Analytical Modeling of Contention-Based Bandwidth Request Mechanism in IEEE 802.16 Wireless Networks

Yaser Pourmohammadi Fallah, *Member, IEEE*, Farshid Agharebparast, *Member, IEEE*, Mahmood R. Minhas, Hussein M. Alnuweiri, *Member, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

***Abstract*—The IEEE 802.16 wireless metropolitan area network (WMAN) standard is a promising and cost-effective, last-mile wireless technology for the provision of broadband Internet access to end users. In this paper, we present an accurate analytical model that describes the contention-based bandwidth (BW) request scheme of the 802.16 standard, which is also known as WiMAX, for the persistent and nonpersistent request generation cases. We first model the contention procedure with a Markov chain, taking into account the exponential back-off procedure as well as the waiting time for a BW assignment and the possible timeout for lost messages. The accuracy of the model is then evaluated by comparing it with simulation results for a wide range of values of the parameters involved. We use this model to accurately calculate the capacity of the contention slots in delivering BW requests, from which the average access delay is also found. These measures are used to determine a proper configuration for the efficient operation of the contention-based BW request scheme. The proposed model provides a useful analytical tool for devising adaptive configuration mechanisms for the contention access mode of the 802.16 medium access control (MAC) layer.**

***Index Terms*—Contention-based access, IEEE 802.16, Markov chain (MC), medium access control (MAC), WiMAX, wireless networks.**

I. Introduction

FOLLOWING the widespread availability of wireless local area networks that provide last-mile wireless access, there was a great deal of interest for similar kinds of networks at a relatively larger geographical scale. This demand for wireless broadband services at a metropolitan or wide-area scale has resulted in collective efforts for the development of wireless metropolitan area network (WMAN) standards in recent years. The IEEE 802.16 standard (whose industry-led equivalent is also referred to as WiMAX) is the WMAN solution set forth by the IEEE in 2001 and later revised in 2004 and 2005 [2]. Low deployment cost is the main reason for the strong drive toward using WiMAX.

Since the 802.16 standard is sometimes viewed as a competitor to the other currently available broadband solutions such as DSL, cable, and WiFi (IEEE 802.11 [4]) in offering high data rates, there is a real need for evaluating its performance, in terms of efficiency and throughput. Some efforts are reported in the literature regarding the performance study of quality-of-service (QoS) scheduling in cable/wireless systems as well as with regard to the performance evaluation of the 802.16 physical (PHY)-layer mechanism. However, to the best of our knowledge, no detailed analytical model has been presented, specifically for the performance evaluation of the contention-based access scheme at the medium access control (MAC) layer in 802.16 WMANs. This access scheme is expected to be the main mode of operation for supporting the best effort (BE) class of traffic generated by most Internet applications (web surfing, FTP, etc.). It is therefore important to study the required configurations for such services in the 802.16 MAC.

The significance of the analysis presented in this paper is that it provides an insight on the roles and effects of various parameters involved in the contention-based bandwidth (BW) request mechanism of the IEEE 802.16-based access. This is particularly important since its implementation has been left up to developers and has not been preset in the standard. For example, we will show in Section IV that the length of the initial contention window and the number of allocated contention slots play important roles in determining the system utilization in contention mode. For practical WiMAX cases with a large number of stations, the choice of the number of allocated contention slots will have a significant bearing on the overall system performance.

We briefly review some related works reported in the literature. Hawa and Petr approximated the contention access scheme of the 802.16 as a slotted Aloha system [5]. However, as we will explain later, the 802.16 contention access process cannot be accurately modeled as a slotted Aloha scheme. Kim and Yeom have presented and then comprehensively evaluated three BW allocation request schemes for the 802.16 BE class through an analytical study and simulations [6]. The authors also show the extension of these schemes to non-real-time polling service (nrtPS)-class traffic. However, their study does not provide an

Manuscript received July 6, 2007; revised October 20, 2007 and November 23, 2007. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The review of this paper was coordinated by Dr. P. Lin.

Y. Pourmohammadi Fallah, F. Agharebparast, M. R. Minhas, and V. C. M. Leung are with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: yasserp@ece.ubc.ca; farshid@ece.ubc.ca; mrminhas@ece.ubc.ca; vleung@ece.ubc.ca).

H. M. Alnuweiri was with the University of British Columbia, Vancouver, BC V6T 1Z4, Canada. He is now with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar (e-mail: hussein.alnuweiri@qatar.tamu.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.2008.914474

analytical model of the contention scheme that considers all the involved parameters, such as the contention window size and the number of contention slots per frame. Cicconetti *et al.* have conducted a performance study to evaluate QoS provisioning available in the 802.16 MAC layer [7]. They simulated two different scenarios to find a number of MAC performance metrics such as average delay and delay variation. Extending their work in [8], they conducted a simulation study to evaluate a number of typical QoS performance metrics such as utilization, delay, throughput, etc., in 802.16 MAC. In their simulation framework, they considered various traffic models (both data and multimedia) under almost all the scheduling classes available in 802.16 MAC. In another work, Cho *et al.* proposed a new MAC architecture with a BW allocation, as well as an admission control technique for better QoS support in 802.16 [9]. Based on their analytical model and simulations, it is stated that the proposed architecture can enhance QoS support and system throughput by balancing the channel utilization among a number of traffic classes. Gusak *et al.* carried out a comparison of two well-known scheduling schemes, namely, weighted fair queuing and weighted round robin, when used in 802.16 MAC in terms of average packet delay [10]. They also proposed an algorithm for dynamic ratio adjustment of *uplink* (UL) and *downlink* (DL) durations within a frame. Based on their simulation of a relatively large 802.16 network, with one base station (BS) and 100 subscriber stations (SSs), the authors show that their propositions can improve performance.

In another related work, Hoymann presented an analytical model for deriving some key performance measures like capacity, throughput, and delay of PHY- and MAC-layer protocols employed in 802.16 [11]. A prototype for the two 802.16 protocol layers was also implemented to validate the theoretical results. This work focused more on MAC operation overhead and did not offer an analysis of the contention BW request process. Ghosh *et al.* presented a simulation study to estimate the achievable MAC throughput in a WiMAX deployment under various PHY parameters such as modulation schemes and antenna configuration [12]. The authors propose a number of enhancements at the PHY layer of 802.16 that, in their opinion, will lead to higher data rates and robustness of the WiMAX systems.

Despite numerous works on analyzing and enhancing the 802.16 MAC, no detailed analysis of the 802.16 contention access scheme is provided. The objective of this work is to develop an accurate analytical model for evaluating the performance of the contention-based access mechanism in 802.16. We propose a 2-D Markov chain (MC) model for this purpose. Such a model can be used for 802.16 MAC parameter selection and configuration. We confine our analysis to the BE and nonreal-time classes of traffic, for which a BW request can be sent using the contention access scheme. Such traffic classes are used for common data applications such as web surfing, FTP, telnet, etc. We use the proposed model to provide performance measures such as the efficiency and capacity of the contention period.

The paper is organized as follows. Section II describes some basic technical details of the IEEE 802.16 standard, including the MAC operation and the contention-based access mechanism. Section III presents a detailed description of the proposed analytical model followed by a discussion on model validation with the help of simulation results in Section IV. Finally, we conclude the paper in Section V.

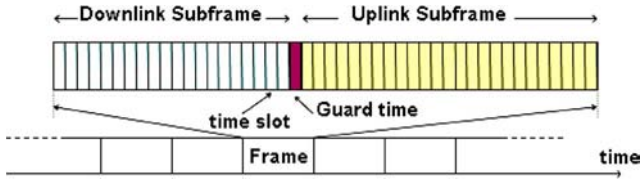

Fig. 1. IEEE 802.16 TDD frame structure.

II. IEEE 802.16 WMAN

The IEEE 802.16 standard describes the MAC and PHY layers of a WMAN. The standard specifies two modes of operation: 1) the point-to-multipoint (PMP) mode and 2) the optional mesh mode. We focus on the PMP mode, in which a BS regulates all the communication in the network and SSs are only allowed to communicate with the BS [2].

In contrast to the IEEE 802.11 standard [4], 802.16 achieves duplex operation through either frequency-division duplexing or time-division duplexing (TDD). Thus, the communication path between an SS and the BS has two directions: UL channel (from an SS to the BS) and DL channel (from the BS to an SS). Time in the UL channel is usually slotted (minislots) and shared using time-division multiple access, whereas on the DL channel, the BS uses a continuous time-division multiplexing scheme, as shown in Fig. 1. Other multiple access schemes such as orthogonal frequency-division multiple access have also been specified in later versions of the standard [2], [3]. This mode is recommended for mobile applications.

The duration of the DL or UL subframes in TDD mode is determined by the BS in a dynamic manner. Each subframe consists of a number of time slots. The UL channel is divided into a sequence of minislots to which SSs can access in a synchronized manner controlled by the BS. Each SS that needs to send data UL has to first request BW from the BS. The BS will then assign UL minislots (subject to availability and scheduling rules) to the SS in the next UL subframes. The synchronization between the BS and SSs is done through UL and DL maps (UL-MAP and DL-MAP) that are broadcast at the beginning of a DL subframe. All SSs listening to the DL frame will parse the received maps and will know in which DL time slots they should receive data from the BS and in which UL time slots they can transmit their data (see Fig. 2 and [7]).

Two modes for transmitting the BW request are defined in IEEE 802.16: 1) contention mode and 2) contention-free mode (polling). In the contention mode, SSs send BW requests during predetermined contention periods in the UL frame. Depending on the number of SSs contending, collision may happen. A collision resolution mechanism is specified in the standard to deal with such situations. In the contention-free mode, the BS assigns BW request opportunities to the stations, and SSs reply by sending BW requests. The standard also allows piggybacking BW requests on data frame transmissions for certain types

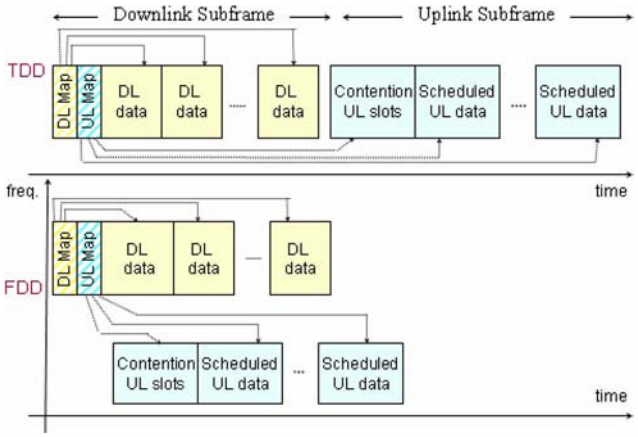

Fig. 2. IEEE 802.16 MAC frames.

of traffic; however, this feature is optional, and SSs do not have to implement it.

The IEEE 802.16 MAC defines four types of service and the rules for the BW request modes of each type. These services are the following: 1) unsolicited grant service (UGS); 2) real-time polling service (rtPS); 3) nrtPS; and 4) BE service. When a UGS is set up, the BS is responsible for assigning fixed-size periodic data grants to the UGS flow; the SS is not allowed to explicitly request BW for an established UGS flow. For rtPS and nrtPS, the SSs are polled through the unicast request polling; normally, rtPS flows are assigned more request opportunities and receive a more predictable service. However, rtPS flows are banned from using any contention requests, whereas nrtPS is allowed to request BW through contention requests. Similar to nrtPS flows, the BE flows are allowed to request BW through the contention mode. However, no periodic unicast polling is done for BE services.

BW requests could be of two types: 1) incremental and 2) aggregate. With incremental requests, the station indicates how much the queue length has increased for the corresponding connection, whereas aggregate requests indicate the total length of the queue for the connection. Since aggregate-type BW requests are naturally self-correcting, the stations should periodically use this method to correct any miscalculation of the queue length (due to requests being lost) at the BS. For this reason, the standard mandates that the BW requests sent in contention modes must be of the aggregate type [2].

It must be noted that the standard does not regulate how often the BW requests are sent. This decision is implementation dependent, and there is no universal choice suitable for all situations. A simple choice could be that stations transmit BW requests as long as the backlog length of the corresponding queue (connection) is increasing. Another option is that regardless of the queue length, as long as data arrives in the queue, the stations send periodic BW requests. However, the length of the period can vary from milliseconds to seconds. Since BW requests are specific to each connection, a station may have several requests for transmission pending at any given time. The standard leaves the design of the request management entity to vendors, as there is no universal solution. For this reason, we model the request generation process with a random Poisson process.

In this paper, we study the performance of the contention mode access of the IEEE 802.16 standard; thus, we only consider the nrtPS and BE classes of service. An evaluation of the MAC performance for the deterministic modes can be found in [7]–[10].

A. Contention Access in IEEE 802.16 MAC

The MAC layer of IEEE 802.16 specifies the rules for the contention-mode BW request. A contention period is a predetermined number of minislots at the beginning of a UL subframe. The contention period is divided into an integer number of transmission opportunities (TOs) and is called an information element. Each TO can be used for transmitting only one BW request. If more than one station tries to transmit in the same TO, collision happens. Since it is not practically possible for SSs to sense the UL channel to detect a collision, the SSs can only know of the success of their BW request transmission if they receive a response in the form of a BW grant in the subsequent frames.

A station that does not receive a response to its BW request by a certain deadline (called T16, with a minimum length of 10 ms) assumes that either a collision happened or resources are not available at the BS. In either case, since the SS cannot determine the cause, it assumes that a collision happened and uses an exponential binary back-off procedure to resolve the collision.

The collision resolution and avoidance regulation of 802.16 MAC requires each SS to wait a random number of TOs before attempting a transmission in the contention period. This number is called the random back-off number and is chosen from an interval of (0, $CW - 1$), in which CW is the contention window size and is initially set to the minimum or initial contention window size of W. After each collision (when the response to the request is not received for T16 duration), the contention window size doubles ($CW = 2^i \cdot W$ after ith collision). The doubling continues until the maximum contention window size is reached after m retransmissions ($CW = 2^m \cdot W$). The values of minimum and maximum contention window sizes can be dynamically adjusted by the BS. When a successful transmission happens or the packet is dropped due to reaching the retransmission limit, the contention window size is reset to W.

The size of the contention period is determined by the BS and may be different in each frame. If the random value of the back-off counter does not reach zero within a contention period, its countdown is frozen at the end of the contention period and resumes in the next contention period.

III. Analytical Modeling of the Contention-Based Access Scheme

In this section, we analyze the performance of the contention access mode of 802.16 in terms of its capacity to deliver BW requests from stations to the BS. The analysis in this section is different from the classic slotted ALOHA system analysis in that we model the exponential back-off procedure for the first transmission and retransmissions and do not assume an infinite

number of stations. In addition, our analysis considers each station rather than the system as a whole on the average. For our analysis, we assume that there are n stations with traffic of either type BE or type nrtPS that use contention access to request BW.

We present a 2-D MC model for the contention BW request operation. The use of the Markov model was initially inspired by a model for the IEEE 802.11 distributed coordination function [13], which was verified and used in numerous other works [14]–[16]. The model in [13] assumes an 802.11 network under saturation condition, where stations always have data to send and persist in contending to access the channel. This model was later extended for nonsaturation or finite-load conditions in several other works [17]–[20]. However, the finite-load models for 802.11 or 802.11e are not directly applicable in our modeling for the 802.16 contention access. This is due to the fact that the pattern of request generation in 802.16 MAC is generally completely different from the pattern of data arrival, whereas for 802.11 nonsaturation models, these patterns are identical or directly related. In 802.16, the requests are generated based on the decisions made by a vendor-defined algorithm, whereas in 802.11, the channel access is attempted for each packet. The model presented in this paper is much more detailed than the models for 802.11 to capture the precise operation of the 802.16 MAC and incorporate its different acknowledgement timeout procedure.

In our model, we do not assume any specific algorithm for BW request generation. Instead, we assume that requests are generated at random intervals, emulating the general behavior of user-defined algorithms for request generation. In particular, we assume a Poisson process to be a good representation of the request generation pattern. Using a Poisson process is generally reasonable because requests for BW are generated for several independent connections within a station and by user-defined algorithms. It is worth emphasizing that the volume of the generated BW requests is not directly correlated with the volume of the actual data, as BW requests are generated based on a vendor-selected algorithm. If the request generation rate is high, requests are generated back to back. We call this mode persistent request generation or the saturation mode. The case where requests are generated with larger intervals is called *infrequent request generation* (also called the *nonsaturation* case).

It is important to note that there is no universal algorithm for BW request generation, and using a model such as the one assumed here is inevitable. In addition, since piggybacking BW requests is optional and may not be implemented in all SSs, we assume that it is not used, and we do not base our model on this feature. It is, however, possible to capture the effect of piggybacking on an average basis for specific request generation algorithms. To do so, one can adjust the mean of the Poisson process that models request generation to account for the reduction of the number of requests when piggybacking is used. However, this requires knowledge of the average interarrival times of the requests in a system that uses piggybacking in addition to contention requests. Such knowledge is implementation dependent; thus, we do not consider it in our model.

We begin our analysis by modeling the back-off procedure and BW request grant waiting periods. For ease of reference, a list of important notations is provided in Table I. The objective of the analysis is to find the efficiency of the contention period in terms of the number of slots that carry a BW request and result in a successful grant of the requested BW. Since the back-off counter value is chosen randomly, we can model it as a random variable $b(t)$. The value that is selected for the back-off counter at the beginning of each back-off process depends on the previous retransmission history of the SS; thus, $b(t)$ is non-Markovian. On the other hand, we can also model the contention window size as a random process $s(t)$. Subsequently, we can model the back-off procedure using a 2-D Markovian process $(s(t), b(t))$. Similarly, when the station is waiting for a response from the BS for its BW request to be granted, it is in a waiting period. In this period, the station does not know whether the BS has received its request or a collision has occurred. To differentiate between these two different aspects of the chain, the states that represent a station in the waiting period under collision are labeled $(s(t), b(t))'$, and those that are collision free are labeled $(s(t), b(t))''$.

TABLE I
LIST OF NOTATIONS

Notation	Description
W	Initial or minimum contention window size, also denoted as W_0
n	Number of active Subscriber Stations
$b(t)$	A random variable used to represent the backoff counter value
$s(t)$	A random variable used to represent the contention window (CW) size
p	Probability of a transmission attempt resulting in collision
q	Probability of BS granting BW to a successfully transmitted request
$(s(t), b(t))$	2-D Markov process to model the backoff procedure
$(s(t), b(t))'$	represents the waiting states of Markov chain after a collision event
$(s(t), b(t))''$	represents the waiting states of Markov chain after a collision-free transmission
N^*	Average number of waiting slots until the end of the current contention period
N	Number of contention slots in each frame
p_f	Probability of failure, due to collision or BW not being granted by the BS.
W_i	Contention window size after the i'th retransmission
$\bar{W}$	Average value of the contention window size
N_D	Expected value of the request generation interval
N_D^*	Average number of slots spent in the holding state, used for modeling infrequent request arrival.
$\tilde{P}\{i,k\}$	Stationary probability of being in any of the states (i,k), $(i,k)'$, or $(i,k)''$
P_s	Probability of any given contention slot (TO) containing a successful request
P_{col}	Probability of any given contention slot (TO) containing a collided request
P_{idle}	Probability of any given contention slot (TO) being idle (containing no request)
P_{tr}	Probability of a given contention slot (TO) containing a transmission
P_{succ}	Probability that a given transmission in a TO being successful in receiving bandwidth grant
C	Number of requests successfully responded to by the BS in one frame

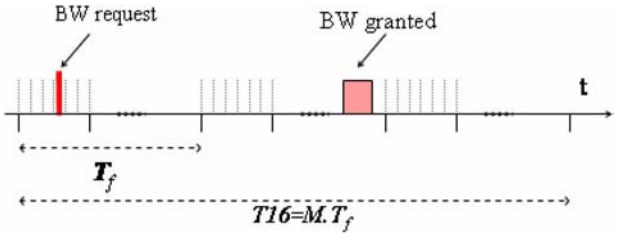

Fig. 3. Timing relationships between the contention period, the frame time, and the response timeout.

This 2-D MC modeling is possible only if we can assume independent transition probabilities between the states. This is achieved through a key approximation by assuming that the probability of collision in each slot, which is represented by p, is constant and is independent of the retransmission history of the system. We refer to p as the *conditional collision probability*. This is the probability of a collision for a request that is transmitted on the channel. We also assume that a transmission in each slot happens with equal probability, independent of the position of the slot in the contention period, and independent of the previous collisions. In Section IV, we show that the Markov model is justified and accurate when the number of stations and the contention window size are not small (e.g., both are greater than five), which is an assumption that is true in practice.

Using the independence assumption, the back-off state represented by the 2-D stochastic process is Markovian. We adopt a discrete MC to express this 2-D process. However, we must emphasize that the transition between different states of this MC does not always happen at fixed intervals. In fact, the transitions happen only at the beginning of TOs in the contention period. During contention-free access (DL and UL scheduled slots), the back-off process and, hence, the MC are frozen for all stations. This waiting period has no effect on the MC model, since the chain is only modeling the contention period, and transitions only happen in the minislots of the contention period. Fig. 3 illustrates the timing relationships as described above.

After each transmission, the transmitting station has to wait for a response from the BS. A BW will be granted if the BW request message is received successfully (no collision) and there is enough BW available. In the case of a collision or unavailability of BW, the station has to wait for some timeout duration (T16 seconds) and then start contending for a BW request again. Without loss of generality, we assume that the timeout value is M frames long. In other words, timeout happens in the Mth frame after the frame in which the request was sent. As discussed earlier, we have modeled this waiting period with states tagged with superscripts $'$ or $''$ to distinguish them from the contention procedure in our MC. We assume that if the BS receives a BW request, it will grant the request with probability q in one of the M frames within timeout period T16. It is reasonable to assume that q is constant for all waiting states of the MC. In fact, q is controlled by the admission control scheme of the network and is independent of the operation of the MAC layer.

The 2-D MC that describes the process $(s(t), b(t))$ is depicted in Fig. 4. This model is governed by the following probabilities for the state transitions that happen at the beginning of each slot in the contention period:

$$P\{i,k|i,k+1\}=1, \qquad k\in(0,W_i-2); \quad i\in(0,m) \tag{1.A}$$

$$P\left\{(i,1)''|i,0\right\}=1-p, \qquad i\in(0,m) \tag{1.B}$$

$$P\left\{(i,1)'|i,0\right\}=p, \qquad i\in(0,m) \tag{1.C}$$

$$P\left\{(i,N^*+j.N+1)''|(i,N^*+j.N)''\right\}=1-q \qquad j\in(0,M-2); \quad i\in(0,m) \tag{1.D}$$

$$P\left\{0,k|(i,N^*+j.N)''\right\}=q/W_0, \qquad j\in(0,M-1) \qquad k\in(0,W_0-1); \quad i\in(0,m) \tag{1.E}$$

$$P\left\{i,k|\,(i-1,(M-1)N+N^*)''\right\}=(1-q)/W_i \qquad k\in(0,W_i-1); \quad i\in(1,m) \tag{1.F}$$

$$P\left\{i,k|\,(i-1,(M-1)N+N^*)'\right\}=1/W_i \qquad k\in(0,W_i-1); \quad i\in(1,m) \tag{1.G}$$

$$P\left\{m,k|\,(m,(M-1)N+N^*)''\right\}=(1-q)/W_m \qquad k\in(0,W_m-1) \tag{1.H}$$

$$P\left\{m,k|\,(m,(M-1)N+N^*)'\right\}=1/W_m \qquad k\in(0,W_m-1). \tag{1.I}$$

We use the following notations for the set of equations in (1):

$$\begin{aligned}
&P\{i_1,k_1|i_0,k_0\}\\
&\quad=P\left\{s(t+1)=i_1,b(t+1)=k_1|s(t)=i_0,b(t)=k_0\right\}\\
&P\left\{i_1,k_1|(i_0,k_0)'\right\}\\
&\quad=P\left\{s(t+1)=i_1,b(t+1)=k_1|\,(s(t)=i_0,b(t)=k_0)'\right\}\\
&P\{i_1,k_1|(i_0,k_0)''\}\\
&\quad=P\left\{s(t+1)=i_1,b(t+1)=k_1|\,(s(t)=i_0,b(t)=k_0)''\right\}.
\end{aligned} \tag{2}$$

The first four probabilities of (1) describe the transitions in the $b(t)$ direction of the MC. The first transition probability, i.e., (1.A), represents the countdown period for a station that decrements from a random number uniformly chosen from (0, W_i) to 0 during the contention period. W_i is the size of the contention window after the ith retransmission ($W_i=2^i\cdot W$). The second and third transition probabilities, i.e., (1.B) and (1.C), describe the transition to the first waiting state. This happens with a probability of one; however, as mentioned

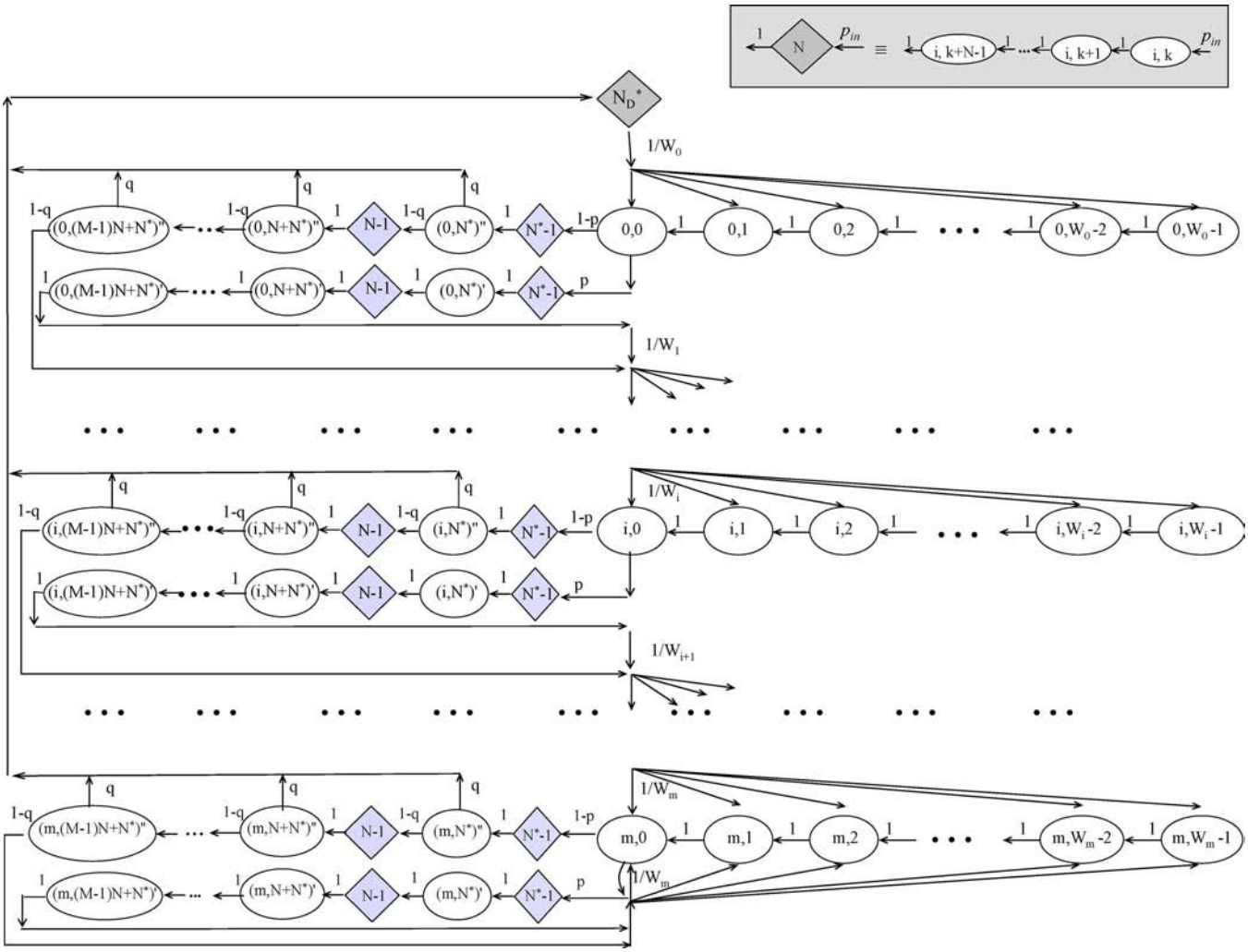

Fig. 4. Markov model for the backoff procedure.

before, two branches are considered in each row of the waiting section of the MC. Here, we account for the no-collision situations separately from the collision situations. The transition probability of the former is $1-p$, and that of the latter is p. The process stays in the waiting period with a transition probability of $1-q$ if a BW has yet to be assigned at each frame, as described by (1.D).

On the other hand, (1.E) accounts for a successful BW assignment where the station starts over again to contend for the next BW request from stage 0. The probability is divided by W_i, since the value of the actual back-off is uniformly chosen from a window of (0, W_i-1).

The process steps down in the $s(t)$ direction, when there is no BW available or when a collision occurs. The former is represented by (1.F) and the latter by (1.G). To consider both of these situations combined, a new probability p_f, called the probability of failure, needs to be used. This parameter is derived and explained later in this section. It can be shown that

$$P\{i,k|i-1,0\}=p_f/W_i, \qquad k\in(0,W_i-1); \quad i\in(1,m).$$

The last two equations, i.e., (1.H) and (1.I), are the special cases of the previous two equations, when the back-off window has reached its maximum allowable value and is not increased anymore.

In (1), N^* denotes the number of "waiting" slots until the end of the current contention period. Note that when a station transmits, it withdraws from the contention process until the end of the current contention period. Since this duration depends on when the back-off counter expired, we use N^* as the average number of waiting slots. After the contention period, if a collision has happened, the station has to wait for $N(M-1)$ additional slots in the timeout period before it can reenter the back-off process. If no collision occurs, the station waits until it receives a TO from the BS, or its BW request timer expires after M frames $(M.T_f)$. A failure to receive the TO allocation, despite successful transmission of the BW request, can be due to unavailability of BW to be allocated by the BS (modeled by probability q).

We can find the average value for N^* using the average value of the contention window size $\overline{W}$. Note that the contention window size is variable and controlled by the truncated exponential back-off process. To calculate the expected value of the contention window, we first need to find $\widehat{P}\{i,k\}$, the stationary probability of being in any of the states (i,k), $(i,k)'$, and $(i,k)''$, for $\forall k$ (row i of the MC). The contention window

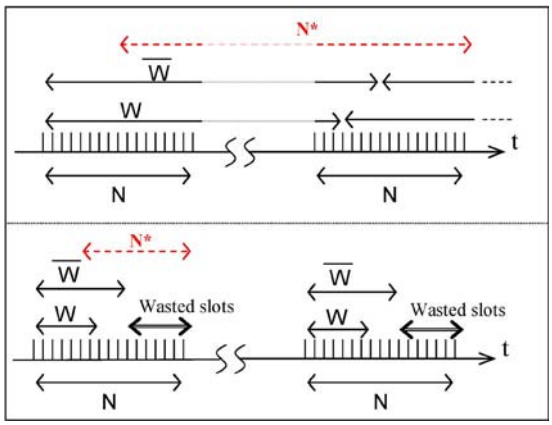

Fig. 5. Relationship between the contention window size and the number of contention slots in each frame.

size is the same for all states in each row of the MC. From (1), we know that $p_1 = \widehat{P}\{i,k\} = p_0 p_f$. Therefore, we have

$$\sum_{i=0}^{m} p_0 p_f^i = 1 \rightarrow p_0 \sum_{i=0}^{m} p_f^i = p_0 \left(\frac{1-p_f^{m+1}}{1-p_f}\right) = 1$$

$$\rightarrow p_0 = \frac{1-p_f}{1-p_f^{m+1}}.$$

Thus, the average value of the contention window is

$$\overline{W} = E\{W\} = \sum_{i=0}^{m} W_i p(W_i)$$

$$= \sum_{i=0}^{m} (W2^i) p_0 p_f^i = W p_0 \frac{1-(2p_f)^{m+1}}{1-2p_f}$$

which results in

$$\overline{W} = W \frac{\left(1-(2p_f)^{m+1}\right)(1-p_f)}{(1-2p_f)\left(1-p_f^{m+1}\right)}.$$

To derive N^*, we note that the value of the back-off counter is uniformly chosen from (1, $\overline{W}$); $\overline{W}$ may span several contention periods, ending in the middle of the last contention period (Fig. 5). If the value of the randomly chosen back-off counter is in any of the contention periods that are completely included in $\overline{W}$, the average number of waiting slots will simply be $N/2$. The probability of this situation is $P_N = [\overline{W}/N]^*(N/\overline{W})$ due to the fact that the back-off counter is uniformly chosen in (1, $\overline{W}$). If the random number falls in the last partially included contention period, with a probability $P_x = (\overline{W} - [\overline{W}/N]^* N)/\overline{W}$, the number of waiting time slots becomes the following:

$$N_x = \left(\overline{W} - [\overline{W}/N] \cdot N\right)/2 + \left(\lceil \overline{W}/N \rceil \cdot N - \overline{W}\right).$$

Taking a weighted average using the probabilities of the two cases, we find N^* as

$$N^* = P_x \cdot N_x + P_N \cdot \frac{N}{2}.$$

After a successful assignment of the requested BW, a station may refrain from sending the next BW request. This can be due to the implementation requirements so that the requests are sent with predetermined intervals or for the station to finish up sending the already buffered data. The intervals are dependent on the amount of the incremented received data and the adopted algorithm and are not set by the standard. To model such practical scenarios, a holding state with duration N_D^* is considered in the MC. The holding state duration, in terms of the number of slots, is found as

$$N_D^* = Max\{0, N_D - \bar{\theta}\} \tag{3}$$

where N_D is the average (mean of the Poisson process) implementation-dependent interrequest intervals for BW requests (presented in terms of the number of contention slots), and $\bar{\theta}$ is the average number of contention slots a station takes to successfully transmit a BW request. Similar to the procedure for the calculation of $\overline{W}$, assuming that θ_i is the average time (in terms of the number of contention slots) that the MC is in any of the states (i,k), $(i,k)'$, and $(i,k)''$, for $\forall k$ (row i of the MC), then

$$\theta_i = \frac{W_i}{2} + p.((M-1)N + N^*) + (1-p)\left(N^* + \sum_{j=1}^{M-1} N(1-q)^j\right).$$

Therefore, we have

$$\bar{\theta} = E\{\theta_i\} = \sum_{i=0}^{m} \theta_i . P(\theta_i)$$

$$= \sum_{i=0}^{m} \left\{ \frac{W_i}{2} + p.((M-1)N + N^* + (1-p)(N^* + \sum_{j=1}^{M-1} N(1-q)^j \right\} p_0 p_f^i$$

$$= p_0 \left\{ \frac{W}{2} . \frac{1-(2p_f)^{m+1}}{1-2p_f} + \left(p.(M-1)N + N^* + (1-p).N \frac{1-q-(1-q)^{m+1}}{q} \right) . \frac{1-p_f^{m+1}}{1-p_f} \right\}$$

$$= \frac{\overline{W}}{2} + p.(M-1)N. + N^* + (1-p).N \frac{1-q-(1-q)^{m+1}}{q}.$$

By adjusting the value of N_D, we can model most practical scenarios and adopted BW request generation algorithms. Our

only assumption is that on the average, the BW requests are generated according to a random process, such as Poisson.

Having found all the transition probabilities and related parameters for the Markov model, we now solve the model to obtain a closed-form solution for the stationary probabilities of the chain, which is denoted by $b_{i,m}$. First, note that from the global balance equation ($b_{i,j}.\sum_k\sum_l P\{i,j|k,l\} = \sum_k\sum_l b_{k,l}P\{k,l|i,j\}$, [21]), we have

$$b''_{i,1}.1 = (1-p)b_{i,0}$$
$$b''_{i,N^*+j.N+1}.1 = (1-q)b''_{i,N^*+j.N}$$

which results in $b''_{i,N^*+(M-1).N}.1 = (1-q)^{M-1}(1-p)b_{i,0}$.

Now, if we use the global balance equation again for each of the $(i,0)$ states, $i \in (1,m)$, we have

$$\begin{aligned} b_{i,0}.1 &= p.b_{i-1,0} + (1-q)^M(1-p)b_{i-1,0} \\ &= \left(p + (1-q)^M(1-p)\right) b_{i-1,0} \end{aligned}$$

thus

$$b_{i,0} = \left(p + (1-q)^M(1-p)\right)^i b_{0,0}. \tag{4}$$

As is seen from the above equation, the probability of transition to the next stage is, in fact, the probability that a collision happens or BW is not available. For this reason and for a clearer notation, we define p_f as the *probability of failure* by $p_f = p + (1-q)^M(1-p)$. Therefore, we have

$$b_{m-1,0}.p_f = (1-p_f)b_{m,0} \rightarrow b_{m,0} = \frac{(p_f)^m}{1-p_f}b_{0,0}. \tag{5}$$

Since the chain flows similarly and due to its regularity for each $k \in (1, W_i - 1)$, we have

$$b_{i,k} = \frac{W_i - k}{W_i}.\begin{cases} (1-p_f)\sum_{j=0}^{m} b_{j,0}, & i = 0 \\ p_f.b_{i-1,0}, & 0 < i < m \\ p_f.(b_{m-1,0} + b_{m,0}), & i = m. \end{cases} \tag{6}$$

From the relationship described in (2) and considering the fact that $\sum_{i=0}^{m} b_{i,0} = b_{0,0}/(1-p_f)$, we can rewrite (3) as

$$b_{i,k} = \frac{W_i - k}{W_i} b_{i,0}, \qquad i \in (0,m); \quad k \in (0, W_i - 1). \tag{7}$$

Now, knowing that the sum of occupancy probabilities in the MC is one, we use (2) and (4) to express $b_{0,0}$ only in terms of system parameters of p_f (or p and q), W, and m. To calculate this sum, we divide the states with occupancy probability $b_{i,k}$ into three groups: 1) the back-off states on the right side of the chain with $k \geq 0$; 2) the timeout states in the left side with $k < 0$; and 3) the waiting states in the left side with $k < 0$. This derivation of $b_{0,0}$ is given in (8), shown at the bottom of the page.

Using the MC model and noting that stations whose back-off counters reach zero will attempt a transmission, we derive the following equation that expresses the probability of transmission by a given station in a given slot:

$$\tau = \sum_{i=0}^{m} b_{i,0} = \frac{b_{0,0}}{1-p_f}. \tag{9}$$

The conditional collision probability p can be independently calculated, assuming that τ is given. This probability is found by noting that its complement $(1-p)$ is the probability that

$$\begin{aligned} 1 &= \sum_{i=0}^{m}\left\{\sum_{k=1}^{W_i-1} b_{i,k} + \sum_{k=1}^{(M-1)\cdot N+N^*} b'_{i,k} + \sum_{k=1}^{(M-1)N+N^*} b''_{i,k} + b_{i,0}\sum_{k=1}^{N_D^*}(1-P_f)\right\} \\ &= \sum_{i=0}^{m}\left\{\sum_{k=0}^{W_i-1} b_{i,k} + \sum_{k=1}^{(M-1)\cdot N+N^*} pb_{i,0} + \sum_{k=1}^{N^*}(1-p)b_{i,0} + \sum_{k=1}^{(M-1)} N\cdot(1-p)(1-q)^k b_{i,0} + N_D^*(1-P_f)b_{i,0}\right\} \\ &= \sum_{i=0}^{m} b_{i,0}\frac{W_i+1}{2} + \left\{(M-1)\cdot N + N^*)\cdot p + N^*(1-p) + (1-p)N.\sum_{j=1}^{M-1}(1-q)^j\right\} \\ &\quad \cdot \sum_{i=0}^{m} b_{i,0} + \sum_{i=0}^{m} N_D^*(1-P_f)b_{i,0} \\ &= \frac{b_{0,0}}{2}\left\{W\left(\sum_{i=0}^{m-1}(2p_f)^i + \frac{(2p_f)^m}{1-p_f}\right) + \frac{1}{1-p_f}\right\} \\ &\quad + \left\{(M-1)N\cdot p + N^* + (1-p)\left(\frac{1-q-(1-q)^M}{q}\right) + N_D^*(1-P_f)\frac{b_{0,0}}{1-p_f}\right\} \end{aligned}$$

$$b_{0,0} = \frac{1}{\frac{W}{2}\left\{\frac{1-(2p_f)^m}{1-2p_f} + \frac{(2p_f)^m}{1-p_f}\right\} + \frac{1/2+(M-1)N\cdot p+N^*}{1-p_f} + (1-p)\left(\frac{1-q-(1-q)^M}{q.(1-p_f)}\right)N + N_D^*} \tag{8}$$

given a station transmits, none of the other $n-1$ stations transmits. Therefore, p is expressed by

$$p = 1 - (1-\tau)^{n-1}. \tag{10}$$

Using these two equations, one can find τ and p based on the known values of maximum and minimum contention window sizes and n (the number of stations). Numeric methods such as bisection or Newton–Raphson can be used to solve the above equations.

A. Analysis of the System Capacity in Granting BW Requests

To analyze the capacity of the contention-mode BW request delivery, we need to find the efficiency of the system in terms of the ratio of TOs that have successfully delivered BW requests in a given contention period. The ratio of successful TOs can be found by determining the probability that any given TO experiences a successful transmission of a BW request. To find this probability—denoted by P_s—we first need to compute a number of related probabilities. These probabilities are the following: 1) P_{tr}, i.e., the probability of a TO containing a transmission, and 2) P_{succ}, i.e., the probability that a given transmission in a TO is successful in receiving a BW grant from the BS.

The *probability of transmission* P_{tr} is the probability that at least one station transmits in a slot and is found by using τ (the probability of a station transmitting)

$$P_{tr} = 1 - (1-\tau)^{n}. \tag{11}$$

The *probability of success* p_{succ} is the probability that only one station transmits and that there is enough BW, given that there is a transmission. Using the conditional probability and Bayes' formula, P_{succ} is equal to the probability that none of the remaining $n-1$ stations transmits (denoted by the event A) "and" there is enough BW to respond to the request in one of the M frames before the deadline (denoted by the event B), given that there is a transmission (denoted by the event C)

$$P_{succ} = P(AB|C) = \frac{P(AB,C)}{P(C)}.$$

Since $P(AB,C) = P(A,C)P(B) = \binom{n}{1}\tau(1-\tau)^{n-1}.(1-(1-q)^{M})$ and $P(C) = p_{tr}$, we have

$$P_{succ} = \frac{n\tau(1-\tau)^{n-1}}{P_{tr}}.\left(1-(1-q)^{M}\right). \tag{12}$$

Now, we can find the probability of a slot containing a successful BW request, which is denoted by P_s, as follows:

$$P_s = P_{succ}.P_{tr} = n\tau(1-\tau)^{n-1}.\left(1-(1-q)^{M}\right). \tag{13}$$

P_s represents the efficiency of the contention access mechanism of the 802.16 MAC. Since P_s specifies the probability of any given TO containing a successful request, we can derive the number of successful requests in any given time frame T by knowing the number of contention TOs available in that time frame. For example, if, in a frame time T_f, the BS assigns N distinct TOs for the contention period, the number of requests that are successfully served are $N.P_s$. Therefore, we can define the system capacity for a successful grant of BW in each contention period (of length N) as

$$C = N.P_s. \tag{14}$$

This means that in one frame, only C requests are successfully responded to by the BS. Therefore, if there are n persistent stations it will take $D = n/C$ frames to respond to all the requests. We call this duration the access delay of stations in saturation, meaning that it will take a station, on average, at most D-frame durations to receive a response from the BS.

In Section IV, we use P_s and C as the considered system performance metrics and present how different system parameters such as N, M, n, and W affect their values.

IV. Model Evaluation and Performance Analysis

To evaluate the model developed in Section III, we numerically solve (9) and (10) and then use (13) to calculate the corresponding values of P_s. These analytical results are then compared with those obtained from simulation experiments. We used an event-driven simulator written in the C language. The simulation model implements the details of the contention access period of the MAC and simulates the timeline of the MAC operation in which two periods of UL and DL are considered in a TDD setting. The UL period begins with the contention slots. All the details of the exponential back off and timeout are implemented in the model. The simulation model also includes a process that simulates BW assignment in response to requests by stations, as well as the actual packet transmissions. For each station, BW requests are generated with exponentially distributed interarrival times. We assume an error-free PHY layer; thus, the only source of failed transmissions is collision. Using this simulator, we have conducted several experiments with different sets of parameters (with 95% confidence interval) and observed that in most cases, the mathematical model is very accurate and is reasonably close for the rest.

As we explained earlier, the expected value of the request generation interval (N_D) is adjustable and is not fixed by the standard. Since the efficiency of the system is dependent on it, in the first set of experiments, we plot its effect on the value of P_s for a wide range of request generation intervals and for four sets of values of the number of stations. As Fig. 6 shows, P_s achieves its maximum value in the persistent BW request generation case (i.e., saturation, when $N_D \to 0$). As a result, in the following analysis, we mostly concentrate on the saturation mode. However, as a final remark we present and compare two figures for the capacity (C) of the contention period for both saturation and nonsaturation cases to demonstrate the validity of the above consideration.

To have insight into the relationship between different contention period states of success, idle, and collision, we solved the analytical model for a typical parameter set ($W = 32$, $m = 5$, $N = 20$, $M = 6$, and $q = 0.7$) and compared the results with the simulation outcome (for different numbers of

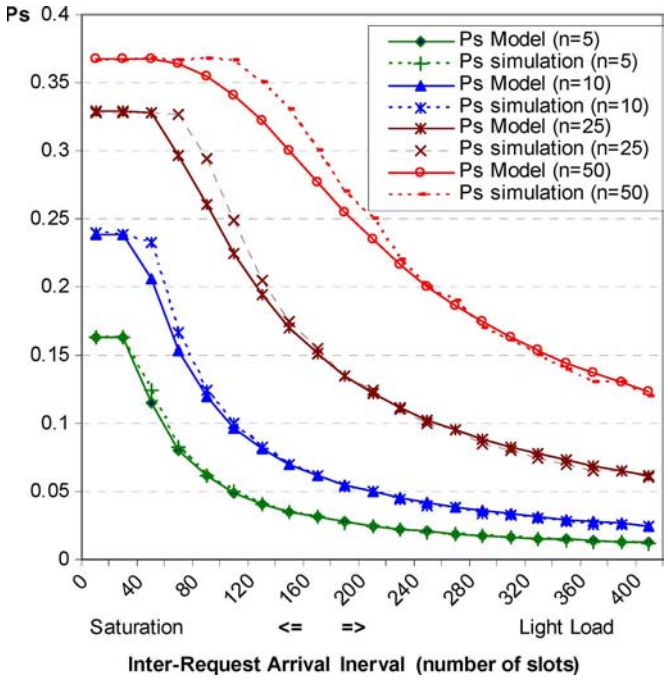

Fig. 6. P_s (efficiency) versus request generation rate for the number of stations ranging from 5 to 50.

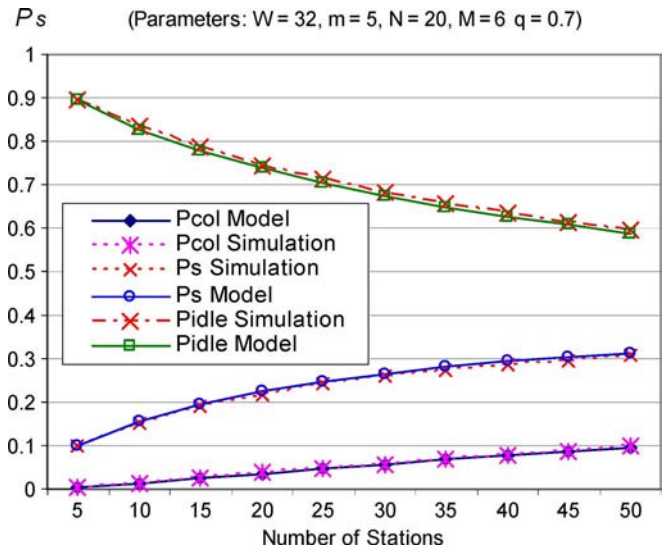

Fig. 7. Model evaluation. P_s versus n for a typical parameter set.

stations with persistent request generation). The results are depicted in Fig. 7, which shows that the model is, in fact, accurately representing the operation of the contention-based BW request period. Fig. 7 depicts three sets of probabilities: 1) the probability that a given slot in the contention period contains a successful transmission (P_s); 2) the probability that a given slot in the contention period contains a collided or corrupted transmission (P_{col}); and 3) the probability that a given slot in the contention period is idle (P_{idle}). As shown, for the given parameters, most of the slots are idle, and the system efficiency is 10% to 30%, depending on the number of contending stations. This suggests that the parameters may be selected in a way that maximizes the efficiency of the system. Note that when the contention period is more efficient, fewer

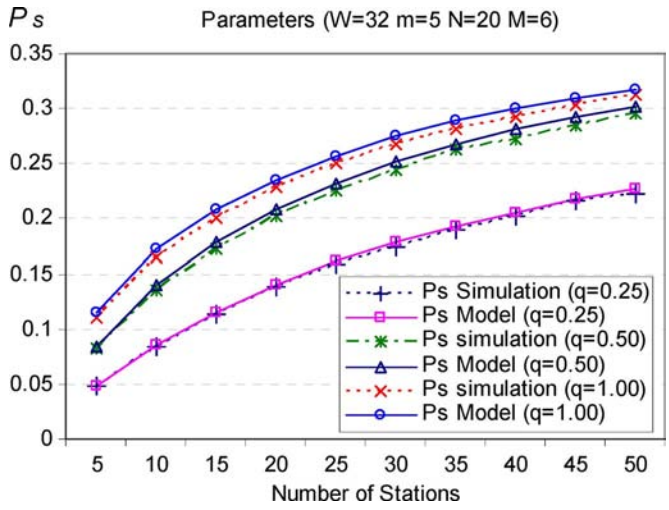

Fig. 8. P_s versus n for different values of q.

contention slots are needed to deliver the same number of BW requests; thus, the entire system becomes more efficient.

To see the effect of different parameters on the efficiency of the system and to further verify the validity of the mathematical model developed in Section III, we repeat the experiments for several different sets of parameters.

To see the effect of the probability of BW being granted by the BS (q) on the efficiency of the system (P_s), we plotted the simulation and numerical results for a typical parameter set of $(W = 32, m = 5, N = 20,$ and $M = 6)$ and three different values of q. The results are depicted in Fig. 8. It is seen that the system efficiency increases at a slower rate as q increases. The peak of efficiency is reached at $q = 1$, as expected. Since the difference between $q = 0.5$ and $q = 1$ is not very large, we conclude that the system efficiency is not significantly affected by q as long as more than 50% of the requests are responded to. This effect becomes more significant, as expected, when q is smaller than 0.5.

One of the parameters that we expect to have a significant effect on the performance of the contention period is the length of the timeout duration, which is partly controlled by parameter M. To see the effect of this parameter, we conducted some experiments with several values of M and plotted the results in Fig. 9. It is seen that the efficiency of the system, in the given conditions, increases when we reduce the value of M. The reason is that most of the capacity of the contention period is wasted by the idle slots (also seen in Fig. 7), partly due to long timeout periods (large values of M). Therefore, it can be concluded that for networks with typical parameters, it is more efficient to use small values of M. However, there are other parameters that also have a significant effect on generating idle slots, such as the initial contention window size (W).

To see the effect of W and to find out whether an optimal value for W exists, we plotted the contention period efficiency in Fig. 10 for two different scenarios: 1) a very crowded network $(n = 100)$ and 2) a modestly crowded network $(n = 25)$. We used $M = 3$ in this experiment to reflect the above finding that smaller values of M yield better efficiency. The results in Fig. 10 show that for $n = 25$, the efficiency is highest

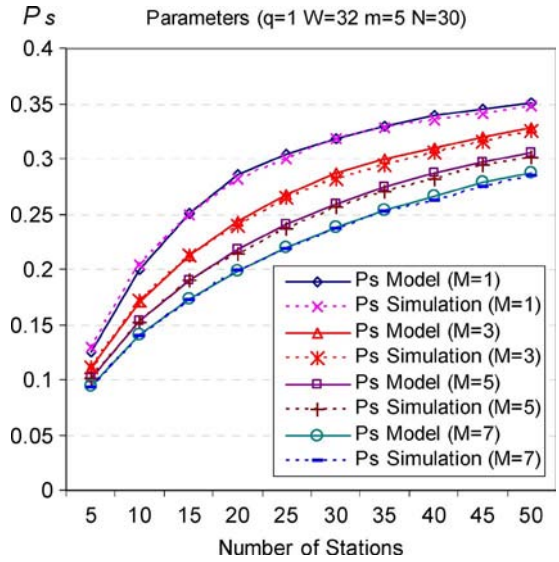

Fig. 9. P_s versus n for different values of M.

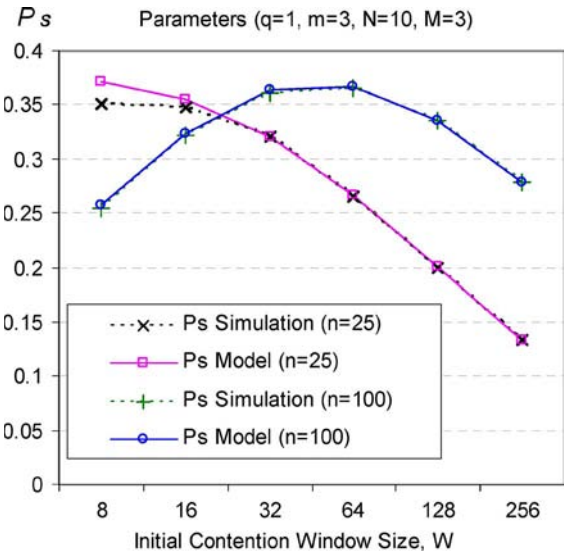

Fig. 10. P_s versus W for different values of n.

when W is small, whereas for $n = 100$, the efficiency peaks at $32 \leq W \leq 64$. This suggests that the value of W should be dynamically adjusted to maximize the efficiency of the system. It is also seen in Fig. 10 that the efficiency peaks at around 0.37.

Another interesting observation in Fig. 10 is that for small values of n and W (e.g., $W = 8$), the analytical model is less accurate. In fact, there are two assumptions important to the accuracy of the model that are affected by small values of n and W. First, we assumed that a transmission in each slot may result in a collision with a constant and independent probability p, regardless of the number of retransmissions that occurred. This assumption becomes less accurate when n and W are small, since a previous retransmission and collision becomes more likely to have come from the same stations that are contending in the given slot. Therefore, when the number of stations or the contention window size is large, the probability of collision can be more accurately assumed to be constant [13].

The other assumption affected by small values of n and W is the assumption that a transmission in each slot happens with equal probability, independent of the position of the slot in the contention period, and independent of the previous collisions. This assumption may not hold for those cases where n is small and W is smaller than N. The reason is that the back-off process always starts counting from the first slot in a contention period; thus, with $W < N$, the last slots in the contention period are more likely to remain empty. Only stations that have previously suffered collision and have increased their contention window sizes to above N will contend in these slots. This condition makes the independence assumption less accurate. Clearly, when $W > N$ or when the number of stations grows, this independence assumption becomes accurate. Fortunately, practical values of W are usually chosen to be greater than N to avoid wasting contention slots. In fact, the slots after $\overline{W}$ are almost certainly not involved in contention and are always empty. As a rule of thumb, the values of N and W should be selected so that $N < \overline{W}$ holds (see Fig. 5).

Finally, we examine the effect of the number of minislots in each contention period (N) on the efficiency of the system. The parameter N along with the initial contention window size (W) and request timeout duration (M) are important system configuration parameters. To see the effect of N, we plotted the efficiency and capacity of the contention period for a typical parameter set of $(q = 1,\ W = 32,\ m = 5,$ and $M = 6)$ and persistent request generation in Fig. 11. We chose the range of N so that it is less than $\overline{W}$ to avoid wasted slots. We observe that the efficiency is highest when N is small, e.g., set to five for the selected system configuration. However, this higher efficiency does not mean higher capacity of the contention period, since this capacity is proportional to the number of minislots N, according to (14). This is shown in the top plot of Fig. 11, where the corresponding plots for P_s are multiplied by N, which results in the plots for C (associated using the arrows on the figure). It is seen that the capacity of the contention period increases as the number of slots N increases (35 for this experiment). Note that N should be selected in a way that all the slots are used in contention, i.e., it must be less than the average contention window size $\overline{W}$. If larger values of N are used, $(N - \overline{W})$ slots will always be wasted and empty. It is also not desirable to arbitrarily increase N and W as this will considerably increase the overhead of the contention period.

In Fig. 11, we can see that, for example, for the case where 50 stations are persistently transmitting requests, if we use $N = 35$, $C = 10$ requests per frame time are successfully responded to by the BS. As defined in Section III, for n persistent stations, it will take n/C frames on the average for the BS to respond to all the requests. In this case, stations have to wait 50/10 or 5 frames for BW grants. When N is set to five, this access delay will increase to approximately 50/2 or 25 frames. Reducing the access delay requires that N and W be considerably increased, for example, to 100 or more in this case. The maximum number of minislots per frame is equal to (Frame_Size/4*Symbol_Duration) [2], which, for a typical Frame_Size $= 1$ ms and Symbol_Duration $= 0.1$ μs, results in 2500 minislots. Therefore, allocating 100 or 200 minislots to support 50 stations is a significant overhead. It must be noted that in WiMAX, it may not be unusual that a single BS serves hundreds of SSs. This simple example shows that proper

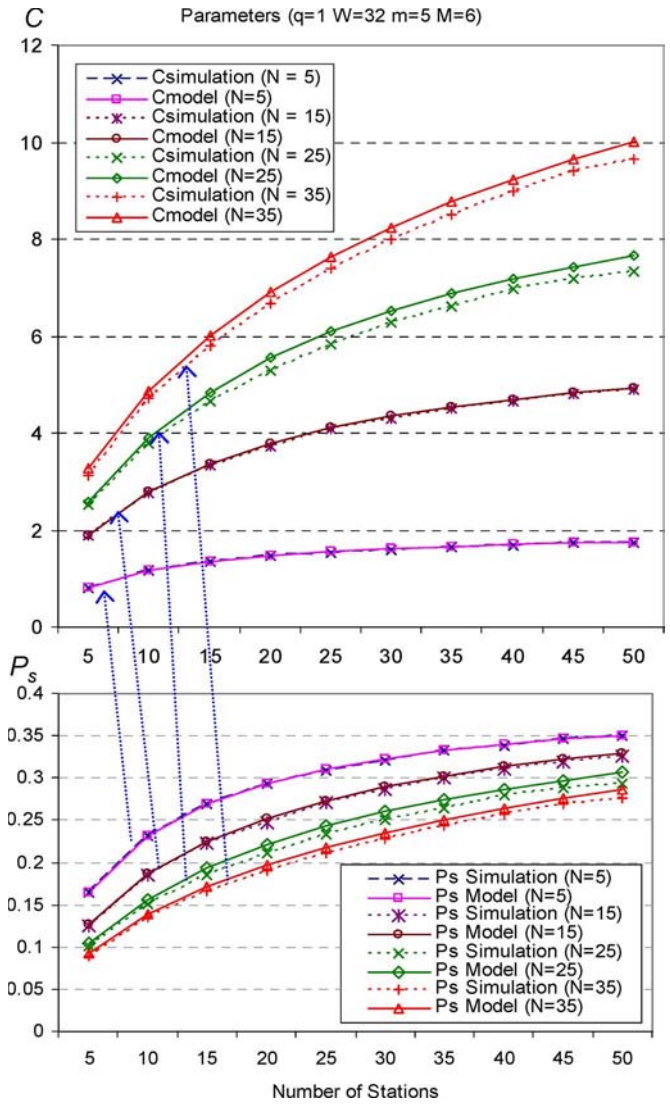

Fig. 11. P_s and capacity (C) versus n for different values of N (persistent request generation case).

setting of contention period parameters is required to improve the overall system performance and access delay of stations. Therefore, our model provides a useful tool for efficiently configuring the MAC without the need for time-consuming simulations or measurements.

We repeated the above experiment with nonpersistent BW request generation (i.e., $N_D = 300$ or $n \cdot (N/N_D)$ requests per frame) and plotted the results in Fig. 12. This demonstrates that the effect of N on the contention period capacity in nonsaturation cases is similar to that in the saturation case. As expected in this case, almost all BW requests are successfully delivered and responded to by the BS in one frame time; therefore, the plots increase in a semilinear manner, in proportion to the increase in the offered load or the number of stations (i.e., according to $n \cdot (N/N_D)$).

V. Concluding Remarks

In this paper, we have presented an accurate analytical model for the contention-based BW request of the MAC layer of the

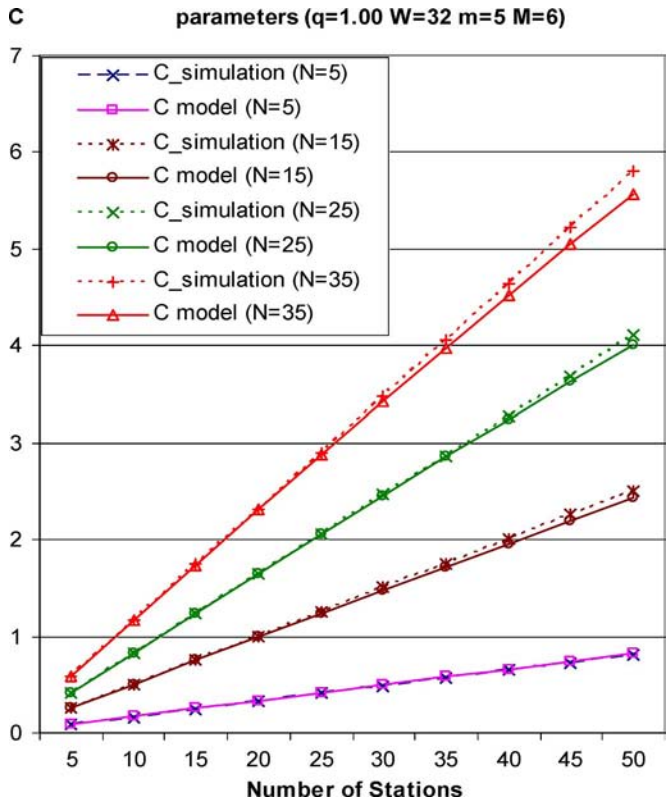

Fig. 12. Capacity (C) versus n for different values of N (infrequent request generation case).

IEEE 802.16 standard. To the best of our knowledge, this is the first detailed analysis of the contention BW request process of the 802.16 MAC. We verified the validity of the model through wide-ranging simulation experiments. The model proved to be very accurate in most cases. Slight inaccuracies were seen when the number of stations and the contention window size were very small. This behavior is also reported in other analyses of similar systems [13]. Nevertheless, such small contention window sizes are not typical in system configurations.

The analytical model developed in this paper can be used to devise optimization schemes that maximize the efficiency of the contention request process. Since the 802.16 standard allows the dynamic adjustment of the contention process parameters, our model can also be used to devise adaptive algorithms that maintain optimal operation of the system under varying network load.

References

[1] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, ANSI/IEEE Std. 802.16-2001, 2002.

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

[3] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, ANSI/IEEE Std. 802.16e-2005, 2005.

[4] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ANSI/IEEE Std. 802.11: 1999 (E) Part 11, ISO/IEC 8802-11, 1999.

[5] M. Hawa and D. W. Petr, "Quality of service scheduling in cable and broadband wireless access systems," in *Proc. 10th IEEE Int. Workshop Quality Service*, May 15–17, 2002, pp. 247–255.
[6] S. Kim and I. Yeom, "Performance analysis of best effort traffic in IEEE 802.16 networks," Dept. Comput. Sci., Korea Advanced Inst. Sci. Technol., 2006. Tech. Rep. CS-TR-2008-285.
[7] C. Cicconetti, C. Eklund, L. Lenzini, and E. Mingozzi, "Quality of service support in IEEE 802.16 networks," *IEEE Netw.*, vol. 20, no. 2, pp. 50–55, Mar./Apr. 2006.
[8] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance evaluation of the IEEE 802.16 MAC for QoS support," *IEEE Trans. Mobile Comput.*, vol. 6, no. 1, pp. 26–38, Jan. 2007.
[9] D.-H. Cho, J.-H. Song, M.-S. Kim, and K.-J. Han, "Performance analysis of the IEEE 802.16 wireless metropolitan area network," in *Proc. 1st Int. Conf. DFMA*, Feb. 2005, pp. 130–137.
[10] O. Gusak, N. Oliver, and K. Sohraby, "Performance evaluation of the 802.16 medium access control layer," in *Comput. and Inform. Sci.*, vol. 3280. Berlin, Germany: Springer-Verlag, 2004, pp. 228–237.
[11] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16," *Comput. Netw.*, vol. 49, no. 3, pp. 341–363, Oct. 2005.
[12] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: Current performance benchmarks and future potential," *IEEE Commun. Mag.*, vol. 43, no. 2, pp. 129–136, Feb. 2005.
[13] 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.
[14] Y. Pourmohammadi Fallah, "Per-session weighted fair scheduling for real time multimedia in multi-rate wireless local area networks," Ph.D. dissertation, Univ. British Columbia, Vancouver, BC, Canada, Mar. 2007.
[15] Y. Pourmohammadi Fallah and H. Alnuweiri, "Modeling and performance evaluation of frame bursting in IEEE 802.11 WLANs," in *Proc. Int. Conf. Commun. Mobile Comput.*, Jul. 2006, pp. 869–874.
[16] X. Yang, "IEEE 802.11e wireless LAN for quality of service," in *Proc. IEEE Wireless Commun. Netw. Conf.*, New Orleans, LA, Mar. 2003, pp. 1291–1296.
[17] G. R. Cantieni, Q. Ni, C. Barakat, and T. Turletti, "Performance analysis under finite load and improvements for multirate 802.11," *Comput. Commun.*, vol. 28, no. 10, pp. 1095–1109, Jun. 2005.
[18] O. Abu-Sharkh and A. H. Tewfik, "Performance analysis of multi-rate 802.11 WLANs under finite load and saturation conditions," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2005, vol. 4, pp. 2652–2657.
[19] J. V. Sudarev, L. B. White, and S. Perreau, "Performance analysis of 802.11 CSMA/CA for infrastructure networks under finite load conditions," in *Proc. 14th IEEE Workshop Local Metrop. Area Netw.*, Sep. 18–25, 2005, 6 pp.
[20] E. Ziouva and T. Antonakopoulos, "CSMA/CA performance under high traffic conditions: Throughput and delay analysis," *Comput. Commun.*, vol. 25, no. 3, pp. 313–321, Feb. 2002.
[21] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Yaser Pourmohammadi Fallah (M'07) received the B.Sc. degree in electrical engineering (electronics) from the Sharif University of Technology, Tehran, Iran, in 1998 and the Ph.D. and M.A.Sc. degrees in electrical and computer engineering from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2007 and 2001, respectively.

Prior to his Ph.D. studies, he was with IBM Canada. He is currently a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, UBC, working in the field of multimedia networking and wireless communications.

Dr. Pourmohammadi Fallah has been a member of the Standards Council of Canada Committee on MPEG Development (ISO/IEC JTC1/SC29) since 2001. He is the recipient of numerous awards and scholarships, including a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship, a Bell Canada graduate scholarship, a UBC Theodore E. Arnold fellowship, and a UBC graduate fellowship.

Farshid Agharebparast (M'07) received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 1995 and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2001 and 2007, respectively. His graduate studies were supported by numerous awards and scholarships, including a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship, a UBC graduate fellowship, and a Canadian Wireless Telecommunications Association graduate scholarship.

He is currently with the Department of Electrical and Computer Engineering, UBC. His current research interests include performance modeling and QoS provisioning techniques in wireless data telecommunication and computer networks.

Mahmood R. Minhas received the M.Sc. degree in computer science from the International Islamic University, Islamabad, Pakistan, in 1999 and the M.Sc. degree in computer engineering from the King Fahd University of Petroleum and Minerals, Dharan, Saudi Arabia, in 2001. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of British Columbia (UBC), Vancouver, Canada. His graduate studies are supported by numerous awards and scholarships, including the UBC Graduate Fellowship and the UBC Graduate Entrance Scholarship.

His current research interests include the design of energy-aware routing algorithms and protocols for wireless sensor networks.

Hussein M. Alnuweiri (S'83–M'89) received the Bachelor's and Master's degrees in electrical engineering from the University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1983 and 1984, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Southern California, Los Angeles, in 1989.

From 1991 to 2007, he was a Professor with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada. He is currently a Professor with the Department of Electrical Engineering, Texas A&M University at Qatar, Doha, Qatar. His broad research interests cover several areas in computer engineering, networking, system design, and real-time multimedia communications. His areas of specific interest include advanced computing structures, application-specific parallel and multiprocessing architectures, configurable computing, high-speed digital design, parallel and distributed algorithms, wireless and sensor networks, scheduling theory, traffic engineering, and quality-of-service mechanisms in wired and wireless networks. He is the author or coauthor of more than 50 refereed journal papers and 100 conference papers. He is the holder of three U.S. patents and one international patent.

Dr. Alnuweiri was a Delegate of the Standards Council of Canada for the MPEG-4 (ISO/IEC JTC1/SC29/WG11) Committee in 2001.

Victor C. M. Leung (S'75–M'79–SM'97–F'03) received the B.A.Sc. (Hons.) and the Ph.D. degrees in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, in 1977 and 1981, respectively. His graduate studies were supported by a Natural Sciences and Engineering Research Council Postgraduate Scholarship.

From 1981 to 1987, he was a Senior Member of Technical Staff with MPR Teltech Ltd., Burnaby, BC, specializing in the planning, design, and analysis of satellite communication systems. In 1988, he was a Lecturer with the Department of Electronics, Chinese University of Hong Kong. He returned to UBC as a Faculty Member in 1989, where he is currently a Professor and the holder of the TELUS Mobility Research Chair in Advanced Telecommunications Engineering with the Department of Electrical and Computer Engineering and a member of the Institute for Computing, Information, and Cognitive Systems. His research interests are in the areas of architectural and protocol design and performance analysis for computer and telecommunication networks, with applications in satellite, mobile, personal communications, and high-speed networks. He serves on the editorial boards of the *International Journal of Sensor Networks* and the *International Journal of Communication Networks and Distributed Systems*.

Dr. Leung is a Voting Member of the Association for Computing Machinery. He serves on the editorial boards of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON COMPUTERS. He has served on the Technical Program Committee (TPC) of numerous conferences. He was the TPC Vice Chair of the 2005 IEEE Wireless Communications and Networking Conference, a General Cochair of 2005 ACM International Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, and the General Chair of QShine 2007. He received the Association of Professional Engineers of British Columbia Gold Medal as the Head of the graduating class in the Faculty of Applied Science.