

Performance Analysis of Best-effort Service in Saturated IEEE 802.16 Networks

Hai L. Vu, *Senior Member, IEEE*, Sammy Chan, *Member, IEEE*,
and Lachlan L. H. Andrew, *Senior Member, IEEE*

Abstract—The IEEE 802.16 standard is a promising technology for the fourth-generation mobile networks and is being actively promoted by the WiMAX Forum, an industry-led consortium. Among the various service classes supported in the standard, best-effort service is expected to be the major service class subscribed to by users due to its operational simplicity and lower charging rate compared with other service classes. In this paper, we investigate the throughput and packet access delay performance of the best-effort service using the contention-based bandwidth request mechanism in a saturated 802.16 network. In particular, we develop a simple fixed point analysis to approximate the failure probability of a bandwidth request and derive analytical expressions for network throughput and packet access delay. Furthermore, the uniqueness of the fixed point is established. The accuracy of the analytical model is validated by comparing with simulation over a wide range of operating conditions. The implications of various different parameter configurations on the system performance are investigated using the analytical model. The utility of the model is further demonstrated by finding the maximum achievable throughput and obtaining its corresponding optimal initial contention window.

I. INTRODUCTION

WiMAX (Worldwide Interoperability for Microwave Access) was first developed for wireless broadband access and has evolved to become 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.

To support multimedia services, the IEEE 802.16 defines five service classes referred to as *scheduling services*, viz.: unsolicited grant service (UGS), extended real time polling

service (ertPS), real-time polling service (rtPS), non-real-time polling service (nrtPS), and best effort service (BE) [3]. The main characteristics of these scheduling services are:

- **UGS** is designed for constant bit rate (CBR) real-time traffic such as E1/T1 circuit emulation and VoIP without silence suppression. The main QoS parameters are maximum sustained rate (MSR), maximum latency and tolerated jitter.
- **ertPS** is designed to support VoIP with silence suppression. It has the same QoS parameters as UGS. However, it is allocated the maximum sustained rate only during active periods, and is not allocated any bandwidth during silent periods.
- **rtPS** is designed to support real-time applications with variable bit rates, such as MPEG videos. Other than the MSR, its QoS parameters also include the minimum reserved rate.
- **nrtPS** is designed for applications without any specific delay requirement but with the need for a minimum amount of bandwidth, such as File Transfer Protocol (FTP).
- **BE** is designed for applications that are delay-tolerant and do not require a minimum bandwidth. Bandwidth will be granted to this service class if and only if there is a left-over bandwidth from other classes.

The basic operating mode of WiMAX is the point-to-multipoint (PMP) mode, where a base station (BS) serves a set of subscriber stations (SSs) within the same antenna sector in a time-division-multiplexed fashion. In the downlink direction, the BS is the only sender. It has full control of which time slot to use to send data to which SS, and data are broadcast to all SSs. In the uplink direction, in which SSs send data to the BS, SSs share the common channel. In order to avoid collision, only one SS should be permitted to transmit at any time slot. According to the 802.16 standard, such an exclusive access is achieved by requiring each SS to have 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 [4]. 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 [5] for contention resolution. It is typically used by BE connections.

BE service is relatively simple to provide because it does not involve QoS negotiation and enforcement of traffic parameters.

Copyright ©2009 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU 111208].

H. Vu and L. Andrew are with the Centre for Advanced Internet Architectures, Faculty of ICT, Swinburne University of Technology, P.O. Box 218, VIC 3122, Australia (email: {h.vu, l.andrew}@ieee.org).

S. Chan is with the Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong (email: eeschan@cityu.edu.hk).

Moreover, since BE service offers no QoS guarantees, its service charge is expected to be cheaper than other scheduling services. As a result, it would be a major scheduling service subscribed to by WiMAX users. Therefore, it is important to understand the network performance when delivering BE service.

However, recent research on MAC layer performance in WiMAX networks just focuses on the performance of bandwidth request mechanisms. In [6], the delay performance of the contention-free unicast polling request mechanism is 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 is investigated. An analytical model based on a Markov chain is used to evaluate the average delay of a transmission request for saturated networks. An alternative modelling approach is also proposed in [8]. The model in [7] is later extended to investigate the case with Bernoulli request arrival [9]. In [10], the capacity of the contention slots in delivering bandwidth requests and the average access delay are derived, taking into account the response time for a bandwidth assignment and the possible timeouts for lost messages. However, the work in [7], [8], [9] and [10] does 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 frames, instead of measuring the time between the first attempt of a request and the completion of the packet transmission. In [11], the performance of bandwidth request mechanisms based on piggyback and broadcast polling are compared by simulation.

The first major contribution in this paper is an accurate analytical model for evaluating the network throughput and packet access delay of BE service. To this end, we study the performance of BE service based on broadcast polling in a *saturated* WiMAX network. Here, a saturated network means that each SS always has a packet to send. Considering such a condition allows us to obtain the upper bound of network performance. Our analytical model is based on a set of fixed point equations which calculates the failure probability of a bandwidth request in broadcast polling as a function of system parameters. Furthermore, we show that these fixed point equations have a unique solution. Based on the obtained failure probability of a bandwidth request, the network throughput and packet access delay are then derived.

Our second major contribution is to show the utility of the proposed analytical model in designing a better network. In particular, we use the model to investigate the optimal achievable throughput under different parameter settings at the MAC layer. This leads us to develop a mechanism to adjust the initial contention window according to system parameters to achieve this optimality. This work provides operators analytical tools to evaluate network performances and to gain insights into the optimal configuration of system parameters.

The rest of the paper is organized as follows. Section II describes the details of the broadcast polling mechanism. Section III presents our analytical model. Section IV validates the model by comprehensive comparisons with simulation results, and evaluates the performance of BE service under

different configurations of system parameters. Section V investigates how the optimal throughput depends on various system parameters and presents a mechanism for achieving optimal throughput adaptively. Finally, conclusions are drawn in Section VI.

II. BROADCAST POLLING

The MAC frame structure defined by the IEEE 802.16 standard for time division duplexing (TDD) in PMP mode is shown in Fig. 1. In this mode, the channel is time slotted into fixed-length frames; each consists of a downlink and uplink subframe. The durations of the downlink and uplink subframes are dynamically controlled by the BS by 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 for bandwidth requests and data packets. Consider a frame i , when an SS has a data packet to send, it first sends a bandwidth request for transmitting its data in one of the transmission opportunities within the request interval of the uplink subframe. Upon receiving the bandwidth requests, the BS then allocates bandwidth or data slots for data transmission in the uplink data interval of frame $i+1$ based on its scheduler.

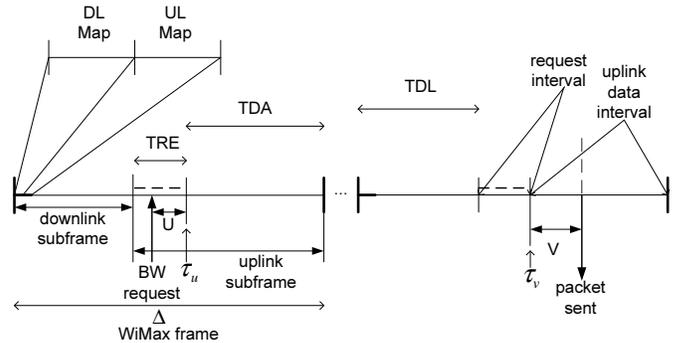


Fig. 1. IEEE 802.16 MAC frame structure with times division duplexing (TDD).

In the following, we consider a scenario where broadcast polling is used by the BS with m (fixed) transmission opportunities for bandwidth requests which are referred to as *request slots*. In this case, if there is only one request submitted to a request slot, the request is successful. On the other hand, if there are two or more SSs submitting their requests in the same request slot, a collision occurs and truncated BEB is used to resolve the contention. In particular, when sending a request at backoff state i , $i \geq 0$, an SS must carry out the backoff process by randomly selecting a backoff time in the range $[0, W_i - 1]$, where W_i is the contention window for backoff state i . Here the backoff time represents the number of request slots that must pass before the request can be submitted. At backoff state 0 (which is the first attempt), W_0 is set equal to W , the initial contention window. If the request is unsuccessful, then the contention window size is multiplied by $\alpha = 2$, and another backoff period is initiated. The process of doubling the contention window continues until the maximum possible

value, $CWmax = \alpha^r W$, $r \geq 1$, is reached. Here, r is referred to as the truncation value. If the request is unsuccessful after r reattempts, the contention window is not expanded anymore, but maintained at $CWmax$ for the remaining attempts until either the request is successful, or the maximum allowable number of attempts, $R \geq r+1$, have been made. If the request is still unsuccessful after R attempts, the packet is discarded. In summary, W_i is given by:

$$W_i = \begin{cases} \alpha^i W, & 0 \leq i \leq r, \\ \alpha^r W, & r < i < R. \end{cases} \quad (1)$$

In this paper the SSs are only allowed to request bandwidth to transmit one packet per request, and all packets are assumed to have the same length. Let t be the length of a request (or backoff) slot. Furthermore, we assume that the BS always allocates the same amount of uplink capacity consisting of $d \leq m$ data slots in every uplink subframe for uplink traffic. Each data slot is of length $T \gg t$ which is the transmission time of a packet. As the standard does not define scheduling algorithms for both BS and SSs, we assume here that the BS uplink scheduler will uniformly allocate bandwidth to SSs whose bandwidth requests were successful in the previous frame. Let j be the number of requests which do not collide. If $j < d$ then in the next frame there will be $(d-j) > 0$ data slots which remain unused and are wasted. However, if $j > d$ then $j-d > 0$ requests must be declined because there are only d slots available in the next frame; these $j-d$ requests are also considered unsuccessful.

III. ANALYTICAL MODEL

Consider an IEEE 802.16 network consisting of N saturated SSs operating in the PMP mode. Our objective is first to develop a fixed point approximation to compute the probability that a request is unsuccessful. The expressions for network performance metrics such as the network throughput and packet delay will then follow.

A. Unsuccessful Request Probability

Let p be the probability that a request sent by an SS is unsuccessful. As in [8], a request is regarded as unsuccessful either when the request experiences collision during transmission (with probability p_c) or when the request is successfully transmitted but the BS could not allocate bandwidth to it due to insufficient data slots (with probability p_d). For simplicity, we model these two events as independent. Then p can be expressed as

$$p = 1 - (1 - p_c)(1 - p_d), \quad (2)$$

Since both p_c and p_d can in turn be expressed as a function of the probability p , fixed point equations can be established to calculate the individual probabilities. The fixed point analysis is detailed as follows.

A request is successful on the first attempt with probability $1 - p$. Recall that the contention window is initially set to W , the average number of elapsed backoff slots before such a request is $\frac{m}{2} + (W - 1)/2$. The first term is due to the fact that an SS cannot 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; because requests slots 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 a request based on the contention resolution mechanism described in Section II.

If its first attempt fails, a request is successful on the second attempt with probability $p(1-p)$. The average number of elapsed backoff slots in this case is $\frac{m}{2} + (\alpha W - 1)/2$. Continuing this argument until the R^{th} attempt yields the average number of elapsed backoff slots (B_{avg}) before the request is successful. We have

$$\begin{aligned} B_{avg} &= \frac{m}{2} + \eta \sum_{i=0}^{r-1} p^i \left(\frac{\alpha^i W - 1}{2} \right) + \eta \left(\frac{\alpha^r W - 1}{2} \right) \sum_{i=r}^{R-1} p^i \\ &= \frac{m}{2} + \frac{\eta W (1 - (\alpha p)^r)}{2(1 - \alpha p)} - \frac{1 - p^r}{2(1 - p^R)} + \\ &\quad \frac{(\alpha^r W - 1)(p^r - p^R)}{2(1 - p^R)}, \end{aligned} \quad (3)$$

where $\eta = (1-p)(1-p^R)^{-1}$, and $(1-p^R)$ is a normalisation factor.

Note that B_{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

$$\tau = 1/(B_{avg} + 1). \quad (4)$$

Given that there are N saturated SSs, the probability p_c that a request sent by an SS collides with other requests can be expressed as

$$p_c = 1 - (1 - \tau)^{N-1}. \quad (5)$$

Let ξ be the probability that a collision-free request is made in a given slot, given that there are N SSs each attempting to send requests with probability τ . Under the approximation that requests are independent, we have

$$\xi = N\tau(1 - \tau)^{N-1}. \quad (6)$$

The probability that there are j , $0 \leq j \leq k = \min(m, N)$, collision-free requests among m request slots is then given by a truncated binomial distribution

$$Q(j) = \frac{\binom{m}{j} \xi^j (1 - \xi)^{m-j}}{\sum_{i=0}^k \binom{m}{i} \xi^i (1 - \xi)^{m-i}}. \quad (7)$$

The probability that a collision-free request is unsuccessful due to lack of bandwidth in the subsequent frame can then be expressed as

$$p_d = \frac{\sum_{j=d+1}^k (j-d)Q(j)}{\sum_{j=0}^k jQ(j)}. \quad (8)$$

Equations (2) to (8) create a fixed point formulation from which p can be computed numerically. The existence and uniqueness of a solution for the above formulation are proved in the next theorem, proved in Appendix A.

Theorem 1: Let $g(\tau(p(p_c, p_d))) : [0, 1] \rightarrow [0, 1]$ be defined by the right-hand side of (8), and let $f(\tau, p_c, p_d) = g(\tau(p(p_c, p_d))) - p_d$.

For $m \leq N$,

- 1) for a given p_d , $0 \leq p_d \leq 1$, the function $\tau(p(p_c, p_d)) : [0, 1] \rightarrow [0, 1]$ defined in (4) and p_c defined in (5), have a unique fixed point solution,
- 2) the equations $f(\tau, p_c, p_d) = 0$, (4) and (5) have a unique fixed point solution.

B. Throughput Analysis

From (7), the probability that there are j collision-free requests in each uplink subframe is given by $Q(j)$. If $j \leq d$, each request is allocated a data slot and thus j packets will be transmitted in the subsequent uplink subframe. On the other hand, if $j > d$, only d requests are allocated data slots. As a result, only d packets will be transmitted in the subsequent uplink subframe. Therefore, the normalized throughput Γ is given by

$$\Gamma = \frac{\sum_{j=1}^d jQ(j) + \sum_{j=d+1}^k dQ(j)}{d}. \quad (9)$$

C. Delay Analysis

The delay of a packet is defined as the time duration from its first bandwidth request until the packet transmission is completed. 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 subframe. Under the 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.

Referring to Fig. 1, for any data packet of a tagged SS, the time epoch at the end of the request interval in which its first request is sent is denoted as τ_u , and the time epoch at the beginning of the uplink data interval in which the data packet is sent is denoted as τ_v . Let U and V be the random variables (RVs) representing the time durations from the time of sending the first request until τ_u , and from τ_v until the packet has been transmitted, respectively. The duration of the uplink subframe is given by $T_{UL} = T_{RE} + T_{DA}$, where $T_{RE} = mt$ and $T_{DA} = dT$ are the length of the request interval and the uplink data interval, respectively.

Given that the tagged SS is successful in its first attempt of sending a bandwidth request, the packet delay $X^{(0)}$ is therefore given by

$$X^{(0)} = U + \Delta + V, \quad (10)$$

where Δ is the duration of an IEEE 802.16 frame (including both downlink (T_{DL}), uplink subframes (T_{UL}) and their guard times). Note that the backoff period of this first attempt is not

included as part of the delay due to the way we have defined the packet delay period, which starts from the time epoch 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 packet delay can be expressed as

$$X^{(1)} = U + Y^{(1)} + V,$$

where $Y^{(1)}$ is the random time that a packet has to wait between the epochs τ_u and τ_v due to the first and second request attempt, and is given by

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

where $K^{(i)}$ is the number of frames delayed in the attempt i . Using this notation, we can rewrite (10) as

$$X^{(0)} = U + Y^{(0)} + V,$$

where, by definition, $Y^{(0)} = K^{(0)} \Delta$ and $K^{(0)}$ is always equal to one. In the second request attempt, the SS will choose its backoff uniformly in $[0, W_1 - 1]$; $W_1 = \alpha W$ and thus $K^{(1)}$ is a discrete random variable with the following probability mass function (pmf)

$$K^{(1)} = \begin{cases} 1 & \text{w.p. } m/W_1, \\ 2 & \text{w.p. } m/W_1, \\ \vdots & \vdots \\ A_1 - 1 & \text{w.p. } m/W_1, \\ A_1 & \text{w.p. } 1 - \frac{(A_1 - 1)m}{W_1}, \end{cases} \quad (11)$$

where $A_1 = \lceil W_1/m \rceil$, and $\lceil x \rceil$ gives a minimum integer value that is greater than or equal to x , and w.p. stands for ‘‘with probability’’.

In general, for a packet transmission requiring i request attempts, the packet delay $X^{(i)}$ is given by

$$X^{(i)} = U + Y^{(i)} + V, \quad 0 \leq i < R, \quad (12)$$

with

$$Y^{(i)} = \sum_{j=0}^i K^{(j)} \Delta,$$

and $K^{(i)}$, $0 < i < R$, has the following probability mass function

$$K^{(i)} = \begin{cases} j & \text{w.p. } m/W_i, \quad j = 1, 2, \dots, A_i - 1, \\ A_i & \text{w.p. } 1 - \frac{(A_i - 1)m}{W_i}, \end{cases} \quad (13)$$

where $A_i = \lceil W_i/m \rceil$. Note that if $A_i = 1$, i.e. $m \geq W_i$, then $K^{(i)} = 1$ with probability one.

Overall, the packet delay X can be expressed as

$$X = X^{(i)} \quad \text{w.p. } \eta p^i, \quad 0 \leq i < R. \quad (14)$$

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 its request, the pmf of U can be approximated as

$$U = it \quad \text{w.p. } 1/m, \quad (15)$$

where $i = 1, 2, \dots, m$.

Equation (15) 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 V can be expressed as

$$V = iT \quad \text{w.p.} \quad \sum_{j=i}^{k'} \frac{q(j-1)}{j}, \quad (16)$$

where $i = 1, 2, \dots, k'$, $k' = \min\{k, d\}$, 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

$$q(j) = Q(j+1)/(1-Q(0)), \quad 0 \leq j \leq k-1, \quad (17)$$

where $Q(j)$ is given in (7).

From (14), we obtain

$$E[X] = \eta \sum_{i=0}^{R-1} p^i E[X^{(i)}], \quad (18)$$

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

The mean and variance of $X^{(i)}$ are derived from (12)

$$\begin{aligned} E[X^{(i)}] &= E[U] + E[Y^{(i)}] + E[V], \\ \text{Var}[X^{(i)}] &= \text{Var}[U] + \text{Var}[Y^{(i)}] + \text{Var}[V], \end{aligned} \quad (19)$$

where

$$\begin{aligned} E[Y^{(i)}] &= \Delta \sum_{j=0}^i E[K^{(j)}], \\ \text{Var}[Y^{(i)}] &= \Delta^2 \sum_{j=0}^i \text{Var}[K^{(j)}]. \end{aligned}$$

From (13), it can be shown that

$$E[K^{(j)}] = \begin{cases} 1 & j = 0, \\ A_j - A_j(A_j - 1) \frac{m}{2\alpha^j W} & j = 1, \dots, r-1, \\ A_j - A_j(A_j - 1) \frac{m}{2\alpha^r W} & j = r, \dots, R-1, \end{cases} \quad (20)$$

and

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

where $\overline{K^{(j)^2}}$ is the second moment of $K^{(j)}$ and is given by

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

It remains to determine $E[U]$, $\text{Var}[U]$, $E[V]$ and $\text{Var}[V]$

from (15) and (16), which can be expressed as

$$\begin{aligned} E[U] &= (m+1)t/2, \\ \text{Var}[U] &= \overline{U^2} - (E[U])^2, \quad \text{where} \\ \overline{U^2} &= (m+1)(2m+1)t^2/6, \quad \text{and} \end{aligned} \quad (21)$$

$$\begin{aligned} E[V] &= T \sum_{j=0}^{k'-1} q(j) \sum_{i=0}^j \frac{i+1}{j+1} = T \sum_{j=0}^{k'-1} \frac{j+2}{2} q(j), \\ \text{Var}[V] &= \overline{V^2} - (E[V])^2, \quad \text{where} \\ \overline{V^2} &= T^2 \sum_{j=0}^{k'-1} q(j) \sum_{i=0}^j \frac{(i+1)^2}{j+1}, \\ &= T^2 \sum_{j=0}^{k'-1} \frac{(j+2)(2j+3)}{6} q(j), \end{aligned}$$

and $q(j)$ is given in (17).

The delay and variance of packet delay are then calculated by substituting (19), (20), (21) into (18).

IV. MODEL VALIDATION AND PERFORMANCE ANALYSIS

In this section we verify our analytical model using simulation and study in detail the network performance including throughput, mean and standard deviation of delay, as functions of N, m, d, r, R and W . To this end, we have developed a simulator [12] to simulate the broadcast polling with the BEB contention resolution mechanism as described in Section II. The simulator is event-driven and developed using C++. The duration of each simulation run is 3,000 seconds, with a warm-up period of 300 seconds. 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 ms consisting of 5000 physical slots or 2500 mini slots each of $0.4 \mu\text{s}$ length. The data rate is 120 Mb/s employing 64-QAM modulation at 25 MHz. Each bandwidth request consists of 6 mini slots including 3 mini slots for subscriber station 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 $37.6 \mu\text{s}$ (i.e. 94 mini slots) which allows the transmission of an approximately 0.5 kB packet per data slot. Using this simulator, we have validated our analytical model with different sets of typical system parameters. As shown by the results presented below, the model accurately predicts the performance of the contention-based best-effort service of 802.16 and is therefore suitable for studying the impact of different parameters on overall system performance.

We first validate our model for the probability of unsuccessful REQ (p) using the above simulation. We set $r = 3$, $R = 5$ and $m = d = 10$, and obtain p for different numbers of SSs with $W = 8, 16, 32, 64$, respectively. The results are shown in Fig. 2. As expected, larger N leads to more request contention and thus larger p . While other system parameters are fixed, p decreases as W increases. This is because when W increases, there are more choices of a request slot in each backoff stage. As a result, the probability that an SS transmits

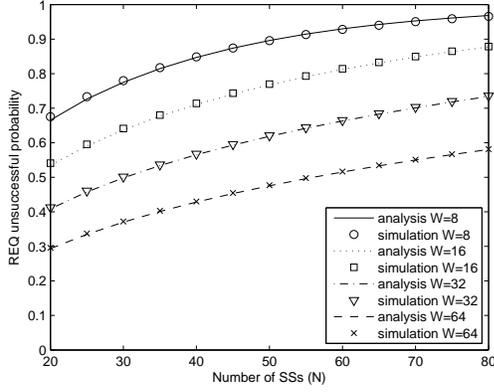


Fig. 2. Unsuccessful request probabilities ($r = 3, R = 5, m = 10, d = 10$).

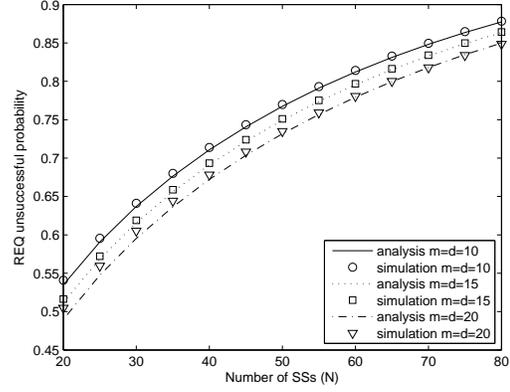


Fig. 4. Unsuccessful request probabilities ($r = 3, R = 5, W = 16, d = m$).

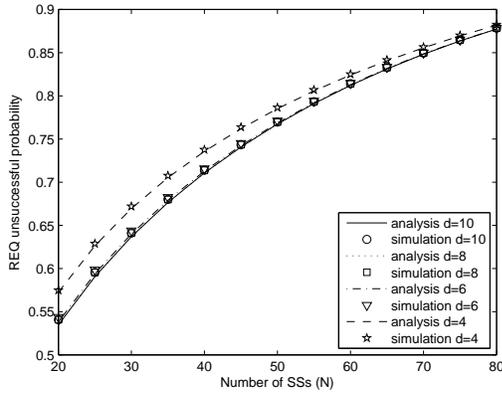


Fig. 3. Unsuccessful request probabilities ($r = 3, R = 5, W = 16, m = 10$).

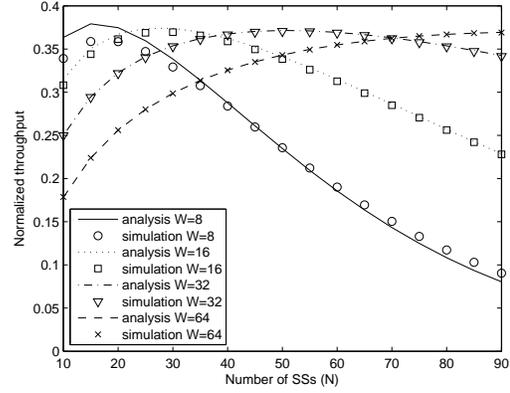


Fig. 5. Normalized throughput ($r = 3, R = 5, m = 10, d = 10$).

a request in a request slot (τ) becomes smaller. Hence, p_c and p decrease. We now set $W = 16, r = 3, R = 5, m = 10$, and obtain p for different numbers of SSs with $d = 4, 6, 8$ and 10 . As can be seen from Fig. 3, the curves for $d = 6, 8$ and 10 essentially overlap with other, while the curve for $d = 4$ exhibits higher p . This is because, given m available request slots, the average number of successful requests is less than m . In this example case, $d = 6$ is quite sufficient to serve the successful requests, therefore increasing d does not change p significantly. However, when d becomes too small, say $d = 4$, there are not enough data slots available in the subsequent subframe which then results in a larger p . Next, we set $W = 16, r = 3, R = 5, m = d$, and obtain p for different numbers of SSs with $m = 10, 15, 20$, respectively. The results are shown in Fig. 4. It can be seen that p increases when m decreases as smaller m causes higher contention probability in each request slot and hence higher p .

We now investigate the throughput performance under different system parameters. Firstly, we set $r = 3, R = 5$ and $m = d = 10$, and plot throughput against different numbers of SSs with $W = 8, 16, 32, 64$, respectively, in Fig. 5. It can be seen that, for a given W , throughput varies with N such that there exists an optimal N which maximizes the throughput. The reason is that when N is small, the demand

for data slots is low and hence the throughput is low. As N gradually increases, the throughput also increases. However, increasing N further will also cause more contention which at some point results in fewer successful requests and lower throughput. Thus there exists an N value where the throughput is maximized. Alternatively, for a fixed N , say at $N = 40$, throughput increases and then decreases when W varies from 8 to 64 , which indicates that there also exists an optimal W at which throughput is maximized. Also, it is interesting to see from the plot that the maximum throughput value is more or less the same, irrespective of the N and W values.

Next, we set $W = 16, r = 3, R = 5, m = 10$, and obtain the throughput for different numbers of SSs with $d = 6, 8, 10$, respectively. The results are shown in Fig. 6. As d decreases, the throughput increases. Recall that the throughput is also a measure of the efficiency of the data slots in each uplink subframe. Since the average number of collision-free requests in a request interval is smaller than m (from (7)), a smaller d gives a higher efficiency. In fact, noticing from Fig. 5 that the throughput does not exceed 37.5% , we then investigate the throughput when d is set to be 40% of m . The result is shown in Fig. 7, with $W = 16, N = 40$, and m varying between 5 and 30 . It can be seen that a throughput of at least 80% can be achieved with a relatively small p_d . However, it should be noted that if d becomes even smaller, p_d and thus packet delay

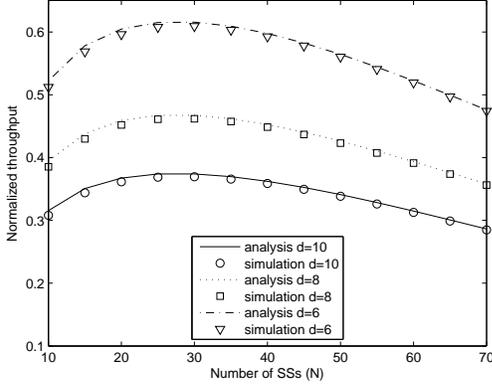


Fig. 6. Normalized throughput ($r = 3, R = 5, W = 16, m = 10$).

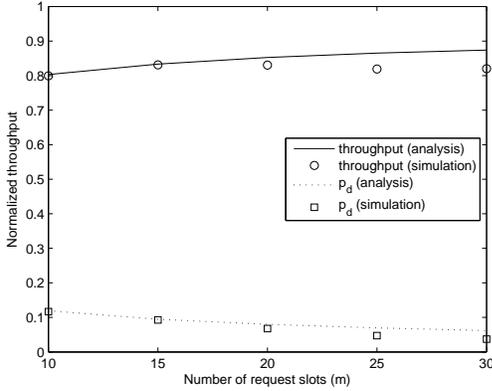


Fig. 7. Normalized throughput ($r = 3, R = 5, W = 16, N = 40, d = 0.4m$).

will increase despite the high throughput value.

Finally, we investigate the delay performance under different system parameters. Firstly, we set $r = 3, R = 5$ and $m = d = 10$, and plot mean packet delay against different numbers of SSs with $W = 8, 16, 32, 64$, respectively, in Fig. 8. Again, as expected, for a given W , larger N incurs larger delay, because p is larger and each request would need more attempts before being successful. For a fixed N , when W increases, the average window size in each backoff stage increases and hence the mean packet delay increases. Also, as W increases, a larger range of request slots across several subframes is available to be chosen resulting in larger standard deviation of packet delay as shown in Fig. 9. Although the increase in W leads to the increase in mean and standard deviation of packet delay, it also means that more requests are successful and thus the probability of packet loss due to exceeding the maximum allowable number of attempts decreases as shown in Fig. 10. Here the probability of packet loss is expressed as p^R where p is given in (2) and R is the maximum allowable number of attempts per packet. Note from Fig. 10 that by doubling W from 8 to 16, or from 16 to 32, the probability of packet loss is reduced significantly.

Next, setting $r = 3, R = 5, W = 16$ and $m = d$, we plot the mean and standard deviation of packet delay against different

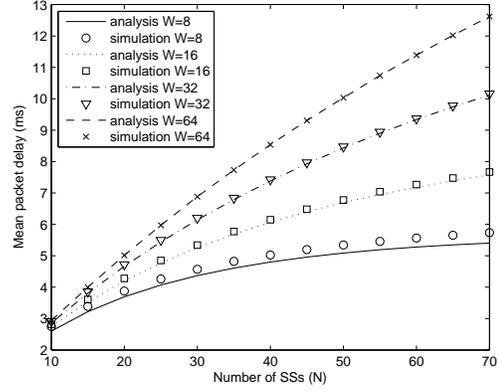


Fig. 8. Mean packet delay ($r = 3, R = 5, m = 10, d = 10$).

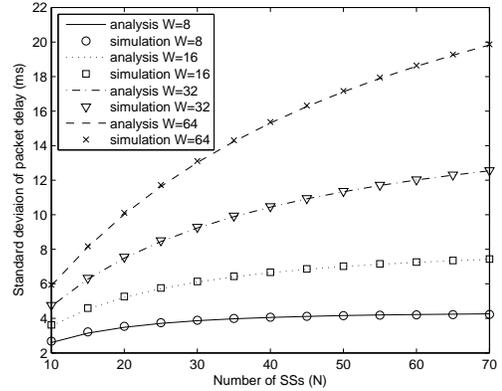


Fig. 9. Standard deviation of packet delay ($r = 3, R = 5, m = 10, d = 10$).

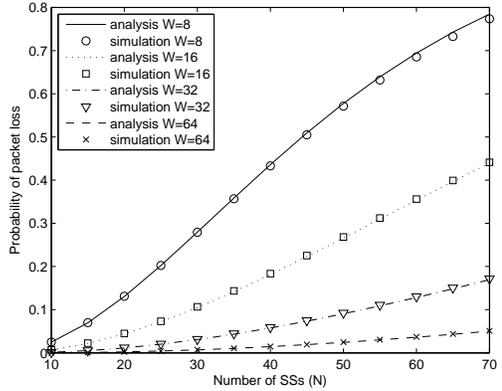


Fig. 10. Probability of packet loss ($r = 3, R = 5, m = 10, d = 10$).

numbers of SSs with $m = 10, 15, 20$, respectively, in Figs. 11 and 12. For a fixed N , when m increases, it is more likely to have a successful request, hence both mean and standard deviation of packet delay decrease. When $r = 3, R = 5, W = 16, m = 10$ and $d = 6, 8, 10$, we can see from Fig. 13 that the mean packet delay does not vary much. This is consistent with the result in Fig. 3 that p is more or less the same over this range of d . But for $d = 4$, p is larger, and so as the delay.

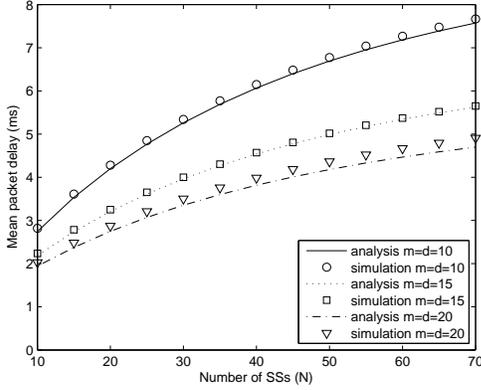


Fig. 11. Mean packet delay ($r = 3, R = 5, W = 16$).

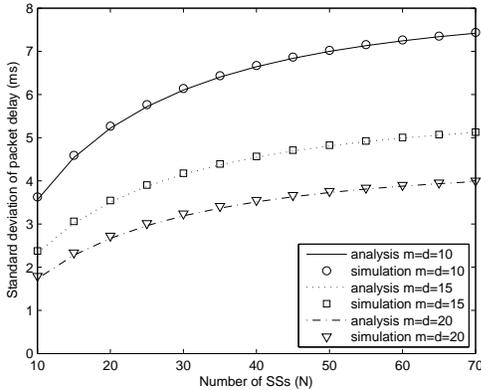


Fig. 12. Standard deviation of packet delay ($r = 3, R = 5, W = 16$).

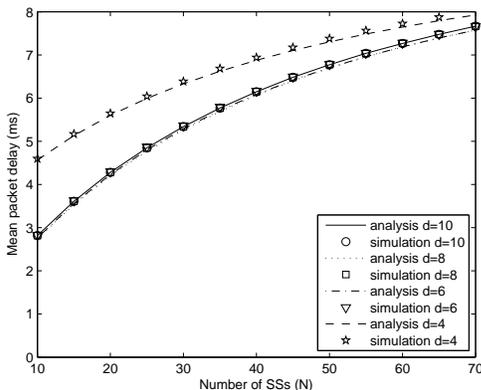


Fig. 13. Mean packet delay ($r = 3, R = 5, W = 16$).

V. THROUGHPUT OPTIMIZATION

From the results shown in Fig. 5, we have observed that for fixed N, m, d, r and R , when W is varied, there exists an optimal value such that throughput is maximized. This observation has motivated us to find the optimal W and the corresponding optimal throughput for a given set of system parameters.

In Appendix B, the optimal W is derived and given by (35). In the special case, where $r = R - 1$ and N is large, the

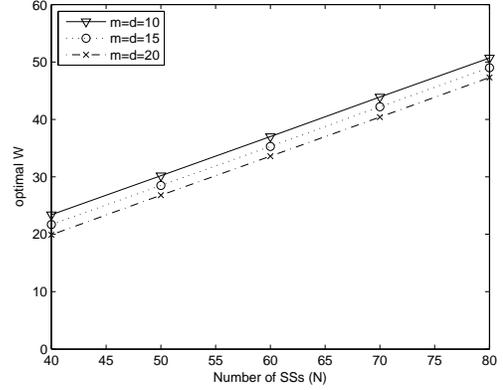


Fig. 14. The optimal W against N , with different m ($r = 3, R = 5$).

optimal W becomes

$$W_{opt} = \frac{(N - m/2)(1 - p^R)(1 - \alpha p)}{(1 - p)(1 - (\alpha p)^R)}. \quad (22)$$

Note that for large R , i.e. $R \rightarrow \infty$, the asymptotic value of W_{opt} in (22) is

$$W_{opt}^{R \rightarrow \infty} = (N - m/2) \frac{(1 - \alpha p)}{(1 - p)}.$$

In the following, we will investigate how the system parameters affect the optimal W and throughput values.

Figure 14 shows the size of the optimal W versus the number of SSs in the system. It can be seen that for a given number of request slots (m), a larger W is required for the throughput to be maximized with an increasing number of subscribers. However, the opposite trend is observed for a fixed number of subscribers but with increasing m . This is because for the same number of subscribers but with larger number of request slots, the attempt probability per request slot should increase to fully utilize all the available request slots. Thus the maximum throughput will occur for smaller W in this case.

The impact of the truncation value (r) defining CW_{max} and the maximum allowable attempts (R) on the optimal W is shown in Fig. 15. Observe that for the same number of subscribers and request slots, smaller (r, R) pair requires larger W to maximize the throughput. This is because the optimal attempt probability that maximizes the throughput would be the same for any (r, R) values provided that the numbers of subscribers and request slots are fixed. Since a smaller (r, R) would yield a smaller average backoff window, which would then result in a greater attempt probability, a larger W is required in order to maximize the throughput as compared to that of larger (r, R).

The impact of various system parameters on the optimal W and the resulting network performances is illustrated in Table I. It is clear that the optimal throughput only depends on the ratio d/m , but is independent of N . However, the corresponding optimal W increases with N . Therefore, for a given d/m , when N is known, W can be adjusted to the optimal value to achieve the optimal throughput. Since the optimal W increases with N , the resulting delay also

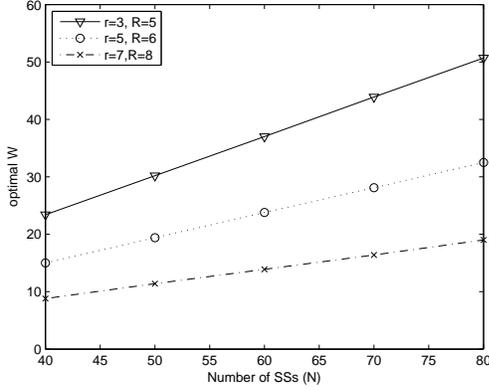


Fig. 15. The optimal W against N , with different r, R ($m = d = 10$).

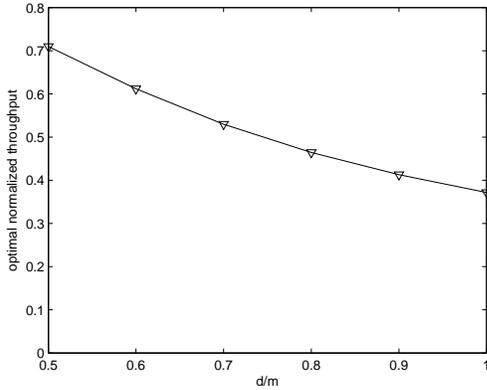


Fig. 16. The optimal throughput against d/m ($N=50, m=10, r=3, R=5$).

increases. On the other hand, as also shown in the table, when N and m are fixed, d can be set smaller than m to achieve higher optimal throughput without greatly affecting the packet delay and loss probability. Finally, the effect of d/m on the optimal throughput is depicted in Fig. 16.

For a given number of SSs (N), Algorithm 1 will optimize the system performance. In particular, the algorithm adjusts the number of available data slots in the uplink sub-frame in a small step size during its iterations and is terminated after the packet loss probability (p^R) exceeds a certain threshold (ϵ). The loss threshold ϵ is a design parameter and can be set according to the actual applications' requirements. Finally the algorithm outputs the optimal initial contention window and the corresponding maximum throughput. A similar algorithm can be developed where the delay is considered as a design parameter. This is because, while achieving the optimal throughput, packets may experience different delays using different (r, R) parameter pairs as described earlier. Note that in both cases (i.e. loss vs. delay), the objective is to adjust the system parameters so that the system performance (e.g. throughput) is optimized based on results of the derivation in this section.

Algorithm 1 Find W_{opt} and maximum throughput

Require: $N, m, (r, R)$ and ϵ

- 1: $d = m$ // Initialization
 - 2: Set $\tau = 1/N$
 - 3: Set $p = 1 - (1 - p_c)(1 - p_d)$, where p_c, p_d are in (4) and (7)
 - 4: **while** $p^R \leq \epsilon$ **do**
 - 5: Set $d = d - 1$
 - 6: Update p according to (2)
 - 7: **end while**
 - 8: Calculate W_{opt} according to (21)
 - 9: Calculate Γ according to (8) // maximum throughput
-

VI. CONCLUSIONS

In this paper, we have developed an analytical model to evaluate the performance of best-effort service in a saturated IEEE 802.16 network. Explicit expressions for network throughput, mean and standard deviation of the packet delay are derived. Comparison with simulation confirms the accuracy of our analytical model over a wide range of operating conditions. Using the proposed model, we have studied the impact of various parameters on the overall network performance such as throughput and packet access delay. We showed that the initial contention window of the broadcast polling mechanism can be suitably chosen to maximize the network throughput. A simple asymptotic value for this optimal window is then provided assuming the number of allowable retransmissions is large. Furthermore, the maximum achievable throughput is found to be only dependent on the ratio between the number of data slots available on the uplink subframe and the number of request slots for bandwidth requests. We have also provided a mechanism which dynamically adjusts the initial contention window to achieve the optimal performance. Further work will be carried out to extend our model to consider the unsaturated case.

APPENDIX A

UNIQUENESS OF THE FIXED POINT

The proof of the uniqueness of the fixed point uses the following lemma.

Lemma 1: Let $Q(j; \xi)$ be given by (7), and let

$$H(\xi; n) = \sum_{j=1}^n jQ(j; \xi) + \sum_{j=n+1}^m nQ(j; \xi).$$

Then, for $\xi \in [0, 1]$, and $d = 1, \dots, m - 1$,

- 1) $H(\xi; d)$ is increasing in ξ
- 2) $H(\xi; d)/H(\xi; m)$ is decreasing in ξ .

Proof: The proof will make repeated use of the fact that, for $a, b, X, X', Y, Y' > 0$,

$$\frac{X}{X'} < \frac{Y}{Y'} \Leftrightarrow \frac{X}{X'} < \frac{aX + bY}{aX' + bY'} \Leftrightarrow \frac{aX + bY}{aX' + bY'} < \frac{Y}{Y'}. \quad (23)$$

To show monotonicity, we will perturb ξ by some amount $\delta \in (0, 1 - \xi]$. We use the notation " \uparrow in (\cdot) " to indicate that a function is increasing in a specified argument.

N	m	d	optimal W	optimal throughput	mean delay (ms)	std dev of delay (ms)	packet loss prob
40	10	6	23	0.6133	6.8110	8.5172	0.1049
		8	23	0.4655	6.7703	8.4848	0.1025
		10	23	0.3725	6.7695	8.4841	0.1025
	20	12	20	0.6179	3.9424	3.9719	0.0998
		16	20	0.4657	3.9400	3.9693	0.0994
		20	20	0.3725	3.9400	3.9693	0.0994
50	10	6	30	0.6119	8.4076	11.1435	0.1035
		8	30	0.4644	8.2607	10.8496	0.1012
		10	30	0.3716	8.2596	10.8487	0.1011
	20	12	27	0.6181	4.7299	5.1985	0.0991
		16	27	0.4645	4.7266	5.1952	0.0987
		20	27	0.3716	4.7266	5.1952	0.0987
60	10	6	37	0.6109	9.8459	13.2976	0.1026
		8	37	0.4636	9.7832	13.2476	0.1003
		10	37	0.3710	9.7819	13.2465	0.1003
	20	12	34	0.6171	5.4802	6.3861	0.0987
		16	34	0.4637	5.4761	6.3821	0.0983
		20	34	0.3710	5.4761	6.3821	0.0983

TABLE I
OPTIMAL W AND NETWORK PERFORMANCES UNDER VARIOUS SYSTEM PARAMETERS.

Let

$$\bar{F}(j; \xi) = \sum_{i=j}^m Q(i; \xi)$$

be the complementary cumulative distribution function (ccdf) corresponding to probabilities Q . Then

$$H(\xi; n) = \sum_{j=1}^n \bar{F}(j; \xi). \quad (24)$$

We first claim that

$$\frac{\bar{F}(j; \xi + \delta)}{\bar{F}(j; \xi)} \uparrow \text{ in } j. \quad (25)$$

To see this, first note that

$$\frac{(\xi + \delta)^j (1 - (\xi + \delta))^{m-j}}{(\xi)^j (1 - \xi)^{m-j}} \uparrow \text{ in } j. \quad (26)$$

This implies

$$\frac{Q(j; \xi + \delta)}{Q(j; \xi)} < \frac{\bar{F}(j + 1; \xi + \delta)}{\bar{F}(j + 1; \xi)}$$

by backward induction on (23) from $j = m - 1$. Applying (23) to the fact that $\bar{F}(j; \xi) = Q(j; \xi) + \bar{F}(j + 1; \xi)$ gives

$$\frac{\bar{F}(j; \xi + \delta)}{\bar{F}(j; \xi)} < \frac{\bar{F}(j + 1; \xi + \delta)}{\bar{F}(j + 1; \xi)}.$$

This implies (25) and establishes the claim.

We can now prove the first part of the lemma. Note that $\bar{F}(0; \xi + \delta) = \bar{F}(0; \xi) = 1$. By the foregoing claim

$$\frac{\bar{F}(j; \xi + \delta)}{\bar{F}(j; \xi)} > \frac{\bar{F}(0; \xi + \delta)}{\bar{F}(0; \xi)} = 1,$$

thus for all $j > 0$, $\bar{F}(j; \xi + \delta) > \bar{F}(j; \xi)$. By (24), this implies that $H(\xi; d) \uparrow$ in ξ , as required.

To prove the second part of the lemma, it is sufficient to show that

$$\frac{H(\xi; m)}{H(\xi; d)} = 1 + \frac{\sum_{j=d+1}^m \bar{F}(j; \xi)}{\sum_{j=1}^d \bar{F}(j; \xi)} \uparrow \text{ in } \xi. \quad (27)$$

Let

$$A(\xi) = \sum_{j=1}^d \bar{F}(j; \xi)$$

$$B(\xi) = \sum_{j=d+1}^m \bar{F}(j; \xi).$$

Then the foregoing claim and (23) together imply that

$$\frac{A(\xi + \delta)}{A(\xi)} \leq \frac{\bar{F}(d; \xi + \delta)}{\bar{F}(d; \xi)} < \frac{B(\xi + \delta)}{B(\xi)},$$

which establishes (27) and hence the lemma. \blacksquare

We are now ready to prove the uniqueness of the fixed point.

Proof: (Theorem 1)

Since $m \leq N$, $k = m$ in the following proof. To prove the first part of Theorem 1, invert (5) to give

$$\hat{\tau}(p_c, p_d) = 1 - (1 - p_c)^{1/(N-1)}. \quad (28)$$

For any given p_d , this is continuous on the domain $p_c \in [0, 1]$, and since

$$\frac{\partial \hat{\tau}}{\partial p_c} = \frac{1}{N-1} (1 - p_c)^{1/(N-1)-1} > 0,$$

(28) is monotonic and increases from $\hat{\tau}(0, p_d) = 0$ to $\hat{\tau}(1, p_d) = 1$. On the other hand, $\tau(p(p_c, p_d))$ defined in (4) is continuous on $p_c \in [0, 1]$, and will now be shown to be monotonic non-increasing. Rewrite (4) as

$$\tau(p(p_c, p_d)) = \left(\frac{m+1}{2} + \frac{W}{2} \frac{\sum_{i=0}^{r-1} (\alpha p)^i + \alpha^r \sum_{i=r}^{R-1} p^i}{\sum_{i=0}^{R-1} p^i} \right)^{-1} \quad (29)$$

$$= \left(\frac{m+1}{2} + \frac{W}{2} \frac{\sum_{i=0}^{R-1} p^i \prod_{j=1}^i \beta_j}{\sum_{i=0}^{R-1} p^i} \right)^{-1},$$

where $\beta_j = \alpha$ if $j \leq r$, and 1 otherwise. It suffices to show that $h(p) := \sum_{i=0}^{R-1} p^i \prod_{j=1}^i \beta_j / \sum_{i=0}^{R-1} p^i$ is non-decreasing

in $p \in [0, 1]$. Let

$$A_k = \sum_{i=0}^{k-1} p^i \prod_{j=R-k+1}^{R-k+i} \beta_j$$

$$B_k = \sum_{i=0}^{k-1} p^i.$$

The monotonicity of h will be shown by induction on the statements (i) A_k/B_k is non-decreasing in p , (ii) $A_k - B_k$ is non-decreasing in p and (iii) $A_k \geq B_k$. The base case $k = 1$ is immediate. For the inductive step, assume the statement is true for some $k \geq 1$, and note that B_k is increasing in p . Then, since $\beta_{R-k} \geq 1$, $A_{k+1} - B_{k+1} = (1 + \beta_{R-k} p A_k) - (1 + p B_k) > 0$ is increasing in p , which establishes (ii) and (iii) for $k + 1$. Moreover,

$$\frac{A_{k+1}}{B_{k+1}} = \frac{1 + \beta_{R-k} p A_k}{1 + p B_k} = \beta - \frac{(\beta - 1) - \beta p (A_k - B_k)}{1 + p B_k},$$

which is increasing in p by (ii) and (iii). This proves (i) for $k + 1$ and completes the induction. The monotonicity of h is implied by the case $k = R$.

Uniqueness of the solution is proven by noting that $\tau(p(p_c, p_d))$ in (29) decreases continuously and monotonically from $\tau(p(0, p_d)) = \left(\frac{m+1}{2} + \frac{W}{2} \frac{\sum_{i=0}^{R-1} (\beta p_d)^i}{\sum_{i=0}^{R-1} p_d^i} \right)^{-1} > \hat{\tau}(0, p_d) = 0$ to $\tau(p(1, p_d)) = \left(\frac{m+1}{2} + \frac{W}{2} \frac{\sum_{i=0}^{R-1} \beta^i}{R} \right)^{-1} < \hat{\tau}(1, p_d) = 1$.

To prove the second part of Theorem 1, note that for any given $p_d \in [0, 1]$, the previous nonlinear equations (28) and (29) give a unique solution for the two unknowns τ and p_c . Denote this solution by $\langle \tau^*(p_d); p_c^*(p_d) \rangle$. Observe that both $\tau^*(p_d)$ and $p_c^*(p_d)$ are monotone decreasing functions of $p_d \in [0, 1]$ since $\tau(p(p_c, p_d)) > \tau(p(p_c, p_d + \delta)) > \hat{\tau}(0, p_d)$ for sufficiently small $\delta > 0$.

It remains to prove that, for a given $d < m$, there is a unique solution to $f(\tau, p_c, p_d) := g(\tau(p(p_c, p_d))) - p_d = 0$, where $g(\tau(p(p_c, p_d))) : [0, 1] \rightarrow [0, 1]$ is defined by the right-hand side of (8). Our task is to prove $f(\tau, p_c, p_d) = 0$ has a unique solution where the function $g(\tau(p(p_c, p_d)))$ is calculated based on the solution $\langle \tau^*(p_d); p_c^*(p_d) \rangle$ which is unique for each p_d value.

To establish the existence of a solution, observe that $f(\tau, p_c, p_d)$ is continuous and $f(\tau, p_c, 0) > 0$. This is because $g(p(p_c, 0)) > 0$ due to the fact that $0 \leq d < m$ and $\langle \tau^*(0); p_c^*(0) \rangle \neq \langle 0; 0 \rangle$. On the other hand, $f(\tau, p_c, 1) < 0$ since $g(p(p_c, 1)) < 1$ for any $0 \leq d < m$. The intermediate value theorem now guarantees a solution. Denote this solution by $\langle p_d^*(d); \tau^*(p_d); p_c^*(p_d) \rangle$ for a given $0 \leq d < m$.

To prove the uniqueness of the solution $\langle p_d^*(d); \tau^*(p_d); p_c^*(p_d) \rangle$, suppose $f(\tau^{*(1)}, p_c^{*(1)}, p_d^{*(1)}) = f(\tau^{*(2)}, p_c^{*(2)}, p_d^{*(2)}) = 0$ for $p_d^{*(2)} > p_d^{*(1)}$, where the corresponding solutions are $\langle p_d^{*(1)}; \tau^{*(1)}; p_c^{*(1)} \rangle$ and $\langle p_d^{*(2)}; \tau^{*(2)}; p_c^{*(2)} \rangle$, respectively. Because $\tau^*(p_d)$ is a monotone decreasing function in p_d we have $\tau^{*(1)} > \tau^{*(2)}$. However, observe that p in (2) is a function of τ and can be

written using (5) and (8) as

$$p = 1 - (1 - \tau)^{N-1} \frac{\sum_{j=0}^d j Q(j) + \sum_{j=d+1}^k d Q(j)}{\sum_{j=0}^k j Q(j)} \quad (30)$$

$$= 1 - \frac{H(\xi; d)}{mN\tau}, \quad (31)$$

where H is defined in Lemma 1, and $H(\xi; m) = m\xi$ is the mean of the binomial distribution. To see that p is increasing in τ , consider first the case that $\tau < 1/N$, in which case $d\xi/d\tau > 0$. By the second part of Lemma 1, the final factor in (30) is decreasing in ξ , and hence decreasing in τ , whence p is increasing in τ . Conversely, if $\tau \geq 1/N$ then $d\xi/d\tau \leq 0$, whence, by the first part of Lemma 1, $H(\xi; d)$ is decreasing in τ , and p is again increasing in τ .

Thus, $\tau^{*(2)} < \tau^{*(1)}$ should correspond to $p^{*(2)} < p^{*(1)}$ which is a contradiction because function $\tau(p)$ in (29) is a monotone increasing function. Therefore, $\langle p_d^{*(1)}; \tau^{*(1)}; p_c^{*(1)} \rangle = \langle p_d^{*(2)}; \tau^{*(2)}; p_c^{*(2)} \rangle$. ■

APPENDIX B

DERIVATION OF OPTIMAL W AND THROUGHPUT

In the following, we derive the optimal W value that maximizes the throughput in (9). Using results from Lemma 1 we know that $H(\xi; d)$ is increasing in ξ , thus maximizing (9) is equivalent to maximizing ξ given in (6). To find the maximum ξ , we set

$$\frac{\partial \xi}{\partial \tau} = N(1 - \tau)^{N-1} - N\tau(N - 1)(1 - \tau)^{N-2} = 0. \quad (32)$$

Solving (32) we obtain the optimal τ , τ_{opt} , as

$$1 - \tau_{opt} = \tau_{opt}(N - 1) \Rightarrow \tau_{opt} = \frac{1}{N}.$$

Using the relationship $\tau = 1/(B_{avg} + 1)$, we can obtain the optimal W , W_{opt} , from (3):

$$\frac{1}{\tau_{opt}} - 1 = \frac{m}{2} + \eta \sum_{i=0}^{r-1} p^i \left(\frac{\alpha^i W_{opt} - 1}{2} \right) + \eta \left(\frac{\alpha^r W_{opt} - 1}{2} \right) \sum_{i=r}^{R-1} p^i. \quad (33)$$

For large W , $W - 1 \approx W$, then (33) becomes

$$\frac{1}{\tau_{opt}} - 1 \approx \frac{m}{2} + \eta \sum_{i=0}^{r-1} \frac{p^i \alpha^i W_{opt}}{2} + \eta \frac{\alpha^r W_{opt}}{2} \sum_{i=r}^{R-1} p^i$$

$$= \frac{m}{2} + \frac{\eta W_{opt}}{2} \left[\sum_{i=0}^{r-1} (\alpha p)^i + \alpha^r \sum_{i=r}^{R-1} p^i \right]. \quad (34)$$

Therefore, W_{opt} is given by

$$W_{opt} \approx \frac{2}{\eta} \frac{1/\tau_{opt} - 1 - m/2}{\sum_{i=0}^{r-1} (\alpha p)^i + \alpha^r \sum_{i=r}^{R-1} p^i}$$

$$= \frac{2}{\eta} \frac{1/\tau_{opt} - 1 - m/2}{\frac{1 - (\alpha p)^r}{1 - \alpha p} + \frac{\alpha^r (p^r - p^R)}{1 - p}}, \quad (35)$$

where p is obtained from (2), (5) and (8) using $\tau = \tau_{opt}$. By substituting τ_{opt} into (35) we obtain (22). Finally, we can obtain the optimal throughput by using (6), (7) and (9).

REFERENCES

- [1] IEEE 802.16-2004, *IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE, October 2004.
- [2] IEEE 802.16e-2005, Amendment and Corrigendum to *IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE, February 2006.
- [3] C. Cicconetti, L. Lenzini, E. Mingozzi and C. Eklund, "Quality of Service Support in IEEE 802.16 Networks", *IEEE Network*, March/April 2006, pp. 50–55.
- [4] 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 Communications Magazine*, May 2007, pp. 132–138.
- [5] B.J. Kwak, N.O. Song and L.E. Miller, "Performance Analysis of Exponential Backoff", *IEEE Trans. on Networking*, vol. 13, no. 2, April 2005, pp. 343–355.
- [6] R. Iyengar, P. Iyer and B. Sikdar, "Delay Analysis of 802.16 based Last Mile Wireless Networks", *Proceedings, IEEE Globecom'05*, St Louis, November 2005, pp. 3123–3127.
- [7] A. Vinel, Y. Zhang, M. Lott and A. Turlikov, "Performance Analysis of the Random Access in IEEE 802.16", *Proceedings, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, September 2005.
- [8] J. He, K. Guild, K. Yang and H.H. Chen, "Modeling Contention Based Bandwidth Request Scheme for IEEE 802.16 Networks", *IEEE Communications Letters*, vol. 11, no. 8, August 2007, pp.698–700.
- [9] A. Vinel, Y. Zhang, Q. Ni and A. Lyakhov, "Efficient Request Mechanism Usage in IEEE 802.16", *Proceedings, IEEE Globecom'06*, San Francisco, November 2006.
- [10] Y.P. Fallah, F. Aghareparast, M. 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. on Vehicular Technology*, vol. 57, no. 5, September 2008, pp. 3094–3107.
- [11] R. Pries, D. Staehle and D. Marsico, "Performance Evaluation of Piggyback Requests in IEEE 802.16", *Proceedings, IEEE VTC-2007 Fall*, Baltimore, 30 September–3 October, 2007.
- [12] Wimax simulator, [Online]. Available: <<http://www.ee.cityu.edu.hk/~schan/wimax.html>>.

PLACE
PHOTO
HERE

Sammy Chan (S'87–M'89) received his B.E. and M.Eng.Sc. degrees in electrical engineering from the University of Melbourne, Australia, in 1988 and 1990, respectively, and a Ph.D. degree in communication engineering from the Royal Melbourne Institute of Technology, Australia, in 1995. From 1989 to 1994, he was with Telecom Australia Research Laboratories, first as a research engineer, and between 1992 and 1994 as a senior research engineer and project leader. Since December 1994, he has been with the Department of Electronic Engineering, City University of Hong Kong, where he is currently an associate professor.

PLACE
PHOTO
HERE

Lachlan Andrew (M'97–SM'05) received the B.Sc. degree in computer science in 1992, the B.E. degree in electrical engineering in 1993, and the Ph.D. degree in engineering in 1997, all from the University of Melbourne, Australia. Since 2008, he has been an Associate Professor at Swinburne University of Technology, Australia. From 2005 to 2008, he was a senior research engineer in the Department of Computer Science at Caltech. Prior to that, he was a senior research fellow at the University of Melbourne and a lecturer at RMIT, Australia. His research interests include performance analysis of congestion control, resource allocation algorithms and energy-efficient networking. He is an associate editor of the journal *Computer Communications* (2009–). He was co-recipient of the best paper award at IEEE MASS'07. Dr. Andrew is a member of the ACM.

PLACE
PHOTO
HERE

Hai L. Vu (S'97–M'98–SM'06) received the B.Sc./M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Budapest, Budapest, Hungary, in 1994 and 1999, respectively. From 1994 to 2000, he was a Research Engineer with Siemens AG, Hungary. During this period, his focus was on performance measurements, Internet quality of service, and IP over ATM. During 2000–2005, he was with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia. In 2005, he joined Swinburne University of Technology and is with the Centre for Advanced Internet Architectures (CAIA). He is currently an Associate Professor at the Faculty of Information and Communication Technologies (FICT), Swinburne University of Technology, Hawthorn, Victoria, Australia. Dr Vu has authored or coauthored over 90 scientific journals and conference papers. His research interests are in data network modeling, performance evaluation of wireless and optical networks, network security and telecommunication networks design.