Dual Threshold Scheduling for VoIP Traffic on Downlink of WiMAX Networks

Chin-ling Chen* and Cheng-yi Pan
National Pingtung University,
Department of Information Management,
Pingtung, Taiwan 900
Email: clchen@mail.npic.edu.tw

***Abstract*—Previous downlink scheduling algorithms of WiMAX (Worldwide Interoperability for Microwave Access), like DRR (Deficit Round-Robin) and WRR (Weighted Round-Robin), usually reserve minimum rate to each type of traffic and cannot consider the status of queue length of each connection, thus making it unsuitable for VoIP on-off traffic model. One well-designed scheduling algorithm is expected to coordinate QoS-related functional entities in WiMAX architecture. We address an efficient downlink scheduling algorithm, which allocate the bandwidth based on dual threshold of queue length. We compare the proposed scheme with DRR and WRR by estimating the system performance such as average delay, loss rate and throughput under several traffic scenario and system parameters value.**

***Index Terms*—Scheduling, WiMAX, VoIP and QoS**

I. INTRODUCTION

Worldwide Interoperability for Microwave Access (WiMAX) [1] has become the alternative of fourth generation (4G) mobile broadband networks due to its availability of providing technological advances to mobile operators to satisfy the mobile broadband service demand of their users. The various services, such as voice, video and data, require different Quality of Services (QoS) requirements. Therefore, we need traffic scheduling to ensure that QoS requirements are met. Scheduling is used to allocate bandwidth and determine the transmission priority when there are many users contend for resources. How to allocate resources in an efficient way and the provision of QoS guarantee are the major issues in delivering delay sensitive traffic, like VoIP service, in WiMAX.

There are three types of scheduling algorithms for WiMAX. Two for base station (BS), namely, downlink (DL) scheduling and uplink (UL) request/grant scheduling. The third is the DL scheduling at subscriber station (SS). In the paper, we especially focus on the DL scheduling at BS.

Manuscript received April 10, 2014; revised May 5, 2014; accepted May 12, xx, 2014.
This paper was sponsored by: NSC 100-2221-E-251-007-
* Corresponding author

Classical DL scheduling at BS can be categorized into three types: Round Robin (RR) [2-3], Weighted Fair Queuing (WFQ) [4-6] and Priority-based (PR) algorithms [7-14]. Round Robin (RR) provides the fairness among the users in the case that the allocation for a given number of bytes or the packet size is fixed. However, it may not assure QoS requirements for various service classes. Weight Round Robin (WRR) is, therefore, proposed for meeting the throughput guarantee by that the weight can be dynamically adjusted in term of queue length, delay and the number of slots. Deficit Round Robin (DRR) [2] and Deficit Weighted Round Robin (DWRR) [3] can be used for the variable length packets. The advantages of these variations of RR are the ability of providing fair queuing with the simplicity of implementation.

In WFQ [4], each connection maintains its own queue. The weight is assigned dynamically for each queue based on the QoS requirement. When the server choose next packet for transmission, it selects the first packet that would complete service. In worst-case fair weighted fair queuing (WF^2Q) [5], the server only considers the set of packets that have started receiving service. Both WFQ and WF^2Q have high complexity O(N). WF^2Q+ [6] is, therefore, introduced to have lower complexity O(log$N$). WF^2Q+ maintains a virtual time function and assigns to each queue a virtual *start* service time and a virtual *finish* service time. WF^2Q+ works by selecting the eligible queue with the *smallest virtual finish time.*

Recently, many methods of PR have been proposed. They can be further classified into two types: delay-based [7-10] and slice-based [11-14]. Earlier Deadline First (EDF) is the fundamental delay-based concept for scheduler to search all the queues for the packet closest to its deadline. The other well-known algorithms include Largest Weighted Delay First (LWDF) [7], Delay Threshold Priority Queuing (DTPQ) [8] and Adaptive Delay Threshold-Based Priority (ADTP) [9]. LWDF scheduling [7], which is parameterized by a weight vector, always chooses for service the longest waiting packet from the queues for which the current *weighted delay* is maximal. Different from a typical priority queuing-based scheduling in which real-time users are always served prior to non-real-time users while non-real-time users are served with the remaining resource, DTPQ [8] takes the

doi:10.4304/jcp.9.11.2712-2719

urgency of the real-time service into account only when their head-of-line (HOL) packet delays exceed a given threshold. Real-time users can be delayed so as to maximize the throughput for the non-real-time users whenever the other QoS requirement (for example, loss rate) is satisfied. ADTP [9], a dynamic version of DTPQ, determines the delay threshold in an adaptive manner as the service scenario varies. The delay threshold is updated based on the varying urgency metric, which is defined as a weighted sum of the delay for the most urgent real-time users and average data rate for real-time users.

Slice-based scheduling usually involves in queue prioritization, flow queuing and the resources need to be reserved per slice basis. Such scheduling can be categorized into two classes: resource-based [11-12] and bandwidth-based [12-15]. Resource-based scheduling allocates a slice in terms of a fraction of the base station's resource slots per OFDMA frame. Bandwidth-based scheduling, on the other hand, allocates a slice in terms of the aggregate throughput that will be obtained by flows of that slice.

In order to guarantee the QoS requirement for various service classes, hybrid priority-based scheme can be used in WiMAX scheduler. In the proposal, queue length is used to set the priority level. We set up two thresholds on the queue, a minimum threshold, T_{min}, and a maximum threshold, T_{max}. The proposed algorithm marks one of three states based on the current queue length. In state 1, all the queue lengths are below minimum threshold. RR fairly allocates the bandwidth slots one by one to all connections. In state 2, more than one queue lengths are greater than minimum threshold, but less than maximum threshold. The scheduler assigns the bandwidth to meet the minimum throughput requirement in term of the number of slots. In state 3, there are more than one queue lengths greater than maximum threshold. The longest one can be considered at the highest priority. More bandwidth is allocated to flows with longer queues. The proposed hybrid scheme combines three classical scheduling algorithms that are suited to specific WiMAX service flows. The paper targets to optimize the total throughput, to minimize delay and loss rate.

The rest of the paper is organized as follows. In section 2, we describe the Queue-Length based Scheduling (QLS). We have the simulation and its results in section 3. Section 4 concludes this paper.

II. The Queue-Length Based Scheduling

Suppose that a base station (BS) maintains n queues for buffer the packets of n flows that share an outgoing link. The proposed scheme, namely, Queue-Length based Scheduling (QLS), monitors the instantaneous occupancy of n queues. The algorithm marks two thresholds on the queue, a minimum threshold, T_{min}, and a maximum threshold, T_{max} (Fig.1), just as RED [15] does.

Suppose QL_i is current queue size of flow i, where $QL_i \geqq 0$, and $i = 1, 2, \ldots, n$. QLS finds flow i^* according to formula (1).

$$i^* = arg_i \max\left\{\frac{QL_i}{\sum_{i=1}^{n} QL_i}\right\} \quad (1)$$

We define j as the states of the queue, where $j \in \{1, 2, 3\}$. QLS classifies flow i^* into one of three states based on the following observation.

$$j = \begin{cases} 1, QL_i \leq T_{min} \\ 2, QL_i > T_{min}, QL_i \leq T_{max} \\ 3, QL_i > T_{max} \end{cases}$$

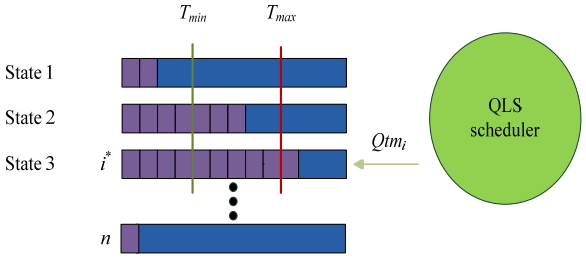

Fig. 1 The proposed model

If the state j of flow i^* is 1, the router is considered to be under-utilization (Fig.2). Suppose Qtm_i is the allocated bandwidth for flow i. QLS initially approximates fair bandwidth allocation for all flows based on the following formula.

$$Qtm_i = \min\left(QL_i, \left\lfloor \frac{w}{n} \right\rfloor\right), \quad (2)$$

where w is the current available bandwidth. After the initial allocation, packets that follows in queue i (QL_i>0) can be transmitted in the remaining space (w>0). We adopt round robin method to assign the queues the time slots that allows it to achieve the objective of fairness. If the state j of flow i^* is not 1, the algorithm dynamically allocates transmission slots to users based on the following formula.

$$Qtm_{i^*} = \begin{cases} min(C, W), & \text{if } T_{min} < QL_{i^*} < T_{max} \\ min(C + QL_{i^*} - T_{max} + 1, W), & \text{if } QL_{i^*} \geq T_{max} \end{cases} \quad (3),$$

where C is the incremental unit of time slots.

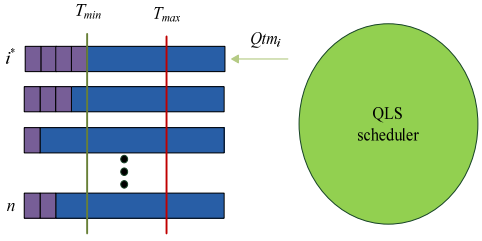

Fig. 2 Under-utilization

If $T_{min} < QL_{i^*} < T_{max}$, the state j of flow i^* is 2. Qtm_{i^*} can be increased by C units by satisfying the minimum constraint for the current available bandwidth ($C < W$). If $QL_{i^*} \geq T_{max}$, the state j of flow i^* is 3. The quantity ($QL_{i^*} - T_{max}$) represents the distance that QLS need to deal with at once. An alternative solution is obtained by assigning additional one unit of time slot ($QL_{i^*} - T_{max}+1$). Actually, the state j of flow i^* is changed from 3 to 2 in this solution. To expedite the allocation process, a larger quantity ($C + QL_{i^*} - T_{max}+1$) is considered as a better solution by satisfying the maximum constraint for the current available bandwidth ($C + QL_{i^*} - T_{max}+1 < W$).

The remaining capacity (W) after allocating flow i^* is obtained by subtracting the allocated time slot (Qtm_{i^*}). The QLS algorithm is shown at Fig.3.

III. Simulation And Its Results

We evaluate the downlink VoIP performance of WiMAX. We use NS-2 [17] as well as CGU-III WiMAX v2.03 [18] for this simulation. For performance analysis, we assume an OFDMA system. A MAC frame consists of 48 symbols and first symbol is used for a preamble. The ratio of the DL symbols to the UL symbols is 36:11. Voice packets are sent by using QPSK 1/2 and we utilize AMR codec in this simulation.

Our G.723.1 VoIP traffic source is a simple On-Off Markov model. 44 bytes of payload is sent for 20 ms during active/on periods and 21 bytes of that for 160 ms during inactive/off periods. All RTP, UDP and IP packet are assumed to be 40 bytes of overhead. Voice activity factor is 40%.All system parameters in this simulation conform to IEEE 802.16 [1]. TABLE I lists the simulation parameters.

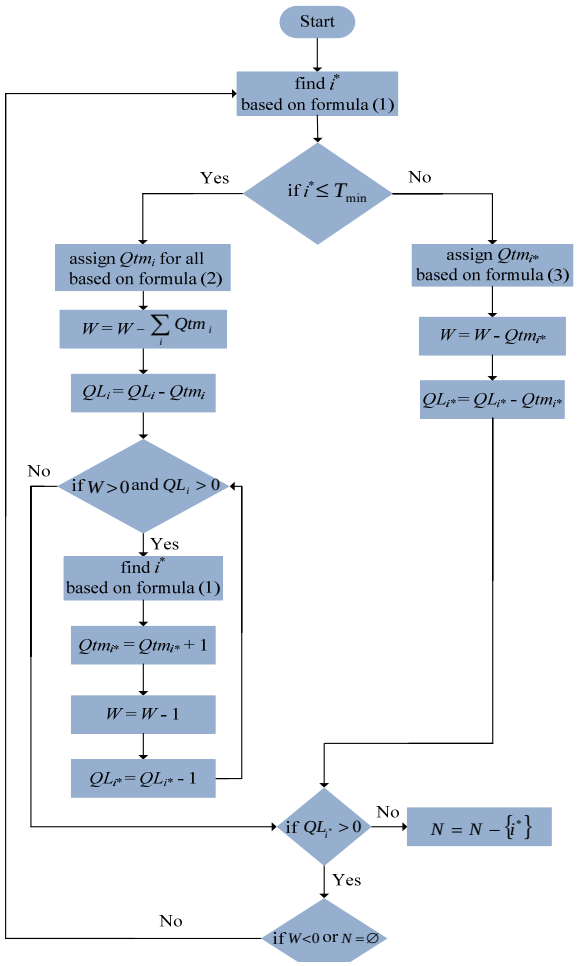

Fig.3 The flowchart of QLS algorithm

TABLE I
SIMULATION PARAMETERS

Parameters	Value
PHY	OFDMA
Duplexing mode	TDD
Frame length (ms)	5
Bandwidth (Mbps)	20
DL modulation	QPSK1/2
No. of DL symbol per frame	36
No. of DL sub-channels	60
Physical queue length (packets)	15
propagation delay (ms)	1

A. Experiment 1: Validity of Parametric Region for QLS

We first study the sensitivity of queue length to two threshold parameters, T_{max} and T_{min}. We first vary the value of T_{max} (8, 10, 12) while fixing the value of T_{min} to be 5 (TABLE II). Fig.4 has shown that all the three scenarios have the same performance irrespective of the network size. However, from Fig.5, higher T_{max} has the effect of increasing packet loss rate once the network size is over 60. A network size N=60 is considered as the

saturated point. The reason is that larger T_{max} has lowered the possibility of being assigned the available bandwidth and has an adverse impact on the performance.

TABLE II
PARAMETER VALIDITY OF QLS WITH VARYING T_{Max}

Scenario	T_{min}	T_{max}
A	5	8
B	5	10
C	5	12

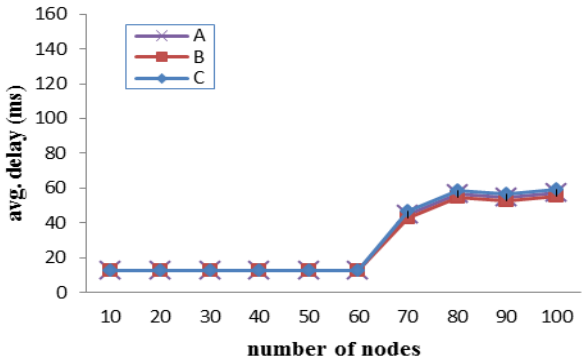

Fig.4 Average delay with varying T_{max}

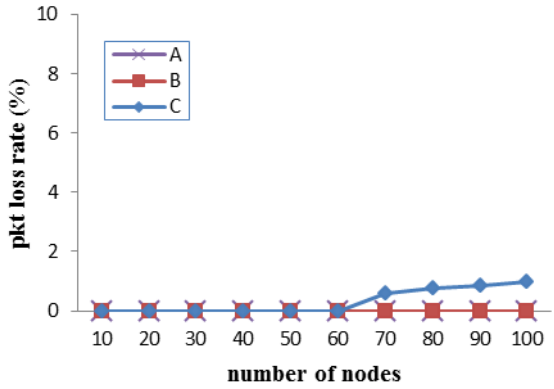

Fig.5 Packet loss rate with varying T_{max}

Alternatively, we fix the value of max to be 10 and use three different values of min (TABLE III) to study the sensitivity of queue length to min. A large min value (Fig.6) will slow down system reaction, thus causing longer delay time. However, from Fig.7, we found that the min value has the minor effect on the packet loss rate. A lower min value will increase the possibility of being state 2 for queue length and thus easily get a larger amount of slots. We call this as Magnetics Effect of state 2 on state 1. A queue in state 1 may easily encounter overflow in case of burst traffic, thus leading to packet loss.

TABLE III
PARAMETER VALIDITY OF QLS WITH VARYING T_{MIN}

Scenario	T_{min}	T_{max}
A	3	10
B	5	10
C	7	10

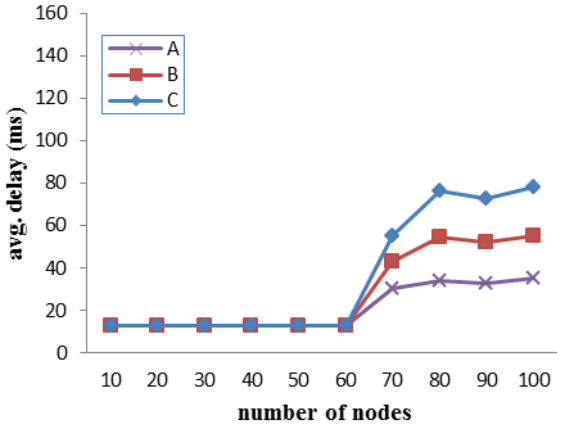

Fig.6 Average delay with varying T_{min}

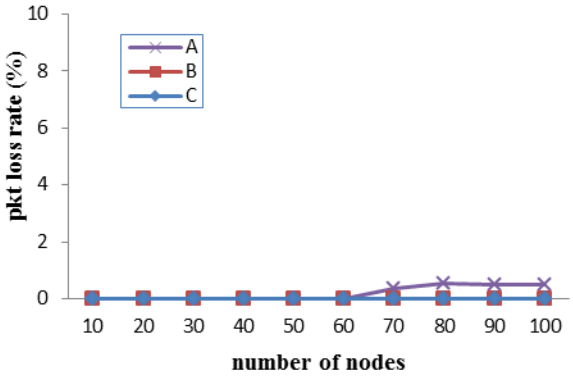

Fig.7 Packet loss rate with varying T_{min}

B. *Experiment 2: Validity of Parametric Region for WRR*

Weighted Round-Robin (WRR) serves flows in round-robin order and transmits a number of packets proportional to its weight for flow i. Let the physical queue length be L. In this experiment, we monitor the queue occupancy of flow i (QL_i) and assign W_j based on the following observation.

$$j = \begin{cases} 1, & QL_i \le \frac{1}{3}L \\ 2, & \frac{1}{3}L < QL_i \le \frac{2}{3}L, \\ 3, & QL_i > \frac{2}{3}L \end{cases}$$

We assume that $W_3 > W_2 > W_1$. TABLE IV lists the combination of W_j. Form Fig.8, we found that The average delay of both scenario A and C are higher than the other two scenarios in case of $N < 20$. However, the performance of scenario B, C, and D have shown the same performance when the network size (N) grows up to 50. Insufficient quantum allocation will be delayed processing till next round, thus leading to higher delay rate. In Fig.9, scenario A tends to discard more packets as the network size (N) increases up to 50. On the other hand, the curves of the other three scenarios have the similar trend. Small quantum allocation (scenario A) has major affected on average delay. This observation is very similar to that of Fig.8.

TABLE IV
PARAMETER VALIDITY OF WRR

Scenario	W_1	W_2	W_3
A	1	2	3
B	2	3	4
C	1	3	5
D	2	4	6

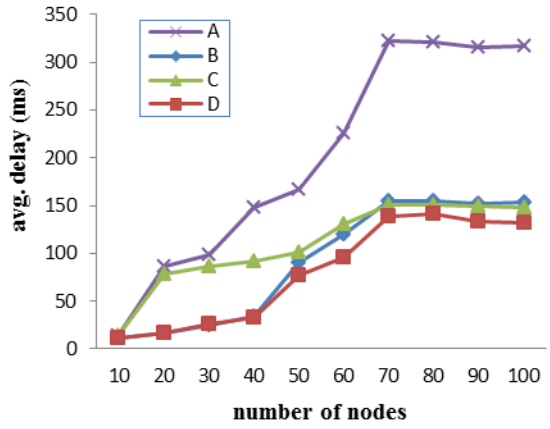

Fig.8 Average delay with combination of quantum W_j

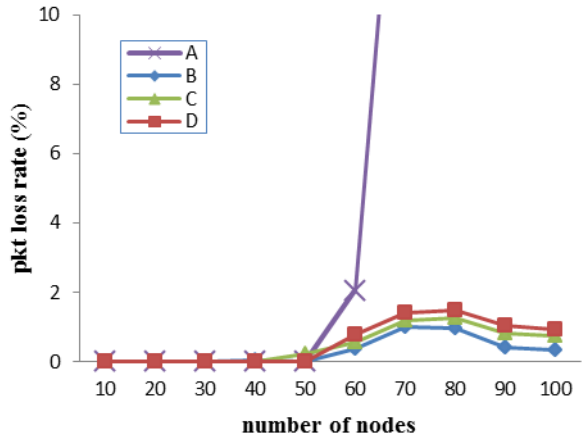

Fig.9 Packet loss rate with combination of quantum W_j

C. *Experiment 3*: *Validity of parametric region for DRR*

We assume that each flow i is allocated Q_i packets in each round. Incoming packets from flow i are stored in queue i. Let $S_i(k)$ be the number of outgoing packets for queue i in round k. We use a deficit counter (DC_i) to indicate the remaining amount ($Q_i - S_i(k)$) for queue i. If queue i is empty, DC is reset to 0. We have an example. In 1st round (Fig.10), two packets are in #1 queue ($S_1(1) = 2$) and three quantum are allocated ($Q_1 = 3$). In 2nd round (Fig.11), deficit counter (DC_1) will be $Q_1 - S_1(1) = 1$

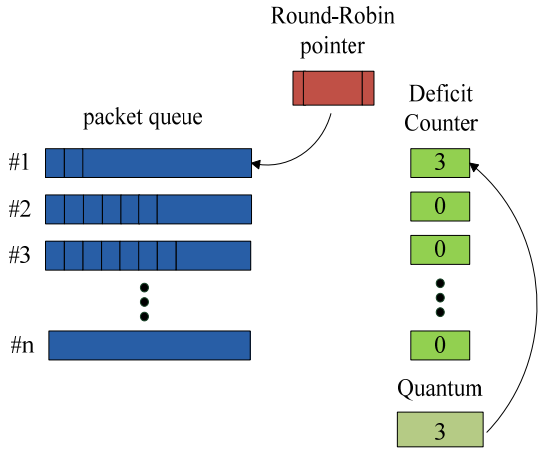

Fig.10 1st round of DRR

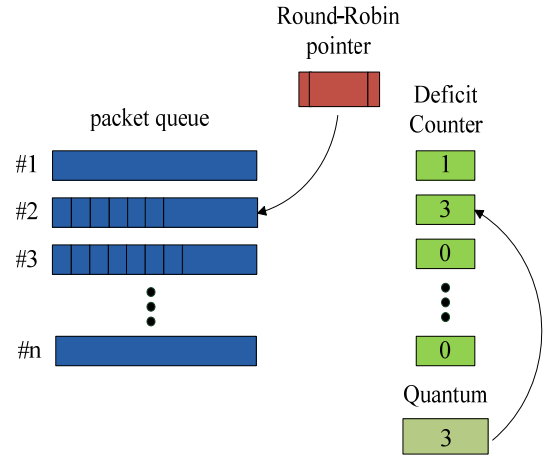

Fig.11 2nd round of DRR

In this experiment, maximum DC is set to be 15. TABLE V lists the varying value of Q_i for this experiment.

TABLE V
PARAMETER VALIDITY OF DRR

Scenario	Q_i
A	1
B	2
C	3
D	4

In Fig.12, the average delay of scenarios C and D grow smoothly as the value of N increases up to 50. The average delay of scenario A is always higher than the other scenarios. The reason is that the remainder from previous quantum is assed to the quantum for next round. Scenario A has to deal with the greater value of DC for next round, therefore causing longer delay.

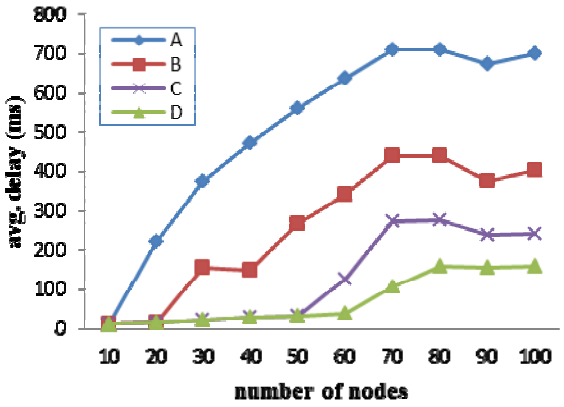

Fig.12 Average delay with varying value of quantum Q_i

The effect of increasing the value of Q_i is similar to that produced by decreasing the possibility of compensation for next round. In Fig.13, we found that the saturation point for scenario A, B, C is 10, 20, 60, respectively. The packet loss rate for the three scenarios increases dramatically in case that N is over the saturation point. On the other hand, scenario D grows smoothly as the value of N increases.

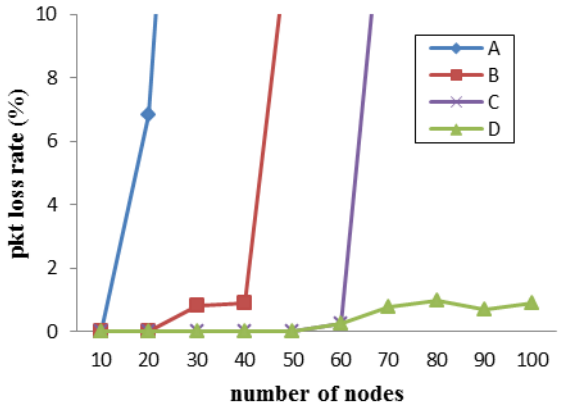

Fig.13 Packet loss rate with varying value of quantum Q_i

D. *Experiment 4: Comparison of Various Scheduling*

Based on experimental results related to DRR and WRR parameters' configuration, scheduling parameters are set as listed in TABLE VI.

TABLE VI
PARAMETER CONFIGURATION

parameters	Value
Quantum of DRR	4
W_1 of WRR	2
W_2 of WRR	3
W_3 of WRR	4
T_{min} of QLS	5
T_{max} of QLS	10
C of QLS	2

To be consistent with this research topic, only downlink end-to-end delay is considered in this simulation. The delay time composes of propagation delay, packet processing delay (queuing delay) and playback buffer delay. Fig.14 shows the average delay time against the number of nodes. The average delay increases linearly as the number of nodes increase. The values of restricted points for WDD, DRR and the proposed algorithm are 40, 60 and 60, respectively. Since resource saturation occurs at the restrict point, the BS cannot assign sufficient downlink resources to surplus users beyond the restrict point. However, the maximum delay bounds (160 ms for both DRR and WRR, and 60ms for QLS) can be obtained due to polling process. Since minimum reserved rate is the basic QoS parameter negotiated by a service flow within a scheduling service, the latency rate scheduling like QLS is ideal.

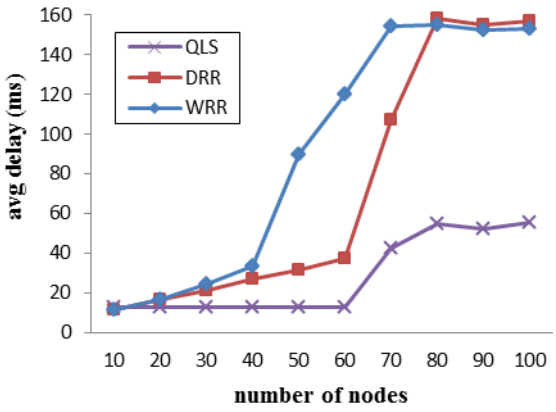

Fig.14 Average delay comparison

Fig.15 shows the average loss rate versus the number of nodes. There is no packet loss for three scheduling algorithms when the system is underutilization ($N \leq 50$). However, the average loss rate for both DRR and WRR increase dramatically in case of $N > 50$. This is because the amount of required resources increases according to the increment in the total number of nodes. DRR requires a minimum rate to be reserved for each service flow, thus making lower loss rate than WRR, up to N=80. The value (N=80) can be considered to be reversed point. Both DRR and WRR have better performance after the reversed point occurs. The system loading could be adjusted owing to polling process. QLS performs very well even under overutilization (N >80). We can find that QLS is an ideal scheduling in any changing environment.

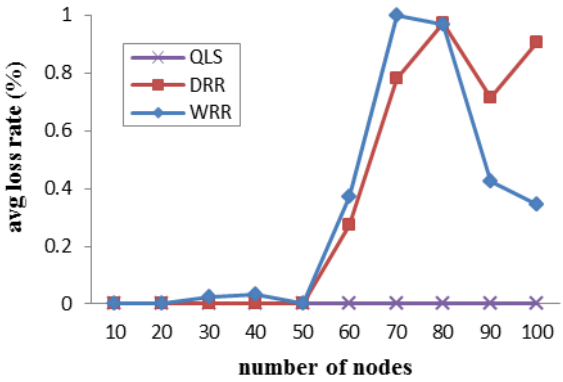

Fig.15 Average loss rate comparison

From Fig.16, it can be seen that the throughputs for three scheduling algorithms grow steadily as the number of nodes increases up to 60. Even though the number of users increases, the average throughput cannot be increased without limit. The value (N=60) can be considered to be the saturation point. After resource saturation occurs, the average throughput decreases slightly according to the increment of the number of nodes (from N=80 to N=90). Since the resource utilization efficiency is greater in our proposed algorithm than in both DRR and WRR, our proposed algorithm has higher throughput compared with both DRR and WRR. QLS takes advantages of excess unreserved bandwidth by other service flows, thereby achieving higher throughput.

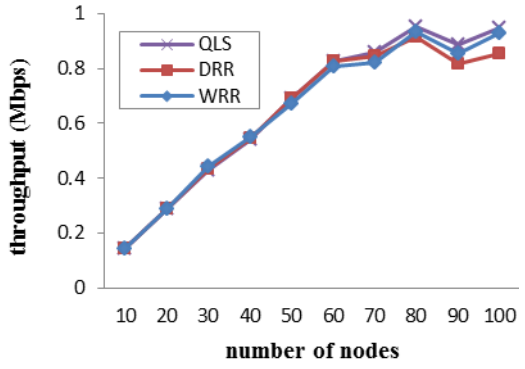

Fig.16 Throughput comparison

IV. Conclusion

The proposed hybrid scheme uses a combination of classic scheduling algorithms in order to better satisfy the QoS requirements. To alleviate problems incurred by classical schemes, the proposed scheme adequately allocates bandwidth to individual traffic flows based on their queue length. Through the performance analysis and simulation results of delay time, packet loss rate and throughput, we have shown that the proposed scheme has better performance than the other scheduling algorithms. The proposed scheme can provide QoS guarantee by ensuring a minimum throughput guarantee and also maintain small delays and loss rate.

Reference

[1] IEEE Std 802.16-2009, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," May 2009.

[2] M. Shreedhar and G. Varghese, "Efficient fair queuing using Deficit Round-Robin," *IEEE/ACM Transactions on Networking*, vol.4, Jun. 1996, pp. 375-385.

[3] C. Cicconetti, L. Lenzini, placeE. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, vol. 20, April 2006, pp. 50-55.

[4] H. Tayyar and H. Alnuweiri, "The complexity of computing virtual-time in weighted fair queuing schedulers," *ICC 2004 - IEEE International Conference on Communications*, no. 1, June 2004, pp. 1996-2002.

[5] X. Fei and A. Marshall, "Delay optimized worst case fair WFQ (WF²Q) packet scheduling," *ICC 2002 - IEEE International Conference on Communications*, no.1, April 2002, pp. 1080-1085.

[6] N. Ciulli and S. Giordano, "Analysis and simulation of WF^{2}Q+ based schedulers: comparisons, compliance with theoretical bounds and influence on end-to-end delay jitter, *Computer Networks*," vol.37, no.5, November 2001, pp. 579-599.

[7] L. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *Annals of Applied Probability*, vol.11, 2001, pp.1-48.

[8] D. H. Kim and C. G. Kang, "Delay Threshold-based Priority Queuing Packet Scheduling for Integrated Services in Mobile Broadband Wireless Access System," in *Proc. IEEE Int. Conf. High Performance Computing and Communications*, Kemer-Antalya, Turkey, 2005, pp. 305-314.

[9] J. M. Ku, S. K. Kim, S. H. Kim, S. Shin, J. H. Kim, and C. G. Kang "Adaptive delay threshold-based priority queuing scheme for packet scheduling in mobile broadband wireless access system," in *Proc. IEEE Wireless Communication and Networking Conf.*, Las Vegas, NV, 2006, vol. 2, pp. 1142-1147.

[10] E. Lucena, F. Lima, W. Freitas and F. Cavalcanti, "Overload prediction based on delay in wireless OFDMA Systems," *GLOBECOM 2010 - IEEE Global Telecommunications Conference*, vol.29, no. 1, December 2010.

[11] T. Choi, S. Kim and D. Sung, "Block scheduling for low-rate, real-time traffic in the downlink mobile WiMAX system," *WCNC 2011 - IEEE Wireless Communications and Networking Conference*, vol.12, no.1, March 2011.

[12] R. Kokku, R. Mahindra, H. Zhang and S. Rangarajan, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," *IEEE/ACM Transactions on Networking*, vol.20, no.5, October 2012.

[13] D. Raychaudhuri, G. Bhanage and R. Daya, "VNTS: A virtual network traffic shaper for air time fairness In 802.16e systems," *ICC 2010 - IEEE International Conference on Communications*, vol.33, no.5, May 2010.

[14] K. Khawam, J. Cohen, D. Marinca and S. Tohme, "Semi-distributed radio resource management for elastic traffic in a hybrid network," *WCNC 2012 - IEEE Wireless Communications and Networking Conference*, vol.13, no.1, April 2012.

[15] Z. T. Sun, A. Gani, X. Y. Sun, N. Liu, "Improving QoS of WiMAX by On_Demand bandwidth allocation based on PMP mode," *Journal of Computers*, vol 6, no 10, pp.2187-2195, October 2011.

[16] S. Floyd and V. Jacobson, "Random Early Detection Gateway for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, vol.1, no.4, 1993, pp. 397-413.

[17] The Network Simulator–2, see *http://www.isi.edu/nsnam/ns/*.

[18] CGU-III WiMAX module, see *http://ndsl.csie.cgu.edu.tw/wimax_ns2.php*.

Chin-ling Chen received his BS degree from National Taiwan University in 1988, the Master degree in Management Information System from University of Wisconsin, Milwaukee, in 1992 and the Ph. D degree in Information Management from National Taiwan University of Science and Technology, 1999. Since 1999 spring, he joined the faculty of Department of Information Management at National Pingtung University, Taiwan. He is currently a Professor at that affiliation. His research interests include Internet QoS, network technology and network management. He is a member of IEEE and IEICE.

Cheng-yi Pan received his Master degree in Information Management from National Pingtung Institute of Commerce, in 2011. Currently, he is in army service

.