An Empirical Study and the Road Ahead of IEEE 802.16

Lenard Zhong Wei Lee, KuokKwee Wee, Tze Hui Liew, Siong Hoe Lau and Keat Keong Phang

***Abstract*— IEEE 802.16, also known as WiMAX is a commonly used broadband wireless access scheme worldwide. This paper presented a survey of research done in the MAC layer of the WiMAX network. A survey was conducted on three main components of the WiMAX MAC layer; scheduling, bandwidth request/granting and Call Admission and Control (CAC) schemes. Various types of scheduling algorithms were surveyed to provide a description on the workings of each of these scheduling algorithms; highlighting their features, advantages and disadvantages. These scheduling algorithms are separated in groups based on their family. Further, various bandwidth request and granting schemes were also surveyed. These bandwidth request and granting schemes were separated into uplink and downlink directions. These schemes are described and their key advantages and disadvantages are also highlighted. Additionally, research done in the past on the WiMAX Call Admission Control (CAC) are also presented. Lastly, a conclusion on the WiMAX research direction is included as a guide for the author's future works. Thus, this survey would be useful to researchers who are keen on acquiring basic knowledge of the WiMAX network, as well as for researchers aiming to understand the past development before they proceed further.**

***Index Terms*— WiMAX, QoS, Scheduling, CAC, Bandwidth Request/Granting**

I. Introduction

WORLDWIDE Interoperability for Microwave Access (WiMAX) is a Broadband Wireless Access (BWA) technology under the IEEE 802.16 standard. WiMAX allows for a higher data rate and a further transmission range and this compliments well with existing last mile wired networks currently in use. Deployment of WiMAX is faster and cheaper compared to existing broadband wireless access. WiMAX's longer transmission range requires fewer base stations to be able to provide coverage to a specific area.

A key feature of WiMAX is the ability to guarantee QoS by classifying data packets into various service classes. Five service classes exists, which are: Unsolicited Grant Service (UGS), Real-Time Polling Service (rtPS), Extended Real-Time Polling Service (ertPS), Non-Real-Time Polling Service (nrtPS) and Best Effort (BE) [1]. These service classes require different QoS perimeters to fulfill the requirements of their respective applications. A summary of the service classes, requirements and applications [2] are listed in Table I.

WiMAX uses a connection orientated mechanism for traffic transmission [3]. Before the SS and BS is able to communicate with each other, the SS must first register itself to the BS. During this process, QoS requirements that are needed by the SS can be negotiated.

TABLE I. APPLICATION SERVICE FLOWS

Service Type	Definition	Requirements	Application
Unsolicited Grant Service (UGS)	Real-time data streams that contain fixed sized data packets.	Maximum sustained rate. Maximum latency tolerance. Jitter tolerance.	VOIP
Real-Time Polling Service (rtPS)	Real-time data streams that contain variable sized data packets.	Maximum sustained rate. Maximum latency tolerance. Traffic priority. Minimum reserved rate.	Audio/Video Streaming
Extended Real-Time Polling Service (ertPS)	Real-time service that contains variable sized data packets that are generated periodically. Designed for voice applications, ertPS behaves like UGS when there is a voice transmission and lowers the grant size to reduce bandwidth during silence.	Maximum sustained rate. Maximum latency tolerance. Jitter tolerance. Traffic priority. Minimum reserved rate.	Voice with Activity Detection
Non-Real-Time Polling Service (nrtPS)	Variable sized data packets which are delay tolerant and requires a minimum data rate.	Maximum sustained rate. Traffic priority. Minimum reserved rate.	File Transfer Protocol (FTP)
Best Effort (BE)	Data streams that do not need any QoS guarantee.	Maximum sustained rate. Traffic priority.	General data transfers, web browsing

A. Development History and Key Milestones of WiMAX

The first WiMAX release, also known as the IEEE Standard IEEE 802.16-2001 defined the WirelessMAN™

Manuscript received May 01, 2015; revised November 30, 2015.

L. Z. W. Lee is pursuing a Masters in Science (IT) in Multimedia University Melaka, Malaysia.

K. K. Wee is a senior lecturer with Faculty of Science, Information and Technology, Multimedia University Melaka, Malaysia. (Phone: 60-06 252 3961; email: wee.kuok.kwee@mmu.edu.my).

T. H. Liew is a lecturer with Faculty of Science, Information and Technology, Multimedia University Melaka, Malaysia.

S.H. Lau is dean of Faculty of Science, Information and Technology, Multimedia University Melaka, Malaysia.

K. K. Phang is associate professor with Department of Computer System & Technology, Faculty of Computer Science & Technology, University of Malaya, Malaysia.

(Advance online publication: 27 August 2016)

air interface specification [4] for wireless metropolitan area networks (MANs). The IEEE 802.16-2001 standard uses frequencies between 10 to 66 GHz with line of sight (LOS) and a single carrier in the physical layer called the WirelessMAN-SC. WiMAX was designed to be used as a backhaul transmission link between fixed stations. A typical scenario was for the WirelessMAN™ to be a medium to bring the network to the building. Then, existing in-building networks such as Ethernet or Wi-Fi is used to connect to individual devices. An amendment IEEE 802.16a defines additional PHY specifications for frequencies between 2 to 11 GHz with NLOS and WirelessMAN-SCa, WirelessMAN-OFDM and WirelessMAN-OFDMA as the options for the physical layer.

The first revision to the original IEEE 802.16 standard was introduced in June 2004 and was numbered as IEEE 802.16-2004. This made the previous 802.16-2001 standard obsolete. Amendments under this revision included a defined Physical and Medium Access Control Layers for combined fixed and mobile operations in licensed bands (IEEE 802.16e-2005, WiMAX System Release 1), a Management Information Base (IEEE 802.16f-2005), and finally for Management Plane Procedures and Services (IEEE 802.16g-2007).

A second revision, the IEEE 802.16-2009 was introduced in May 2009.Changes in this revision includes the removal of the WirelessMAN-SCa physical specifications due to the lack of interest by vendors in a single carrier physical layer for 2 to 11 GHz frequencies. A mobile multi hop relay specification was also introduced in the amendment (IEEE 802.16j-2009) to provide multi hop wireless connectivity where traffic between the BS and SS can be relayed through a relay station. A second amendment, the IEEE 802.16h-2010 defines an Improved Coexistence Mechanism for License-except Operation. Finally a third amendment, the IEEE 802.16m-2011 for Advanced Air Interface formed the basis for WiMAX System Release 2.

The current revision is numbered IEEE802.16-2012 and it is also known as WiMAX2. The IEEE 802.16m-2011, introduced three key items which are multicarrier operation, extended MIMO support and superframe structure which is aimed at increasing the data rates. Further development are still underway for the future of WiMAX.

B. *QoS and Quality of Experience (QoE) in WiMAX*

Quality of Services or QoS is the performance measurement of the WiMAX network using objective and easily quantifiable variables of the network. These variables include delay, jitter, packet loss rate, and throughput. Packet delay is the amount of time taken by a packet to reach the receiver after it is sent from the sender. Jitter is the variation of the packet arrival time at the destination. Jitter may be caused by a few factors such as the variations in the queue length, variation in processing time needed for the packets arriving at the queue and etc. Packet loss happens when a transmitted data packet is not received by its intended receiver. When a packet is loss on the network, the receiver may request for a retransmission of the packet. Some packet types such as VoIP packets may just be discarded to maintain the quality of the call. Packet loss rate is the ratio of total lost packet to the total transmitted IP packets. Throughput is the amount of data that can pass through the network and is usually measured in bits per second (bps).

On the other hand, Quality of Experience (QoE) is defined as the subjective measure of a user's experience with the service [5]. QoE is not as easily measured compared to QoS as it requires the subjective perceptive of users which are different to each particular user. Two quality evaluation methodologies exist to measure QoE, which are: subjective performance assessment and objective performance assessment. In subjective performance assessment, human subjects will be asked to measure their overall perceived quality of network in a controlled environment. A commonly used measurement technique is the Mean Opinion Score (MOS), recommended by the International Telecommunication Union (ITU). MOS is measured on a scale of 1 through 5 where 1 is bad, 2 is poor, 3 is fair, 4 is good and 5 is excellent. Objective methods are based on algorithms, mathematical or comparative techniques that generate a quantitative measure of the service provided.

C. *WiMAX Architecture*

The WiMAX architecture is made up of two fixed stations which are the subscriber stations (SS) and base stations (BS). Two modes of operations are possible in WiMAX 802.16-2004, namely, the Point-to-Multipoint (PMP) as shown in Figure 1 and the mesh mode[6] as shown in Figure 2. In the PMP mode, transmissions between the base stations (BS) and subscriber stations (SS) are regulated by the BS and there are no peer-to-peer exchanges between SSs. As for the mesh mode, transmissions between SSs are possible. However, each node must coordinate its transmission with other SSs in its extended neighborhood [7]. For the communication between the SS and BS, there are two directions, namely the uplink (UL) and downlink (DL) transmissions. The UL transmission happens in the direction of SS to BS, while the DL transmission happens from the direction of BS to SS[8].

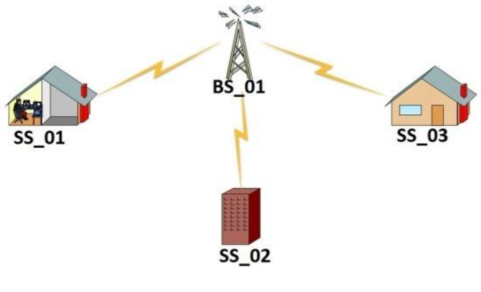

Fig. 1. PMP Mode in WiMAX 802.16

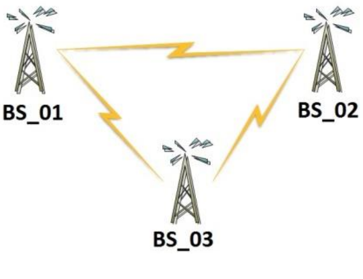

Fig. 2. Mesh Mode in WiMAX 802.16

D. *WiMAX Frame Structure*

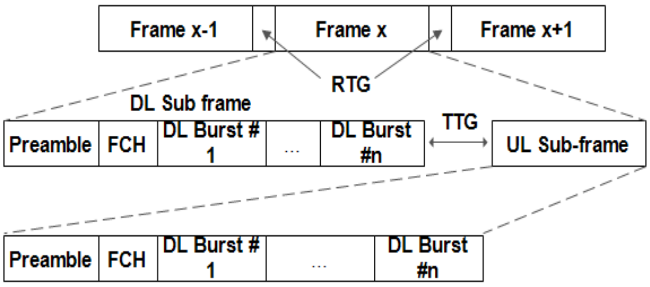

Fig. 3. Frame Structure of TDD WiMAX 802.16

The UL and DL sub-frames can be duplexed by using either the Time Division Duplex (TDD) or Frequency Division Duplex (FDD) [9]. In TDD, UL and DL sub-frames are transmitted at different times and usually in the same frequency while in FDD, the UL and DL can be transmitted simultaneously but at different frequencies [10]. TDD and FDD WiMAX offers its own set of advantages and disadvantages. In TDD WiMAX, a key advantage is in the cost and efficiency of the given spectrum. Traffic for UL and DL usually does not require the same amount of resources at any given time. TDD WiMAX can accommodate both symmetrical and asymmetrical traffic by dynamically modifying the balance of the UL and DL subframes [11] whilst utilizing the full spectrum of bandwidth provided. In FDD, this is not possible because the spectrum needs to be separated into two sub-channels, one for uplink, and another for the downlink. Additionally, these two channels needs to be separated sufficiently in order for the receiver and transmitter to work without interfering with each other. Figure 3 shows the frame structure of the TDD WiMAX frame. Contrary to traditional packet based networks, each frame in WiMAX is measured in terms of a time duration [12]. Frame length varies over releases in the 802.16 standard. Based on figure 3, the frames are separated by a Receive-Transmit Gap (RTG) and individual frame consist of a DL sub-frame and a UL sub-frame. The DL sub-frame consist of a preamble, Frame Control Header (FCH) and the downlink data transmission bursts. On the other hand, the UL sub-frame consist of the ranging slots, Contention for Bandwidth Request (CBR) and uplink data transmission bursts.

II. WIMAX SCHEDULING ALGORITHMS

Scheduling algorithms provides a mechanism to distribute the packets to the users in the network. Packets which arrive at the BS are classified into various queues based on the priority of those packets [13]. Implemented correctly, this classification of packets ensures that the service requirements of the applications are fulfilled efficiently. The goal of a good scheduling algorithm is to maximize the utilization of the network while providing fairness among all users [14].

A. *Priority Schedulers*

In Strict Priority (SP), packets are first classified by their QoS class and placed into queues of varying priorities. Non-empty queues which are of higher priorities will be served first until it is empty. Then the scheduler will serve the next highest non-empty queue [15]. The advantage of this scheduling is that it is always able to guarantee good QoS for high priority packets.

Authors in [16] introduced a Priority-based Fair Scheduling for IEEE802.16-2005 which handles only rtPS, nrtPS and BE traffics. UGS traffic is given a dedicated bandwidth and does not share with the other three types of traffic. In this scheme, rtPS connections will always be served first. The nrtPS connections will be served after all rtPS connections are served. Finally, BE traffics are served. If there are multiple connections of the same class of traffic, then a RR format is used to serve one packet from each connection. A common problem in priority based schedulers is that low priority queues may face starvation when the bandwidth is inadequate.

An uplink scheduling scheme called the Random Early Detection based Deficit Fair Priority Queue (RED-based DFPQ) [17] provides for some adaptation of the network scheduler based on the queue length. The *deficitCounter* is adjusted adaptively during every round of scheduling based on the queue length of rtPS flows. If the current queue length, $QL_{current}$is less than the low threshold of the rtPS queue, $QL_{threshold1}$, then the *deficitCounter* is set to the minimum *deficitCounter* value. If the $QL_{current}$is between the values of$QL_{threshold1}$ and the highest threshold of the rtPS queue, $QL_{threshold2}$ then the *deficit counter* would be set to a dynamic value based on an algorithm. Else, it sets the *deficitCounter* to the maximum value. As the RED-based DFPQ takes into account the current queue length of the rtPS queue in the scheduling process, more transmission slots can be provided to guarantee QoS in rtPS flows. An optimum maximum deficit counter which is not too large also ensures that lower priority service classes are not starved.

B. *Round Robin*

The RR scheduler works by servicing the first packet of the highest priority queue first. Then, the next highest priority queue is served and moves on until all the first packet of each queue is served. The process then restarts at the subsequent packets of the highest priority queue [15]. Every queue will get its turn but QoS requirements cannot be guaranteed. Higher priority queue will get the same allocation with lower priority queue. Enhanced versions of the RR algorithm such as the Weighted RR (WRR), Deficit RR (DRR) and Weighted Deficit RR (WDRR) partly addressed this problem.

In WRR, packets will be segregated into queues based on its service class. The queues can be assigned with a weighted percentage of the bandwidth and then it will be served in a RR order [18]. WRR is able to provide some differentiation between queues by assigning a larger weight to queues which has a higher priority.

DRR is another derivation of the RR algorithm. In WiMAX, scheduling is not done in terms of packets but in slots [19]. Virtual and real queue sized needs to first be converted from bytes into slots. A *Quantum* parameter in terms of slots is given to each queue. Additionally, a *deficitCounter* is used to keep track of the credit (deficit) available to each queue. The *deficitCounter* of every active connection is increased by a fixed amount when that specific connection is being served. In the case of the size of the head-of-line packet is smaller or equal to the *deficitCounter*, the packet is then sent and the *deficitCounter* is decreased by the size of the packet. When the *deficitCounter* is smaller than the head-of-line packet, it moves on to the subsequent connection. The deficit that is stored in the *deficitCounter* of that connection is saved for the following round. The *deficitCounter* is reset to zero when all the packets are served successfully.

Authors in [19] also performed some studies on WDRR for IEEE802.16-2004, which is a variant of DRR where the Quantum can be adjusted according to the current MCS.

C. Earliest Deadline First

The EDF scheduler is commonly used in WiMAX for real-time applications due to the fact that the selection of the SS is based on their delay requirements. This algorithm assigns a deadline to the arriving packets. Priority can then be provided depending on the deadline which has been assigned. Queues which has a smaller deadline will be served earlier compared to queues with a larger deadline. EDF scheduler gives priority to real time traffic because these type of traffics have deadline requirement [20]. Additional modifications or applications of the EDF algorithm has been studied in [8], [21], [22] and [23].

In [8], an enhancement of the EDF method is implemented for IEEE802.16-2005. The researchers use a hybrid of EDF and Weighted Fair Queuing (WFQ) in the scheduling of different types of traffic. UGS traffic requires a fixed bandwidth capacity, rtPS traffic is scheduled by the EDF algorithm, nrtPS via the WFQ and finally the other traffic is allocated the remaining bandwidth equally. The scheduling algorithm also breaks apart BE packets into smaller pieces and stuffs them into the last packet (if it is not at maximum size) of other slots. Hence, BE traffic can be successfully transmitted over the combination of a few of these packets. This method reduces the occurrences of starvation especially for BE traffics. However, there is a need for additional buffer at the BS, additional computation at the BS and SS and extra overhead in the header.

Weighted Earlier Deadline First (WEDF) proposed in [21] for IEEE802.16-2005 considers the destination distance when separating traffics into queues. This algorithm gives higher priority to packets which has a longer travel distance (e.g., number of hops) even if the deadline of the longer distance packet is greater than the shorter distance packet. WEDF uses two classifications which are the IP2GEO and the original WiMAX traffic classification. IP2GEO determines the packet destination distance while the WiMAX traffic classification supports the separation the traffic into different flows. WEDF is found to be able to balance the delay between long distance and short distance packets.

Authors in [22] adopts EDF scheduling in the uplink of rtPS packets in IEEE802.16-2005. Additionally, they also designed an adaptive bandwidth scheduling scheme to maximize the utilization of the remaining bandwidth. In the uplink scheduler, there are three modules which are the information module, database module and service assignment module. The information module extracts from the BW-request message the queue size information and the size of each connection's packet. The scheduling database module stores information for all the connections. Lastly the service assignment module determines the uplink sub-frame allocation in terms of number of bits per SS. In the adaptive scheduling scheme, a Batch Markovian Arrival Process (BMAP) and Newton's interpolation polynomial function is used to predict the rtPS packet's bandwidth requirement. This reduces MAP and MAC SDUs sub-header overhead while increasing throughput over existing algorithms such as WRR and Weighted Fair Queuing.

Authors in [23] introduced an enhancement of EDF called Heuristic Earlier Deadline First (H-EDF) in the uplink scheduler of the WiMAX system. A heuristic function is used to estimate the deadline of packets arriving in queue. In the heuristic model, fairness and efficient utilization of bandwidth is enhanced by assigning queues with a dynamic priority calculated based on the following formula in (1).

$$Priority_i = \frac{c_i}{1+th_i} \tag{1}$$

Where c_i is the transmission capacity and th_i is the historical throughput of the subscriber station.

D. WFQ

WFQ assigns a different weight to each flow so that the flow can be given differing percentage of the total available bandwidth. This prevents any one flow to be allocated all of the bandwidth and thus prevents starvation of the lower priority flows [24]. WFQ has an advantage when the scheduler has to serve many connections or multiple classes of service [20].

An improved version of WFQ, Channel and Duration aware WFQ (CD-WFQ) is proposed in [25] for IEEE802.16-2005.The proposed CD-WFQ consists of two parts of work; the opportunistic feature which allows the scheduling algorithm to exploit the variation of the wireless channel and a duration awareness parameter, Queue Period (QP), to prevent starvations which happens when the opportunistic features gives priority to packets with better channel quality. The scheduler is designed to consider two kinds of awareness, which are the channel and packet waiting duration awareness into consideration. Simulation shows that CD-WFQ is able to provide a lower end-to-end delay for rtPS delay, especially when the mobility speed of the receiver nodes increases. A higher degree of fairness is also achieved because the QP reduces the dominance of opportunistic mechanism in deciding packet priority.

Research on the aging method on the WFQ algorithm [26] has also been proposed with the aim of reducing the packet loss rate in IEEE802.16-2005 transmissions. This scheduling scheme combines a regular WFQ scheduler with an aging method. In order to implement an aging method, two parameters, an aging period and aging time needs to be taken into consideration. An aging period is the periodic time which the observation is done on the packet inside the priority queue of the WFQ algorithm. Aging time is the time limit of the packet aging in the WFQ queue. These parameters are statically set and further optimization is required to obtain better results.

E. Cross Layer Schedulers

DCLASA [27] is able to dynamically adjust the holdoff exponent in the MAC layer by taking into consideration the physical layer's channel quality. The main idea is that links with better channel quality is allocated more timeslots while lower quality links are given reduced timeslots to increase the overall efficiency of the network. DCLASA retrieves information such as the modulation scheme, sending queue length and user QoS requirements from the other layers and assigns timeslots according to these information gathered. With this, DCLASA is able to improve the total throughput of the network and reduce the delay of link transmission. However, DCLASA "stepchilds" the lower quality links by allocating it less timeslots.

Dynamic MCS and Interference Aware Scheduling Algorithm (DMIA) [28] for IEEE802.16-2005 is a two-stage algorithm which takes into account information from the physical layer and makes the necessary adjustments in the MAC layer to satisfy the QoS requirements. In the first stage, the value of bandwidth request of each queue is retrieved by the scheduler in order to calculate the Quantum[i]. The service flows are scheduled in the order or UGS>ertPS>rtPS>nrtPS>BE, consistent with the QoS requirements of these flows. The Quantum[i] is given by Equation (2):

$$Quantum[i] = \sum_{j=0}^{J_i} BW_{max}(i,j) \qquad (2)$$

Where
i = 0,1,2 which UGS, ERTPS, RTPS class.
j = connection index
J_i = total number of connections for *i*th service class.
$BW_{max}(i,j)$ = amount of bandwidth to satisfy maximum sustained traffic rate of one connection within *i*th service class. In the second stage, priority functions are developed and used to determine the scheduling order of each connection and the connections with a higher priority will be served first.

In [29], a cross layer scheduling algorithm with QoS support is proposed by the authors for IEEE802.16-2005. It involves a scheduling algorithm at the MAC layer for multiple connections. The connections uses an Adaptive Modulation and Coding (AMC) scheme at the physical layer where it is assigned a priority based on its channel and service status. The AMC layer strives to maximize the data rate by making adjustments to the transmission modes to channel variations.

F. Other Recommendations

The Differentiated Service algorithm in [15] uses a 6-bit Differentiated Service Code Point (DSCP) field in the header of IP packets to classify the packets and indicate the per hop behavior (PHB). It is a simple algorithm which can provide low latency and guaranteed service to critical and non-critical traffic.

Channel Aware Uplink Scheduler [30]for 802.16-2005is a scheduler in the SS which optimizes the resource allocation based on information obtained from the AMC. Various Modulation and Coding Scheme (MCS) can be assigned depending on the channel quality. There are three channels condition; g*ood*, *intermediate* and *bad* and these depends on the instantaneous SNR and MCS for the SS. 64QAM is used in good channel condition, 16QAM for intermediate channel condition and QPSK for bad channel condition.

The Deficit Fair Priority Queue scheduler [31] for IEEE802.16-2005uses a hierarchical scheduling architecture for the allocation of bandwidth to support all types of service flows. It uses a combination of Deficit Fair Priority Queue (DPFQ) when there are multiple service flows, RR for BE flows, EDF for rtPS flows and WFQ for nrtPS flows.

The scheduling strategy in [32] for IEEE802.16-2004 involves one scheduler scheme at the BS and another scheduling scheme at the SS. At the BS, two types of queues are defined which are the type I and type II queues. Type I, which is processed first, schedules data grants for UGS and allocates dedicated requests for rtPS and nrtPS queues. Type II schedules data grants for rtPS, nrtPS and BE queues based on the information from the bandwidth request message. A fair queuing algorithm is proposed here to ensure fairness as well as to ensure that each service flow receive a minimum allocation. At the SS, an additional scheduler is proposed to be implemented in each SS to reassign the received transmission opportunities among the various connections. The SS scheduler can be adapted to meet the different requirements of each service flow. This proposed architecture is able to fulfill the basic requirements of QoS for the various traffic types. When virtual timestamps for nrtPS are introduced, the SS scheduler helps to increase the overall throughput of nrtPS traffic. However, at the same time, there is a tradeoff involved for BE traffic.

A modified Latency Rate (LR) schedulerforIEEE802.16-2004 which uses a token bucket algorithm is proposed in [33]. The token bucket is used to limit the incoming traffic while the modified LR scheduler allocates the rate for each user. This LR scheduler's behavior is dependent upon two parameters which are the *latency*, $\vartheta_i$and Allocated Rate, r_i.The *latency* which is in a Time Frame (TF) period needs to be optimized in order to increase the number of connections that can be accommodated by the CAC. By considering an ideal TF in the 802.16 standard, the system is able to optimize the allocations of users in the system. An optimal TF can cater to a bigger number of users. This scheduler is also able to ensure a guaranteed upper limit of the delay.

A scheduler with compensation algorithm is proposed by the authors in [34]. This proposed scheduler consists of five parts which are the packet classifier, bandwidth requests with carrier to interference and noise ratio (CINR) reports, packet scheduler, channel aware compensator and buffer manager. In the packet classifier, packets arriving at the BS are sorted into their respective traffic class queue which will be managed by the scheduler. The bandwidth request with CINR monitors the channel quality between the BS and SS, attempts to predict the future channel state and then sends this information to the BS scheduler. The packet scheduler at the BS then uses an adapted WF2Q+ algorithm to schedule the packets. While these processes are ongoing, the channel aware compensator assists the scheduler in selecting the substitute flow via the management of a *debit/credit counter* for each flow. When meeting poor channel conditions which prevents a head of line (HOL) packet to be transmitted, this mechanism chooses the substitute flow with the highest *debit counter* among unmarked flows. At each frame's beginning, the buffer manager will check the queues for packet which has expired past the tolerated delay and purge the packets. The advantage of this process is, there will be fair sharing of bandwidth as long as there is no persisting poor channel condition. Higher *goodput* is also achieved via the use of a buffer manager. Although there is an increase in delay accumulated by the queued packets which are awaiting for a clean channel, it is under the maximum tolerated QoS delay values for each traffic class.

Service Class Downlink Scheduling (SCDS) [35] for IEEE802.16-2005 is a scheduling scheme at the BS which makes two segregations before service the traffic classes. The packets are first segregated into queues meant for a particular SS. Then it segregates the packets meant for each SS into service type queues. Service classes with a higher priority as UGS, ertPS and rtPS are given more time slots than lower priority service classes like nrtPS and BE.

An uplink scheduler[36] is developed to distribute the uplink bandwidth in a way that the QoS requirements of each connection in the SS are satisfied. The policing of the maximum sustained traffic rate and maximum traffic burst is performed by a dual leaky bucket regulator. The maximum latency to the flows that do not exceed their minimum reserved traffic rates can be guaranteed.

A two-layer classification of packets is proposed in this scheduling scheme [37].In the first layer of classification, packets are classified based on the types of data that it is. This will either be UGS, rtPS, ertPS, nrtPS or BE. In the second layer of classification, packets in each of the data types are classified again based on their destination nodes. The BS will maintain a separate queue for each of the service class of each destination node. These queues are then scheduled by the first classification (service classes) using a specific scheduling algorithm for each service class. For UGS, an EDF scheduling algorithm is used to provide priority based on the packet's arrival time and maximum latency. Subsequently, a modified EDF algorithm is used for rtPS class, WFQ for ertPS class, RR for nrtPS class and FIFO for BE class.

Enhanced Adaptive Proportional Fairness (E-APF) [38] scheduling algorithm considers QoE instead of QoS. A delay outage concept is introduced. Delay outage happens when packets transmitted experiences a delay greater than its specified allowable threshold. A parameter called the Packet Delay Outage Ratio (PDOR) is the maximum ratio of packets that are delivered for packets surpassing the allowable threshold.

Two-Rate-Based Scheduler (TRBS) [39] for IEEE802.16-2012 aims to provide fairness to the pool of users based on the radio resources left after satisfying the guaranteed demands. Before doing the scheduling, the proposed scheme used a traffic shaper to provide priority to traffic types that has higher priority. After shaping the traffic, the scheduler tries to achieve a Max-Min fairness by allocation users with equal requirements, a same share of the airtime regardless of the mobile station's CQI. Two important QoS parameter used in shaping the traffic are Minimum Reserved Traffic Rate (MRTR) and Maximum Sustained Traffic Rate (MSTR). The first stage of the scheduler schedules the S_{MRTR}before the second stage performs the S_{MSTR}scheduling. In the first stage, scheduling is done based on a SP rule between guaranteed and permissible traffic demands whereas the second stage schedules the packets in equal amounts until their demand are satisfied.

A scheduling strategy which takes into account the QoE of the user is proposed [5]. This scheduling strategy consist of three QoE levels where each user is given an initial max data rate, minimum subjective requirement and mean subjective threshold value. A specific packet loss rate threshold has to be set before transmission can occur. Each user will perform transmission on their maximum transmission rate until the packet loss rate is higher than the initially set threshold. In this stage, the user will check if the transmission rate exceeds their subjective requirement. If it is so, then transmission rate will be decreased. Else, transmission rates remain unchanged.

Researchers proposed a scheduling algorithm which uses a weight equation [40] to allocate bandwidth among queues. The weight function, $W_i(t)$ is defined in terms of two parameters which is the minimum reserved traffic rate and average packet size.

$$W_i(t) = \frac{\max(P_{javg}(t))}{P_{iavg}(t)} * \ w_i \tag{3}$$

Where $P_{javg}(t)$ is the maximum average packet size at time *t* and $P_{iavg}(t)$is the time varying average packet size of queue *i*. The steps taken by the scheduling scheme is as follows:

1. Classify packet into queues based on service flows.
2. Calculate the $P_{iavg}(t)$.
3. Calculate the $w_i$based on Equation (4):
   $$\sum_{i=1}^{n} W_i(t) = 1 \tag{4}$$
4. Calculate $W_i(t)$.
5. Distribute the UL sub-frame bandwidth based on Equation (5):
   $$BW_i = \ W_i * UL_{BW} \tag{5}$$
   Where BW_i is the reserved bandwidth of queue *i* and $UL_{BW}$is the total bandwidth of UL sub-frame.
6. Send bandwidth value of each queue to SS.
7. Continue servicing the queue until bandwidth is insufficient.
8. Move between queues using a round robin scheme.

A hybrid downlink scheduler [41] which uses a modified Greedy Latency algorithm with a Shortest Job First scheduler is proposed to reduce the latency, packet loss rate while increasing throughput.

G. Summary of Scheduling Algorithms Surveyed

TABLE II. SCHEDULING ALGORITHMS SURVEYED

No.	Algorithm	Advantages	Disadvantages
1.	SP [15]	Guarantee good QoS for high priority packets.	Low priority packets will experience starvation.
2.	Priority-based Fair Scheduling [16]	Simple implementation. Delay for real-time services such as multimedia streaming can be maintained at an acceptable level.	Low priority packets will experience starvation.
3.	RED-based DFPQ [17]	Decreased delay in rtPS queues. Good overall throughput for rtPS	Reduced throughput in nrtPS queues.

No.	Algorithm	Advantages	Disadvantages
		queues.	
4.	RR [15]	Each queue will get equal scheduling time.	No QoS guarantee.
5.	WRR [18]	Reduction in packet loss.	Average delay of all QoS classes are not reduced. Not suitable to be used for multimedia applications.
6.	DRR [19]	Good for real time traffics.	Lower overall throughput compared to WDRR.
7.	WDRR [19]	Outperforms DRR in MAC throughput and TCP goodput.	Starvation avoidance feature is required.
8.	EDF [20]	Low delay for real time traffic.	Higher delay for lower priority traffic such as FTP and HTTP.
9.	Enhanced EDF [8]	Reduced starvation of low priority traffics.	Added implementation complexity in form of additional computation at BS and SS as well as extra overhead in packet header.
10.	WEDF [21]	Delay between long distance and short distance packets are balanced.	Increased complexity because additional techniques are required to differentiate distances.
11.	Low Overhead Uplink Scheduling [22]	Reduced overhead, leading to increased overall system throughput.	Some increased calculation complexity because the scheduler needs to be able to predict future bandwidth requirements.
12.	H-EDF [23]	Reduces bandwidth wastage and lower end to end delay.	Simulation results shows a lower overall system throughput compared to other existing algorithms.
13.	WFQ [24]	Low and constant delay for voice, FTP and HTTP queues if traffic is within load.	High delay for video queues if load is over 100%.
14.	CD-WFQ [25]	Better throughput, delay and fairness for rtPS compared to WFQ.	Added complexity in the form of a opportunistic mechanism and queue duration awareness.
15.	Aging Method on WFQ [26]	Outperforms WFQ in terms of throughput when there is a large proportion of UGS traffic.	More optimization for *aging period* and *aging time* is required for better performance.
16.	DCLASA [27]	Improve total throughput and reduced delay.	Lower quality links gets fewer timeslot allocation.
17.	DMIA [28]	Outperforms WRR and RR in terms of fairness, throughput delay and packet loss.	Requires additional complexity to determine the priority function.
18.	Cross Layer Scheduling with QoS Support [29]	Efficient bandwidth utilization. Throughput guarantee for nrtPS as long as there is enough bandwidth. Low implementation complexity.	Rate performance for BE connections is negatively affected when the bandwidth is inadequate.
19.	Differentiated Service [15]	Simple algorithm. Low latency and guaranteed service to both critical and non-critical traffic.	Lower throughput compared to round robin and may have higher latency when there is a large number of mobile stations.
20.	Channel Aware Uplink Scheduler [30]	Improves delay and throughput for real time services and rtPS.	Overall system throughput may be affected because it gives priority to lower channel condition users first when scheduling the packets.
21.	Deficit Fair Priority Queue [31]	Generally improved throughput and fairness under unbalanced uplink and downlink traffic.	Traffic rate for nrtPS flows are lower compared to the PQ scheduler when the nrtPS flow rises.
22.	Two Steps Scheduler [32]	Increased throughput in nrtPS flows.	Decreased quality of BE traffic.
23.	Modified LR Scheduler [33]	Able to cater to a larger number of users. Guaranteed upper limit of delay.	Choice of a non-optimal TF may reduce the number of SS that is able to be served.
24.	Scheduler with Compensatio n Algorithm [34]	Fairer bandwidth scheduling.	Increased delay of packets in the overall system.
25.	SCDS [35]	Reduced delay and increased throughput for UGS, ertPS and rtPS flows compared to RR.	Does not consider channel quality, hence it is unclear if channel quality will affect performance.
26.	Dual Leaky Bucket Uplink Scheduler [36]	Provide maximum latency and minimum rate guarantees for intermediate and high priority queues.	Higher latency values and rejected connections for rtPS flows under heavy load.
27.	Two Layer Classification Scheduling [37]	Scheduler able to handle large amount of data traffic. Better throughput and delay for all service flows.	Use of higher modulation rates to increase throughput may cause the transmission to be susceptible to noise and interference.
28.	E-APF [38]	Maximize overall system throughput while maintaining QoS requirements.	Increased complexity due to the requirement of an estimation model.
29.	TRBS [39]	Performs up to 75% better compared to existing PF scheduler in terms of goodput.	Usage of 64-QAM 3/4 coding scheme may be susceptible to noise and interference.
30.	QoE Scheduler [5]	Lower packet loss, delay and jitter compared to existing QoS schedulers.	Throughput is lower compared to existing QoS schedulers.
31.	Weight Equation Scheduler [40]	Higher throughput, lower delay, lower jitter compared to WRR at higher number of SSs.	WRR performs better in throughput, delay and jitter when the number of SS low.
32.	Modified Greedy Latency and SJF [41].	Low latency and low packet loss rate.	Simulation results show high packet loss ratio when the number of processes or service are low.

III. Bandwidth Allocation and Request

In the 802.16 standard, bandwidth is allocated to each of the SS based on each SS's bandwidth requirements. The BS receives bandwidth request from each of the SS in the form of a MAC frame before it is able to allocate bandwidth. Bandwidth request is done via two main methods, a stand-alone bandwidth request header or from a piggyback request.

The stand-alone method uses a dedicated MAC frame to indicate the number of bytes it requires for uplink. Comparisons in performance between stand-alone bandwidth request and piggyback request have been studied in and it was concluded that piggyback requests is best used in scenarios where there is a large number of users and when packets have a short inter-arrival time [42]. This is due to the fact that in these two scenarios, the chances of collisions to occur is high and piggybacking avoids these collisions.

There are two main mechanisms in the 802.16 standard for bandwidth request transmission: contention based random access and contention free polling [43]. Both methods has its advantages and disadvantage [44]. In contention based random access, all SSs contend to obtain transmission opportunities for sending request using contention resolution mechanisms. Multiple SS may transmit their bandwidth request message at the same time and this will cause collisions. Therefore this method does not always guarantee the success of resource reservation. To resolve contentions that occur, a random backoff mechanism based on a truncated binary exponential backoff (BEB) is used [45]. From this method, the SS only performs this procedure when it wants to transmit a BR. This may lead to higher bandwidth utilization. As for contention free polling, the SS only sends its bandwidth request when it is polled by the BS. The BS will maintain a list of SSs and will poll them one by one to allow them to transmit their bandwidth request message. This scheme is able to provide a guarantee to a successful resource reservation. However, if the polled SS does not need a BR, then transaction opportunities will be wasted and this leads to reduced bandwidth utilization.

Figure 4 shows the direction for UL and DL transmission in WiMAX. UL is traffic sent from the SS to the BS while DL is traffic sent from the BS to the SS.

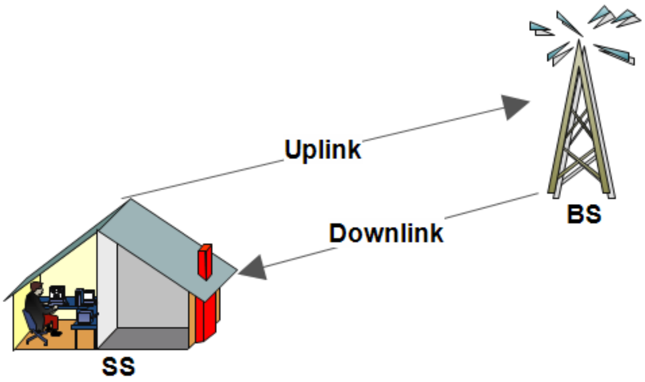

Fig. 4. Uplink and Downlink Transmission in WiMAX 802.16

A. Uplink Bandwidth Allocation and Request

This section surveys the research done to optimize the uplink bandwidth allocation and request in WiMAX systems.

In [46], CDMA-based bandwidth request method, during the generation of uplink data, the SS generates a BR code that consists of the channel quality information (CQI) and the amount of slots required for transmission. The 8-bit BR code is made up of 3 bits for the MCS code and 5 bits for the slot code. The BR code will then be sent to the BS and it will allocate the required uplink bandwidth. This proposed method provides a higher probability of success of a BR transmission.

Intelligent Bandwidth Allocation of Uplink (IBAU) [47] introduces two modules; a Service Flow Management (SFM) module and an Uplink Bandwidth Management (ULBM) module in the SS. SFM send the Dynamic Service Addition (DSA), Dynamic Service Change (DSC) and Dynamic Service Deletion (DSD) to the ULBM. ULBM will calculate the total required bandwidth and reply with information of UL bandwidth size. Then, the service flows will be scheduled by the UL scheduler. AMC will be dynamically adjusted. This method improves the throughput, reduces the delay and reduces the occurrences of starvation in the service flow by using the available resources more efficiently.

A bandwidth request and handling scheme called Intensive Bandwidth Request and Handling introduces a sub-module called the rtPS Rapid Request (rRR) in the SS and nrtPS Redundant Detection and Normalized (rRDN) sub-module in the BS. The Rapid Request (rRR) sub-module [48] minimizes the interval gap for rtPS polling to improve on QoS performance. At the same time, the rRDN detects and removes redundant nrtPS bandwidth requests. Additionally, the rRDN also allocates the bandwidth to the SS in a way that is not more than what is requested.

In [49], bandwidth request slots are allocated during a contention free (CF) period of the polling. This scheme gives priority to ertPS flows by letting it take over some of the contention time of other lower priority flows such as the nrtPS and BE. Bandwidth request can be transmitted immediately during the CF period if it is an ertPS flow. Else, it participates in the contention period of nrtPS or BE flows. This reduces the delay of bandwidth requests while providing a slight increase in throughput.

Efficient Uplink Bandwidth Request with Delay Regulation [50] introduces two new concepts which are target delay and dual feedback. Compared to traditional BW-REQ scheme which aims to minimize delay to a minimum, the target delay is a parameter which denotes the maximum allowable delay for the service flow. The target delay is calculated using Equation (6):

$$T_q = T_{ref} - T_o \qquad (6)$$

Where T_{ref} is the target MAC-to-MAC delay and$T_o$is the additional delay encountered at the MAC except queuing delay.

The target length of transmission queue in SS,Q_{ref} is denoted by Equation (7):

$$Q_{ref} = l\frac{T_{ref}-T_o}{T_a} \qquad (7)$$

The rate of additional bandwidth request, $\Delta b_q(t)$, needs to be increased when the queue length,$q(t)$ increases over Q_{ref}. However, these dynamic changes react slowly and hence the Dual Feedback Approach is introduced. There will be two feedback loops, one for queue length and one for the rate of additional bandwidth request. The rate feedback mechanism provides a form of predictive information about the queue length and this allows quicker change in the bandwidth request control. Although computational and packet header overhead is present, it is negligible due to its simple and minimalist nature. This scheme minimizes delay jitter while maintaining the QoS for the queues.

Adaptive Bandwidth Allocation Scheme (ABAS) in IEEE802.16e-2005 [51] adjusts the bandwidth ratio according to current traffic profile while working together with the scheduler to throttle TCP source when acknowledgements received are infrequent. Specifically on the BA side, the *schedulablebandwidth* that can be allocated to upstream and downlink traffic must be determined to be able to maximize the throughput of the network. The steps taken by the BA is as follows:

1. BS retrieves information of number of n_{dSS} andn_{uSS}, whereby $n_{dSS}$is the number of SS which has downlink TCP transfers and n_{uSS} is the number of SS which has uplink TCP transfers.
2. The download asymmetry ratio is ensured to be equal to one.
3. Split between uplink and downlink is adjusted.
4. BS informs SS about adjustment results. BS also informs scheduler about new values of allocated schedulable bandwidths in UL and DL direction.

Results show that ABAS has a higher aggregate downstream throughput and a smaller average uplink access delay.

In Priority Based Bandwidth Allocation Scheme [52], the researchers proposed an uplink scheduler in the SS which considers the service classes when assigning priority is proposed. UGS is given a guaranteed bandwidth and is assigned to the highest priority queue. The rtPS and nrtPS are assigned to the next highest priority queue. Both the rtPS and nrtPS traffic classes needs to request for the bandwidth allocation of the next frame according to the periodical interval given. The calculation of the bandwidth required takes into consideration the size of the queue and the arrival time of the packet. Simulated results show a slight improvement in the overall network throughput. However, as more nodes are present, the margin of improvement drops. The results also show that ertPS traffic will be given more bandwidth and this is due to the higher QoS priority at the expense of rtPS traffic. In this situation, the rtPS traffic will suffer big average delay.

B. Downlink Bandwidth Allocation and Request

A downlink bandwidth allocation and request scheme [11] is proposed to make use of the remaining bandwidth after each round of scheduling. The BS is able to adaptively decide on which mechanism of bandwidth request to choose based on the remaining bandwidth. Contention free scheme is chosen when the amount of remaining bandwidth is sufficient for at least one additional bandwidth request message. The BS selects some SSs to transmit their BR messages based on information about the SS's backlogged packets. Otherwise, the contention based mechanism is chosen. The algorithm used in the BS side is as follows:

SSList (list of all SSs which have nrtPS or BE traffics)
freeBW (free uplink BW after scheduling of UL sub-frame is finished)
theSS (selected SS allocated the BW for transmitting the bandwidth request message in contention-free scheme)

1. BS creates *SSList* based on QoS parameter information of each SS.
2. IF BW is available for transmitting at least one bandwidth request message AND *SSList* is not empty, select contention-free scheme.
3. IF contention-free scheme is chosen, *theSS* from *SSList* is chosen and bandwidth is allocated to *theSS*.
4. Remove *theSS* from *SSList*.
5. Repeat until UL bandwidth is not sufficient or *SSList* is empty.
6. IF BW available is not enough for a bandwidth request message, contention-based scheme is selected.
7. If contention-based scheme is selected, remove all SS from *SSList*.
8. All SSs contend with other SSs to transmit bandwidth request message.

In this scheme, bandwidth wastage from unassigned allocation can be reduced. This leads to higher throughput in the overall system. Average delay is also found to be shorter. Additionally, nrtPS and BE flows can also receive more bandwidth (from the remaining unallocated bandwidth) without affecting the UGS and rtPS flows.

A dynamic bandwidth allocation scheme [37] used after a destination classification step is proposed. In the destination classification step, packets are segregated into queues based on their destination nodes. To perform the dynamic bandwidth allocation, the available bandwidth of the channel is first estimated before the transmission of the next frame. During the transmission of a packet with higher priority, bandwidth allocation for lower priority packets are throttled to accommodate the higher priority packet. A parameter called the residual bandwidth, $B_r$is measured dynamically and this parameter is dependent on two factors which are the *fairness* and *bandwidth utilization* factor. *Fairness* is defined as the ratio of allocated bandwidth to the requested bandwidth while *utilization factor* is defined as the ratio of throughput to the allocated bandwidth. After the bandwidth allocation scheme, the packets are scheduled using a Dynamic Priority (DP) scheduler which transmits packets from the BS to SSs based on the available bandwidth to each traffic queue.

In Efficient Downlink Bandwidth Allocation Scheme (EDBA) [53], a concept called the burst allocation problem (BAP) is introduced. The solution of the problem aims at solving the question of how the parameters of shape, location and modulation scheme can be adjusted to obtain a high sub-channel utilization. The determination of these parameters are known to be a nondeterministic polynomial (NP) problem. EDBA has four steps which aims to determine these parameters. The first step is to determine the sequence of burst allocation so that the BS can serve the bursts to candidates with the best modulation levels. Step two uses the sub-channel efficiently by determining the width and length of the burst. A third step only executes when an External Bandwidth Wastage (EBW) occur and this step checks for unallocated transmission slots and allocates them to a burst to the SS. The last step is the determination of the burst location. After the burst shape is determined and the EBW is alleviated, BS now makes calculations to determine the location of the burst. EDBA provides better sub-channel utilization which increases the overall throughput while eliminating EBW.

Adaptive Bandwidth Allocation (ABA) [54] works by first assigning bandwidth to the UGS traffic. Next, for rtPS, nrtPS and BE flows, an initial bandwidth requirement is estimated and the bandwidth is assigned based on that estimation. To prevent starvation of BE flows, rtPS and nrtPS flows are only given the bandwidth it needs to meet its QoS delay constraints and minimum throughput requirement respectively. If there are any remaining bandwidth, it is then further assigned to all the traffic flows to prevent wastage. ABA is able to reduce the possibility of starvation of BE traffic while maintaining the QoS requirements for rtPS and nrtPS flows. The ABA scheme is not greedy in taking bandwidth from the lowest priority traffic.

Dynamic BA (DBA) [55] is able to adjust allocations based on the traffic characteristics and network conditions. The traffic arrival rate is used to characterize the traffic behavior while two metrics, fairness (*Fr*) and utilization (*Ut*) is used to characterize the network conditions. Fairness is defined as the ratio of *allocated bandwidth* over *requested bandwidth* for each of the service flows. Utilization is the ratio of *throughput over allocated bandwidth*. BA is then done according to a specific formula which is given:

$$bw_{alloc}^{T} = \left(\frac{dCap^{n}}{dCap^{T}}\right) \times bw_{avail}^{T} \qquad (8)$$

$$bw_{avail}^{T} = \sum_{n \in N} \sum_{i=1}^{I} bw_{res,i}^{j-1} + bw_{rem}^{T} \qquad (9)$$

$$bw_{res,i}^{n,j} = bw_{alloc,i}^{n,j-1} - bw_{req,i}^{n,j} \qquad (10)$$

$$bw_{rem}^{T} = BW - bw^{T} \qquad (11)$$

Where

BW = total link capacity

bw^{T} = total occupied bandwidth

bw_{rem} = remaining bandwidth

bw_{res} = residual bandwidth

The allocated bandwidth, bw_{alloc}^{T} is given to each of the service flows based on the available bandwidth bw_{avail}^{T} to the service request.

A Multi-Round Resource Allocation scheme [56] for the downlink is proposed to provide resource allocation of the multiple service classes in WiMAX. This algorithm takes into account five key parameters which are the CQI, inherent priority of each service flow, QoS requirement of each service flow, fairness among users and fairness among the service flows in a multi-round scheme. Before performing the first round of allocation, the BS should be aware of the MCS of each SS. This can be measured periodically using the channel estimation module of each SS and the SS should report this to the BS periodically. Additionally, the resources available in the OFDMA frame also need to be computed according to the CQI as well as the minimum and maximum amount of traffic in each flow. In the first round of scheduling, the resources are allocated to the flows based on their priority. The priority rank is UGS first, followed by ertPS, rtPS and nrtPS. The first round provides allocation up to the minimum rate of each of the service flows. The Multi-Round Resource Allocation scheme of that frame stops on the first round if the resources are not enough to satisfy the minimum traffic rate of these flows. However, if resources remain, a second round is initiated. In the second round, an algorithm similar to WFQ is used to provide allocation for the BE service flows. A *weight* is computed based on a few parameters, namely the MCS of each SS, amount of resources allocated already in the previous round, and the maximum amount of data that needs to be sent. Then with the *weight*, resource is allocated to the service flow for each SS. If there are still resources remaining, then the third round of allocation will allocate the resources according to the inherent priority of the service classes. In this round, MCS of each SS will need to be considered and results will be adjusted to prevent the resources allocated exceeding the packet queue length and maximum amount of traffic.

IV. WIMAX ADMISSION CONTROL

The WiMAX Connection Admission Control (CAC) plays an important part in the 802.16 standard to provide QoS guarantee. CAC is able to provide QoS guarantee via two main ways. Firstly, the CAC prevents the system from being overloaded. Secondly, by blocking certain connections, different traffic loads can assigned to different priority [57]. Since the IEEE 802.16 standard is a connection oriented system [58], all applications must first establish a connection with the BS and be classified into a specific service class before data transmission can happen. In the WiMAX architecture shown in Figure 5, the applications from the SS will send its service flow parameters through a Dynamic Service Addition (DSC), Dynamic Service Change (DSC) or Dynamic Service Delete (DSD) request. At the BS, the admission control will check the connections with QoS requirements (UGS, rtPS and nrtPS) and decide if the connection's QoS can be satisfied by the bandwidth available in the BS [59]. If the admission control mechanism accepts the connection, it will assign the connection with a 16-bit Connection Identifier (CID) and informs the scheduler to allocate the required bandwidth. At the SS side, an uplink scheduler retrieves the packets from the queues and transmits them based on slots which are specified in the Uplink Map Message (UL-MAP).

A Novel CAC Algorithm called the Greedy Choice with Bandwidth Availability aware Defragmentation (GCAD-CAC) [7] is proposed. In GCAD, each mesh node is designed to support three data traffic classes with priority values of "1", "2", and "3". The value "1" is the highest priority class while "3" is the lowest priority class. In GCAD, when a source tries to transmit data, it must first identify a path to send its intended data to the destination.

The admission control designed in [36] admits connections which fulfils Equation (12):

$$C_{reserved} + TR_{i}^{service} \leq C \qquad (12)$$

Where

$TR_{i}^{service}$ = traffic rate that should be guaranteed to connection *i* which has the service type service.

$C_{reserved}$ = capacity already assigned to connections admitted into the system.

C = capacity available for the uplink scheduler.

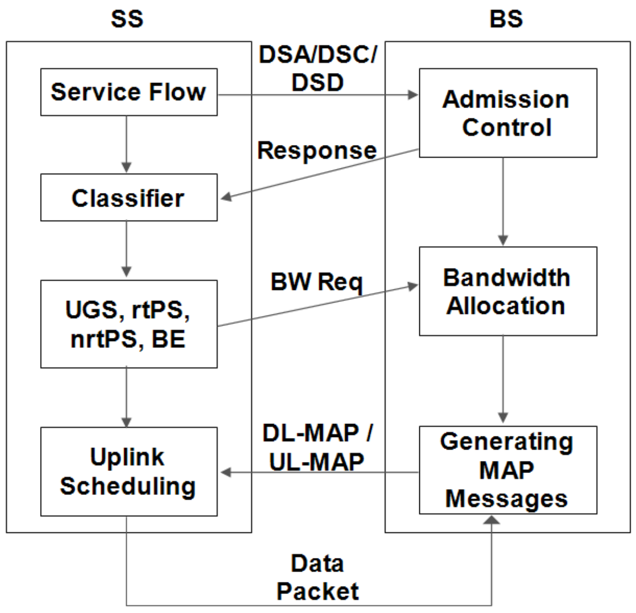

Fig. 5. QoS in WiMAX 802.16

Researchers also presented two admission control methods [60] for real time services. The first method, Measurement-Based Admission Control (MBAC) utilizes the average number of free slots as inputs in the admission decisions. For each frame i, the $freeSlots_i$, which is the number of remaining DL and UP slots are checked to get the $freeSlotAv_i$, which is the exponentially weighted moving average. The second method, Measurement-Aided Admission Control (MAAC) has a bookkeeping mechanism which keeps track of the updates of reservation limits for DL and UL traffic. When a new connection arrives, the sum of the current reserved DL/UL bandwidth and the Minimum Reserved Traffic Rate (MRTR) is checked if it is below the corresponding limit. If so, this connection is added and the MRTR is added to the reserved DL/UL bandwidth. When the connection is removed, the MRTR is subtracted from the reserved DL/UL bandwidth.

V. Conclusion

This study presented a survey of the WiMAX network, specifically on modules in the MAC layer. Three main focus of this survey are the scheduling algorithms, bandwidth request and granting scheme, and the CAC in WiMAX. WiMAX scheduling algorithms such as the RR and EDF family of schedulers has been extensively covered in this paper. This study noted the fact that many legacy schedulers in older 2.5G or 3G technology have been modified to be used WiMAX. Often, there is a tradeoff between a few performance metrics such as overall throughput, fairness and packet loss rates. However, QoS requirements that existed during the introduction of these legacy schedulers are different compared to today's QoS requirements.

Of late, there is an exponential growth of demand for higher bandwidth which is caused by the increase in the number of internet enabled devices as well as the applications requirements such as from High Definition Voice and Video. This has led to a strong interest to develop a more capable network which can support the increase in bandwidth demand especially for multimedia services. Future enhancement of the WiMAX MAC layer must be designed to take into consideration this factor. WiMAX is seen as one of the candidates in 4G networks to be used worldwide due to its high speed, efficiency and flexibilities.

It is also noted that most of the researches done are focused on improving the overall QoS of the network. We opine that more research can be done from a different perspective whereby the researchers give more consideration to satisfying QoE rather than QoS. In the end of the day, user's experience of the network is also an important concern.

References

[1] H. J. E. Blanco and I. P. P. Parra, "Evaluation of scheduling algorithms in WiMAX networks," Proc. ANDESCON, Bogota, 2010, pp. 1–4.

[2] C. So-in, R. Jain, and A. K. Al-Tamimi, "Resource Allocation in IEEE 802 . 16 Mobile WiMAX," in Orthogonal Freq. Div. Mult. Access, T. Jiang, L, Song. Y. Zhang, Boca Raton, FL: Auerbach, 2010, pp. 1–48.

[3] O. Alanen, "Multicast Polling and Efficient VoIP Connections in IEEE," Proc.10th ACM Symp. on Modeling, Analysis and Simulation of Wireless and Mobile Systems, New York, 2007, pp. 289-295.

[4] C. Eklund et al., "IEEE Standard 802.16: A Technical Overview of the WirelessMAN TM Air Interface for Broadband Wireless Access," IEEE Commun. Mag., vol. 40, no. 6, June 2002, pp. 98–107.

[5] T. Anouari, andA. Haqiq, "Improved QoE-Based Scheduling Algorithm in WiMAX Network," Proc.2014 Int. Conf. on Multimedia Computing and Systems (ICMCS), Marrakech, 2014, pp. 878-883.

[6] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support,"IEEE Transactions on Mobile Computing,vol. 6, pp. 26–38, Nov. 2007.

[7] F. De Rango, A. Malfitano, and S. Marano, "GCAD: A Novel Call Admission Control Algorithm in IEEE 802.16 based Wireless Mesh Networks," J. Networks, vol. 6, no. 4, pp. 595–606, Apr. 2011.

[8] C. Nagaraju and M. Sarkar, "A Packet Scheduling To Enhance Quality of Service in IEEE 802.16," Proc. World Congress on Engineering and Computer Science.San Francisco, 2009, pp. 6–11.

[9] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of Service Support in IEEE 802.16 Networks," IEEE Network, vol. 20, no. 2, pp. 50–55, Apr. 2006.

[10] D. Chuck and J. M. Chang, "Bandwidth Recycling in IEEE 802 . 16 Networks," IEEE Trans. Mobile Computing, vol. 9, no. 10, pp. 1451-1464, Oct. 2010.

[11] S. Kim, W. Kim, and Y. Suh, "An Efficient Bandwidth Request Mechanism for Non-Real-Time Services in IEEE 802 . 16 Systems," Proc. 2nd Intl. Conf. Commun. Systems Software and Middleware, Bangalore, 2007, pp. 1-9.

[12] Q. Ni, L. Hu, A. Vinel, Y. Xiao and M, Hadjinicolaou, "Performance Analysis of Contention Based Bandwidth Request Mechanisms in WiMAX Networks," IEEE Syst. J. vol. 4, no. 4, pp. 477–486, 2010.

[13] A. Asadi and T. S. Wei, "An Enhanced Cross Layer Downlink Scheduling Algorithm for IEEE 802 . 16 Networks," Proc. Intl. Conf. Information Networking, Barcelona, 2011, pp. 212–217.

[14] A. Z. Al-Howaide, A. S. Doulat, and Y. M. Khamayseh, "Performance evaluation of different scheduling algorithms in WiMAX,"International Journal of Computer Science, Engineering and Applications, vol. 1, no. 5, pp. 81-94, 2011.

[15] M. S. Arhaif, "Comparative Study of Scheduling Algorithms in WiMAX," International Journal of Scientific and Engineering Research, vol. 2, no. 2, pp. 1–7, 2011.

[16] Y. Wang, S. Chan, M. Zukerman, and R. J. Harris, "Priority-Based fair Scheduling for Multimedia WiMAX Uplink Traffic," 2008 IEEE Intl. Conf. Commun., Beijing, 2008, pp. 301–305.

[17] P. Ting, C. Yu, N. Chilamkurti, W. Tung-hsien, and C. Shieh, "A Proposed RED-based Scheduling Scheme for QoS in WiMAX Networks,"4th Intl. Symp. Wireless Pervasive Computing,Melbourne, 2009, pp. 1–5.

[18] R. F. Sari, I. D. Gde, N. Mukhayaroh, and D. Laksmiati, "Performance Evaluation of Weighted Round Robin based Scheduler over Wimax."Proc. Quality in Research Conf., Jakarta, 2007, pp. 1-5.

[19] J. Lakkakorpi, A. Sayenko, and J. Moilanen, “Comparison of Different Scheduling Algorithms for WiMAX Base Station,” Proc. IEEE Wireless Commun. and Networking Conf. Las Vegas, 2008, pp. 1991–1996.

[20] N. Ruangchaijatupont, L. Wangt, and Y. Jit, “Study on the Performance of Scheduling Schemes for Broadband Wireless Access Networks,” Proc. Intl Symp. Commun. and Information Technology, Bangkok, 2006, pp. 1008–1012.

[21] M. Oktay and H. A. Mantar, “A Distance-Aware Scheduler for Real-Time Applications in 802 . 16 Wimax Networks,” in 1st Intl. Conf. Networked Digital Technologies, Ostrava, 2009, pp. 435–440.

[22] N. Xiong, H. Wang, and W. Nie, “Low-Overhead Uplink Scheduling through Load Prediction for WiMAX Real-Time Services,” IET Commun., vol. 5, no. 8, pp. 1060–1067, 2011.

[23] N. Lal, A. P. Singh, S. Kumar, S. Mittal, and M. Singh, “A Heuristic EDF Uplink Scheduler for Real Time Application in WiMAX Communication,” in ArXiv preprint, arXiv: 1501.04553, 2015.

[24] I. Khirwar, A. Yadav, and P. Trivedi, “Comparative Assessment of WiMAX Scheduler in Fixed and Mobile WiMAX Networks for VoIP using QualNet,” in Intl. Conf. Comput. Commun. Technology, Allahabad, 2010, pp. 15–21.

[25] S. Y. You, A. Sali, N. K. Noordin, and C. K. Ng, “An Enhanced Weighted Fair Queuing Scheduling for WIMAX Using QualNet,” in4th Int. Conf. Comput. Intell. Model. Simul., Kuantan, 2012, pp. 387–392.

[26] G. B. Satrya, I. K. G. W. Jaya, and N. D. W. Cahyani, “The Analysis and Implementation of Aging Method on Packet Scheduling Algorithm in WIMAX Network,” in IEEEAsia Pacific Conf. Wireless and Mobile, Bali, 2014, pp. 340–345.

[27] H. Guo, Z. Zhang, J. Guo, and J. Zhang, “A Distributed Cross-layer Adaptive Scheduling Algorithm for Wireless Mesh Networks,” in Intl. Conf. Control, Automation and Syst. Engineering, Singapore, 2011, pp. 1-5.

[28] W. Gan, J. Xian, X. Xie, and J. Ran, “A cross-layer designed scheduling algorithm for WiMAX uplink,” in 9th Intl. Conf. Electron. Meas. Instruments, Beijing, 2009, pp. 122-127.

[29] Q. Liu, X. Wang, and G. B. Giannakis, “A Cross-Layer Scheduling Algorithm With QoS Support in Wireless Networks,” IEEE Trans. Veh. Technol. vol. 55, no. 3, pp. 839–847, 2006.

[30] S. Kulkarni, D. Shwetha, J. T. Devaraju, and D. Das, “Channel Aware Uplink Scheduler for a Mobile Subscriber Station of IEEE 802 . 16e,” Intl. Journal of Computer Applications, vol. 35, no. 6, pp. 15–22, 2011.

[31] J. Chen, W. Jiao, and H. Wang, “A Service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode,” in Intl. Conf. Commun. 2005, pp. 3422–3426.

[32] J. Sun, Y. Yao, and H. Zhu, “Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems,” in IEEE 63rd Veh. Technol. Conf., Melbourne, 2006, pp. 1221–1225.

[33] E. R. Dosciatti and W. G. Junior, “Scheduling Mechanisms with Call Admission Control (CAC) and an Approach with Guaranteed Maximum Delay for Fixed WiMAX Networks,” in Quality of Service and Resource Allocation in WiMAX, R. Hincapie, Ed. Rijeka: InTech, 2004, pp 59-84.

[34] A. Lera, A. Molinaro, and S. Pizzi, “Channel-Aware Scheduling for QoS and Fairness Provisioning in IEEE 802.16/WiMAX Broadband Wireless Access Systems,” IEEE Network, vol. 21, no. 5, pp. 34–41, 2007.

[35] J. Thaliath, M. M. Joy, E. P. John, and D. Das, “Service Class Downlink Scheduling in WiMAX,” in 3rd Intl. Conf. Commun. Syst. Software and Middleware and Workshops, Bangalore, 2008, pp. 196-199.

[36] J. F. Borin and N. L. S. Fonseca, “Uplink Scheduler and Admission Control for the IEEE 802 . 16 standard,” in IEEE Global Telecommunications conf., Honolulu, 2009, pp. 1–6.

[37] D. Pradishta, B. Jothimohan, and A. Ponraj, “Dynamic QoS-Based Optimized Video Transmission in WiMAX Networks,” in Intl. Conf. Commun and Signal Processing. Melmaruvathur, 2014, pp. 1209–1213.

[38] K. Mnif, R. Khdhir, and L. Kamoun, “Evaluation and Comparison of Scheduling Algorithms in Wimax Networks,” in Intl. Conf. Multimedia Computing and Syst. Marrakech, 2014, pp. 884-889.

[39] V. Richter, R. Radeke, and R. Lehnert, “QoS Concept for IEEE 802.16-2012 Based WiMAX Networks,” in IEEE 10th Intl. Conf. Wireless and Mobile Computing, Networking and Commun. Larnaca, 2014, pp. 371–377.

[40] N. M. El-shennawy, M. A. Youssef, M. N. El-Derini, and M. M. Fahmy, “A Dynamic Uplink Scheduling Scheme for WiMAX Networks,” in 8th Intl. Conf. Computer Engineering & Syst., Cairo, 2013, pp. 111–115.

[41] S. Sharma, N. S. Randhawa, and D. Kaur, “An Efficient Hybrid Downlink BS Scheduler for WiMAX Netowrk,” in Intl. Conf. Pervasive Computing, Pune, 2015, pp. 1-6.

[42] N. A. Ali, P. Dhrona, and H. Hassanein, “A Performance Study of Uplink Scheduling Algorithms in Point-to-Multipoint WiMAX Networks,” Comput. Commun., vol. 32, no. 3, pp. 511–521, 2009.

[43] R. Pries, D. Staehle, and D. Marsico, “Performance Evaluation of Piggyback Requests in IEEE 802.16,” inIEEE 66th Veh. Technol. Conf., Baltimore, 2007, pp. 1892–1896.

[44] Q. Ni, A. Vinel, Y. Xiao, A. Turlikov, and T. Jiang, “Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMAX Networks,” IEEE Commun. Magazine, vol. 45, no. 5, pp. 132-138, 2007.

[45] D. Chuck, K.-Y. C. K.-Y. Chen, and J. M. Chang, “A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks,” IEEE Trans. Veh. Technol., vol. 59, no. 4, pp. 2046-2056, 2010.

[46] N. Lee, Y. Choi, S. Lee, and N. Kim, “A New CDMA-Based Bandwidth Request Method for,”IEEE Commun. Letters, vol. 14, no. 2, pp. 124–126, 2010.

[47] S. Z. Tao and A. Gani, “Intelligent Uplink Bandwidth Allocation Based on PMP Mode for WiMAX,” in Intl. Conf. Comput. Technol. Dev., Kota Kinabalu, 2009, pp. 86–90.

[48] K. Wee, W. K. New, Y. Y. Wee, and C. Wong, “Intensive Bandwidth Request and Handling Design in PMP,” IJCSNS, vol. 12, no. 2, pp. 27–32, 2012.

[49] C.-Y. Liu and Y.-C. Chen, “An Adaptive Bandwidth Request Scheme for QoS Support in WiMAX Polling Services,” in 28th Int. Conf. Distrib. Comput. Syst. Work., Beijing, 2008, pp. 60–65.

[50] E. Park, “Efficient Uplink Bandwidth Request with Delay Regulation for Real-Time Service in Mobile WiMAX Networks,” IEEE Trans. Mob. Comput., vol. 8, no. 9, pp. 1235–1249, 2009.

[51] C. Chiang, W. Liao, and T. Liu, “Adaptive Downlink / Uplink Bandwidth Allocation in IEEE 802 . 16 (WiMAX) Wireless Networks : A Cross-Layer Approach,” in IEEE Global Telecommunications Conf. Washington DC, 2007, vol. 16, pp. 4775–4779.

[52] K. K. Wee and S. W. Lee, “Priority based bandwidth allocation scheme for WIMAX systems,” in 2nd IEEE Int. Conf. Broadband Netw. Multimed. Technol., Beijing, 2009, pp. 15–18.

[53] H. Chen, K. Shih, S. Wang, and C. Chiang, “An Efficient Downlink Bandwidth Allocation Scheme for Improving Subchannel Utilization in IEEE 802.16e WiMAX Networks,” in IEEE 71st Veh. Technol. Conf. Taipei, 2010, pp. 1-5.

[54] T.-L. Sheu and K.-C. Huang, “Adaptive bandwidth allocation model for multiple traffic classes in IEEE 802.16 worldwide interoperability for microwave access networks,” IET Commun., vol. 5, no. 1, pp. 90–98, 2011.

[55] A. Esmailpour and N. Nasser, “Dynamic QoS-Based Bandwidth Allocation Framework for Broadband Wireless Networks,” IEEE Trans. Veh. Technol., vol. 60, no. 6, pp. 2690–2700, 2011.

[56] L. Chen, Q. Guo, J. Wang, and G. Liu, “A Multi-Round Resources Allocation Scheme for OFDMA-Based WiMAX Based on Multiple Service Classes,” in Intl. Conf. Information Science, Electronics, Electrical Engineering, Sapporo, 2014, pp. 260-264.

[57] B. Rong, Y. Qian, K. Lu, H. Chen, and M. Guizani, “Call Admission Control Optimization in WiMAX Networks,” IEEE Trans. Veh. Technol.,vol. 57, no. 4, pp. 2509–2522, 2008.

[58] W. Lilei and X. Huimin, “A New Management Strategy of Service Flow in IEEE 802.16 systems,” in 3rd IEEE Conf. Industrial Electronics Applications,Singapore, 2008, pp. 1716–1719.

[59] S. Chandra and A. Sahoo, “An Efficient Call Admission Control for IEEE 802.16 Networks,” in 15th IEEE Works. Local & Metropolitant Area Networks, Princeton, 2007, pp. 188-193.

[60] J. Lakkakorpi and A. Sayenko, “Measurement-Based Connection Admission Control Methods for Real-Time Services in IEEE 802.16e,” in 2ndIntl. Conf. Commun. Theory, Reliab. Qual. Serv., Colmar, 2009, pp. 37–41.