

8.3. *On-line NBAF training*

The real-time adjustment of the trained NBAF is now under investigation. As outlined in Section 7.3, traffic conditions not considered during the off-line training can be matched by updating the trained NBAF using (8) according to the IPA-based gradient values computed on-line.

To test the performance in this case, the first simulation scenario is considered generating VoIP calls interarrival frequencies (λ_{VoIP}) different from the one used during off-line training. As a consequence, the trained NBAFs $^{\mathrm{PLP}}\hat{f}(\mathrm{I}(\cdot),{}^{\mathrm{PLP}}\mathrm{w}^*)$, $^{\mathrm{AD}}\hat{f}(\mathrm{I}(\cdot),{}^{\mathrm{AD}}\mathrm{w}^*)$ receive on-line slightly different inputs (in the information vector $\mathrm{I}(\cdot)$) from the ones used during the first training.

Figure 7 shows the values of the cost function (2) for PLP and AD during on-line training. The NBAFs have been trained off-line with respect to $\lambda_{\mathrm{VoIP}} = 3$ calls per minute. The new values of λ_{VoIP} introduced are: 4 calls per minute (at the beginning of the simulation) and 5 calls per

DOI: 10.1002/dac

minute (after about 1000 s of simulation). To obtain fast adaptation to the new traffic conditions, the observation horizon T is reduced to 1 min. The NBAFs need two further training periods (of about 500 s of duration each) to find out optimality in the presence of new system condition. The steady states achieved by the cost function around zero in Figure 7 denote the equalization of the QoS measured at the MAC layer with respect to the QoS^* targets for both VoIP and video.

The analysis suggests that new traffic conditions may be deduced on the basis of the current values of the cost function. However, its gradient and the measured QoS levels themselves are good indicators of the presence of new conditions as well. Thus, the network operator responsible for the WiMAX core performance has to periodically supervise (at least one) of the mentioned quantities in real time, in order to exploit other possible instances of traffic conditions, not previously used in the training phase, and restarts a new training to update the NBAF, if necessary. The key point is that the training applied in real time is faster than the one performed for the first time off-line, because an 'already partially trained' NBAF is adopted. Real-time adaptation of the NBAF weights may be applied when needed until a new steady state of the cost function is reached (as shown in Figure 7). Reallocations and CAC may be driven by the EqB heuristic (11) to assure QoS when the NBAF is not optimized with respect to unexpected system conditions.

9. CONCLUSIONS AND FUTURE WORK

The bandwidth allocation problem has been investigated in relation to the QoS mapping in a WiMAX environment. A novel control mechanism has been developed to this aim, in presence of heterogeneous traffic trunks. The control methodology has been obtained by reducing a proper functional optimization problem to an approximating scheme, suited to neural network training. The performance evaluation confirmed the good performance of the proposed methodology.

Directions for future research may rely on a deep investigation of the WiMAX system performance by considering the overall architecture (e.g. [5, 17]) in dependence of different traffic categories and emphasizing the impact of the different points where traffic aggregation is applied. Special attention may be devoted to modelling the physical layer and considering cross-layer issues triggered by power scheme and redundancy codes applicable at the physical layer [6, 7]. The implementation of specific scheduling algorithms (*Deficit Round Robin*, *Weighted Round Robin*, etc.) [16] in dependence of traffic categories may be under investigation, too.

REFERENCES

1. IEEE. *Standard for Local and Metropolitan Area Networks— Part 16: Air Interface for Fixed Broadband Wireless Access Systems.* Revision of IEEE Sts 802.16-2001, 1 October 2004.
2. Ganz A, Ganz Z, Wongthavarawat K. *Multimedia Wireless Networks: Technologies, Standards, and QoS.* Prentice-Hall PTR: Upper Saddle River, NJ, 2004.
3. WiMAX Forum. Business case models for fixed broadband wireless access based on WiMAX technology and the 802.16 standard. *WiMAX Forum Technical Report*, vol. 10. October 2004.
4. ETSI. Satellite Earth Stations and Systems (SES). Broadband satellite multimedia (BSM) services and architectures: diffserv mappings and negotiations. *Technical Specification, Draft ETSI TS 102 464 V 0.0.1*, January 2006.
5. Chen J, Jiao W, Guo Q. Providing integrated Qos control for IEEE 802.16 broadband wireless access. *Proceedings of 62nd IEEE Vehicular Technology Conference*, Dallas, TX, 25–28 September 2005.

DOI: 10.1002/dac

6. Hossain E, Fantacci R, Karmouch A, Kota SL. Cross-layer protocol engineering for wireless mobile networks: part 1. *IEEE Communication Magazine* 2005; **43**(12):110–111.
7. Shakkottai S, Rappaport TS, Karlsson PC. Cross-layer design for wireless networks. *IEEE Communication Magazine* 2003; **41**(10):74–80.
8. ETSI. Satellite Earth Stations and Systems (SES). Broadband satellite multimedia (BSM) services and architectures: QoS functional architecture. *Technical Specification, Draft ETSI TS 102 462 V0.4.2*, January 2006.
9. Marchese M, Mongelli M. On-line bandwidth control for quality of service mapping over satellite independent service access points. *Computer Networks* 2006; **50**(12):1885–2126.
10. Eklund C, Marks RB, Stanwood KL, Wang S. IEEE standard 802.16: a technical overview of the wirelessman air interface for broadband wireless access. *IEEE Communication Magazine* 2002; **40**(6):98–107.
11. Ghosh A, Wolter DR, Andrews JG, Chen R. Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential. *IEEE Communication Magazine* 2005; **43**(2):129–136.
12. Maheshwari S. An efficient QoS scheduling architecture of IEEE 802.16 wireless MANs. *M.Tech Dissertation*, Indian Institute of Information Technology, 2005.
13. Chu G, Wang D, Mei S. A QoS architecture for the MAC protocol of IEEE 802.16 BWA system. *Proceedings of IEEE International Conference on Communications 2002 (ICC)*, New York, April–May 2002: 435–439.
14. Hawa M, Petr DW. Quality of service scheduling in cable and broadband wireless access systems. *Proceedings of 10th IEEE International Workshop on Quality of Service 2002*, Miami, FL, May 2002; 247–255.
15. Chen J, Jiao W, Wang H. A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD Mode. *Proceedings of IEEE International Conference on Communication 2005 (ICC)*, Seoul, 16–20 May 2005; 3422–3426.
16. Cicconetti C, Lenzini L, Mingozzi E. Quality of service support in IEEE 802.16 networks. *IEEE Network* 2006; **20**(2):50–55.
17. Wongthavarawat K, Ganz A. Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems. *International Journal of Communication Systems* 2003; **16**(1):81–96.
18. Schmitt J. Translation of specification units between IP and ATM quality of service declarations. *International Journal of Communication Systems* 2003; **16**(4):291–310.
19. Su CF, de Veciana G. Statistical multiplexing and mix-dependent alternative routing in multiservice VP networks. *IEEE/ACM Transactions on Networking* 2000; **8**(1):99–108.
20. Georgoulas S, Trimintzios P, Pavlou G, Ho KH. Heterogeneous real-time traffic admission control in differentiated services domains. *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM) 2005*, St Louis, MO, U.S.A., 28 November–2 December 2005; 523–528.
21. Guérin R, Ahmadi H, Naghshineh M. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications* 1991; **9**(7):968–981.
22. Knightly EW, Shroff NB. Admission control for statistical QoS: theory and practice. *IEEE Network* 1999; **13**(2): 20–29.
23. ITU-T. Series Y: Global information infrastructure and internet protocol aspects. Network Performance Objectives for IP-based services. *ITU-T Recommendation Y.1541*, May 2002.
24. Eun D, Shroff N. A measurement-analytic approach for QoS estimation in a network based on the dominant time scale. *IEEE/ACM Transactions on Networking* 2003; **11**(2):222–235.
25. Bertsekas D. *Dynamic Programming and Optimal Control* (2nd edn.). Athena Scientific: Belmont, MA, 2001.
26. Baglietto M, Davoli F, Marchese M, Mongelli M. Neural approximation of open-loop-feedback rate control in satellite networks. *IEEE Transactions on Neural Networks* 2005; **16**(5):1195–1211.
27. Zoppoli R, Sanguineti M, Parisini T. Approximating networks and extended ritz method for solution of functional optimization problems. *Journal of Optimization Theory and Applications* 2002; **112**(2):403–439.
28. Kushner HJ, Yin GG. *Stochastic Approximation Algorithms and Applications*. Springer: New York, NY, 1997.
29. Haykin S. *Neural Networks. A Comprehensive Foundation*. Macmillan Publishing: New York, NY, 1994.
30. Wardi Y, Melamed B, Cassandras CG, Panayiotou CG. Online IPA gradient estimators in stochastic continuous fluid models. *Journal of Optimization Theory and Applications* 2002; **115**(2):369–405.
31. Boggia G, Camarda P, Grieco LA, Mascolo S. Feedback based bandwidth allocation with call admission control for providing delay guarantees in IEEE 802.11e networks. *Computer Communications* 2005; **28**(3):325–337.
32. Aweya J, Ouelette M, Montuno DY. A self-regulating TCP acknowledgement (ACK) pacing scheme. *International Journal of Network Management* 2002; **12**(1):145–163.
33. Boggia G, Camarda P, Grieco LA, Mascolo S. Energy efficient feedback-based scheduler for delay guarantees in IEEE 802.11e networks. *Computer Communications* 2006; **29**(13–14):2680–2692.
34. Chisci L, Pecorella T, Fantacci R. Dynamic bandwidth allocation in GEO satellite networks: a predictive control approach. *Control Engineering Practice* 2006; **14**(9):1057–1067.
35. Fortunato E, Marchese M, Mongelli M, Raviola A. QoS guarantee in telecommunication networks: technologies and solutions. *International Journal of Communication Systems* 2004; **17**(10):935–962.
36. IEEE. 802.11 WG: *Draft Amendment to Standard for Information Technology—LAN/MAN Specific Requirements —Part 11: Wireless MAC and PHY Specifications: MAC QoS Enhancements*. IEEE 802.11e/D10.0, 2004.

37. Bolotin VA. Modeling calling holding time distributions for CCS network design and performance analysis. *IEEE Journal on Selected Areas in Communications* 1994; **12**(3):433–438.
38. http://www-tkn.ee.tu-berlin.de/research/trace/trace.html
39. Ross K. *Multiservice Loss models for Broadband Telecommunication Networks*. Springer: Berlin, 1995.

AUTHORS' BIOGRAPHIES

Mario Marchese was born in Genoa, Italy in 1967. He got his 'Laurea' degree cum laude at the University of Genoa, Italy in 1992 and the Qualification as Professional Engineer in April 1992. He obtained his PhD (Italian 'Dottorato di Ricerca') degree in 'Telecommunications' at the University of Genoa in 1996. From 1999 to 2004, he worked with the Italian Consortium of Telecommunications (CNIT), the University of Genoa Research Unit, where he was Head of Research. From February 2005 he has been Associate Professor at the University of Genoa, Department of Communication, Computer and Systems Science (DIST). He is the founder and still the technical responsible of CNIT/DIST Satellite Communication and Networking Laboratory (SCNL), the University of Genoa, which contains high value devices and tools and implies the management of different units of specialized scientific and technical personnel. He is Vice-Chair of the IEEE Satellite and Space Communications Technical Committee and Senior Member of the IEEE. He is author and co-author of more than 80 scientific works, including international magazines, international conferences and book chapters. His main research activity concerns: Satellite Networks, Transport Layer over Satellite and Wireless Networks, Quality of Service over ATM, IP and MPLS, and Data Transport over Heterogeneous Networks.

Maurizio Mongelli was born in Savona, Italy in 1975. He got his 'Laurea' degree cum laude at the University of Genoa, Italy in 2000 and the Qualification as Professional Engineer in April 2002. He obtained his PhD (Italian 'Dottorato di Ricerca') degree in 'Electronic and Computer Engineering' at the University of Genoa in 2004. His PhD was funded by Selenia S.p.A. He worked for both Selenia S.p.A and the Italian Consortium of Telecommunications (CNIT), by the University of Genoa Research Unit from 2000 to 2004. He is now a member of the research staff of the Telecommunication Networking Research Group by the University of Genoa, with a post-doctoral scholarship founded by Selenia S.p.A. His main research activity concerns: QoS architectures, resource allocation and optimization algorithms for telecommunication systems.

DOI: 10.1002/dac