International Journal of the Physical Sciences Vol. 6(14), pp. 3487-3501, 18 July, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.539
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Modeling and performance analysis of uplink scheduling algorithm in mobile WiMAX systems

D. Mohd Ali[1,2]*, K. A. Noordin[1], K. Dimyati[3] and A. Idris[1, 2]

[1]Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[2]Faculty of Electrical Engineering, Mara University of Technology, UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia.
[3]Electrical and Electronic Engineering Department, Faculty of Engineering, National Defense University of Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia.

Accepted 1 June, 2011

Extensive numbers of the IEEE 802.16e Mobile WiMAX scheduling proposals focus on simulation for obtaining performance measures of the proposed scheduling algorithm. There is also handful of article that focused on the mathematical modeling. However, the model is derived assuming the traffic from different quality of service (QoS) classes is separated, which allow for independent analysis. In this paper, we propose an uplink (UL) scheduling algorithm that considers the adaptive modulation and coding scheme and QoS parameters of all five service classes. The resources are distributed in two stages. The first stage allocates the resources to different classes of service in accordance to the threshold based priority while the second stage allocates the resources within the same class with the exhaustive service strategy. A mathematical model that combine the different QoS classes of Mobile WiMAX is formulated and it is derived based on the weighted sum of the mean waiting time for the individual queues or known as pseudo-conservation law.

Key words: Uplink scheduling, mobile WiMAX (IEEE 802.16e), threshold based cyclic polling (TbCP) algorithm, pseudo-conservation law, QoS.

INTRODUCTION

Mobile WiMAX is one of the promising alternatives to the cable modem and Digital Subscriber Line in providing last mile broadband access at a cheaper cost. The mobile WiMAX network has the capability of delivering different types of traffic with the support from the resilient medium access control (MAC) as delineated by the standard. In order to achieve the various QoS constraint, the MAC layer offers the differentiation through five defined service classes: 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). The standard provides details of QoS parameters; however the algorithm implementation was left undefined. Thus, a suitable UL scheduling algorithm is required to schedule and allocate resources effectively to achieve satisfactory outcome in conformity to the QoS parameters. For instance, delay is the most critical to the quality of real-time traffic and should be properly satisfied. We indicate that the UL scheduling has become necessary for the multimedia services, interactive video applications and online gaming which are becoming more prevalent recently.

In this paper, we present an UL scheduling based on the cyclic poling model, where a threshold-based priority and an exhaustive scheme are proposed as the first and second stage of the algorithm respectively. The proposed work is different from the work in Zsolt and Miklos (2010), where the analysis of a globally gated Markovian limited cyclic polling model is introduced for the nrtPS. The mathematical model is applied to the subscriber station (SS) of IEEE 802.16. An exhaustive time-limited polling model with vacation is proposed in Haitham et al. (2008). A queueing model is presented for the polling service traffic (rtPS and nrtPS in general are labelled the polling traffic) which takes into account the polling periods of the SS stations which is served by a single base station (BS). In Howon et al. (2004), the authors proposed a scheduling

*Corresponding author. E-mail: darma504@yahoo.com.my Tel: +603 – 79675328. Fax: +603 – 79675316.

algorithm for the VoIP (voice over internet protocol) services. The system model is represented as a one dimensional Markov chain with an on-off model for the voice traffic. A novel scheduling scheme was presented in Fen et al. (2007) to provide QoS satisfaction and service differentiation in terms of delay. The time window, T_i of the Proportional Fairness is manipulated to distinguish the service and QoS assurance of each queue. The mean queue length and mean waiting time of the system is derived using the M/M/1 Markov model. Jae (2008) formulated the mathematical model taking into account the modulation and coding scheme as well as the signaling overhead in the downlink (DL). The VoIP packets are scheduled based on the First-In-First-Out policy and the traffic is modeled as an exponentially distributed on-off model. The aggregate VoIP traffic from the N SSs are represented as the two-state Markov-modulated Poisson process. The analysis is done independently for the ertPS, UGS and rtPS.

Most of the works involving the mathematical modeling have assumed that only one service class is employed at one time which is not always true in the actual condition. Mobile WiMAX networks are designed to allow for the coexistence of various types of traffic, ranging from real-time traffics such as VoIP and video conferencing to non real-time traffic such as file transfer and e-mail service through the various service classes. These would cause certain applications not to function properly. For instance, delay is the most critical issue for the real-time traffic and will not be suitable for service class without the QoS parameter considering that the performance has the tendency to deteriorate. Thus, this paper will attempt to combine the different service classes of mobile WiMAX in formulating the mathematical model. We take into account the correlation between the service classes by introducing the switch-over time. Owing to the extreme complexity of the analysis which is due to the amount of asymmetric characteristics of the service classes, an approximation model referred as pseudo-conservation law is used. In the case of symmetric queues, the law gives exact expression of the mean waiting times. On the contrary, an accurate approximation model can be obtained for the case of asymmetric queues.

The rest of the work is organized as follows. Subsequently, the Physical and MAC layers of the IEEE 802.16e are briefly given, after which the proposed algorithm with detailed explanation and mathematical derivation is presented. This is followed by a discussion of the performance analysis. Finally, the conclusion was given, as well as a determination of our future directions in this particular area of study.

PHYSICAL LAYER AND MAC LAYER OF THE MOBILE WIMAX

The accessing technique for the IEEE 802.16e has been specified as the orthogonal frequency division multiple access (OFDMA). OFDMA subcarriers are divided into subset of subcarriers and each subset represents one subchannel. The number of subcarriers changes with the channel bandwidth in order to provide adaptive frequency bandwidth and data rate (Nuaymi and Loutfi, 2007). The IEEE 802.16e also supports time division duplexing (TDD) as the duplexing technique. A TDD frame has fixed duration and contains one DL and UL subframes. Figure 1 illustrates the two-dimensional mapping between the frequency and time domains. A slot is the minimum unit of frequency-time resource with which a user may be granted, which contains 48 data sub-carriers and the amount of data carried varies with different modulations and coding schemes.

The MAC protocol is connection-oriented. Upon entering the network, each mobile station (MS) creates one or more connections over which its data are transmitted to and from the BS. The MAC layer schedules the usage of the air link resources and provides QoS differentiation for the different types of applications through the five defined service classes. The BS is responsible for coordinating the communication in the network since there is no direct communication between the MS in the Point-to-Multipoint mode. The transmission from the MS to BS is called UL whereas DL is from BS to MS.

The mobile WiMAX architecture

The architecture of mobile WiMAX consists of a number of MSs and a BS. The BS manages the communication between the MSs in the network which take place in two directions, UL and DL. The MS that attempts to communicate with the BS will be identified with a unique connection identifier (CID). Due to the centralized architecture, the BS scheduler controls the system parameters. The transmission in DL is uncomplicated since the BS is the only one that transmits and schedules all the connections It should be noted that for an. UL transmission, packets are queued at the MSs and specifically the UL scheduler works on a request-grant basis. For this purpose, each MS will send a bandwidth request message to the BS. Several bandwidth requests mechanisms are available such as unicast polls, broad-cast and multicast polls, contention and piggybacking. Subsequently, after acquiring the bandwidth request messages, the messages are then classified according to the service class and QoS parameter in the scheduler for the scheduling process. An information element (IE) is created in the UL-MAP to indicate the control region and new resource assignments that MSs should transmit (Sze et al., 2009). The UL-MAP is placed at the beginning of the DL subframe of each frame and the completed MAP is then broadcast to all MSs in the network. Each MS listens to the broadcasted MAP message for their CID and

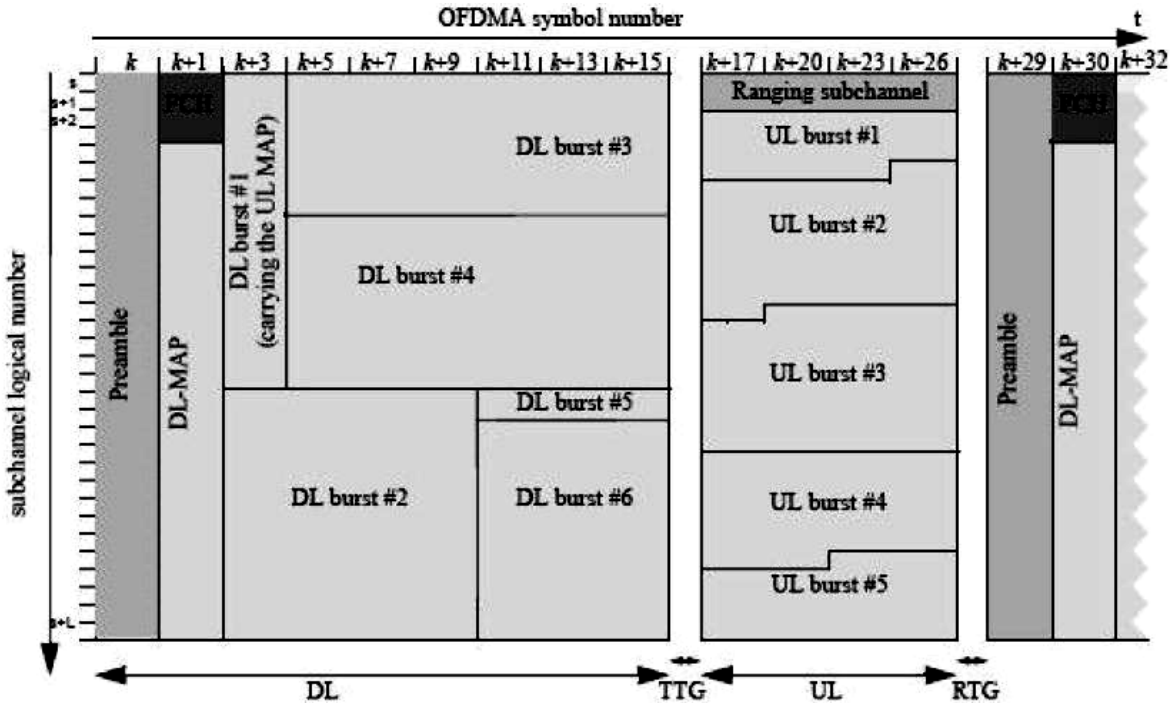


Figure 1. The OFDMA frame structure (Nuaymi and Loutfi, 2007).

decodes the UL-MAP IE so that packets are transmitted in accordance to the allocated slots.

After the initialization process takes place, the MS will send a Dynamic Service Addition – Request message to demand for an UL request opportunity. The Dynamic Service Addition – Request contains the service flow (SF) parameter for the request connection and MAC address. The BS that receives the message will call the admission control to decide whether to accept or reject the request based on the available capacity. If there are several SF arriving simultaneously, the SF is handled according to the priority of each class. Then the available bandwidth is updated accordingly, following the acceptance of the SF.

The available bandwidth can be calculated as:

$$C_{avail} = C_{total} - \sum_{i=0}^{I} \sum_{j=0}^{J_i} SF \quad (1)$$

where C_{total} represents the total bandwidth. I, represents the total number of SF or connections and J_i is the total number of service classes of the i_{th} connection. Assume that each MS carries a single SF which eliminates the effect of packet scheduling at MSs. Each SF is fixed to maximum sustained traffic rate (MSTR) for UGS and ertPS. As for rtPS and nrtPS, it is set to a minimum reserved traffic rate (MRTR), and BE is given the available capacity after all the service classes are considered. A Dynamic Service Addition – Response will be sent from BS to inform the acceptance of the connection and a CID will then be assigned to the MS.

Five service classes are supported in the IEEE 802.16e. The UGS is appropriate for the applications that generate constant bit rate traffic hence it requires a fixed bandwidth at periodic intervals. The BS pre-allocates periodic slots to such MS prior to the QoS agreement (observing the MSTR) negotiated at the connection setup. The ertPS is formed through the combination of UGS and rtPS. It suits the real-time application with varying bit rates and delay requirements. In this case, the BS will provide unicast grants in an unsolicited manner to the MSs with the ertPS flows. Applications such as audio and video streaming are classified under the rtPS service class. These types of applications generate real-time packets stream which are variable in size. Due to that, the BS will provide unicast request opportunities in which the MS will stipulate the size of the grant required. The nrtPS is designed for a delayed tolerant variable size applications such as file transfer. The MS with such flow will be polled by the BS on a regular basis. Finally, BE is suitable for the application with no throughput or delay guarantees. MSs with the BE connections have to contend for their bandwidth allocation.

THRESHOLD BASED CYCLIC POLLING (TBCP) ALGORITHM

In order to initiate a graceful trade-off between the real-time and non-real time traffic, a threshold based priority algorithm is presented. This is to ensure that the delay property of the real-time traffic (UGS, ertPS and rtPS) is

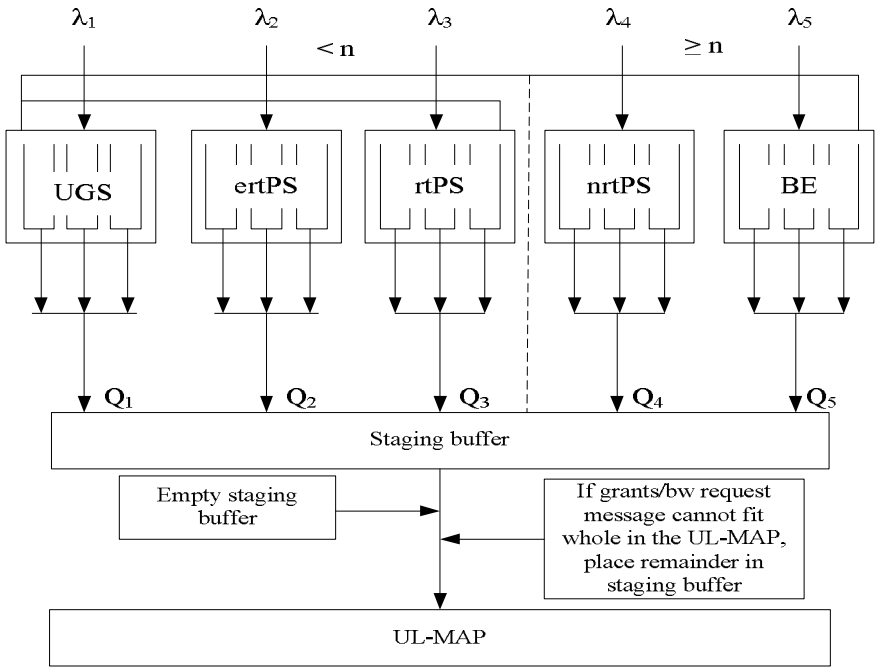


Figure 2. The architecture of the TbCP algorithm.

properly satisfied without sacrificing the throughput of the non-real time traffic (lower priority class –BE). A priority queueng (PQ) and deficit fair priority queuing (DFPQ) (Po et al., 2007) algorithms seems capable of guaranteeing the QoS parameters of the different service classes. However, the direct outcome of these concept is low throughput being delivered from the BE service class. Thus, to mitigate with this issue, a threshold-based priority is proposed. Our approach aims at adjusting the threshold value, n which represents the number of bandwidth request messages in the nrtPS queue. We performed the simulation rigorously while varying the threshold value in order to determine the suitable one. Through this, we observed that the optimal value of the threshold is equal to 10 and the smallest is 1. We have discovered the importance of the threshold value and its undeniable influence towards the configuration of the service classes. When the network is configured with a higher concentration of latency-guaranteed QoS classes (UGS, ertPS and rtPS), then the threshold value should be set higher which is 10. In contrast, lower threshold value should be used when the network is configured with higher concentration of non-real time service classes (nrtPS and BE).

The scheduling scheme starts with the scheduler visits to the UGS queue (If the UGS and ertPS service classes are unoccupied, the scheduler will start the allocation with the rtPS). The UGS queue is served until no more grants are available. Once the queue is empty, the scheduler will move to the ertPS queue which is also attended until the queue is empty. The rtPS queue is served next and attended until the queue is empty. Before continuing the service to nrtPS, the scheduler will check on the amount of bandwidth requests available at the nrtPS queue. If the amount of bandwidth request message exceeds the threshold assigned, n, then the service will be carried out to nrtPS and subsequently the BE. On the other hand, the scheduler will return to serve UGS (rtPS if UGS and ertPS are unoccupied) if the amount of the bandwidth request is less than the threshold assigned.

Immediately after each queue is served, the grants and bandwidth request messages are then sent to the staging buffer area for the mapping of IE to the UL-MAP process (to decide on the time slots that should be allocated to each MS). However, the mapping of IE to the UL-MAP and DL-MAP to the OFDMA frame is not included in the scope of this work. The staging buffer is emptied as needed (if there are spaces in the UL-MAP) to fill up the empty slots in the UL-MAP. The UL scheduler explained above is depicted in Figure 2. Fundamentally,the network operators in the real environment do not provide all thefive service classes simultaneously (as define theoretically by the standard) to the end users so as not to burden the scheduler. Thus, to reduce the complexity (Chakchai et al., 2009), mainly two or the maximum of three QoS classes are assigned. The UGS and ertPS classes are designed for the VoIP traffic while the nrtPSand BE both represent the non-real time traffic, thus choosing either one to represent the traffic is considerably sufficient. Therefore, we have chosen the UGS, rtPS and nrtPS for Scenario 1 while the rtPS, nrtPS and BE are selected for Scenario 2. The architecture of

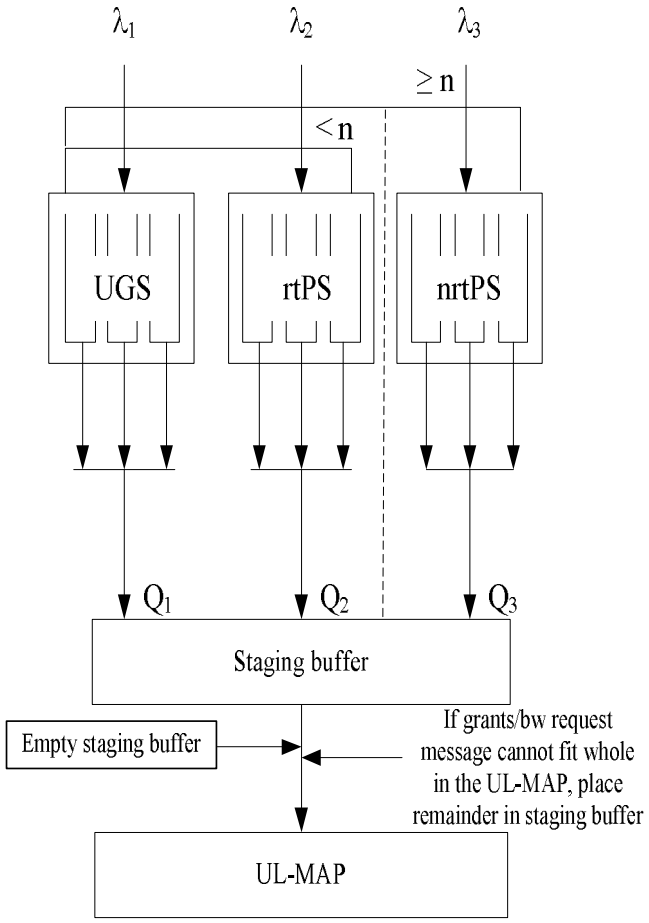


Figure 3. The architecture for Scenario 1.

the UL scheduler for Scenarios 1 and 2 are depicted in Figures 3 and 4 respectively.

Mathematical modeling

System model

In this study, we will derive the weighted sum of the Scenario 1. The same approach can be applied for Scenario 2. As mentioned earlier in mobile WiMAX architecture, the UL scheduler schedules the grants/bandwidth request messages in accordance to their service classes and not the individual packet. However, the lower the mean waiting time of the grants/bandwidth request messages, the packets which reside in the MS can be transmitted faster.

The scheduler serves three service classes (queues) in Scenario 1. Q_1, Q_2 and Q_3 are represented as UGS, rtPS and nrtPS. The grants and bandwidth request messages arrive at all queues according to the independent Poisson processes with an arrival rate of $\lambda_1, \lambda_2, \lambda_3$. The compound arrival rate is given as follows:

$$\Lambda = \sum_{i=1}^{3} \lambda_i \quad (2)$$

Though the arrival of UGS and ertPS grants are modeled as a constant bit rate, however, in the case of multiple session arrival, the Poisson process provides adequate accuracy (Fen et al., 2006). Similarly, let b_i be the service time of Q_i with the first two moments:

$$\beta_i = E[b_i] \text{ and } \beta_i^{(2)} = E[b_i^{(2)}] \quad (3)$$

The server switches from Q_i to the next queue with non-zero walk time with first moment, s_i and the second moment $s_i^{(2)}$. The first moment of total walk time, s is given as:

$$s = \sum_{i=1}^{3} s_i \quad (4)$$

The switch over process is independent, the arrival and service time is assumed to be independent, and identically distributed (IID). The offered traffic at ith queue is, $\rho_i = \lambda_i \beta_i$ and the total offered traffic is:

$$\rho = \sum_{i=1}^{3} \rho_i \quad (5)$$

The order of service within each queue is first-come-first-serve. The queues are attended by a scheduler in cyclic order and incurs non-zero switch over time. Each queue will be served under exhaustive service discipline in which the scheduler continues to work until the queue becomes empty. When Q_i is empty, scheduler immediately begins to switch to Q_{i+1}.

Pseudo-conservation law

In the situation of zero switch-over times, the server works when there is work in the system and becomes idle if no work is present. Thus, the conservation law holds for the total amount of work in the system. The amount of in the system is independent of the service policy work and hence, equals to the amount of work in an M/G/1 queue (Hanoch and Moshe, 1990; Onno and Wim, 1987; Onno and Wim, 1988; Yoshitaka and Krishna, 1995) with the arrival rate of λ_i and service time of β_i

$$E[V_{M/G/1}] = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)}$$

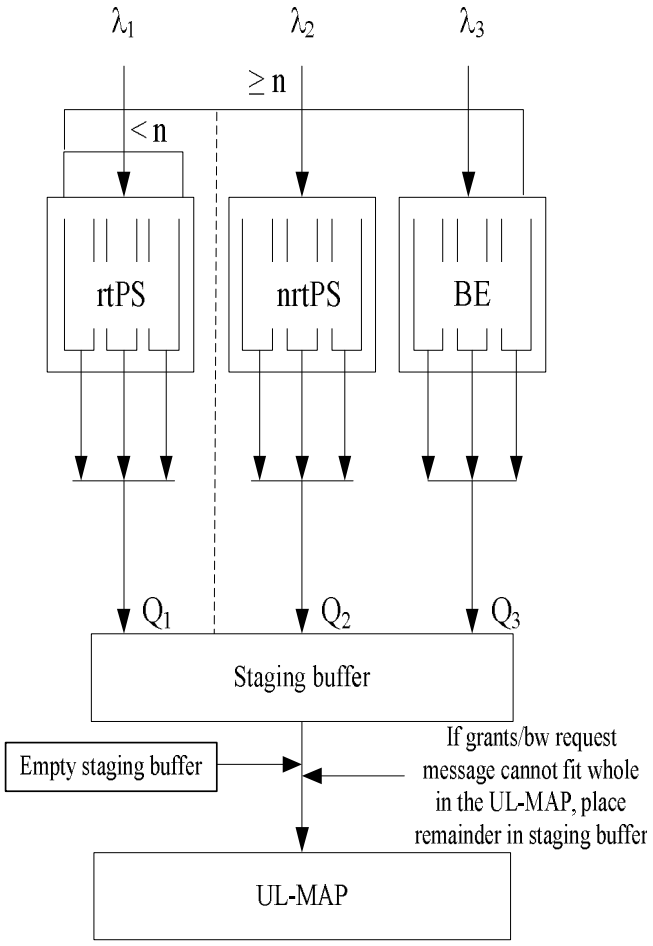


Figure 4. The architecture for Scenario 2.

$$\sum_{i=1}^{N} \rho_i EW_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} \quad (6)$$

On the contrary, in the case of non-zero switch-over times, the amount of work (the weighted sum of the waiting times is $\rho_i EW_i$. Let W_i be the waiting time and define as time between the arrival of grants/ bandwidth request messages at Q_i and the moment at which it starts to receive service with mean EW_i) is no longer independent of the service policy. This is called the pseudo-conservation law and established as follows (Hanoch and Moshe, 1990; Onno and Wim, 1987; Onno and Wim, 1988; Yoshitaka and Krishna, 1995):

$$V_c \underline{d} V_{M/G/1} + Y$$

where : $\underline{d}$ = equality in distribution. V_c = amount of work in a cyclic system at an arbitrary epoch, $V_{M/G/1}$ = amount of work in corresponding M/G/1 system at an arbitrary epoch, Y = amount of work in a cyclic-service system at an arbitrary epoch in a switching period/non-serving interval, $V_{M/G/1}$ and Y are independent.

Here, epoch designates an instant in time. It follows that:

$$E[V_c] = E[V_{M/G/1}] + E[Y] \quad (7)$$

Moreover, $E[Y]$ can be defined as:

$$E[Y] = \sum_{i=1}^{N} \frac{s_i}{s} E[Y_i] \quad (8)$$

From Onno and Wim (1987), $E[Y]$ is composed of three terms: Where: $EM_i^{(1)}$ = mean amount of work in Q_i at a departure epoch of the server from Q_i. $EM_i^{(2)}$ = mean amount of work in the rest of the system at a departure epoch of server from Q_i. $\rho \frac{s_i^{(2)}}{2s_i}$ = mean amount of work arrived in the system in the preceding part of the switching interval under consideration.

$$E[Y_i] = \sum_{i=1}^{N} EM_i^{(1)} + \sum_{i=1}^{N} EM_i^{(2)} + \sum_{i=1}^{N} \rho \frac{s_i^{(2)}}{2s_i} \quad (9)$$

$EM_i^{(1)}$ is totally dependent on the choice of service strategies applied to the queue and can only be determined when the service strategy is identified. Therefore, parameter $EM_i^{(2)}$ will be investigated later in the study that is, the mean amount of work in the rest of the system at a departure epoch of server from Q_i when the requests in nrtPS are less than n and greater than or equals to n respectively. Derivation of $EM_i^{(2)}$ when requests in nrtPS is less than n: As discussed earlier, when the bandwidth request in Q_3 is less than n, the scheduler will only carry out the service for Q_1 and Q_2 cyclically with an exhaustive service policy. In this case, Q_3 will not be attended by the scheduler as if the queues do not exist in the scheduler. Thus, a cycle in the condition when requests in nrtPS is less than n is consists of a scheduler visits to Q_1 and Q_2. The total switch over time and total offered traffic are defined as $s = \sum_{i=1}^{2} s_i$ and $\rho = \sum_{i=1}^{2} \rho_i$ respectively. The mean visit time of Q_i is given as:

$$E[V_i] = \rho_i \frac{s}{1-\rho} \quad (10)$$

that is Q_i is visited once in a cycle. Thus, the total amount of work at the departure epoch of server from Q_1 and Q_2 can be derived as follows:

$$\frac{s_1}{s} EM_1^{(2)} + \frac{s_2}{s} EM_2^{(2)} \quad (11)$$

where a detailed derivation is given as (Darmawaty and Kaharudin, 2010):

$$EM_i^{(2)} = \rho_{i-1}(s_{i-1} + \rho_i \frac{s}{1-\rho}) + \rho_{i-2}(s_{i-2} + \rho_{i-1}\frac{s}{1-\rho} + s_{i-1} + \rho_i \frac{s}{1-\rho}) + \ldots + \rho_{i+1}(s_{i+1} + \rho_{i+2}\frac{s}{1-\rho} + \rho_{i+3}\frac{s}{1-\rho} + \ldots + s_{i-1} + \rho_i \frac{s}{1-\rho}) \quad (12)$$

Solving Equation (12) and expanding Equation (11) which can then be simplified into:

$$\sum_{i=1}^{2} \frac{s_i}{s} EM_i^{(2)} = \frac{\rho}{s} \sum_h \sum_{<k} s_h s_k + \frac{s}{1-\rho} \sum_h \sum_{<k} \rho_h \rho_k \quad (13)$$

Derivation of $EM_i^{(2)}$ when request in nrtPS is greater than or equals to n: When the bandwidth request in Q_3 is greater than or equals to n, the scheduler will visit Q_3 after the completion of Q_2. Q_3 will be serviced according to the exhaustive service strategy. Thus, a cycle in the condition when requests in nrtPS is greater than or equals to n is consists of a server visits to Q_1, Q_2 and Q_3 in a cyclic order. The total switch over time and total offered traffic in the condition of requests in nrtPS is greater than or equals to n are $s = \sum_{i=1}^{3} s_i$ and $\rho = \sum_{i=1}^{3} \rho_i$ respectively. The amount of work in Q_3 is denoted as (Darmawaty and Kaharudin, 2010):

$$\rho_3 \frac{s}{1-\rho}$$

Thus the total amount of work at the departure epoch of server from Q_1, Q_2 and Q_3 can be obtained from:

$$\frac{s_1}{s} EM_1^{(2)} + \frac{s_2}{s} EM_2^{(2)} + \frac{s_3}{s} EM_3^{(2)} \quad (14)$$

By noting that $EM_i^{(2)}$ is given by Equation (12) and solving Equation (14) which then can be simplified to:

$$\sum_{i=1}^{3} \frac{s_i}{s} EM_i^{(2)} = \frac{\rho}{s} \sum_h \sum_{<k} s_h s_k + \frac{s}{1-\rho} \sum_h \sum_{<k} \rho_h \rho_k \quad (15)$$

Derivation of pseudo-conservation law for the uplink scheduler

With reference to Equation (7), $E[V_c]$ is defined as (Onno and Wim, 1988; Borst SC, 2006):

$$E[V_c] = \sum_{i=1}^{N} \rho_i EW_i + \frac{1}{2} \sum_{i=1}^{N} \lambda_i \beta_i^{(2)} \quad (16)$$

Solving Equation (7) from Equation (6) and Equation (16):

$$\sum_{i=1}^{N} \rho_i EW_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + E[Y] \quad (17)$$

From Equation (8), $E[Y]$:

$$E[Y] = \sum_{i=1}^{N} \frac{s_i}{s} E[Y_i]$$

which is consequently composed of:

$$E[Y_i] = \sum_{i=1}^{3} EM_i^{(1)} + \sum_{i=1}^{2} EM_i^{(2)} + \sum_{i=1}^{3} EM_i^{(2)} + \sum_{i=1}^{3} \rho \frac{s_i^{(2)}}{2s_i} \quad (18)$$

where $EM_i^{(2)}$ is derived on the condition that the amount of bandwidth requests in nrtPS are less than n and greater than or equals to n. EM_j^i depends on the service strategy applied to the queues. For exhaustive service, $EM_i^{(1)} = 0$ since there are no bandwidth requests left behind the server leaving the queue. Thus, E[Y] is defined as follows:

$$E[Y] = \sum_{i=1}^{2} \frac{s_i}{2(1-\rho_i)} \left[\left(\sum_{i=1}^{2} \rho \right)^2 - \sum_{i=1}^{2} \rho_i^2 \right] + \sum_{i=1}^{3} \frac{s_i}{2(1-\rho_i)} \left[\left(\sum_{i=1}^{3} \rho \right)^2 - \sum_{i=1}^{3} \rho_i^2 \right] + \sum_{i=1}^{3} \rho \frac{s_i^{(2)}}{2s_i} \quad (19)$$

By inserting the Equation (19) into Equation (17), finally, the weighted sum of the mean waiting times for the individual queues or pseudo-conservation law can be written as follows:

$$\sum_{i=1}^{3} \rho_i EW_i = \rho \frac{\sum_{i=1}^{3} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \sum_{i=1}^{2} \frac{s_i}{2(1-\rho_i)} \left[\left(\sum_{i=1}^{2} \rho \right)^2 - \sum_{i=1}^{2} \rho_i^2 \right] + \sum_{i=1}^{3} \frac{s_i}{2(1-\rho_i)} \left[\left(\sum_{i=1}^{3} \rho \right)^2 - \sum_{i=1}^{3} \rho_i^2 \right] + \sum_{i=1}^{3} \rho \frac{s_i^{(2)}}{2s_i} \quad (20)$$

SIMULATION ENVIRONMENT

The proposed algorithm is implemented and evaluated using the OPNET simulator to show the correctness and the performance of the algorithm, respectively. Threshold value of 1 (th 1) and 10 (th 10) is selected for the TbCP algorithm.The IEEE 802.16e standard enables the optimization of each MS data rate by allowing the BS to set the modulation and coding scheme with regards to the channel condition (Jae, 2008). Thus, to account for the adaptive modulation and coding scheme, each SF is translated in accordance to the coding rate and bits per symbol carried for each modulation. Table 1 summarised the coding rates and bits per symbol supported.

A standard admission control that distributes bandwidth among the service classes is employed as discussed in The mobile WiMAX architecture.

We simulate here two scenarios to demonstrate the efficiency of the proposed algorithm. The simulation parameters and traffic parameters are summarised in Tables 2 and 3 respectively.

SIMULATION RESULTS

Results obtained through series of simulations are presented in this study. We define the QoS metric that we are looking for from the various traffic which provide the best indications from the QoS perspective. It can be understood that applications possess the real-time traffic, the QoS parameters serve to be the delay as specified by the WiMAX forum. For applications with the non real-time traffic, throughput (WiMAX, 2008) is the best indicator for the applications. Thus, in this simulation we choose delay and throughput since they are the most general criteria used for evaluating the performance of the scheduling algorithm. The mean waiting time (queueing delay) of the grants/bandwidth request messages and QoS metric of fairness is highlighted too.

We have chosen to simulate simultaneously the UGS, rtPS and nrtPS for Scenario 1 while the rtPS, nrtPS and BE for Scenario 2. Through this, we highlight the importance of the threshold setting which represents the number of bandwidth request message in the nrtPS queue.

Scenario 1

In scenario 1, ten MSs are configured to HTTP using the nrtPS while an increasing number of MSs performing VoIP and video conferencing which are associated with the UGS and rtPS service classes respectively are adopted for the purpose of expanding the load. The VoIP server, video conferencing server and HTTP server are linked up to the BS via the core network as shown in Figure 5. We compare the performance of the proposed algorithm with reference to Settembre et al. (2006) and Jani and Jani (2008) for the purpose of evalution and comparison. Both Settembre et al. (2006) and Jani and Jani (2008) used the PQ algorithm as the first stage of the allocation. We have chosen the work of Settembre et al. (2006) and Jani and Jani (2008) with the condition that our proposed algorithm falls into the category of priority-based algorithms. As for the second stage, the fixed bandwidth implementation, weighted round robin (WRR) (Settembre et al., 2006) and deficit round robin (DRR) (Jani and Jani, 2008) are involved. Fixed bandwidth implementation is specified in the IEEE 802.16e as the algorithm for the UGS and ertPS service classes.

We can observe that generally the average throughput increases as more MSs of UGS participate in it. However, the average throughput of fixed bandwidth implementation does not increase after the offered load overwhelms the system capacity. The TbCP of Thresholds 1 and 10 outperforms the fixed bandwidth implementation when the number of MSs reaches 40.

Figure 6 (b) further examines the average delay of VoIP application. The TbCP of Thresholds 1 and 10 has shown an increase in delay for TbCP as compared to fixed bandwidth especially when MS approaches 40. This is due to the additional delay introduced in the scheduler which is the time that the grants have to wait (waiting time or queuing delay) to be serviced (Table 4).

However, TbCP with Threshold 10 has reduced the mean waiting time as compared to Threshold 1.

In UGS, the BS provides fixed-size grants (IE) in the UL-MAP to the MSs at periodic intervals, thus as the number of MS increases, higher amount of packets are transmitted. The incoming VoIP packets are dropped upon arrival if the buffer holding the packets is full. Therefore, the average delay of VoIP is very small in the low loading condition because most packets are immediately serviced without dropping but steeply increases if the buffer is about to be full.

Figure 7 (a) examines the rtPS throughput of the algorithms. The WRR algorithm is executed on the basis of weight value that specifies the number of bandwidth request messages that can be delivered from each queue during a round. The rtPS is allotted with a bandwidth which is equal to the weight assigned and calculated as:

$$weight\ [j] = \frac{MRTR\ (i,j)}{\sum_{j=0}^{J_i} MRTR\ (i,j)}\ x\ Total\ Capacity \qquad (21)$$

where weight [j] is the weight of the jth connection of the ith service class and $\sum_{j=0}^{J_i} MRTR(i,j)$ is the total of the minimum reserved traffic rate of the ith service class.

As for DRR, the bandwidth request message is transmitted if the deficit counter is greater than the size of the head-of-line message. Upon transmission, the deficit counter is decreased. If the deficit counter is smaller than the size of the head-of-line message, the queue is unable to be sent. The quantum is saved and added to the deficit counter in the following round. If the queue is empty, the

Table 1. Coding rate and bits per symbol for different MCS.

MCS	QPSK		16-QAM		64-QAM		
Code rate	1/2	3/4	1/2	3/4	1/2	2/3	3/4
Bits per symbol	2		4		6		

Table 2. Simulation parameters.

Parameter	Value
PHY profile	OFDMA
Bandwidth	10 MHz
Base frequency	2.5 GHz
TTG (transmit-receive transition gap)	106 µs
RTG (receive-transmit transition gap)	60 µs
OFDMA symbol duration	100.8 µs
Frame preamble	1 symbol
Duplexing mode	TDD
FFT size	1024
Frame duration	5 ms
Subframe ratio (DL/UL)	1:1
MSTR UGS	96000 b/s
MRTR rtPS	80000 b/s
MRTR nrtPS	50000 b/s
MRTR BE	40000 b/s
Polling time (rtPS)	2 ms
Polling time (nrtPS)	10 ms

Table 3. Traffic parameters.

Application	Parameters
Voice	- T_{on} = 0.352 - T_{off} = 0.65 - Size = 10 ms - Coding rate = 64 kbps
Video Conference (Yi et al., 2009; Hua and Lars, 2007)	Frame size: - Lognormal distribution - Average : 4.9 bytes - Standard deviation : 0.75 bytes Inter-arrival time: - Normal distribution - Mean : 33 msec - Standard deviation : 10 msec
FTP	Inter request time: - Constant distribution (30 sec) File size : - Constant distribution - 10000 bytes
Web Browsing (HTTP)	Page interarrival time: - Exponential (exp) distribution (30 sec) Page properties: - Object size:exp 1000 bytes - Object per page: exp 4

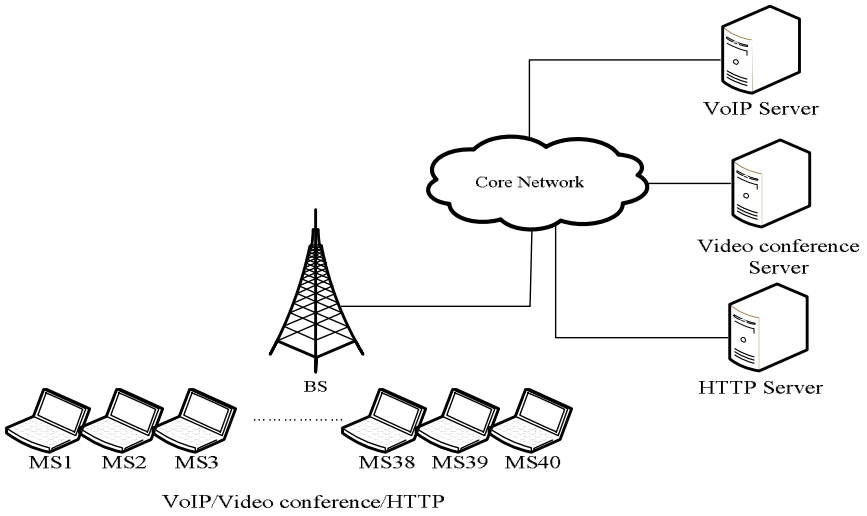
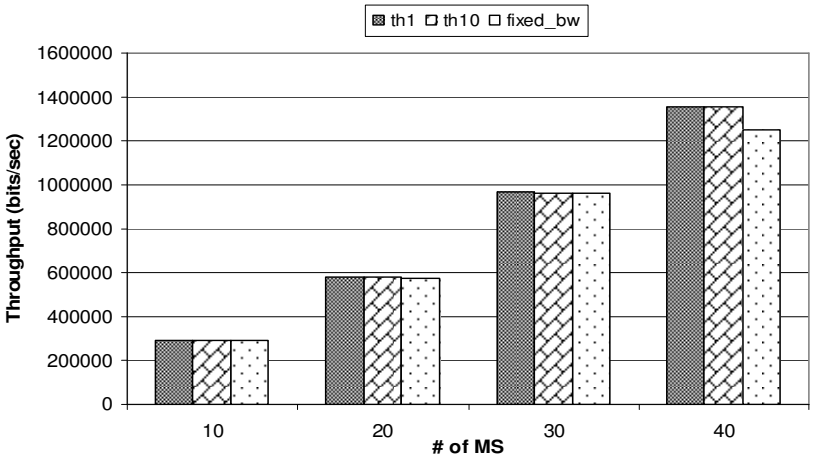
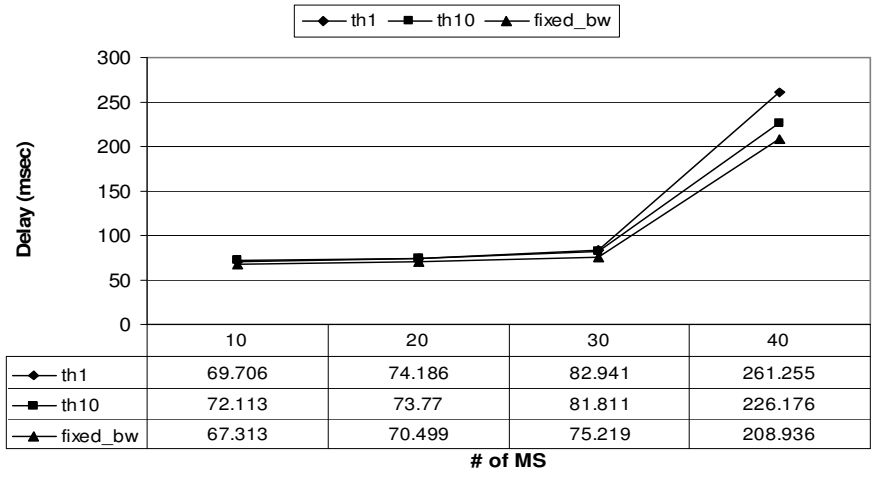


Figure 5. Network simulation topology for Scenario 1.



(a)



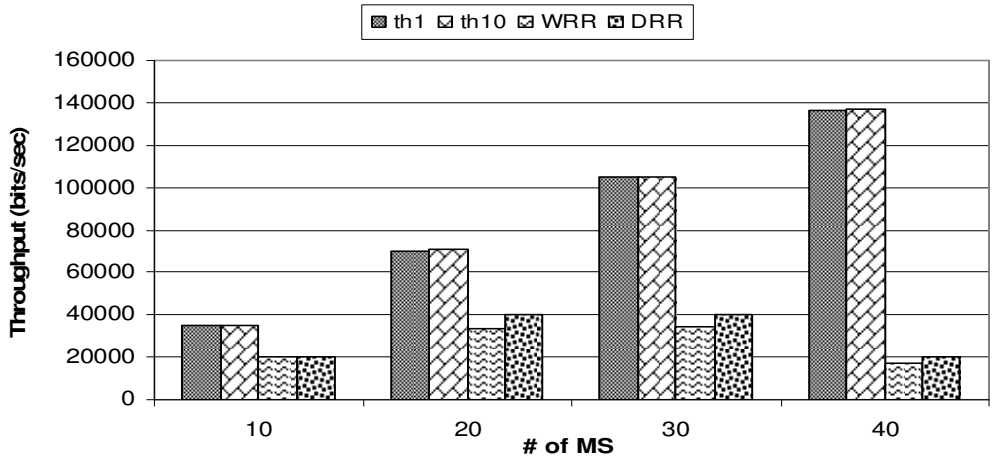
	10	20	30	40
th1	69.706	74.186	82.941	261.255
th10	72.113	73.77	81.811	226.176
fixed_bw	67.313	70.499	75.219	208.936

(b)

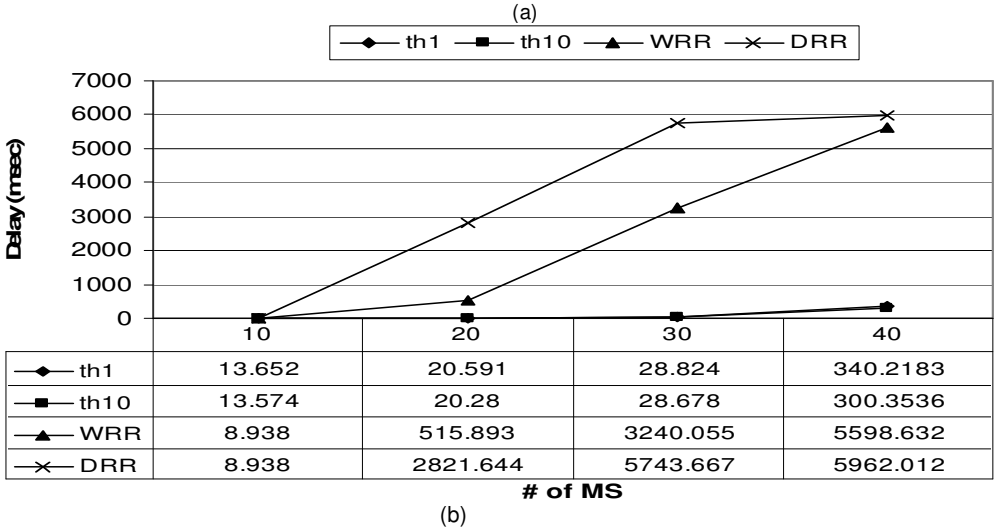
Figure 6. (a) UGS throughput, (b) average delay.

Table 4. Mean waiting time (s) of the grants in the UGS queue.

# of MS	th1	th10	fixed_bw
10	0.001907	0.001906	-
20	0.005241	0.005239	-
30	0.010257	0.010244	-
40	0.019586	0.019002	-



(a)



	10	20	30	40
th1	13.652	20.591	28.824	340.2183
th10	13.574	20.28	28.678	300.3536
WRR	8.938	515.893	3240.055	5598.632
DRR	8.938	2821.644	5743.667	5962.012

(b)

Figure 7. (a) rtPS throughput, (b) average delay.

deficit counter is then set to zero. We continue to serve the queues as long as the deficit counter allows the bandwidth request messages to be transmitted to the UL-MAP. A quantum is defined as an amount in bits or bytes given to a queue whenever it is served and defined as follows:

$$quantum\ [\,j\,] = \frac{MRTR\ (i,j)}{\sum_{j=0}^{J_i} MRTR\ (i,j)}\ x\ Total\ \ Capacity \qquad (22)$$

where quantum [j] is the quantum value for the jth connection of the ith service class and $\sum_{j=0}^{J_i} MRTR(i,j)$ is the total minimum reserved traffic rate of the ith service class.

We observe that WRR starts to deliver the lowest throughput when the number of MS advances to 20 because the algorithm cannot perform well in the presence of variable size packets. However, we can see

that the throughput delivered using the DRR algorithm is much improved as compared to WRR. This is mainly because the DRR suits packets that vary in size; although the algorithm needs accurate information of the packet size. The BS is only capable of estimating the UL traffic (Ahmet et al., 2009) through the bandwidth request messages sent from the MS and the MRTR assigned for each service class. The actual packets reside in the MS in which there is no mechanism that serves to transmit the information of the individual packet sizes to the BS. Thus, the use of MRTR in determining the values of weight and quantum thus affect the throughput and delay. Figure 7 (b) demonstrates the average delay in which the DRR displays the highest delay while TbCP produces the lowest delay. The TbCP with the threshold of 10 has reduced the delay as compared to threshold of 1 which is in line with the mean waiting time. We observed that the mean waiting time (Table 5) of the DRR is higher than the WRR. The scheduler continues to serve the queues as long as the deficit counter is greater than the size of the head-of-line of the bandwidth request message. On the other hand, if the deficit counter is too small than the size of the head-of-line message, thus the queue is unable to be send, which subsequently has caused an increased in the mean waiting time.

The throughput of nrtPS is presented in Figure 8. TbCP of Threshold 10 delivers the highest throughput among the algorithms. This is because, when the threshold value is set to 1, the scheduler will service the nrtPS queue after serving the rtPS queue, that is when there is ≥ 1 bandwith request message available in the buffer. At this point in time, the number of bandwidth request message that is serviced is basically less than the number of bandwidth request message when the threshold value is set to 10. Furthermore, the non-real time traffic load is configured to be lower than the real-time traffic load and the MS with nrtPS connections is polled every 10 ms which is much later than the MS with rtPS connections thus, resulting in smaller number of incoming bandwidth request messages directed to the nrtPS queue.

Scenario 2

In Scenario 2 eight MSs are configured to the FTP and HTTP server using the nrtPS and BE service classes respectively while an increasing number of MSs performing video conferencing using rtPS are employed to enlarge the load. The video conferencing server, FTP server and HTTP server are linked up to the BS via the core network. The network topology is illustrated in Figure 9. In this evaluation, we compare the proposed TbCP with the work of Settembre et al. (2006) and Po et al. (2009). Red-based DFPQ (Po et al., 2009) is a type of priority-based algorithm in which the highest priority class is fixed with a larger quantum value than the lower priority class. As for the work of Settembre et al. (2006), the round robin (RR) is chosen to be the intra-class scheduling for the BE service class.

Figure 10 (a) shows the throughput of the rtPS with the TbCP of Thresholds 1and 10 deliver the highest for the QoS class. In spite of the fact that the delay of Figure 10 rendered low and bounded as compared to PQ+WRR and Red-based DFPQ algorithms.

Figure 11 (a) compares the throughput for the BE service class. The comparison shows that our proposed algorithm with the threshold of 1 delivers the highest throughput for the BE. The exhaustive service strategy employed helps to improve the amount of throughput delivered. However, for the threshold of 10, the scheduler needs to satisfy the higher priority service class as long as the bandwidth request messages are less than the threshold assigned, causing the throughput to be the lowest.

The Red-based DFPQ delivers a lower average throughput as compared to PQ+RR when the MS approaches 20 because the rtPS is allowed higher transmission opportunities when the number of bandwidth request messages of rtPS increases. This is because the counter of the rtPS is adjusted adaptively according to the queue length of the rtPS.

The combination of the PQ algorithm and RR reduces the throughput of the BE class because the scheduler needs to satisfy the higher priority class before the BE is served. Furthermore, the RR scheduling algorithm would allow only one bandwidth request message to be served in each round. The intra-class fairness between rtPS and BE can be calculated as (Yi et al., 2009):

$$Fairness = \left| \frac{Th_{rtPS}}{S_{rtPS}} - \frac{Th_{BE}}{S_{BE}} \right| \quad (23)$$

where S_{rtPS} and Th_{rtPS} are the total traffic and corresponding throughput of rtPS, S_{BE} and Th_{BE} are those of BE. A scheduling algorithm is said to be fair if the difference in the normalized services received by different service flows in the scheduler is bounded (Chen et al., 2005). Figure 11(b) shows that the TbCP of Threshold 1 can still keep the fairness boundary under 0.1. In the Red-based DFPQ, the counter of BE is set to a fixed quantum which is equal to the MRTR. However, the BE transmission opportunity is sacrificed as the rtPS becomes demanding causing the throughput to be lower and thus affecting the fairness index. Furthermore, determining the appropriate value of the counter is crucial as it will cause the fairness to suffer if the value is not configured properly.

Conclusion

In this paper, we introduced a threshold based priority

Table 5. Mean waiting time (s) of the bandwidth request messages in the rtPS queue.

Number of MS	th 1	th10	WRR	DRR
10	0.007245	0.007234	0.001708	0.002446
20	0.014234	0.014227	6.516451	6.887139
30	0.021912	0.021876	7.392593	11.86776
40	0.344019	0.304334	12.75495	14.18106

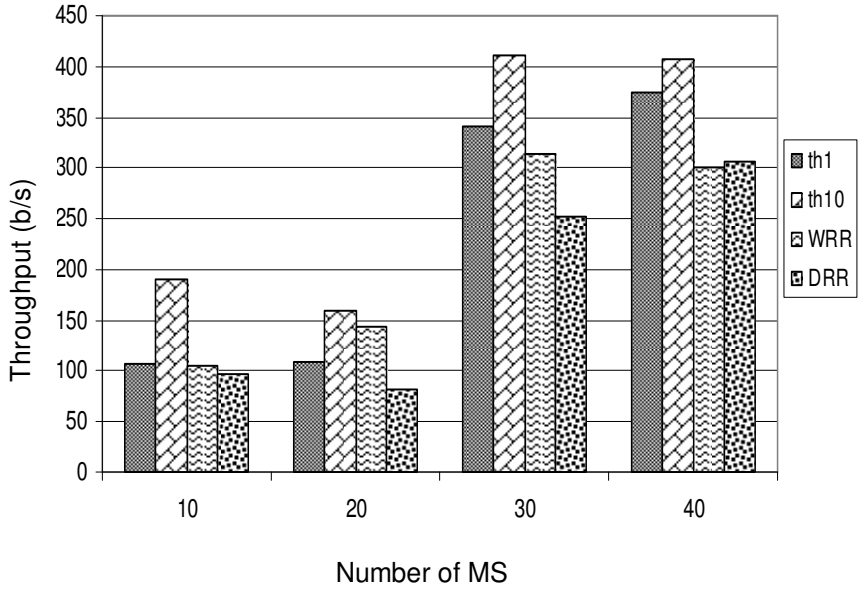


Figure 8. nrtPS throughput.

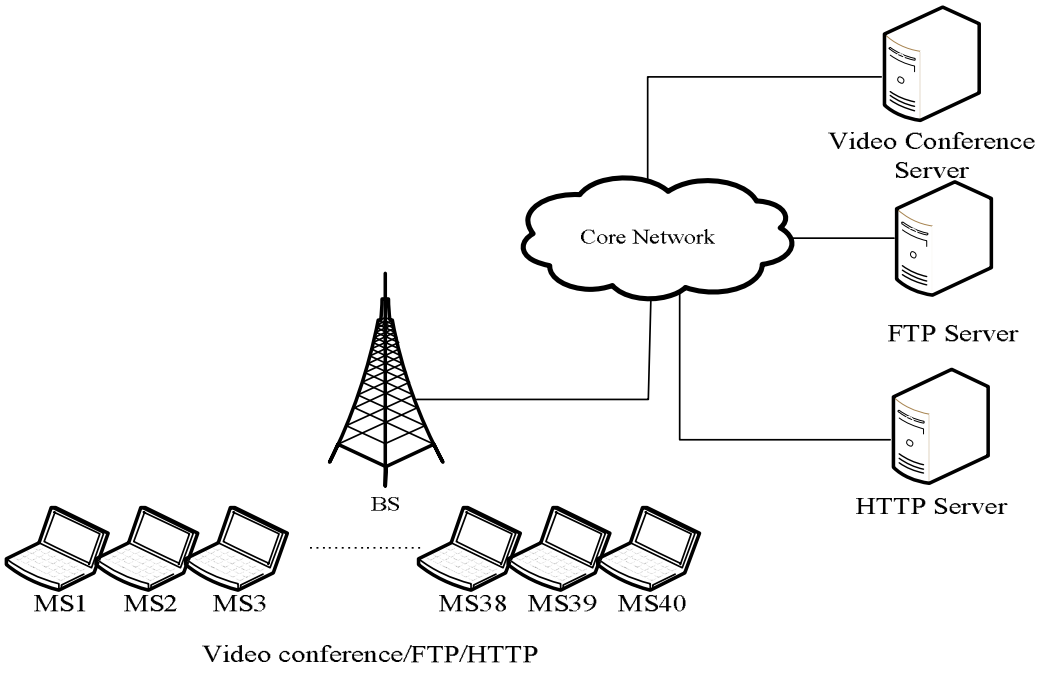
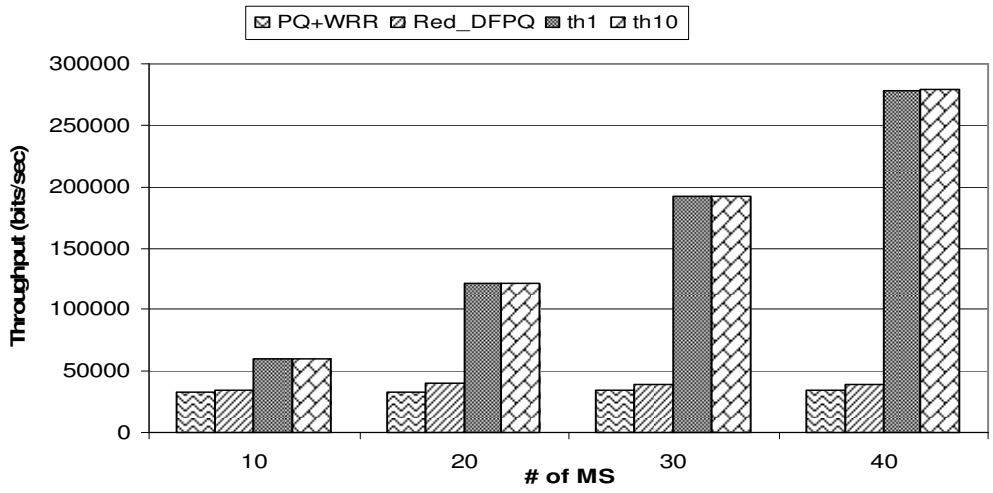
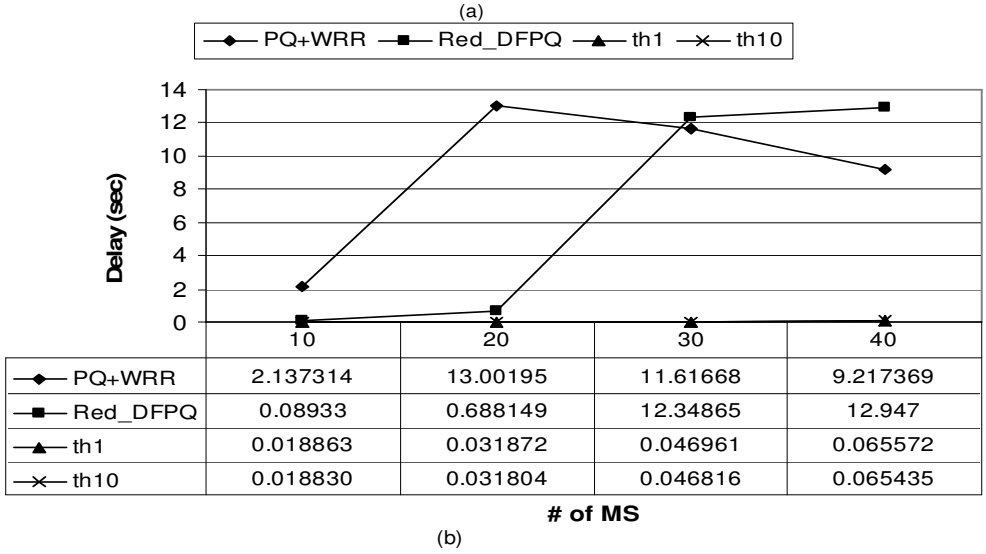


Figure 9. Network simulation topology.





	10	20	30	40
PQ+WRR	2.137314	13.00195	11.61668	9.217369
Red_DFPQ	0.08933	0.688149	12.34865	12.947
th1	0.018863	0.031872	0.046961	0.065572
th10	0.018830	0.031804	0.046816	0.065435

(b)

Figure 10. (a) rtPS throughput, (b) average delay.

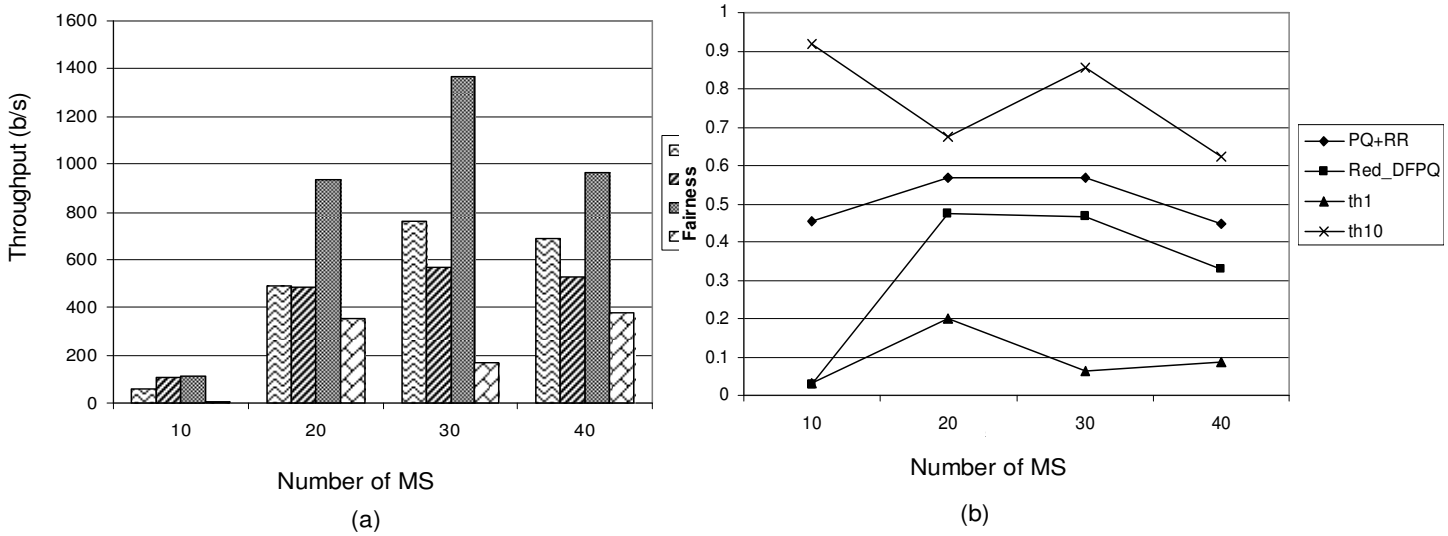


Figure 11. (a) BE throughput, (b) fairness.

algorithm for the UL scheduler of the mobile WiMAX. The algorithm supports the QoS classes as defined by the standard. We have manipulated the threshold value set upon the nrtPS bandwidth request message to take advantage of the delay-tolerant feature demonstrated by this service class. We have showed the importance of the threshold value and its outcome towards the delay of the real-time traffic and throughput of the non real-time traffic. Results from simulations show that when the network is configured with a higher concentration of latency-guaranteed QoS classes (UGS, ertPS and rtPS), then the threshold value should be set higher which is 10. This can be demonstrated from the results of Scenario 1 with low bounded delay and higher throughput delivered from the UGS, lowest delay of the rtPS and highest throughput achieved of the nrtPS. On the other hand, results from scenario 2 of the simulations prove that the threshold of 1 is the most appropriate value. This is specially so, when the network is configured with higher concentration of non-real time service classes and confirmed through the results of higher rtPS throughput and low bounded delay, highest throughput delivered by the BE and finally is observed to be fairer than PQ+RR, Red-based DFPQ and TbCP with the threshold of 10.

We have formulated the mathematical modeling of our proposed UL scheduling algorithm using the pseudo-conservation law (Darmawaty and Kaharudin, 2010). However, since the analyses do not result in explicit expressions of the mean waiting time, this has impeded us from working on the numerical analysis along with the simulation in this paper. The results of the analysis is relatively a simple expression for the weighted sum of the mean waiting times, which may be used as a first indication of overall system performance especially when queues are asymmetric (Alex, 2007) (where each queue may have different stochastic characteristic such as arrival rate and service time distribution) and when their number is large (more than 2 queues where numerical computation of the exact formulas can become very cumbersome).

As for our future direction, we recommend an admission control scheme that adaptively admits the connection, by taking into account the threshold setting and the QoS parameters. Furthermore, the threshold value should be self-configuring, implying that the adjustment of the value should be done intelligently based on the configuration of the network instead of being adjusted manually.

REFERENCES

Ahmet Sekercioglu Y, Milosh I, Alper Y (2009). A survey of MAC based QoS implementation for WiMAX networks. Comput. Commun., 53: 2517-2536.

Alex R (2007). Polling systems and their applications. Master dissertation, Vrije Universiteit Amsterdam.

Borst SC (2006). Polling systems. CWI Amsterdam: CWI Tract. pp. 1-51.

Chakchai SI, Raj J, Abdel KT (2009). Scheduling in IEEE 802.16e Mobile WiMAX networks: key Issues and a survey. IEEE J. Sel. Area Comm., 27: 156-171.

Chen J, Jiao W, Wang H (2005). A service flow management strategy for IEEE 802.16 broadband wireless access syst. in TDD Mode. In Proceedings of the IEEE Int. Conf. in Commun. held at Seoul, Korea, pp. 3422-3426.

Darmawaty MA, Kaharudin D (2010). On the modelling of the Mobile WiMAX (IEEE 802.16e) uplink scheduler. Modelling and Simulation in Engin.,2010:(804939). doi:10.1155/2010/804939.

Fen H, Pin HH, Xuemin (Sherman) S (2006). Performance evaluation for unsolicited grant service flows in 802.16 networks. In Proceedings of an Int. Conf. on Communications and Mobile Computing held at Vancouver, Canada, pp. 991-996.

Fen H, Pin HH, Xuemin (Sherman) S, An YC (2007). A novel QoS scheduling scheme in IEEE 802.16 networks. In Proceedings of the Wireless Commun. Networking Conf., pp. 2457-2462.

Haitham AG, Jun C, Attahiru SA (2008). Performance analysis for polling service in IEEE 802.16 networks under PMP mode. In Proceedings of the Wireless Communications and Mobile Computing Conf. held at Crete Island, pp. 819-824.

Hanoch L, Moshe S (1990). Polling systems: applications, modeling and optimization. IEEE T. Commun., 38:1750-1760.

Howon L, Taesoo K, Dong HC (2004). An efficient uplink scheduling algorithm for VoIP services in IEEE 802.16 BWA systems. In Proceedings of the Vehicular Techno. Conf. held at Los Angeles, California. pp. 3070-3074.

Hua W, Lars D (2007). Adaptive radio resource allocation in hierarchical Qos scheduling for IEEE 802.16 systems. In Proceedings of the IEEE Global Telecommun. Conf held at Washington DC, USA, pp. 4769-4774.

Jae WS (2008). Performance analysis of uplink scheduling algorithms for VoIP services in the IEEE 802.16e OFDMA syst. Wireless Pers. Commun.: Ann. Int. J., 47: 247-263.

Jani LAS, Jani M (2008). Comparison of different scheduling algorithms for WiMAX base station. In Proceedings of the Wireless Commun. and Networking Conf. held at Las Vegas, USA. pp. 1991-1996.

Nuaymi, Loutfi (2007). WiMAX technology for broadband wireless access, England: Wiley.

Onno JB, Wim PG (1987). Pseudo-conservation laws in cyclic-service systems. J. Appl. Probab., 24: 949-964.

Onno JB, Wim PG (1988).Waiting times in discrete-time cyclic-service syst. IEEE T. Commun., 36: 164-170.

Po CT, Chia YY, Naveen Ci, Wang TH, Ce KS (2007). A proposed RED-based scheduling scheme for QoS in WiMAX Networks. In Proceedings of an Int. Symposium on Wireless Pervasive Computing held at Melbourne, Australia, pp. 1-5.

Settembre M, Puleri M, Garritano S, Testa, Albanese MMR, Lo CV (2006). Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding. In Proceedings of an Int. Symposium on Comput. Networks held at Istanbul, pp. 11-16.

Sze YT, David C, Yoong CC (2009). Uplink traffic scheduling with QoS support in broadband wireless access networks. In Proceedings of the IEEE 9th Malaysia Int. Conf. on Commun. held at K. Lumpur, Malaysia, pp. 800-805.

WiMAX Forum (2008). WiMAX system evaluation methodology Version 2.1, WiMAX Forum.

Yi NL, Ying DL, Yuan CL, Che WW (2009). Highest Urgency First (HUF): A latency and modulation aware bandwidth allocation algorithm for WiMAX base stations. Comput. Commun., 32: 332-342.

Yoshitaka T, Krishna KB (1995). Pseudo-conservation law for a priority polling system with mixed service strategies. Perform. Evaluation, 23: 107-120.

Zsolt S, Miklos T (2010). Analysis of globally gated Markovian Limited cyclic polling model and its application to IEEE 802.16 network. In Proceedings of the 5th Int. Conf. on Queueing Theory and Network Applications held at Beijing, China, pp.1-8.