

## Research Article

# Performance Optimization for Delay-Tolerant and Contention-Based Application in IEEE 802.16 Networks

Fei Yin and Guy Pujolle

*LIP6, Pierre et Marie Curie University, 104 Avenue du President Kennedy, 75016 Paris, France*

Correspondence should be addressed to Fei Yin, fei.yin@lip6.fr

Received 25 January 2008; Accepted 5 June 2008

Recommended by Jong Hyuk Park

IEEE 802.16 standard suite defines the air interface specifications for fixed and mobile broadband access in wireless metropolitan area networks. Although the IEEE 802.16 MAC has been well defined by various bandwidth allocation and scheduling mechanisms to support QoS for different applications, efficient bandwidth allocation still remains as an open issue. We analyze and develop a mathematical model to evaluate the performance of the contention-based and delay-tolerant applications in IEEE 802.16 networks. We focus our attentions on allocating the uplink bandwidth efficiently, the basic goal is to optimize the performance with an optimal bandwidth allocation mechanism. The results of our analysis lay out clearly that a maximum uplink throughput and a minimum number of pending bandwidth request transmission can always be acquired by optimizing the contention period size in a frame. This optimal size is also influenced by the number of terminals in the network, which is also analyzed in the later part of the paper. Our results can be used for providing probabilistic throughput guarantee and determining the optimal contention period.

Copyright © 2008 F. Yin and G. Pujolle. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1. INTRODUCTION

Broadband wireless access (BWA) has gained a particular attention during the past few years. The widely successful IEEE 802.11 wireless LAN (WLAN) technologies are suitable for an indoor BWA solution but are not well suited for outdoor BWA applications. In response to this need, the IEEE 802.16 is set up to develop a new standard for BWA applications. IEEE 802.16 is an emerging suite of air interface standards combining fixed, portable, and mobile BWA specifications. The first IEEE 802.16 standard, 802.16-2001, is the original fixed wireless broadband air-interface specification in the 10–66 GHz frequency band for line-of-sight (LOS) only wireless services. The 802.16a was completed in 2003 to extend the standard in the 2–11 GHz for non-line-of-sight (NLOS) wireless broadband services. The final revision of fixed BWA standard, IEEE 802.16-2004 [1], which appeared in 2004, defines the air interface and medium access control (MAC) protocol for a current fixed wireless metropolitan area network, intended for providing high bandwidth wireless voice, video, and data for residential

and enterprise in licensed and license-exempt frequencies bands for both line-of-sight and non-line-of-sight. IEEE 802.16e [2] amendment, appeared in 2005, extends the 802.16 to support not just fixed, but also portable and mobile operation.

In IEEE 802.16 system, two kinds of stations (fixed or mobile) are defined: base station (BS) and subscriber station (SS). The BS coordinates all the communication in the network. The SS can deliver voice, video, and data using common interface. IEEE 802.16 standards support two operational modes: a mandatory point-to-multipoint (PMP) mode, and an optional mesh mode. In a PMP topology network, a centralized BS is capable of connecting multiple SSs to various public networks linked to the BS, the traffics can only occur between the BS and SSs. In the mesh mode, the SSs can also serve as routers by cooperative access control in a distributed manner. The communication between BS and SSs has two directions: uplink (from SSs to BS) and downlink (from BS to SSs). The downlink transmission is on a broadcast basis from the BS to all SSs, while the uplink bandwidth is shared by SSs on a demand basis. Both uplink

and downlink can operate in different frequencies using frequency division duplexing (FDD) or at different time using time division duplexing (TDD). Figure 1 illustrates an example of general architecture of IEEE 802.16 networks. The fixed or mobile customer premise equipments (CPEs) connect to the central BS, the BS receives transmissions from multiple sites and sends to internet directly or via other BSs. End users (laptop, telephone, computer, . . . , etc.) inside the building, through inbuilding networks such as Ethernet or WLAN, can connect to an outside CPE and then link to the IEEE 802.16 network.

Resource management and allocation mechanisms are crucial to guarantee quality-of-service (QoS) performance in IEEE 802.16 networks. A polling-request-grant mechanism is defined in IEEE 802.16 MAC for efficient bandwidth allocation in uplink channel from multiple SSs to a central BS. In a PMP network, the SS first has to utilize an allocated polling interval to request uplink bandwidth before transmitting data in a corresponding bandwidth grant. This means that if an SS wants to do uplink transmission, it first sends a request to BS during the polling interval. On receiving the request from an SS, the BS should determine and grant to the SS the bandwidth, which is used by the SS to transmit the data. The IEEE 802.16 defines two main methods for SSs to send their bandwidth request messages: unicast polling, and contention-based polling including multicast or broadcast polling. In the first case each SS station is polled individually by the BS to send the request; in the latter all SSs contend to obtain transmission opportunities for sending requests using contention resolution mechanisms.

The IEEE 802.16 MAC is designed to be capable of accommodating a variety of traffics, including data, voice, and video. Then four scheduling service classes are defined to support different QoS requirements for kinds of applications: unsolicited grant service (UGS), real-time polling service (rtPS), non-real-time polling service (nrtPS), and best effort (BE). The IEEE 802.16 physical layer (PHY) supports time division multiple access (TDMA) for uplink channel access and each uplink channel is divided into a number of time minislots. These minislots are allocated in a MAP message to the SSs for the different propositions.

Even rounded bandwidth allocation and scheduling mechanisms are defined in the IEEE 802.16 standard, the efficiency of the mechanisms are still left to deliberate. A scheme that can efficiently allocate bandwidth to guarantee the QoS performance is essential in IEEE 802.16 networks. In this paper, we focus on evaluating IEEE 802.16 performance with efficient bandwidth allocation mechanisms for nrtPS and BE services. We want to find out an optimal bandwidth allocation to optimize the performance. Besides, the performance influenced by the network size is also investigated in our analysis.

The rest of the paper is organized as follows. In Section 2 we give some general insights on the MAC operation of IEEE 802.16. Based on the overview, in Section 3, we discuss the problem of maximizing the uplink data throughput by setting an optimal contention period in a frame. We also present the related works in performance optimization in IEEE 802.16 in Section 4. In Section 5 we present the system model

while Section 6 addresses performance analysis through our mathematic model. We also address the simulation results in Section 7 and present the concluding remarks in Section 8.

## 2. OVERVIEW OF IEEE 802.16 MAC

In this section, we give a brief overview of some technical aspects of IEEE 802.16 MAC protocols: the frame structure, the bandwidth allocation process, and the uplink service classes. The problem statement and the analysis in the later sections largely depend on these basic operations of the MAC protocol.

### 2.1. Frame structure

The frame in IEEE 802.16 standard is modeled as a stream of minislots, which help to partition the bandwidth easily, and is divided into two subframes: downlink subframe and uplink subframe. According to different duplexing techniques, the downlink and uplink transmission occur in FDD mode or TDD mode. Figure 2 shows the overall frame structure of the IEEE 802.16 MAC with TDD.

A TDD frame has a fixed duration and contains one downlink and one uplink subframe. The downlink subframe is generally broadcast and starts with preamble, downlink MAP (DL-MAP), and UL-MAP (Uplink MAP). The preamble is used by the PHY for synchronization and equalization. The DL-MAP and UL-MAP contain the correlative information of the intervals' usage in the following downlink and uplink subframes, respectively, and are broadcast to all SSs. The following downlink bursts carry the data to transmit to SSs, and a transmit/receive time gap (TTG) in the end of the bursts to separate the downlink subframe from the uplink subframe.

In a UL-MAP, the BS may specify some uplink intervals as the opportunities for new SSs to join the channel by requesting the basic management connection identifiers (CIDs), by adjusting its power level and frequency offsets and by correcting its time offset; other intervals as the request information elements (IEs) for authorized SSs to competitively request uplink bandwidth; and another intervals as the uplink bandwidth grants in which particular SSs transmit data or uniquely request bandwidth. Correspondingly, the uplink subframe is divided into chunks of minislots for the purpose of initial ranging, bandwidth request, and data transmission. Specifically, the request IEs allocated for contention request are composed by transmission opportunities (TOs). A TO is defined as an allocation provided in a UL-MAP or part thereof intended for a group of SSs authorized to transmit bandwidth requests [1]. This group may include either all SSs or a multicast polling group of SSs having bandwidth request for a transmission. The number of TOs associated with a particular IE in a MAP depends on the total size of the allocation as well as the size of an individual transmission. The BS will always allocate bandwidth for contention IEs in integer multiple TOs.

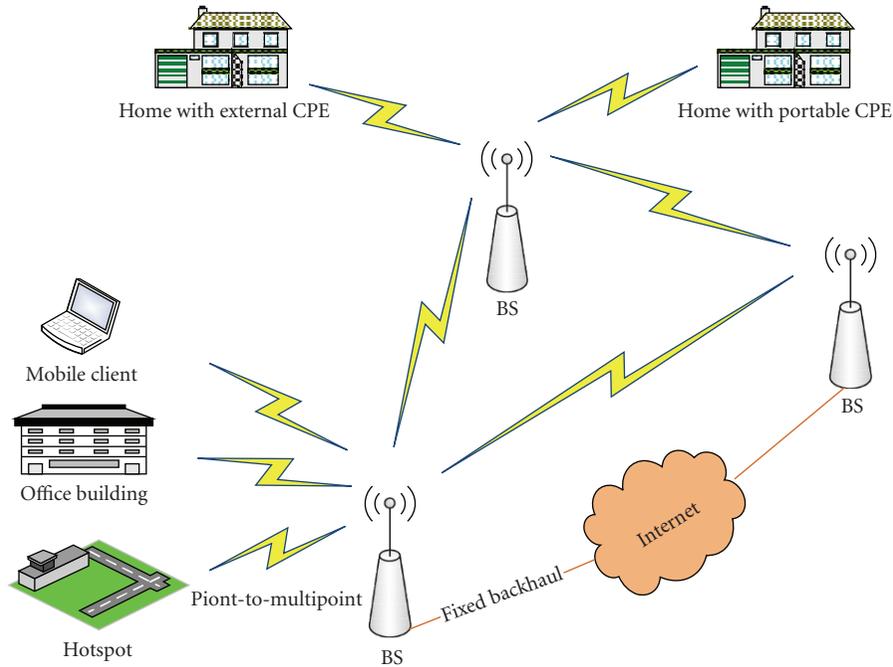


FIGURE 1: IEEE 802.16 network architecture.

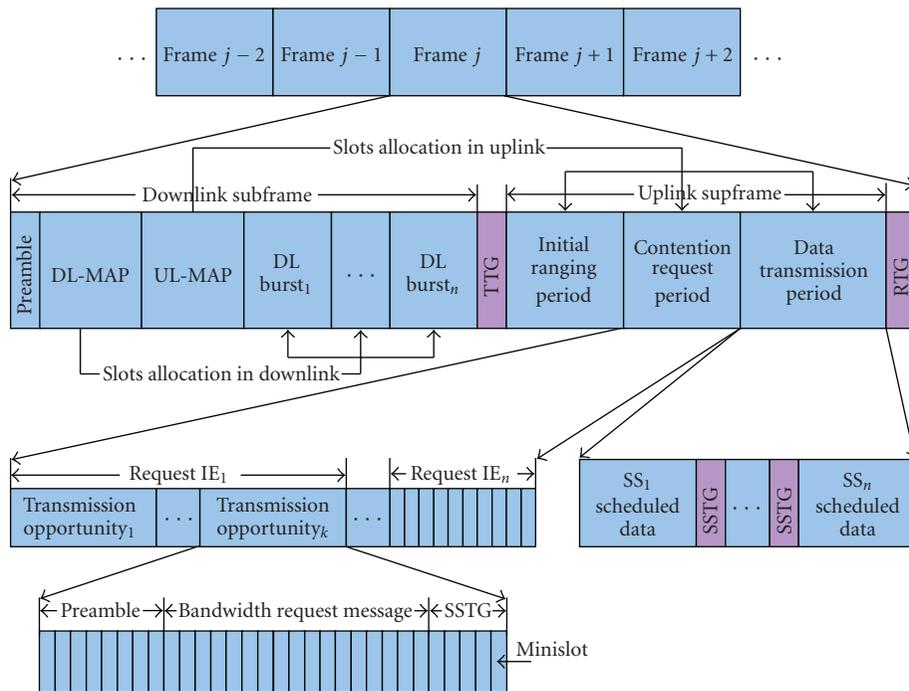


FIGURE 2: Frame structure of IEEE 802.16 MAC.

**2.2. Polling-request-grant bandwidth allocation procedure**

In a PMP network, the BS controls all transmissions in the uplink and the downlink. For uplink access, a Polling-Request-Grant mechanism is defined for bandwidth alloca-

tion during a certain duration. Figure 3 shows the overall procedure.

Initially, the SSs who have new connections need to get an admission into the network from the BS through an admission control mechanism which is vendor-dependent and does not specify in this paper. According to the QoS

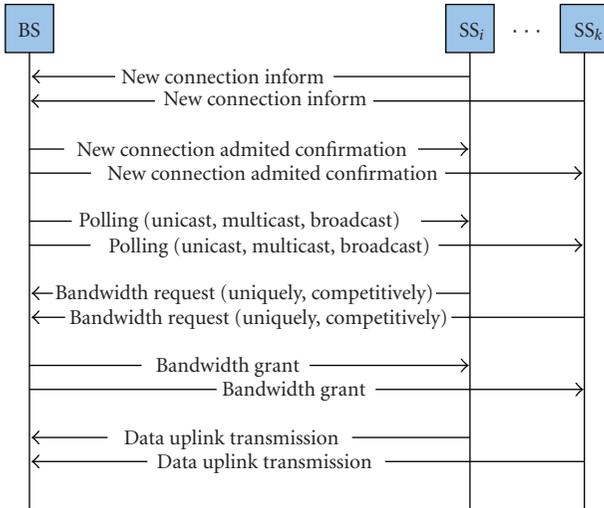


FIGURE 3: Polling-request-grant process.

parameters of connections in SSs, the BS has to poll the admitted SSs by allocating bandwidth specifically for the purpose of making bandwidth requests. Depending on the connections' service classes and the residual bandwidth in the BS side, these polling intervals may address to individual SSs (unicast polling) or to groups of SSs (multicast/broadcast polling).

The SSs then utilize these request IEs to uniquely or competitively request uplink bandwidth for each connection to make a reservation with the BS. On receiving the requests, the BS allocates chunks of minislots in the coming MAP as the bandwidth grants to the SSs, which should take into account the requirements from all authorized SSs and the available bandwidth in the uplink subframe. The bandwidth grant is aggregated into a single grant to the SS and not to the on-requesting connections. Typically, the SS decodes the received UL-MAP and determines the honored connections to transmit data. This bandwidth allocation technique in the IEEE 802.16 standard is called Polling-Request-Grant procedure.

The SS will assume that the transmission has been unsuccessful if no bandwidth grant has been received in a specific timeout,  $T_{16}$  [1], or if a shorter one than expected is received. Then a contention resolution process is started and retransmission of unsuccessful bandwidth request will be implemented by the SS.

### 2.3. Scheduling service classes

Scheduling services represent the data-handling mechanisms supported by the MAC scheduler for data transport on a given connection. The IEEE 802.16 MAC provides QoS differentiation for the different types of applications that operate over 802.16 networks, through four defined scheduling service types. This classification facilitates bandwidth sharing between different users as follows.

- (i) The UGS is designed to support real-time service flows that generate fixed-size data packets on a periodic basis, such as T1/E1 and Voice over IP (VoIP) without silence suppression. In this service, the BS offers fixed-size data grants on a real-time periodic basis, which eliminate the overhead and latency of SS requests and assure that grants are available to meet the flow's real-time needs. The mandatory QoS service parameters are maximum sustained traffic rate, maximum latency, tolerated jitter, and request/transmission policy [1].
- (ii) The rtPS is designed to support real-time service flows that generate variable-size data packets that are issued at a periodic intervals, such as Moving Pictures Experts Group (MPEG) video. The BS provides periodic unicast polling opportunities, which meet the flow's real-time needs and allow the SS to specify the size of the desired grant. This service requires more request overhead than UGS, but supports variable grant sizes for optimum data transport efficiency. The mandatory QoS service parameters are maximum reserved traffic rate, maximum sustained traffic rate, maximum latency and request/transmission policy [1].
- (iii) The nrtPS is designed to support delay-tolerant service flows that generate variable-size data packets for which a minimum data rate is requested. The BS offers unicast polling opportunities on a regular basis, which assures that the service flow receives request opportunities even during network congestion. In addition, the SS is also allowed to use contention request opportunities. The mandatory QoS service parameters are maximum reserved traffic rate, maximum sustained traffic rate, traffic priority, and request/transmission policy [1].
- (iv) The BE is designed to support data streams for which no minimum service level is required and therefore may be handled on a space-available basis. The SS may use contention request opportunities as well as unicast request opportunities when the BS sends any. The mandatory QoS service parameters are maximum sustained traffic rate, traffic priority, and request/transmission policy [1].

As shown in Figure 4, among these four service classes, UGS is prohibited from any polling, rtPS connections can only use unicast polling intervals to transmit bandwidth requests, nrtPS connections may adopt a mandatory unicast polling and an optional contention-based polling, while BE connections adopt a mandatory contention-based polling and do not have any unicast polling obligation. Specially, in nrtPS, the BS first has to poll the SSs by unicast polling, and then switch to contention-based polling only when no sufficient residual bandwidth to support unicast polling. Then we refine these four service classes into two major types: the UGS and the rtPS are delay-sensitive and contention-free services; the nrtPS and the BE are delay-tolerant and

contention-based services. In this paper, we only consider the delay-tolerant and contention-based applications.

### 3. PROBLEM STATEMENT

In wireless network, because of the limited bandwidth and the expensive radio spectrum, the demand of performance optimization became more and more critical. In most situations, the optimization target is to get a higher throughput and/or lower delay system. For the delay-tolerant and contention-based applications in IEEE 802.16, the delay is not a key QoS parameter, then the uplink throughput across a set of fixed or mobile SSs stands out as the most important performance figure.

According to the Polling-Request-Grant mechanism, in a fix-size uplink subframe, we know that uplink throughput is affected by the number of bandwidth grants allocated to SSs, which is controlled by the size of contention request period and the available bandwidth in data transmission period. It is well known that the different allocation of minislots exhibit very different efficiency. Based on Figure 2, we know that the contention request period and data transmission period are interactional.

- (i) If the size of contention request period is very small, there are few TOs that can be utilized to transmit the bandwidth requests. Many requests may be queued in the buffer, and might be dropped depending on the implementation policy of the request queue. In this case, only a small quantity of bandwidth requests are successfully received by the BS. On the other hand, though the data transmission period is very large, the BS can only allocate few bandwidth grants to SSs in one frame based on the received requests. There are some minislots in data transmission period are idle and make a waste of bandwidth. The result of this allocation leads to a very low uplink throughput.
- (ii) If the size of contention request period is very large, numerous bandwidth requests are successfully received by the BS. However, the available bandwidth in the data transmission period is very small and is deficient to fit all the bandwidth requests. Then, only a few bandwidth requests could be granted. And then little bandwidth might be allocated to SSs to do the uplink transmission in one frame, which results in a low uplink throughput.

We must be absorbed in the efficiency of the bandwidth allocation to get a performance optimization in IEEE 802.16 networks. To do it, an efficient combination of polling, request, and grant mechanisms to optimize the contention request period size is necessary, which may achieve a tradeoff between the number of bandwidth requests and the number of bandwidth grants and then maximize the uplink throughput. Furthermore, the optimal size of contention request period is dependent on the network size, since the tradeoff varies with the number of SSs in the network.

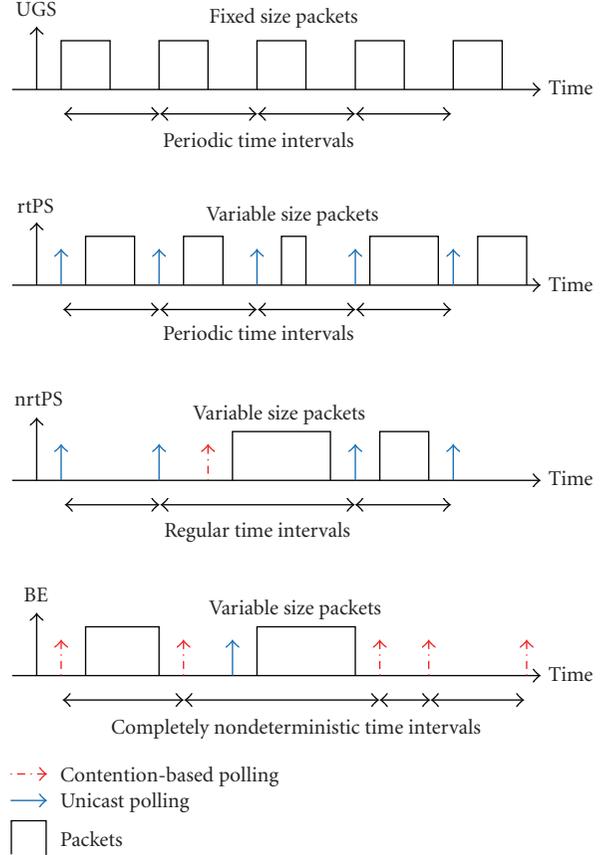


FIGURE 4: Uplink scheduling service classes.

### 4. RELATED WORKS

In recent years, many related works have studied the performance optimization for wireless networks. Benelli et al. [3] analyzed the optimal frame size according to the number of users and the number of collided packets in a slotted aloha multiple access radio mobile network. Bianchi [4] contributed in a simple but extremely accurate analytical model to compute the maximum and saturation throughput in IEEE 802.11 distributed coordination function (DCF) networks. His conclusion shows that maximum performance can be achieved by adaptively tuning the value of backoff window size depending on the network size. This is also proved in Bianchi et al. [5].

As for the IEEE 802.16 networks, by now few scientific results have been obtained to optimize the performance by efficient bandwidth allocation mechanisms. Chu et al. [6] proposed an efficient QoS architecture to provide QoS guarantees for IEEE 802.16 system. An idea of adopting a contention slot allocator (CSA) to dynamically adjust the ratio of the contention request period and the data transmission period is presented in the article. Their study analysed, in theory, that the CSA had significant impact on the system performance and there should be a tradeoff in the design of the CSA. They also suggested that an algorithm to fully utilize the bandwidth needs to be developed. But the

authors did not give out any algorithms of how CSA works and how to calculate the ratio.

Cho et al. [7] also proposed a new QoS architecture of IEEE 802.16, in which an uplink bandwidth allocation scheduling mechanism is adopted. Their work focuses on the request throughput optimization mostly, whose objective is to maximize the number of bandwidth requests successfully transmitted in the contention request period. In order to obtain a maximum number of requests, the authors take the backoff windows size into account and want to find an optimum value. After mathematic deductions, the authors concluded that the maximum request throughput could be achieved with the backoff windows size which is equal to the number of competing Ss. However, the conclusion is biased and not exactly right in IEEE 802.16 networks, in which maximizing the number of successful bandwidth requests could not always lead to an optimum uplink throughput. As discussed in Section 3, we know that, in IEEE 802.16 network, the uplink throughput depends on not only the number of bandwidth requests, but also the number of bandwidth grants. An optimal tradeoff between requests and grants should be found in order to get a maximum uplink throughput.

The research on optimizing the performance of nrtPS and BE applications running in IEEE 802.16 network is supported by Oh and Kim [8]. The main objective is to find out an optimal contention request period for the number of users in the system, in order to guarantee the throughput and delay. In their article, the authors first explained the relation between the contention request period and the delay, and stated that it is essential to find an optimal period. In order to stochastically analyze the performance, they redid the definition of the throughput and delay. They also analyzed that the throughput and delay are tradeoff of each other and it is difficult to find the optimal point, then a new parameter, cost function, is introduced by the author to evaluate the performance.

- (i) Throughput is the ratio of the number of successfully transmitted requests and the total number of transmitted requests, which in fact is the request throughput for newly generated requests.
- (ii) Delay is the time spent until a new bandwidth request successfully transmits, which in fact is the request delay for newly generated requests.
- (iii) Cost function is the ratio of throughput and delay which indicates that the optimal value could be obtained when the throughput is large and the delay is small.

After some mathematic deduction, the authors concluded that the optimal size of the contention period is achieved when the cost function reaches a maximum value. In their study, the value is  $2M - 1$  slots, where  $M$  is the number of Ss in network. However, there are some points need to be improved and be more accurate in their research. First, the objective of the authors is to find out an optimal contention request period to get an optimum tradeoff between throughput and delay, and then

to optimize the performance, but their analysis results in an optimal frame size. They did the analysis by assuming that the bandwidth requests can be uniformly distributed in a frame duration, which actually took the whole frame duration as the contention request period. Consequentially, the following mathematic deductions are based on the frame size but not a particular contention request period size. Deducing by the mathematic formulas in the article, we got the results that the  $2M - 1$  slots are the frame size. Second, the throughput and delay defined in the article are request throughput and request delay, which cannot accurately evaluate the performance in IEEE 802.16 network. Same as the problem in [7], the uplink data throughput and delay cannot be represented only by the number of successfully transmitted requests, but should involve the affection caused by the number of bandwidth grants into consideration. Third, the request throughput and delay should be composed of two parts: the throughput and delay caused by newly generated requests and by unsuccessful requests. The unsuccessful requests include the collision and non-granted requests produced in the former frame duration, which should be retransmitted in the current frame and highly influence the performance and then affects the optimal frame size. In this article, the author only focuses on the newly generated requests but ignores the others, which is not right in real situation. Fourth, in their work, the authors assumed that  $M$  Ss produced  $M$  bandwidth requests to transmit, which is far from being realistic because in practice the Ss will sporadically generate such packets of nrtPS and BE applications. It would be necessary to relax this assumption. Fifth, as shown in Section 2.3, in nrtPS and BE applications, the throughput is a mandatory QoS parameter but the delay is not, there is no necessary to introduce cost function as a key parameter to evaluate the performance.

In our paper, we concentrate on the performance optimization only for the delay-tolerant and contention-based applications runs in IEEE 802.16 networks, in the assumption of fixed and finite number of Ss. We analyze the uplink data throughput and the pending competitive bandwidth requests in uplink channel by efficiently combining request and grant allocation strategies together. In our analysis, we introduce a random process for bandwidth request generation in the network during a frame time horizon, and the bandwidth requests caused by newly generated and unsuccessful transmitted are both considered. We also provide a simple, nevertheless accurate, analytical model to compute an optimal contention request period, by which maximum uplink throughput and minimum pending competitive transmission are obtained. The influences of different network size on the optimal contention period are also investigated in our analysis.

## 5. SYSTEM MODEL

Let us consider a PMP system in which there are one BS and  $V$  Ss. An example of model of the uplink subframe structure with a realization of request arrivals occurring over a frame  $j$  is presented in Figure 5. The interarrival time of requests for

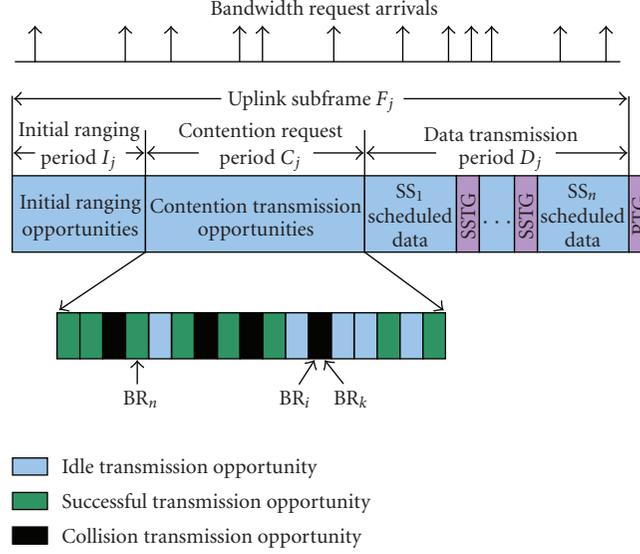


FIGURE 5: Analysis system model.

uplink bandwidth reservation is assumed to follow a general distribution with a positive and finite mean.

We denote  $F_j$ , in minislots, the size of the uplink subframe of frame  $j$ . The uplink subframe is divided into chunks of minislots as initial ranging period  $I_j$ , contention request period  $C_j$ , and data transmission period  $D_j$ :

$$F_j = I_j + C_j + D_j. \quad (1)$$

The contention minislots are all clustered adjacently at each uplink subframe. This allows easier implementation at both the BS and the SSSs because both devices have to switch to the contention mode only once at each frame period. In the system model, there are three possible TO allocations when there are multiple SSSs.

- (i) The collision TOs is a certain number of TOs on which more than two bandwidth requests are simultaneously transmitted, means that a transmission collision will occur on these TOs.
- (ii) The successful TOs are a part of TOs on which only one request is transmitted, means that the bandwidth request will be successfully transmitted.
- (iii) The idle TOs are some empty TOs on which all bandwidth requests refrain from transmitting.

The competing SSSs randomly select TOs to transmit bandwidth requests. When more than one SS start simultaneously bandwidth request transmission in the same TOs, a collision occurs. The collided bandwidth requests should be retransmitted. After receiving the successful bandwidth requests, the BS allocates chunks of minislots in the data transmission period as the bandwidth grants to the SSSs. If the available bandwidth in data transmission period is not sufficient to fit all the bandwidth requests received by the BS, some requests may not be granted in the coming frame, and are scheduled in BS side. These non-granted requests should

be retransmitted when the SSSs do not receive any grants in a predefined time. Finally, the SSSs transmit uplink data during the corresponding allocated intervals.

For ease of analysis, we assume the following.

- (i) The number of SSSs in system is fixed to  $V$  during operation period.
- (ii) The  $V$  SSSs is divided into  $M$  groups for the purpose of multicast polling during operation period.
- (iii) The uplink subframe size is fixed to  $F$  minislots during operation period.
- (iv) The bandwidth requests issued from the SSSs in the network are Poisson distributed during a frame time horizon.
- (v) The uplink bandwidth requests are fixed to  $R$  minislots (occupying one TO) and are uniformly distributed during contention request period in a frame.
- (vi) The collided and non-granted bandwidth requests are retransmitted in the next frame if SSSs do not receive any grants in the coming frame.
- (vii) All bandwidth requests successfully transmitted during contention period should be received by the BS.

## 6. PERFORMANCE ANALYSIS

The core contribution of this paper is the analytical evaluation of the optimal contention period to optimize the performance. The analysis is divided into two distinct parts. First, we analyse the uplink throughput and the pending bandwidth requests, and obtain the optimal contention period  $C_{opt}$ . Then, we study the influence to the  $C_{opt}$  which is depending on the network size.

Since we want to analyse the performance for delay-tolerant and contention-based application, then we take

contention-based polling mechanism into account, where the SSs should be multicast or broadcast polled. In generalization, we divide the  $V$  SSs into  $M$  multicast groups, then each group includes  $\lceil V/M \rceil$  SSs. In multicast polling, the BS averagely assigns TOs to each group and the probability that each group attains the TOs is  $1/M$ . Let  $N_j$  be the number of TOs the BS assigns to each group in frame  $j$ , then we have

$$N_j = \left\lceil \frac{C_j}{R * M} \right\rceil. \quad (2)$$

Let  $t_j$  be the number of pending bandwidth requests transmits in the contention period during frame  $j$  in a multicast group. Since the bandwidth requests are uniformly distributed in contention request period, its distribution will converge to a binomial distribution for  $N_j$ . Based on the definition in [9], the probability that  $r$  in  $t_j$  bandwidth requests transmit in one TO during frame  $j$  is

$$p(r) = \binom{t_j}{r} p^r (1-p)^{t_j-r}, \quad (3)$$

where  $p$  is the probability that a request is assigned to a particular TO. Since the distribution is uniform, we find

$$p = \frac{1}{N_j} \quad (4)$$

substitute the  $p$  in (3), and we get

$$p(r) = \binom{t_j}{r} \left(\frac{1}{N_j}\right)^r \left(1 - \frac{1}{N_j}\right)^{t_j-r}. \quad (5)$$

The expected number of contention TOs in which  $r$  bandwidth requests transmit are

$$\begin{aligned} E(r) &= N_j * p(r) \\ &= N_j \binom{t_j}{r} \left(\frac{1}{N_j}\right)^r \left(1 - \frac{1}{N_j}\right)^{t_j-r}. \end{aligned} \quad (6)$$

We can thus identify three contributions in TO allocation.

- (i) The successful TOs in which bandwidth requests are said to be successfully transmitted, where  $r = 1$ . Then, during frame  $j$ , in a multicast group, the number of TOs in which bandwidth requests are successfully transmitted is

$$\begin{aligned} S_j &= N_j * p(1) \\ &= N_j \binom{t_j}{1} \left(\frac{1}{N_j}\right)^1 \left(1 - \frac{1}{N_j}\right)^{t_j-1} \\ &= t_j \left(1 - \frac{1}{N_j}\right)^{t_j-1}. \end{aligned} \quad (7)$$

- (ii) The collision TOs in which requests are said to be in collision, where  $r \geq 2$ . The number of TOs in which

transmission collision generated is

$$\begin{aligned} B_j &= N_j \sum_{r=2}^{t_j} p(r) \\ &= N_j \sum_{r=2}^{t_j} \binom{t_j}{r} \left(\frac{1}{N_j}\right)^r \left(1 - \frac{1}{N_j}\right)^{t_j-r}. \end{aligned} \quad (8)$$

- (iii) The idle TOs is that in which bandwidth requests refrain from transmitting, where  $r = 0$ . The number of TOs in which no bandwidth request transmitted is

$$\begin{aligned} H_j &= N_j * p(0) \\ &= N_j \binom{t_j}{0} \left(\frac{1}{N_j}\right)^0 \left(1 - \frac{1}{N_j}\right)^{t_j} \\ &= N_j \left(1 - \frac{1}{N_j}\right)^{t_j}. \end{aligned} \quad (9)$$

Since one successful TO only holds one bandwidth request, then we can get the number of successfully transmitted bandwidth request equal to  $S_j$ .

One collision TO contains simultaneously more than two bandwidth requests. To obtain the number of collided bandwidth requests in the frame  $j$ ,  $x_j$ , we multiply (8) by the number of bandwidth request in a particular TO in which collision occurs. Hence,

$$\begin{aligned} x_j &= \sum_{r=2}^{t_j} r B_j \\ &= \sum_{r=2}^{t_j} r N_j \binom{t_j}{r} \left(\frac{1}{N_j}\right)^r \left(1 - \frac{1}{N_j}\right)^{t_j-r} \\ &= t_j - t_j \left(1 - \frac{1}{N_j}\right)^{t_j-1}. \end{aligned} \quad (10)$$

Since the requests newly generated in network are Poisson distributed during a frame duration, then we get the probability  $l$  requests issued by the SSs during a minislot time horizon:

$$p_{in}(l) = \frac{\lambda^l e^{-\lambda}}{l!}, \quad (11)$$

where  $\lambda$  is the average number of requests generated by the SSs per minislot duration in network. Consequently, the probabilities 0, 1 and more than 2 requests issued in a minislot duration are  $p_{in}(l = 0)$ ,  $p_{in}(l = 1)$  and  $p_{in}(l > 1)$ , respectively,

$$\begin{aligned} p_{in}(l = 0) &= e^{-\lambda} \\ p_{in}(l = 1) &= \lambda e^{-\lambda} \end{aligned} \quad (12)$$

$$p_{in}(l > 1) = 1 - e^{-\lambda} - \lambda e^{-\lambda}.$$

Let  $n$  be the number of bandwidth requests newly generated in network during a frame, then we get

$$\begin{aligned} n &= \sum_{l=0}^{\infty} l F p_{in}(l) \\ &= F \left( 0 p_{in}(0) + 1 p_{in}(1) + \sum_{l=2}^{\infty} l p_{in}(l) \right) \\ &= F \left( \lambda e^{-\lambda} + \sum_{l=2}^{\infty} l \frac{\lambda^l e^{-\lambda}}{l!} \right). \end{aligned} \quad (13)$$

Since,

$$\begin{aligned} \sum_{l=2}^{\infty} l \frac{\lambda^l e^{-\lambda}}{l!} &= \lambda \sum_{(l-1)=1}^{\infty} \frac{\lambda^{(l-1)} e^{-\lambda}}{(l-1)!} \\ &= \lambda \left( 1 - \frac{\lambda^0 e^{-\lambda}}{0!} \right) \\ &= \lambda (1 - e^{-\lambda}) \end{aligned} \quad (14)$$

and hence,

$$n = F(\lambda e^{-\lambda} + \lambda(1 - e^{-\lambda})) = F\lambda. \quad (15)$$

Then we get the number of pending bandwidth requests in a multicast group during frame  $j + 1$  is

$$t_{j+1} = F\lambda + x_j + k_j, \quad (16)$$

where  $k_j$  is the number of non-granted bandwidth requests in one group during frame  $j + 1$ .

The corresponding bandwidth grants to the  $S_j$  bandwidth requests in frame  $j$  will be allocated in frame  $j + 1$ . Let  $C_{j+1}$  and  $D_{j+1}$  be the contention request period and the data transmission period in the frame  $j + 1$ , respectively; let  $X_{j+1}$  be the number of bandwidth grants allocated to all Ss in  $D_{j+1}$ , and let  $Q_i$  be the size of uplink data packet corresponding to the  $i$ th grant (in minislot), then we get

$$X_{j+1} = \sum_{z=1}^M G_z, \quad (17)$$

where  $G_z$  is the number of grants in the  $z$ th multicast group.

We can thus identify three cases in performance evaluation related to the history of bandwidth requests  $S_j$  that are successfully transmitted in the previous frame and to the number of bandwidth grants  $G_{j+1}$  produced in the current frame.

### 6.1. $S_j > G_{j+1}$

In this case, let  $C_{j+1}^1$  and  $D_{j+1}^1$  be the contention request period and the data transmission period in frame  $j + 1$ , respectively. The  $D_{j+1}^1$  is deficient to grant all the  $M * S_j$  bandwidth requests, only  $X_{j+1}^1$  requests can be granted by BS.

Then, the uplink data throughput is equal to the total size of  $X_{j+1}^1$  bandwidth grants, and then equals the data transmission period  $D_{j+1}^1$  in frame  $j + 1$ :

$$\begin{aligned} T_{S_j > G_{j+1}} &= D_{j+1}^1 = \sum_{i=1}^{X_{j+1}^1} Q_i, \\ X_{j+1}^1 &< M * S_j. \end{aligned} \quad (18)$$

Then, the number of non-granted bandwidth requests in a multicast group during frame  $j + 1$  is

$$k_j^1 = S_j - G_{j+1}^1. \quad (19)$$

We can get that the number of pending bandwidth request transmitting in a group during frame  $j + 1$  is

$$t_{j+1}^1 = F\lambda + x_j + k_j^1 = t_j + (F\lambda - G_{j+1}^1). \quad (20)$$

### 6.2. $S_j < G_{j+1}$

In this case, the contention period and the data transmission period in frame  $j + 1$  are  $C_{j+1}^2$  and  $D_{j+1}^2$ , respectively. The  $D_{j+1}^2$  is so large that all  $M * S_j$  bandwidth requests are granted by BS. Furthermore, some minislots ( $D_{j+1}^2 - \sum_{i=1}^{M * S_j} Q_i$ ) are idle and do not use to grant the requests.

Then, the uplink data throughput is equal to the total size of  $M * S_j$  bandwidth grants, and then less than the data transmission period  $D_{j+1}^2$  in frame  $j + 1$ :

$$T_{S_j < G_{j+1}} = \sum_{i=1}^{M * S_j} Q_i < D_{j+1}^2. \quad (21)$$

There are no non-granted bandwidth requests,  $k_j^2 = 0$ . Then, the number of pending bandwidth request transmission in one group during frame  $j + 1$  is

$$t_{j+1}^2 = F\lambda + x_j + k_j^2 = t_j + (F\lambda - S_j). \quad (22)$$

### 6.3. $S_j = G_{j+1}$

In this case, the contention period in frame  $j + 1$  is  $C_{j+1}^3$ , the data transmission period  $D_{j+1}^3$  is large enough that all minislots exactly are used to grant all  $M * S_j$  bandwidth requests.

Then, the uplink data throughput is equal to the total size of  $M * S_j$  bandwidth grants, and then then equals the data transmission period  $D_{j+1}^3$  in frame  $j + 1$ :

$$\begin{aligned} T_{S_j = G_{j+1}} &= \sum_{i=1}^{M * S_j} Q_i, \\ D_{j+1}^3 &= \sum_{i=1}^{X_{j+1}^3} Q_i, \\ M * S_j &= X_{j+1}^3. \end{aligned} \quad (23)$$

There is no non-granted bandwidth requests  $k_j^3 = 0$ , and no idle minislots in  $D_{j+1}^3$ . Then, the number of pending bandwidth request transmission  $t_{j+1}^3$  in one group during frame  $j + 1$  is the same as in Section 6.2,

$$t_{j+1}^3 = F\lambda + x_j + k_j^3 = t_j + (F\lambda - S_j). \quad (24)$$

#### 6.4. Results analysis

Based on the above analysis, we got the results of the uplink data throughput and pending bandwidth requests in three different situations. Comparing (18), (21), and (23), we can get that

$$\begin{aligned} M * S_j &> X_{j+1}^1, \\ T_{S_j = G_{j+1}} &= T_{S_j < G_{j+1}} > T_{S_j > G_{j+1}}. \end{aligned} \quad (25)$$

Comparing (20), (22), and (24), we can get that

$$\begin{aligned} S_j &> G_{j+1}^1, \\ t_{j+1}^3 &= t_{j+1}^2 < t_{j+1}^1. \end{aligned} \quad (26)$$

Then, higher uplink throughput and less pending bandwidth request transmission in frame  $j + 1$  can be achieved in Sections 6.2 and 6.3. However, a further analysis of the uplink throughput and pending bandwidth request transmission in frame  $j + 2$  leads to a strong difference between Sections 6.2 and 6.3. During frame  $j + 1$ , among the three cases, we can obviously get the result based on (18), (21), and (23),

$$D_{j+1}^1 < D_{j+1}^3 < D_{j+1}^2. \quad (27)$$

Based on (1) and (2), we can get

$$\begin{aligned} C_{j+1}^1 &> C_{j+1}^3 > C_{j+1}^2, \\ N_{j+1}^1 &> N_{j+1}^3 > N_{j+1}^2. \end{aligned} \quad (28)$$

As we know, the function  $S = t(1 - (1/N))^{t-1}$  is a continuous and monotone increasing function with respect to  $N$ . Then, we can get  $S_{j+1}^3 > S_{j+1}^2$ . Applying (21), (22), (23), and (24), we can get the following results:

$$\begin{aligned} t_{j+2}^3 &= F\lambda + t_{j+1}^3 - S_{j+1}^3 < t_{j+2}^2 = F\lambda + t_{j+1}^2 - S_{j+1}^2, \\ T_{S_{j+1} = G_{j+2}} &= \sum_{i=1}^{M * S_{j+1}^3} Q_i > T_{S_{j+1} < G_{j+2}} = \sum_{i=1}^{M * S_{j+1}^2} Q_i. \end{aligned} \quad (29)$$

Then, we concluded that the maximum uplink throughput and the minimum pending transmission can be obtained by optimizing the contention request period size  $C_{opt}$ . This optimal size could make the number of bandwidth requests successfully received by BS in former frame be equal to the number of bandwidth grants allocated to SSs in current frame:

$$S_j = t_j \left(1 - \frac{R * M}{C_{opt_j}}\right)^{t_j - 1} = G_{j+1}, \quad (30)$$

where  $G_{j+1} = \lceil X_{j+1}/M \rceil$ .

Since the bandwidth requests indicate the uplink bandwidth, the SSs make the reservation from the BS, then the BS knows the bandwidth grants size  $Q_i$  for each request, and then the  $G_{j+1}$  can be calculated based on the information of the size of data transmission period  $D_{j+1}$  and the size of bandwidth grant  $Q_i$ :

$$D_{j+1} = F - C_{opt_{j+1}} - I_{j+1} = \sum_{i=1}^{X_{j+1}} Q_i. \quad (31)$$

Furthermore, Abi-Nassif et al. [10] developed an estimation scheme to measure the number of requests in a data over cable service interface specification (DOCSIS) [11] system. (DOCSIS specification is developed by Cable Television Laboratories as the major industry standard for two-way communication over hybrid fiber/coax (HFC) cable plants. The DOCSIS MAC is strikingly similar to IEEE 802.16 MAC, since IEEE 802.16 standard is developed based on IEEE 802.14 and DOCSIS.) However, their study assumed the number of retransmitted requests to be negligibly small compared to the number of new requests, which do not reflect the real situation. To solve the problem, Yin and Lin [12] proposed a statistically optimized minislot allocation algorithm to maximize the request minislot throughput by estimating the number of new requests with a time-proportional scheme and the number of collided requests by looking up a statistical most likelihood number of requests table. The scheme drives the request minislot throughput to the optimal bound by accurately estimating the number of requests and allocating that number of minislots to resolve them. The schemes from the above research can also be used here to estimate the  $t_j$  in our analysis.

Then, we can calculate and thus set the optimal contention period in any frame  $j$ :

$$C_{opt_j} = \frac{R * M}{1 - (X_{j+1}/M * t_j)^{1/(t_j - 1)}}. \quad (32)$$

#### 6.5. Optimal contention period for the different number of SSs

Now we know that the optimal contention period  $C_{opt}$  is achieved when  $S_j = G_{j+1}$ . However, the  $S_j$  varies with the  $t_j$  when a  $C_{opt}$  is configured and exhibits an unstable behavior. In particular, as shown in Figure 6, as  $t$  increases, the  $S_j$  increase to a maximum value, further increases of  $t$  lead to an eventually significant decrease of  $S_j$ . In order to find out this maximum value, we take the derivative of  $S_j$  with respect to  $t$ , and imposing the derivation equal to 0:

$$\frac{dS}{dt} = \left(1 - \frac{1}{N_{opt}}\right)^{t-1} + t \left(1 - \frac{1}{N_{opt}}\right)^{t-1} \ln \left(1 - \frac{1}{N_{opt}}\right) = 0. \quad (33)$$

Then, we get the  $t$  when the  $S_j$  is maximum:

$$t_{max} = \frac{1}{\ln(N_{opt}/(N_{opt} - 1))}. \quad (34)$$

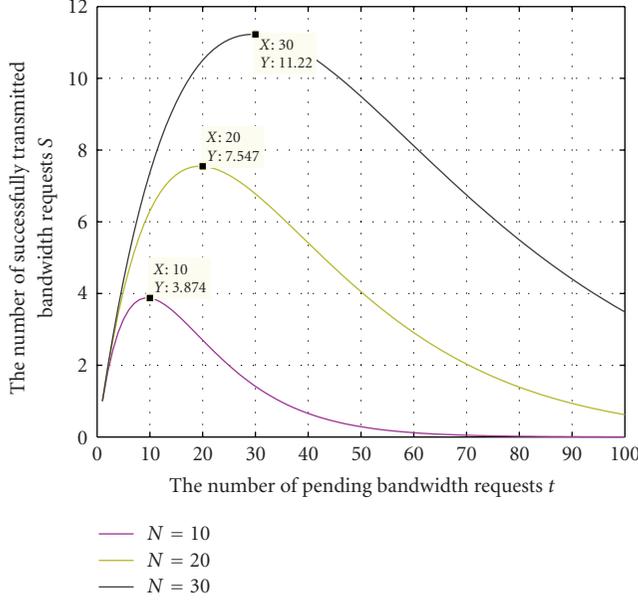


FIGURE 6: The number of successful transmission with  $N = 10, 20, 30$ .

If  $\ln(N_{\text{opt}}/(N_{\text{opt}} - 1))$  is developed in Taylor series and assume that  $N_{\text{opt}}$  is large enough, then we can get

$$t_{\max} = N_{\text{opt}} = \frac{C_{\text{opt}}}{R * M}. \quad (35)$$

We know that the variety of network size will change the value of  $\lambda$ , which will change the  $n$ , and then changes the  $t$ . As the  $S$  varies with the  $t$  and the  $C_{\text{opt}}$  depends on the condition  $S_j = G_{j+1}$ , then the optimal contention period size should be dynamically adapted with different network size in order to get better performance. In the following parts, we will analyse the influence to the  $C_{\text{opt}}$  and the maximum uplink data throughput caused by the network size.

Suppose  $C_{\text{opt}_j}$  and  $C_{\text{opt}_{j+1}}$  are the size of optimal contention period in frame  $j$  and  $j + 1$ , respectively, then we know  $S_j = G_{j+1}$  and the maximum uplink throughput is  $T = \sum_{i=1}^{M * S_j} Q_i$ . We thus analyse the influence from the network size in two cases as follows.

#### 6.5.1. $t < C_{\text{opt}}/(R * M)$

In this case, new SSs entry during the time of frame  $j$  increases the value of  $t_j$  and makes the  $S_j$  increase. Then, we get the situation where  $S'_j > G_{j+1}$ . In order to keep the best performance, the  $G_{j+1}$  in the frame  $j + 1$  has to increase to get a new tradeoff  $S'_j = G'_{j+1}$ , which makes the configured  $C_{\text{opt}_{j+1}}$  to decrease to a new value  $C'_{\text{opt}_{j+1}} < C_{\text{opt}_{j+1}}$ . Under this condition, the maximum uplink throughput in frame  $j + 1$  also increased to  $T' = \sum_{i=1}^{M * S'_j} Q_i$ . The above process will continue with the new SSs entering the network, until the condition  $t = C_{\text{opt}}/R$  is reached.

TABLE 1: The simulation parameters.

Upstream channel capacity	QPSK, 2.56 Mbps
Time covered by MAP	10 milliseconds
Minislot size	50 $\mu\text{sec}$
Contention slots	0–100 minislots
Bytes per minislot	16
Backoff window start	1
Backoff window end	8
Upstream scheduling service	Nonreal time polling
Piggyback	Disabled

#### 6.5.2. $t > C_{\text{opt}}/(R * M)$

In this case, new SSs entry during the time of frame  $j$  makes the  $S_j$  decrease. Then, we know that  $S'_j < G_{j+1}$ . In order to maintain the best performance, the  $C_{\text{opt}_{j+1}}$  in frame  $j + 1$  has to increase to a new value  $C'_{\text{opt}_{j+1}} > C_{\text{opt}_{j+1}}$ , making  $G_{j+1}$  to decrease to get a new tradeoff  $S'_j = G'_{j+1}$ . The maximum uplink throughput is also decreased to  $T'' = \sum_{i=1}^{M * S'_j} Q_i$ . The above process will continue with the new SSs entry.

## 7. SIMULATION RESULT

This section presents the performance evaluation of IEEE 802.16 networks by simulation and then validates our analytical model. We report on the simulation results and make the observation. We only present a limited number of cases here. However, we find that the conclusions we draw here are generally true for many other cases we have evaluated. We build an IEEE 802.16 simulation network by using OPNET [13] simulator with DOCSIS module. Table 1 shows the simulation parameters. The nrtPS service class is adopted, the unicast polling is defined impossible during the simulation time and the piggybacking of bandwidth request is prohibited, which makes the SSs must only use contention request TOs to request bandwidth.

It has to indicate that, during the simulation with DOCSIS module, the actual time covered by MAP may be different with the configured MAP size specified in Table 1. The actual time covered by MAP during simulation is calculated as the sum of time required for the configured contention slots, UGS grants, and reservation for requests. During the simulation, if the actual time calculated by the BS is less than the configured time, the BS will make up the missing time by padding some slots to the actual MAP size [13]. This additional slots are considered as contention slots and are added to the contention request period. That makes the actual size of contention request period bigger than the configured size in our simulation. Figure 7 shows the actual size of contention request period in our simulation. Unless otherwise specified, the contention slots in the following figures represent the actual slots in contention request period.

We first do the simulation to calculate the number of bandwidth requests successfully transmitted and the number

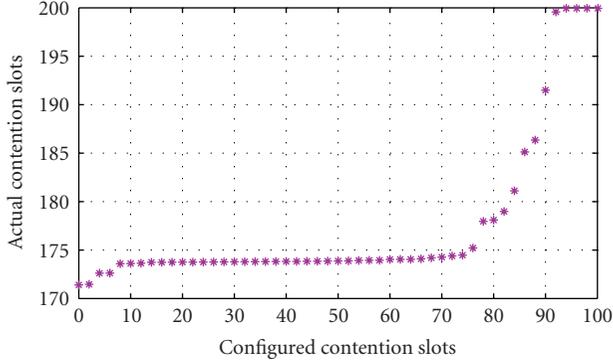


FIGURE 7: The actual contention slots versus the configured contention slots,  $V = 10$ .

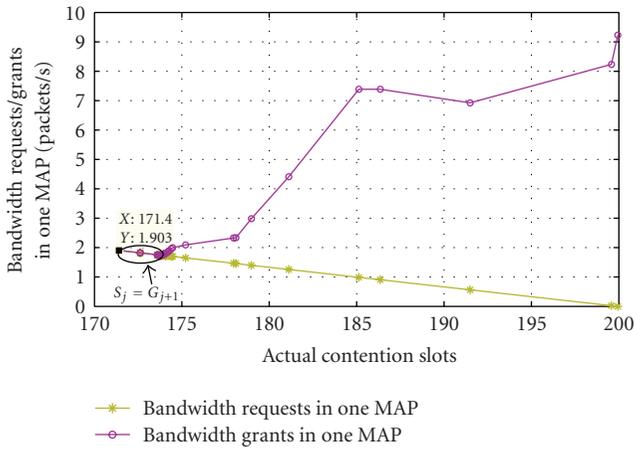


FIGURE 8: The bandwidth request versus the bandwidth grants of one multicast group during one MAP,  $V = 10$ .

of bandwidth grants in a network with 10 SSs. Figure 8 shows results. The  $x$ -axis is the actual contention slots and the  $y$ -axis is the number of requests and grants during one MAP time. The figure displays the sensitivity of the requests and grants with contention slots in one MAP. We can observe that the number of bandwidth requests monotonously increases with the increase of contention slots, while the number of bandwidth grants monotonously decreases with the increase of contention slots. This indicates that, in IEEE 802.16 networks, the unilateral increase of the number of bandwidth requests may not always increase the uplink throughput and sometimes make it worse, which is similar to what we discussed in Section 4. We can also observe that the number of bandwidth grants reaches the maximum, where the number of corresponding bandwidth requests is equal to the number of bandwidth grants,  $S_j = G_{j+1}$ . Shown in the figure, this size of contention request period is 171 slots.

Note that, in Figure 8, there is no zone where the number of bandwidth requests less than the number of bandwidth grants, which are discussed in our mathematical model as the case  $S_j < G_{j+1}$ . On the contrary, we find that there is a zone but not a point where  $S_j = G_{j+1}$ . We also notice that

the contention slots begin from 171 slots but not 0 slot as it is shown in Table 1.

That is reasonable. The case  $S_j < G_{j+1}$  is not a stable status in our simulation. If we set the configured size of contention request period  $C_{j+1}^{\text{con}}$  and data transmission period  $D_{j+1}^{\text{con}}$  to make  $S_j < G_{j+1}$ , then the BS only allocates  $S_j$  bandwidth grants to SSs. This means that the actual data transmission period  $D_{j+1}^{\text{act}} = \sum_{i=1}^{M \cdot S_j} Q_i$  is less than the configured size  $D_{j+1}^{\text{con}} = \sum_{i=1}^{M \cdot G_{j+1}} Q_i$ . Then, the actual frame size  $F^{\text{act}} = I_{j+1} + C_{j+1}^{\text{con}} + D_{j+1}^{\text{act}}$  less than the configured size  $F^{\text{con}} = I_{j+1} + C_{j+1}^{\text{con}} + D_{j+1}^{\text{con}}$ . According to the definition of OPNET DOCSIS module, some minislots should be added into the actual MAP as the contention slots to make up the missing time, where  $F^{\text{act}} = F^{\text{con}} = I_{j+1} + C_{j+1}^{\text{act}} + D_{j+1}^{\text{act}}$ . These additional contention slots increases the configured contention request period, then increases the value of  $S_j$ . Along with the simulation going, the value of  $S_j$  increases until  $S'_j = G_{j+1}$  or  $S'_j > G_{j+1}$ . At that time, all the slots in data transmission period are utilized as the bandwidth grants and no idle slots exist. The value of  $S_j$  does not increase any more, the simulation system is steady. That makes clear why the case  $S_j < G_{j+1}$  does not exist but a portion of  $S_j = G_{j+1}$  exists, and it also well explains why the actual contention slot in the  $x$ -axis begins from 171 slots instead of 0 slot. After the portion of  $S_j = G_{j+1}$ , we go into the period  $S_j > G_{j+1}$ , in which the number of bandwidth requests increases but the number of bandwidth grants decreases with the increase of contention slots.

To investigate the uplink throughput with different size of contention request period, we present the simulation results in Figure 9. The  $x$ -axis is the actual contention slots and the  $y$ -axis is the uplink throughput which is represented by the number of packets transmitted in uplink channels per second. The uplink throughput monotonously decreases with the increase of contention slots, and the maximum value is arrived when the contention period is 171 slots. It shows that the maximum uplink throughput could always be achieved by determining an optimal contention request period, where we can also get  $S_j = G_{j+1}$ .

For the sake of finding out the pending bandwidth request transmission on a contention period basis, we do the simulation to calculate the number of collision requests and the number of non-granted requests with different contention periods. We present the simulation results in Figure 10, the  $x$ -axis is the actual contention slots and the  $y$ -axis is the sum of the number of collision and non-granted requests. We observed that the number of retransmitted bandwidth request monotonously increases with the increase of contention slot, and the minimum value is got when the contention period is 171 slots. It lays out that the minimum number of pending bandwidth requests might always be achieved when an optimal contention period is adopted.

A conclusion can be made based on the above simulation results. In a PMP IEEE 802.16 network, the maximum uplink data throughput and the minimum number of pending bandwidth requests could always be achieved when an optimal contention period is determined by the BS. Furthermore, during this optimal contention period, the number of

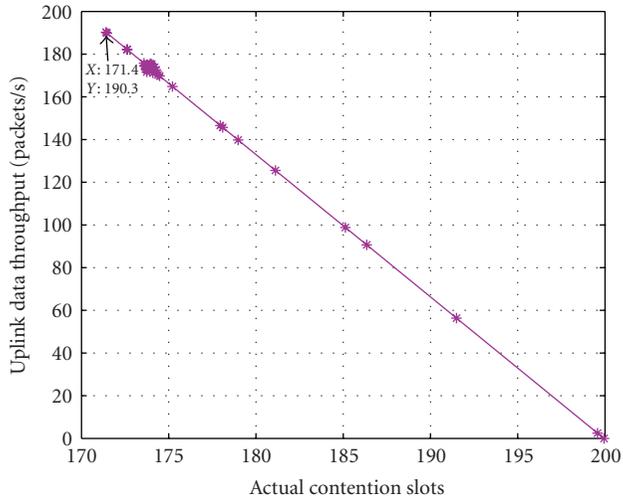


FIGURE 9: The uplink throughput with variable contention slots,  $V = 10$ .

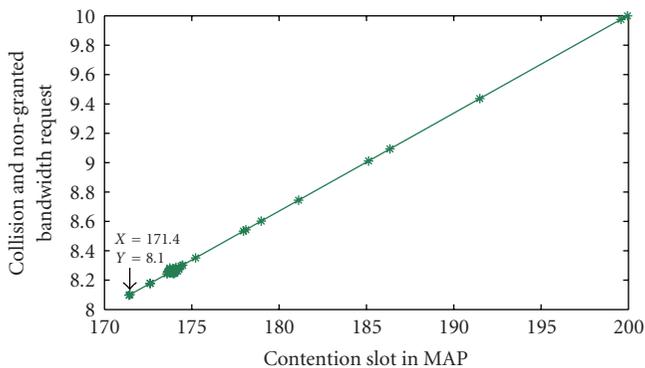


FIGURE 10: The retransmitted bandwidth request with variable contention slots,  $V = 10$ .

bandwidth requests successfully received by the BS is equal to the number of bandwidth grants the BS allocated.

In order to observe the influence of performance caused by the different network size, we repeat the simulation by changing the number of SSs in the networks. Figure 11 shows the simulation result, the x-axis is the number of SSs and the y-axis is the value of maximum uplink throughput (packets/sec) and the corresponding size of optimal contention request period (slots). We can observe that the network size definitely influences the maximum uplink throughput and the corresponding optimal contention request period size, and leads to very different trend. Shown in the figure, as the network size increases, the maximum uplink throughput increases to a maximum value, while the optimal contention request period decreases to a minimum value. However, further increase of network size leads the uplink throughput to decrease but the optimal contention period to increase, which shows that there is a critical point (mentioned in our analysis  $t = C_{opt}/(R * M)$ ) where an optimal network size is got with the optimum performance.

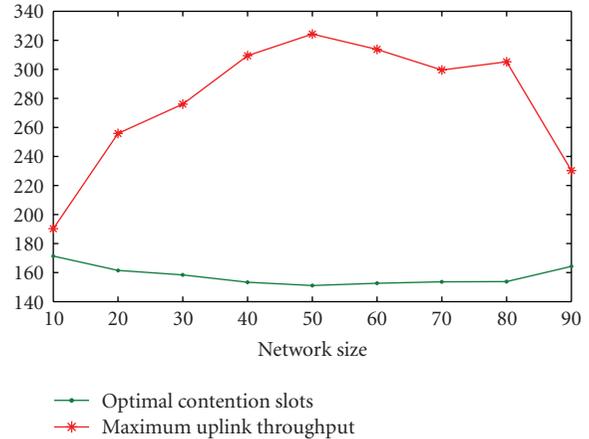


FIGURE 11: The maximum uplink throughput and the optimal-contention-request period with different network size.

## 8. CONCLUSION

In this paper, we presented an analytical model to evaluate and optimize the performance of the delay-tolerant and contention-based applications in IEEE 802.16 broadband wireless access networks. An optimal contention period based on a certain number of SSs proposed in our proposed model can maximize the uplink data throughput and minimize the pending bandwidth requests. This optimal contention period varies with the number of SSs in the network in order to keep the best performance. Further more, this best performance could be optimum with an optimal network size. Our analytical results were verified by the simulations using the OPNET DOCSIS module. The analytic models and results in this paper can be used to optimize the performance of IEEE 802.16-based MAC protocol, such as fixed WiMAX, Mobile WiMAX, DOCSIS, WiBro, and IEEE 802.20 MAC protocols.

We only analysed the performance for the delay-tolerant and contention-based applications in the IEEE 802.16 network in our research. Involving the delay-sensitive and contention-free applications into the analysis model, finding out the optimal contention period to maximize the uplink throughput and minimize the delay is our future works. Scheduling in the BS side to effectively guarantee the bandwidth grants and scheduling in the SS side to choose the appropriate honor connections to transmit data are also the research interests in the future. And an efficient polling mechanism is also the future work under consideration.

## REFERENCES

- [1] "IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems," 2004.
- [2] "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 and corrigendum 1," 2005.

- [3] G. Benelli, G. R. Cau, and A. Radaelli, "Performance evaluation of slotted Aloha multiple access algorithms with fixed and variable frames for radiomobile networks," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 181–193, 1994.
- [4] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.
- [5] G. Bianchi, L. Fratta, and M. Oliveri, "Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs," in *Proceedings of the 7th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '96)*, vol. 2, pp. 392–396, Taipei, Taiwan, October 1996.
- [6] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system," in *Proceedings of the IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions (ICCCAS & WeSino Expo '02)*, vol. 1, pp. 435–439, Chengdu, China, July 2002.
- [7] 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 *Proceedings of the 1st International Conference on Distributed Frameworks for Multimedia Applications (DFMA '05)*, pp. 130–136, Besancon, France, February 2005.
- [8] S.-M. Oh and J.-H. Kim, "The analysis of the optimal contention period for broadband wireless access network," in *Proceedings of the 3rd IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom '05)*, pp. 215–219, Kauai Island, Hawaii, USA, March 2005.
- [9] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, American Mathematical Society, Providence, RI, USA, 2nd edition, 1997.
- [10] F. Abi-Nassif, W. C. Lee, and I. Stavrakakis, "Offered load estimation in a multimedia cable network system," in *Proceedings of the IEEE International Conference on Communications (ICC '99)*, vol. 1, pp. 140–146, Vancouver, Canada, May 1999.
- [11] Cable Television Laboratories Inc, "Data-overcable service interface specifications docsis: Mac and upper layer protocols interface specification cm-sp-mulpiv3.0-i04-070518," 2007.
- [12] W.-M. Yin and Y.-D. Lin, "Statistically optimized minislots allocation for initial and collision resolution in hybrid fiber coaxial networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 9, pp. 1764–1773, 2000.
- [13] OPNET Technologies Inc, <http://www.opnet.com/>.