Performance Evaluation for Saturated 802.16 Networks with Binary Backoff Mechanism

Hai L. Vu∗ and Sammy Chan†

∗Centre for Advanced Internet Architectures, Faculty of ICT, Swinburne Univ. of Technology, P.O. Box 218, VIC 3122, Australia. Email: h.vu@ieee.org
†Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong. Email: eeschan@cityu.edu.hk

Abstract—We propose in this paper an analytical model to evaluate the average packet delay and its standard deviation in a saturated IEEE 802.16 network with the binary exponential backoff mechanism. In particular, we develop a simple fixed point analysis to approximate the probability that a bandwidth request is unsuccessful in such a network. A closed form expression for the mean and standard deviation of the delay are then derived from the model as a function of the system parameters. The accuracy of the analytical model is evaluated by comparing with simulation results over a wide range of operating conditions.

I. INTRODUCTION

WiMAX (Worldwide Interoperability for Microwave Access) is one of the candidate technologies for fourth-generation mobile communication networks. Based on the IEEE 802.16-2004 Air Interface standard [1] and the IEEE 802.16e amendment [2], WiMAX can support: 1) A very high capacity; for example, a theoretical peak data rate of 60 Mb/s for a total downlink operation (without uplink) and 28 Mb/s for a total uplink operation can be achieved with the use of two antennas at 10-MHz channel bandwidth. 2) Wide area mobility, a speed of up to 120 km/h with handoff. 3) Multimedia services with different traffic characteristics and various quality of service (QoS) requirements.

The basic operating mode of WiMAX is called point-to-multipoint (PMP) mode, where a base station (BS) serves a set of subscribers (SSs) within the same antenna sector in a broadcast manner. Since multiple SSs share the uplink, to avoid collision, only one SS is permitted to transmit at any given time. To achieve such exclusive access at a given time, each SS needs to be granted time slots before it can transmit. For this purpose, several request/grant mechanisms at the MAC layer have been specified for various scheduling services. They are unsolicited granting, unicast polling and broadcast polling. Among these bandwidth reservation mechanisms, the broadcast polling mechanism is contention based and requires SSs to use the binary exponential backoff (BEB) algorithm for contention resolution. Note that collisions between bandwidth requests from different SSs would lead to unexpected delay and under utilization of the wireless link. Characterizing and evaluating this delay are thus important in order to meet QoS requirements across service classes and to better utilize the resources in WiMAX networks.

There have been several works in the literature studying the performance of 802.16 networks. In [5], the delay performance of the contention-free unicast polling request mechanism was studied analytically. The cases in which nodes are polled sequentially at the beginning or end of each uplink subframe are both modelled. In [7], the delay performance under broadcast polling was investigated. An analytical model based on Markov chain was proposed to evaluate the average delay of a transmission request for saturated networks. An alternate modelling approach was also proposed in [4]. The model in [7] was later extended to investigate the case with Bernoulli request arrival [8]. However, the work in [4], [7] and [8] did not take into account the delay incurred by data packet transmission after a request is successfully transmitted. Also, the delay is measured in terms of transmission frame, instead of measuring the time between the first attempt of a request and the completion of the packet transmission.

In this paper, we develop an analytical model to evaluate the average packet delay and its standard deviation in a saturated WiMAX network when broadcast polling is used for uplink channel allocation. In particular, we propose a simple fixed point analysis to approximate the probability that a bandwidth request is unsuccessful. A detailed delay model is then introduced to study the statistics of the packet delay. A closed form expression for the mean and standard deviation of the delay are derived from the model as a function of the system parameters. We verify the accuracy of our model by comparison with extensive simulation results.

II. IEEE 802.16 MAC PROTOCOL

The MAC frame structure defined by the IEEE 802.16 standard with time division duplexing (TDD) channel allocation in PMP mode is shown in Fig. 1. In this scheme, the channel is time slotted into fixed-length frames; each consists of a downlink and uplink sub-frame. The duration of the downlink and uplink sub-frames is dynamically controlled by the BS via broadcasting the so-called downlink map (DL Map) and uplink map (UL Map) messages at the beginning of each frame (as indicated in Fig. 1). The UL Map contains data or information element that informs the SSs about transmission opportunities, that is, the time slots in the uplink sub-frame where an SS could send a bandwidth request for transmitting its data in the next frame. Upon receiving the bandwidth requests, the BS then allocates bandwidth or data slots for data transmission in the uplink sub-frame based on its scheduler.
III. Analytical Model

Consider an IEEE 802.16 network consisting of $N$ subscribers (SSs) operating in the point-to-multipoint (PMP) mode. Assume that all the SSs are in saturated condition, i.e., they always have packets to send. The packet delay is defined as the time duration from its first bandwidth request until the packet transmission has finished. Note that if the bandwidth reservation of a packet is successful, the packet will be removed from the head of the queue into a transmission queue waiting for its transmission in the coming uplink sub-frame. Under saturation condition an SS will then start a new backoff process for its next packet (which is now at the head of the queue) in the subsequent frame.

A. Unsuccessful Request Probability

Let $p$ be the probability that a request sent by an SS is unsuccessful which can be expressed as follows [4]:

$$p = 1 - (1 - p_c)(1 - p_d),$$

where $p_c$ and $p_d$ are probabilities that a request sent by an SS is unsuccessful due to collision with other requests, or lack of bandwidth (data slots) in the next coming uplink sub-frame, respectively. Note that (1) assumes independency between $p_c$ and $p_d$ which will be validated through simulation in Section IV. And since both $p_c$ and $p_d$ can be expressed as a function of the probability $p$, fixed point equations can be established to calculate the individual probabilities. To this end, our approach to obtain the fixed point analysis is very different from that in [4] and is detailed as follows.

A request is successfully transmitted on the first attempt with probability $1 - p$ (ignoring a normalisation factor that we will introduce later). Recall that the contention window is initially set to $W$, the average number of elapsed backoff slots before such a request is $\frac{mW}{2} + (W - 1)/2$. The first term is due to the fact that an SS can not start a new backoff period for its next request immediately after the previous one in the same frame but has to wait until the next frame. And because requests are uniformly chosen among the $m$ request opportunities in each frame, the average number of backoff slots wasted until the next frame is $\frac{m}{2}$. The second term represents the average number of backoff slots an SS has to wait before attempting to send request based on the contention avoidance mechanism described in Section II.

If the first transmission fails, the request is successfully transmitted on the second attempt with probability $p(1 - p)$. The average number of elapsed backoff slots in this case is $\frac{mW}{2} + (2W - 1)/2$. Continuing this argument until the $n$th attempt yields the overall average number of elapsed backoff
slots before a successful request transmission:

\[
B_{\text{avg}} = \frac{m}{2} + \eta \sum_{i=0}^{r-1} p^i \left( \frac{\lambda \xi W - 1}{2} \right) + \eta \left( \frac{\lambda \xi W - 1}{2} \right) \sum_{i=r}^{R-1} p^i
\]

\[
= \frac{m}{2} + \frac{\eta W(1-(\lambda \xi)^r)}{2(1-\lambda \xi)} - \frac{1-p^r}{2(1-p^R)} + \frac{(\lambda \xi W - 1)(p^r - p^R)}{2(1-p^R)}
\]

where

\[
\eta = (1-p)(1-p^R)^{-1},
\]

and \((1-p^R)\) is a normalisation factor.

Note that assuming a request will be eventually successful, then \(B_{\text{avg}}\) is the average number of backoff slots an SS has to wait before sending requests, i.e. it is an average inter-arrival time of requests in this system. And therefore the probability that an SS attempts to send the request in a slot is given by

\[
\tau = 1/(B_{\text{avg}} + 1).
\]

Given that there are \(N\) saturated SSs, the probability \(p_c\) that a request sent by an SS is unsuccessful due to collision with other requests can be expressed as

\[
p_c = 1 - (1-\tau)^{N-1}.
\]

The probability that there are \(j\), \(0 \leq j \leq k = \min(m,N)\) successful requests among \(m\) request slots/opportunities can be approximated based on a truncated binomial distribution

\[
Q(j) = \sum_{i=0}^{k-j} \binom{m}{i} \xi^i (1-\xi)^{m-i},
\]

where \(\xi = N\tau(1-\tau)^{N-1}\) is the probability that a request sent in a request slot will be successful given that there are \(N\) SSs each attempting to send requests with probability \(\tau\).

The probability that a request is unsuccessful due to lack of bandwidth (data slots) in a subsequent frame can then be expressed as

\[
p_d = \frac{\sum_{j=d+1}^{k} (j-d)Q(j)}{\sum_{j=0}^{k} jQ(j)},
\]

where \(d < k\) is the number of data slots set by BS in the uplink data frame. Equations (1), (3) and (5) create a fixed point formulation from which \(p\) can be computed numerically.

**B. Delay Model**

Let \(U\) and \(V\) be the random variables (RVs) representing the time durations from the time of sending the request until the end of the reservation interval, and from the start of the data transmission in the uplink direction until a packet of the tagged SS has been transmitted, respectively. As shown in Fig. 1 the duration of the uplink sub-frame is given by \(T_{UL} = T_{RE} + T_{DA}\), where \(T_{RE} = mt\) and \(T_{DA} = dT\) are the length of the reservation interval and the uplink data frame, respectively.

Given that the tagged SS is successful in its first attempt of sending bandwidth request, the packet delay \(X\) is therefore given by

\[
X^{(0)} = U + \Delta + V \quad \text{w.p.} \quad 1 - p,
\]

where \(\Delta\) is the duration of an IEEE 802.16 frame (including both downlink (\(T_{DL}\), uplink sub-frames (\(T_{UL}\) and their guard times), and w.p. stands for “with probability”. Note that the backoff period will not be included in the delay at this first attempt due to the way we defined the packet delay which starts when the first request is sent.

For the case when the tagged SS is not successful in its first attempt but is successful in the second attempt of sending bandwidth request, the service time can be expressed as

\[
X^{(1)} = U + Y^{(1)} + V \quad \text{w.p.} \quad p(1-p),
\]

where \(Y^{(1)}\) is a random sum of a frame duration \(\Delta\), and \(p(1-p)\) is the probability that the request is successful in the second attempt. The variable \(Y^{(1)}\) is originated from the fact that an SS will have to wait a random backoff period before sending its request which is uniformly chosen from the new backoff (contention) window. In this second attempt of sending request, an SS will choose its backoff uniformly in \([0,W_1-1]\); \(W_1 = \lambda W\) and \(Y^{(1)}\) can be calculated as

\[
Y^{(1)} = \sum_{i=0}^{1} K^{(i)} \Delta,
\]

where \(K^{(i)}\) is a discrete random variable with the following distribution:

\[
K^{(i)} = \begin{cases} 
1 & \text{w.p.} \ m/W_i, \\
2 & \text{w.p.} \ m/W_i, \\
\ldots \ & \ldots \\
A_i - 1 & \text{w.p.} \ m/W_i, \\
A_i & \text{w.p.} \ 1 - \frac{(A_i-1)m}{W_i}, 
\end{cases}
\]

where \(A_i = \lfloor W_i/m \rfloor, i = 1, \ldots, R - 1\). The \(\lfloor z \rfloor\) operator gives a minimum integer value that is greater than or equal to \(z\). For \(i = 0\), we define \(K^{(0)} = 1\) w.p. one. Note that if \(A_i = 1\), i.e. \(m \geq W_i\), then \(K^{(i)} = 1\) with probability one.

In general, the packet delay \(X\) can be expressed as

\[
X = X^{(i)} \quad \text{w.p.} \quad \eta p^i, \quad 0 \leq i \leq R - 1,
\]

where

\[
X^{(i)} = U + Y^{(i)} + V,
\]

and

\[
Y^{(i)} = \sum_{j=0}^{i} K^{(j)} \Delta.
\]

To complete the expression of \(X\), we now determine the pmf of the \(U,V\) RVs. As the tagged SS uniformly chooses
the backoff before sending request, the pmf of the RV $U$ can be approximated as below

$$U = \begin{cases} mt & \text{w.p. } 1/m, \\ (m-1)t & \text{w.p. } 1/m, \\ \vdots & \vdots \\ t & \text{w.p. } 1/m, \end{cases}$$

(9)

where $t$ is the length of one transmission opportunity (or the request slot's length). Equation (9) assumes that an SS always sends its request at the beginning of the chosen slot.

As the BS uniformly allocates data slots among successful requests, the pmf of the RV $V$ can be expressed as

$$V = \begin{cases} T & \text{w.p. } \sum_{j=0}^{k'-1} \frac{1}{j+1} q(j), \\ 2T & \text{w.p. } \sum_{j=1}^{k'-1} \frac{1}{j+1} q(j), \\ \vdots & \vdots \\ k'T & \text{w.p. } \frac{1}{k'} q(k'-1), \end{cases}$$

(10)

where $T$ is the length of one data slot (i.e., the packet transmission time), and $q(j)$ is the probability that there are $j \geq 0$ successful requests other than the tagged SS in a frame. The probability $q(j)$ follows a truncated binomial distribution

$$q(j) = Q(j+1)/(1-Q(0)), \quad 0 \leq j \leq k-1,$$

(11)

where $k = \min(m, N)$ and $k' = \min(k, d)$ and $Q(j)$ is given in (4).

C. Mean and Standard deviation

From (7), we obtain

$$E[X] = \eta \sum_{i=0}^{R-1} p^i E[X^{(i)}],$$

(12)

$$\text{Var}[X] = \eta \sum_{i=0}^{R-1} p^i (\text{Var}[X^{(i)}] + (E[X^{(i)}] - E[X])^2).$$

The mean and variance of $X^{(i)}$ are derived from (8) as

$$E[X^{(i)}] = E[U] + E[Y^{(i)}] + E[V],$$

$$\text{Var}[X^{(i)}] = \text{Var}[U] + \text{Var}[Y^{(i)}] + \text{Var}[V],$$

(13)

where

$$E[Y_i] = \Delta \sum_{j=0}^{i} E[K^{(j)}],$$

$$\text{Var}[Y_i] = \Delta^2 \sum_{j=0}^{i} \text{Var}[K^{(j)}].$$

From (6), it can be shown that

$$E[K^{(j)}] = \begin{cases} 1 & \text{if } j = 0, \\ A_j - A_j(A_j - 1) \frac{m}{2\lambda W} & \text{if } j = 1, \ldots, r-1, \\ A_j - A_j(A_j - 1) \frac{m}{2\lambda W} & \text{if } j = r, \ldots, R-1. \end{cases}$$

(14)

and

$$\text{Var}[K^{(j)}] = (E[K^{(j)}])^2 - (E[K^{(j)}])^2,$$

where $K^{(j)}$ is the second moment of $K^{(j)}$ and is given by

$$K^{(j)} = \begin{cases} 1 & \text{if } j = 0, \\ A_j^2 - A_j(A_j - 1)(1 + 4A_j) \frac{m}{6\lambda W} & \text{if } j = 1, \ldots, r-1, \\ A_j^2 - A_j(A_j - 1)(1 + 4A_j) \frac{m}{6\lambda W} & \text{if } j = r, \ldots, R-1. \end{cases}$$

It remains to determine $E[U]$, $\text{Var}[U]$, $E[V]$ and $\text{Var}[V]$ from (9) and (10), which can be expressed as

$$E[U] = (m+1)t/2,$$

$$\text{Var}[U] = \bar{U}^2 - (E[U])^2,$$

where $\bar{U}^2 = (m+1)(2m+1)t^2/6,$ and

$$E[V] = T \sum_{j=0}^{k-1} q(j) \frac{j+1}{j+1},$$

$$\text{Var}[V] = \bar{V}^2 - (E[V])^2,$$

where

$$\bar{V}^2 = T^2 \sum_{j=0}^{k-1} q(j) \frac{(i+1)^2}{i+1},$$

and $q(j)$ is given in (11).

The delay and variance of packet delay are then calculated by substituting (13), (14), (15) into (12).

IV. MODEL VALIDATION AND DISCUSSION

In this section we verify our analytical model using simulation and study the characteristics of the access delay as a function of $N, m, d$ parameters. To this end, we have developed a simulator [6] to simulate the broadcast polling with a BEB contention resolution mechanism in IEEE 802.6 as described in Section II. The simulator is event-driven and developed using C++. The duration of each simulation run was 3,000 seconds, with a warm-up period of 300 seconds. All the simulation results are plotted with 95% confidence intervals resulting from six runs for each point in the graphs. The MAC and physical layer parameters were configured in accordance with default parameters taken from the standard [1]. In particular, the frame duration is 1 msec consisting of 5000 physical slots or 2500 mini slots each of 0.4 msec length. The data rate is 120 Mbps employing 64-QAM modulator scheme at 25 MHz. Each bandwidth request consists of 6 mini slots for preamble and transition gap (SSTG), 2 mini slots for preamble and one mini slot for a bandwidth request message of 48 bits. The length of a data slot including the preamble and transition gap is 37.6 msec (i.e. 94 mini slots) which allows the transmission of approximately 0.5
number of SSs

\[ \text{REQ Unsuccessful probability} \]

analysis \( d=2 \)
simulation \( d=6 \)

Figure 2. Unsuccessful request probabilities \( (W = 8, r = 3, R = 5, m = 6) \).

KB packet per data slot. The initial contention window \( W \) is set to 8, and the maximum window is \( CW_{\text{max}} = 2^3 \times W = 64 \) with \( R = 5 \) maximum attempts allowed for each packet.

Figure 2 shows the simulation and analytical results for the probability of unsuccessful request for different data slots available in the up-link frame using the same number of transmission opportunities for bandwidth requests \( (m = 6) \). Observe that the analytical results obtained from the fixed point analysis based on (1), (3) and (5) match with the values obtained from the simulation. The agreement has been achieved within two-percentage error for both cases where the probability of an unsuccessful bandwidth request is due to collision with other requests only \( (d = 6) \), or due to the lack of data slots in a subsequent up-link sub-frame as well \( (d = 2 < m) \). Having the probability of unsuccessful request, the mean and standard deviation of the packet delay are computed as described in Section III-C. These analytical values together with the simulation results are plotted against the number of SSs in Figs. 3 and 4. In general the model gives good estimations of the mean delay and its standard deviation, and as shown in the figures, reducing the number of data slots available in the uplink frame causes an increasing in both mean and standard deviation of the packet delay. We have observed similar agreements using a range of different \( \{m,d\} \) pair values which are omitted here due to space limitation.

V. CONCLUSION

In this paper we have proposed a simple fixed point analysis for the probability of an unsuccessful bandwidth request using binary exponential backoff, and developed an analytical model for performance evaluation of a saturated IEEE 802.16 network in terms of the mean and standard deviation of the packet delay. Explicit forms of the above performance metrics have been derived from the analytical model and validated against simulation. The model can be used to gain insights into the effectiveness of the binary backoff mechanism, and to choose suitable parameters for 802.16 networks to minimize request unsuccessful probability and packet delay.

REFERENCES