ARTICLE IN PRESS

Computer Networks xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A survey of MAC based QoS implementations for WiMAX networks

Y. Ahmet Şekercioğlu [a,*], Milosh Ivanovich [a], Alper Yeğin [b]

[a] Department of Electrical and Computer Systems Engineering, Monash University, Australia
[b] Standards and Industry Initiatives Group, Samsung Electronics, Korea

ARTICLE INFO

Article history:
Received 9 September 2008
Received in revised form 20 April 2009
Accepted 10 May 2009
Available online xxxx

Responsible Editor: L. Jiang Xie

Keywords:
Wireless networks
WiMAX
Quality of Service
QoS
MAC
Media Access Control

ABSTRACT

We present a comprehensive survey of proposed Quality of Service (QoS) mechanisms in the Media Access Control (MAC) sublayer of WiMAX based wireless networks. QoS support in WiMAX is a fundamental design requirement, and is considerably more difficult than in wired networks, mainly because of the variable and unpredictable characteristics of wireless links.

We discuss various QoS architectures, signaling mechanisms and admission control techniques proposed in the WiMAX research literature, summarizing the operation of each, and providing comparative evaluations that include advantages and limitations.

1. Introduction

IEEE 802.16 Wireless Metropolitan Area Network Air Interface Standard [17,19] provides the details of physical layer and Media Access Control (MAC) sublayer of an advanced wireless communication system which aims to build a cost effective, multi-service network with WiMAX (Worldwide Interoperability for Microwave Access) technology. The standard published in 2004 [17] describes the physical and MAC sublayer specifications for fixed wireless access systems supporting multiple services. It consolidates the IEEE Standards 802.16, 802.16a, and 802.16c. The WiMAX Forum describes WiMAX as "a standards-based technology enabling the delivery of last mile wireless broadband access as an alternative to cable and DSL (Digital Subscriber Line)". The newer version of the standard, IEEE 802.16e-2005 [19], published in 2005 contains numerous revisions, adds higher layer handovers between base stations, as well as support for mobile terminals at vehicular speeds.

This standard is the outcome of a convergence of the market need and current wireless technological achievements, and is considered a benchmark solution for wireless metropolitan area networks (WMANs), as opposed to Wi-Fi wireless local area networks (WLANs). High data rates, large area of coverage, ease and cost effectiveness of deployments makes WiMAX suitable for a number of applications. This includes connecting multiple Wi-Fi hotspots, backhaul services and high speed mobile data communication.

When the aim is to provide a multi-service wireless network, a key challenge is the optimal allocation and utilization of the available raw data transmission capacity of shared wireless links among users and services. In this survey, we use the term 'bandwidth' to refer to the data transmission capacity of the links. Bandwidth utilization is considered optimum when there is no over- or under-allocation of capacity for a particular service type. Data transmission requirements depend on the type of services requested by a subscriber, and suboptimal distribution of

* Corresponding author. Tel.: +61 3 99053503; fax: +61 3 99053454.
E-mail address: Asekerci@ieee.org (Y. A. Şekercioğlu).
URL: http://titania.ctie.monash.edu.au (Y. A. Şekercioğlu).

1389-1286/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2009.05.001

available transmission capacity inevitably affects the service quality. Another factor is latency. Some services have greater tolerance for latency (e.g., FTP or e-mail), while others (like VoIP or video-conferencing) have strict delay bounds. Taking such QoS requirements into account, packet flows need to be prioritized via appropriate QoS management and scheduling methods. In broad terms, these always seek to achieve the optimal trade-off between the conflicting goals of maximized user performance and maximized system utilization.

WiMAX has a built-in QoS framework for real-time applications as well as data, which can take advantage of its polling architecture, and dynamically adaptable modulation in the physical layer. It should be noted that, while the standardized WiMAX QoS framework provides the details about the types of service flows that are supported, it does not explicitly define the actual packet mechanisms for achieving QoS differentiation in the MAC sublayer. As in other standards of this kind, such mechanisms are left open for vendor implementation, as long as they conform to the stated WiMAX QoS framework.

In Section 2, we provide an overview of this WiMAX-specific QoS framework, and separately consider point-to-multipoint and Mesh Network variants. Our work examines QoS implementation in WiMAX by subdividing the relevant issues into three distinct categories: packet scheduling and admission control (Section 3.1), signalling and internetworking (Section 3.2) and Mesh Network (Section 4). In each of these three sections we summarize the state of the art research activity along with results, possible implementation, drawbacks and scope for further development. In Section 6, an overall analysis is provided along with our concluding remarks.

2. QoS and the MAC sublayer of WiMAX networks

2.1. Definitions of Quality of Service

There are two broad definitions of Quality of Service (QoS):

> **User-Centric QoS** is "the collective effect of service performances which determine the degree of satisfaction of a user of the service" [22].
> **Network-Centric QoS** comprises "the mechanisms that give network managers the ability to control the mix of bandwidth, delay, variances in delay (jitter), and packet loss in the network in order to deliver a network service (e.g., voice over IP)" [9].

Our paper is primarily concerned with the second definition of QoS, and in the rest of this work, the term QoS shall be taken to mean "Network-Centric QoS".

Network engineers can use the existing resources efficiently by implementing a QoS mechanism. Early packet networks typically catered for one service type and all packets were treated equally. There was no QoS differentiation or guarantee of reliability, minimum latency, jitter or other performance characteristics (Table 1) for any set of packets. As a result of such a regime, a single bandwidth intensive application may cause the performance of other applications to degrade significantly. In a multi-service network, the QoS mechanism has to ensure that it can provide preferential delivery service to packets according to their performance requirements and QoS priority level, while maintaining a high network utilization. QoS differentiation can be implemented on either a per-application or per-user basis. With this in mind, QoS mechanisms can broadly be grouped under the following two categories:

Table 1
Network performance parameters and their characteristics maintained by QoS.

Network performance parameter	Characteristics
Latency	Delivery delay of a packet from source to destination
Jitter	Variation in latency
Reliability	The percentage of traffic that should be successfully delivered from source to destination to maintain the service quality
Data transmission rate	The amount of data that should be carried from source to destination in a given period of time to maintain the service quality

> **Admission Control** determines how and when the traffic generated by a given application or user can have access to the network resources. Typically operates at a session or flow timescale (i.e. decisions relate to admission of user sessions or flows).
> **Traffic Control** determines how packet marking, scheduling and shaping (flow rate control) is performed for packet traffic generated by a given application or user. Typically operates at packet timescales (i.e. decisions relate to which packet from which flow is to be transmitted next).

By implementing a well functioning QoS mechanism (or a combination thereof), network engineers can control available network resources to suit a particular requirement model, and to ensure that critical services are not affected by services of lower-priority. The end result is improved user experience, and reduced system cost due to more efficient and targeted use of available resources

2.2. QoS architecture of WiMAX

In wireless networks, including WiMAX, QoS support usually resides in the MAC sublayer because of the need to interact with radio resource management and physical layer dynamics. Fig. 1 shows the WiMAX QoS architecture as defined by the standard [19]. The base station (BS) has the responsibility of managing and maintaining the QoS for all packet transmissions. The BS manages this by dynamically distributing usage time to subscriber stations (SSs) through information embedded in the transmitted frames (Fig. 2). The figure only shows the TDD (time-division duplex) mode of operation in which BS-to-SS broadcasts (downlink subframe) are followed by SS-to-BS (uplink subframe) transmissions. It is also possible to use FDD (frequency-division duplex) mode of operation in

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

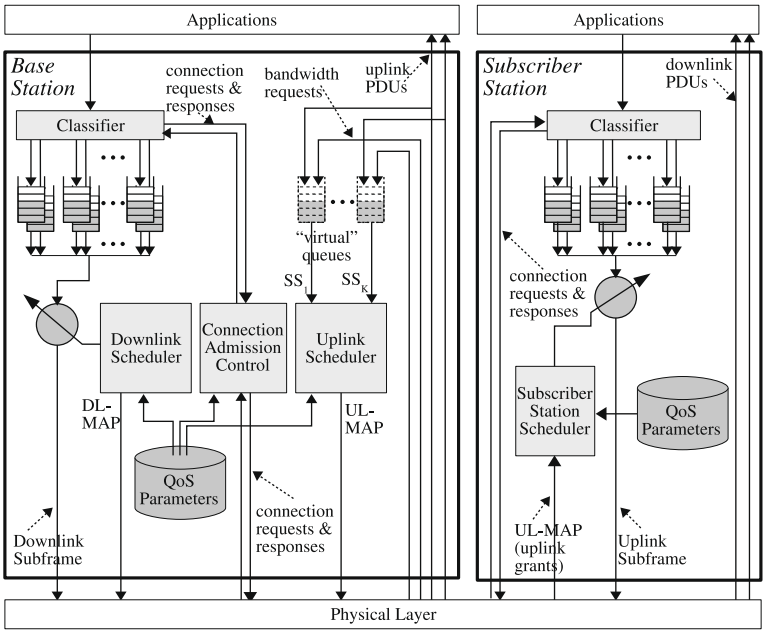


Fig. 1. Overall structure of the WiMAX QoS architecture.

which downlink and uplink subframes are transmitted in separate frequency bands, but the QoS management principles remain the same.

The downlink subframe contains two fields for managing allocation of the wireless medium: DL-MAP (downlink bandwidth allocation map) to tell the SSs of the timetable and physical layer properties for transmitting subsequent bursts of packets (the latter is referred to as the "Downlink Burst Profile" in WiMAX literature), and UL-MAP (uplink bandwidth allocation map) for regulating the uplink transmission rights of each SS. That is, the UL-MAP controls the amount of time each SS is given access to the channel in the immediately following or the next uplink subframe(s). A parameter called *Uplink Allocation Start Time* specifies for which uplink subframe the UL-MAP contents should be applied for. This flexibility allows an SS to have sufficient time to schedule uplink transmissions and prepare for the actual physical stream of data to be filled in the assigned uplink resource.

Uplink subframes contain three categories of fields:

- Initial ranging contention slot (denoted as "initial rang." in Fig. 2) is used by SSs to discover the optimum transmission power as well as timing and frequency offset to communicate with the BS. An SS begins the ranging process by sending a ranging request MAC sublayer message using the minimum transmission power. If it does not receive a response from the BS, it resends the message in the same field of a subsequent UL subframe using a higher transmission power.
- Bandwidth requests contention slot is used by SSs for transmitting bandwidth request (BW-REQ) MAC sublayer messages.
- Slots specifically allocated to the individual SSs for transmitting data.

The overall operation of the system can be summarized as follows:

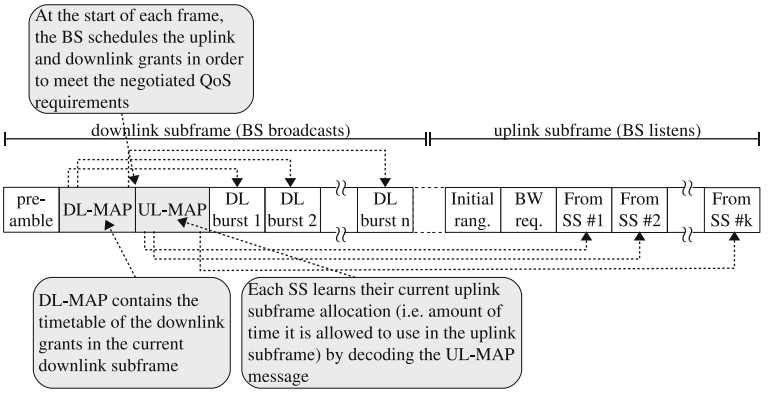


Fig. 2. Simplified WiMAX frame structure emphasizing its QoS management aspects.

The 802.16 MAC protocol is connection oriented. Signaling messages between the BS and an SS need to be exchanged in order to establish a "service flow"[1] between them. Service flows can be requested by the BS (according to the standard this is a mandatory capability), or by an SS (an optional capability). Each service flow is characterized by a range of parameters including three sets of QoS parameters indicating the required latency, jitter, and throughput assurances. These correspond to the three possible service flow states (*provisioned*, *admitted* and *active*), and are thus called *ProvisionedQoSParamSet*, *AdmittedQoSParamSet*, and *ActiveQoSParamSet*. Furthermore, each service flow is assigned a unique 32-bit long SFID (Service Flow Identifier) by the BS.

Service flows can be requested with any of these parameter sets being null. The 802.16 standard has two different kinds of call activation processes: a flow can be dynamically set up through DSA (dynamic service activation) transactions, or through a two-phase activation model similar to telephony applications. The former DSA-based approach is not expected to be available until dynamic QoS is introduced in the Network Release 1.5 of the IEEE 802.16e-2005 standard. The latter telephony-based approach supports the notion of static (i.e. nonprovisioned) QoS, and is available from the earlier Network Release 1.0 of the standard. Given static QoS, an arriving service flow typically has a non-null *ProvisionedQoSParamSet*, enters the *provisioned* state, and is allocated an SFID by the BS without being able to carry data packets until it is "activated".

Like a telephony call, a service flow goes through a transient "admitted" state, and changes to the "activated" state only after the end-to-end negotiation is completed. For each of the three states of a service flow different QoS parameter sets can be defined, but the set relationship

$$\textit{ActiveQoSParamSet} \subset \textit{AdmittedQoSParamSet} \subset \textit{ProvisionedQoSParamSet}$$

should always hold. Provisioned service flows become admitted or activated when their QoS requirements become known through the subsequently sent update messages containing non-null *AdmittedQoSParamSet* and *ActiveQoSParamSet* fields (though, it is not clear in the specifications that whether both sets should be non-null). If the QoS requirements of an active flow are included in an initial request, such a flow can be provisioned and immediately admitted/activated by the BS. For an activated service flow, the BS allocates a unique 16-bit connection identifier (CID). In this way, each BS-to-SS connection will have assigned to it as many CIDs as it has activated service flows (i.e. typically generated by separate active applications on the SS).

For SS-initiated communications, an SS, on behalf of an application, first requests a connection. The CAC (connection admission control) module located in the BS then checks whether the requested QoS parameters are within the limits of available resources, and if this is the case, the BS then responds with a unique SFID. BS-initiated communications work similarly, but in addition to the CAC checks, the BS, after allocating a unique SFID, also waits for the response of the SS indicating whether it can support the requested communication (the standard does not describe the internal criteria for an SS supporting or rejecting requests).

Various higher layer packet properties (such as IP addresses and protocol ports) are used for assigning the packets generated by the individual applications to specific SFIDs, and following activation, to corresponding CIDs. The assigned CIDs are then used to classify the resulting MAC frames into appropriate SS transmission queues. This functionality is encapsulated in the Convergence Sublayer (CS).

The scheduler of an SS visits the queues and selects packets for transmission. Selected packets are transmitted to the BS in the allocated time slots as defined in the UL-MAP, which is constructed by the BS Uplink Scheduler and broadcast by the BS to the SSs. It should be noted that in the IEEE 802.16e-2005 standard, the UL bandwidth grants do not specify the CID. That is, an SS is delegated decisions about scheduling multiple service flows belonging to it. This approach ensures that scheduling is left to the most appropriate node – an SS has queue state information which is more timely and accurate than the delayed "virtual queue" estimates available to the BS. Importantly, the absence of the CID in the UL bandwidth grants does not diminish the scheduling effectiveness of the SS for the following reasons. Firstly, the traffic priority QoS parameter specified at service flow creation/modification, governs the scheduling priority among service flow types other than UGS and system signalling messages (i.e., MAC management). Based on this QoS parameter, the SS and BS always have a common understanding of the order in which the station's flows should be scheduled, thus making the CID in the UL bandwidth grant unnecessary. Secondly, the BS knows when it expects system signalling messages, and so at these times will ensure that enough UL resources are assigned to an SS to cater for both the UGS and system signalling traffic. In terms of the relative priorities between the two, the BS and SS have a similarly common understanding, such that system signalling is generally afforded the highest priority. Again, the additional specification of CID in the UL grants is not needed. Similarly, the BS Downlink Scheduler selects the order in which it will transmit packets to the SSs and constructs a corresponding DL-MAP, as shown in Fig. 2.

2.2.1. QoS parameters, scheduling and data delivery services

After the admission of service flows, arguably the most complex aspect of the provision of QoS to individual packets is performed by the three schedulers: (i) the Downlink Scheduler which manages the BS-to-SS flows, and (ii) the Uplink and (iii) Subscriber Station Schedulers, which together manage the SS-to-BS flows. The Downlink Scheduler's task is relatively simple as compared to the Uplink Scheduler, since all downlink queues reside in the BS and their state is locally accessible to the scheduler. On the other hand, as the queues of uplink packet flows are distributed among the SSs, and their states and QoS requirements need

[1] A service flow signifies a unidirectional flow of packets that is provided a particular QoS, or in 802.16 terms, it "is a MAC transport service that provides unidirectional transport of uplink packets transmitted by the SS or to downlink packets transmitted by the BS".

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

to be obtained through bandwidth requests, the task of the Uplink Scheduler is much more complex (for this reason, we only include the uplink behavior in Table 2). The information gathered from the remote queues forms the operational basis of the Uplink Scheduler and is depicted as "virtual" queues in Fig. 1. None of the actual algorithms for the three schedulers are defined in the standard, and are instead left open to proprietary implementations.

In order to deal with the complexities associated with QoS provision to various applications in an ecosystem of different vendors' scheduler implementations, the standard defines a number of "scheduling service" and "data delivery service" classes. One of these classes is typically requested by an application, when its traffic flow goes through the stages outlined in Section 2.2. For each scheduling service class there is a corresponding data delivery service class (Table 2). The data delivery service classes are defined for and used with both uplink and downlink flows. In contrast, scheduling service classes are only used for uplink flows. The version of the standard published in 2004 [17] only covered the definitions of the scheduling services for uplink flows. During the discussions of the 802.16e [19] standardization process, the need to also define a scheduling service capability for downlink flows was brought up, and resulted in the introduction of data delivery services. For backward compatibility, the scheduling services already defined in the 2004 version of 802.16 [17] were also retained. The set of QoS parameters associated with a scheduling service and/or a data delivery service are almost identical, and the only reason both still remain in the specifications is that of historical standard evolution.

When a specific scheduling or data delivery service is associated with a service flow, that flow is further associated with a certain pre-defined set of QoS parameters. However, according to the standard this does not include assignment of specific values to the parameters, which is managed using dynamic service addition (DSA) and dynamic service change (DSC) messages.

2.3. Point-to-multipoint vs mesh WiMAX networks

In a mesh WiMAX network, a "Mesh BS" (mesh base station) provides the external backhaul link. The backhaul links connect the WiMAX network to other communication networks. There can be multiple Mesh BSs in a network; other nodes are known as "Mesh SSs" (mesh subscriber stations).

The sectorized antenna used by the BS in a WiMAX cell is capable of splitting its coverage area into separate subfields and managing transmissions simultaneously and independently in each. The MAC sublayer uses these antenna properties to control data transmission between the BS and SSs to optimize the channel utilization.

As discussed earlier, in point-to-multipoint mode, the SS transmissions are controlled directly by the BS. In Mesh mode the uplink and downlink is not clearly separated, and SSs can communicate with each other without communicating with the BS. Fig. 3 shows the frame structure in Mesh WiMAX networks. Similar to point-to-multipoint WiMAX networks, data transfer is connection oriented. Connection setup can be achieved using either of the following two scheduling schemes

Table 2
WiMAX scheduling and data delivery service classes, their typical usage, and BS and SS behaviors for uplink. The rtPS, nrtPS and BE scheduling services can use the piggybacked bandwidth (BW) request method in addition to the special purpose BW request PDUs. Piggybacked BW requests are signalled by populating the grant management subheader fields in the generic MAC PDUs [19].

Scheduling service	Corresponding data delivery service	Typical applications	Uplink		QoS specifications
			BS behaviour	SS behaviour	
Unsolicited grant service (UGS)	Unsolicited grant service (UGS)	Voice (VoIP) without silence suppression	The BS uplink scheduler offers fixed size UL BW grants on a real-time periodic basis	An SS does not need to send any explicit UL BW requests	Maximum sustained rate Maximum latency tolerance Jitter tolerance
Extended real-time polling service (ertPS)	Extended real-time variable-rate service (ERT-VR)	VoIP with silence suppression	The BS uplink scheduler offers real-time, periodic, UL BW request opportunities (similar to UGS, but ertPS allocations are dynamic, not fixed)	An SS uses the offered opportunity to specify the desired UL BW grant	Maximum sustained rate Minimum reserved rate Maximum latency tolerance Jitter tolerance Traffic priority
Real-time polling service (rtPS)	Real-time variable-rate service (RT-VR)	Streaming audio or video	The BS uplink scheduler offers real-time, periodic, UL BW request opportunities	An SS can use (a) the offered opportunity to specify the desired UL BW grant, or (b) piggybacked BW request opportunities (an SS can not use contention based BW requests)	Maximum sustained rate Minimum reserved rate Maximum latency tolerance Traffic priority
Non-real-time polling service (nrtPS)	Non-real-time variable rate service (NRT-VR)	File transfers	The BS uplink scheduler provides timely (in the order of a second or less) UL BW request opportunities	An SS can use (a) offered uplink, or (b) contention-based, or (c) piggybacked BW request opportunities	Maximum sustained rate Minimum reserved rate Traffic priority
Best-effort service (BE)	Best-effort service (BE)	Web browsing, email	The BS does not specifically offer any UL BW opportunity	An SS can use (a) contention-based, or (b) piggybacked BW request opportunities	Maximum sustained rate Traffic priority

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

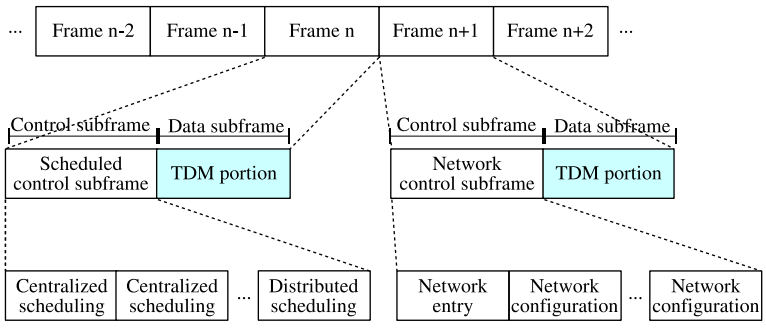


Fig. 3. Frame structure of the mesh WiMAX networks.

Centralized Scheduling (**Mesh CS**): the Mesh-BS has the responsibility of granting resources for each link in response to resource requests. Mesh centralized scheduling messages transmitted in a scheduled control subframe (Fig. 3) are used for this purpose.
Distributed Scheduling (**Mesh DS**): The neighboring Mesh SS responds to a request with a corresponding grant for a link between two Mesh SSs. Mesh distributed scheduling messages are exchanged to perform this operation.

In contrast with point-to-multipoint WiMAX networks, the standard does not define scheduling services for Mesh WiMAX networks.

Network control subframes periodically appear and are used for servicing the new nodes which want to gain access to the network. The transmission opportunities in a control subframe and the data minislots in a data subframe are separated. The nodes compete for the control channel access and the contention outcome does not effect the data transmission.

3. Mechanisms for QoS provision in point-to-multipoint WiMAX networks

Research studies conducted in point-to-multipoint WiMAX networks can be classified into two broad categories:

Admission control and packet scheduling research focuses on the implementation of the admission control and scheduling services (Uplink and Downlink Schedulers at the BS, as well as the SS scheduler).
Signaling and internetworking research into methods through which signalling can be improved (Section 3.2) and internetworking between WiMAX and other networks (e.g., fiber backhaul and Wi-Fi access).

In the following sections, we discuss the research work proposed for each category.

3.1. Admission control and packet scheduling

A considerable number of studies may be found in the research literature on algorithms and methods for scheduling services and admission control, in the context of WiMAX point-to-multipoint networks.

3.1.1. A study of QoS support in 802.16 networks

In one of the early studies on QoS support in WiMAX networks, Cicconetti et al. focus on the available QoS support mechanisms in the MAC sublayer and evaluate their effectiveness through simulation [7]. They conduct the performance evaluation based on two common application scenarios conceived by the WiMAX Forum [12]: residential, and small to medium-size enterprises (SME). The test case uses 7 MHz channel bandwidth with carrier frequency between 2 and 11 GHz and operating in FDD mode. In the study, it is also assumed that frame duration is 10 ms, all SSs have full duplex capability, and channel conditions are ideal.

Since the actual implementations of the SS scheduler, and downlink and uplink schedulers of the BS are not included in the standard, the authors needed to choose appropriate algorithms for them. They note that the basic QoS parameter negotiated for a connection within a scheduling service is the minimum reserved rate, and because of this, they argue that the class of rate-latency scheduling algorithms [35] are suitable for implementing the schedulers.

Within this class, the authors have chosen Deficit Round Robin (DRR) [34] algorithm for implementation of the downlink scheduler of the BS. They justify this selection to DRR's ability to maintain fair queueing when packet size is variable and its ease of implementation. But, DRR can not be used for the uplink scheduler since it needs to know the size of the packet at the head of each queue for its operation. The BS, through the virtual queues (see Fig. 1) can only estimate the uplink load but not the packet sizes, which is not sufficient for operation of DRR. Because of this, the authors have selected Weighted Round Robin (WRR) [25] algorithm (which also belongs to the class of rate-latency scheduling algorithms). Their choice for the implementation of the SS scheduler remains as DRR, because an SS always knows the sizes of the packets waiting at the head of its packet queues.

In the Residential Scenario, the BS only provides Internet connectivity to the SSs and all traffic is of BE class. The results show that as long the network is lightly loaded the connection queues are almost empty. The average delay increases sharply as soon as the system starts to get overloaded. When overloaded, the average delay of uplink traffic becomes greater than the downlink traffic.

In the SME Scenario, the BS caters for various types of services like VoIP, video or data. It assumes that VoIP and

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

video traffic is classified as rtPS and data as best-effort. VoIP is provided with a greater reserved rate than video. The results show that as the number of active SSs increases, the downlink delay increases smoothly for all classes of traffic. However, as the network gets overloaded (i.e., the number of subscribers is more than 30), there is a sharp rise in the delay for BE traffic, but delay for VoIP and video is unchanged. This happens because of the way in which capacity has been provisioned to different connections. The scheduling algorithm is configured such that rtPS connections have a reserved rate equal to the mean rate of VoIP and video traffic respectively. The guaranteed rate for BE is negligible compared to rtPS connections. Further increases in the load show a rise for delay in video traffic but not for VoIP, due to its greater reserved rate. the same behavior is observed in the uplink.

The uplink traffic delay variation is greater than downlink traffic when the system is not in overload (fewer than 24 SSs), but lower when the system is overloaded (number of SSs is between 24 and 36). This happens for the following reason: when the system is not overloaded, the BS issues an uplink grant as soon as it receives the bandwidth request. But, when the system is overloaded, applications at the SSs generate the next packet before the uplink grant arrives from the BS for the previous packet. Therefore, the SSs are able to piggyback the bandwidth request for the next packet on the current outgoing packet and reduce the delay (and delay variation). The results show that when the number of subscribers exceeds 36, this phenomenon cannot compensate further and the delay variation curve begins to increase.

3.1.2. A scheduling algorithm and admission control method

Wongthavarawat and Ganz propose an implementation of an uplink packet scheduling (UPS) and admission control framework [38] at the BS, and a Traffic Policing module at the SS (Fig. 4). Using simulation methods, the authors show that their proposal yields an improvement in system performance over a "default" case without this functionality.

As mentioned earlier, in the standard WiMAX QoS architecture (Fig. 1), details of both the admission control and uplink scheduling at the BS are undefined, with their implementation left to vendors. The same holds at the SS, where neither traffic policing module nor its interaction with the BS admission control are defined. The SS scheduler receives the UL-MAP from the BS after a bandwidth request is made to the BS UPS module; however, the specific policy that will be used in the UPS module is undefined in the standard WiMAX QoS architecture.

Fig. 4 shows a sketch of the proposed implementation. in the context of the original WiMAX QoS architecture, with key interactions and information flows clearly marked. At the BS, an admission control module and detailed UPS module are introduced. A traffic policing module is included at the SS. When an application at the SS originates a connection request to the BS, it includes the bandwidth and delay requirement in the request message. The admission control module accepts or rejects this request based on its traffic policy. If the request is accepted, it notifies the BS UPS module and provides appropriate parameters. After receiving the parameters, the SS traffic policing module ensures that traffic is classified based on the traffic contract. The information module of the UPS collects the queue size information from the BW-request messages received from the previous time frame. This is used by the information module to update the scheduling database module. The UL-MAP is generated by the service assignment module after information is received from the scheduling database module. The UL-MAP is broadcast to all SSs

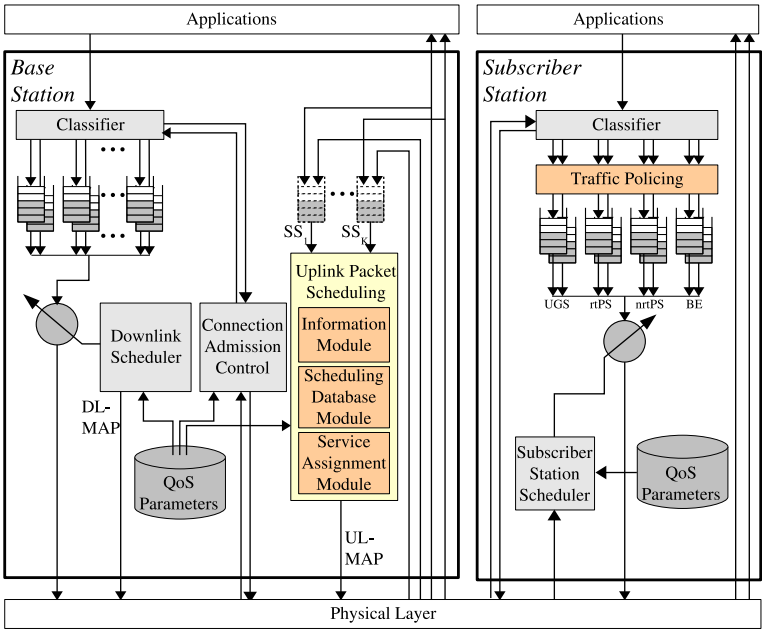


Fig. 4. QoS architecture proposed by Wongthavarawat and Ganz [38].

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

and based on this, the UPS of an individual SS transmits the packets.

The authors explain in detail the information, scheduling database and service assignment modules, describing the steps taken for each class of traffic considering various factors like queue size, arrival time and delay threshold. Appropriate algorithms are also proposed as implementations of each of these three modules.

To admit a new connection, the admission control mechanism enforces procedures based on the scheduling class of traffic, which we summarize below:

Procedure for UGS: On arrival of a new request, it checks for the available bandwidth. There is no check required for delay. However, it checks whether accepting this request will cause any delay violation for the existing rtPS connections. If there is no violation, the connection is accepted.
Procedure for rtPS: First it checks for the available bandwidth. If the bandwidth is available, then checks if delay guarantees can be maintained. It also checks for any delay violations for the existing rtPS connections. If these conditions are met, then the connection is accepted.
Procedure for nrtPS: It only checks for the available bandwidth. There is no need to check for the delay violation for existing rtPS (or the lower-priority BE) connections.
Procedure for BE: No admission control process is required. They are always admitted, but do not receive QoS support.

The simulation study only assumes that there are two kinds of traffic, rtPS and BE. Each connection has specific QoS parameters in terms of average bandwidth requirement which is equal to the token bucket rate, and maximum delay requirement. The authors present the outcomes of the study in three graphs: the arrival curve which depicts the arrival pattern of the input traffic, the service curve which shows the service pattern provided by UPS, and the percentage of packets that miss their deadline. The downlink and uplink capacity is set to 5 MBps each, frame size is set to 10 ms. For rtPS, there are three sessions each with a bandwidth of 3 Mbps.

For the first experiment, the combined bandwidth for rtPS and BE connection is 5 MBps and the results show that none of the packets miss their deadline. The second experiment shows the arrival and service curves of all three rtPS connections. The graphs show that the service curve adapts and follows the arrival curve for all three sessions. As none of the packets miss their deadline, the delay is also guaranteed.

3.1.3. A hard and soft server scheduling mechanism

Inspired by an earlier study [3], Shang and Cheng propose a hierarchical packet scheduling model for WiMAX uplink by introducing the "soft-QoS" and "hard-QoS" concepts [32]. rtPS and nrtPS traffic are classified as soft-QoS because their bandwidth requirement varies between the minimum and maximum bandwidth available for a connection. UGS traffic is classified as hard-QoS since it requires the maximum bandwidth available for the connection. By allowing the BE traffic to be scheduled by the BS, the model is able to distribute bandwidth between BE and other classes of traffic efficiently and guarantees fairness among the QoS-supported traffic (UGS, rtPS and nrtPS). A delay comparison performance evaluation is provided between the models.

The study by Bennett and Zhang [3] proposes the worst case fair weighted fair queueing (H-WF2Q+) scheduling framework. Based on some criteria (the authors do not specify the criteria) it distributes weighted bandwidth to different sets of flows. However, this model is not suitable for multimedia traffic as it does not take into consideration its diverse traffic requirements. In the proposed model packet scheduling takes place in the BS uplink. As shown in Fig. 5, each traffic class is assigned to three logical scheduling servers (hard-QoS server, soft-QoS server and best-effort server). UGS traffic is routed through the hard-QoS server, rtPS and nrtPS through the soft-QoS server and BE through the best-effort server. The capacity of each server is allocated by a pre-defined algorithm. There is a provision for soft-QoS traffic to be scheduled by best-effort server. This enables it to obtain additional bandwidth. The packet scheduling algorithm comprises of four parts:

(1) hard-QoS server scheduling
(2) soft-QoS server scheduling
(3) best-effort server scheduling
(4) co-scheduling among the above three servers

A detailed algorithm for each server is shown along with a delay comparison between the initial and this developed model. The difference between the two models is the treatment of the soft-QoS traffic. This changes the tree-like structure to a two-level hierarchical structure. The results show an improvement in delay and the soft-QoS and BE traffic is able to obtain greater share of bandwidth by minimizing bandwidth wastage. Based on the network dynamics, the servers are able to change their weights for different traffic loads. It also proves that the

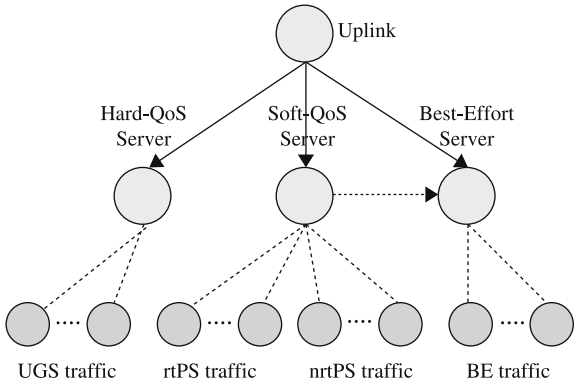


Fig. 5. The hierarchical packet scheduling model of the uplink in IEEE 802.16 as proposed in [32]. Each traffic class is assigned to three logical scheduling servers. There is a provision for soft-QoS traffic to be scheduled by the best-effort server to obtain additional bandwidth.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

hierarchical model can guarantee lower delay and delay jitter for variable bit-rate traffic unlike the model presented in [3].

3.1.4. A method for MAC flexibility exploitation for multimedia streaming

The study of Sengupta et al. [33] investigates the mechanisms through which MAC PDUs can be continuously modified based on the feedback obtained through the channel state information. The scheme changes the payload size by aggregation or fragmentation of the upper layer PDUs. By adopting the dynamic MAC PDU approach, the study shows performance enhancements for streaming various types of media.

The idea of a dynamic payload size based on near-instantaneous channel state information has been used in other wireless broadband technologies, such as 3GPP High Speed Downlink Packet Access (HSDPA) [28], albeit at lower layers of the protocol stack. Most typically, this approach is adopted at the physical layer, an example being the Adaptive Modulation and Coding technique employed in HSDPA.

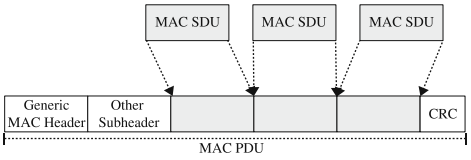
The scheme presented in the study works this way: when an SS requests media content, the media server located in the core network transmits the raw data to the WiMAX gateway. The encoder at the BS receives this raw data and pushes it to the MAC sublayer. Depending on the channel state of the SS, the scheduler at the MAC sublayer manipulates the MAC SDUs to construct the MAC PDUs. A feedback mechanism placed at a receiver's MAC sublayer is the core of this scheme. Based on the feedback signals generated, the transmitting side modifies the MAC PDU payload size. By changing the MAC PDU size dynamically, the system attempts to match packet transmissions to the underlying radio channel conditions. This results in reduction of the number of dropped or corrupted packets and retransmissions, and eventually achieves reduced delays and increased overall network throughput. In the authors' scheme, ARQ mechanism is used for recovering the corrupted transmissions and is an integral part of estimating the channel conditions.

Fig. 6 illustrates how multiple MAC SDUs can be concatenated to a single MAC PDU or how a single MAC SDU can be fragmented and distributed over multiple MAC PDUs.

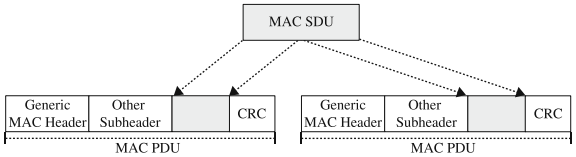
The connection setup and transmission takes place in three phases. First, the SS makes a connection request. This enables the BS to detect the initial ranging, and measure the timing/power offset. This is followed by the service flow parameter request, and at this point, the variable length MAC SDU indicators are turned on. Second, the BS confirms the connection by responding with a response message that has the initial ranging, power adjustment information for the SS. The service flow adjustments are negotiated and the SS is provided with a CID. Finally, MAC SDUs obtained from the MAC convergence sublayer are transmitted through the MAC PDU payload. Depending on the channel requirements, the MAC SDUs can be fragmented or aggregated at the start of transmission. Feedback is received after the first transmission, and the next MAC PDU payload size is changed accordingly. There are six different feedback possibilities and Table 3 shows the actions taken by the BS when each type of feedback is received.

Simulation based experiments were conducted over a channel model with various bit error probabilities for experimenting with a range of channel conditions from "good", "fair", "medium" and "bad" (with simulated bit error rates of 0.045, 0.06, 0.07 and 0.085, respectively). The experiments consider mechanisms with or without feedback, and comparative results are presented in the paper. The authors first compare the packet restore probability (PRP) over time for MAC PDUs whose sizes are either kept constant or adaptively modified as described in the paper. Although the authors do not provide a quantitative analysis of the results, the graphs show 70–80% improvement for the adaptive scheme. By studying the graph we can observe that, for the non-adaptive scheme, the PRP reaches zero in a 30 ms time frame 15 times, whereas for the adaptive scheme, the worse case scenario occurs just once.

The goodput (the ratio of information bits to total bits transmitted) for the non-adaptive scheme is about 77% when the channel error rate is approximately 1% and this gradually drops to around 63% as the channel error rate increases to 20%. For the adaptive scheme, the goodput is 85% and 82%, respectively, showing an improvement of 8–20%. The most significant improvement is observed with the MAC PDU drop rate. With the non-adaptive scheme, as the channel error rate increases from 1% to 20%, the MAC PDU drop increases from 1.5% to 18%. However, with the



(a) Multiple SDUs (Service Data Units) in a single PDU (Protocol Data Unit).



(b) Single SDU split into multiple PDUs.

Fig. 6. Packet aggregation and segregation in the adaptive multimedia streaming scheme for WiMAX networks proposed by Sengupta et al. [33].

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

Table 3
Feedback types and associated BS responses in the adaptive multimedia streaming scheme for WiMAX networks proposed by Sengupta et al. [33].

Feedback type	Feedback classification	Action taken by base station
1	MAC PDU received correctly	(1) Increase MAC PDU payload (2) Decrease CRC for not so important MAC PDU
2	MAC PDU received with errors, and uncorrectable	(1) Increase CEC for important MAC PDU (2) Keep payload and CRC fixed for not-so-important MAC PDU
3	MAC PDU received with errors, but correctable	(1) Decrease payload for MAC PDU (2) Increase CRC of MAC PDU
4	MAC PDU dropped, timeout in receiver MAC occurred	Same as feedback type 3, but the increment/decrement is more pronounced
5	Receiver MAC buffer full, last stored frame is important	Stall transmission until further request received
6	Receiver MAC buffer full, last stored frame is not so important	(1) Skip transmission of next few not so important frames (2) Important frame(s) is/are transmitted

adaptive scheme, the MAC PDU drop rate never increases beyond 1.5%.

3.1.5. A two-tier scheduling algorithm

A hierarchical QoS architecture is proposed in Chan et al.'s study [4] that implements a two-tier scheduling algorithm (2TSA) at the BS. The first tier is based on the connection category and the second tier is weight-based. The study considers TDD operating mode of WiMAX physical layer and assumes the uplink subframe occupies half a frame time. As UGS connection is automatically allocated per frame, 2TSA does the scheduling for rtPS, nrtPS and BE traffic.

2TSA implements a simple service category for each connection that is based on the allocated bandwidth. The categories are:

Unsatisfied: A connection receives less bandwidth than the minimum requirement or reserved rate.
Satisfied: A connection receives bandwidth that is more than the minimum requirement, but less than the maximum bandwidth sustained rate.
Over-Satisfied: A connection receives more bandwidth than the specified maximum requirement.

Based on the service category, each connection is given a weight between 0 and 1. For example, if the allocated bandwidth of a connection is less than its minimum demand, its weight indicates the shortage compared to this demand. Similarly, weights of the other two categories indicate the corresponding satisfaction degree. Fig. 7 shows the flowchart of the proposed 2TSA. The functionality of each tier can be summarized as below:

First-Tier Allocation: The BS classifies all connections into three categories based on the collected bandwidth request and updated weights. 2TSA then allocates the bandwidth first to the "unsatisfied", followed by the "satisfied" and finally to the "over-satisfied" categories.
Second-Tier Allocation: For each specific category, the received bandwidth is further distributed to the connections based on the value of the weight parameter. Connections with smaller weights are given higher priority.

After completing the two-tier bandwidth allocation, the BS generates the corresponding UL-MAP and broadcasts to all SSs.

The authors investigated the performance of the scheme through the simulation of a WiMAX network which has 5 UGS and 7 rtPS, nrtPS and BE connections served by a BS. The simulation has two scenarios.

- In the first scenario the total available uplink bandwidth is 8 Mbps, and the sum of all connections' maximum sustained rate exceeds 8 Mbps. The results show that no matter how much traffic a connection generates, 2TSA can guarantee each connection its minimum bandwidth demand and fairly distribute the residual bandwidth to all connections (0.1 Mbps to each). This is in contrast with strict-priority scheduling (SPS) proposed in [37], where nrtPS and BE connections begin to starve under same conditions. This is because SPS always allocates rtPS connections first.
- In the second scenario, the total available bandwidth is 12 Mbps (which is greater than the total maximum sustained rate of the connections). This experiment were conducted to evaluate how fairly the residual bandwidth is allocated in 2TSA compared to SPS. The results presented demonstrate that the residual bandwidth is distributed to all connections after maximum sustained rates are allocated. In contrast, nrtPS and BE connections get starved when SPS algorithm is used.

3.1.6. A scheduling architecture for improving delay and throughput

In the study [31], the authors propose a scheduling architecture in order to improve the delay and throughput for rtPS connections, which is an extension of an earlier research work [6]. The previous work implemented a two-layer scheduling structure for bandwidth allocation to support all types of service flows. Direct Fair Priority Queue (DFPQ) was used in the first layer to distribute total bandwidth among flow services in different queues (6 in total depending on service class and direction) as shown in Fig. 8. In the second layer of [6] various scheduling algorithms are used for each class of traffic. For rtPS connections the packet with the earliest deadline is scheduled first [14]. Weight based scheduling algorithm [8] is used for nrtPS connections and round robin scheduling algorithm [15] for BE traffic. The paper then proposes the new scheduling technique and presents three different scenarios for its implementation.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

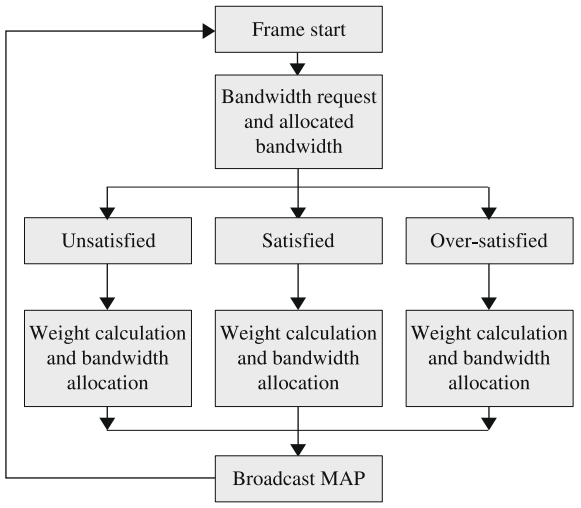


Fig. 7. Operational flowchart of the 2TSA scheduling algorithm [4].

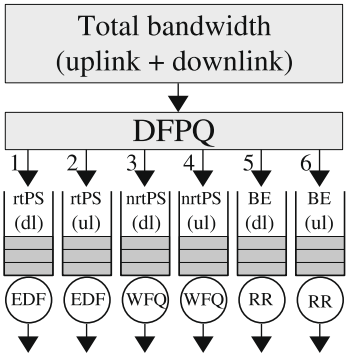


Fig. 8. Deficit Fair Priority Queue (DFPQ) bandwidth allocation method proposed by Chen et al. [6]. For supporting all types of service flows, a hierarchical scheduling structure of the bandwidth allocation is proposed for TDD mode (RR: round robin, EDF: earliest deadline first, WFQ: weighted fair queue).

The proposed architecture is designed to provide rtPS service flow packets more chance to meet their deadline and decrease the delay. Apart from checking if the available bandwidth is enough for granting a request, the system has to monitor nominal polling interval, nominal polling jitter and reference time (the time used as a reference to calculate both the generation time and the deadline of the rtPS data grants) related to the rtPS service flows that are admitted. The information gathered from this monitoring is used to approximate the expected delay of each rtPS connection and the proposed scheduling algorithm, similar to [16], is used to calculate the deadline. This deadline is used by the scheduler to determine if an rtPS packet is critical or not. Preemptive Direct Fair Priority Queue (PDFPQ) is proposed for the first layer scheduling and total bandwidth distribution. The structure is almost identical to the one shown in Fig. 8, the only difference being the DFPQ in the first layer is replaced by PDFPQ, maintaining four lists:

Active List contains non-empty queues whose deficit counter values are greater than zero.
Blocked List contains non-empty queues whose deficit counter values are either zero or negative.
Waiting List contains queues that are empty and their deficit counter values are greater than zero.
Non-active List contains queues that are empty and their deficit counter values are either zero or negative.

The rtPS queues, both uplink and downlink, are non-preemptive queues. Other queues can be preempted under certain conditions. If an rtPS packet has a deadline to meet, but will probably fail, then that packet is considered *critical*. The PDFPQ defines a "quantum critical" value for each non-preemptive queue. Queues are allowed to use this value to serve critical packets only. This gives a queue another chance to service critical packets. There are three scenarios that are not handled in the original DFPQ method:

Scenario 1: *A critical packet arrives to the waiting list of the non-preemptive queue, while the scheduler is serving packets from one of the preemptive queues.* Under this situation the DFPQ algorithm will most likely service the critical packet with the next frame. This will cause the packet to fail meeting its deadline.
Scenario 2: *The deficit counter becomes less than or equal to zero while the scheduler is processing the packets of a non-preemptive queue.* If a critical packet is waiting to be serviced at the head of the queue, the DFPQ algorithm will not service the packet in the current round.
Scenario 3: *A critical packet arrives to the inactive list of the non-preemptive queue, while the scheduler is serving packets from one of the preemptive queues.* The packet will be served by the DFPQ algorithm. However, PDFPQ will not serve packets in the inactive list.

These scenarios are addressed in the Preemptive DFPQ algorithm proposed by the authors.

The simulation compares the improvement in delay and throughput when using PDFPQ over DFPQ. Some assumptions are made, such as total bandwidth is 10 Mbps and each frame duration is 1 ms. The authors simulated the behavior for four frames, each divided into several rtPS and BE packets. DFPQ and PDFPQ were applied to all the above mentioned scenarios and the minimum, maximum and average delay were measured and reported for 4 ms (four frames). There is no change in the maximum delay for both the algorithms. Minimum delay improves by 800 μs in frame number two and four when PDFPQ is used. This 800 μs is a significant amount considering the maximum delay recorded is 3600 μs. For the first and third frames, the minimum delay improves by 200 μs. This change in minimum delay naturally affects the average delay accordingly. Consequently, the results show that PDFPQ algorithm reduces the delay of critical packets that could not have possibly been serviced using the DFPQ algorithm.

Throughput of rtPS and BE service flows were also compared for both DFPQ and PDFPQ algorithms. The results

show that, for DFPQ, the throughput for rtPS and BE service flows are almost at a constant level (negligible change) for the simulation duration. However, when PDFPQ algorithm is implemented, throughput for rtPS in the first and third frames increases. This increase is directly proportional to the decrease in throughput of BE service flow for the respective frames. The authors claim that this decrease in BE service flow is insignificantly small and it will never experience starvation.

The simulation results are convincing at face value, but the simulation is run only for four frames. To observe the improvement in average delay, simulations should be conducted over a large number of frames. Experiments conducted over a longer period will also demonstrate if BE service flow actually survive starvation when PDFPQ is implemented.

3.2. Signaling and internetworking

In this section, we discuss the research efforts focusing on the QoS signaling mechanism in the MAC sublayer and internetworking issues with other networks (such as optical and Wi-Fi). The studies covered here propose various ways to improve QoS signaling and create hybrid architectures for improving inter-connectivity with existing networks.

3.2.1. An integrated signaling mechanism

A fast signaling mechanism proposed by Chen et al. [5] modifies the default signaling mechanism of WiMAX to enable the system to reduce the initial connection setup time. The WiMAX standard specifies that service flows can be dynamically added, changed or deleted (DSA, DSC and DSD messages) and these actions can take a number of handshakes between an SS and the BS. In contrast to the default architecture, in the authors' proposed system, the SS sends the DSA message embedded with the BW request messages. This is illustrated in Fig. 9, where, in the context of IntServ architecture [2], the sender initially transmits a PATH message that includes the traffic specification (TSpec) information, consisting of bandwidth, jitter and delay requirements. This information then can be embedded in the subsequent DSA request message. Similarly, the DSA response message can contain additional information such as allocated bandwidth. When a new service flow arrives, the admission control mechanism accepts it if the requested bandwidth is less than the available bandwidth (the difference between the total capacity and the sum of all current connections). Under the default architecture, the negotiation of QoS parameters between the BS and an SS takes place twice – a situation which is avoided in the authors' enhanced signalling proposal.

The authors developed a simulation platform for evaluating their proposal. The simulated network used for evaluation consists of one BS and three SSs. The total bandwidth is 10 Mbps and frame duration is 10 ms, which is divided into 256 minislots. For management, basic, primary and secondary connections, 1 Mbps of bandwidth is reserved. DSA, DSC and DSD message transmission delays are set to 10 ms (even though admission control and reservation related processing time can vary due to performance of the BSs and SSs, for the purpose of this simulation work it is fixed at 10 ms).

The graphs presented in the study illustrate that the setup time for the proposed signalling implementation is insensitive to offered load; it remains unchanged at approximately 75 ms as the rate of frame arrival increases from one to three per time unit. Conversely, with unmodified (traditional) WiMAX signalling, the minimum setup time starts at just over 100 ms, and increases to around 200 ms and 700 ms for frame arrival rates of two and three per time unit, respectively. This shows a significant improvement in the setup time.

3.2.2. WiMAX and optical network integration

The study presented in [27] proposes a bandwidth allocation scheme for Video-on-Demand (VoD) services over an experimental integrated optical and WiMAX network. The end-to-end connection between the VoD client and server is distributed over Synchronous Optical Network (SONET) and WiMAX links. The SONET ring is the backbone used for connecting the WiMAX BSs and VoD clients.

As WiMAX BSs can cater for up to 75 Mbps data rate (shared among all users), if only one STS-1 link is provided to each BS node, congestion will be experienced whenever total user demand per BS exceeds the STS-1 data rate of 51.84 Mbps. If two links are provided, that will make the system less efficient and not cost effective. This research proposes a solution that overcomes these obstacles: to use one STS-1 link per BS and shift system operation between an Erlang-C and an Erlang-B queueing model, depending on the network load. The three possible scenarios are:

(1) Average offered load is less than the link capacity (single STS-1 circuit): All requests are queued and served accordingly. The behavior of the BS subnet is characterized by the Erlang-C delay model.

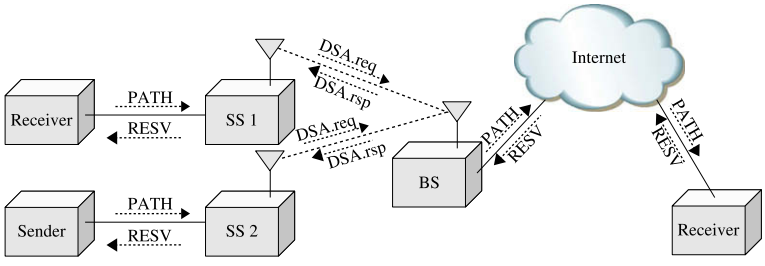


Fig. 9. Traffic specification (TSpec) information is embedded in PATH and DSA.req messages in the proposed mechanism by Chen et al. [5].

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

(2) Average offered load is greater than the capacity of a single STS-1 circuit: Queueing of the infinite number of requests makes the system unstable. Hence, extra packets need to be dropped, and the behavior of the BS subnet is characterized by the Erlang-B delay model.
(3) Average offered load is greater than the capacity of a single STS-1 circuit but less than two STS-1 circuits: It is reasonable now to queue all unsatisfied requests. The behavior of the BS subnet then follows the Erlang-C model.

The experiment looks into two possible heuristic solutions:

Maximum Utilization: The algorithm picks the BS in the non-increasing order of utility and sequentially allocate sufficient bandwidth to it.
Maximum Efficiency: The algorithm picks the BS with the maximum cost effectiveness first. Cost effectiveness is defined by the larger ratio of the utility over the bandwidth needed between the two types of bandwidth allocation.

In the simulation, 1000 nodes are used and the arrival requests follow the M/M/m models and each VoD request consumes 1 Mbps. The results are presented in a graph (aggregate utility value against the VoD server capacity). The results show that when the capacity is small, simple greedy approach does not work well but the algorithm proposed optimizes the utility function and performs better. Although the authors did not provide a quantitative analysis, our study of the graph shows an improvement by 25% when the capacity is small. The results also show that Maximum Efficiency heuristic is not sensitive to capacity variations and outperforms other greedy algorithms.

3.2.3. WiMAX and Wi-Fi integration via mapped QoS classes

The study of Gakhar et al. [13] proposes an architecture to achieve differentiated QoS for end-to-end services in a hybrid WiMAX and Wi-Fi (802.11e) network. It maps QoS requirements of an application that originates from a Wi-Fi network to a WiMAX network and assures transfer of data with appropriate QoS.

802.11 a/b/g offers best-effort service only. In contrast, the 802.11e [18] was designed to ensure QoS differentiation among packet flows generated by applications. It introduces the Hybrid Coordination Function (HCF) which enhances the DCF and PCF access schemes of 802.11. HCF multiplexes between two channel access methods for sharing the medium: Enhanced Distributed Channel Access (EDCA) which is a decentralized algorithm, and a centralized algorithm called HCF Controlled Access (HCCA) for tightly controlled frame transmissions. Varying degrees of QoS at the MAC sublayer of 802.11e can be provided by either of these mechanisms [24]:

Prioritized QoS through service differentiation with EDCA: Frames are segregated into classes, and frames belonging to the same class receive best-effort-within-class service while different classes receive different grades of service in aggregate. Absolute guarantees of QoS parameters like delay or loss are not provided. Thus, this is also called "better than best-effort" service and is suitable for elastic traffic.
Parameterized QoS through per-flow time reservation with HCCA: MAC-level flows are defined and each flow is guaranteed a certain fraction of time during which a node (AP or Wi-Fi client) can transmit. The standard also includes means for admission control and reservation signaling at MAC level between a Wi-Fi client and AP. This mechanism provides tightly controlled QoS.

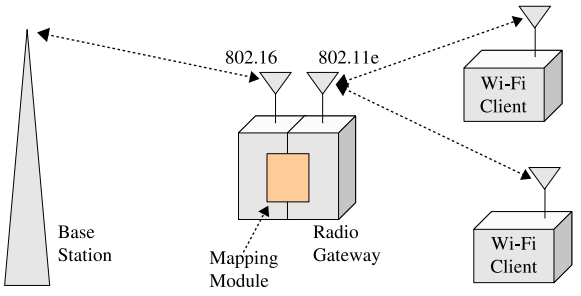


Fig. 10. The architecture proposed by Gakhar et al. [13]. The radio gateway module functions as an SS for the IEEE 802.16 network and an access point for the IEEE 802.11e wireless LAN.

In 802.11e, classification of traffic is achieved through the introduction of access categories for EDCA, and traffic streams for HCCA. HCF defines four access category queues and eight traffic stream queues at MAC sublayer. When a frame arrives at MAC sublayer, it is tagged with a traffic priority identifier (TID) according to its QoS requirements. A frame with TID value between 0 and 7 is assigned to one of the four[2] access category queues. Similarly, a frame with a TID value of 8 to 15 is assigned to one of the eight traffic stream queues [26].

The authors, in their study, create a mapping mechanism between the traffic parameters of the IEEE 802.16 and IEEE 802.11e networks. Their approach is depicted in Fig. 10. The Radio Gateway simultaneously performs the functions of an IEEE 802.16 SS and an IEEE 802.11e access point, and the Mapping Module is responsible for choosing the most suitable class for traffic flowing between the two systems. For the QoS mapping, the authors propose two approaches. The first one, called "prioritized mapping", is similar to the Differentiated Services architecture [1]. In this kind of mapping, application flows coming from an IEEE 802.11e network are mapped to a corresponding traffic class in an IEEE 802.16 network and vice versa. In the second kind of mapping, called per-flow "parameterized mapping", which resembles the Integrated Services architecture [2], optional/mandatory traffic parameter requirements for a traffic stream are used to find the most suitable traffic class (C1 to C4, as shown in Table 4).

[2] TID value 1 and 2 are assigned to access category queue 0, 0 and 3 to queue 1, 4 and 5 to queue 2, and 6 and 7 are assigned to queue 3 [20, Table 20i].

Table 4
Parameterized mapping function performed by the Mapping Module, traffic classes, and their typical usage.

Traffic class	Typical usage	IEEE 802.11e	IEEE 802.16	Remarks
C1	Constant bit rate (CBR) with real-time traffic	Peak data rate Delay bound Data rate + delay bound	Maximum sustained traffic rate Maximum latency Tolerated jitter	Applications like real-time audio/video. The desirable characteristics for this class are very limited packet losses, minimum latency delays and very little jitter
C2	Variable bit rate (VBR) with real-time traffic	Maximum data rate Peak data rate Delay bound Burst size	Minimum reserved traffic rate Maximum sustained traffic rate Maximum latency Maximum traffic burst	Examples of traffic for this class include video on demand (streaming) and variable rate voice-over-IP. Packet loss, minimum latency delay and jitter limits apply to such traffic within more relaxed bounds as compared to Class C1
C3	VBR with precious data	Minimum data rate Peak data rate User priority Burst size	Minimum reserved traffic rate Maximum sustained traffic rate Traffic priority Maximum traffic burst	Can be used for traffic types like large data file transfers
C4	Unspecified type	Peak data rate User priority	Maximum sustained traffic rate Traffic priority	Caters for best-effort type traffic such as Web access, email communication, etc.

The authors do not elaborate the circumstances under which each mapping model should be used, pointing out that handling of these mappings is implementation dependent. It may be assumed that the type of traffic expected to be carried in such a hybrid network (e.g., predominantly elastic or a mix with a sizeable component of delay-sensitive traffic) would be the determining factor for the choice of mapping model.

In the paper, the authors also discuss the further improvements required for their approach. Unfortunately, there are no experimental results provided in the study to quantify the performance of the architecture proposed.

3.2.4. A QoS integration model for WLANs and WiMAX using media independent handover

Focusing on a heterogeneous network consisting of IEEE 802.11e WLAN and IEEE 802.16d WiMAX nodes, Roy et al. [30] propose a mechanism that supports Always Best-Connected (ABC) QoS integration. In this ABC QoS integration method, a client can seamlessly switch between WLAN and WiMAX networks and vice-versa without compromising QoS during the handover. The work also incorporates the IEEE 802.21 draft standard where, a "L2.5" layer is defined to execute the media independent handover (MIH) that occurs between multiple access networks. The standard also specifies MIH to attain service continuity with guaranteed QoS during handover. The paper proposes a mechanism where a drop in measured user QoS parameters in one network will trigger a MIH to switch to the other network. The architecture places a Generic Virtual Link Layer (GVLL) to reside over the MAC sublayer. The GVLL triggers the MIH based on various user QoS parameters such as throughput, packet loss and delay. The user is always connected to the network with the best QoS support.

Fig. 11 illustrates the proposed architecture: the user equipment is equipped with multiple interfaces to support both access networks. Layers including Mobile IP and above do not have the knowledge that there are multiple MAC sublayers with their corresponding MAC addresses. When a higher layer packet arrives destined for the MAC

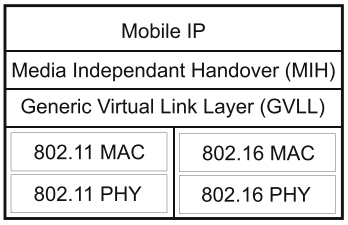


Fig. 11. User equipment protocol stack proposed in [30].

sublayer, the GVLL sends the packet to the appropriate MAC sublayer depending on the best available QoS. Packet loss and delay are the chosen QoS parameters to determine the superior network. The GVLL has three primary functions:

(1) It is the virtual MAC sublayer interface to the higher layers.
(2) Collects information from multiple MAC sublayers and triggers the MIH if the handover condition is achieved.
(3) Receives higher layer packets and forwards them to any particular MAC to which it is attached at that moment.

According to the proposed architecture, the GVLL triggers MIH under two circumstances:

Whenever a new connection (data or voice) has been admitted: When a new connection admission is requested, the GVLL simultaneously sends requests to both access networks. If the response is from only one network, then the user decides if the connection should be established. If the response is from both networks, then the GVLL decides between the available interfaces depending on the best QoS support. If the QoS support from both networks is the same, then the signal to noise ratio (SNR) is determined to select the interface. As the study does not implement SNR in the simulation, under such circumstances it defaults to WLAN network.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

Whenever the QoS guarantee falls beyond the threshold: QoS parameters are continuously monitored and if they fall below the threshold, a request to other available networks is sent out: the one with the best QoS support is selected and MIH is triggered to initiate the handover.

HCF MAC functionality is used to guarantee QoS support in the WLAN domain and TDMA based MAC has been used in the WiMAX domain. Admission control policy accepts a call if there are sufficient resources available to support the mean data rate of the call which is assumed to be of VBR type.

The simulation scenario in the study consists of two WLAN access points inside a single WiMAX coverage area with 18 user stations. Each network is connected to a backbone individually. Link capacities for each WLAN and WiMAX network are 12 Mbps and 24 Mbps, respectively.

In the first simulation scenario, handovers between WLAN and WiMAX are not supported. Users can only attach to their respective networks, and WLAN users can roam between the two WLAN access points but a WLAN call can not be handed over to the WiMAX BS and vice-versa. 9 users are directly connected to the WiMAX network and the remaining 9 are WLAN users. This is the single interface scenario where GVLL is not implemented.

In the second simulation scenario the GVLL scheme is adopted to support interoperability between the two technologies. During the call initiation the GVLL sends request to both a WLAN access point and WiMAX BS and chooses the best among them according to the responses it gets. Handovers between WLAN and WiMAX are also supported.

The results compare the performance of the two scenarios, and show noticeable improvement on GVLL implementation when the network is heavily loaded, an improvement of roughly 18%. As the number of calls admitted is more in case of GVLL, the system throughput reflects likewise. The results further show that as the network gets saturated, with the implementation of GVLL, the average delay improves by approximately 10%.

4. Mechanisms for QoS provision in WiMAX based mesh networks

In a mesh WiMAX network, a "mesh base station" (Mesh BS) undertakes the role of a BS and provides the connection to other communication networks. There can be multiple mesh BSs in a network and other nodes are known as mesh subscriber stations (Mesh SSs). In contrast with point-to-multipoint WiMAX networks, the standard does not define scheduling services for mesh WiMAX networks. In point-to-multipoint mode, the SSs are under the direct control of the BS. In Mesh mode the uplink and downlink is not clearly separated and SSs can communicate with each other without communicating with the BS. The transmission opportunities in the control subframe and the data minislots in the data subframe are separated. The nodes compete for the control channel access and the contention outcome does not effect the data transmission.

The QoS provision in mesh WiMAX networks is more challenging and very few researchers have thus far focused their efforts on this area. In the following sections we present a couple of representative studies which propose ways to improve QoS signaling mechanisms and create hybrid architectures for improving inter-connectivity with existing networks.

4.1. Routing and admission control for mesh WiMAX networks

In [36] Tsai and Wang propose a routing method using Shortest–Widest Efficient Bandwidth (SWEB) as a metric for distributed, coordinated WiMAX mesh mode along with a token bucket based admission control (TAC) algorithm. The study uses the token bucket mechanism as it works well for smoothing the burstiness of packet flows and helps in estimating the required bandwidth.

The SWEB metric considers three parameters:

Packet Error Rate can be retrieved by exchanging the MSH-DSCH messages. Each MSH-DSCH message is associated with a unique sequence number, there any lost or damaged messages can be detected.
Link Capacity can be determined by the burst profile indicated in the MSH-NCFG message.
Hop Count is included also in the MSH-NCFG messages from a station to the BS.

Based on these parameters, SWEB is retrieved and the path with the largest SWEB is chosen.

TAC has two essential components:

Bandwidth Estimation: It is estimated using the token bucket based admission control, and it is dependent on token rate and bucket size associated with a given connection and frame length.
Algorithm Determination: The estimated bandwidth is used to determine the admission control algorithm. To prevent starvation of lower-priority traffic, minimum usage of timeslots by each connection is defined. The algorithm is determined through the following procedure:

(1) When a new bandwidth request occurs, the source node computes its available bandwidth as the total empty slot number.
(2) The station that handles the request checks if requested bandwidth is less than available bandwidth. If yes, it goes to next step, otherwise goes to Step 4.
(3) By comparing the current and minimum usage of other traffic classes, the station determines if the flow should be downgraded.
(4) If the current usage exceeds the minimum usage of the traffic class, the station rejects the flow. Or else, it goes to next step.
(5) The station checks the timeslots used by downgraded flows in the order of BE, VBR or CBR. The request is rejected if there are no such timeslots.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

Else, it sets these timeslots empty, which means to preempt these timeslots. It then grants the timeslots and updates the value of available bandwidth.

The study reports the results of simulation based experiments conducted on a 16 node topology with various types of traffic (BE, VBR and CBR). In terms of the physical and data link layer parameters, QPSK modulation is assumed, the simulation area is 16 km^2, the radio range radius is set to 1.5 km, while frame length is chosen to be 8 ms. The data rate used for CBR traffic is 64 kbps with 960 bit packet size and a packet interval of 15 ms. VBR traffic data rate is 400 kbps with a mean packet size of 16,000 bits and a packet interval of 40 ms. Finally, the offered BE data rate is 1000 kbps, with a packet size 8000 bits and packet interval of 8 ms. The results based on SWEB are compared against previous studies that used Expected Transmission Count (ETX) and Shortest Path First algorithm. The graphs show that, the throughput for ETX is highest because it selects a route with the lowest packet error rate. However, as ETX does not take the hop count into account, it causes higher packet delays. SWEB had the best performance with respect to jitter, but from our observation, the improvement is significant only after the number of flows is more than 20.

Under normal circumstances, best-effort traffic experiences preemption from higher priority traffic classes. But, when TAC is used the best effort flows gain the advantage of having the guaranteed minimum throughput. However, the observations on the simulation results presented as a graph in the study reveal that to prevent the starvation of best-effort flows, variable bit rate traffic throughput is sacrificed. The results further show that, when TAC is not used, 12% of variable bit rate packets exceed the delay requirements when the number of flows is 25. This is reduced to 7% when TAC is used.

4.2. A QoS differentiation scheme for mesh WiMAX networks

Zhang et al. propose a scheme to achieve QoS differentiation in the WiMAX mesh mode [40]. In their work, the authors introduce the distributed scheduling concept, and also develop a new formula for its theoretical evaluation in random topologies.

In distributed scheduling, a node can transmit in any slot during the eligibility interval, and has to contend with other nodes. This contention is irrespective of the service type and its priority. To overcome this drawback, the study proposes a scheme to prioritize traffic and enable the QoS differentiation by varying the eligibility intervals for different traffic classes.

The formula for evaluating the scheme is derived for two different topologies: co-located scenario (all nodes are one-hop neighbors of each other), and general topology (multihop neighborhood). The numerical results show the effectiveness of achieving differentiated QoS in both of these topologies: with all nodes equally partitioned into three priority classes (1, 2 and 3), the proposed scheme is able to ensure that class 1 has the shortest and 3 the longest delay.

5. QoS Issues in evolutions of the WiMAX standard

Several important evolutions of the WiMAX standard are currently in progress, and for each of these we now examine the issues relating to QoS support.

5.1. WiMAX's road to 4G

With the goal of improving performance of the current release of Mobile WiMAX [19], two separate evolution efforts have been under way since the beginning of 2007.

5.1.1. WiMAX Forum: Release 1.5

The WiMAX Forum, with its Release 1.5 evolution project, is aiming for a short time horizon (targeting systems deployed in 2009/10 timeframe) by trying to minimize the changes to the current IEEE Network Release 1.0 specification, which supports the IEEE 802.16e-2005 standard [39].

With regards to QoS support, Release 1.0 only offers basic functionality, in the form of static (i.e., pre-provisioned) QoS and an optional rather than mandatory radio resource manager. Static QoS implies that the SS may not modify parameters of the service flows already provisioned by the system, nor create any service flows dynamically. This issue is addressed in Release 1.5 through the incorporation of dynamic QoS functionality, whereby an SS may dynamically set up a flow through DSA transactions as discussed in Section 2.2 [10].

Another QoS enhancement in Network Release 1.5 is the policy and charging (PCC) functionality, planned to be fully compatible with the 3GPP Release 7 specification [29]. Policies are rules which are triggered by certain types of traffic or user behavior in the network. Combining such rules with the ability to dynamically assign QoS to user flows, PCC becomes a powerful enabler of differentiated QoS features such as (i) QoS based on accumulated usage, and (ii) QoS based on aggregate network load. In (i) a particular user or application could be dynamically assigned an inferior QoS class (e.g., lower-priority or a small traffic-shaped data rate) after reaching a volume threshold in bytes. Similar dynamic de-prioritization of a targeted user or application could be undertaken in (ii), with the policy trigger in this case being aggregate network load (e.g., protecting higher priority users when network utilization exceeds 80%).

The final enhancement in Network Release 1.5, which may be considered a *direct* enabler in the provision of differentiated QoS, is the inclusion of telephony VoIP. On the air interface, this is supported in the 802.16REV2 revision of the standard by a VoIP specific optimization called "persistent scheduling" [11]. More broadly speaking, all of the various fixes and minor amendments necessary to support the Network Release 1.5 specification are incorporated in the 802.16REV2 revision of the mobile WiMAX standard, which "combines the IEEE 802.16-2004 base standard plus IEEE 802.16e/f/g amendments and related corrigenda" [10]. Compared to use of the default MAC protocol, the authors of [11] report an increase in WiMAX VoIP user capacity of approximately 16% under this MAC sublayer persistent allocation modification, due to the significant reduction of signalling message overheads.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

In summary, while Network Release 1.0 was primarily aimed at carriage of BE data traffic or static QoS-enabled flows, the Network Release 1.5 and associated 802.16REV2 air interface enhancements described above, directly enable the *dynamic* provision of application- and user-based QoS differentiation, while maintaining efficient network utilization.

5.1.2. IEEE 802.16 Working Group: 802.16m

The IEEE 802.16 Working Group, with its IEEE 802.16m project and proposed future standard, has its sights set on a more ambitious longer-term (2011/12) goal: to fundamentally enhance the performance of mobile WiMAX so that it meets the requirements of the ITU's international 4G standard, known as IMT-Advanced (the successor of the IMT-2000 3G standard) [23]. As such, most of the focus in IEEE 802.16m is on deriving raw physical layer performance improvements [10], which would only indirectly impact QoS by improving the performance of all QoS classes. Such performance improvements would be achieved using techniques such as increased spectral efficiency through more advanced and higher-order Multiple Input Multiple Output (MIMO) antenna systems, lower framing overheads at the physical and data link layers, and wider band carriers (e.g., 20 MHz).

From the set of IEEE 802.16m enhancements which will deliver better performance for all users, we single out two in particular [10,39], because they may be viewed as "direct enablers" for improved QoS differentiation in WiMAX. The first of these is lower latency which will be achieved in IEEE 802.16m by a leaner, faster MAC and signalling framework. The expectation is that this will lead to less complex and lower-latency QoS signalling and hence a better ability to provide differentiated QoS. The second of these enablers is the planned provision of seamless low-latency handovers between WiMAX and other radio access technologies such as Wi-Fi, allowing true "multimedia session continuity" – an important aspect of providing ent-to-end QoS for multimedia services.

5.2. WiMAX Multihop Relays: IEEE 802.16j

The goal of the emerging IEEE 802.16j WiMAX Multihop Relay (MR) standard [21] is to increase radio coverage, user throughput and capacity of traditional 802.16e-2005 WiMAX networks. The proposed 802.16j standard aims to achieve this goal by specifying PHY and MAC sublayer enhancements for licensed bands of spectrum that enable the operation of relay stations (RS). Note that the SS specifications are not changed.

The two main difficulties which are found to be common to each of the impacted QoS features of the proposed 802.16j standard (discussed below), are (i) the increased complexity of the whole-of-path signalling, as opposed to a single message exchange in single-hop networks, and (ii) the increased latency associated with relaying information (both signalling and user data) across multiple hops.

A challenge for designers of future MR WiMAX networks will be to compute near-optimal tradeoffs between the increased complexity and latency associated with relaying information across multiple hops, and the benefits afforded by increased radio coverage without the cost of rolling out full base stations.

What follows is a summary of the impacts of multihop relaying on the key WiMAX QoS features, as described in the latest baseline document of the proposed IEEE 802.16j standard.

5.2.1. Impact on scheduling services

Unlike in single-hop networks where bandwidth is granted by a BS directly to its attached SSs, in an MR system this allocation is cascaded down in hierarchical fashion. In the case of the UGS scheduling service, this means that to meet a UGS service flow's need, the Multihop Relay base station (MR-BS) and RSs along the path have to grant fixed size bandwidth to their subordinate nodes on a real-time periodic basis.

Similarly, unlike single-hop networks where a BS directly polls its attached SSs, an MR system requires that the polling must be cascaded down in hierarchical fashion. In the case of the rtPS and ertPS scheduling services, this means that in order to meet an rtPS/ertPS service flow's need, the MR-BS and RSs along the path must poll their subordinate nodes on a real-time periodic basis.

5.2.2. Impact on bandwidth allocation and request mechanisms

Another distinguishing feature of MR WiMAX networks is that an RS may combine (i) bandwidth requests arriving from its subordinate neighbor RSs during a given period of time, and (ii) bandwidth needs of packets in its local queue, into one "aggregated" bandwidth request header per QoS class. In order to minimize the additional delays introduced by this relay-based procedure, the RS is allowed to transmit a bandwidth request header shortly after it receives a bandwidth request header from one of its downstream stations, instead of waiting for the actual packets to arrive. The timing is chosen to yield an uplink allocation at the RS, which immediately follows the arrival of the relayed packets from the downstream station.

5.2.3. Impact on dynamic QoS procedures

In an MR WiMAX network with distributed scheduling, a BS cannot immediately admit a service flow and send a DSA-RSP message to the requesting SS, as in the case of traditional single-hop WiMAX networks. Instead, the procedure becomes considerably more complex due to the need for the BS to discover if all of the RSs in the path to the SS have sufficient resources to support the dynamically requested QoS. The discovery procedure begins with the BS sending a DSA-REQ message to its subordinate RS. This RS then sends its own DSA-REQ message to its subordinated neighboring RSs, with this hierarchical cascade continuing down until the access RS is reached.

6. Analysis and concluding remarks

The studies discussed in this paper examine various aspects of QoS architecture and QoS differentiation for two

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

key types of WiMAX networks: point-to-multipoint and mesh. The paper by Cicconetti et al. [7] provides an implementation of a QoS mechanism with basic traffic management. Significant improvement with regards to traffic management and admission control is proposed by Wongthavarawat and Ganz [38], with a focus on uplink packet scheduling and traffic policing at the SS. Although the simulation results only take rtPS and BE traffic into consideration, the research provides adequate information for dealing with other classes of traffic. However, one aspect of the admission control implementation in [38] which has room for improvement is maintaining fairness between all classes of traffic. The current implementation fails to prevent instances where one service class can dominate the entire link bandwidth.

A successful implementation of a WiMAX-customized WFQ2+ algorithm is reported by Shang and Cheng [32]. Their approach of implementing "hard" or "soft" QoS can be integrated with [38] for further optimization. There is ample scope for further research into an optimal scheduling algorithm from the many available candidates.

Using the fragmentation and aggregation capabilities of MAC SDUs in multiple PDUs, Sengupta et al. [33] provide a very good solution for maintaining differentiated QoS for streaming media. Their approach of rearranging MAC SDUs before transmission, along with a feedback mechanism, provided significant improvement in performance.

The two-tier scheduling algorithm (2TSA) proposed by Chan et al. [4] improves network performance significantly compared to earlier approaches which used strict-priority scheduling (such as [37]). In 2TSA, the first-tier allocation algorithm is category based and the second-tier allocation is weight based. When compared with the [37] algorithm, the simulation results show that 2TSA can guarantee connections' bandwidth demands, avoid starvation of lower-priority service class, and achieve a better degree of fairness. Other QoS metrics such as delay or delay jitter are left for future research.

The Preemptive Direct Fair Priority Queue (PDFPQ) scheduling method implemented by Safa et al. [31] improves minimum and average delay for rtPS traffic, as compared to a previous proposal [6] that used the non-Preemptive version (Direct Fair Priority Queue, DFPQ). However, one significant drawback is the drop in throughput of BE traffic. Although BE traffic does not face starvation, PDFPQ will cause a slower BE traffic response than in DPFQ. The study therefore leaves room for future work on methods which simultaneously seek to minimize the throughput degradation of BE traffic, while still improving delay for rtPS traffic.

Chen et al. [5] presented a technique embedding DSA, DSC and DSD messages inside the BW-Request message, which showed a significant improvement in connection setup time. However, their approach can compromise other potential capabilities of the network. If a network is to provide multiple services, like VoIP, video and data, it is important for admission control to know the service request from each SS before it receives the BW-Request. To provide multiple services, the admission control needs to consider fairness for all classes of traffic. Therefore, if a SS has multiple service requests, it should be able to partially accept some of the requests (to maintain fairness). As BW-Request messages only deal with aggregates, it will not be possible for the WiMAX admission control to partially accept some of the requests. This would prevent the system from providing differentiated admission control, running contrary to our stated goal of QoS differentiation in a multi-service wireless network. Conversely, in a network that caters for only one class of traffic, such an embedded signaling approach would work without any problems.

A successful internetworking solution between SONET and WiMAX is provided by Lin et al. [27]. They overcome the problem of bandwidth over- or under-utilization (due to mismatch of a WiMAX BS and an STS-1 backhaul link), by implementing a heuristic approach. The heuristic is based on maximizing utilization and efficiency, dependent on the measured network saturation level.

A WiMAX and QoS-enabled Wi-Fi (IEEE802.11e) internetworking solution is illustrated in a paper by Gakhar et al. [13]. The QoS management facility provided by 802.11e is successfully exploited by implementing a Mapping Module. Although the paper does not provide any simulation results to verify the possible outcome, in theory the solution sounds plausible. The implementation is not available for other popular Wi-Fi variants, which do not support QoS at the MAC sublayer (i.e., 802.11a/b/g).

QoS integration model for WLAN and WiMAX of Roy et al. [30] is another promising WiMAX internetworking study that provides scope for further developments. The study shows how Generic Virtual Link Layer (GVLL) can be used for interoperability between multiple standards. Factors such as high speed mobility and coverage present opportunities for future work. The study can also be expanded by including other wireless networks such as HSDPA, EDGE and EV-DO.

Zhang et al. [40] make a significant contribution with their QoS Differentiation Scheme for WiMAX mesh mode. The probabilistic methodology evaluating the scheduling performance in a general topology is a novel idea. The numerical results illustrate performance improvements in both the collocated and general topologies.

In closing, in this survey paper, we illustrated the general framework as well as many specific approaches for implementing QoS differentiation in the MAC sublayer of a WiMAX network. A brief explanation of the WiMAX MAC architecture was given before a number of research studies were explored. Each of these studies was placed into one of three categories. The "Packet scheduling and admission control" category looked into the way QoS implementation improves user service quality and network efficiency. The "Signaling and integration" category focused on how WiMAX networks can be deployed alongside other networks to meet various requirements. The third category "QoS in WiMAX mesh networks" focused on research into the distributed methods of signalling and scheduling required to achieve QoS differentiation in the mesh variant of WiMAX networks. We also examined the issues associated with provision of differentiated QoS services in future evolution of the WiMAX standard. Finally, we compared and contrasted the various studies, analyzing the potential and limitations of each, including options for future work in this important area of networking research.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001

Acknowledgement

Many thanks to Hyoung-Kyu Lim and Jungshin Park of Samsung for their valuable comments on improving the contents of the paper.

References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, RFC 2475 an architecture for differentiated services, 1998, URL reference <http://www.ietf.org/rfc/rfc2475.txt>.
[2] R. Braden, D. Clark, S. Shenker, Integrated services in the internet architecture: an overview, 1994, URL reference <http://www.ietf.org/rfc/rfc1633.txt>.
[3] J.C.R. Bennett, H. Zhang, Hierarchical packet fair queueing algorithms, IEEE/ACM Transactions on Networking 5 (5) (1997) 675–689.
[4] L. Chan, H. Chao, Z. Chou, Two-tier scheduling algorithm for uplink transmissions in IEEE 802.16 broadband wireless access systems, in: Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM'06), September 2006, pp. 1–4.
[5] J. Chen, W. Jiao, Q. Guo, An integrated QoS control architecture for IEEE 802.16 broadband wireless access systems, in: Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM'05), St. Louis, USA, IEEE Communications Society, November 2005.
[6] J. Chen, W. Jiao, H. Wang, A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode, in: Proceedings of the 2005 IEEE International Conference on Communications (ICC'05), Seoul, Korea, IEEE Communications Society, May 2005, pp. 3422–3426.
[7] C. Cicconetti, L. Lenzini, E. Mingozzi, C. Eklund, Quality of service support in IEEE 802.16 networks, IEEE Network 20 (2006) 50–55.
[8] A. Demers, S. Keshav, S. Shenker, Analysis and simulation of a fair queueing algorithm, in: Proceedings of the Communications Architectures and Protocols Symposium, ACM, September 1989, pp. 1–12.
[9] H. Dewing, S. Potter, Implementing QoS solutions in enterprise networks, February 2002, URL reference <http://www.tmcnet.com/it/0202/0202inim.htm>.
[10] K. Etemad, Overview of WiMAX technology and evolution, IEEE Communications Magazine 46 (10) (2008) 31–36.
[11] M. Fong, R. Novak, S. McBeath, R. Srinivasan, Improved VoIP capacity in mobile WiMAX systems using persistent resource allocation, IEEE Communications Magazine 46 (10) (2008) 50–56.
[12] WiMAX Forum, Business case models for fixed broadband wireless access based on WiMAX technology and the 802.16 standard, October 2004, URL reference <http://www.wimaxforum.org/technology/downloads/WiMAX-The_Business_Case-Rev3.pdf>.
[13] K. Gakhar, A. Gravey, A. Leroy, IROISE: a new QoS architecture for IEEE 802.16 and IEEE 802.11e interworking, in: Proceedings of the Second International Conference on Broadband Networks (Broadnets'05), Boston, USA, October 2005, pp. 607–612.
[14] L. Georgiadis, R. Guerin, A. Parekh, Optimal multiplexing on a single link: delay and buffer requirements, IEEE Transactions on Information Theory 43 (5) (1997) 1518–1535.
[15] E.L. Hahne, R.G. Gallager, Round Robin scheduling for fair flow control in data communication networks, in: Proceedings of the IEEE International Conference on Communications (ICC'86), Toronto, Canada, IEEE Communications Society, March 1986, pp. 103–107.
[16] M. Hawa, D.W. Petr, Quality of service scheduling in cable and broadband wireless access systems, in: Proceedings of the 10th IEEE International Workshop on Quality of Service, IEEE, May 2002, pp. 247–255.
[17] IEEE, IEEE standard for local and metropolitan area networks Part 16: Air interface for fixed broadband wireless access systems, 2004, URL reference <http://standards.ieee.org/getieee802/download/802.16-2004.pdf>.
[18] IEEE, Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Part 11, Amendment 7: medium access control (MAC) quality of service (QoS) enhancements, 2004.
[19] IEEE, IEEE standard for local and metropolitan area networks Part 16: Air interface for fixed and mobile broadband wireless access systems (amendment and corrigendum to IEEE Std 802.16-2004), 2005, URL reference <http://standards.ieee.org/getieee802/download/802.16e-2005.pdf>.
[20] IEEE, IEEE standard: information technology – telecommunication and information exchange between systems – local and metropolitan area networks – specific requirements – Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications – amendment 8: medium access control (MAC) quality of service enhancements, 2005, URL reference <http://standards.ieee.org/getieee802/download/802.11e-2005.pdf>.
[21] IEEE, Baseline document for draft standard for local and metropolitan area networks, Part 16: Air interface for fixed and mobile broadband wireless access systems (Multihop Relay specification), 2007, URL reference <http://www.ieee802.org/16/relay/docs/80216j-06_026r4.zip>.
[22] European Telecommunications Standards Institute, General aspects of quality of service and network performance in digital networks, including ISDN, Technical report ETR 003 ed.1, ETSI, 1990.
[23] ITU, ITU-R recommendation M.1645, framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000, 2003, URL reference <http://www.itu.int/rec/R-REC-M.1645/e>.
[24] A. Kumar, D. Manjunath, J. Kuri, Wireless Networking, Morgan Kaufmann (2008).
[25] M. Katevenis, S. Sidiropoulos, C. Courcoubetis, Weighted round-Robin cell multiplexing in a general-purpose ATM switch chip, IEEE Journal on Selected Areas in Communications 9 (8) (1991) 1265–1279.
[26] H. Labiod, H. Afifi, C. De Santis, Wi-Fi, Bluetooth, Zigbee and WiMAX, Springer, 2007.
[27] P. Lin, C. Qiao, T. Wang, J. Hu, Optimal utility-based bandwidth allocation over integrated optical and WiMAX networks, in: Proceedings of the Optical Fiber Communication Conference and the 2006 National Fiber Optic Engineers Conference, March 2006.
[28] Third Generation Partnership Project, 3GPP TS 25.308 high speed downlink packet access (HSDPA), overall description stage 2, URL reference <http://www.3gpp.org/ftp/specs/html-info/25308.htm>.
[29] Third Generation Partnership Project, Technical specification group services and system aspects; policy and charging control architecture (release 7) 3GPP TS 23.203 V7.5.0 (2007-12), URL reference <http://www.3gpp.org/FTP/Specs/html-info/23203.htm>.
[30] R.J. Roy, V. Vaidehi, S. Srikanth, Always best-connected QoS integration model for the WLAN, WiMAX heterogeneous network, in: Proceedings of the First International Conference on Industrial and Information Systems, August 2006, pp. 361–366.
[31] H. Safa, H. Artail, M. Karam, R. Soudah, S. Khayat, New scheduling architecture for IEEE 802.16 wireless metropolitan area network, in: Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'07), May 2007, pp. 203–210.
[32] Y. Shang, S. Cheng. An enhanced packet scheduling algorithm for QoS support in IEEE 802.16 wireless network, in: Third International Conference on Networking and Mobile Computing (ICCNMC'05), Zhangjiajie, China, August 2005, pp. 652–661.
[33] S. Sengupta, M. Chatterjee, S. Ganguly, R. Izmailov, Exploiting MAC flexibility in WiMAX for media streaming, in: Proceedings of the Sixth IEEE International Symposium World of Wireless Mobile and Multimedia Networks (WoWMoM 2005), Taormina, Italy, IEEE Computer Society, June 2005, pp. 338–343.
[34] M. Shreedhar, G. Varghese, Efficient fair queuing using deficit round Robin, IEEE Transactions on Networking 4 (3) (1996) 375–685.
[35] D. Stiliadis, A. Varma, Latency-rate servers: a general model for analysis of traffic scheduling algorithms, IEEE/ACM Transactions on Networking 6 (5) (1998) 611–624.
[36] T.C. Tsai, C.Y. Wang, Routing and admission control in IEEE 802.16 distributed mesh networks, in: IFIP International Conference on Wireless and Optical Communications Networks (WOCN'07), Singapore, 2007, pp. 1–5.
[37] K. Wongthavarawat, A. Ganz, IEEE 802.16 based last mile broadband wireless military networks with quality of service support, in: Proceedings of the IEEE Military Communications Conference, vol. 2, October 2003, pp. 779–784.
[38] K. Wongthavarawat, A. Ganz, Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems, International Journal of Communication Systems 16 (1) (2003) 81–96.
[39] F. Wang, A. Ghosh, C. Sankaran, P. Fleming, F. Hsieh, S. Benes, Mobile WiMAX systems: performance and evolution, IEEE Communications Magazine 46 (10) (2008) 41–47.
[40] Y. Zhang, J. Zheng, W. Li, A simple and effective QoS differentiation scheme in IEEE 802.16 WiMAX mesh networking, in: Proceedings of

the IEEE Wireless Communications & Networking Conference (WCNC'07), Hong Kong, China, March 2007.

Ahmet Sekercioglu is a researcher at the Centre for Telecommunications and Information Engineering (CTIE) and a Senior Lecturer at the Department of Electrical and Computer Systems Engineering of Monash University. He was the leader of the Applications Program of Australian Telecommunications CRC until the end of the centre's research activities (December 2007). He has completed his Ph.D. degree at Swinburne University of Technology, and B.Sc., M.Sc. (all in Electrical and Electronics Engineering) degrees at Middle East Technical University. He has lectured at Swinburne University of Technology for 8 years, and has had numerous positions as a research engineer in private industry.

His more recent work focuses on distributed algorithms for self-organization in wireless networks. He is also interested in application of intelligent control techniques for multi-service networks as complex, distributed systems.

Milosh Ivanovich fills the role of Senior Emerging Technology Specialist within the Chief Technology Office of Telstra, and is an Honorary Research Fellow at Melbourne and Monash Universities in Australia. A Senior Member of IEEE, Milosh's interests lie in queuing theory, teletraffic modeling, performance analysis of wireless networks, and the study and enhancement of TCP/IP in hybrid fixed/wireless environments. He obtained a B.E. (1st class Hons) in Electrical and Computer Systems Engineering (1995), a Master of Computing (1996) and a Ph.D. in Information Technology (1998), all at Monash University Australia. He is an author of two edited book chapters, a patent, and over 40 international journal and conference publications.

Alper Yegin is an architect at the Standards and Industry Initiatives Group of Samsung Electronics. He currently chairs IETF PANA Working Group and Security Team of WiMAX Forum Network Working Group. In the past he served as members of IETF Wireless Directorate and IPv6 Forum Technical Directorate. He has received his M.Sc. in Computer Science degree at University of Illinois, Urbana-Champaign, and B.Sc. in Computer Engineering at Bogazici University. His recent work focuses on IP-based end-to-end 4G architectures, especially in the network security and mobility management areas.

Please cite this article in press as: Y. A. Şekercioğlu et al., A survey of MAC based QoS implementations for WiMAX networks, Comput. Netw. (2009), doi:10.1016/j.comnet.2009.05.001