QoS Management in a Multilayer WiMAX System

Marie-Paule Gakuba, Mjumo Mzyece, Anish Kurien
French South African Institute of Technology (F'SATI)
Tshwane University of Technology (TUT)
Private Bag X680, Pretoria 0001, South Africa
Tel: +27 12 382 4191
Fax: +27 12 382 5294
Email: mgakuba3@gmail.com, mzyecem@tut.ac.za, kurienam@tut.ac.za

Abstract— In 2008, telecommunication market trends in Africa reported 0.2% penetration for broadband wireless access (BWA) technologies. IEEE 802.16 (WiMAX) is recognized as a promising technology to provide BWA for heterogeneous traffic. Among the various features defined in the standard, some features such as the scheduling algorithm is undefined. A new MDRR-based (Modified Deficit Round Robin) scheduling algorithm is proposed in this paper to maintain QoS requirements among heterogeneous traffic. The performance of the enhanced MDRR-based algorithm is compared to the existing MDRR algorithm in terms of throughput in an evenly-distributed and mixed traffic scenario. Maximum throughput of 68 kbps and 112 kbps were attained by the 20% and 40 % latency MDRR-based algorithm respectively while only 56 kbps was achieved by the MDRR algorithm.

Index Terms— IEEE 802.16, Heterogeneous traffic, QoS, Scheduling algorithm.

I. Introduction

After the installation and expansion of wired systems in many areas worldwide, the need for wireless access has rapidly grown in an attempt to bridge the digital divide as well as provide access in unconnected areas. However, wireless networks are challenged by bandwidth scarcity and limitations when compared to wired networks. Thus, the provisioning of QoS remains an issue for the transmission of multimedia services to support high data rates [1]. Statistics recorded Internet penetration growth in Africa of only 4.8 % which stood as the lowest penetration in the world in 2008. High data rate provisioning known as Broadband Wireless has seen a growth in demand to improve Internet usage. WiMAX Broadband Wireless Access, a technology based on the IEEE 802.16 standard is considered in this paper for broadband access. WiMAX is deployed in Wide Metropolitan Area Networks (WMAN) to offer sufficient throughput of up to 70 Mbps and a maximum range of 50 km. Its support of fixed wireless Internet Access, mobile networks as well as backhaul applications provides transmission of multimedia services at high data rates. With its set of rich features, the WiMAX standard is known as one of the most promising telecommunication systems. The main challenge is to allocate QoS requirements of each application in WiMAX networks. The QoS management and realization is performed in different layers. WiMAX defines two main variants in the IEEE 802.16 standard namely the fixed and the mobile variants, both mainly differed by the mobility support in the mobile variants [2].

In wireless systems, there is scarcity and limitation of radio resources; as a result, QoS support for different applications remains crucial. The need for a suitable scheduling algorithm to maintain QoS among applications is required. Scheduling is required in the downlink (DL) as well as in the uplink (UL) channels. In the DL channel, the communication is established from the base station (BS) to the subscriber stations (SSs) whereas in the UL channel, the SSs utilize the channel to transmit packets to the BS. Since the BS does not synchronize with other stations, the downlink channel faces little congestion. The scheduling algorithm in this paper is implemented in the uplink channel for efficient radio resource management and to reduce congestion caused by the multitude of SSs.

The rest of the paper is structured as follows. Section II provides the protocol layers defined in IEEE 802.16 standard. QoS management is reviewed in Section III. Section IV describes the scheduling algorithms. In Section V, an enhanced MDRR-based scheduling algorithm is implemented. The simulation model and results are presented in Section VI followed by a conclusion and recommendation for future work in Section VII.

II. Protocol Layers

In comparison with the OSI model, the IEEE 802.16 standard specifies the Physical layer or Layer 1 of the OSI model and the MAC layer which is included in the Data Link Layer (Layer 2) of the OSI model. The Physical (PHY) layer is responsible for physical connections of the entities while the Medium Access Control (MAC) layer establishes and maintains connections. The MAC layer supports a Point to Multipoint (PMP) topology with an optional mesh

topology and five PHY layers variants as specified in the PHY layer sub-section.

A. Physical Layer

Five variants in the air interface are defined in the IEEE 802.16 standard at the PHY layer. A range of 2-66 GHz band is divided into NLOS and LOS transmissions. The NLOS utilizes 2-11 GHz while the LOS makes use of 11-66 GHz. These air interfaces include WirelessMAN-SC (10-66 GHz), WirelessMAN-SCa (below 11 GHz), WirelessMAN-OFDM (below 11 GHz), WirelessMAN-OFDMA (below 11 GHz) and WirelessHUMAN (below 11 GHz). TDD and FDD duplexing modes are applied to the first four variants whereas the last variant uses only TDD [2, 3].

B. MAC Layer

The MAC layer on the other hand is made of three sublayers that include the CS (Convergence Sublayer), the CPS (Common Part Sublayer) and the Security sublayer. The communication between layers is defined as follows. When Layer X addresses an XPDU (Layer X Protocol Data Unit) to Layer Y, this XPDU is received as an (X-1) SDU (Layer X-1 Service Data Unit) which is the lower layer [2]. The CS is above the CPS Sublayer and makes use of the CPS services through Service Access Point (SAP) in the form of a SDU. The CS classifies and maps the SDU into corresponding CIDs (Connection Identifier) as well as respective QoS; and if required, it suppresses the payload header by means of PHS (Payload header Suppression) and delivers CS PDU to the lower sublayer. By means of a classifier, a connection with a CID is assigned to a scheduling service type. The CPS is located in the middle of the MAC layer and is considered as the core because of its responsibilities such as bandwidth allocation, connection establishment and maintenance of connections between two entities. The CPS includes among other functions, bandwidth demands and allocation, scheduling, QoS management. After classifying the various connections to the scheduling classes, the base station maps the results to their destination addresses. Subsequently, the BS verifies the available resource before granting an amount of bandwidth to the connection which is then broadcasted through the UL-MAP. The UL-MAP message contains the QoS specifications of the specific connection and the transmission schedule. The Security Sublayer provides data encryption and authentication in the system through different keys such as Digital Encryption System (DES) and Advanced Encryption System (AES) [1, 2].

III. QoS MANAGEMENT

This section describes different scheduling classes defined in the IEEE 802.16 standard and the request/grant mechanism for QoS provisioning. It is to be noted that the IEEE 802.16 standard defines an 802.16 MAC that is connection-oriented. This means that the necessary QoS parameters are predetermined before the connections are established. The QoS parameters include the scheduling service class, Minimum Reserved traffic rate (MRTR), Maximum sustained traffic rate (MSTR) and scheduling algorithm. The Minimum Reserved Traffic Rate (MRTR guarantees QoS by defining the minimum data rate for a service flow. The Maximum Sustained Traffic Rate (MSTR) is similar to MRTR but guarantees QoS by defining a maximum peak data rate that a service cannot exceed [3].

It is important to note that the QoS mechanism operates under the bandwidth request mode and the bandwidth grant mode. Either the SS or the BS initiates the connection to establish any sort of packets transmission in the network.

A. QoS scheduling Class

There are five types of QoS classes defined in the IEEE 802.16 standard and are described as follows [4]-[6].

1. Unsolicited Grant Service (UGS): the UGS class is designed to support real-time application such as VoIP with silence suppression. A grant of fixed size uplink is granted by the base station periodically. However, bandwidth allocation is processed in its own dedicated channel. Since no overhead are included in the request by a subscriber station, no bandwidth request is needed to be sent. Thus, piggyback and contention request are permitted by the UGS class. Some of its QoS specifications include MSTR, jitter and latency.
2. Extended real-time polling service (ertPS): the ertPS is mainly defined for IEEE 802.16e version where handovers are involved. Similar to the UGS, the ertPS present the same BS performance in the uplink, but only differs in the dynamic allocation. The QoS specifications defined for ertPS include MSTR, MRTR and jitter.
3. Real-time Polling service (rtPS): Implemented for real-time application such as VoIP without silence suppression or MPEG video, rtPS supports variable packets size issued periodically. Thus, a desired grant size can be indicated by the subscriber station. Piggyback and contention request are allowed in rtPS class although excessive overheads are included. The QoS specifications are MSTR, MRTR and traffic priority.
4. Non real-time Polling Service (nrtPS): The nrtPS class is implemented to support non real-time (delay-tolerant) applications with variable size issued periodically; File Transfer Protocol (FTP) is a typical case for nrtPS. Similar to rtPS, piggyback request and contention are permitted. The QoS specifications are the MSTR, the MRTR and the traffic priority.
5. Best Effort (BE): The BE class is designed to support packets streams with no minimum QoS guaranteed; this would be the case for email and web browsing applications. The BE class also makes use of contention or piggyback requests. The QoS parameter could be defined by MSTR for better treatment but generally no specific requirements are defined for this scheduling class.

B. Bandwidth Request Mechanism

In the IEEE 802.16 standard, the following types of bandwidth requests are defined [2],[4].

The Bandwidth Request Message, this type of bandwidth request defines the incremental mode and the aggregate mode described. The *Incremental Bandwidth Request* (IBR) is sent by the subscriber when the amount of bandwidth acquired is not enough for transmission. There is no specific amount of bandwidth indicated. The IBR introduces unfairness when multiple subscribers are to be satisfied regardless of the existing amount of bandwidth. Unlike the IBR, the subscriber specifies the amount of bandwidth through the *Aggregate Bandwidth Request* (ABR) mode. Thus, ABR is considered as a fair and accurate mode of request in the sharing of bandwidth.

The Implicit Request, commonly designed for UGS scheduling class, is utilized for flows that do not demand for a request message at the connection setup as the bandwidth is automatically assigned by the base station. Thus, a bandwidth request message is not required. It is at this point that the implicit request is sent when a better treatment is required.

The *piggyback request* is designed for other scheduling classes other than UGS. It consists of a piggyback grant-request sub-header and an extended piggyback request.

C. Bandwidth Grant Mechanism

In Figure 1, a request/grant mechanism is illustrated. After being admitted into the network and for security purposes, a service flow is assigned an authentication number before the base station establishes a connection. The granting mechanism is processed by the base station to various subscriber stations in the network.

A grant is defined as the opportunity of transmitting packets within a specific duration of time. Through the grant mechanism, an amount of bandwidth is allotted to subscribers. The base station however ensures that the bandwidth grant would not affect the connection established before the grant process by maintaining the same quality of service after bandwidth grant. The IEEE

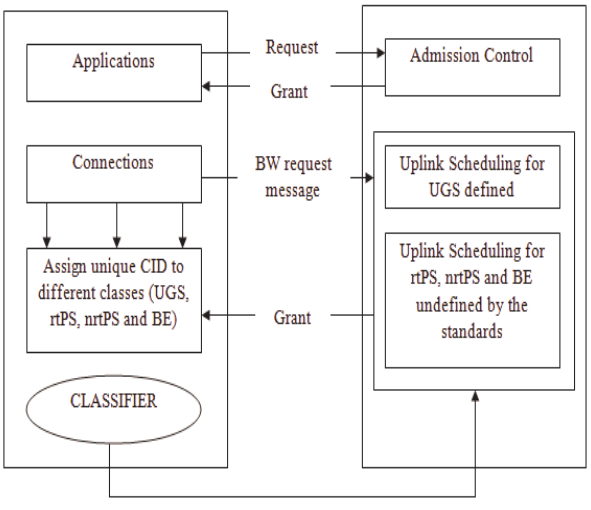

Figure. 1: Bandwidth request/grant mechanism [2]

802.16 standard defines two types of bandwidth grant: Grant per Connection (GpC) used for an amount of bandwidth granted for a specific requested connection and Grant per Subscriber Station (GpSS) is provided to a large number of subscribers which needs to be shared among multiple connections.

IV. Scheduling Algorithms

This section is based on related work of scheduling algorithms implemented by a number of researchers.

The First-in-First-out (FIFO) scheduling algorithm is known for its simplicity of serving packets with no priority considered. The FIFO algorithm [7] does not consider the QoS requirements of all packets in the network which makes it inappropriate for real-time applications. Strictly designed for delay-sensitive packets, the Earliest Deadline First (EDF) prioritizes only packets with deadline regardless of other packets in the network. It thus creates some level of unfairness for delay-insensitive packets especially when a continuous set of delay-sensitive packets are to be served.

Implemented to reduce unfairness in some of the scheduling algorithms, Round Robin (RR) scheduling [8] serves packets in a round robin way (a queue after another) until the complete round is made regardless of the QoS requirements of the applications. RR however is not suitable for variable packets size. An enhancement of the RR algorithm known as Weighted Round Robin (WRR) [8] was developed based on a weight basis to generate some level of priority amongst packets. The heavier the packets weight, the higher the priority to get served at first. In cases when a stream of continuous packets with heavier weight is involved, packets with lower weight suffer from bandwidth allocation.

Based on RR, the Deficit Round Robin (DRR) [7, 9] attempts to produce the QoS as expected by the user by generating a deficit counter and a quantum value to de-queue its packets. Yet, in a round robin manner, a queue de-queues its packets when the packet size located at the head of the queue is less than the deficit counter; otherwise, the former is retained in the queue and the later is incremented by the quantum. After serving a queue, the deficit counter is decreased by the number of bits served in the specific queue. In other words, the deficit counter keeps track of the served bits in a circular fashion whereas the quantum defines the number of bits acquired by a queue. On analysis of the DRR performance, it is seen that the algorithm tries to satisfy all packets in a round robin way. However, if many queues that contain delay-sensitive packets are to get served, these would suffer from delay as queues have to wait for the complete round to transmit another bit.

A well established scheduling algorithm defined as Modified Deficit Round Robin (MDRR) scheduling

algorithm was designed for Cisco routers. MDRR differs from DRR by the quantum; the former algorithm is assigned a value as given in [8], [10] while the latter algorithm is assigned the packet size in a queue. MDRR was implemented to schedule rtPS and nrtPS scheduling classes and offers same bandwidth sharing to both regardless of their QoS specifications. It improves the DRR algorithm by diminishing the delay present in DRR in serving each queue whenever the queue is visited. No priority of any kind is determined in the MDRR algorithm since the same amount of bandwidth is given to rtPS and nrtPS class. This process stops only under the following conditions; when the deficit counter is zero or when no more bandwidth is available for allocation or when the UL-MAP (Uplink-Mobile Application Part) message is ready to be broadcasted to different users in the network [7]. Newer scheduling algorithms are of great importance to satisfy as many constraints as possible in the network.

TABLE I
ADVANTAGES AND DISADVANTAGES OF EXISTING ALGORITHMS

Scheduling algorithm	Advantages	Disadvantages
FIFO	Simple principle, Fast dequeueing process	Same treatment to all packets
PQ	Based on priority level	Unfair to a continuous set of priority with higher level
EDF	Based on deadline Suitable for real-time applications	Unfair to non-real-time applications
RR	Based on round robin	Allocate same amount of bandwidth Unfair to variable packets, Unfair to real-time applications
WRR	Based on weight principle	Packets dropped Unfair to lower weight
DRR	Based on deficit counter and quantum Solves the unfairness in RR	Unsuitable for delay-sensitive packets when many queues are involved Delay
MDRR	Similar to DRR Reduces delay in DRR	Packets dropped Unsuitable for delay-sensitive packets

As seen in the Table I, each of the algorithms described above presents some advantages and disadvantages that may influence the QoS of heterogeneous traffic. Thus, an enhanced MDRR-based scheduling algorithm realizes the QoS of real-time application while maintaining QoS of non real-time applications in WiMAX networks.

V. AN ENHANCED MDRR-BASED ALGORITTHM

From the literature above, five scheduling classes are defined in the IEEE 802.16 standard. However, in the development of the enhanced scheduling algorithm below, only three scheduling classes are considered (rtPS, nrtPS and BE). This is mainly due to the fact that the UGS and the ertPS classes do not share the same bandwidth channel neither with other scheduling classes (eg: UGS to rtPS) nor with scheduling classes of the same nature (eg: UGS to UGS).

Known for its support of polling service, the MDRR scheduling algorithm is enhanced in this section to improve the QoS support that it offers to rtPS and nrtPS services classes. The important challenge is to prioritize real-time applications (rtPS) while maintaining QoS for non real-time applications in the MDRR algorithm. The new MDRR-based algorithm is defined as a priority-based algorithm that allots higher importance to real-time applications. The implementation design of the enhanced MDRR algorithm is based on separating delay-sensitive packets to delay-insensitive packets. Every rtPS packets stamped with a deadline value is referred to as delay-sensitive packets in the next frame. A packet is considered as delay-sensitive packets whenever its latency is within the deadline the range indicates as given in Equation (1). The specific packet with the latency detected to be within the range is therefore assigned bandwidth. A latency queue is defined to strictly hold packets with high priority (delay-sensitive). If there is any real-time packet whose deadline is outside the range of deadline, the packet contains either an expired time of latency or it is far from reaching its expiring time in the enhanced MDRR scheduling algorithm. Else, the packet is retained in the polling service. The deadline equation is given as follows.

$$\text{Deadline} = [(I_t + L) - C_t > 0 \,\&\&\, (I_t + L) - C_t < (20\%L)] \quad (1)$$

where I_t stands for the insertion time of the packet ın the queue, L represents the latency and C_t represents the current time. If the difference between the time spent in the queue and the instantaneous time is greater than 0 and the same time is less than 20% of the latency time, the packet is still suitable for transmission. Thus, bandwidth is fairly allocated to the specific packet. After extensive simulations, 20 % of latency was found to be suitable for real-time packets where delay becomes unnoticeable. Similar to the existing MDRR algorithm, the deficit and quantum parameters are used to update the status of queues. If no queue was previously attended, the scheduler commences by serving the first queue in the sub-pool and serves one queue after another until the deficit counter (value used to track the number of transmitted bits) becomes negative or the queue is empty. When the deficit counter reaches a negative value, the queue is not considered to be scheduled and only the awaited queues with deficit counter greater than zero are attended to so as to serve the pending packets inside queues. To adjust the deficit counter's value, a follow up on the MDRR update is performed after a packet has been granted a certain amount of bandwidth. When all polling service queues are empty, the scheduler then settles on de-queueing the BE class request. Finally, a staging buffer inside the scheduler accumulates all the information to generate an UL-MAP which is then transmitted by the scheduler in the downlink channel. The UL-MAP consists of a number of **PLs** in which an information element (IE) to specify the bandwidth allocations and the CID to indicate the minimum number of slots through unicast, multicast or broadcast are located.

As the number of subscribers increases in the network, more packets are to be transmitted in the network. Since subscribers increase unexpectedly in the network, an approach that manages as many delay-sensitive packets as possible in the new rtPS queue is introduced. This approach is similar to the previous enhanced MDRR scheduling algorithm but allows a large number of sensitive-delay packets in the queue. In doing so, the 20 % to the 40 % algorithms were distinguished by considering a higher number of delay-sensitive packets in the 40 % latency than in the 20 % latency. Unlike the algorithms discussed in the literature, the enhanced MDRR algorithm maintains fairness by considering priority for both delay-sensitive and delay-insensitive packets.

VI. Simulations And Results

In a point to multipoint (PMP) topology of IEEE 802.16d WiMAX network, 30 nodes were used for the simulation purpose of the enhanced MDRR algorithm with FDD multiplexing mode. At a carrier frequency of 3.5 GHz, a bandwidth of 7 MHz was applied to a framing module with a frame duration of 20 ms and symbol duration of 102.86 ms in the OPNET simulator.

This section analyses the network throughput for 3 scheduling algorithms. In the first case, an even distributed scenario where 30 nodes are evenly supporting 3 classes, 10 subscribers for each scheduling class. A mixed traffic scenario on the other hand is made of a total of 30 nodes distributed as follows. When 50 % of 30 nodes (15 subscribers) support the rtPS class (VoIP), the remaining 15 subscribers are divided into 7 subscribers to support FTP, the other 7 to support email application and one subscriber

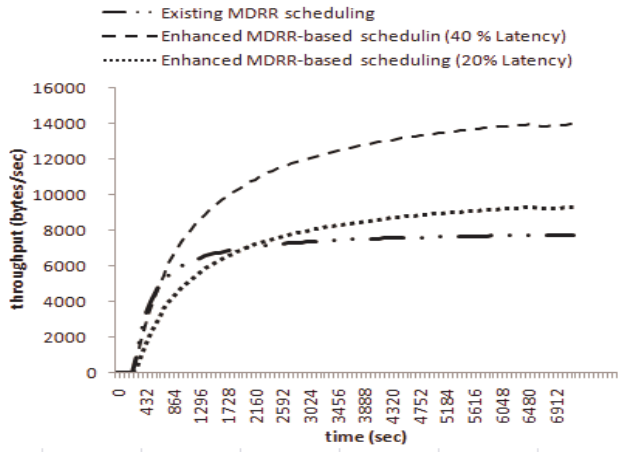

Figure. 2: Average throughput comparison - 30 nodes

simultaneously transmits both FTP and email applications.

As observed from Figure. 2, an average throughput comparison is conducted for the three scheduling algorithms, among which, the 20 % and the 40 % of latency constraints are specified. As mentioned above, the 20% of latency constraint is designed to select among all real-time packets, the delay-sensitive ones which are considered as high priority packets. Same case is applied to the 40 % of latency. When compared to other scheduling algorithms, the 40% enhanced algorithm is translated almost to double that of 20% of latency constraint, an

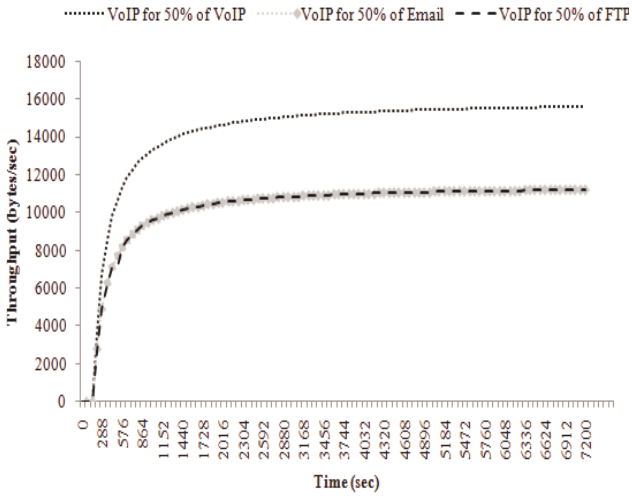

Fig. 3: Impact of mixed traffic on rtPS class

average throughput of 14,000 bytes/sec (112 kbps) was used by the rtPS class. This higher performance results from the fact that the enhanced MDRR-based scheduling algorithm with 40% latency constraint accumulates a maximum or nearly all rtPS packets compared to the 20% latency constraint. Thus, many rtPS packets stamped with higher priority get to be served at first. However, in the existing MDRR algorithm, rtPS resource allocation is achieved at an average throughput of 7,000 bytes/sec (56 kbps).

Figure 3 nevertheless depicts the impact of mixed traffic on real-time application (VoIP) supported by rtPS class. VoIP subscribers are the dominant (leading) subscribers in the network reaching an average throughput of up to 16,000 bytes/sec or 128 kbps. A throughput of 10,000 bytes/sec (80 kbps) is maintained for FTP and email applications regardless of the number of subscribers that support VoIP application. This is mainly due by the high priority constraint defined by the enhanced algorithm. It serves any delay-sensitive packet of VoIP as the delay-sensitive packets are of higher priority than the remaining packets.

VII. CONCLUSION

This paper presented a number of simulations conducted for three scheduling algorithms, the existing MDRR, the enhanced 20% latency MDRR and the 40 % latency MDRR scheduling algorithms. It is worthy of note that the enhanced MDRR scheduling algorithm performs better than the existing MDRR scheduling in terms of the throughput in an even-distributed and mixed traffic scenario. Through the enhanced MDRR algorithm, Quality of Service (QoS) for heterogeneous traffic was achieved and maintained. A 40 % MDRR algorithm was implemented for a wider range of accumulating extra delay-sensitive packets when a network is overcrowded with subscribers. It is to be

concluded that the enhanced MDRR-based scheduling algorithm fairly allocates bandwidth resource since it takes into consideration the QoS requirements from the highest priority packets to the lowest priority packets whereas the existing MDRR algorithm offers the same resource to all applications regardless of their QoS requirements. It is recommended that an implementation of a downlink MDRR-based scheduling is conducted so that a performance analysis from the BS to the SS(s) can be determined. Another aspect of the study that requires further work is the design of the enhanced MDRR-based algorithm for mobile WiMAX where dynamic channel and handovers are involved for bandwidth allocation.

REFERENCES

[1] IEEE 802.16-2001, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems", October 1, 2004.

[2] L. Nuaymi, 2007 “WiMAX: Technology for Broadband Wireless Access”, John Wiley & Sons, the Atrium, Southern June 2007. Available at: http://www.wiley.com.

[3] W. Jakub, “ The IEEE 802.16 WiMAX Broadband Wireless Access; Physical layer (PHY), Medium, Access Control Layer (MAC), Radio Resource Management (RRM)”, Master of Science in Communication Engineering, Institute for Communication Engineering and Networks., Munich University of Technology, January 2005

[4] C. Kwang-Cheng and J. R. B. d. MArca, "MOBILE WiMAX", C. D. Marca, Ed.: John Wiley, John Wiley& Sons, 2008.

[5] C. Claudio, E. Alessandro, L. Luciano, and M. Enzo, "Performance Evaluation of the IEEE 802.16 Mac for QoS Support", IEEE Transaction on Mobile Computing, vol. 6, pp. 1-13, January 2007.

[6] S. Baban, “Design and Implementation of a Scheduling Algorithm for the IEEE 802.16e (Mobile WiMAX) Network”, MSc dissertation Dept. Electronic Systems., Univ Westminster, September 2008.

[7] V. Rangel, J. Ortiz, J. Gomez,“ Performance analysis of QoS scheduling in broadband IEEE 802.15 based networks”, Proceeding of OPNETWORK 2006 Technology Conference, Washington D.C, August 2006, pp. 1-13.

[8] M. Xiaojing, “ An Efficient Scheduling For Diverse QoS Requirements in WiMAX”, Master of Applied Science dissertation Dept. Electronic and computer Engineering., Univ of Waterloo, September 2007.

[9] M. Settembre, M. Puleri, S. Garritano, P.Testa, R. Albanese, M.Mancini, V.Locurto, “Performance analysis of an Efficient Packet-based IEEE 802.16 MAC Supporting Adaptive Modulation and Coding”, International Symposium on Computer Networks, June 2006, pp.11-16.

[10] N.A. Ali, P. Dhrona, H. Hassanein, “A Performance study of uplink scheduling algorithms in point-to-multipoint WiMAX networks”, computer communications, 2008, pp. 1-11

[11] Marie-Paule Gakuba, Mjumo Mzyece and Anish Kurien “An Enhanced WiMAX Scheduling Algorithm for QoS Guaranteed”, Southern Africa Telecommunication Networks and Applications Conference (SATNAC) 2009, 30 August-02 September 2009, Royal Swazi Spa, Ezulwini, Swaziland.

MP Gakuba: received a BTech degree in Electrical Engineering (High Frequency) at Tshwane University of Technology (TUT) in South Africa. She is currently pursuing an M’Tech degree in Telecommunication Engineering at TUT and an MSc degree in Electronics Engineering at the French South African Technical Institute in Electronics (F’SATIE). Her research interest is on Improved QoS for Heterogeneous Traffic Using Enhanced MDRR-Based Scheduling Algorithm in IEEE 802.16d WiMAX networks.