Published in IET Communications
Received on 2nd January 2010
Revised on 8th April 2010
doi: 10.1049/iet-com.2010.0005

ISSN 1751-8628

Adaptive bandwidth allocation model for multiple traffic classes in IEEE 802.16 worldwide interoperability for microwave access networks

T.-L. Sheu   K.-C. Huang

Department of Electrical Engineering, Center for Wireless Multimedia Communications, National Sun Yat-Sen University, Kaohsiung, Taiwan
E-mail: sheu@ee.nsysu.edu.tw

Abstract: This study presents an adaptive bandwidth allocation model for multiple classes in wireless broadband network, worldwide interoperability for microwave access (WiMAX) (IEEE 802.16). Data to and from Internet are all conveyed through WiMAX links to its final destination. In order to promise the quality of real-time traffic and allow more transmission opportunity for other traffic types, the proposed adaptive bandwidth allocation (ABA) algorithm would first reserve the unsolicited bandwidth for constant-bit-rate traffic (unsolicited grant services). Then, polling bandwidth is allocated for real-time traffic (rtPS) to meet their end-to-end delay constraints and for non-real-time traffic (nrtPS) to meet their minimum throughput requirements. One of the novelties presented by this study is right in that ABA does not greedily overtake too much bandwidth from the lowest-priority class, best effort (BE). Instead, it is very intelligent to only meet the delay constraint of rtPS and the minimum throughput requirement of nrtPS, while it endeavours to avoid any possible starvation of BE traffic. For the purpose of performance evaluation, a four-dimensional Markov chains is built to analyse the proposed ABA. The analytical results are validated by a simulation. Finally, from the comparison with a previous work, the authors observe the performance superiority of the ABA in satisfying the delay constraints (for rtPS), meeting the minimum throughput requirements (for nrtPS) and reducing the average packet drop ratio (for BE), when traffic environments are varied.

1 Introduction

The fast development of wireless communications and networking technologies has greatly improved the mobile and roaming capability of handset devices, such as personal data acquisition (PDA) and cellular phones. As a result, the popularity of multimedia applications on handset devices is overwhelming. Worldwide interoperability for microwave access (WiMAX), first standardised in 2004 known as IEEE 802.16, can provide broadband communications over wireless for various types of multimedia traffic, such as video streaming, VoIP, FTP etc. In comparison to wireless fidelity (WiFi) (IEEE 802.11), WiMAX not only offers larger and dedicated bandwidth, but it also provides different traffic types with different priority classes for meeting individual quality of service (QoS) requirements.

Among the different traffic classes, unsolicited grant services (UGS) is for transferring delay-sensitive, constant-bit-rate (CBR) traffic, such as VoIP, E1/T1 and ATM CBR. Hence, it is unsolicited to allocate fixed bandwidth for UGS traffic during their data transfer. Second to UGS, real-time polling services (rtPS) traffic is polled for sufficient bandwidth to meet end-to-end delay constraint; examples of rtPS are real-time, variable-bit-rate traffic, such as video streaming. Without specifying any delay requirements, on the other hand, non-real-time polling services (nrtPS) traffic is polled for sustaining a minimum throughput; examples of nrtPS are conventional Internet traffic, such as FTP, WWW etc. The last and also the least important class is best effort (BE). BE traffics do not possess any delay or throughput requirements except that they have to maintain a minimum threshold of packet loss ratio for preventing any possible starvation.

Bandwidth allocation (BA) operated by a base station (BS) in a WiMAX network can be divided into two categories, request-based and polling-based. Previous works in request-based BA tried to adjust contention window to reduce contention probability among different requests [1–3]. Polling-based BA, on the other hand, has been focused on two research subjects, connection-oriented and packet-oriented. In connection-oriented research, call admission control (CAC) is developed for different traffic classes to meet different QoS requirements [4–6]. Based on CAC, Rong *et al.* [7, 8] built a performance model by separating uplink and downlink paths in a WiMAX network so that the model becomes two independent optimisation problems. In packet-oriented research, most previous works investigated the methodology of how to effectively allocate bandwidth to different traffic classes. For example, in Hou *et al.* [9] and Sengupta *et al.* [10], they first allocate bandwidth for traffic classes to meet their minimum QoS requirements. If there is remaining bandwidth, then it will be equally allocated to the classes that still have traffic in the queues. Niyato and Hossaiin [11] and Jeon and Jeong [12] developed a model consisting of two levels, CAC and packet scheduling, with which different QoS requirements can be satisfied more easily. In Cicconetti *et al.* [13, 14] backlog records were used to reduce unnecessary polling. Other previous works [15–17] in contrast to the above-mentioned papers, are more straightforward; they simply allocate fixed bandwidth to UGS and the remaining bandwidth to the rest of traffic classes. Finally, Niyato and Hossain [18] presented a Q-Aware model where they purposely control arrival rates to meet different levels of QoS requirements. Specifically, except for UGS, traffic of polling services (rtPS and nrtPS) is rigorously restricted based on the following equations

$$\lambda_{\min} = \lambda_0/2 \tag{1}$$

$$\Gamma(\lambda_0, x) = \lambda_0 - \frac{\lambda_0(x - \sigma_{\min})}{2(\sigma_{\max} - \sigma_{\min})} \tag{2}$$

$$\lambda(x, \lambda_0, \lambda_{\min}) = \begin{cases} \lambda_0, & x \le \sigma_{\min} \\ \Gamma(\lambda_0, x), & \sigma_{\min} < x < \sigma_{\max} \\ \lambda_{\min}, & \sigma_{\max} \le x \end{cases} \tag{3}$$

where $\lambda_{\min}$ is the minimum guaranteed amount for polling service traffic which is restricted to be one-half of λ_0, the data generation rate. $\sigma_{\min}$ and $\sigma_{\max}$, respectively, represent the minimum and the maximum buffer threshold and x is the queue length. In short, the Q-Aware model applies CAC first and then schedules the departure rates based on different priorities of traffic classes.

Unlike the previous works, in this paper, we consider a WiMAX network that delivers traffic not only for its subscriber station (SS) but also for the traffic from the attached Internet. Thus, it is highly possible that all the arriving traffic may not be accommodated in a WiMAX frame. The proposed adaptive bandwidth allocation (ABA) is different from a previous work, the Q-Aware [18], where BA is not based on the actual QoS requests but simply based on the variations of queue length. In other words, when network is overloaded, low-priority traffic, such as BE, is very easy to obtain starvation in the Q-Aware. However, in ABA, bandwidth is allocated based on the actual bandwidth requests. That is, after reserving a portion of bandwidth for UGS traffic, rtPS and nrtPS traffic are polled to meet their end-to-end delay constraints and minimum throughput requirements, respectively. For the purpose of performance evaluation, we build four-dimensional (4-D) Markov chains to analyse the proposed ABA. The analytical results are validated by a simulation. From the comparison with a previous work, the Q-Aware, we observe the superiority of the ABA in satisfying the delay constraints (for rtPS), meeting the minimum throughput requirements (for nrtPS) and reducing the average packet drop ratio (PDR) (for BE), when traffic environments are varied.

The remainder of this paper is organised as follows. In Section 2, we introduce the ABA algorithms for an Internet-attached WiMAX network. In Section 3, a performance evaluation model with 4-D Markov chains is built. Section 4 presents the analytical and the simulation results. Finally, we give concluding remarks in Section 5.

2 Adaptive bandwidth allocation

In a WiMAX network, four classes of queue, UGS, rtPS, nrtPS and BE are specified. Without adequately scheduling the traffic flows, QoS of higher-priority traffic, such as UGS and rtPS, may be seriously affected by non-real-time traffic, such as nrtPS and BE. In the Q-Aware model [18], BS first examines the queue status and then dynamically assigns bandwidth to the four types of traffic. This approach can certainly guarantee QoS for higher-priority traffic, but it may over-assign bandwidth to real-time traffic and accidentally starve the non-real-time traffic.

2.1 Internet-attached WiMAX networks

Unlike the previous work, a WiMAX network in our study consists of a BS, a number of SS, and an attached Internet. As shown in Fig. 1, since a BS in this environment may receive traffic simultaneously from SS (wireless) and the attached Internet (wired), it is highly possible that the total arriving traffic cannot be accommodated in a single WiMAX frame. Thus, an adaptive BA model is required. Fig. 2 shows the traffic aggregation for the four classes of queue within a BS. Traffic of the same type from wired

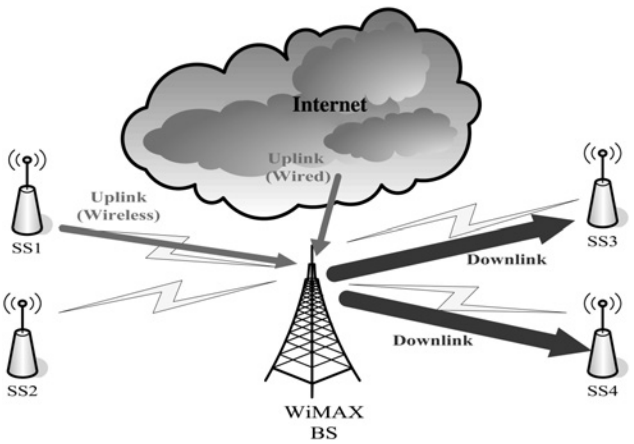

Figure 1 *WiMAX network with Internet backbone*

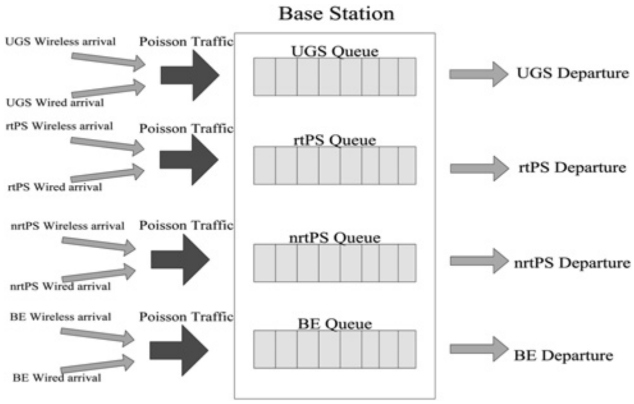

Figure 2 *Aggregation of four uplink traffic classes in BS queues*

and wireless network is aggregated into a single Poisson arrival process.

2.2 Proposed ABA algorithm

In the proposed ABA scheme, we first assign a fixed departure rate to UGS to fulfil its stringent delay requirements. One of the novelties in the ABA is to guarantee a specified delay constraint for rtPS traffic and provide minimum throughput requirement for nrtPS traffic. For the least important traffic class (BE), ABA will prevent them from starvation. Fig. 3 shows the proposed ABA algorithm. First, the ABA assigns initial bandwidth, $r^{(\mathrm{UGS})}$, $r^{(\mathrm{rtPS})}$, $r^{(\mathrm{nrtPS})}$ and $r^{(\mathrm{BE})}$, based on the requested bandwidth of UGS, the required minimum bandwidth of rtPS and nrtPS, and the queue length of BE traffic, respectively. If remaining bandwidth (A) exists, the ABA then assigns extra bandwidth, $a^{(\mathrm{rtPS})}$, $a^{(\mathrm{nrtPS})}$ and $a^{(\mathrm{BE})}$, to rtPS, nrtPS and BE, respectively, if any of the traffic classes have packets still waiting in the queues.

Through the ABA, the final allocated bandwidth for the four different traffic classes becomes $\mu^{(\mathrm{UGS})}$, $\mu^{(\mathrm{rtPS})}$, $\mu^{(\mathrm{nrtPS})}$ and $\mu^{(\mathrm{BE})}$. Note that $\mu^{(\mathrm{UGS})} + \mu^{(\mathrm{rtPS})} + \mu^{(\mathrm{nrtPS})} + \mu^{(\mathrm{BE})} \le \mu$, where μ is the upper bound of WiMAX departure rate.

//allocation of initial bandwidth
$r^{(UGS)} \leftarrow$ required(UGS); // the requested bandwidth of UGS
$r^{(rtPS)} \leftarrow$ Min(rtPS); // the required minimum bandwidth of rtPS
$r^{(nrtPS)} \leftarrow$ Min(nrtPS); // the required minimum bandwidth of nrtPS
If (BE's Queue is full)
$r^{(BE)} \leftarrow$ Min(BE); // the minimum bandwidth for BE to avoid starvation
Else
$r^{(BE)} \leftarrow 0$;
End If
$A \leftarrow \mu - r^{(UGS)} - r^{(rtPS)} - r^{(nrtPS)} - r^{(BE)}$; // calculate the remaining bandwidth
// allocation of the extra bandwidth
$a^{(rtPS)} \leftarrow$ Min(rtPS packets in queue per frame time, A);
$a^{(nrtPS)} \leftarrow$ Min(nrtPS packets in queue per frame time, $A - a^{(rtPS)}$);
$a^{(BE)} \leftarrow$ Min(BE packets in queue per frame time, $A - a^{(rtPS)} - a^{(nrtPS)}$);
// sum of the initial allocated and the extra bandwidth
$\mu^{(UGS)} = r^{(UGS)}$;
$\mu^{(rtPS)} = r^{(rtPS)} + a^{(rtPS)}$;
$\mu^{(nrtPS)} = r^{(nrtPS)} + a^{(nrtPS)}$;
$\mu^{(BE)} = r^{(BE)} + a^{(BE)}$;

Figure 3 *Proposed ABA algorithm*

2.3 Initial bandwidth estimation

The initial bandwidth allocated to UGS is simply the requested bandwidth, that is, $\mu^{(\mathrm{UGS})} = r^{(\mathrm{UGS})}$. Yet, for rtPS, nrtPS and BE traffic, we need to estimate their required initial bandwidth. The throughput of rtPS queue ($\mathrm{Thr}^{(\mathrm{rtPS})}$) can be calculated from (4), where $\lambda^{(\mathrm{rtPS})}$ is the mean packet arrivals in a single WiMAX frame time (in reality, it is about 1–2 ms) and $P_{\mathrm{drop}}^{(\mathrm{rtPS})}$ is the packet drop ratio. The delay time of rtPS queue (i.e. $\mathrm{DT}^{(\mathrm{rtPS})}$), which should be smaller than the rtPS allowable delay constraint (τ), can then be calculated from (5), where i represents that there are exactly i packets waiting in the queue during a single WiMAX frame time, and $r^{(\mathrm{rtPS})}$ denotes the mean packet departure of rtPS in a single WiMAX frame time. Specifically, speaking, (5) computes the delay time of rtPS queue by considering the packet arrivals, the packets waiting in the queue, and the packet departures.

$$\mathrm{Thr}^{(\mathrm{rtPS})} = (1 - P_{\mathrm{drop}}^{(\mathrm{rtPS})}) \times \lambda^{(\mathrm{rtPS})} \tag{4}$$

$$\mathrm{DT}^{(\mathrm{rtPS})} = \frac{\lambda^{(\mathrm{rtPS})} + i - r^{(\mathrm{rtPS})}}{\mathrm{Thr}^{(\mathrm{rtPS})}} \le \tau \tag{5}$$

Since the throughput will be limited by the required minimum bandwidth, (5) can be rewritten to (6). Finally, the initial bandwidth allocated to rtPS class is shown in (7).

$$\mathrm{DT}^{(\mathrm{rtPS})} = \frac{\lambda^{(\mathrm{rtPS})} + i - r^{(\mathrm{rtPS})}}{r^{(\mathrm{rtPS})}} \le \tau \tag{6}$$

$$r^{(\mathrm{rtPS})} = \mathrm{Min}\left(\mathrm{Min}\left(\frac{i + \lambda^{(\mathrm{rtPS})}}{1 + \tau},\ i\right),\ \mu - r^{(\mathrm{UGS})}\right) \tag{7}$$

Similarly, the throughput of nrtPS can be expressed in (9) based on the number of packet dropping ($N_{\mathrm{drop}}^{(\mathrm{nrtPS})}$) computed from (8). Thus, the initial bandwidth allocated

IET Commun., 2011, Vol. 5, Iss. 1, pp. 90–98
doi: 10.1049/iet-com.2010.0005

to nrtPS and BE class are calculated from (10) and (11), respectively. Note that from (8) to (11), we assume that buffer sizes of different traffic classes are the same and they are all equal to b. Also, j represents the number of packets waiting in the queue during a WiMAX frame time and α is the minimum throughput requirement provided for nrtPS traffic. Equation (10) computes the initial bandwidth of nrtPS in the following way. The Min of the inner parentheses takes buffer size, b, as the lower bound in case of buffer overflow. The Max of the middle parentheses uses zero to guarantee no negative number. The Min of the outer parentheses uses $\mu - r^{(\mathrm{UGS})} - r^{(\mathrm{rtPS})}$ as the lower bound to confine the request to the remaining bandwidth

$$N_{\mathrm{drop}}^{(\mathrm{nrtPS})} = \begin{cases} \lambda^{(\mathrm{nrtPS})} + j - \gamma^{(\mathrm{nrtPS})} & \text{if}(\lambda^{(\mathrm{nrtPS})} + j - \gamma^{(\mathrm{nrtPS})} > b) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\begin{aligned} \mathrm{Thr}^{(\mathrm{nrtPS})} &= (1 - P_{\mathrm{drop}}^{(\mathrm{nrtPS})}) \times \lambda^{(\mathrm{nrtPS})} \\ &= \left(1 - \frac{N_{\mathrm{drop}}^{(\mathrm{nrtPS})}}{\lambda^{(\mathrm{nrtPS})}}\right) \times \lambda^{(\mathrm{nrtPS})} \geq \alpha \end{aligned} \tag{9}$$

$$\begin{aligned} r^{(\mathrm{nrtPS})} = {} & \mathrm{Min}(\mathrm{Max}(\mathrm{Min}(\lambda^{(\mathrm{nrtPS})} + j - \alpha,\ b)\ ,0), \\ & \mu - r^{(\mathrm{UGS})} - r^{(\mathrm{rtPS})}) \end{aligned} \tag{10}$$

$$r^{(\mathrm{BE})} = \begin{cases} \mu - \gamma^{(\mathrm{UGS})} - \gamma^{(\mathrm{rtPS})} - \gamma^{(\mathrm{nrtPS})} & \text{if } k = b \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

3 Mathematical model

The scenario under our study is a WiMAX network that delivers four traffic classes, UGS, rtPS, nrtPS and BE, not only for its SS but also for the attached Internet. Since it is highly possible that all the arriving traffic may not be accommodated in a WiMAX frame, the ABA algorithm is developed to adaptively allocate the bandwidth. The performance of the proposed ABA is evaluated through 4-D Markov model. Each state in the Markov model is represented by four parameters as shown in Fig. 4.

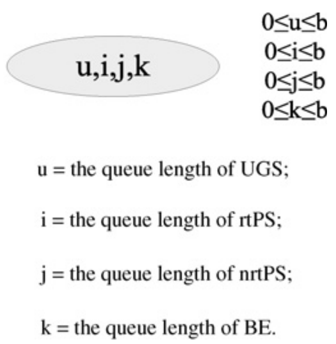

Figure 4 *Parameters of a 4-D Markovian state*

3.1 Probabilities of state transition

A WiMAX frame consists of two phases, uplink and down links. Hence, a transition between current state and next state in the 4-D Markov model must be converted to downlink and uplink transitions as shown in Fig. 5, where the four parameters in the intermediate states are represented by (u', i', j', k'). In the traffic model, we assume UGS traffic is CBR at on−off basis; it generates N packets at on-state and 0 packets at off-state during a single WiMAX frame time. We also assume that rtPS, nrtPS and BE traffic are generated by Poisson process, respectively. Thus, the transition probabilities of intermediate states in the Markov model can be derived by investigating the following two different cases of queue lengths at BS. Note that in the following equations, b denotes the maximum number of packets that can be buffered per traffic class at BS.

Case 1: Queue length is less than buffer size.

$$P^{(\mathrm{UGS})}\{u' | u' + N\} = P_a^{(\mathrm{UGS})}$$

$$P^{(\mathrm{rtPS})}\{i' | i' + x\} = P\{n^{(\mathrm{rtPS})} = x\} = \frac{(\lambda^{(\mathrm{rtPS})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{rtPS})}}, \quad x = 0, 1, 2, \ldots, (b - i' - 1)$$

$$P^{(\mathrm{nrtPS})}\{j' | j' + x\} = P\{n^{(\mathrm{nrtPS})} = x\} = \frac{(\lambda^{(\mathrm{nrtPS})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{nrtPS})}}, \quad x = 0, 1, 2, \ldots, (b - j' - 1)$$

$$P^{(\mathrm{BE})}\{k' | k' + x\} = P\{n^{(\mathrm{BE})} = x\} = \frac{(\lambda^{(\mathrm{BE})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{BE})}}, \quad x = 0, 1, 2, \ldots, (b - k' - 1)$$

Case 2: Queue length exceeds buffer size.

$$P^{(\mathrm{UGS})}\{u' | u' + N\} = P_a^{(\mathrm{UGS})}$$

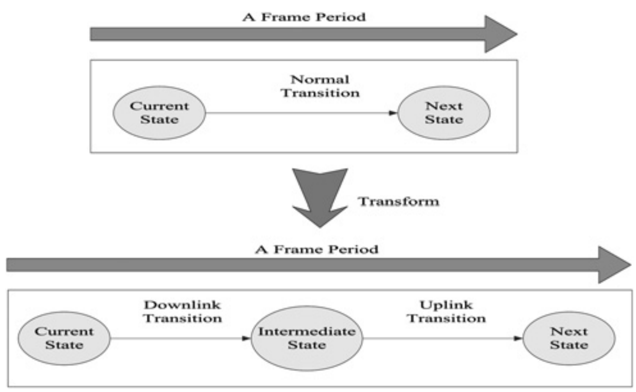

Figure 5 *State transition with intermediate states*

$$P^{(\mathrm{rtPS})}\{i'|b\} = \sum_{x=b-i}^{\infty} P\{n^{(\mathrm{rtPS})} = x\} = \sum_{x=b-i'}^{\infty} \frac{(\lambda^{(\mathrm{rtPS})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{rtPS})}}$$

$$P^{(\mathrm{nrtPS})}\{j'|b\} = \sum_{x=b-j}^{\infty} P\{n^{(\mathrm{nrtPS})} = x\} = \sum_{x=b-j'}^{\infty} \frac{(\lambda^{(\mathrm{nrtPS})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{nrtPS})}}$$

$$P^{(\mathrm{BE})}\{k'|b\} = \sum_{x=b-k}^{\infty} P\{n^{(\mathrm{BE})} = x\} = \sum_{x=b-k'}^{\infty} \frac{(\lambda^{(\mathrm{BE})})^x}{x!} \times \mathrm{e}^{-\lambda^{(\mathrm{BE})}}$$

As to down link, the current state (u, i, j, k) will be converted to intermediate state $(u - \mu^{(\mathrm{UGS})}, i - \mu^{(\mathrm{rtPS})}, j - \mu^{(\mathrm{nrtPS})}, k - \mu^{(\mathrm{BE})})$. Finally, by integrating the transition probabilities of uplinks and downlinks, we can derive the probability of state transition (from current state to next state), based on the assumption that the arrivals of the four traffic classes are mutually independent. The probability of state transition is denoted as $P_{(u,i,j,k),(U,I,J,K)}$ and it can be expressed as in (12). To reduce the size of the state-transition matrix as shown in Fig. 6, the UGS state is simplified into either 0 or 1, where 0 represents no packets are waiting in the UGS queue, and 1 denotes N packets are waiting in the UGS queue. Thus, from the state-transition matrix, we can derive the steady-state probabilities as shown in

$$\begin{aligned} P_{(u,i,j,k),(U,I,J,K)} &= P_{(u',i',j',k'),(U,I,J,K)} \\ &= P^{(\mathrm{UGS})}\{u'|U\} \times P^{(\mathrm{rtPS})}\{i'|I\} \\ &\quad \times P^{(\mathrm{nrtPS})}\{j'|J\} \times P^{(\mathrm{BE})}\{k'|K\} \end{aligned} \tag{12}$$

$$\pi_m = (\pi(0, 0, 0, 0),\ \pi(0, 0, 0, 1), \ldots,\ \pi(1, b, b, b)) \tag{13}$$

3.2 Performance metrics

3.2.1 Average delay time of rtPS ($DT^{(rtPS)}$): First, from the Markovian states and (13), we can calculate the average queue length in (14). Then, the dropping probability in (16) can be computed after we calculated the number of packet drops of rtPS traffic, $N_{\mathrm{drop}}^{(\mathrm{rtPS})}(i)$, in (15). Finally, the average delay time of rtPS traffic as shown

$$P = \begin{bmatrix} P_{(0,0,0,0)(0,0,0,0)} & \cdots & P_{(0,0,0,0)(1,0,0,0)} & \cdots & P_{(0,0,0,0)(1,b,0,0)} & \cdots & P_{(0,0,0,0)(1,b,b,0)} & \cdots & P_{(0,0,0,0)(1,b,b,b)} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{(1,0,0,0)(0,0,0,0)} & \cdots & P_{(1,0,0,0)(1,0,0,0)} & \cdots & & & & & \vdots \\ \vdots & \ddots & \vdots & \ddots & & & & & \vdots \\ P_{(1,b,0,0)(0,0,0,0)} & \cdots & P_{(1,b,0,0)(1,0,0,0)} & \cdots & & & & & \vdots \\ \vdots & \ddots & \vdots & \ddots & & & & & \vdots \\ P_{(1,b,b,0)(0,0,0,0)} & \cdots & P_{(1,b,b,0)(1,0,0,0)} & \cdots & & & & & \vdots \\ \vdots & \ddots & \vdots & \ddots & & & & \ddots & \vdots \\ P_{(1,b,b,b)(0,0,0,0)} & \cdots & P_{(1,b,b,b)(1,0,0,0)} & \cdots & \cdots & \cdots & \cdots & \cdots & P_{(1,b,b,b)(1,b,b,b)} \end{bmatrix}$$

Figure 6 *State transition matrix*

in (17) is very easy to derive simply by Little's law

$$QL^{(\mathrm{rtPS})} = \sum_{i=0}^{b}\left(i \times \left(\sum_{u=0}^{1}\sum_{j=0}^{b}\sum_{k=0}^{b} \pi(u, i, j, k)\right)\right) \tag{14}$$

$$\begin{aligned} N_{\mathrm{drop}}^{(\mathrm{rtPS})}(i) &= \lambda^{(\mathrm{rtPS})} - \sum_{n=0}^{b-(i-\mu^{(\mathrm{rtPS})})} (n \times P\{n^{(\mathrm{rtPS})} = n\}) \\ &\quad - (b - (i - \mu^{(\mathrm{rtPS})})) \\ &\quad \times \left(1 - \sum_{n=0}^{b-(i-\mu^{(\mathrm{rtPS})})} P\{n^{(\mathrm{rtPS})} = n\}\right) \end{aligned} \tag{15}$$

$$P_{\mathrm{drop}}^{(\mathrm{rtPS})} = \sum_{i=0}^{b}\sum_{u=0}^{1}\sum_{j=0}^{b}\sum_{k=0}^{b}\left(\pi(u, i, j, k) \times \frac{N_{\mathrm{drop}}^{(\mathrm{rtPS})}(i)}{\lambda^{(\mathrm{rtPS})}}\right) \tag{16}$$

$$DT^{(\mathrm{rtPS})} = \frac{QL^{(\mathrm{rtPS})}}{(1 - P_{\mathrm{drop}}^{(\mathrm{rtPS})})} \tag{17}$$

3.2.2 Average throughput of nrtPS ($Thr^{(nrtPS)}$): Similarly, the throughput of nrtPS traffic as shown in (20) can be derived by calculating the number of packet drops and the packet dropping probability as shown in (18) and (19), respectively

$$\begin{aligned} N_{\mathrm{drop}}^{(\mathrm{nrtPS})}(j) &= \lambda^{(\mathrm{nrtPS})} - \sum_{n=0}^{b-(j-\mu^{(\mathrm{nrtPS})})} (n \times P\{n^{(\mathrm{nrtPS})} = n\}) \\ &\quad - (b - (j - \mu^{(\mathrm{nrtPS})})) \\ &\quad \times \left(1 - \sum_{n=0}^{b-(j-\mu^{(\mathrm{nrtPS})})} P\{n^{(\mathrm{nrtPS})} = n\}\right) \end{aligned} \tag{18}$$

$$P_{\mathrm{drop}}^{(\mathrm{nrtPS})} = \sum_{j=0}^{b}\sum_{u=0}^{1}\sum_{i=0}^{b}\sum_{k=0}^{b}\left(\pi(u, i, j, k) \times \frac{N_{\mathrm{drop}}^{(\mathrm{nrtPS})}(j)}{\lambda^{(\mathrm{nrtPS})}}\right) \tag{19}$$

$$\mathrm{Thr}^{(\mathrm{nrtPS})} = (1 - P_{\mathrm{drop}}^{(\mathrm{nrtPS})}) \times \lambda^{(\mathrm{nrtPS})} \tag{20}$$

4 Analytical and simulation results

By running on the MATLAB, we can obtain the analytical results of the proposed ABA algorithm. The analytical results are compared with a previous work, the Q-Aware [18]. By referring to (1)–(3), in the Q-Aware model, we assume $\sigma_{\max}^{(\mathrm{rtPS})} = \lfloor(2 \times \lambda_0^{(\mathrm{rtPS})}/\tau)\rfloor$ and $\sigma_{\min}^{(\mathrm{rtPS})} = \lfloor(\lambda_0^{(\mathrm{rtPS})}/\tau)\rfloor$ for rtPS traffic, and $\sigma_{\max}^{(\mathrm{nrtPS})} = b - \alpha$ and $\sigma_{\min}^{(\mathrm{nrtPS})} = \alpha$ for nrtPS traffic.

To validate the analytical results, we develop a simulator written in Visual C/C++. Table 1 shows the parameters used in the analytical and the simulation models. The

doi: 10.1049/iet-com.2010.0005

Table 1 Parameter settings in the analytical and the simulation models

Parameters	Values
μ (total downlink bandwidth)	10 packets/ frame
N (UGS packets at on-state)	2 packets/frame
$P_a^{(\mathrm{UGS})}$ (UGS traffic-on probability)	0.5
τ (rtPS allowable delay constraint)	1.5 frames
α (nrtPS throughput requirement)	2 packets/frame
b (buffer size per class at BS)	10 packets
number of SS in the simulation	20
traffic ratio of UGS, rtPS, nrtPS and BE	1:1:1:1

topology used in the simulation includes a BS and 20 SS. Every SS is assumed to be able to transmit and receive four different types of traffic, UGS, rtPS, nrtPS and BE. The traffic ratio among the four traffic types is 1:1:1:1. All the simulation results are averaged of up to ten repeated runs

$$\rho = \frac{N \times P_a^{(\mathrm{UGS})} + \lambda^{(\mathrm{rtPS})} + \lambda^{(\mathrm{nrtPS})} + \lambda^{(\mathrm{BE})}}{\mu} \tag{21}$$

To study the impact of traffic load on rtPS delay time, nrtPS throughput, and PDR of BE traffic, we define traffic load (ρ) in (21). To further study the variations between the actual delay time and the allowable delay constraints of rtPS traffic, we define delay-time shift ratio (DTSR) in (22). Furthermore, to observe the deviation between the actual achieved throughput and the minimum throughput requirements, we define throughput shift ratio (TSR) in (23)

$$\mathrm{DTSR} = \frac{\mathrm{DT}^{(\mathrm{rtPS})} - \tau}{\tau} \tag{22}$$

$$\mathrm{TSR} = \frac{\mathrm{Thr}^{(\mathrm{nrtPS})} - \alpha}{\alpha} \tag{23}$$

Fig. 7 shows the DTSR of rtPS against traffic load (ρ) between the proposed ABA and the Q-Aware. As it can be observed, DTSR in the ABA is much closer to zero than that in the Q-Aware when the network is overloaded. This phenomenon demonstrates that the proposed ABA can better meet the rtPS allowable delay constraint (τ) than the Q-Aware because of too much bandwidth consumed in the Q-Aware than really required. Additionally, we observe that the simulation results quite match with the analytical results when traffic load (ρ) is below 1.0; yet the two curves have about 7–10% deviation when it is overloaded.

Fig. 8 shows the TSR of nrtPS against traffic load. When traffic load is increased from 1.0 to 2.2, we observe that both the analytical and the simulation results of TSR in the ABA

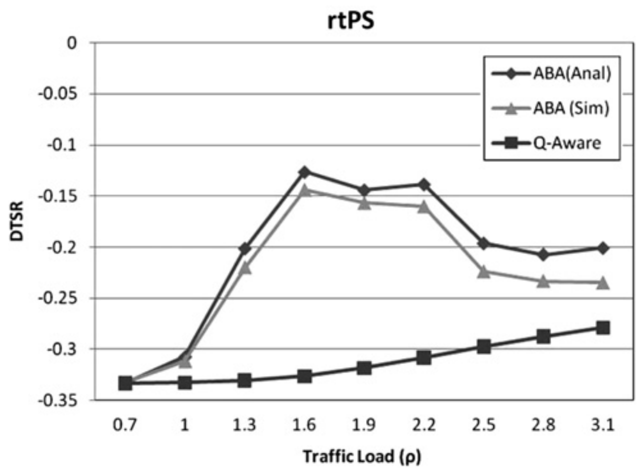

Figure 7 *DTSR of rtPS against traffic load*

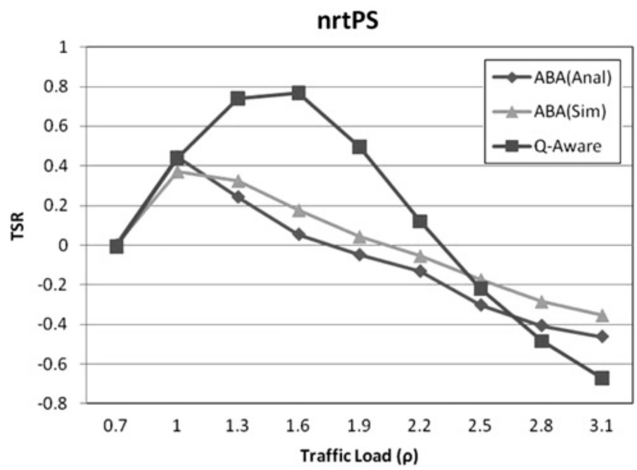

Figure 8 *TSR of nrtPS against traffic load*

are much closer to zero than that in the Q-Aware. Yet when traffic load continuously increases and exceeds 2.5, both TSR curves gradually move away from zero. This phenomenon demonstrates that proposed ABA not only can meet the minimum throughput requirement of nrtPS, but it also endeavours to avoid any possible starvation of BE traffic. This logical predication can be verified from Figs. 9 and 10. As it can be seen from Fig. 9, the ABA can substantially reduce the average PDR of BE traffic as compared to the Q-Aware as traffic load is increased from 1.0 to 2.2. From Fig. 10, the average delay time of BE traffic in the Q-Aware has increased radically as traffic load exceeds 1.6. To the contrast, the average delay time in the proposed ABA remains almost unchanged. The analytical results in Figs. 9 and 10 are again validated by the simulation results. As it can be observed, the two curves in the average PDR are very close except when traffic load is in the range between 1.3 and 2.5. Similar situation appears in the delay time as it can be observed from Fig. 10.

We may observe the ABA performance from another perspective. In fact, the proposed ABA does end up with

doi: 10.1049/iet-com.2010.0005

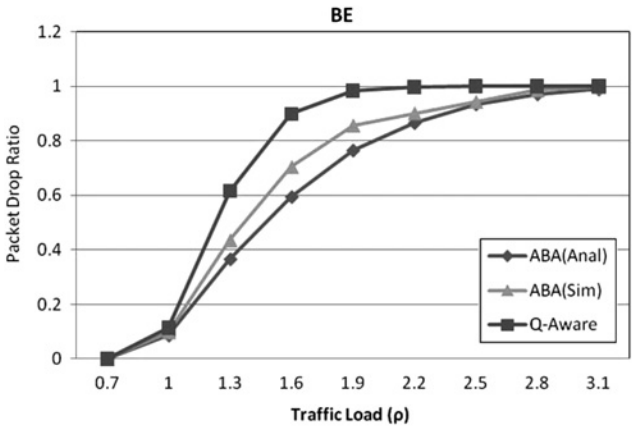

Figure 9 *Average PDR of BE against traffic load*

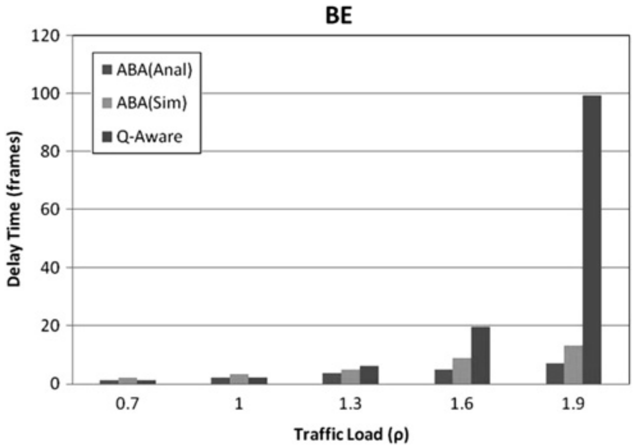

Figure 10 *Average delay time of BE against traffic load*

more delay time than the Q-Aware, even though both of them can utterly meet their rtPS delay constraints. Fig. 11 shows the average delay time of rtPS when the delay time constraint (in frames) is varied. As it can be seen, the Q-Aware maintains almost constant delay time along with the increase of delay time constraints. To the contrast, in

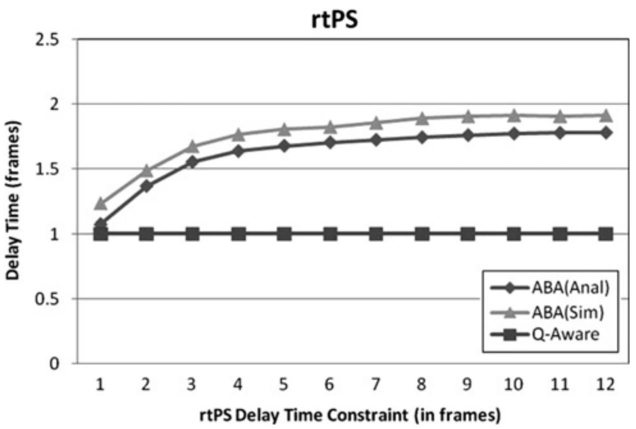

Figure 11 *Average delay time of rtPS against delay time constraints*

the ABA, the delay times from the analytical and the simulation results increase slightly when the specified rtPS time constraint is small (one tothree frames), and then, like in the Q-Aware, both curves become stable when the rtPS time constraint is relaxed to a great extent. Similar situation can be observed from Fig. 12. Both the ABA and the Q-Aware can meet the throughput requirements of nrtPS (the curve of α) quite easily, no matter how we vary the rtPS delay time constraints. However, it is noticed that the Q-Aware has offered too much bandwidth to nrtPS than it is really required. On the other hand, the throughput accomplished by the ABA is just slightly higher than the required one. Also, we observe that the simulation results quite match with the analytical results except when the rtPS delay time constraints are small (one to three frames).

To further observe the performance merits of the proposed ABA, Fig. 13 shows the average nrtPS throughput by varying their minimum throughput requirements (packets/frame). When the nrtPS minimum throughput requirement is small (one to three packets/frame), the proposed ABA does

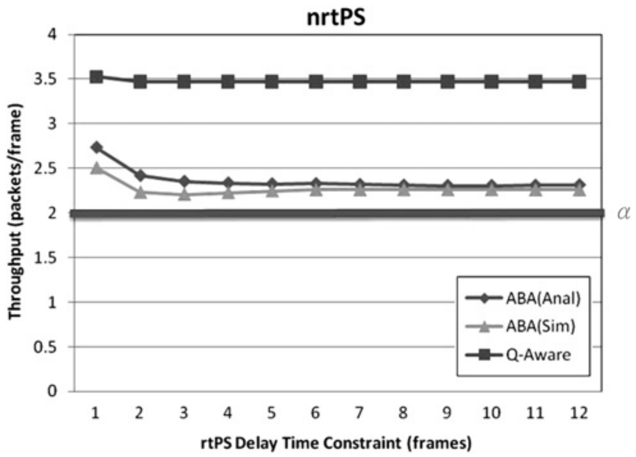

Figure 12 *Average throughput of nrtPS against rtPS delay time constraints*

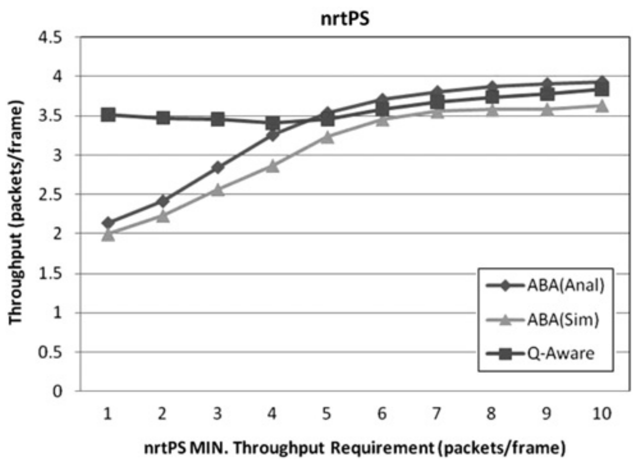

Figure 13 *Average throughput of nrtPS against minimum throughput requirements*

doi: 10.1049/iet-com.2010.0005

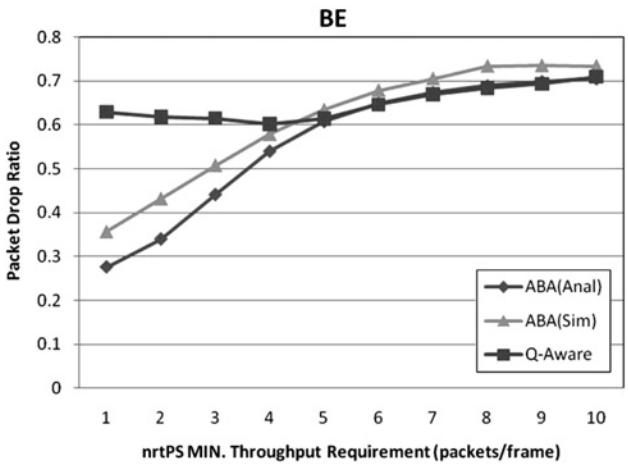

Figure 14 *Average PDR of BE against nrtPS minimum throughput requirements*

not greedily take too much bandwidth. Moreover, we can observe that when the nrtPS minimum throughput requirement exceeds four packets per frame, neither the ABA nor the Q-Aware can meet the throughput requirements, but the former can provide considerably higher throughput than the latter. Finally, we can certify Fig. 13 using the results in Fig. 14. As it can be seen from Fig. 14, when the demand for nrtPS throughput requirement is still small (one to three packets per frame), the average PDR of BE traffic in the ABA is much lower than that in the Q-Aware. This again demonstrates that nrtPS traffic in the proposed ABA does not egoistically take too much wireless bandwidth, so that BE traffic can avoid from any possible starvation. It is also interesting to notice that when the demand for nrtPS throughput requirement exceeds four packets per frame, the average PDR of BE traffic in the ABA is about the same as that in the Q-Aware. This implies that even for higher-throughput requirements in nrtPS, the proposed ABA still perform better than the Q-Aware in preventing from packet drops of BE traffic. From Figs. 13 and 14, it is also noticed that the nrtPS throughput is lower and the PDR of BE is higher as we compare the simulation results to the analytical ones. This is because in the simulation model we do consider those dominate factors in a real network, such as the contention delay among SS and the scheduling delay at BS.

5 Conclusions

This paper has presented an ABA model for multiple classes in IEEE 802.16 WiMAX networks. Four traffic classes of a WiMAX network, UGS, rtPS, nrtPS and BE, are scheduled within a BS. First, ABA reserves the unsolicited bandwidth for UGS. Then, polling bandwidth is allocated for rtPS to meet their end-to-end delay constraints and for nrtPS to meet their minimum throughput requirements. Finally, the remaining bandwidth, if any, is allocated for BE traffic. The major contribution and novelty presented by this paper is right in that the proposed ABA does not greedily overtake too much bandwidth from the lowest-priority class. Instead, it is very intelligent to only meet the delay constraints of rtPS and the minimum throughput requirements of nrtPS, while it endeavours to avoid any possible starvation of BE traffic.

For the purpose of evaluation, a mathematical model with 4-D Markov chains was built to analyse the ABA performance and to compare the performance with a previous work, the Q-Aware model. The analytical results were validated through a simulation, written in Visual C/C++. From the analytical and the simulation results, we have demonstrated that the proposed ABA performs much better than the Q-Aware in the following three aspects: (i) the average delay time of rtPS in the ABA is much closer to the allowable delay time constraint, (ii) the accomplished throughput of nrtPS in the ABA is only slightly higher than the specified minimum throughput requirement and (iii) the average PDR of BE traffic contributed by the ABA is significantly lower than that contributed by the Q-Aware.

6 References

[1] FALLAH Y.P., AGHAREBPARAST F., MINHAS M., ALNUWEIRI H., LEUNG V.C.M.: 'Analytical modeling of contention-based bandwidth request mechanism in IEEE 802.16 wireless networks', *IEEE Trans. Veh. Technol.*, 2008, **57**, (5), pp. 3094–3107

[2] HE J., GUILD K., YANG K., CHEN H.H.: 'Modeling contention based bandwidth request scheme for IEEE 802.16 networks', *IEEE Commun. Lett.*, 2007, **11**, (8), pp. 689–700

[3] LIN L., JIA W., HAN B., ZHANG L.: 'Performance improvement using dynamic contention window adjustment for initial ranging in IEEE 802.16 P2MP networks'. IEEE Wireless Communications and Networking Conf., 11–15 March 2007, pp. 1877–1882

[4] CHANDRA S., SAHOO A.: 'An efficient call admission control for IEEE 802.16 networks'. IEEE Workshop on Local and Metropolitan Area Networks, 10–13 June 2007, pp. 188–193

[5] LIU Q., WANG X., GIANNAKIS G.B.: 'A cross-layer scheduling algorithm with QoS support in wireless networks', *IEEE Trans. Veh. Technol.*, 2006, **55**, (3), pp. 839–847

[6] MSADAA I.C., FILALI F., KAMOUN F.: 'An adaptive QoS architecture for IEEE 802.16 broadband wireless networks'. IEEE Int. Conf. on Mobile Ad hoc and Sensor Systems, 8–11 October 2007, pp. 1–3

[7] RONG B., QIAN Y., LU K.: 'Integrated downlink resource management for multiservice WiMAX networks', *IEEE Trans. Mob. Comput.*, 2007, **6**, (6), pp. 621–632

[8] RONG B., QIAN Y., LU K., CHEN H.H., GUIZANI M.: 'Call admission control optimization in WiMAX networks', *IEEE Trans. Veh. Technol.*, 2008, **57**, (4), pp. 2509–2522

[9] HOU F., HO P.H., SHEN X., CHEN A.Y.: 'A Novel QoS scheduling scheme in IEEE 802.16 networks'. IEEE Wireless Communications and Networking Conf., 11–15 March 2007, pp. 2457–2462

[10] SENGUPTA S., CHATTERJEE M., GANGULY S.: 'Improving quality of VoIP streams over WiMAX', *IEEE Trans. Comput.*, 2008, **57**, (2), pp. 145–156

[11] NIYATO D., HOSSAIIN E.: 'A queuing-theoretic and optimization-based model for radio resource Management in IEEE 802.16 broadband wireless networks', *IEEE Trans. Comput.*, 2006, **55**, (11), pp. 1473–1488

[12] JEON W.S., JEONG D.G.: 'Combined connection admission control and packet transmission scheduling for mobile internet services', *IEEE Trans. Veh. Technol.*, 2006, **55**, (5), pp. 1582–1593

[13] CICCONETTI C., ERTA A., LENZINI L., MINGOZZI E.: 'Performance evaluation of the IEEE 802.16 MAC for QoS support', *IEEE Trans. Mob. Comput.*, 2007, **6**, (1), pp. 26–38

[14] CICCONETTI C., LENZINI L., MINGOZZI E.: 'Quality of service support in IEEE 802.16 networks', *IEEE Netw.*, 2007, **20**, (2), pp. 50–55

[15] KIM S.J., KIM W.J., SUH Y.J.: 'An efficient bandwidth request mechanism for non-real-time service in IEEE 802.16 systems'. Int. Conf. on Communication Systems Software and Middleware, January 2007, pp. 1–9

[16] MUKUL R., SINGH P., JAYARAM D., *ET AL.*: 'An adaptive bandwidth request mechanism for QoS enhancement in WiMAX real-time communication'. Int. Conf. on Wireless and Optical Communications Networks, April 2006, pp. 11–13

[17] PENG Z., GUANGXI Z., HONGZHI L., HAIBIN S.: 'Adaptive scheduling strategy for WiMAX real-time communication'. Int. Symp. Intelligent Signal Processing and Communication Systems, November 28–December 1, 2007, pp. 718–721

[18] NIYATO D., HOSSAIN E.: 'Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks', *IEEE Trans. Mob. Comput.*, 2006, **5**, (6), pp. 668–679

doi: 10.1049/iet-com.2010.0005