Dynamic Bandwidth Allocation for QoS Provisioning in IEEE 802.16 Networks with ARQ-SA

Weiwei Wang, Zihua Guo, *Senior Member, IEEE,* Xuemin (Sherman) Shen, *Senior Member, IEEE,* Changjia Chen, and Jun Cai, *Member, IEEE*

***Abstract*—In this paper, bandwidth allocation, in terms of distributing available data slots among different users, is studied for QoS provisioning in IEEE 802.16 networks. By considering the Automatic Repeat reQuest with Selective Acknowledgement (ARQ-SA) scheme for erroneous wireless channels, a mathematical model is established to theoretically analyze the delay performance of transmitting Service Data Unit (SDU) under a multiuser environment. The analytical results indicate that the delivery delay of the SDU is dominated by the time spent for the first transmission of all its Protocol Data Units (PDUs). Based on this observation, a novel dynamic bandwidth allocation algorithm is proposed and a detailed performance analysis is provided. Simulation results show that the proposed bandwidth allocation algorithm can significantly improve the delay performance of SDUs and ensure the fairness among different users.**

***Index Terms*—IEEE 802.16, WiMAX, bandwidth allocation, ARQ-SA, PDU, SDU.**

I. Introduction

IEEE 802.16, which is usually referred as WiMax (Worldwide Interoperability for Microwave Access), provides both Fixed Broadband Wireless Access (FBWA) and Mobile BWA (MBWA). It has gained significant attention from both industry and academia in recent years. WiMax tends to provide transmission rate of around 10Mbps in the range of few kilometers. In 2004 and 2005, the FBWA (IEEE 802.16d) and MBWA (IEEE 802.16e) versions were ratified, respectively, where the medium access control (MAC) layer and the physical (PHY) layer are clearly defined [1]. There exist several PHY specifications for the 2-11GHz and 10-66GHz in IEEE 802.16d, such as Single Carrier (SC) and Orthogonal Frequency-Division Multiplexing (OFDM). In MAC, the standard supports Time Division Duplexing (TDD) and Frequency Division Duplexing (FDD), and defines two different air-interfaces: Point-to-Multi-Point (PMP) and Mesh. In PMP mode, two Subscriber Stations (SSs) can only communicate through Base Station (BS); while in Mesh mode, two SSs can communicate directly.

In order to provide reliable communications over dynamic wireless channels, Automatic Repeat reQuest (ARQ) has been defined as an option at the MAC layer in IEEE 802.16 standards. Data from the upper layer, which is called Service Data Unit (SDU), is partitioned into ARQ blocks. Several ARQ blocks are then encapsulated into one or more Protocol Data Units (PDUs). As the response of receiving a PDU, different kinds of acknowledgement (ACK) messages can be fed back from the receiver, such as selective ACK, cumulative ACK, cumulative with selective ACK, and cumulative ACK with block sequence ACK. Among them, selective ACK is more commonly used. In selective ACK, once a PDU is not received or is received in errors, an ARQ feedback will be used to provide the receipt status (i.e., ACK or NACK) and only the negative acknowledged PDU will be retransmitted. In this paper, selective ACK is chosen as the ACK message and the corresponding ARQ scheme is called ARQ with Selective ACK (ARQ-SA). Compared to the traditional ARQ applying selection ACK, called SR-ARQ, the ARQ-SA takes into account the specific frame structure of IEEE 802.16 networks and highlights the time delay between the transmission and the retransmission.

Evaluating effects of ARQ on the network performance is important to provide insight on network operation and guideline for designing effective network management schemes [2]. There are several existing research works on the delay analysis of selective repeat ARQ. An exact analysis of PDU delivery delay over two-state Markov channel is provided in [3], [4], which is extended to more general N-state Markov channel in [5]. The re-sequencing delay is considered in [6], [7], while in [8], the overall end-to-end PDU delay is discussed. More description about the delay analysis is given in [9], [10]. However, all these works focus on the delay performance of PDUs only, while from the viewpoint of upper layer applications, the delay of SDU is more important. In [11], [12], the analysis of SDU delivery delay under selective repeat ARQ (SR-ARQ) is presented, while the analysis is limited for a single-user network only and the data is assumed to be continuously transmitted. It is well known that in IEEE

Manuscript received March 8, 2007; revised July 15, 2007; accepted September 9, 2007. The associate editor coordinating the review of this paper and approving it for publication was X. Zhang. This work was supported in part by the Chinese NSFC under Grant 60672069 and 60772043, Chinese Ministry of Education under grant 20050004033 and Beijing Jiaotong University under grant 2005SM006.

W. Wang and C. Chen are with the School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China (e-mail: comeonwww1981@yahoo.com.cn, changjiachen@sina.com.cn).

Z. Guo is with Lenovo Corporate Research, Beijing, China (e-mail: guozh@lenovo.com).

X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1 (e-mail: xshen@bbcr.uwaterloo.ca).

J. Cai is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 5V6 (e-mail: jcai@ee.umanitoba.ca).

Digital Object Identifier 10.1109/TWC.2008.070277.

802.16 networks, the bandwidth is shared by multiple users. Therefore, the analysis of SDU delivery delay in a multiuser IEEE 802.16 network should be carried out.

Scheduling is one of the most important issues in IEEE 802.16 networks. A well designed scheduling scheme should be effective in quality of service (QoS) provisioning, efficient in resource utilization, and fair in resource allocation [13]. Since the scheduling is not specified in the standards, it has become one of the hottest research topics in this area. An uplink scheduling scheme is proposed for supporting all types of service flows defined in IEEE802.16 [14]. Another uplink scheduling scheme for VoIP services is presented in [15] by considering the characteristics of voice data. Both of them focus on the bandwidth allocation of the UL subframe. In [16], a framework is provided for scheduling different types of service flows in both uplink and downlink. The bandwidth of DL/UL is allocated dynamically in PMP mode and the fairness among different flows becomes the main target. However, the discussion did not take the SDU delivery delay into account. Weighted Round Robin (WRR), as a standard and simple scheduling scheme, is commonly adopted in wireless communication networks [17]. It allocates the bandwidth according to the QoS requirements of each service flow so that in the same QoS class, the allocated bandwidth to each flow is equal and fixed. We term such scheduler as the traditional scheme in this paper and let it as the performance benchmark. Obviously, such bandwidth allocation scheme is by no means the best solution in terms of delivery delay. Intuitively, when multiple users compete for the resource, the bandwidth should be dynamically allocated even in the same QoS class according to the SDU buffer status. In addition, ARQ scheme does play important role on the delay performance. Therefore, it is important to design new bandwidth allocation algorithms, which should consider the following two key features:

- the SDU delay instead of an individual PDU delay;
- the effects of the ARQ scheme on the scheduling.

In this paper, bandwidth allocation issues have been studied for downlink IEEE 802.16 networks under PMP mode. The focus on downlink results from the fact that the downlink may have to transport more traffic than the uplink and may become the bottleneck of the networks. A mathematical model is first established to theoretically analyze the delivery delay of a SDU with ARQ-SR. Here, the delivery delay is defined as the time duration from the first transmission of the first PDU in the SDU to the time when all PDUs have been successfully received so that the SDU can be delivered to the upper layer. Analytical results indicate that the delivery delay of one SDU is dominated by the time spent for the first transmission of all its PDUs. By taking this property into account, a novel downlink dynamic bandwidth allocation algorithm, in terms of distributing available data slots among different users, is introduced. The proposed algorithm is based on the priority allocation principle and assigns higher priority to the user which may experience longer time to transmit all its PDUs for the first time. Theoretical analysis indicates that the proposed algorithm can significantly reduce the delivery delay of SDUs and at the same time, hold a similar fairness performance as the traditional scheme. Simulation results are finally provided to further demonstrate the effectiveness and efficiency of the proposed dynamic bandwidth allocation algorithm.

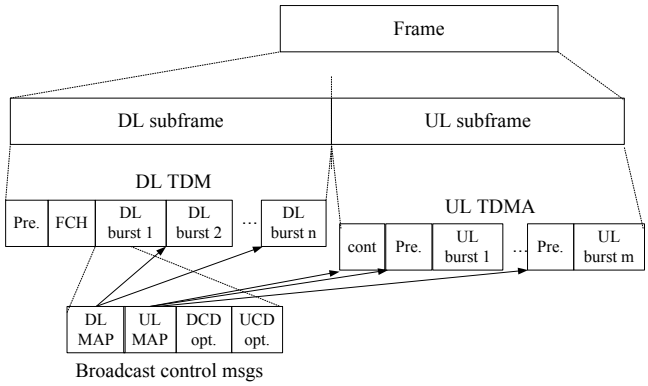


Fig. 1. OFDM frame structure with TDD.

The remainder of this paper is organized as follows. Section II defines the system model of an IEEE 802.16 network under consideration. In Section III, the analysis of SDU delivery delay under ARQ-SA is presented. Section IV presents the proposed dynamic bandwidth allocation algorithm in detail. Performance analysis of the proposed algorithm is also provided. Numerical results are given in Section V, followed by the conclusions in Section VI.

II. System Model

An IEEE 802.16 network operating under PMP mode with OFDM and TDD is considered. The network has a base station (BS) located at the center of the covered area. The data transmission at the MAC layer is frame-by-frame based. Each frame consists of one downlink (DL) subframe and one uplink (UL) subframe as shown in Fig. 1. In this paper, we focus on the DL subframe only. Each DL subframe begins with a preamble followed by a Frame Control Header (FCH). The FCH specifies the burst profile, which defines the coding algorithm, code rate and modulation level used for data transmission, and the duration of one or more DL bursts immediately following the FCH. After that, broadcast messages, such as DL-MAP, UL-MAP, DL Channel Descriptor (DCD), and UL Channel Descriptor (UCD), can be transmitted. The remainder of the subframe contributes to the pay load, which is further divided into a number of data bursts (or slots). In this paper, each slot is assumed to hold a same time duration, which is long enough to support the transmission of one PDU. The details of other components in Fig. 1 can be found in [1].

Data from the upper layer, called SDU, is partitioned into ARQ blocks, and several ARQ blocks are encapsulated into one or multiple PDUs with equal length. In this paper, ARQ-SA is applied to compensate the possible transmission errors from the physical layer. With ARQ-SA, once a PDU is lost in DL/UL subframe, the ACK should be sent to the transmitter in the following UL/DL subframe, and the PDU can be retransmitted in the next DL/UL subframe. Obviously, there is at least one subframe (UL/DL) between the transmission and the retransmission of the same PDU. In this paper, we assume that the retransmission has higher priority than the

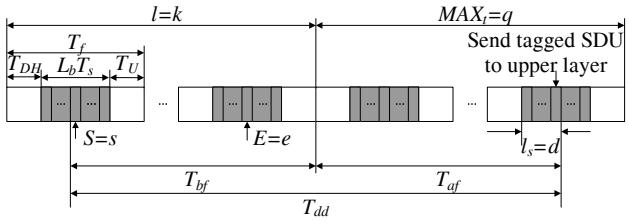


Fig. 2. Definition of variables.

transmission of new PDUs, and at the receiver end, each SDU is delivered to the higher layer only if all SDUs with lower identifiers have been correctly received. Each PDU experiences an independent error with probability p, while ACK/NACK messages are error-free since, in the real world, they are shorter than data packets and are transmitted by more robust modulation and coding schemes. The situation that the ACK/NACK messages are erroneous and delivered after several subframes from the transmission of the PDU will be left for our future works.

III. Delivery Delay Performance Analysis

In this section, delivery delay of a tagged SDU is analyzed in the downlink IEEE 802.16 network with ARQ-SR. However, we'd like to point out that the method could also be applied to the uplink. The delivery delay of the tagged SDU, T_{dd}, is defined as the time interval from the first transmission of the first PDU to the time when the SDU is delivered to the upper layer in the receiver.

A. Delivery Delay of SDU

Consider a tagged SDU which belongs to a tagged user and consists of L tagged PDUs. Note that L is a random variable for different SDUs. The definitions of variables used in the analysis are shown in Fig. 2.

Each frame has a duration of T_f. From the tagged user point of view, each frame can be separated into three portions, denoted by T_{DH}, L_bT_s, and T_U. T_{DH} and T_U represent the total time in the frame before and after the transmission of the tagged user, respectively. T_{DH} and T_U take into account the transmission of control messages, data from other active users, and UL subframe. L_bT_s, called tagged burst, is the time actually reversed for the transmission of the tagged user in each frame, which is shown as the shadowed areas in Fig. 2. Here L_b denotes the number of slots in the tagged burst and T_s denotes the slot duration. In other words, total of L_b PDUs from the tagged user can be transmitted in each DL subframe. In this paper, L_b is supposed to be fixed during the transmission of the tagged SDU. According to Fig. 2, we have

$$T_f = T_{DH} + L_bT_s + T_U. \tag{1}$$

By further defining

- S ($S \in [1, L_b]$): the first transmission of the first tagged PDU happens at the S-th slot in the tagged burst;
- E ($E \in [1, L_b]$): the first transmission of the last tagged PDU happens at the E-th slot in the tagged burst;
- l ($l \in [1, \infty)$): the number of frames from the position of S to the position of E;
- MAX_t ($MAX_t \in [0, \infty)$): the number of frames after the first transmission of the last tagged PDU to the delivery of the tagged SDU;
- l_s ($l_s \in [1, E]$): the number of PDUs before the delivery of the tagged SDU in the last frame;
- T_{bf}: the time duration from the first transmission of the first tagged PDU till the end of the frame where the first transmission of the last tagged PDU happens;
- T_{af}: the time duration from the frame after the first transmission of the last tagged PDU to the delivery of SDU to the upper layer,

T_{dd} can be calculated in terms of S, E, l, MAX_t, and l_s by considering the following three cases.

- Case 1: $MAX_t = 0$ and $l = 1$
  If $MAX_t = 0$ and $l = 1$, all tagged PDUs are successfully transmitted in one frame. Otherwise, the retransmission happening in the next frame will result in non-zero MAX_t. Under this case, we have
  $$T_{dd} = LT_s. \tag{2}$$
- Case 2: $MAX_t = 0$ and $l > 1$
  Under this case, the delivery of the tagged SDU to the upper layer occurs in the same frame where the position E happens. Therefore,
  $$T_{dd} = (L_b - S + 1)T_s + T_U + (l-2)T_f + T_{DH} + ET_s. \tag{3}$$
- Case 3: $MAX_t \neq 0$ and $l \geq 1$
  Under this case, from Fig. 2, we have
  $$T_{bf} = (L_b - S + 1)T_s + T_U + (l-1)T_f \tag{4}$$
  $$T_{af} = (MAX_t - 1)T_f + T_{DH} + l_sT_s. \tag{5}$$
  Therefore,
  $$\begin{aligned} T_{dd} &= T_{bf} + T_{af} \\ &= (L_b - S + 1)T_s + T_U + (l-1)T_f \\ &\quad + (MAX_t - 1)T_f + T_{DH} + l_sT_s. \end{aligned} \tag{6}$$

In summary, the delivery delay T_{dd} can be written as

$$T_{dd} = \begin{cases} LT_s, & \text{if } MAX_t = 0, l = 1 \\ (L_b - S + 1)T_s + T_U + (l-2)T_f + T_{DH} + ET_s, & \text{if } MAX_t = 0, l > 1 \\ (L_b - S + 1)T_s + T_U + (l-1)T_f \\ \quad + (MAX_t - 1)T_f + T_{DH} + l_sT_s, & \text{if } MAX_t \neq 0, l \geq 1. \end{cases} \tag{7}$$

From (7) and Fig. 2, T_{dd} is determined by several variables, i.e., S, E, l, MAX_t, l_s; thus, the information of the joint probability $P(S = s, E = e, l = k, MAX_t = q, l_s = d)$ should be derived. Since

$$\begin{aligned} & P(S = s, E = e, l = k, MAX_t = q, l_s = d) \\ = \; & P(MAX_t = q, l_s = d | S = s, E = e, l = k) \\ & \times P(S = s, E = e, l = k), \end{aligned}$$

we discuss how to derive these two probabilities, $P(MAX_t = q, l_s = d|S = s, E = e, l = k)$ and $P(S = s, E = e, l = k)$, separately.

B. Calculation of $P(S = s, E = e, l = k)$

Let the tagged SDU be the $(i+1)$-th SDU of the tagged user, denoted as SDU_{i+1}. Then, the previous SDU can be denoted as SDU_i. In order to do the calculation, we introduce two new variables.

- X_i ($X_i \in [0, L_b - 1]$): the number of slots left in the tagged burst after the first transmission of the last PDU from the SDU_i;
- R: the number of retransmitted tagged PDUs in the same frame as the first transmission of the last tagged PDU if the first transmission of SDU_{i+1} cannot be finished in one frame.

We first derive the conditional probability, $P(X_{i+1} = j, l = k|X_i = m)$, with $m \neq 0$ and $m = 0$, respectively.

- $m \neq 0$
  If $m \neq 0$, after the first transmission of the last PDU from SDU_i, there are slots left in the current tagged burst, which can be used for the first transmission of the tagged PDUs from SDU_{i+1}. If $k = 1$, the first transmission of all tagged PDUs can be finished in the same frame so that $r = 0$ and $j = m - L$, i.e.,

$$P(X_{i+1} = j, l = k, R = r|X_i = m) = 1, \quad \text{if } k = 1, j = m - L, r = 0. \tag{8}$$

  Otherwise, if $k > 1$, as shown in Fig. 3, define positions 1-3 as the beginnings of the first frame, the $k-1$th frame, and the kth frame of the tagged SDU, respectively, and define position 4 as the end of the first transmission of all tagged PDUs. Then, from position 1 to position 4, $L + (L_b - m)$ different PDUs of the tagged user are transmitted. From position 1 to position 3, C_{k-1} PDUs of the tagged user are transmitted correctly, where

$$C_{k-1} = L + (L_b - m) - (L_b - j), \quad C_{k-1} \in [0, (k-1)L_b]. \tag{9}$$

  Because r PDUs are retransmitted in the k-th tagged burst, from position 2 to position 3, $L_b - r$ PDUs are transmitted correctly. Thus, from position 1 to position 2, C_{k-2} PDUs of the tagged user are successfully transmitted, where

$$C_{k-2} = C_{k-1} - (L_b - r), \quad C_{k-2} \in [0, (k-2)L_b]. \tag{10}$$

  Therefore, for $k > 1$,

$$\begin{aligned} &P(X_{i+1} = j, l = k, R = r|X_i = m) \\ &= \binom{(k-2)L_b}{C_{k-2}} p^{((k-2)L_b - C_{k-2})}(1-p)^{C_{k-2}} \\ &\quad \times \binom{L_b}{L_b - r} p^r (1-p)^{L_b - r}. \end{aligned} \tag{11}$$

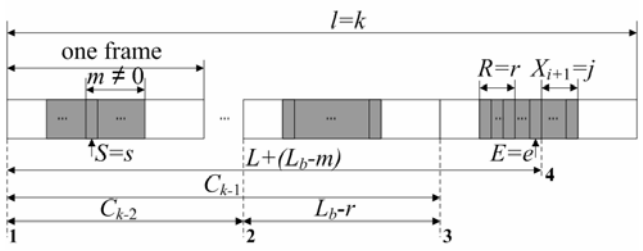


Fig. 3. Calculation of $P(X_{i+1} = j, l = k, R = r|X_i = m)$ with $m \neq 0$.

Combining (8) and (11), $P(X_{i+1} = j, l = k, R = r|X_i = m)$ for $m \neq 0$ can be written as

$$\begin{aligned} &P(X_{i+1} = j, l = k, R = r|X_i = m) \\ &= \begin{cases} 1, & \text{if } k = 1, j = m - L, r = 0 \\ \binom{(k-2)L_b}{C_{k-2}} p^{(k-2)L_b - C_{k-2}}(1-p)^{C_{k-2}} & \\ \times \binom{L_b}{L_b - r} p^r (1-p)^{L_b - r}, & \\ \quad \text{if } k > 1, C_{k-1} \in [0, (k-1)L_b], & \\ \quad C_{k-2} \in [0, (k-2)L_b] & \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

Finally, we have

$$\begin{aligned} &P(X_{i+1} = j, l = k|X_i = m) \\ &= \sum_{r=0}^{L_b - j - 1} P(X_{i+1} = j, l = k, R = r|X_i = m). \end{aligned} \tag{13}$$

- $m = 0$
  If $m = 0$, as shown in Fig. 4, the first transmission of the last PDU from SDU_i happens at position 2. By considering possible retransmissions, assume that the first transmission of the first tagged PDU happens at the B_i-th slot of the A_i-th tagged burst after the first transmission of the last PDU of SDU_i. Obviously, according to Fig. 4, from position 1 to position 3, all PDUs are transmitted with errors, while from position 3 to position 4, $b - 1$ PDUs are transmitted with errors. Following the similar way used in the $m \neq 0$ case, we have

$$\begin{aligned} &P(X_{i+1} = j, l = k, A_i = a, B_i = b, R = r|X_i = 0) \\ &= \begin{cases} (p^{L_b})^{a-1} \binom{L_b}{b-1} p^{b-1}(1-p)^{L_b - (b-1)}, & \\ \quad \text{if } k = 1, j = L_b - (b + L - 1), r = 0 & \\ (p^{L_b})^{a-1} \binom{L_b}{b-1} p^{b-1}(1-p)^{L_b - (b-1)} & \\ \times \binom{(k-2)L_b}{C'_{k-2}} p^{(k-2)L_b - C'_{k-2}}(1-p)^{C'_{k-2}} & \\ \times \binom{L_b}{L_b - r} p^r (1-p)^{L_b - r} & \\ \quad \text{if } k > 1, C'_{k-1} \in [0, (k-1)L_b], & \\ \quad C'_{k-2} \in [0, (k-2)L_b] & \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{14}$$

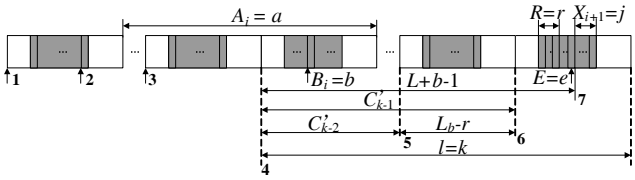


Fig. 4. Calculation of $P(X_{i+1} = j, l = k, R = r | X_i = m)$ with $m = 0$.

where

$a \in [1, \infty)$
$b \in [1, L_b]$
$C'_{k-1} = (L + b - 1) - (L_b - j), C'_{k-1} \in [0, (k-1)L_b]$
$C'_{k-2} = C'_{k-1} - (L_b - r), C'_{k-2} \in [0, (k-2)L_b]$.

Finally,

$$P(X_{i+1} = j, l = k, A_i = a, B_i = b | X_i = 0) = \sum_{r=0}^{L_b - j - 1} P(X_{i+1} = j, l = k, A_i = a, B_i = b, R = r | X_i = 0). \quad (15)$$

Summing (13) by all values of k, and (15) by all values of k, a, and b, we can get $P(X_{i+1} = j | X_i = m)$, $\forall j, m \in [0, L_b - 1]$, for the given value of L. Since the SDU length L is also a random variable, $P(X_{i+1} = j | X_i = m)$ can finally be written as

$$P(X_{i+1} = j | X_i = m) = \sum_{\lambda=1}^{L_{max}} P(X_{i+1} = j | X_i = m, L = \lambda) P(L = \lambda). \quad (16)$$

where $P(L = \lambda)$ is the distribution of the tagged SDU's length and L_{max} denotes the maximum value. Obviously, (16) defines the transition probability of X_i from state m to state j. Let $\mathbf{P}$ be a transition probability matrix of the state variable X_i and define the steady-state probability vector as $\mathbf{\Pi} = (\pi_0, \ldots, \pi_{L_b - 1})$, where $\pi_j = P(X_i = j)$. $\mathbf{\Pi}$ can be calculated by solving the following equation system

$$\begin{cases} \mathbf{\Pi} \times \mathbf{P} = \mathbf{\Pi} \\ \sum_{j=0}^{L_b - 1} \pi_j = 1. \end{cases} \quad (17)$$

Given $\mathbf{\Pi}$, we can calculate $P(S = s, E = e, l = k)$.

If $S = 1$, $(S = 1, E = e, l = k)$ means $(X_{i+1} = L_b - e, l = k, B_i = 1, X_i = 0)$. Then, we have

$$P(S = 1, E = e, l = k) = \sum_{a=1}^{\infty} P(X_{i+1} = L_b - e, l = k, A_i = a, B_i = 1 | X_i = 0) \times P(X_i = 0). \quad (18)$$

Otherwise, if $S \neq 1$, $(S = s, E = e, l = k)$ means $(X_{i+1} = L_b - e, l = k, B_i = s, X_i = 0)$ or $(X_{i+1} = L_b - e, l = k, X_i = L_b - (s-1))$. Then from (13) and (15), we can get $P(S = s, E = e, l = k)$ $(s > 1)$ as

$$P(S = s, E = e, l = k) = \sum_{a=1}^{\infty} P(X_{i+1} = L_b - e, l = k, A_i = a, B_i = s | X_i = 0) \times P(X_i = 0) + P(X_{i+1} = L_b - e, l = k | X_i = L_b - (s-1)) \times P(X_i = L_b - (s-1)). \quad (19)$$

C. Calculation of $P(MAX_t = q, l_s = d | S = s, E = e, l = k)$

From Fig. 2, we can deduce that the successful transmission of e PDUs appeared in the k-th frame means the successful delivery of the tagged SDU to the upper layer. Assume that the α-th PDU of those e PDUs is transmitted t_α times during the last MAX_t frames. Then, $\{t_\alpha,\ \alpha = 1, 2, \ldots, e\}$ is a random variable with an independent and identical distribution (i.i.d). By considering the fact that the α-th PDU has already experienced one erroneous transmission in the kth frame, the probability of $P(t_\alpha = y)$ and $P(t_\alpha \leq y)$ can be obtained, respectively, as

$$P(t_\alpha = y) = p^y(1-p) \quad (20)$$

$$P(t_\alpha \leq y) = \sum_{i=0}^{y} P(t_\alpha = i) = \sum_{i=0}^{y} p^i(1-p) = 1 - p^{y+1}. \quad (21)$$

If $q = 0$, all e PDUs must be successfully transmitted in the k-th frame. Therefore,

$$P(MAX_t = 0, l_s = d | S = s, E = e, l = k) = \begin{cases} (1-p)^e, & \text{if } d = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

If $q > 0$, d PDUs are transmitted for q times and $(e - d)$ PDUs are transmitted for $q - 1$ times at most during the last MAX_t frames. Therefore, according to (20) and (21), we get $P(MAX_t = q, l_s = d | S = s, E = e, l = k)$, $(q > 0, d > 0)$ as

$$P(MAX_t = q, l_s = d | S = s, E = e, l = k) = \binom{e}{d} (p^q(1-p))^d (1 - p^q)^{e-d}. \quad (23)$$

Finally, combining (18), (19), (22), and (23), we can obtain the distribution of $P(S = s, E = e, l = k, MAX_t = q, l_s = d)$.

IV. Bandwidth Allocation Algorithm

According to the analysis in Section III, and the analytical results given in Section V later, for IEEE 802.16 networks with ARQ-SR, the delivery delay of a SDU is mainly determined by the time used for the first transmission of all its PDUs. By taking this property into account, in this section, a novel downlink bandwidth allocation algorithm is proposed to reduce the time spent by the first transmission of one SDU, which can achieve

fair bandwidth sharing and reduction on the delivery delay. Performance analysis of the proposed bandwidth allocation algorithm is also provided.

A. Bandwidth Allocation Algorithm

Before presenting the details of the proposed bandwidth allocation algorithm, we first introduce two parameters which are related to the first transmission of each SDU.

- L_p: the number of PDUs which haven't been transmitted for the first time at the beginning of a given DL subframe. L_p has an initial value equal to the SDU length, L;
- L_f: its initial and minimum values are L_{f0} and 0, respectively, where L_{f0} denotes that, with the highest probability, the number of frames that are needed for the first transmission of a SDU based on the traditional scheme, WRR. Given the length of the SDU, there exists a corresponding L_{f0}.

When the first transmission of the first PDU in one SDU begins, the values of L_p and L_f are initialized to L and L_{f0}, respectively. L_p will be reduced by one after the first transmission of one PDU, and L_f will be reduced by one if the first transmission of all PDUs from one SDU can not be finished in one DL subframe.

Let S be the total number of slots available in the DL subframe. Due to the existence of retransmissions, only some of S slots can be used for the transmission of new PDUs, the number of which is denoted by N ($N \in [0, S]$). The value of N can be calculated at the beginning of any DL subframe. In this paper, the system bandwidth is defined in terms of N and the proposed algorithm focuses on how to allocate these N slots to different users for better delay and fairness performance.

From the definitions of L_f and L_p, the former decreased by one means the delay is increased by the length of one frame; while the latter decreased by one means the delay is increased by the length of one PDU. Obviously, L_f is more important than L_p on indicating the delay performance of SDU. Therefore, we define a two-level priority system for each user in the network as follows:

$$P_1 : L_f \tag{24}$$

$$P_2 : \min(L_p, N) \tag{25}$$

When allocating bandwidth, we first consider L_f to determine the allocation priority of each user. If some users have the same L_f, $\min(L_p, N)$ will then be considered. In each priority level, the smaller the value, the higher the priority. In (24), the first priority level, P_1, will be reduced by one if the first transmission of one user's SDU cannot be finished in one frame. Therefore, with the decreasing of P_1, the probability of allocating bandwidth to this user will be increased. As a result, it could ensure that no user will suffer from starvation. On the other hand, the second priority level shows that the SDU, which has smaller number of PDUs waiting for their first transmission (i.e., smaller L_p) should be allocated bandwidth with higher priority. The second priority level also considers the situation where the SDUs from different users have very different lengths. Under this case, the shorter SDU should be transmitted more quickly than the larger one to minimize the total delivery delay.

Without loss of generality, we consider a two-user network (userA and userB) to illustrate the proposed bandwidth allocation algorithm. At the BS, each user has its own transmitting queue. Let L_A and L_B be the lengths of SDUs from userA and userB, respectively. At the initial state, e.g., the beginning of the i-th frame, the corresponding parameters of userA and userB are denoted by $(L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$. Let $\zeta = (N, L_{pA}, L_{fA}, L_{pB}, L_{fB})$ denote the network state. Then, at the beginning of the i-th frame, we can set the initial values of ζ as $(N^i, L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$, where N^i means the number of slots which can be used for the first transmission of new PDUs in the i-th DL subframe. We can also calculate the two-level priority of each user based on (24) and (25), respectively, and denote them as $(P^1_A, P^2_A, P^1_B, P^2_B)$. Finally, we denote L_{sA} and L_{sB} the number of slots allocated to userA and userB, respectively, in the current frame, and the initial values of them are set to be zero. The proposed algorithm follows two steps:

Step 1: Bandwidth allocation.

Both users are sorted by ascending order based on the two-level priority. The bandwidth is allocated to the user with the highest priority as follows.

$$(A, B) = \begin{cases} (\min(L_{pA}, N), 0), \\ \qquad \text{if } P^1_A < P^1_B \text{ or } (P^1_A = P^1_B, P^2_A < P^2_B) \\ (0, \min(L_{pB}, N)), \\ \qquad \text{if } P^1_A > P^1_B \text{ or } (P^1_A = P^1_B, P^2_A > P^2_B) \\ (\min(L_{pA}, N), 0) \text{or } (0, \min(L_{pB}, N)) \\ \qquad \text{if } P^1_A = P^1_B \text{ and } P^2_A = P^2_B \end{cases} \tag{26}$$

where A and B denote the allocated bandwidth to userA and userB, respectively.

Step 2: Parameter update.

After the bandwidth allocation, system parameters are updated as follows:

$$\begin{cases} N = N - A - B \\ L_{sA} = L_{sA} + A \\ L_{sB} = L_{sB} + B. \end{cases} \tag{27}$$

In addition, (L_{pA}, L_{fA}) can be updated as

$$(L_{pA}, L_{fA}) = \begin{cases} (L_{pA} - A, \max(L_{fA} - 1, 0)), \\ \qquad \text{if } A < L_{pA} \\ (L_A, L_{fA0}), \qquad \text{if } A = L_{pA}. \end{cases} \tag{28}$$

In (28), if the allocated bandwidth is smaller than the left number of PDUs from the SDU at the head of userA's transmitting buffer, the number of the left PDUs will be reduced by A, and L_{fA} is reduced by one till zero; if all the left new PDUs can be transmitted in the current frame, L_{pA} and L_{fA} are updated to the length (L_A) and L_{fA0} of the next SDU, respectively. Likewise, (L_{pB}, L_{fB}) can be updated in the similar way.

After updating the parameters, the priority of each user can be recalculated. Then steps 1 and 2 are repeated until all N^i

slots have been allocated. The ultimate values of L_{sA} and L_{sB} are the final bandwidth allocation for the first transmission of the new PDUs of userA and userB, respectively, in the current DL subframe.

After bandwidth allocation for the i-th frame, $(L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$ are updated to $(L^{i+1}_{pA}, L^{i+1}_{fA}, L^{i+1}_{pB}, L^{i+1}_{fB})$ for the $(i+1)$-th frame, which satisfies

$$(L^{i+1}_{pA}, L^{i+1}_{fA}) = \begin{cases} (L^i_{pA} - L_{sA}, \max(L^i_{fA} - 1, 0)), \\ \qquad \text{if } L_{sA} < L^i_{pA} \\ (L_A, L_{fA0}), \\ \qquad \text{if } L_{sA} > L^i_{pA},\ \bmod(L_{sA} - L^i_{pA}, L_A) = 0 \\ (L_A - \bmod(L_{sA} - L^i_{pA}, L_A), \max(L_{fA0} - 1, 0)), \\ \qquad \text{if } L_{sA} > L^i_{pA},\ \bmod(L_{sA} - L^i_{pA}, L_A) \neq 0 \end{cases} \tag{29}$$

where $\bmod(c, C)$ means the remainder of c being divided by C. From (29), if the allocated bandwidth is not enough for the first transmission of the left PDU of the SDU, L^{i+1}_{pA} is updated to the left number of PDUs after the i-th frame transmission and L^{i+1}_{fA} equals $L^i_{fA} - 1$ till zero. If the first transmission of a SDU is just finished at the end of the i-th frame, L^{i+1}_{pA} and L^{i+1}_{fA} are updated to their initial values. Otherwise, the parameter updates following the case when $L_{sA} < L^i_{pA}$. The difference is that the number of new PDUs which has been served in the i-th frame equals $\bmod(L_{sA} - L^i_{pA}, L_A)$. The parameters of userB can be updated following the same way. The pseudo-code of the proposed algorithm is summarized in Appendix A. Note that the similar principle can also be applied for the multi-user scenario. In the multi-user scenario, the two-level priority system will be kept. The difference from the two-user scenario is that at any step, the status of more than two users needs to be updated.

Moreover, since the definition of L_f is related to the SDU transmission with WRR scheme, which is only determined by the bandwidth allocation of WRR and the length of SDU, the proposed algorithm can provide similar fairness performance as WRR.

B. Performance Analysis

In this subsection, the performance of the proposed bandwidth allocation algorithm is analyzed. Since the time spent for the first transmission of the SDU is equivalent to the value of l as shown in Fig. 2, the analysis will focus on deriving the distribution of l.

Let state variable ζ ($\zeta \in \mathbf{\Omega}$) be defined as $(N, L_{pA}, L_{fA}, L_{pB}, L_{fB})$ at the beginning of the frame, where $\mathbf{\Omega}$ is the set of all possible values of ζ (how to obtain $\mathbf{\Omega}$ is given in the Appendix B). Let $P(\zeta_2|\zeta_1)$ denote the one-step transition probability from state $\zeta_1 = (N_1, L_{pA1}, L_{fA1}, L_{pB1}, L_{fB1})$ at the beginning of the i-th frame to state $\zeta_2 = (N_2, L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})$ at the beginning of the $(i+1)$-th frame. Since N_2 is independent of ζ_1, the transition probability can be calculated by

$$P(\zeta_2|\zeta_1) = P((L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})|\zeta_1)P(N_2) \tag{30}$$

where $P(N_2)$ is the probability of $S - N_2$ erroneous transmissions among S PDUs. For independent PDU error probability p,

$$P(N_2) = \binom{S}{N_2} p^{S-N_2}(1-p)^{N_2}. \tag{31}$$

Due to the possible random choosing in (26), different $(L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})$ may be available based on the given ζ_1. By assuming each $(L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})$ can be obtained by m ways and each way has n_k ($k \in [1, m]$) random choosing,

$$P((L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})|\zeta_1) = \sum_{k=1}^{m} 0.5^{n_k}. \tag{32}$$

Combining (30)-(32), we can generate a transition probability matrix, $\mathbf{P_\Omega}$, for any $\zeta \in \mathbf{\Omega}$. Let the steady-state probability of ζ be $\mathbf{\Theta} = (\theta_1, \ldots, \theta_\varepsilon)$, where ε is the total number of elements in $\mathbf{\Omega}$, $\theta_j = P(\zeta^j)$, and ζ^j means the j-th element of $\mathbf{\Omega}$. Vector $\mathbf{\Theta}$ can be obtained by solving the following linear equations of a Markov chain:

$$\begin{cases} \mathbf{\Theta} \times \mathbf{P_\Omega} = \mathbf{\Theta} \\ \sum_{j=1}^{\varepsilon} \theta_j = 1. \end{cases} \tag{33}$$

Given $\mathbf{\Theta}$, we have:

$$P(L_{sA1}, L_{sB1}, \zeta_1) = P(L_{sA1}, L_{sB1}|\zeta_1)P(\zeta_1). \tag{34}$$

We now use userA as an example to show how to calculate the distribution of l, i.e., $P(l_A \leq k)$, for all $k \in [1, \infty)$, which means the first transmission of one SDU for userA is finished in no more than k frames. Let Y_1 denote $(L_{sA1}, L_{sB1}, \zeta_1)$ in the i-th frame, and Y_2 denote $(L_{sA2}, L_{sB2}, \zeta_2)$ in $(i+1)$-th frame. Let $\mathbf{F}_{A1}$ be the set of all possible values of Y_1, which satisfy the condition so that the first transmission of the SDU can be finished in one frame for userA. Then, $P(l_A \leq 1)$ can be obtained by using (34) as

$$P(l_A \leq 1) = \sum_{\mathbf{F}_{A1}} P(Y_1). \tag{35}$$

In order to obtain $P(l_A \leq 2)$, which means the first transmission of one SDU for userA is finished in one or two frames, we calculate

$$\begin{aligned} & P(Y_2, Y_1) \\ = \ & P(Y_2|Y_1)P(Y_1) \\ = \ & P(L_{sA2}, L_{sB2}|\zeta_2, Y_1)P(\zeta_2|Y_1)P(Y_1). \end{aligned} \tag{36}$$

Since L_{sA2} and L_{sB2} are only determined by ζ_2,

$$P(L_{sA2}, L_{sB2}|\zeta_2, Y1) = P(L_{sA2}, L_{sB2}|\zeta_2). \tag{37}$$

Moreover, since $(L_{pA2}, L_{fA2}, L_{pB2}, L_{fB2})$ are deterministic given Y_1, $P(\zeta_2|Y_1)$ only depends on the number of slots which are used for the first transmission of PDUs in the $(i+1)$-th frame, i.e.,

$$P(\zeta_2|Y_1) = \binom{S}{N_2} p^{S-N_2}(1-p)^{N_2}. \tag{38}$$

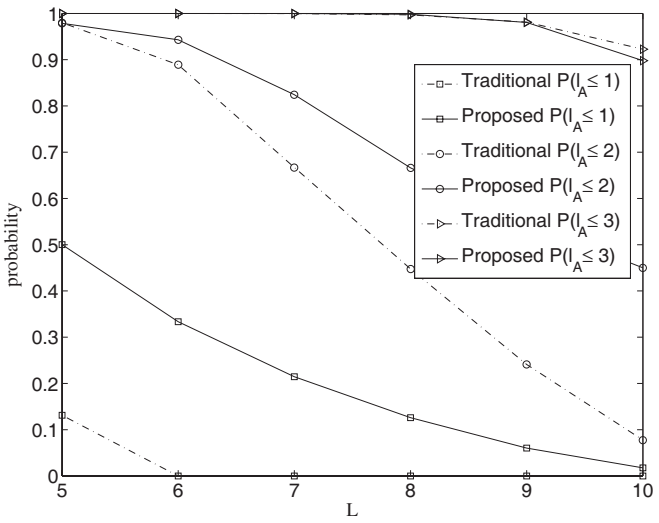


Fig. 5. Probability comparison.

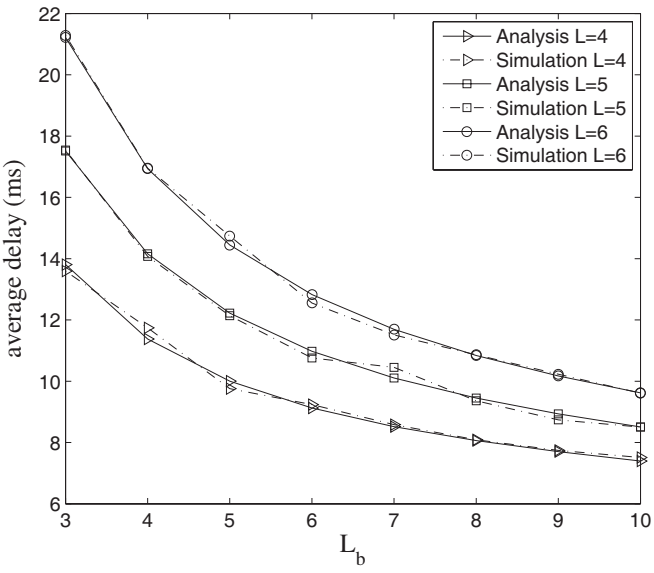


Fig. 6. Comparison of analysis and simulation results (a static channel model).

Substituting (37) and (38) into (36), we have

$$P(Y_2, Y_1) = P(L_{sA2}, L_{sB2}|\zeta_2) \begin{pmatrix} S \\ N_2 \end{pmatrix} p^{S-N_2}(1-p)^{N_2} P(Y_1). \tag{39}$$

Let $\mathbf{F}_{A2}$ be the set of all possible values of (Y_2, Y_1), which satisfies the condition so that the first transmission of one SDU for userA can be finished in one or two frames. Then,

$$P(l_A \leq 2) = \sum_{\mathbf{F}_{A2}} P(Y_2, Y_1). \tag{40}$$

Similarly, we can get $P(l_A \leq k)$, for any $k \in [1, \infty)$.

Fig. 5 shows the numerical results based on the previous analysis. Each DL subframe consists of total 10 slots for both userA and userB. Two users have a same bandwidth requirement. Therefore, in the traditional scheme, each user are allocated 5 PDUs in one DL subframe. The results of the traditional scheme are obtained by using the similar way as our analysis except with the fixed bandwidth allocation. In Fig. 5, the probability of finishing the first transmission of SDU in one, two and three frames under different SDU lengths are compared. It can be seen that $P(l_A \leq 1)$ and $P(l_A \leq 2)$ of the proposed algorithm are much larger than those of the traditional scheme, and $P(l_A \leq 3)$ almost reaches 100% for both schemes. When L=10, the traditional scheme outperforms our scheme a little. That is because, in our algorithm, the PDU transmission of userB will influence that of userA and such influence will increase with the increment of SDU length; while in the traditional scheme, the transmissions of two users' SDUs are independent. In other words, in the traditional scheme, no matter how bad the channel of userB is, userA can always transmit 5 PDUs in one frame; while in our scheme, with the increase of the erroneous PDUs of userB, the number of PDUs transmitted by userA may be decreased to be smaller than 5 in the frame. Nevertheless, due to much smaller probability for lots of erroneous PDUs in one frame, the degradation is very small. Overall, the analysis implies that in the proposed algorithm, the first transmission of all PDUs in one SDU will be finished in one or two frames with higher probability; thus, it brings the significant reduction on the delivery delay of the SDU.

V. Simulation Results

In this section, simulation results are presented to evaluate the introduced theoretical analysis model and to demonstrate the performance of the proposed bandwidth allocation algorithm.

In the simulation, an IEEE 802.16 network with OFDM as PHY specification has been considered. The applied OFDM consists of 256 subcarriers, 192 of which is used for data transmission. Each PDU experiences i.i.d. transmission error with probability $p = 0.1$. At the MAC layer, the length of SDU, L, varies from 4 to10 PDUs. Each frame has duration $T_f = 10$ms and each data burst has duration $T_s = 685\mu$s for the transmission of one PDU with 1200 bytes, i.e., the transmission symbol duration is 13.7μs. The simulation duration is 100s, and 10 slots in each DL subframe are used for users' data.

Fig. 6 shows both analytical and simulation results of the average delivery delay for one tagged SDU. In the simulation, T_{DH} is set to be 1ms. The x-axis represents the number of PDUs which can be transmitted in the tagged burst, i.e., L_b. Actually, it is the bandwidth allocated to the tagged user. Three curves are corresponding to results based on three different SDU lengths, L=4, 5, and 6. The simulation results show that the analysis and the simulation match pretty well. Given the SDU length L, the transmission of one SDU will be finished in fewer frames by increasing the allocated bandwidth L_b. Meanwhile, the delay increases with the increasing of L.

In order to test the analytical model with time variant channel error rate, a similar simulation is carried out by employing a two-state Markov channel, which is represent by a 4-tuple $(p_0, p_1, r_{01}, r_{10})$. p_0 and p_1 denote the error rates at state 0 and state 1, respectively, and r_{01} and r_{10} are corresponding state transition probabilities. Given $(p_0, p_1, r_{01}, r_{10})$, the average

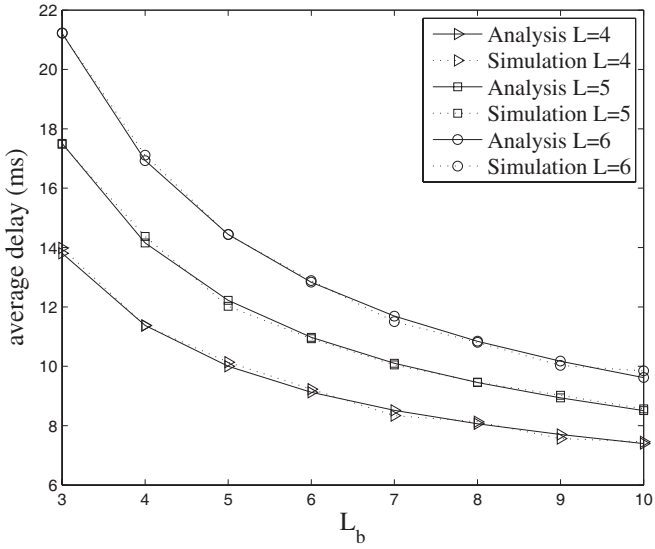


Fig. 7. Comparison of analysis and simulation results (a two-state Markov model).

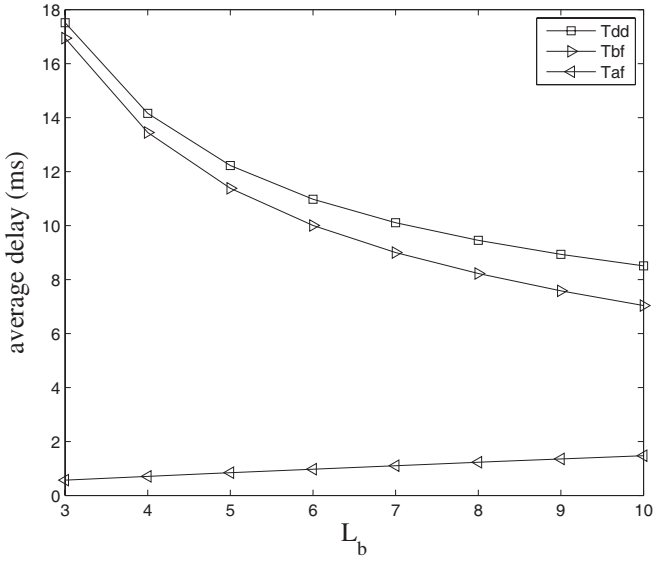


Fig. 8. Comparison of the delivery delay.

error rate of PDU is given by

$$p = \frac{r_{01}p_1 + r_{10}p_0}{r_{01} + r_{10}}. \tag{41}$$

Fig. 7 shows the performance comparison between the analytical results based on the static channel assumption and the simulation results from a two-state Markov channel. In the figure, the numerical results are obtained based on the two-state Markov channel with $(p_0, p_1, r_{01}, r_{10}) = (0, 1, 0.1, 0.9)$, while the analytical results are calculated by replacing average error rate from (41) in the proposed mathematical model. From the figure, it can be seen that the analytical results match pretty well with the simulation ones.

To further understand the delay performance of ARQ-SA, in Fig. 8, the two components of the delivery delay, T_{bf} and T_{af}, are demonstrated separately with $L = 5$. It can be observed that when increasing L_b, both T_{dd} and T_{bf} decrease significantly, while T_{af} only increases a little. It

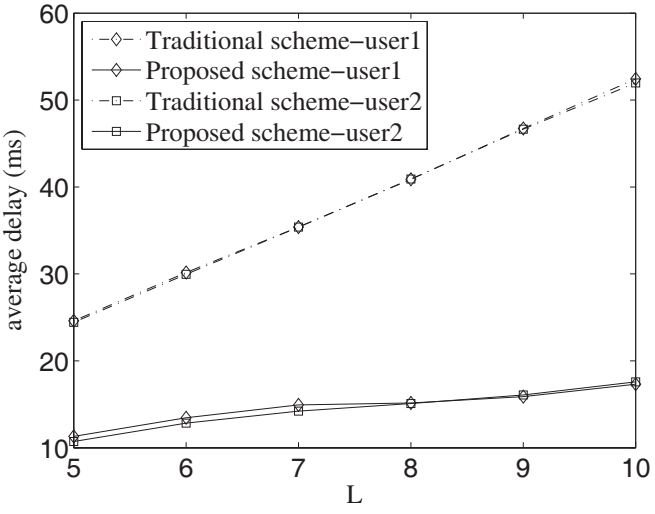


Fig. 9. Delay comparison (homogeneous case).

is because when increasing L_b, the first transmission of one SDU will be finished in fewer frames so that T_{bf} can be decreased significantly. However, T_{af} is determined by the retransmission of the last E PDUs in the frame where the first transmission of the last tagged PDU happens. Therefore, when increasing L_b, E will be probably increased, and it results in the increment of T_{af}, although the amount is not significant. In summary, we can conclude that T_{bf}, i.e., the time used for the first transmission of SDU, plays a key role in determining the delivery delay of SDU.

The performance of the proposed bandwidth allocation algorithm are shown in Fig. 9. Five users are simulated but only the average SDU delay of user1 and user2 are presented. Similar results are obtained for other users. The bandwidth requirements of all users are equal, i.e., a homogeneous case. Thus, each user is allocated 2 slots in the traditional scheme. It can be seen that the SDU delivery delay is significantly improved with the proposed algorithm compared to the traditional one. Such performance gain increases with the increasing of L. In addition, since two users experience almost same delivery delay, the proposed algorithm can provide a similar fairness performance as that of the traditional scheme.

To demonstrate the performance of the proposed algorithm under heterogeneous case, where each user has different bandwidth requirement, similar simulation is carried out and the results are shown in Fig. 10. The average required bandwidth for user1 to user5 are 5, 2, 1, 1 and 1 slots, respectively. From the figure, the delay of SDU for the proposed algorithm still significantly outperforms the traditional scheme. In addition, it can be seen that the fewer the bandwidth required, the more the delay reduction is observed in the proposed algorithm.

VI. Conclusion

In this paper, theoretical model for analyzing delivery delay performance in an IEEE 802.16 network under PMP mode has been derived. The effects of ARQ protocol on the system performance is evaluated, which indicates that the delivery delay of the SDU is mainly determined by the time duration

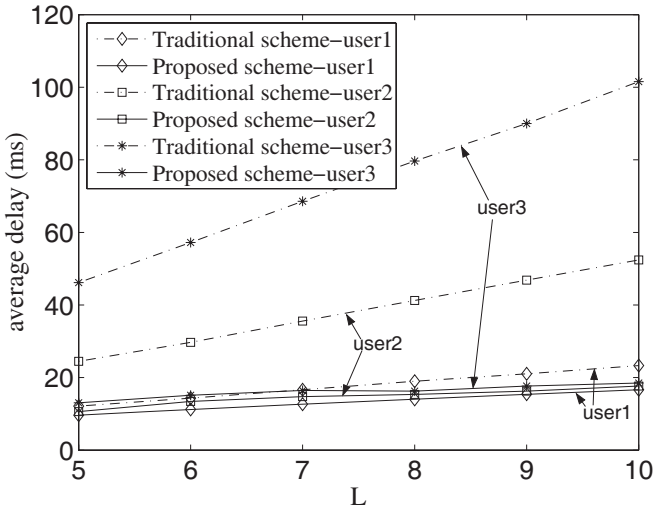


Fig. 10. Delay comparison (heterogeneous case).

used for the first transmission of all PDUs from this SDU. Based on this, a dynamic bandwidth allocation algorithm has been proposed. Both analytical and simulation results demonstrate that the proposed bandwidth allocation algorithm can significantly reduce the delivery delay, while keeping the fairness among different users.

APPENDIX A
PSEUDO-CODE OF THE PROPOSED BANDWIDTH ALLOCATION ALGORITHM

input parameters:
$L_{sA}=0,\ L_{sB}=0,\ (N^i, L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$
set
$(N, L_{pA}, L_{fA}, L_{pB}, L_{fB}) = (N^i, L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$
while $(N>0)$
  using (24) **and** (25)
    get the two-level priority of two users: $(P^1_A, P^2_A, P^1_B, P^2_B)$
  using (26)
    allocate bandwidth for two users (A, B)
  using (27) **and** (28)
    refresh $(N, L_{pA}, L_{fA}, L_{pB}, L_{fB})$
end while
output parameters: $L_{sA},\ L_{sB}$
using (29)
  get $(L^{i+1}_{pA}, L^{i+1}_{fA}, L^{i+1}_{pB}, L^{i+1}_{fB})$

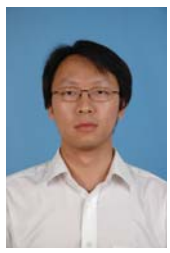
APPENDIX B
DERIVATION OF $\mathbf{\Omega}$

We use an iterative method to obtain all elements of $\mathbf{\Omega}$. The details are as follows.

1) Set $\mathbf{\Omega} = \{(S, L_A, L_{fA0}, L_B, L_{fB0})\}$;
2) Set $\mathbf{\Omega}_{tmp} = \mathbf{\Omega}$;
3) Choose an element from $\mathbf{\Omega}_{tmp}$ one by one, and use it as the input to the proposed algorithm (as shown in Appendix A), i.e., $(N^i, L^i_{pA}, L^i_{fA}, L^i_{pB}, L^i_{fB})$;
4) The possible output parameters, i.e., $(L^{i+1}_{pA}, L^{i+1}_{fA}, L^{i+1}_{pB}, L^{i+1}_{fB})$, with each possible value of N^{i+1} $(N^{i+1} \in [0, S])$ are $((N^{i+1}, L^{i+1}_{pA}, L^{i+1}_{fA}, L^{i+1}_{pB}, L^{i+1}_{fB}))$. Those that missed in $\mathbf{\Omega}$ are added into $\mathbf{\Omega}$ as the new elements;
5) Repeat steps 3 and 4 until all elements in $\mathbf{\Omega}_{tmp}$ are chosen;
6) If the updated $\mathbf{\Omega}$ is not equal to $\mathbf{\Omega}_{tmp}$, go back to step2; otherwise, stop.

REFERENCES

[1] IEEE Standard 802.16-2004, "IEEE Standard for local and metropolitan area networks–part 16: air interface for fixed broadband wireless access systems," 2004.
[2] Y. Xiao, X. Shen, and H. Jiang, "Optimal adaptive ACK mechanism of the IEEE 802.15.3 MAC for ultra-wideband system," *IEEE J. Select. Areas Commun.*, vol. 24, no. 4, pp. 836–842, 2006.
[3] M. Rossi, L. Badia, and M. Zorzi, "Exact statistics of ARQ packet delivery delay over Markov channel with finite round-trip delay," in *Proc. IEEE Globecom 2003*, San Francisco, CA, pp. 3356–3360, Dec. 2003.
[4] M. Rossi, L. Badia, and M. Zorzi, "On the delay statistics of SR ARQ over Markov channels with finite round-trip delay," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1858–1868, July 2005.
[5] M. Rossi, L. Badia, and M. Zorzi, "SR ARQ delay statistics on N-state markov channels with non-instantaneous feedback," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1526–1536, June 2006.
[6] Z. Rosberg and N. Shacham, "Resequencing delay and buffer occupancy under the Selective Repeat ARQ," *IEEE Trans. Inform. Theory*, vol. 35, pp. 166–173, Jan. 1989.
[7] Z. Rosberg and M. Sidi, "Selective-Repeat ARQ: the joint distribution of the transmitter and the receiver resequencing buffer occupancies," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1430–1438, Sept. 1990.
[8] J. Chang and T. Yang, "End-to-end delay of an adaptive selective repeat ARQ protocol," *IEEE Trans. Commun.*, vol. 42, no. 11, pp. 2926–2928, Nov. 1994.
[9] J. G. Kim and M. M. Krunz, "Delay analysis of selective repeat ARQ for a markovian source over a wireless channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1968–1981, Sept. 2000.
[10] M. E. Anagnostou and E. N. Protonotarios, "Performance analysis of the selective repeat ARQ protocol," *IEEE Trans. Commun.*, vol. 34, no. 2, pp. 127–135, Feb. 1986.
[11] W. Luo, K. Balachandran, S. Nanda, and K. Chang, "Delay analysis of selective-repeat ARQ with applications to link adaptation in wireless packet data systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1017–1029, May 2005.
[12] M. Rossi and M. Zorzi, "Analysis and heuristics for the characterization of selective repeat ARQ delay statistic over wireless channels," *IEEE Trans. Veh. Technol.*, vol. 52, no. 5, pp. 1365–1377, Sept. 2003
[13] J. Cai, X. Shen, and J. W. Mark, "Downlink resource management for packet transmission in OFDM wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1688–1703, 2005.
[14] K. Wongthavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *International J. Commun. Syst.*, vol. 16, pp. 81–96, Feb. 2003.
[15] H. Lee, T. Kwon, and D.-H. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 691–693, Aug. 2005.
[16] J. Chen, W. Jiao, and H. Wang, "A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode," in *Proc. ICC'05*, Seoul, Korea, pp. 3422–3426, May 2005.
[17] L. Le, E. Hossian, and A. Alfa,"Service differentiation in multirate wireless networks with weighted round-robin scheduling and ARQ-based error control," *IEEE Trans. Commun.*, vol. 54, no. 2, pp. 208–215, Feb. 2006.



Weiwei Wang received the B.S. degree in communication engineering from the Beijing Jiaotong University (BJTU), China, in 2003 and is working toward his Ph.D in the field of the future wireless communication at the same university. His research interests include WiMax, B3G, and 4G.

Zihua Guo (S'99-M'01-SM'07) received his Ph.D degree from the Hong Kong University of Science and Technology in August 2001. He received his B.S. and M.S. from the University of Science and Technology of China in 1995 and 1998, respectively. His research interests include wireless communications and networking, 3G and beyond technologies, multimedia communications, signal processing, etc. He has published more than 30 IEEE journal papers and conference papers and hold more than 20 patents in wireless communications and networking. Dr. Guo is the Senior Member of IEEE and he is now with Lenovo Corporate Research.

Xuemin(sherman) Shen (M'97-SM'02) received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and University Research Chair, and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of three books, and has published more than 300 papers and book chapters in wireless communications and networks, control and filtering. Dr. Shen serves as the Technical Program Committee Chair for IEEE Globecom'07, General Co-Chair for Chinacom'07 and QShine'06, the Founding Chair for IEEE Communications Society Technical Committee on P2P Communications and Networking. He also serves as a Founding Area Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS; Editor-in-Chief for PEER-TO-PEER NETWORKING AND APPLICATION; Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY; KICS/IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS, COMPUTER NETWORKS; ACM/WIRELESS NETWORKS; and WIRELESS COMMUNICATIONS AND MOBILE COMPUTING (Wiley), etc. He has also served as Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE WIRELESS COMMUNICATIONS, and IEEE COMMUNICATIONS MAGAZINE. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo. Dr. Shen is a registered Professional Engineer of Ontario, Canada.

Changjia Chen received the M.S. degree in the Electronics Institute, Academia Sinica China in 1982 and the Ph.D degree in electrical and electronics engineering from University of Hawaii at Manoa, HI in 1986. He is currently a Professor at Beijing Jiaotong University. His research interests are in the fields of communication networks and communication protocols. His research now focuses on the measurement and modeling of p2p networks. Dr Chen is fellow of CIC and CIE.

Jun Cai (M'04) received the B.Sc. (1996) and the M.Sc. (1999) degrees from Xi'an Jiaotong University (China) and Ph.D. degree (2004) from University of Waterloo, Ontario (Canada), all in electrical engineering. From June 2004 to April 2006, he was with McMaster University as NSERC Postdoctoral Fellow. Since July 2006, he has been with the Department of Electrical and Computer Engineering, University of Manitoba, Canada, where he is an Assistant Professor. His current research interests include multimedia communication systems, mobility and resource management in 3G beyond wireless communication networks, and ad hoc and mesh networks. He is currently a holder of NSERC Associated Industrial Research Chair.