Adaptive Radio Resource Allocation in Hierarchical QoS Scheduling for IEEE 802.16 Systems

Wang, Hua; Dittmann, Lars

Published in:
Proceeding of IEEE GLOBECOM 2007

Link to article, DOI:
10.1109/GLOCOM.2007.905

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Wang, H., & Dittmann, L. (2007). Adaptive Radio Resource Allocation in Hierarchical QoS Scheduling for IEEE 802.16 Systems. In Proceeding of IEEE GLOBECOM 2007 (pp. 4769-4774). Washington, DC, USA: IEEE. https://doi.org/10.1109/GLOCOM.2007.905

Adaptive Radio Resource Allocation in Hierarchical QoS Scheduling for IEEE 802.16 Systems

Hua Wang and Lars Dittmann

Department of Communications, Optics & Materials
Technical University of Denmark, Lyngby, Denmark
Email: {huw, ld}@com.dtu.dk

***Abstract*— Future mobile communication systems such as IEEE 802.16 are expected to deliver a variety of multimedia services with diverse QoS requirements. To guarantee the QoS provision, appropriate scheduler architecture and scheduling algorithms have to be carefully designed. In this paper, we propose an adaptive bandwidth distribution algorithm for the aggregate scheduler in a two-level hierarchical scheduler, which can provide more organized service differentiation among different service classes. By taking the backlogged traffic, the spectral efficiency in terms of modulation efficiency, and the QoS satisfaction into account, the proposed algorithm adaptively allocates bandwidth to each service class with the objective of increasing the spectral efficiency while satisfying the QoS requirements. Through system-level simulation, it is shown that the proposed algorithm can adapt to the performance of the class schedulers and distribute the bandwidth among them more efficiently than the conventional schemes.**

I. Introduction

Throughout the world, the demand for broadband wireless access has increased exponentially in the last few years. Future mobile communication systems are designed towards a high-date-rate, low-latency and packet-optimized radio access technology. One example is the IEEE 802.16 wireless metropolitan area networks. The key feature of future mobile communication systems is the ability to deliver a variety of multimedia services with different Quality-of-Service (QoS) requirements, such as throughput, delay, delay jitter, fairness and loss rate. Radio resource allocation and scheduling algorithms play an important role in QoS provision.

Many packet scheduling algorithms have been proposed to support real-time and non-real-time traffics for mobile and wireless networks. Max C/I and Proportional Fair (PF) [4] are first proposed with the design objective of improving the overall system throughput and proportional fairness among users, respectively. Since neither Max C/I nor PF can guarantee any QoS requirements, they can not support RT services such as voice and video streaming. Instead, Modified-Largest Weighed Delay First (M-LWDF) [5] and Exponential (EXP) [3] scheduling algorithms are proposed to support a mixed service of RT and NRT traffics. In addition, other scheduling algorithms with different design objectives have been proposed in [6]-[9].

The above mentioned works can be categorized into one-level priority-based scheduling algorithms. In such approach, each connection is assigned a priority value based on some criterion and the connection with the highest priority is scheduled each time. This approach has the advantage of low implementation complexity. However, due to different traffic characteristics and diverse QoS requirements among RT, NRT and BE service classes, it is hard to well define a unified priority criterion. Thus, it is desirable to individually design the scheduling algorithm for each service class and separate the resource allocation from the packet scheduling. The first paper proposing the idea of a two-level hierarchical scheduler is in [1]. Performance comparisons between one-level and two-level schedulers are evaluated in [2]. However, so far little work has been done in the design of an efficient aggregate scheduler, which is critical on the performance of a two-level hierarchical scheduler and should be carefully designed.

In this paper, we propose an adaptive resource allocation algorithm of the aggregate scheduler. For each service class, the proposed algorithm first estimates the required amount of bandwidths based on the backlogged traffic and the modulation efficiency. Then with respect to the QoS satisfaction, an exponentially smoothed curve is applied to adjust the estimated amount of bandwidth in order to increase the spectral efficiency while maintaining a guaranteed QoS performance. After the bandwidth estimation procedure is done in each service class, the aggregate scheduler distributes the bandwidth among the class schedulers according to the class priority.

The rest of the paper is organized as follows. In Section II, the structure of a two-level hierarchical scheduler is introduced, followed by the design of the class scheduler and the proposed algorithm of the aggregate scheduler. Section III presents the system and traffic models used in the simulation. The simulation results and discussions are presented in Section IV. Finally, a conclusion is drawn in Section V.

II. Structure of a Two-level Hierarchical Scheduler

Fig. 1 depicts the structure of a two-level hierarchical scheduler in a base station (BS) for IEEE 802.16 systems. Arriving packets from the upper layer are classified by the connection classifier according to their connection identifications (CID), and traffic types, and are sent to the corresponding service class and get queued. The scheduler consists of an aggregate scheduler and four class schedulers. The aggregate scheduler distributes bandwidth to each class scheduler. When the class

1930-529X/07/$25.00 © 2007 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

scheduler receives bandwidths from the aggregate scheduler, it serves packets of its flow queues. As the incoming flows in each class scheduler have similar traffic patterns and QoS requirements, the class scheduler can independently choose its own scheduling algorithm which can best meet the QoS requirements. Therefore, the two-level scheduler can have multiple scheduling criteria and better schedule packets in each service class than the one-level scheduler.

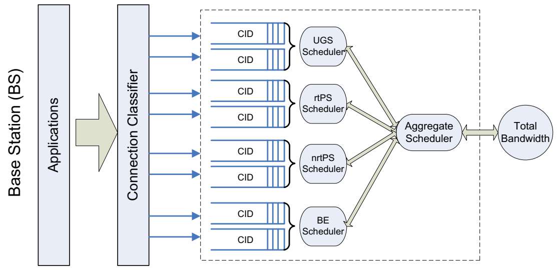

Fig. 1. Structure of a two-level hierarchical scheduler for IEEE 802.16

A. Class Scheduler Design

In this section, we apply the appropriate packet scheduling algorithm to each class scheduler.

1) Scheduling UGS connections: In UGS service, the transmission mode at the PHY layer is fixed during the whole service time. The Adaptive Modulation & Coding (AMC) scheme is not adopted for UGS connections. The time slots allocated for UGS connections are fixed, based on their constant bit-rate requirements negotiated in the initial service access phase [6].

2) Scheduling rtPS connections: The rtPS service is delay-sensitive and has strict delay requirement. We apply the Exponential Rule (EXP) algorithm to schedule rtPS connections. It was proposed to provide QoS guarantees over a shared wireless link in terms of the average packet delay, expressed as $\mathrm{P_r}(W_k > T_{k,max}) \leq \delta_k$, where W_k is the head-of-line packet delay of the k^{th} user, $T_{k,max}$ is the maximum allowable delay, and δ_k is the maximum outage probability. It has been analytically proved that the EXP algorithm is throughput-optimal [3]. At each scheduling time-slot, the EXP algorithm selects user i with the highest priority value as follows:

$$i = \arg\max_k \left\{ \gamma_k \cdot \mu_k(t) \cdot \exp\left(\frac{a_k W_k(t) - \overline{aW}}{1 + \sqrt{\overline{aW}}} \right) \right\} \qquad (1)$$

where $\overline{aW} = \frac{1}{N}\sum_k a_k W_k(t)$, with $\gamma_k = a_k/\overline{\mu}_k$, and $a_k = -\log \delta_k / T_{k,max}$. $\mu_k(t)$ is the instantaneous channel rate at time t, $\overline{\mu}_k$ is the mean channel rate, and subscript k denotes parameters of the k^{th} user.

3) Scheduling nrtPS connections: The nrtPS service can tolerate longer delays, but requires a minimum throughput. We apply the M-LWDF algorithm to schedule nrtPS connections. The design objective of M-LWDF is to maintain the packet delay within a predefined threshold value with certain probability. If the M-LWDF algorithm is used in conjunction with a token bucket control, it can be used to guarantee a minimum throughput $r_{k,req}$ to user k, expressed as $\mathrm{P_r}(\overline{R}_k < r_{k,req}) \leq \delta_k$. We associate each queue with a virtual token bucket. Tokens in each bucket k arrive at a constant rate $r_{k,req}$, which is the guaranteed minimum throughput to the k^{th} user. $W_k(t)$ is the delay of the longest waiting token in token bucket k, calculated as $W_k(t) =$ [Number of tokens in bucket $k/r_{k,req}$]. It has also been analytically proved that the M-LWDF algorithm is throughput-optimal [5]. At each scheduling time-slot, the M-LWDF algorithm selects user i with the highest priority value as follows:

$$i = \arg\max_k \left\{ \gamma_k \cdot \mu_k(t) \cdot W_k(t) \right\} \qquad (2)$$

where γ_k is the same as that of the EXP rule.

4) Scheduling BE connections: Since there is no QoS guarantees for BE connections, we apply the Proportional Fair (PF) algorithm to schedule BE connections. The PF algorithm attempts to serve each user at his peak channel condition. Hence the PF algorithm can utilize the radio resource efficiently and give proportional fairness among users [4]. At each scheduling time-slot, the PF algorithm selects user i with the highest priority value as follows:

$$i = \arg\max_k \frac{\mu_k(t)}{\overline{\mu}_k} \qquad (3)$$

B. Radio Resource Allocation in the Aggregate Scheduler

The bandwidth distribution algorithm of the aggregate scheduler is a critical factor on the performance of the class scheduler. If the aggregate scheduler does not allocate enough bandwidth to the class scheduler, the QoS requirements in the corresponding service class may not be guaranteed. On the other hand, if the aggregate scheduler allocates too much bandwidth to the class scheduler, the allocated radio resource may not be utilized efficiently or even be wasted. So the bandwidth distribution algorithm of the aggregate scheduler has to be carefully designed.

1) Conventional Radio Resource Allocation Algorithms: One possible resource allocation algorithm is that the aggregate scheduler distributes bandwidth among service classes following strict class priority, from highest to lowest: UGS, rtPS, nrtPS and BE. By doing so, the aggregate scheduler can differentiate the service class based on their priority. The strict class priority discipline is simple, but one disadvantage of this algorithm is that higher priority classes may starve the bandwidth for lower priority classes.

To overcome this problem, the aggregate scheduler may separate the total bandwidth into four portions to satisfy proportional fairness among service classes. This method can prevent the starvation of low priority classes. There are static and dynamic bandwidth allocation schemes in this method. In the static scheme, the aggregate scheduler distributes a fixed amount of bandwidth to each class scheduler, thus is suitable when the traffic pattern in each service class is regular and stable, which is not always the case in data communications. Therefore, the dynamic scheme which can adapt to the traffic pattern dynamically is believed to be a better solution.

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

2) Proposed Adaptive Resource Allocation Algorithm: The design objective of our proposed resource allocation algorithm is to adaptively allocate bandwidth to each service class in order to increase the spectral efficiency while satisfying the diverse QoS requirements. In designing our resource allocation algorithm, we have taken the following aspects in each service class into account: (i) the amount of backlogged traffic; (ii) the satisfaction of QoS requirement; (iii) the average spectral efficiency in term of modulation efficiency.

We separate the bandwidth allocation of UGS class from the others as UGS scheduling has been defined by the standard. At the beginning of each frame, the aggregate scheduler allocates a fixed amount of time slots $N_{\mathrm{UGS}} = \sum_{i \in \{\mathrm{UGS}\}} d_i$ to UGS class based on their constant-bit-rate requirements, where d_i is the number of time slots required by UGS connection i. Let N_{total} be the total number of time slots in each frame, then the residual time slots after serving UGS class $N_{\mathrm{rest}} = N_{\mathrm{total}} - N_{\mathrm{UGS}}$ are distributed among rtPS, nrtPS and BE classes, which employs AMC scheme at the PHY layer.

For rtPS class, the amount of bandwidth is estimated upon the backlogged traffic $B_{\mathrm{rtPS}}(t)$ and the average modulation efficiency $\overline{\mu}_{\mathrm{rtPS}}(t)$. As each packet in rtPS has rigid delay requirement, the current queue size in rtPS class scheduler $Q_{\mathrm{rtPS}}(t) = \sum_{i \in \{\mathrm{rtPS}\}} q_i(t)$ is an appropriate measure for the backlogged traffic, where $q_i(t)$ is the number of bits in queue i at time t. The average modulation efficiency $\overline{\mu}_{\mathrm{rtPS}}(t)$ is defined as the number of bits carried per symbol over a sliding time window t_c. Then the estimated number of time slots for rtPS class can be expressed as follows:

$$E_{\mathrm{rtPS}} = \alpha(t) \cdot \frac{B_{\mathrm{rtPS}}(t)}{\overline{\mu}_{\mathrm{rtPS}}(t)} \tag{4}$$

where $\alpha(t)$ is a QoS-aware heuristic control parameter that is updated on a frame by frame basis to adapt to the QoS satisfaction of the class scheduler. The basic idea in adjusting $\alpha(t)$ is that when the class scheduler experience good QoS satisfaction, the value of $\alpha(t)$ is decreased to save the bandwidth for other classes. Otherwise, the value of $\alpha(t)$ is increased to guarantee the required QoS. Towards this end, an exponentially smoothed curve is applied to adjust the value of $\alpha(t)$. The adjustment, which is $|\Delta\alpha(t)| = |\alpha(t) - \alpha(t-1)|$, is small if the QoS outage probability is around a predefined threshold. Otherwise, $|\Delta\alpha(t)|$ is exponentially increased as either to increase or reduce the allocated bandwidth to the class scheduler. The calculation of $\Delta\alpha(t)$ is specified as follows:

$$\Delta\alpha(t) = \begin{cases} \xi \cdot \frac{\exp(\beta \cdot d(t)) - 1}{\exp(\beta \cdot D_{\max}) - 1} & \text{if } \mathrm{P_r}(t) \geq T_h \\ -\xi \cdot \frac{\exp(\beta \cdot d(t)) - 1}{\exp(\beta \cdot D_{\max}) - 1} & \text{if } \mathrm{P_r}(t) < T_h \end{cases} \tag{5}$$

where $d(t) = \min\{|\mathrm{P_r}(t) - T_h|, D_{\max}\}$, $\mathrm{P_r}(t)$ is the delay outage probability at time t, defined as the proportion of packets with delay exceeding the maximum allowable delay $T_{\max}$ within a certain time window, T_h is the outage threshold, $D_{\max}$ is the truncated maximum value of $d(t)$, β is a shape factor which is used to tune the adaptation degree, and ξ is the maximum value of $\Delta\alpha(t)$. Term $(\exp(\beta \cdot d(t)) - 1)/(\exp(\beta \cdot D_{\max}) - 1)$ is a normalized utility function of $(\mathrm{P_r}(t) - T_h)$. When $\mathrm{P_r}(t)$ is close to T_h, the normalized value is close to zero; when $|\mathrm{P_r}(t) - T_h|$ is large, it increases exponentially to one. In real implementation, we set a maximum and minimum value of $\alpha(t)$ to optimize the performance. In general, the bandwidth estimation procedure for rtPS class is as follows:

- **Step 1:** At each scheduling instant, calculate the backlogged traffic $B_{\mathrm{rtPS}}(t)$, the average modulation efficiency $\overline{\mu}_{\mathrm{rtPS}}(t)$, and the delay outage probability $\mathrm{P_r}(t)$. Update the value of $\alpha(t)$:

$$\alpha(t) = \begin{cases} \min\{\alpha(t-1) + \Delta\alpha(t), \alpha_{\max}\} & \text{if } \mathrm{P_r}(t) \geq T_h \\ \max\{\alpha(t-1) + \Delta\alpha(t), \alpha_{\min}\} & \text{if } \mathrm{P_r}(t) < T_h \end{cases} \tag{6}$$

  where $\Delta\alpha(t)$ is specified in Exp. (5).
- **Step 2:** Calculate the estimated bandwidth for rtPS class according to Exp. (4).

For nrtPS class, the bandwidth estimation procedure is the same as rtPS class, except the definition of the backlogged traffic and the outage probability. Packets in nrtPS can tolerate longer delays, but need QoS guarantees in terms of the minimum throughput. Hence we use the total number of virtual tokens associated with each queue $V_{\mathrm{nrtPS}} = \sum_{i \in \{\mathrm{nrtPS}\}} v_i(t)$ as the backlogged traffic, where $v_i(t)$ is the number of virtual tokens in bucket i at time t. $\mathrm{P_r}(t)$ in nrtPS is defined as the probability that the average throughput is less than the predefined minimum throughput within a certain time window.

For BE class, as there is no QoS guarantees, after serving UGS, rtPS, and nrtPS classes, the residual bandwidth is allocated to BE class.

After the aggregate scheduler calculates the estimated amount of bandwidth for rtPS and nrtPS classes, it checks the remaining bandwidth. If the remaining bandwidth is larger than the estimated sum of rtPS and nrtPS, the aggregate scheduler allocates E_{rtPS} and E_{nrtPS} to rtPS and nrtPS classes respectively. Then the residual bandwidth is distributed to each service class proportional to their queue size Q_{rtPS}, Q_{nrtPS}, and Q_{BE}. Otherwise, if the remaining bandwidth is smaller than the estimated sum of rtPS and nrtPS, the aggregate scheduler first allocates E_{rtPS} to rtPS class, the residual bandwidth is allocated to nrtPS class. A detailed description of the proposed algorithm is listed in pseudocode 1.

III. Simulation Model

To evaluate the performance of the proposed resource allocation algorithm with other conventional algorithms, a system-level simulation is performed in OPNET.

A. System Model

In this paper, we consider the downlink of a single-cell IEEE 802.16 OFDM/TDD system with cell radius of 2 km, where subscriber stations (SSs) are randomly placed in the cell with uniform distribution, and move with a speed of 3 km/h in a random direction. The duration of a frame is set to be 1 ms as recommended by the standard so that the channel quality of each connection almost remains constant per frame, but is

Algorithm 1 Adaptive Radio Resource Allocation in the Aggregate Scheduler

```
Set initial $N_{\text{total}}$ in each round
$N_{\text{UGS}} \leftarrow \sum_{i \in \{\text{UGS}\}} d_i$
$N_{\text{rest}} \leftarrow N_{\text{total}} - N_{\text{UGS}}$
**if** $N_{\text{rest}} > 0$ **then**
    Update the heuristic value $\alpha_{\text{rtPS}}(t)$ by Exp. (6)
    Estimate the number of time slots allocated to rtPS class scheduler $E_{\text{rtPS}}$ by Exp. (4)
    Update the heuristic value $\alpha_{\text{nrtPS}}(t)$ by Exp. (6)
    Estimate the number of time slots allocated to nrtPS class scheduler $E_{\text{nrtPS}}$ by Exp. (4)
    **if** $N_{\text{rest}} \geq (E_{\text{rtPS}} + E_{\text{nrtPS}})$ **then**
        $N_{\text{rest}} \leftarrow N_{\text{rest}} - E_{\text{rtPS}} - E_{\text{nrtPS}}$
        $N_{\text{rtPS}} \leftarrow E_{\text{rtPS}} + N_{\text{rest}} \cdot \frac{Q_{\text{rtPS}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
        $N_{\text{nrtPS}} \leftarrow E_{\text{nrtPS}} + N_{\text{rest}} \cdot \frac{Q_{\text{nrtPS}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
        $N_{\text{BE}} \leftarrow N_{\text{rest}} \cdot \frac{Q_{\text{BE}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
    **else**
        $N_{\text{rtPS}} \leftarrow \min\{E_{\text{rtPS}}, N_{\text{rest}}\}$
        $N_{\text{nrtPS}} \leftarrow N_{\text{rest}} - N_{\text{rtPS}}$
    **end if**
**end if**
```

Parameters	Value
System	OFDM/TDD, TDM
Central frequency	3500 MHz
Channel bandwidth	10 MHz
Physical slots (downlink)	2000 PS/frame
User distribution	Uniform
User speed	3 km/h in random direction
Beam pattern	Omni-directional
Cell radius	2 km
Frame duration	1 ms
BS transmit power	10 W
Pass loss model	Okumura-Hata model
Large-scale shadowing	Log-normal distribution with mean: 0, standard deviation: 8 dB
Maximum MAC PDU size	56 bytes

TABLE I

A SUMMARY OF SIMULATION PARAMETERS FOR SYSTEM MODEL

Modulation scheme	Coding rate	bits/symbol	Target SNR for 1% PER (dB)
BPSK	1/2	0.5	1.5
QPSK	1/2	1	6.4
QPSK	3/4	1.5	8.2
16QAM	1/2	2	13.4
16QAM	3/4	3	16.2
64QAM	1/2	4	21.7
64QAM	3/4	4.5	24.4

TABLE II

MODULATION AND CODING SCHEMES FOR 802.16 [10]

allowed to vary from frame to frame [11]. The propagation model consists of path loss and large-scale shadowing. Path loss is modeled according to the Okumura-Hata model. Large-scale shadowing is modeled by log-normal distribution with zero mean and a standard deviation of 8 dB.

Table I summarizes the system parameters used in the simulation. We assume that all packets are transmitted and received without errors and the transmission delay is negligible. We also assume that the BS has perfect knowledge of channel state information (CSI). The modulation order and coding rate is determined by the instantaneous SNR. We follow the AMC table shown in Table II, which specifies the minimum SNR required to meet a target frame error rate, e.g., 1%.

B. Traffic Model

In the simulation, three types of traffic streams are generated: VoIP, videoconference, and internet traffic. VoIP and videoconference are served in UGS class and rtPS class, respectively. Internet traffic is served in nrtPS class and BE class. Each user generates one or several traffic types independently. VoIP traffic is modeled as a two-state Markov ON/OFF source [8]. A videoconference consists of a VoIP and a video source [8]. Internet traffic can be web browsing that requires large bandwidth and variable size bursty data. We apply the WWW browsing model [9]. A summary of traffic parameters for different traffic types are listed in Table III.

IV. PERFORMANCE EVALUATION

A. Performance Metrics

Since the performance of fixed bandwidth allocation for UGS connections is well defined by the standard and BE connections do not have any specific QoS requirements, here we only focus on the performance evaluation of rtPS and nrtPS connections. For rtPS service, the *average packet delay* and the *delay outage probability* are the main performance metrics, with QoS requirements of packet delay < 100 ms and outage probability $< 5\%$. For nrtPS service, the *average throughput* and the *throughput outage probability* are the main performance metrics, with QoS requirements of throughput ≥ 100 Kbits/sec and outage probability $< 5\%$. In order to evaluate the spectral efficiency, the *modulation efficiency* in each class scheduler is also evaluated.

B. Results and Discussion

The performance of the proposed adaptive resource allocation algorithm is compared with that of conventional schemes, e.g., strict priority-based scheme and static scheme. Some of the parameters used in the adaptive scheme are set as follows: $\beta = 80$, $T_h = 0.03$, $\xi = 0.05$, $D_{\max} = 0.1$, $\alpha_{\min} = 0.1$, and $\alpha_{\max} = 0.4$. In the static scheme, the proportions of the total available bandwidth allocated to rtPS, nrtPS, and BE classes are 40%, 40%, and 20% respectively.

Fig. 2 shows the average packet delay in rtPS of three bandwidth allocation schemes. The average packet delay of the proposed adaptive scheme and the priority-based scheme remains almost constant regardless of the number of users in the system, while in the static scheme, the average packet delay increases sharply when the number of users is above 45. Similar phenomenon can be observed for the delay outage

Type	Characteristics	Distribution	Parameters
VoIP	ON period	Exponential	Mean = 1.34 sec
VoIP	OFF period	Exponential	Mean = 1.67 sec
VoIP	Packet size	Constant	66 bytes
VoIP	Inter-arrival time between packets	Constant	20 ms
Video	Packet size	Log-normal	Mean = 4.9 bytes Std. dev. = 0.75 bytes
Video	Inter-arrival time between packets	Normal	Mean = 33 ms Std. dev. = 10 ms
Web	Reading time between sessions	Exponential	Mean = 5 sec
Web	Number of packets within a packet call	Geometric	Mean = 25 packets
Web	Inter-arrival time between packets	Geometric	Mean = 0.0277 sec
Web	Packet size	Truncated Pareto	k = 81.5 bytes α = 1.1 m = 2 M bytes

TABLE III

A SUMMARY OF TRAFFIC PARAMETERS

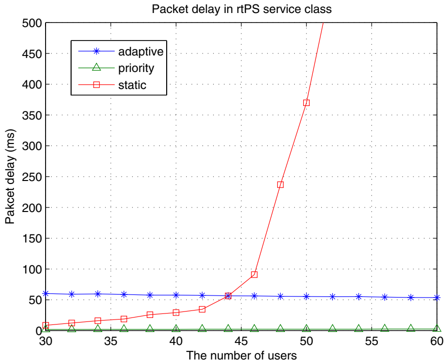

Fig. 2. Average packet delay in rtPS

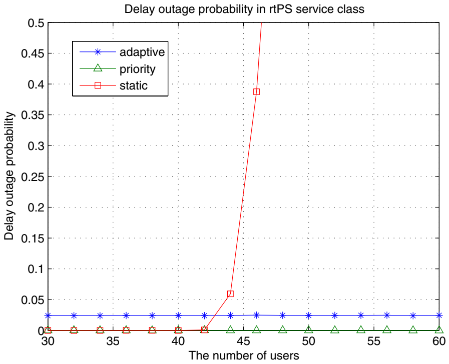

Fig. 3. Delay outage probability in rtPS

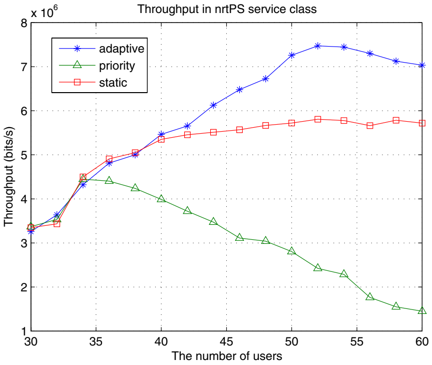

Fig. 4. Average throughput in nrtPS

probability shown in Fig. 3. From Fig. 2 & 3, we can obviously see that both the priority-based scheme and dynamic scheme can meet the QoS requirement in rtPS. On the other hand, the performance of the static scheme can not adapt to the traffic load, thus is not suitable for load varying systems. The advantage of the adaptive scheme over the priority-based scheme is depicted in Fig. 6, which shows the spectral efficiency of different bandwidth allocation schemes. For rtPS, it is shown that the spectral efficiency of the adaptive scheme is about two times than that of the priority scheme. This is achieved due to the reason that instead of allocating all the available bandwidth to rtPS in the priority scheme, the proposed algorithm adaptively allocates a ”necessary” amount of bandwidth to the class scheduler to keep its outage probability around a predefined threshold, which is 2.5% in our scenario, so that the performance of the class scheduler can be maximized. By doing so, the channel and QoS aware class scheduler has more chances to serve a user in a good channel state without sacrificing the QoS requirement (as we can see in Fig. 2 that the average delay in the adaptive scheme is higher than that of the priority-based scheme, but is well kept below a threshold), thus significantly increase the efficiency of bandwidth utilization.

Fig. 4 shows the throughput in nrtPS of three bandwidth allocation schemes. When the traffic load is light (the number of users in the system is less than 34), the throughput in all three schemes increases proportional to the number of users. After that point, the priority-based scheme experiences bandwidth starvation and the throughput is inverse proportional to the number of users. In the adaptive scheme, the throughput keeps increasing proportional to the number of users when there are less than 52 users in the system. After that point, the throughput decreases as the number of users increases. This is because in the adaptive scheme, the aggregate scheduler tries to balance the bandwidth distribution among different class schedulers so as to increase the spectral efficiency while satisfying the QoS requirements. If the system is unsaturated, e.g., the total available bandwidth is larger than the estimated sum (the number of users in less than 52), the aggregate scheduler allocates the estimated bandwidth to each class scheduler, thus the throughput in each service class increases in proportion to the number of users. If saturation occurs, e.g., the total available bandwidth is not enough to serve all classes

1930-529X/07/$25.00 © 2007 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

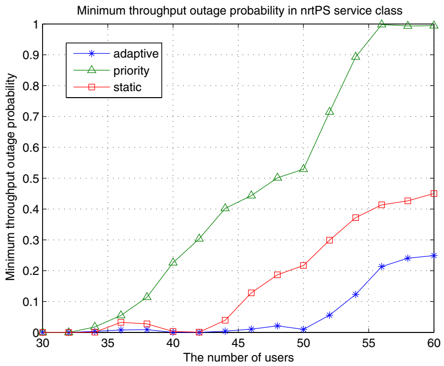

Fig. 5. Throughput outage probability in nrtPS

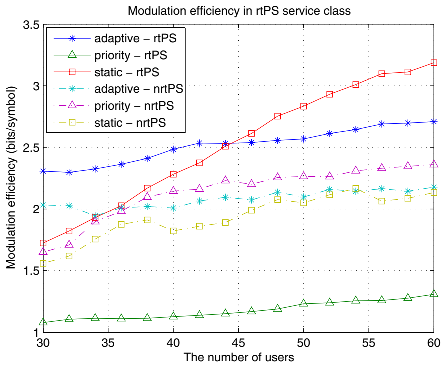

Fig. 6. Modulation efficiency in rtPS and nrtPS

(the number of users in larger than 52), the aggregate scheduler allocates the estimated bandwidth to the class scheduler from high priority to low priority, thus the throughput in service class with low priority decreases as the bandwidth balancing scheme in the aggregate scheduler favors service class with high priority when congestion occurs. However, in the priority-based scheme, nrtPS class experiences severe bandwidth starvation due to the reason that much of the bandwidth allocated to rtPS class is utilized of low spectral efficiency, thus the residual bandwidth allocated to nrtPS class is not sufficient to serve nrtPS connections. While in the static scheme, as the amount of bandwidth allocated to each class is fixed, the throughput remains on a steady level regardless of the number of uses as expected. Fig. 5 shows the throughput outage probability in nrtPS of three bandwidth allocation schemes. It is obvious that the proposed adaptive scheme outperforms over the other two schemes. The number of supportable users under a predefined 5% outage probability in priority-based, static and adaptive scheme are 36, 44 and 52 respectively. From Fig. 6, we notice that for nrtPS, the spectral efficiency of the adaptive scheme is lower than the priority-based scheme when the number of users is greater than 34. This is because in the priority-based scheme, the traffic in nrtPS class starts experiencing congestion when there are more than 34 users, which means that the nrtPS class scheduler in the priority-based scheme can't get enough bandwidth as in the adaptive scheme, that in turn results a higher modulation efficiency due to the scheduling mechanism of EXP algorithm.

V. Conclusions

In this paper, an adaptive resource allocation algorithm of the aggregate scheduler in two-level hierarchical QoS scheduling for IEEE 802.16 systems is proposed to increase the spectral efficiency while satisfying the diverse QoS requirements in each service class. The proposed algorithm takes the backlogged traffic, the modulation efficiency, as well as the QoS satisfaction into account when estimating the amount of bandwidths required in each service class. Through system-level simulation, the performance of the proposed algorithm is evaluated in terms of packet delay, throughput, outage probability and modulation efficiency in rtPS and nrtPS service classes. It is shown that the overall performance of the proposed algorithm can be significantly improved compared with the conventional schemes in terms of the maximum number of supportable users under a predetermined outage probability. As the bandwidth allocation module and the packet scheduling module are loosely separated in our scheduler, the design of the class scheduler is independent to the aggregate scheduler. Our proposed algorithm of the aggregate scheduler is aware of the performance of the class scheduler, thus can support and adapt to various scheduling algorithms in the class scheduler.

References

[1] Kitti W. and Aura G.: *Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems*, International Journal of Communication Systems, Vol.16 Issue.1, pp. 81–96, 2003.

[2] Woo J.K., Joo Y.B., Sun D.L., Young J.S., Yun S.K., and Jin A K.: *Efficient Uplink Scheduler Architecture of Subscriber Station in IEEE 802.16 System*, Lecture Notes in Computer Science, Vol.3823, pp. 734–743, 2005.

[3] Sanjay S., and Alexander L.S.: *Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR*, Proceedings of International Teletraffic Congress (ITC), 2001.

[4] A. Jalali, R. Padovani, and R. Pankaj: *Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System*, Vehicular Technology Conference Proceedings, Vol.3, pp. 1854–1858, 2000.

[5] Matthew A., Krishnan K., Kavita R., Alexander L. S., and Phil W.: *Providing Quality of Service over a Shared Wireless Link*, IEEE Communications Magazine, Vol.39 Issue.2, pp. 150–154, 2001.

[6] Qingwen L., Xin W., and Georgios B.G.: *Cross-Layer Scheduler Design with QoS support for Wireless Access Networks*, 2nd International Conference on QoS in Heterogeneous Wired/Wireless Networks, 2005.

[7] Muhammad K., and Niclas W.: *Scheduling Algorithms for HS-DSCH in a WCDMA Mixed Traffic Scenario*, PIMRC, Vol.2, 2003 Sep.

[8] Claudio C., Luciano L. and Enzo M.: *Quality of Service Support in IEEE 802.16 Networks*, IEEE Network, Vol.20 Issue.2, pp. 50–55, 2006.

[9] Dong H. K., Byung H. R., and Chung G. K., *Packet Scheduling Algorithm Considering a Minimum Bit Rate for Non-real-time Traffic in an OFDMA/FDD-Based Mobile Internet Access System*, ETRI Journal, Vol.26, Issue.1, pp. 48–52, 2004

[10] Christian H.: *Analysis and performance evaluation of the OFDM-based metropolitan area networks IEEE 802.16*, Computer Networks, Vol.49 Issue.3, pp. 341–363, 2005.

[11] IEEE 802.16-2004, *IEEE standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, 2004

1930-529X/07/$25.00 © 2007 IEEE