A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks

David Chuck, Kuan-Yu Chen, and J. Morris Chang, *Senior Member, IEEE*

***Abstract*—The IEEE 802.16 standard is considered to be one of the most promising technologies. Bandwidth reservation is employed to provide quality of service (QoS)-guaranteeing services. A request/grant scheme is defined in the IEEE 802.16 standard. There are two types of bandwidth request (BR) mechanisms, i.e., *unicast polling* and *contention resolution*, which are defined in the standard. As specified, connections belonging to scheduling classes of extended real-time polling service, non-real-time polling service, and best effort have options to make BRs via both mechanisms, depending on the scheduling decision made by the base station (BS). However, most research assumes that only one of them is available and do not take both of them into account. A comprehensive study of both mechanisms is critical for the BS to make an appropriate decision for those connections to achieve better system performance. To the best of our knowledge, this is the first attempt to analyze this issue. There are two major contributions presented in this paper. First, a comprehensive study of both BR mechanisms in terms of bandwidth utilization and delay is provided. Additionally, we propose two practical performance objectives: When the expected delay or target bandwidth utilization is given, how does the BS make a scheduling decision such that the performance of the other metric (either delay or bandwidth utilization) is optimized? As our second contribution, we proposed two scheduling algorithms to find the combination of both mechanisms to meet our objectives. The simulation results show that our scheduling algorithms can always help the BS make a scheduling decision to reach better system performance.**

***Index Terms*—Bandwidth request, contention resolution, IEEE 802.16, unicast polling.**

I. Introduction

The IEEE 802.16 standards (e.g., 802.16-2004 [1]) are considered to be among the critical broadband wireless access (BWA) technologies in fourth-generation networks. The Worldwide Interoperability for Microwave Access (WiMAX), which is based on this family of standards, is designed to facilitate services with high transmission rates for data and multimedia applications in metropolitan areas. The physical (PHY) and medium access control (MAC) layers of WiMAX have been specified in the IEEE 802.16 standard. Many advanced communication technologies such as orthogonal frequency-division multiplexing/orthogonal frequency-division multiple access and multiple input–multiple output are embraced in the standards. Supported by these modern technologies, WiMAX is able to provide a large service coverage, a high-speed data rate, and quality of service (QoS)-guaranteeing services. Because of these features, WiMAX is considered to be a promising alternative for last-mile BWA.

To provide the QoS-guaranteeing services, bandwidth reservation is adopted in the WiMAX network. A request/grant bandwidth allocation is employed for reserving bandwidth. The subscriber station (SS) is required to reserve a sufficient amount of bandwidth from the base station (BS) before any data transmissions. The amount of reserved bandwidth can be reserved or adjusted by the SS via sending bandwidth requests (BRs). There are two types of BRs specified in the IEEE 802.16 standard: *unicast polling* and *contention resolution*. In unicast polling, the BS allocates a small piece of bandwidth to the target SS. This small piece of bandwidth is on the top of reserved bandwidth and contains one or more transmission opportunities (TxOPs), depending on the scheduling policy that the BS enforces. These TxOPs are called *unicast polling TxOPs* in this paper. The target SS can use these TxOPs to send BRs. Moreover, for simplicity, we assume that the unicast polling TxOP is only used to transmit a BR. Contention resolution, on the other hand, requires that each SS independently contends for a TxOP to transmit a BR. The BS schedules an amount of bandwidth, which is divided into several TxOPs, for a group of SSs to send BRs. These TxOPs are called *contention TxOPs*. If the attempt of contention failed, then the SS enters the backoff procedure to prepare the next attempt until reaching the maximum number of attempts.

Each type of BR mechanism (i.e., unicast polling or contention resolution) has its own advantages and disadvantages. In unicast polling, the unicast polling TxOPs are exclusively allocated for the target SS, which guarantees that this SS has opportunities to make BRs successful. Therefore, the delay of the SS to transmit a BR can be bounded within a certain range. However, because of the exclusive usage, the allocated unicast polling TxOPs are wasted if the target SS does not make BRs. This may reduce the bandwidth utilization of the system. In contention resolution, on the other hand, the allocated bandwidth is shared by a group of SSs. The SS contends with each other to get a contention TxOP for the BR. In the contention resolution, each SS actively contends for a contention TxOP. Therefore, the SS performs the contention procedure only if the SS wants to transmit a BR, which may lead to higher bandwidth utilization. However, each SS cannot be guaranteed to have contention TxOPs to send BRs. Thus, the delay to request bandwidth cannot be ensured.

Manuscript received July 25, 2009. First published January 15, 2010; current version published May 14, 2010. The review of this paper was coordinated by Dr. T. Taleb.

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: chuang@iastate.edu; kychen@iastate.edu; morris@iastate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2010.2040642

Support for QoS is a fundamental part of the IEEE 802.16 MAC-layer design. When the service data unit arrives in the IEEE 802.16 MAC layer, the classification process is performed. The classification process is the process that maps the service data unit to the appropriate scheduling class based on the QoS constraints of the service data unit. As specified in the IEEE 802.16 standard, only connections belonging to three scheduling classes [i.e., extended real-time polling service (ertPS), non-real-time polling service (nrtPS), and best effort (BE)] are allowed to have the option to choose between unicast polling and contention resolution for making BRs. Because of the features of each BR mechanism, a scheduling decision made by the BS for the connections in these scheduling classes to transmit BRs may affect the overall bandwidth utilization and delay. For example, unicast polling may result in low bandwidth utilization when the probability of an SS to make a BR is low. Similarly, contention resolution may lead to a large number of collisions when the probability that an SS makes BRs is high. The motivation of this research is "how does the BS schedule those two types of BR mechanisms to serve the SS while maintaining good system performance?" An appropriate decision made by the BS is needed to achieve the desired performance objectives. Thus, the impact of this research is to help the BS make scheduling decisions between the two types of BR mechanisms specified in the standard to meet our performance objectives.

There are two proposed performance objectives considered in this paper: 1) maximizing the bandwidth utilization while satisfying the desired delay and 2) minimizing the delay while maintaining the target bandwidth utilization. To achieve the performance objectives, respectively, two scheduling algorithms are proposed in Section V: $MAX-U$ (for the first objective) and $MIN-D$ (for the second objective). Much research related to those two BR mechanisms has only focused on the optimization of one type of BR mechanisms based on the assumption that only one type of BR mechanisms is available to be used. A comprehensive study considering both mechanisms is desired for the BS to schedule the connections that are allowed to send BRs via both mechanisms. In this paper, we provide a mathematical analysis for both BR mechanisms. Based on the analysis, we proposed two scheduling algorithms for performance objectives to help the BS make an appropriate scheduling decision such that the system can have better performance.

The rest of this paper is organized as follows. An overview of IEEE 802.16 and the related work are provided in Sections II and III, respectively. Our mathematical analysis of both BR mechanisms is given in Section IV. In Section V, we introduce the objectives and proposed scheduling algorithms. Section VI presents the simulation, and Section VII concludes our discussion.

II. Overview of IEEE 802.16

An IEEE 802.16 network is composed of a number of SSs and at least one BS. There are two operational modes, i.e., point to multipoint (PMP) and mesh, defined in the IEEE 802.16 standard. This paper is focused on the PMP mode, which defines that transmissions are only allowed between the BS and the SSs. All transmissions can be classified into downlink (DL) and uplink (UL) transmissions based on the direction of transmissions. The DL transmission is defined as the transmission from the BS to an SS. Conversely, the UL transmission is the transmission in the opposite direction. According to the IEEE 802.16 standard, the BS is responsible for scheduling both UL and DL transmissions. All scheduling behavior is expressed in a MAC frame.

The structure of a MAC frame defined in the IEEE 802.16 standard can be divided into the UL subframe and the DL subframe. The UL subframe is for UL transmissions. Similarly, the DL subframe is for DL transmissions. In an IEEE 802.16 network, all SSs should be coordinated by the BS. All coordinating information, including burst profiles and offsets, is resided in the DL and UL maps, which are broadcast at the beginning of the MAC frame.

The IEEE 802.16 network is connection oriented. It requires SSs to establish connections with the BS before any data transmissions. To support a wide variety of applications, the IEEE 802.16 standard classifies all traffic into five scheduling classes based on the different QoS requirements: unsolicited grant service (UGS), real-time polling service (rtPS), ertPS, nrtPS, and BE.

The mechanism to make BRs for each scheduling class has been specified in the IEEE 802.16 standard. For example, a fixed amount of bandwidth is given to UGS connections, and BRs are prohibited to be made for this type of connections. All connections in other scheduling classes (i.e., rtPS, ertPS, nrtPS, and BE) are allowed to make BRs via unicast polling opportunities. However, ertPS, nrtPS, and BE connections are the only connections that are allowed to request bandwidth via contention resolution.

The operation procedure of unicast polling defined in the IEEE 802.16 standard is straightforward. The BS allocates an extra piece of bandwidth to the target SS. This extra piece of bandwidth can be considered to be one or more unicast polling TxOPs. The target SS makes BRs by utilizing these TxOPs. Since these TxOPs are exclusively allocated to this particular SS, it can ensure that this SS has opportunities to request bandwidth if needed. However, the drawback is that these TxOPs are wasted if this SS does not make BRs.

The contention resolution, on the other hand, is not TxOP guaranteed, which means that the SS may have no opportunities to transmit BRs due to failures of contention. The BS schedules a few contention TxOPs for a group of SSs. Each SS within this group is required to contend for a contention TxOP with each other to transmit a BR. Note that each contention TxOP can only carry one BR. If the SS fails in the contention procedure, it enters the back-off procedure for preparing the next attempt. In this paper, the binary exponential back-off (BEB) algorithm [11] is employed as the back-off procedure. The initial back-off window size and the maximum back-off window size are controlled by the BS and specified in the UL map. The value of the contention window size is represented as a power-of-two value. For example, a value of 4 indicates that the contention window size is 16.

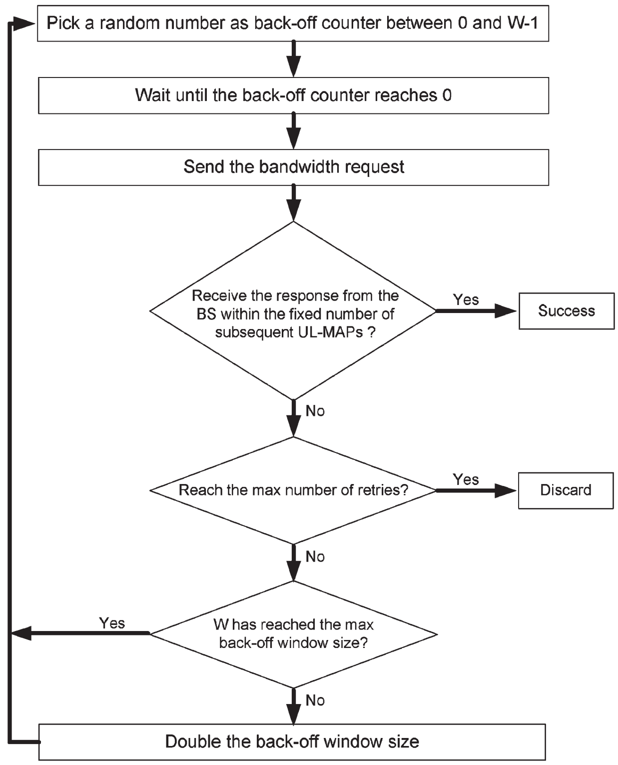


Fig. 1. Operation of contention resolution.

The operation procedure of contention resolution is summarized in Fig. 1. When an SS tends to contend for a TxOP, it selects a random number from 0 to $W-1$, where W is the current back-off window size. This random number is called the back-off counter and indicates the number of contention TxOPs that the SS shall defer before transmitting. The number of contention TxOPs is determined by the BS and may be different in each frame. If the back-off counter does not reach zero within a contention period, its countdown should be frozen at the end of the contention period and resume at the beginning of the next coming contention period.

When the back-off counter reaches zero, the SS attempts to send a BR. It is possible that there is more than one SS whose back-off counter reaches zero at the same time, which means that there are more than one SS trying to send a BR in the same TxOP. In this case, collision happens. Since it is not practically possible for SSs to sense the UL channel to detect a collision, the SS can only know the success of the BR transmission if it receives a response from the BS in the form of a bandwidth grant within a fixed number of subsequent UL map messages. If the SS fails to receive the response, it considers that the BR was not successfully delivered. The SS shall double its back-off window size if the current contention window size is smaller than the maximum back-off window size, which is controlled by the BS. The SS selects a fresh random number from 0 to $W'-1$, where W' indicates the new back-off window size and repeats the deferring procedure described earlier. The SS can attempt to transmit BRs until the maximum number of retries is reached.

III. Related Work

Much research related to unicast polling and contention resolution have been proposed in the literature. In [3], an adaptive polling scheme for ON/OFF traffic was proposed to improve the bandwidth utilization for unicast polling. During ON periods, polling intervals are fixed and short, while during OFF periods, polling intervals are exponentially lengthened. Therefore, adaptive polling reduces the signaling overhead without significantly compromising delay performance. A Markov chain (MC) model for unicast polling is proposed in [4]. The authors proposed the MC analysis that aims to minimize average polling delay while increasing network throughput. Based on the QoS requirements of each scheduling class, the priorities can be given between scheduling classes. However, this obtains a reward only from high-class services because the priority does not differentiate the priorities of nodes.

Contention resolution has been discussed not only in IEEE 802.16 but also in IEEE 802.11. A classic MC model to analyze contention resolution in IEEE 802.11 has been proposed in [7]. Because the bandwidth reservation is employed in the IEEE 802.16 standard, it is not practical for the SS in the IEEE 802.16 network to sense the medium status. Instead, the SS in the IEEE 802.16 network waits for a fixed number of subsequent UL maps to receive the response from the BS before entering the back-off procedure. By considering this difference, a Markov model of contention in the IEEE 802.16 network is proposed in [6]. This model consists of two types of states: back-off states and waiting states. The former illustrate the contention procedure. The latter represent the status that the SS waits for the response from the BS before entering the back-off procedure. The parameters that control the contention resolution in the IEEE 802.16 network, such as back-off start/end values, have been investigated in [2]. Moreover, the connections belonging to three types of scheduling classes (i.e., ertPS, nrtPS, and BE) are able to join the contention resolution. The connection in each scheduling class has its own QoS requirements. However, there are no priorities employed in the contention resolution since the BS fixes the initial and maximum back-off windows, and each SS in the system uses the same value for all connections. To distinguish the priorities between the connections in different scheduling classes, a modified contention-resolution process is proposed [8] to improve system performance, including end-to-end delay and throughput, by assigning different initial window sizes to the connection in different scheduling classes.

The research summarized earlier provides the investigation of either unicast polling or contention resolution. However, the connections in the scheduling classes of ertPS, nrtPS, and BE are allowed to use both BR mechanisms (i.e., unicast polling and contention resolution). Unfortunately, none of the research shown earlier takes this option into considerations. Their research is based on the assumption that only one BR mechanism is available. Research considering both BR mechanisms is presented in [10]. The authors first compare two bandwidth request mechanisms specified in the standard. Their results demonstrate that the contention resolution outperforms unicast polling when the probability of making bandwidth requests is low. However, the authors do not provide detailed analysis for all types of bandwidth request mechanisms. Moreover, the scheduling algorithms to help the BS make scheduling decisions are desired.

TABLE I
LIST OF NOTATIONS FOR UNICAST POLLING

Notation	Description
N	Total number of SSs
N_p	Number of SSs with unicast polling TxOPs
FPS	Number of frames per second
M_p	The minimum average number of unicast polling TxOPs per frame
T_p	The expected delay
Pr	The probability of each SS to send BR
U_p	Bandwidth Utilization of unicast polling

In this paper, two major contributions are included. First, a comprehensive study of both BR mechanisms is provided. We perform the performance analysis of each BR mechanism in terms of the bandwidth utilization and delay. Second, two performance objectives are proposed. To achieve each of our proposed performance objectives, two scheduling algorithms are proposed to reach them individually. The simulation results presented in Section VI show that our scheduling algorithms can also have better performance while the corresponding performance objectives are satisfied.

IV. ANALYTICAL MODELING

In this section, we analyze the performance of each BR mechanism in terms of the bandwidth utilization and the delay of delivering a BR. The network model used for analyzing both BR mechanisms is composed of a BS residing at the center of geographical area and N SSs randomly distributed in the service coverage of the BS. Each SS serves one identical variable-bit-rate (VBR) traffic, based on the traffic model introduced in [5], which is classified as a BE connection with the average probability Pr to transmit BRs. Additionally, we assume that each SS transmits at most one BR during the expected delay. This assumption is reasonable since there is no maximum delay requirement in BE connections, and our objective is to make sure that the average delay is no more than the expected delay. Although piggybacking defined in the IEEE 802.16 standard is another way for SSs to transmit BRs, it is optional and not able to carry all types of BRs. Consequently, we do not consider piggybacking in this paper.

A. Unicast Polling

We begin our analysis of unicast polling by investigating the minimum average number of unicast polling TxOPs allocated in each frame and the average delay of transmitting a BR. For ease of reference, a list of important notations is summarized in Table I.

Assume that N_p is the total number of SSs assigned with unicast polling TxOPs, where $0 \leq N_p \leq N$. Since it is not necessary to schedule a unicast polling TxOP to each SS in every frame, we focus on the minimum number of unicast polling TxOPs that should be scheduled per frame to achieve the expected delay. Assume that the probability of the SS to make a BR is uniformly distributed between two consecutive unicast polling TxOPs. To maintain the expected delay, which is denoted as T_p, the BS has to schedule at least one unicast polling TxOP to the SS in every $2T_p$. Consequently, the minimum average number of unicast polling TxOPs assigned to each frame can be expressed as

$$M_p \geq \frac{N_p}{2\mathtt{FPS}T_p} \tag{1}$$

where M_p stands for the minimum average number of unicast polling TxOPs scheduled in each frame. Because of the nature of unicast polling, the unicast polling TxOP is wasted if the assigned SS does not transmit a BR. Therefore, the bandwidth utilization of unicast polling is same as the probability of an SS to transmit a BR (i.e., $U_p = Pr$).

B. Contention Resolution

We analyze the contention resolution in the IEEE 802.16 network by using a 2-D MC model in Fig. 2. As shown in the figure, each SS attempts to transmit a BR until the number of attempts reaches the maximum retry limit R. If the SS cannot successfully transmit a BR in R attempts, this BR shall be discarded. A list of important notations is summarized in Table II.

According to the specification of the contention resolution procedure described in the IEEE 802.16 standard, the SS shall select a random value within its back-off window. This random number indicates the number of contention TxOPs that the SS shall defer before transmitting a BR. After the contention transmission, the SS has to wait for a fixed number of subsequent UL maps before entering the back-off procedure. Therefore, the contention resolution procedure is classified into two planes: the back-off plane and the waiting plane. The back-off plane describes how the SS transmits a BR (i.e., BEB in this paper). After transmitting a BR, the SS should wait for the response from the BS. The waiting plane is used to represent this waiting period. In Fig. 2, all states in the back-off plane and waiting plane are denoted as ellipses and rectangles, respectively.

In back-off plane, each back-off state, which is denoted as $b(i, r_i)$, represents the ith attempt of sending a BR with a randomly chosen back-off counter r_i. This 2-D MC modeling is possible if we assume an independent and constant probability of an unsuccessful request p for each attempt. It is intuitive that this assumption results in better accuracy as long as the back-off window size W and the number of SSs with contention resolution TxOPs N_c get larger. The correctness of this assumption has been proven in [6]. We refer to p as the *conditional collision probability* [7]. An SS starts to transmit a BR when its back-off counter is equal to 0, regardless of the back-off stage. Once the independence is assumed, p is supported to be a constant value.

After a BR is transmitted, the SS enters the waiting plane, which represents that the SS waits for a response from the BS. According to the IEEE 802.16 standard, the SS should consider that the transmission failed if it does not receive a response from the BS within the number of subsequent UL-MAP messages specified by the parameter of contention-based reservation timeout. Here, we use T_w to represent the

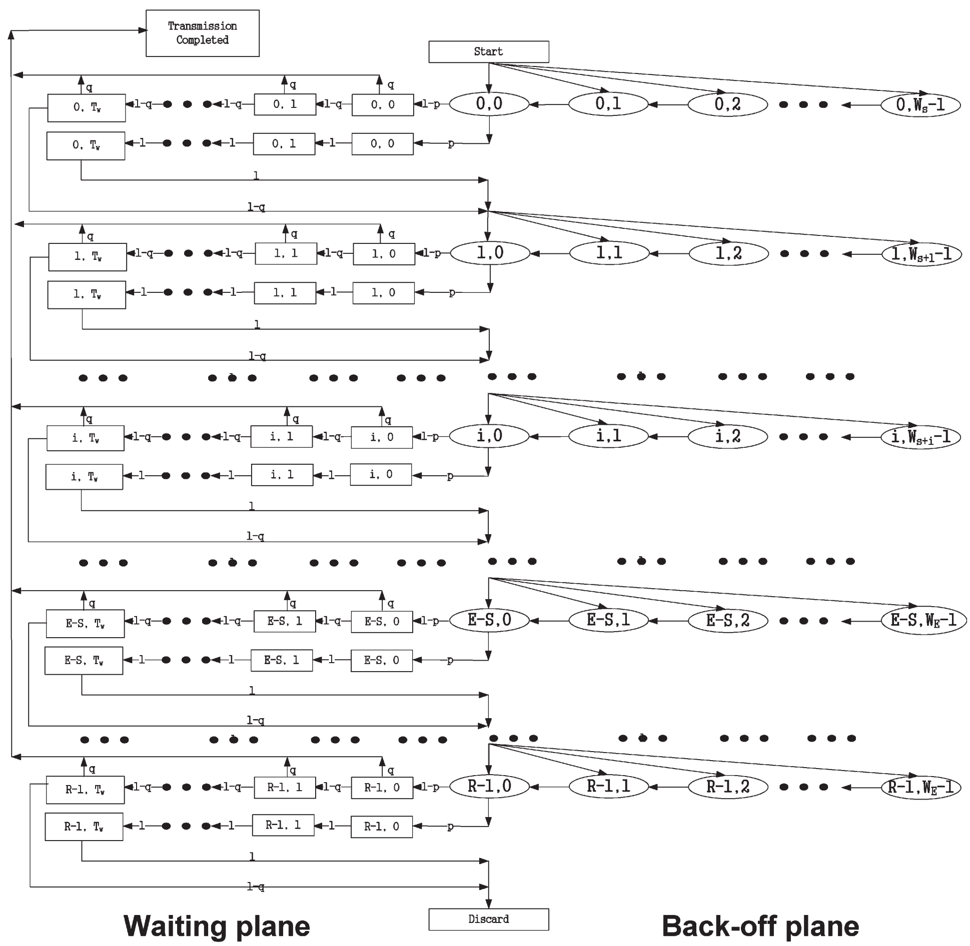


Fig. 2. MC model for contention resolution.

TABLE II
LIST OF NOTATIONS FOR CONTENTION RESOLUTION

Notation	Description
N	Total number of SSs
N_c	Number of SSs with contention TxOPs
FPS	The number of frames per second
M_c	The minimum average number of TxOPs for contention resolution per frame
T_c	The target delay of contention resolution
Pr	The probability of a SS to send BR
S	Back-off start value
E	Back-off end value
p	Probability of a unsuccessful transmission
W_S	Initial back-off window size
W_E	Maximum back-off window size
R	Maximum number of retries
q	Probability of the BS to accept a BR
$b(i, r_i)$	A back-off state in i-th attempt with random back-off counter r_i
$^\dagger w_1(i, t_i)$ $^\ddagger w_2(i, t_i)$	A waiting state in the branch of †collision/ ‡non-collision in i-th attempt and the SS has waited for t_i frames
p_f	the probability of failures
τ	The probability of a SS to transmit a BR in a randomly chosen TxOP
$^\dagger T_1$ $^\ddagger T_2$	The expected delay †before/‡after the contention window size reaches the W_E.
T_w	The maximum number of subsequent UL-MAP messages that a SS waits for a response from the BS

maximum number of subsequent UL-MAP messages for which the SS can wait before entering into the back-off procedure. There are two possibilities in which the SS cannot receive a response within T_w subsequent UL-MAPs: 1) The BR collided with another BR sent from other SSs, and 2) the BR is rejected by the BS. Based on these two possibilities, the waiting states are classified into two branches: collision and noncollision. The states in the collision branch and noncollision branch are represented by $w_1(i, t_i)$ and $w_2(i, t_i)$, respectively, where i is the ith attempt, and t_i is the number of subsequent UL-MAP messages for which the SS has waited after transmitting a BR.

As mentioned, p is the probability of an unsuccessful request. Thus, the probability of entering the branch of collision is also p. It can be obtained that the probability of transition between all states in the branch of collision is 1 due to the failure of the BR transmission. Intuitively, the probability of entering the states in the branch of noncollision is $1 - p$. It is possible that the BS successfully receives a BR but rejects it due to the lack of radio resources or violation of its scheduling policies. Suppose q is the probability of the BS to *accept* a BR in each frame. It is reasonable to assume that q is a constant for the waiting states of this 2-D MC model. In fact, q is controlled by the policy of admission control and is independent of the operation of the MAC layer.

By combining these two factors that may cause failures of BR transmissions (collision and rejection by the BS), the *probability of failures*, which is denoted as p_f, can be represented as

$$p_f = p + (1-p)(1-q)^{T_w}. \tag{2}$$

Here, p is the probability of entering the branch of collision. $(1-p)(1-q)^{T_w}$ denotes the probability of entering the branch of noncollision but having no response received from the BS. It leads to the following observation:

$$b(i,0) = p_f \cdot b(i-1,0), \qquad 0 < i \le R \tag{3}$$

$$\begin{cases} P\{b(i,k) \mid b(i,k+1)\} = 1, & k \in (0, W_i - 2) \quad \text{(4a)} \\ P\{b(i+1,k) \mid b(i,0)\} = \frac{p_f}{W_{i+1}} & \\ \qquad k \in (0, W_{i+1} - 1), & i \in (0, E-S-1) \quad \text{(4b)} \\ P\{b(i+1,k) \mid b(i,0)\} = \frac{p_f}{W_E} & \\ \qquad k \in (0, W_E - 1), & i \in (E-S, R-1) \quad \text{(4c)} \\ P\{w_1(i,0) \mid b(i,0)\} = p, & i \in (0, R) \quad \text{(4d)} \\ P\{w_2(i,0) \mid b(i,0)\} = 1-p, & i \in (0, R) \quad \text{(4e)} \\ P\{w_1(i,t_i+1) \mid w_1(i,t_i)\} = 1 & \\ \qquad & i \in (0,R), t_i \in (0, T_w - 1) \quad \text{(4f)} \\ P\{w_2(i,t_i+1) \mid w_2(i,t_i)\} = 1-q & \\ \qquad & i \in (0,R),\ t_i \in (0, T_w - 1) \quad \text{(4g)} \\ P\{b(i+1,r_i) \mid w_1(i,T_w)\} = 1 & \\ \qquad & i \in (0, E-S-1), r_i \in (0, W_i - 1) \quad \text{(4h)} \\ P\{b(i+1,r_i) \mid w_1(i,T_w)\} = 1 & \\ \qquad & i \in (E-S, R-1), r_i \in (0, W_E - 1) \quad \text{(4i)} \\ P\{b(i+1,r_i) \mid w_2(i,T_w)\} = 1-q & \\ \qquad & i \in (0, E-S-1),\ r_i \in (0, W_i - 1) \quad \text{(4j)} \\ P\{b(i+1,r_i) \mid w_2(i,T_w)\} = 1-q & \\ \qquad & i \in (E-S, R-1),\ r_i \in (0, W_E - 1) \quad \text{(4k)} \end{cases}$$

Based on (2) and (3), the probabilities of transition between states shown in Fig. 2 are summarized in (4a)–(4k). Equation (4a) represents the countdown of the back-off counter. Equations (4b) and (4c) illustrate the probability of entering each back-off state while the window size has and has not reached the maximum window size, respectively. The probabilities of entering the branch of collision and the branch of noncollision are shown in (4d) and (4e), respectively. Equations (4f) and (4g) are the probability between states between the branch of collision and the branch of noncollision, respectively. Equations (4h) and (4i) express that the SS enters the back-off procedure from the branch of collision with different contention window sizes. Similarly, (4j) and (4k) express that the SS enters the back-off procedure from the branch of noncollision with different contention window sizes.

Based on the size of contention window, the back-off states can be classified into two types: **Type 1**, in which the size of contention window is smaller than W_E; and **Type 2**, in which the size of contention window has reached W_E. Suppose that $b(i,k_1)$ and $b(j,k_2)$ denote the back-off states in Type 1 and Type 2, respectively. Additionally, $w_1(i,t_i)$ and $w_2(i,t_i)$ stand for the waiting states in the branch of collision and the branch of noncollision, respectively. Suppose that P_{dis} represents the probability that an SS discards a BR because this BR cannot successfully be transmitted in R attempts. Thus, the sum of probabilities in all the states plus P_{dis} must be equal to 1, as shown in the following:

$$\begin{aligned} 1 &= \sum_{i=0}^{E-S} \sum_{k_1=0}^{W_{S+i}-1} b(i,k_1) + \sum_{j=E-S+1}^{R-1} \sum_{k_2=0}^{W_E-1} b(j,k_2) \\ &\quad + \sum_{i=0}^{R-1} \sum_{t_i=0}^{T_w} w_1(i,t_i) \\ &\quad + \sum_{i=0}^{R-1} \sum_{t_i=0}^{T_w} w_2(i,t_i) + p_f \cdot b(R-1,0) \\ &= \sum_{i=0}^{E-S} \sum_{k_1=0}^{W_{S+i}-1} \left(\frac{W_{S+i} - k_1}{W_{S+i}} \cdot b(i,0) \right) \\ &\quad + \sum_{j=E-S+1}^{R-1} \sum_{k_2=0}^{W_E-1} \left(\frac{W_E - k_2}{W_E} \cdot b(j,0) \right) \\ &\quad + \sum_{i=0}^{R-1} \sum_{t_1=0}^{T_w} p \cdot b(i,0) \\ &\quad + \sum_{i=0}^{R-1} \sum_{t_i=0}^{T_w} (1-p)(1-(1-q)^{t_i}) b(i,0) \\ &\quad + p_f \cdot b(R-1,0) \\ &= \frac{b(0,0)}{2} \left\{ \sum_{i=0}^{E-S} p_f^i (1 + W_{s+i}) \right. \\ &\qquad + \sum_{j=E-S+1}^{R-1} p_f^j (1+W_E) + \sum_{i=0}^{R-1} 2T_w \cdot p \cdot p_f^i \\ &\qquad \left. + \sum_{i=0}^{R-1} \frac{2(1-p)\left(1-(1-q)^{T_w}\right)}{q} p_f^i + 2P_f^{R+1} \right\} \\ &= \frac{b(0,0)}{2} \left\{ \left[1 + 2p \cdot T_w + \frac{2(1-p)\left(1-(1-q)^{T_w}\right)}{q} \right] \right. \\ &\qquad \cdot \frac{1-p_f^R}{1-p_f} + \frac{1-(2p_f)^{E-S+1}}{1-2p_f} \cdot W_S \\ &\qquad \left. + 2^{E-S} W_S \frac{p_f^{E-S+1} - p_f^R}{1-p_f} + 2p_f^{R+1} \right\}. \end{aligned} \tag{5}$$

By simplifying (5), we can derive $b(0,0)$, as shown in the following:

$$\begin{aligned} b(0,0) &= 2 \left\{ \left[1 + 2p \cdot T_w + \frac{2(1-p)\left(1-(1-q)^{T_w}\right)}{q} \right] \right. \\ &\qquad \cdot \frac{1-p_f^R}{1-p_f} + \frac{1-(2p_f)^{E-S+1}}{1-2p_f} \cdot W_S \\ &\qquad \left. + 2^{E-S} W_S \frac{p_f^{E-S+1} - p_f^R}{1-p_f} + 2p_f^{R+1} \right\}^{-1}. \end{aligned} \tag{6}$$

The probability that an SS transmits a BR in a randomly chosen contention TxOP can be calculated as the sum of $b(i,0)$, where $0 \le i \le R-1$. This probability, which is denoted as τ, is expressed as

$$\tau = \sum_{i=0}^{R-1} b(i,0) = \frac{b(0,0)}{1-p_f}. \tag{7}$$

As shown in (3), $b(0,0)$ is represented as a function of P_f, which is a function of p presented in (2). Thus, the value of τ stated in (7) can be expressed as a function of the conditional collision probability p, which is unknown in our model. Again, p is the probability that a collision occurs, which is equivalent to the probability of at least two SSs transmitting BRs at the same contention TxOP. Thus, p can be represented as

$$p = 1 - (1-\tau)^{(N_c \cdot Pr)-1} \tag{8}$$

where τ is the probability that an SS transmits a BR at the randomly chosen contention TxOP shown in (7).

By using (7) and (8), we can solve these two unknown values p and τ based on the known values of back-off start and end (i.e., S and E), the probability of an SS to send a BR (i.e., Pr), the probability of a BS to accept a BR (i.e., q), and the number of SSs with contention TxOPs (i.e., N_c).

To analyze the bandwidth usage of contention TxOPs, it is necessary to find the bandwidth utilization U_c, which is defined as the ratio of the number of TxOPs that successfully deliver BRs to the total number of contention TxOPs. To get this ratio, first, we investigate the *probability of transmission*, which is denoted as p_{tx}, which is referred to the probability that at least one SS transmits a BR at a TxOP. This probability can be obtained as

$$p_{tx} = 1 - (1-\tau)^{N_c \cdot Pr}. \tag{9}$$

The probability of a successful transmission, which is denoted as p_{st}, is the probability that a BR is successfully delivered and that the BS grants this BR. This probability can be achieved by using the conditional probability that only one SS transmits a BR at a TxOP and that the BS has enough bandwidth to serve this BR under the condition that at least one transmission is transmitted at this TxOP. Therefore, the probability of a successful transmission can be addressed as

$$p_{st} = \frac{n\tau(1-\tau)^{(N_c P_r)-1}}{p_{tx}} \left(1-(1-q)^{T_w}\right). \tag{10}$$

From (9) and (10), the probability of a TxOP that successfully delivers a BR, which is represented as p_{sbr}, is derived as

$$p_{sbr} = p_{st} \cdot p_{tx} = n\tau(1-\tau)^{(N_c Pr)-1}\left(1-(1-q)^{T_w}\right). \tag{11}$$

Intuitively, the probability that a BR is delivered in a given TxOP is equivalent to the probability of a TxOP being successfully utilized. Consequently, the bandwidth utilization of contention TxOPs U_c is the same as p_{sbr}.

Although the maximum delay requirement is not a necessary requirement for BE connections, in practice, we still hope that the delay can be limited into a certain bound, which is considered as our expected delay T_c. Here, the delay is calculated as the time difference between the time that the SS intends to send a BR and the time that the SS receives a response from the BS. One of the important factors that affect the delay is the number of contention TxOPs scheduled by the BS in each frame. In this paper, we focus on the relation between the minimum average number of contention TxOPs assigned per frame (which is denoted as M_c) and the target delay (which is denoted as T_c). Based on the contention window size, the expected delay can be calculated into two sections: 1) $i \le E-S$, and 2) $E-S < i \le R$, where i is the ith attempt. Let T_1 stand for the expected delay in the first section. It can be calculated as (12). Similarly, the delay of the second section, which is denoted as T_2, can be derived as (13). It is intuitive that the sum of the delay of two sections is at most the target delay, which is represented as T_c. Moreover, in (12) and (13), everything is known except M_c and p, which are the minimum average number of contention TxOPs assigned per frame and the probability of an unsuccessful transmission, respectively. Therefore, we can use $H(M_c, p)$ to represent the total delay as the sum of delay in these two sections. By writing formally, it can be expressed as (14), shown below. These equations are expressed as follows:

$$T_1 = \sum_{j=S}^{E} p_f^{j-S}(1-p_f) \times \left[\frac{1}{W_j \mathtt{FPS}} \cdot \sum_{k=0}^{W_j-1}\left\lceil\frac{k}{M_c}\right\rceil + \sum_{i=0}^{T_w-1}\frac{i}{\mathtt{FPS}}q(1-q)^i + \sum_{m=S}^{j-1}\left(\frac{1}{W_m \mathtt{FPS}}\sum_{k=0}^{W_m-1}\left\lceil\frac{k}{M_c}\right\rceil + \frac{T_w}{\mathtt{FPS}}\right)\right]$$

$$T_2 = \sum_{n=E-S+2}^{R} p_f^{n-1}(1-P_f) \times \left[\frac{1}{W_E \mathtt{FPS}}\sum_{k=0}^{W_E-1}\left\lceil\frac{k}{M_c}\right\rceil + \sum_{i=0}^{T_w-1}\frac{i}{\mathtt{FPS}}q(1-q)^i + \sum_{d=E_S+2}^{n-1}\left(\frac{1}{W_E \mathtt{FPS}}\sum_{k=0}^{W_E-1}\left\lceil\frac{k}{M_c}\right\rceil + \frac{T_w}{\mathtt{FPS}}\right) + T_b\right] \tag{12}$$

Step 1	Find all combinations of (N_p, N_c) such that $N_p + N_c = N$.
Step 2	For each (N_p, N_c), calculate the corresponding (M_p, M_c) and the bandwidth utilization while the target delay, T_D, is satisfied.
Step 3	Return the (N_p, N_c) and the corresponding (M_p, M_c) such that the bandwidth utilization is maximized.

Fig. 3. Steps of MAX-U.

where

$$T_b = \sum_{m=S}^{E-S+1} \left(\frac{1}{W_m \mathtt{FPS}} \sum_{k=0}^{W_m - 1} \left\lceil \frac{k}{M_c} \right\rceil + \frac{T_w}{\mathtt{FPS}} \right) \quad (13)$$

$$T_c \geq T_1 + T_2 = H(M_c, p). \quad (14)$$

V. Scheduling Algorithms for Performance Objectives

Based on the analysis shown in Section IV, we proposed two scheduling algorithms to meet the two performance objectives proposed in this paper, respectively.

1) Maximize the bandwidth utilization under the condition of satisfying a given target delay requirement (represented as *Fixed-delay-MAX-Utilization* in the rest of this section).
2) Minimize the target delay when a given bandwidth utilization as a constraint is given (represented as *Fixed-Utilization-MIN-delay* in the rest of this section).

To meet the first objective, a scheduling algorithm, which is called the *Maximum Bandwidth Utilization Scheduling Algorithm* (MAX-U), is proposed. It helps the BS to schedule the number of TxOPs and the number of participating SSs for each BR mechanism to maximize the bandwidth utilization while satisfying the target delay. Similarly, the scheduling algorithm proposed for the second objective is called the *Minimize Delay Scheduling Algorithm* (MIN-D). It helps the BS to find the combination of TxOPs assigned for each BR mechanism such that the system delay is minimized while maintaining the desired utilization. Note that both scheduling algorithms help the BS schedule either unicast polling TxOPs or contention TxOPs to each SS to achieve the corresponding performance objective. No SSs receive both types of TxOPs at the same time.

In this paper, we only focus on the BR mechanisms. Thus, the bandwidth utilization indicated in this paper is the bandwidth utilization of TxOPs assigned for both BR mechanisms (i.e., unicast polling and contention resolution). Moreover, the TxOPs scheduled for each mechanism are only used to transmit BR messages.

A. MAX-U

This algorithm is designed to satisfy our first performance objective: Maximize the bandwidth utilization while satisfying the fixed delay requirement. The flow of this algorithm is shown in Fig. 3. Suppose that T_D is the given achievable target delay. Our objective is to find the number of unicast polling TxOPs and contention TxOPs scheduled in each frame such that the bandwidth utilization is maximized.

Step 1	Find all combination of (N_p, N_c) such that $N_p + N_c = N$.
Step 2	$\forall$ (N_p, N_c), find all corresponding (M_p, M_c) such that $M_p(1 - P_r) + M_c(1 - p_{sbr}) \leq S_u$ and $M_p \leq N_p$.
Step 3	$\forall$ (N_p, N_c), find the corresponding delay of each (M_p, M_c) and select one with minimum delay.
Step 4	$\forall$ (N_p, N_c), set the delay as the corresponding delay of the picked (M_p, M_c).
Step 5	Return the (N_p, N_c) with minimum delay.

Fig. 4. Steps of MIN-D.

Algorithm 1 MAX-U
Input: All variables specified in Tables I and II
Output:
N_p number of SSs with unicast polling TxOP;
N_c number of SSs with contention resolution TxOP;
M_p average number of TxOPs scheduled for unicast polling per frame;
M_c average number of TxOPs scheduled for contention resolution per frame.
For $i = 0$ to N **do**
$N_p^i \leftarrow i$, $N_c^i \leftarrow N - N_p$
Unicast Polling:
$M_p^i \leftarrow M_p^i$ calculate by (1)
Contention Resolution:
a. Solve τ^i and p^i by using (7) and (8) with given N_c^i.
b. $M_c^i \leftarrow M_c^i$ calculated by (14) with a known p^i.
c. $p_{sbr}^i \leftarrow p_{sbr}^i$ calculated by (11).
Finalize:
$U_t^i \leftarrow (M_p^i Pr + M_c^i p_{sbr}^i / M_p^i + M_c^i)$
End For
$U_t \leftarrow Max\{U_t^i\}$ with $Min\{M_c^i + M_p^i\}$
$M_p \leftarrow M_p^i$, $N_p \leftarrow N_p^i$, $M_c \leftarrow M_c^i$, $N_c \leftarrow N_c^i$
Return $M_p \quad N_p \quad M_c \quad N_c$

In this algorithm, each SS is scheduled with either one of the BR mechanisms: unicast polling TxOPs and contention TxOPs. Suppose that N_c and N_p are the number of SSs scheduled with contention and unicast polling TxOPs, respectively, such that $N_c + N_p = N$. The objective of the algorithm is to maximize the bandwidth utilization while achieving the given target delay. For all combinations of (N_p, N_c), we calculate the corresponding value of M_p and M_c for each combination and select a combination of (N_p, N_c) and the corresponding M_p and M_c that can maximize the bandwidth utilization. Note that the overall throughput may be higher if we can minimize the number of TxOPs assigned for BR mechanisms because there is more bandwidth that can be assigned for data transmissions. Therefore, if there are multiple combinations that result in the same maximum bandwidth utilization, the one with the minimum number of TxOPs (i.e., $M_c + M_p$) is selected.

B. MIN-D

This algorithm focus on achieving our second performance objective: minimizing the delay while satisfying a given bandwidth utilization requirement. The detailed steps of this algorithm are presented in Fig. 4. Assume that U_t is the given bandwidth utilization with a fixed number of TxOPs S_t for both BR mechanisms (i.e., $M_p + M_c = S_t$). Thus, the number of unused TxOPs, which is denoted as S_u, can be represented as

$$S_u = (1 - U_t)S_t.$$

It is intuitive that the total unused TxOPs of both BR mechanisms are at most S_u. Formally, it can be expressed as

$$M_p(1 - Pr) + M_c(1 - p_{sbr}) \leq S_u. \quad (15)$$

Similar to Algorithm 1, we examine all combinations of (N_p, N_c) such that $N_p + N_c = N$. Our objective is to find a combination of (N_p, N_c) with the minimum overall expected delay while (15) is satisfied.

Algorithm 2 MIN-D
Input: All variables specified in Tables I and II
Output:

N_p number of SSs with unicast polling TxOPs;
N_c number of SSs with contention resolution TxOPs;
M_p average number of unicast polling TxOPs per frame;
M_c average number of contention resolution TxOPs per frame.

For $i = 0$ to N **do**
  $N_p^i \leftarrow i \quad N_c^i \leftarrow N - N_p$
  $K^i \leftarrow$ the set of all combinations of (M_p^i, M_c^i) such that (15) is satisfied and
$M_p^i \leq N_p^i$.
  For j = 1 to $|K^i|$
    Unicast Polling:
    a. $T_p^j \leftarrow$ Calculated by (1).
    Contention Resolution:
    a. Solve τ^i and p^i by using (7) and (8) with given N_c^i.
    b. $T_c^j \leftarrow T_c^j$ Calculated by (14).
  End For
  $T_D^i \leftarrow Min\{Max\{T_p^j, T_c^j\}\}$
  $M_p^i \leftarrow M_p^j$
  $M_c^i \leftarrow M_c^j$
End For
$T_D \leftarrow Min\{T_D^i\}, \quad M_p \leftarrow M_p^i, \quad N_p \leftarrow \quad N_p^i, \quad M_c \leftarrow M_c^i,$
$N_c \leftarrow N_c^i$
Return $M_p \quad N_p \quad M_c \quad N_c$

For each pair of (N_p, N_c), there exist several pairs of (M_p, M_c) that satisfy the constraint stated in (15). Suppose M' is the set of qualified (M_p, M_c) for each pair of (N_p, N_c). Therefore, we check all combinations of $(M_p, M_c) \in M'$ and find the combinations resulting in the delay being minimized as our candidates. Here, delay is defined as $\max\{T_p, T_c\}$, where T_p and T_c are the delay caused by unicast polling and contention resolution, respectively. Consequently, for each pair of (N_p, N_c), there is at least one pair of (M_p, M_c) as our candidates. Among these candidates, we pick one candidate with the minimum delay as our solution for the scheduling decision.

TABLE III
SIMULATION PARAMETERS

Parameters	Value
Number of BS	1
Number of SS	200, 300, 400, 500
Frame Duration	20 ms
Modulation	BPSK, QPSK, 16QAM, 64QAM
TTG/RTG	10 μs
SSTG	4 μs
Application	HTTP
Traffic Type	VBR
Scheduling Class	BE
Mean Packet Size	512 $\sim$ 1024(byte)
Mean Traffic Rate	2Kbps

VI. NUMERICAL AND SIMULATION RESULTS

A. System Setup

In this section, we validate the theoretical results with our simulation results. The theoretical results are made by *Matlab 2009a*. The simulation results are conducted by our simulator. The simulator is written in C and closely followed the IEEE 802.16 standard. Both analytical and simulation results are also compared with two ordinary schemes: 1) unicast polling only and 2) contention resolution only. Table III summarizes the system parameters used in our numerical analysis and simulation. In our simulation, each SS serves one HTTP web browsing traffic [12], [13], which is classified as a BE connection. To increase the variety of BE traffic, the mean packet size is randomly selected from 512 to 1024 B. Because the mean traffic rate is fixed, the mean traffic rate can be calculated based on the selected mean packet size.

B. MAX-U

The target delay used in this simulation is 1 s, which is the most common delay used for BE connections. The results of bandwidth utilization under different Pr are shown in Fig. 5(a) and (b). It is easy to observe that the results of bandwidth utilization are similar with different numbers of SSs. It shows that the bandwidth utilization does not strongly relate to the number of SSs in the system. The utilization of contention resolution only is always around 35%, regardless of the value of Pr. On the other hand, the utilization of unicast polling only is very close to the value of Pr. By these results, we can conclude that unicast polling can achieve better bandwidth utilization if Pr is larger than 0.35. Therefore, it is impossible to always reach better performance if only one BR mechanism is considered.

As shown in the figures, our analytic and simulation results are very close to each other. This validates this analysis presented in Section IV. Additionally, our results always achieve the better bandwidth utilization produced by either unicast polling only or contention resolution only. For example, in Fig. 5(a), our algorithm achieves around 35% of the bandwidth utilization when $Pr = 0.1$, which is similar to the one that

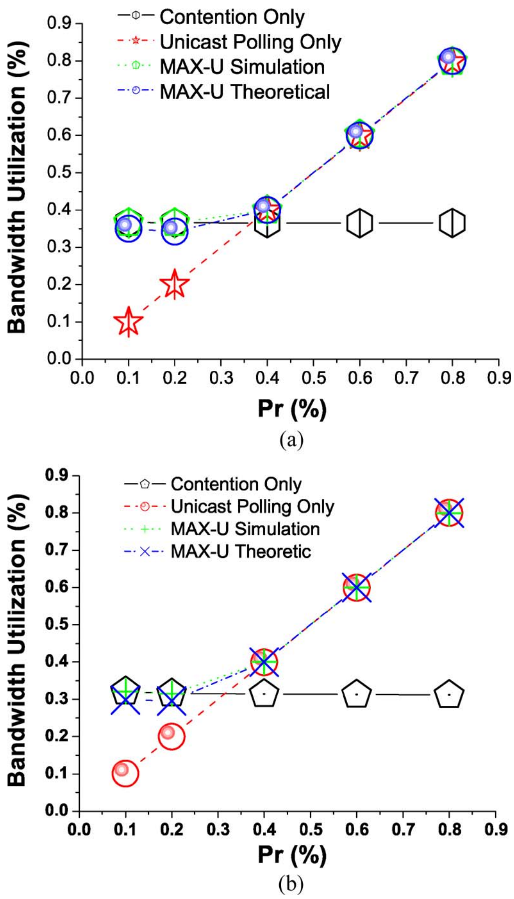


Fig. 5. Simulation results of MAX-U. (a) Number of SSs = 200. (b) Number of SSs = 500.

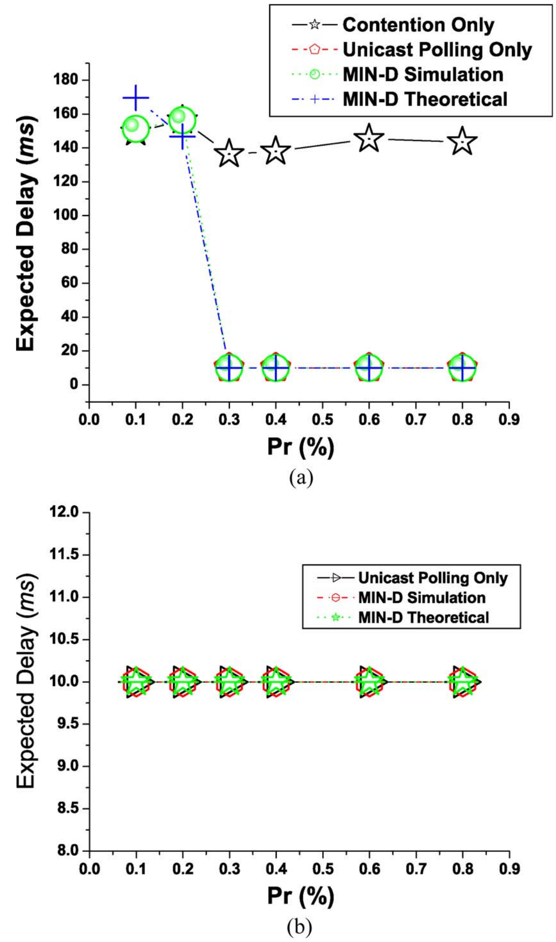


Fig. 6. Simulation results of MIN-D. (a) U = 0.3. (b) U = 0.5.

contention only achieves. However, unicast polling only results in 10% bandwidth utilization. When $Pr = 0.8$, both unicast polling and our algorithm reach 80% bandwidth utilization. Contention only still keeps its bandwidth utilization around 35%. It is because our scheduling algorithm (i.e., MAX-U) can help the BS schedule one type of BR mechanisms that can achieve better performance according to the current network status. It is worth to note that our scheduling algorithm (MAX-U) schedules all SSs with either unicast polling TxOPs or contention TxOPs. The combinations in between (i.e., part of SSs with unicast polling TxOPs and the rest of them with contention TxOPs) do not exist. This is because contention resolution can always give 35% bandwidth utilization, and it will be chosen if unicast polling cannot contribute a bandwidth utilization value as high as it does. On the other hand, unicast polling will always be chosen when it can have more than 35% bandwidth utilization (i.e., $Pr > 35\%$).

C. MIN-D

Fig. 6(a) and (b) show the relationship between the expected delay and Pr, while the target bandwidth utilization is 0.3 and 0.5, respectively. From the figures, we obtain that our scheduling algorithm (i.e., MIN-D) always picks a BR mechanism, resulting in better performance (i.e., shorter delay). For instance, in Fig. 6(a), both unicast polling and our algorithm reach 10-ms delay when $Pr = 0.4$. However, contention only keeps the delay around 145 ms in all values of Pr. In Fig. 6(a), there are no results for unicast polling only when $Pr = 0.1$ and 0.2. This is because the bandwidth utilization cannot achieve 0.3 if only unicast polling is used. Similarly, there are no results for contention only in Fig. 6(b) since the contention resolution cannot reach 50% bandwidth utilization.

TABLE IV
SIMULATION RESULTS OF MIN-D

Pr	N_p	M_p	N_c	M_c
0.1	2	1	498	15
0.2	0	0	500	15
0.3	500	500	0	0
0.4	500	500	0	0
0.6	500	500	0	0
0.8	500	500	0	0

Table IV shows the simulation results of the scheduling algorithm in terms of the number of SSs and the number of TxOPs assigned to each BR mechanism. Here, the target bandwidth utilization is 0.3. It is worth noting that both BR mechanisms are scheduled for BR transmissions when $Pr = 0.1$. This is because the performance requirement (i.e., U = 0.3) cannot be achieved if only one BR mechanism is considered. This result shows an example where better performance can be achieved by scheduling both types of BR mechanisms.

VII. Conclusion

According to the IEEE 802.16 standard, the connections belonging to ertPS, nrtPS, and BE are allowed to make BRs via both BR mechanisms (i.e., unicast polling and contention resolution). The mechanism that the BS schedules to those connections may result in different system performance values because of the nature of each BR mechanism. However, most conventional research works limit the option to consider only one type of BR mechanism. A scheduling scheme that considers both types of BR mechanisms is desired for the BS to optimize system performance. In addition, it is not necessary for the BS to perform either unicast polling or contention resolution to all SSs within one frame. Instead, the BS needs to schedule the appropriate number of contention resolution or unicast polling TxOPs to the SS to meet the delay requirement. Therefore, the scheduling decision should be made in a multiframe basis.

In this paper, we have provided the performance analysis of each BR mechanism in terms of bandwidth utilization and expected delay. Based on the analysis, we have taken both BR mechanisms into account and proposed two scheduling algorithms to help the BS make the scheduling decision based on the current network status such that the corresponding performance objectives are achieved. There are two performance objectives proposed in this paper: 1) maximizing the bandwidth utilization under the condition that the target delay is satisfied and 2) minimizing the delay while the desired bandwidth utilization is reached. Our numerical and simulation have confirmed that the scheduling algorithms can always have better performance by scheduling the number of TxOPs to one of the BR mechanisms. Additionally, when the probability of making BR (i.e., Pr) is 0.1, a hybrid decision (i.e., scheduling SSs with two BR mechanisms) can conduct the minimum delay while satisfying the desired bandwidth utilization.

References

[1] *IEEE Standard for Local and Metropolitan Area Network Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Std. 802.16-2004, 2004.

[2] A. Sayenko, O. Alanen, and T. Hämäläinen, "On contention resolution parameters for the IEEE 802.16 base station," in *Proc. GLOBECOM*, 2007, pp. 4957–4962.

[3] C. Nie, M. Venkatachalam, and X. Yang, "Adaptive polling service for next-generation IEEE 802.16 WiMAX networks," in *Proc. GLOBECOM*, 2007, pp. 4754–4758.

[4] B.-J. Chang, C.-M. Chou, and Y.-H. Liang, "Markov chain analysis of uplink subframe in polling-based WiMAX networks," *Comput. Commun.*, vol. 31, no. 10, pp. 2381–2390, Jun. 2008.

[5] T. G. Robertazzi, *Computer Networks and Systems: Theory and Performance Evaluation*. New York: Springer-Verlag, 1990.

[6] Y. Pourmohammadi, F. Agharebparast, M. R. Minhas, H. M. Alnuweiri, and V. C. M. Leung, "Analytical modeling of contention-based bandwidth request mechanism in IEEE 802.16 wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 5, pp. 3094–3107, Sep. 2008.

[7] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[8] J. Delicado, Q. Ni, F. M. Delicado, and L. Orozco-Barbosa, "New contention resolution schemes for WiMAX," in *Proc. WCNC*, 2009, pp. 1–6.

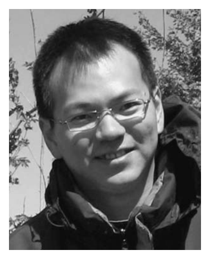
[9] H. Fattah and H. Alnuweiri, "Performance evaluation of contention-based access in IEEE 802.16 networks with subchannelization," in *Proc. ICC*, 2009, pp. 1–6.

[10] Q. Ni, A. Vinel, Y. Xiao, A. Turlikov, and T. Jiang, "Investigation of bandwidth request mechanisms under point-to-multipoint mode of WiMAX networks," *IEEE Commun. Mag.*, vol. 45, no. 5, pp. 132–138, May 2007.

[11] R. M. Metcalfe and D. R. Boggs, "Ethernet: Distributed packet switching for local computer networks," *Commun. ACM*, vol. 19, no. 7, pp. 395–404, Jul. 1976.

[12] M. Molina, P. Castelli, and G. Foodies, "Web traffic modeling exploiting TCP connection temporal clustering through HTML-REDUCE," *IEEE Netw. Mag.*, vol. 14, no. 3, pp. 46–55, May/Jun. 2000.

[13] Z. Sun, D. He, L. Liang, and H. Cruickshank, "Internet QoS and traffic modeling," *Proc. Inst. Elect. Eng.—Softw.: Special Issue Perform. Eng.*, vol. 151, no. 5, pp. 248–255, Oct. 2004.



David Chuck received the B.S. degree in applied mathematics from Tatung University, Taipei, Taiwan, in 2004 and the M.S. degree in computer science from Illinois Institute of Technology, Chicago, in 2007. He is currently working toward the Ph.D. degree in computer engineering with the Department of Electrical and Computer Engineering, Iowa State University, Ames.

In 2009, he was a Research Intern with NTT Network Innovation Laboratories, Yokosuka, Japan. His research interests include wireless networking, network virtualization, resource allocation, and game-theoretical study of networks.

Kuan-Yu Chen received the B.S. degree in electrical engineering from National Chung Cheng University, Chiayi, Taiwan, in 2005. He is currently working toward the M.S. degree in computer engineering with the Department of Electrical and Computer Engineering, Iowa State University, Ames.

In 2008, he was a Research Assistant with Iowa State University. His research interests include voice over Internet protocol, resource allocation, quality of service, and WiMAX technology.

J. Morris Chang (SM'08) received the Ph.D. degree in computer engineering from North Carolina State University, Raleigh.

His industrial experience includes positions at Texas Instruments Incorporated, Microelectronic Center of North Carolina, and AT&T Bell Laboratories. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Iowa State University, Ames. His research interests include wireless networks, performance study of Java virtual machines, and computer architecture.

Dr. Chang received the University Excellence in Teaching Award from Illinois Institute of Technology in 1999. He currently serves as the Handling Editor of the *Journal of Microprocessors and Microsystems* and the Middleware and Wireless Networks Subject Area Editor of *IEEE IT Professional*.