Performance Evaluation and Parameterization of the IEEE 802.16 Contention-Based CDMA Bandwidth Request Mechanism for the OFDMA Physical Layer

Dirk Staehle and Rastin Pries
University of Würzburg
Institute of Computer Science, Chair of Distributed Systems Würzburg, Germany
dstaehle,pries@informatik.uni-wuerzburg.de

Alexey Vinel
Saint-Petersburg Institute for Informatics and Automation, Russian Academy of Sciences
St. Petersburg, Russia
vinel@ieee.org

Andreas Mäder
NEC Network Laboratories Europe
Heidelberg, Germany
andreas.maeder@nw.neclab.eu

ABSTRACT

The IEEE 802.16 standard specifies two contention-based mechanisms for the OFDMA physical layer to transmit bandwidth requests from subscriber station to base station: the standard mechanism is based on Slotted Aloha with a truncated binary exponential backoff; the alternative one is based on CDMA. This paper describes the CDMA-based contention mechanism and presents an analytic model to compute its performance in terms of delay and consumed resources. The tunable parameters for the CDMA-based random access procedure are the number of ranging subchannels, the number of codes per ranging subchannels, and the detection threshold. An optimal configuration is derived for a given load in terms of the request arrival rate.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Performance attributes

General Terms

Performance, Design

Keywords

Performance evaluation, Dimensioning, WiMAX, IEEE 802.16, Random access, CDMA

1. INTRODUCTION

The IEEE 802.16 standard [4] defines physical and medium access control (MAC) layer functionality for fixed and mobile broadband wireless access networks. The WiMAX Forum (Worldwide interoperability for Microwave Access) is a consortium with the goal to bring IEEE 802.16 products to market. Currently, the WiMAX forum certifies products built on the OFDM 256 physical layer under the label "Fixed WiMAX", and products built on the OFDMA physical layer under the label "Mobile WiMAX". In order to support mobility in Mobile WiMAX networks an entire network architecture is developed and standardized by the Network Working Group NWG.

The IEEE 802.16 MAC layer is strictly connection oriented and specifies a detailed set of QoS parameters as the scheduling service of a transport connection. The base station is responsible for scheduling data transmissions in the downlink and for resource allocations in the uplink, and must by these means ensure that the QoS parameters of every connection are maintained. On the uplink, a scheduling type with associated parameter set defines how the base station assigns resources for data transmissions (grants) or resources for requesting bandwidth (polling) to a subscriber station. Polling can either allocate a request transmission to a single subscriber station (unicast polling) or to a set of subscriber stations (multicast/broadcast polling). The available scheduling types are the Unsolicited Grant Services (UGS), the extended real-time polling service (ertPS), the real-time polling service (rtPS), the non real-time polling service (nrtPS), and the best-effort service (BE).

Transport connections with UGS, ertPS, or rtPS scheduling type rely mainly on direct grants without prior request or on unicast polling while transport connections with nrtPS or BE scheduling type rely mostly on broadcast polling with contention. Besides the mandatory Slotted Aloha type contention mechanism, the IEEE 802.16 standard specifies an alternative CDMA-based random access (CRA) mechanism for the OFDMA physical layer. This mechanism allows a subscriber station (SS) to request unicast polling by transmitting a ranging code. Therefore, it randomly selects one out of a predefined set of ranging codes and transmits it on a ranging subchannel. If the base station (BS) detects the transmission of a code on a ranging subchannel it assigns a

MSWiM 2009 Tenerife, Canary Islands, Spain

bandwidth request transmission opportunity (BRTO) to the ranging code that the SS may then use to transmit its actual bandwidth request (BR). The ranging codes are quasi-orthogonal which means that they are not entirely orthogonal but possess a low cross-correlation. Accordingly, the SSs share the ranging subchannel by a CDMA scheme such that up to a certain degree the parallel transmission of multiple different ranging codes on the same subchannel can be identified by the BS. The BS detects a ranging code as transmitted if the cross-correlation of the code and the received superposition of different ranging codes exceeds a certain threshold. Consequently, using a low detection threshold leads to a high possibility that all transmitted codes are detected by the BS. On the other hand, a low detection threshold also increases the probability that codes are detected though they were not transmitted by any SS. In the latter case, one BRTO is wasted. The tunable parameters of the CRA mechanism are the number of ranging subchannels per frame, the number of ranging codes per subchannel, and the detection threshold. The contribution of this paper is first, an analytical performance model for the CRA scheme, and second, some clues on how to choose good or optimal CRA parameters for a certain request load.

In Section 2 we give a more detailed description of the IEEE 802.16 CRA scheme and explain which problems occur and how they are solved. Furthermore, we discuss the problem on how to set the tunable CRA parameters. In Section 3, we give an overview of the existing literature on IEEE 802.16 ranging codes and on the different IEEE 802.16 random access schemes. In Section 4, we specify the detailed system model and present an analytical performance model. In Section 5, we study the impact of the CRA scheme and derive some clues on how to obtain good parameters for a certain request load. In Section 6, we summarize the main contribution of this paper.

2. CDMA BASED RANDOM ACCESS (CRA)

The CRA scheme is exactly referred to as *Contention-based CDMA bandwidth requests for WirelessMAN-OFDMA*. The principle of the operation is that the BS broadcasts the location of one or more ranging subchannels within the UL-MAP. A ranging subchannel consists of 144 subcarriers. A ranging code is a binary pseudo-noise sequence of 144 bits which is transmitted on the ranging subchannel by BPSK modulating each of the 144 subcarriers. Every BS generates 256 ranging codes using polynomial generator $1+x+x^4+x^7+x^{15}$ with a BS specific seed. Out of these 256 ranging codes, subgroups are used for initial ranging, periodic ranging, bandwidth requests, and handover ranging.

When a SS intends to transmit data it first selects a random ranging subchannel and a random ranging code out of the set of ranging codes available for bandwidth requests. The BS receives the superposition of ranging codes on every ranging subchannel. By correlating the received signal with all available ranging codes it detects whether a code is present or not. A more detailed description of the detection procedure is given in Section 4.1. As a response to every detected ranging code, the BS assigns a BRTO in the UL-MAP of the next frame. Since the BS does not know which SS transmitted the ranging code, the allocation is not addressed to a certain connection ID but to the ranging subchannel-ranging code pair. If the SS recognizes that an allocation to the ranging subchannel-ranging code pair it used in the respective frame is present in one of the following UL-MAPs, it uses this allocation to transmit a BR which informs the BS about the amount of data it has presently to transmit. The BS receives the BR and starts to grant the requested data volume in the following frames. The principle of the CRA is illustrated in Fig. 1. The random access procedure takes at least two frames until the data transmission process starts which means that avoiding the collision of BRs by requesting exclusive BRTOs via CDMA codes is achieved at the cost of an additional delay of one frame.

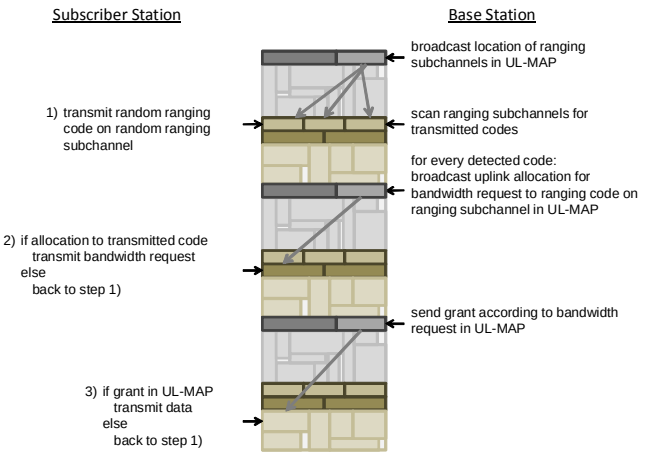


Figure 1: CDMA random access procedure

However, there are still two cases in which the CRA procedure is corrupted. The first case are collisions that occur when two SSs select the same ranging subchannel and ranging code. The BS detects the ranging code and schedules a BRTO in one of the next UL-MAPs. Both SSs recognize their ranging subchannel-ranging code pair in the UL-MAP and their BRs collide. Consequently, the BS is not able to assign grants to any of the SSs. After a certain time which might be already the next UL-MAP the SSs notice that they did not receive a grant and restart the procedure by transmitting a ranging code again. Alternatively, a binary exponential backoff algorithm might be initiated. The second case are failed detections, i.e. the BS fails to detect a ranging code that is not collided. Many parallel ranging codes and received power impairments may cause failed detections. As a consequence of a failed detection, the SS recognizes no BRTO allocation to its ranging-subchannel ranging-code pair in the next UL-MAP and reacts by restarting the procedure, i.e. it transmits a random ranging code on a random ranging subchannel and waits for a corresponding allocation in the next UL-MAP. In this paper, we assume that the random access delay is minimized, i.e. the BS assigns a BRTO always in the next UL-MAP and if it is not present, the SSs assumes that its ranging code was not detected and immediately retransmits a new code. Also, the BS assigns a grant directly after receiving a BR and if it is not present in the next UL-MAP, the SS assumes that the BR failed and retransmits a ranging code in the next frame.

Another problem are false detections, i.e. the BS detects a ranging code that is not transmitted by any SS. False detections may result from many parallel ranging code transmissions and received power impairments. The consequence of a false detection is that the BS schedules a BRTO in the next frame which remains unused.

The CRA procedure gives us three tunable parameters: the number of ranging subchannels, the number of ranging codes, and the detection threshold that tunes the trade-off between failed and false detections. The key performance parameters are the probability of code collisions and the probability of failed detections as these values define the random access delay. The amount of resources consumed by the random access procedure consists of the ranging subchannels, the BRTO for the successful, the collided, and the unused BRs. For every BR, we have to take into account both the entry in the UL-MAP and the actual BRTO.

Let us now formulate the problem how to optimize the CRA parameters. The objective is to minimize the average consumed resources while keeping the random access delay below a desired threshold. Increasing the number of ranging subchannels decreases all negative effects, i.e. code collisions, detection failures, and false detections while increasing the amount of consumed resources. Increasing the number of codes per ranging subchannel decreases the probability of code collisions but also increases the number of false detections. Increasing the detection threshold decreases the probability of failed detections but increases the probability of false detections. We first intend to derive the performance of a single ranging subchannel and optimize the number of codes and the detection threshold. Then, the number of ranging subchannels can be determined according to the estimated request arrival rate.

3. RELATED WORK

Performance evaluation and parameter optimization of contention-based or contention-free bandwidth request schemes of the IEEE 802.16 standard are researched intensely since the publication of the standard. One of the first works on IEEE 802.16 random access was [11], that presents an analysis of the mean random access delay for a given number of saturated stations and shows how the delay can be minimized by choosing appropriate values for the minimum and maximum contention window. This analysis is further refined in a couple of studies [2, 15]. In particular the work of Fallah et al [2], presents a precise analytic model for the mandatory contention-based BR mechanism for both persistent and non-persistent request generation. He et al [3] propose an analytic model for the bandwidth efficiency of IEEE 802.16 considering both BR process and grant allocation scheme. They make the simplifying assumption that a fixed number of grants can be served per frame and BRs are either served directly or dropped. A comparison of contention-free (unicast polling) and contention-based (random access) BR schemes can be found in [6, 12] with the conclusion that random access is preferable when the request rate is low and polling when the request rate is high. The efficiency of piggybacking requests to data transmissions is studied in [1, 7].

The related work described so far considers the mandatory contention-based access schemes. In the following, we describe the more closely related work that focuses on the CRA mechanism. Here, we distinguish the work on ranging [5, 14, 16] and the work on the CRA mechanism using the ranging subchannel [8, 9, 10]. The work on ranging focuses on multi-user ranging code detection, timing, frequency, and power synchronization. Lee and Morikawa, [5] provide an analysis of the initial ranging process of IEEE 802.16e with different multi-path channel profiles. The key difference of detection during initial ranging and code detection for bandwidth requests is that in the latter case the signals are already synchronized and received at an adjusted power level. Zhou et al, [16], propose two different approaches for the IEEE 802.16e initial ranging procedure. In the first method, timing offset estimation is done in time domain while in the second method it is done utilizing the correlation in frequency domain. Simulations show that both methods perform well and are superior to other existing methods. You et al, [14] analyze the capacity of a ranging subchannel in terms of the ranging code error probability depending on the number of time-synchronous distinct code requests. They consider both an AWGN and a two-ray Raleigh fading channel.

In [10], Staehle et al compare by means of simulations the different random access variants specified in the IEEE 802.16 standard, i.e. the mandatory one for OFDM256, the extension with subchannelization, and the CRA for OFDMA. A key result is that the CRA mechanism for OFDMA provides a comparable performance in terms of delay while consuming less resources. Seo et al, [8] provide an analysis of the IEEE 802.16a BR procedure based on ranging codes which is similar though not identical to the CRA scheme presented here. It is assumed that the transmitted codes are orthogonal such that no multiple access interference occurs and retransmissions are caused by collisions only while failed detections are neglected. The steady-state probability and the resulting mean delay are found by solving a Discrete Time Markov chain for the number of collided codes per frame.

In [9], Seo et al present an analytic model for the performance of the CRA mechanism including the actual data transmission. They focus on a single SS and model its queue length by a Markov chain. The SS may be either in idle mode, set-up mode, or transmission mode. Transitions occur from idle mode to set-up mode when new packets arrive. Transitions from set-up mode to transmission mode occur after the successful delivery of a BR, and transitions from transmission mode to idle mode occur when the queue becomes empty. Transitions from transmission mode to set-up mode can not occur since request piggybacking is allowed. One key part of the analytic model is the derivation of the request success probability, i.e. the probability that a request is transmitted and successfully received by the BS. The success probability is determined for a given number of requests per ranging subchannel and includes both the case of request collisions and failed detections. This derivation is in principle similar to the approach presented in Section 4.3 though the formulation is different. The main distinction of this paper is that retransmissions due to collisions and failed detections are strictly separated since they are resolved in different frames which is not considered in [9]. The second key component of the analytic model is the derivation of the steady-state probabilities. The steady-state balance equations are formulated using the success probability which again depends on the probability that a random SSs transmits a code request in steady-state. The solution of this mutual dependency does not become entirely clear. The main result of the analytic model is the mean and variance of a SS's queue length. The main enhancements of this paper compared to [9] are a) a more sophisticated derivation of the detection probability, b) the distinction of collided and not detected code requests, c) a derivation of the request delay distribution and resource consumption of the random access mechanism. Furthermore, detailed and ex-

tensive simulations show the accuracy of our model and we present guidelines how to optimize parameters for the random access mechanism. The main features considered in [9] but not in this paper are a) the exponential backoff for retransmissions and b) the transmission of grants using an ARQ process.

4. PERFORMANCE MODEL FOR A SINGLE RANGING SUBCHANNEL

This section describes an analytic model to compute the performance of a ranging subchannel with a given load in terms of requests arriving according to a Poisson process with rate λ. The ranging subchannel is characterized by the number of available ranging codes C and the detection threshold θ. The performance metrics are the random access delay and the amount of resources consumed by the random access process. The analytic model is described in the following way: In Section 4.1, the mechanism by which the BS detects a code is specified mathematically and, based on this description, Section 4.2 provides a simple formula for the probability that a transmitted code is actually detected. In Section 4.3, a performance model for a subchannel with a given number of requests is provided that is used in Section 4.4 to determine the steady-state distribution of the number of retransmitted requests per frame. In Section 4.5, the distribution of the random access delay and the consumed resources are derived from the steady-state distribution.

4.1 Code Detection Mechanism

In the following we will explain how the BS detects whether a ranging code is present in the received signal or not. Therefore, let us consider the following situation: a candidate SS uses code j and there are N other SSs on the same ranging subchannel that transmit different codes $x \neq j$ in parallel. Assuming ideal conditions every bit of every code is received with equal power that we normalize to one, i.e. $c_{x,k} \in \{-1, +1\}$ for all bits $k = 1, ..., 144$ of all transmitted codes. For detecting whether a code j is present, the BS correlates the received superposition of $N + 1$ codes with code j. If the correlation yields a value larger than the detection threshold θ the BS assumes the code to be present. Mathematically, code j is detected if

$$\frac{1}{144} \sum_{k=1}^{144} \left(c_{j,k} + \sum_{i=1}^{N} c_{i,k} \right) \cdot c_{j,k} \geq \theta. \quad (1)$$

In more realistic propagation conditions, not all bits are received with equal power. Instead, power control imperfections lead to different average bit powers among the SSs and frequency-selective fading channels lead to different powers for the 144 bits of the code of one user. Consequently, the propagation channel changes the kth bit of code x to

$$r_{x,k} = \gamma_x \cdot \beta_{x,k} \cdot c_{x,k}, \quad (2)$$

where γ_x is a random variable with mean $\mathrm{E}[\gamma_x] = 1$ for the relative signal power of SS x. Following the widely used models introduced by Viterbi et al [13] for power control imperfections in CDMA systems, we model γ_x by a Lognormal random variable and scale the degree of received power impairments by its standard deviation $\mathrm{STD}[\gamma_x]$. The random variable $\beta_{x,k}$ describes the relative power of the kth bit of SS x which is determined by the frequency-selective fading channel. Note that the bit powers $\beta_{x,k}$ are correlated for different bits k and their sum always equals one, i.e., $\sum_{k=1}^{144} \beta_{x,k} = 1$. Let $\psi_{j,x}$ be a random variable for the correlation of code j with the code received from SS x, i.e.

$$\psi_{j,x} = \sum_{k=1}^{144} r_{x,k} \cdot c_{j,k} = \gamma_x \sum_{k=1}^{144} \beta_{x,k} \cdot c_{x,k} \cdot c_{j,k}. \quad (3)$$

Substituting Eq. (3) into Eq. (1), we obtain that ranging code j is detected if

$$\varphi_j = \psi_{j,j} + \sum_{x=1}^{N} \psi_{j,x} \geq \theta. \quad (4)$$

Let us further introduce the variable $I_j = \sum_{x=1}^{N} \psi_{j,x}$ for the interference of code j and the variable φ_j for the correlation factor of code j, i.e. the result when correlating the received signal with code j.

4.2 Detection Probability

Let us now determine the probability that the candidate code j is detected if it does not collide with the other N requests that share the $C - 1$ remaining ranging codes. The code usage vector $\bar{z} = z_1, ..., z_K$ describes how the other N SSs share these codes, i.e. there are z_k codes that are used by exactly k of the N interfering SSs and $\sum_{k=1}^{K} k \cdot z_k = N$. We are interested in the probability that code j is actually detected when additionally $\bar{z}$ codes are transmitted.

In order to determine the probability that code j is detected, we model the interference I_j as a zero-mean Gaussian random variable and interpret its variance as a function of the code usage vector $\bar{z}$

$$\sigma_I^2(\bar{z}) = \sum_{k=1}^{K} k \cdot z_k \cdot \left(\mathrm{E}\left[\gamma^2\right] \cdot \Gamma_f + (k-1) \right). \quad (5)$$

Please refer to the Appendix for the proof that the zero-mean assumption is justified and a derivation of the variance. The variable Γ_f characterizes the impact of the multipath propagation profiles. Values for exemplary profiles are obtained by Monte Carlo simulations and shown in Tab. 1.

The variable $\psi_{j,j}$ stands for the correlation of a code j with its received copy, i.e.

$$\psi_{j,j} = \gamma_j \cdot \sum_{k=1}^{144} \beta_{j,k} \cdot c_{j,k} \cdot c_{j,k} = \gamma_j \quad (6)$$

is exactly one when we assume equal mean powers and Lognormal if we consider power control imperfections. Consequently, the probability that the BS detects ranging code j is

$$p_d(\bar{z}) = P\{\psi_{j,j} + N(0, \sigma_I^2(\bar{z})) > \theta\} \quad (7)$$

$$= \begin{cases} Q\left(\frac{1-\theta}{\sigma_I(\bar{z})}\right) & \text{for equal mean powers} \\ \int_0^\infty a_\gamma(x) Q\left(\frac{x-\theta}{\sigma_I(\bar{z})}\right) dx & \text{for LN mean powers.} \end{cases}$$

The function $a_\gamma(x)$ denotes the probability density function of γ.

4.3 Deterministic Number of Code Requests

Let us now consider a single frame, in which N code requests are sent on a ranging subchannel with C ranging codes. From the last section we know the probability that a

Table 1: Channel specific values for Γ_f

		ITU			
	flat	Veh. A	Veh. B	Ped. A	Ped. B
Γ_f	0	1.50	1.47	1.25	1.57

code is actually detected when the code usage of the other $N-1$ requests is expressed by a vector $\bar{z}$. We are know interested in the probability $p_F(u,c|n)$ to have c collided requests and additionally u failed detections when in total n requests are sent.

Let us first consider the probability $p_Z(\bar{z}|n)$ that n requests result in code usage vector $\bar{z}$. For the computation, the length of the vector $\bar{z}$ is limited to length K though theoretically more than K SSs might use a single code. Consequently, z_K is not the number of codes used by exactly K but by at least K SSs. We determine this probability recursively

$$p_Z(\bar{z}|n) = p_Z(\bar{z}-\bar{1}_1|n-1)\cdot\frac{C-(z_1-1)-\sum_{k=2}^{K} z_k}{C} \quad (8)$$

$$+\sum_{k=1}^{K-1} p_Z(\bar{z}+\bar{1}_k-\bar{1}_{k+1})|n-1)\cdot\frac{z_k+1}{C} \quad (9)$$

$$+\,p_Z(\bar{z}|n-1)\cdot\frac{z_K}{C}, \quad (10)$$

where $\bar{1}_k = (0,...,0,1,0,...,0)$ is a vector of zeros with a single one at position k. The recursion distinguishes three cases according to the lines of the equation. The nth SSs can either select a previously unused code or select a code which was previously selected by $k<K$ other SSs or select a code which was previously selected by at least K other SSs.

Now, the probability to have c requests involved in collisions is

$$p_C(c|n) = \sum_{\bar{z}|z_1=n-c} p_Z(\bar{z}|n). \quad (11)$$

The number of undetected requests when the code usage is $\bar{z}$ follows a binomial distribution and the probability to have u undetected requests is

$$p_U(u|\bar{z}) = \binom{z_1}{u}(1-p_d(\bar{z}-\bar{1}_1))^u\cdot p_d(\bar{z}-\bar{1}_1)^{z_1-u}. \quad (12)$$

Putting the two equations together, we obtain

$$p_F(u,c|n) = \sum_{\bar{z}|z_1=n-c} p_Z(\bar{z}|n)\cdot p_U(u|\bar{z}). \quad (13)$$

4.4 Steady-State Distribution

The result of the last section is a formula to derive the joint distribution of collided and undetected requests from the number of requests transmitted on a ranging subchannel. In this section, we consider requests that arrive according to a Poisson process and are interested in the steady-state distribution of retransmitted requests per frame. Please remember that an undetected request is retransmitted in the next frame while a collided request is retransmitted two frames after the collision. Consequently, we define a state (r,c) of the Markov chain by the number r of requests that are retransmitted in the current frame and the number c of requests that collided in the previous frame and will be retransmitted in the next frame. The distributions of successful, collided, and undetected requests in the current frame are entirely determined by the number r of retransmissions.

Let $p^{(i)}(r,c)$ be the probability to be in state (r,c) just before the ith frame. It is derived from state distribution $p^{(i-1)}(r,c)$ just before frame $i-1$ as

$$p^{(i)}(r,c) = \sum_{(r^-,c^-)} p^{(i-1)}(r^-,c^-)\cdot q\left\{(r^-,c^-),(r,c)\right\}. \quad (14)$$

The probability of the transition from state (r^-,c^-) to state (r,c) is

$$q\left\{(r^-,c^-),(r,c)\right\} = \sum_{m=0}^{M_{max}} p_M(m)\cdot p_F(r-c^-,c|r^-+m) \quad (15)$$

where $p_M(m)$ is the probability to have m new requests which is

$$p_M(m) = \frac{(\lambda\tau)^m}{m!}\cdot e^{-\lambda\tau} \quad (16)$$

for a Poisson process with arrival rate λ and frame length τ. The variable M_{max} is an upper bound for the maximum number of requests per ranging subchannels and set to 100 for all numerical studies in this paper. The steady-state distribution $p^\infty(r,c)$ is determined by numerically iterating Eq. (14) until convergence is reached. The iteration is initiated without retransmissions and collisions and convergence is achieved when the mean number of retransmissions changes by less than one-tenth of a percent. From the two-dimensional state distribution $p^\infty(r,c)$ we can easily determine the probability

$$p_R(i) = \sum_{c=0}^{M_{max}} p^\infty(i,c) \quad (17)$$

to have i retransmissions per frame.

4.5 Performance Metrics

The performance metrics for the CRA procedure are on the one hand the random access delay and on the other hand the consumed resources. These resources include the ranging subchannels and the slots reserved for BRs. Obviously, every successful request requires at least one BRTO so we are mainly interested in the wasted resources, i.e. the BRTO reserved for collided ranging requests and those for false detections which means for codes that are detected by the BS though they have not been transmitted.

The random access delay is defined as the time from the start of the frame in which the first code request is sent until the end of the frame in which the BR is successfully received at the BS. Thus, the minimum random access delay is two frames. In the first one the code request is transmitted and detected by the BS. In the second one, the BS polls a BR in the UL-Map and it is delivered successfully. Let us now determine the distribution of the random access delay. The minimum random access delay of two frames is increased by two for every time the code request is involved in a collision and by one for every time a code request is not detected. Then, the probability $p_D(d)$ that the random access delay consists of two plus d additional frames is given by

$$p_D(d) = \sum_{c=0}^{\lfloor d/2\rfloor}\binom{d-c}{c}\cdot(p_C)^c\cdot(p_U)^{d-2\cdot c}\cdot p_S, \quad (18)$$

where p_C is the probability that a request collides, p_U is the probability that a code does not collide but is not detected, and p_S is the probability that a request is successful, i.e. it does not collide and is detected. We obtain these probabilities from the distribution $p_N(i) = p_M(i) \circledast p_R(i)$ of requests per frame in the steady-state by

$$p_C = \sum_{n=0}^{M_{max}} p_N(n) \cdot \sum_{(u,c)} p_F(u,c|n) \cdot \frac{c}{n} \quad (19)$$

$$p_U = \sum_{n=0}^{N_{max}} p_N(n) \cdot \sum_{(u,c)} p_F(u,c|n) \cdot \frac{u}{n} \quad (20)$$

$$p_S = \sum_{n=0}^{N_{max}} p_N(n) \cdot \sum_{(u,c)} p(u,c|n) \cdot \left(1 - \frac{c+u}{n}\right). \quad (21)$$

The number of wasted resources is equal to the number of collisions and false detections per frame. The joint distribution of collisions and false detections is determined similar to the joint distribution of collisions and retransmissions. Let $p_W(f,c)$ be the probability to have f false detections and c collisions in a frame in steady-state. We obtain

$$p_W(f,c) = \sum_{n=0}^{N_{max}} p_N(n) \cdot p_W(f,c|n) \text{ with} \quad (22)$$

$$p_W(f,c|n) = \sum_{\bar{z}|\sum_{k=2}^{K} z_k = c} p(\bar{z}|n) \cdot p_F(f|\bar{z}) \text{ and} \quad (23)$$

$$p_F(f|\bar{z}) = \binom{F(\bar{z})}{f} \cdot p_{false}(\bar{z})^f \cdot (1 - p_{false}(\bar{z}))^{F(\bar{z})-f}. \quad (24)$$

Here, $F(\bar{z}) = C - \sum_k z_k$ corresponds to the number of codes that are not used by any SS and $p_{false}(\bar{z})$ is the probability that one of these codes is detected by the BS if the code usage is $\bar{z}$. Such a code is detected when the interference I_x for a code x alone is larger than the detection threshold and we obtain

$$p_{false}(\bar{z}) = P(I_x \geq \theta) = Q\left(\frac{\theta}{\sigma_I(\bar{z})}\right). \quad (25)$$

The total number of wasted resources is the sum of the collisions and false detections and the probability $p_W(i)$ to have i wasted BRTOs per frame is

$$p_W(i) = \sum_{(f,c)|f+c=i} p_W(f,c). \quad (26)$$

5. VALIDATION AND PARAMETRIZATION

This section first presents a validation of the detection probability approximation based on a Monte Carlo simulation. The impact of different parameters like detection threshold, power control imperfections, and multi-path channels are studied. Then, the steady-state distribution and the resulting delay distribution are validated by a time-dynamic simulation. Furthermore, the trade-off between increased delay and wasted resources are shown by tuning the number of ranging codes and the detection threshold. Finally, the optimal number of ranging subchannels and the ranging subchannel parameters is determined depending on the total request load.

5.1 Detection Probability Approximation

In order to evaluate the accuracy of the detection probability computation we determine the detection probabilities for perfect and imperfect power control (STD $[\gamma] \in \{0, 0.25\}$ and for a flat and the frequency-selective ITU Ped. A channel. The evaluation is done by Monte Carlo simulation technique using 250000 repetitions. The codes, the power per user, and the power per subcarrier are independently generated per repetition and user. The selection of codes is done

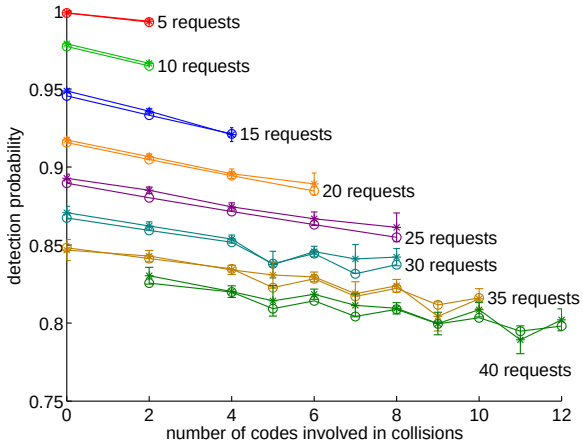


Figure 2: Accuracy of detection probability approximation for idealistic conditions.

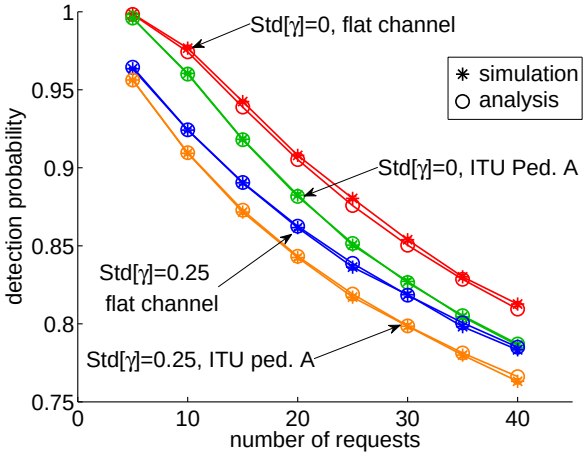


Figure 3: Impact of realistic propagation conditions.

in such a way that first the code of the candidate request is selected and then the N interfering codes are chosen randomly from the remaining codes. The confidence intervals indicate a confidence level of 95%.

Fig. 2 shows the detection probabilities for 192 codes and up to 40 requests with detection threshold $\theta = 0.5$, equal mean powers and a frequency-flat channel. On the x-axis, we have the number of codes that are involved in collisions and the different curves represent different numbers of requests, from 5 to 40. For reasons of clearness, the detection probabilities are not shown depending on the entire code usage but only depending on the number of total requests, i.e. including the candidate request, and depending on the number of codes involved in collisions. The analytic results are marked by a circle and the simulation results by a star plus 95% confidence interval. The simulation and analytic results match exaclty for a small number of requests, and still show acceptable accuracy for more requests. The larger confidence intervals and also the increased inaccuracy of the analysis for higher loads comes from the rare cases that many (more than four) codes are involved in one collisions. For reasons of computational complexity, the analysis was re-

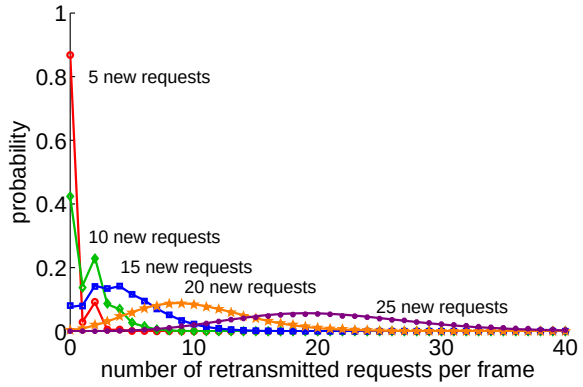
stricted to a maximum of four requests per code, i.e. the vector $\bar{z}$ has a length of $K = 4$.

In order to demonstrate the impact of power control imperfections and the ITU Ped. A channel, the detection probability is shown depending only on the number of requests aggregating all different code usage cases. In Fig. 3 the detection probabilities are shown for the four cases introduced above. First of all, we can see that the accuracy is very good. In fact, it even improves if imperfections are considered. This is due to the fact that the interference for the idealistic case assumes only discrete values while the interference with imperfections assumes continuous values and is consequently better approximated by the Gaussian distribution. Both types of imperfections decrease the detection probability. Unequal mean powers reduce the detection probability by about 4% more or less independent of the number of requests. On the contrary, the impact of the frequency-selective ITU Ped. A channel increases with the number of requests.

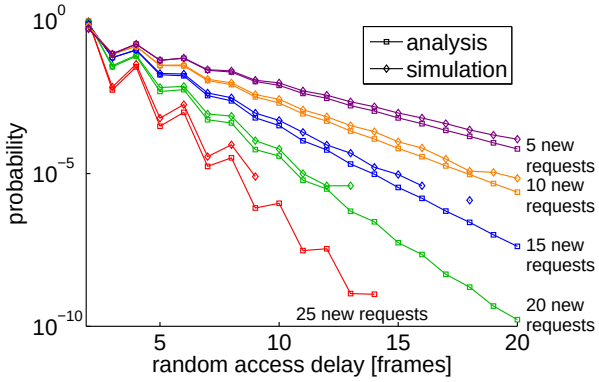
5.2 Steady-State Distribution

The steady-state distribution is validated by means of a time-dynamic simulation that generates new requests for every frame and transmits them together with the retransmissions waiting from the last two frames. For every frame and SS i an independent mean power γ_i and bit powers $\beta_{i,k}$ are generated. The simulation is performed for 50000 frames, the first 10000 frames are truncated as transient phase. The analytic model is validated by comparing the distribution of retransmitted requests per frame, the distribution of the random access delay, and the distribution of wasted resources with simulation results. The validation is done for loads from 5 to 25 new requests per frame and a single ranging subchannel with detection threshold 0.5 and 128 ranging codes. Fig. 4(a) depicts the distribution of retransmitted requests for the different loads. The analytic results are shown by solid lines and the belonging simulations by markers. Again, we can state a very good match of simulation and analysis. For low loads of 5 to 10 requests, the distribution has a minimum at one retransmission since the probability for a collision is higher than the probability for a failed detection. For a higher number of requests, the distribution becomes symmetric and the mean number of retransmissions approaches the mean number of new requests. The analysis still converges for a load of 30 requests per frame and diverges for 35 requests per frame though the cases of an actually overloaded ranging subchannel are not of primary interest here, since they lead to an extremely bad delay performance and an enormous waste of resources.

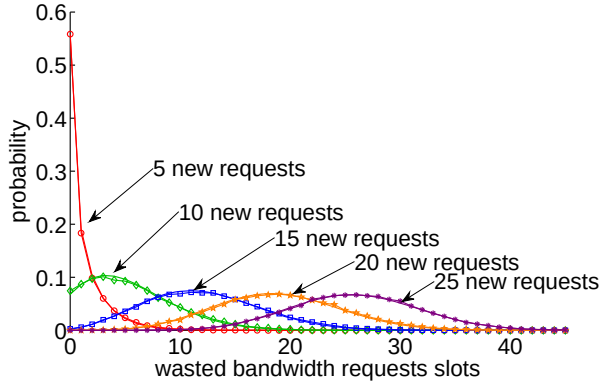
Fig. 4(b) shows the corresponding distribution of the extra random access delay, i.e. excluding the two always required frames. The y-axis is shown in logarithmic scale in order to make small probabilities visible. The minimum values obtainable for the simulation are about 5e-6 for 5 requests and 1e-6 for 25 requests since the total number of requests contributing to the statistics are the 40000 frames times the 5 to 25 mean requests per frame. The value for 15 requests and 17 frames is missing since in the simulation with 15 new requests no random access delay of 17 frames occurred whereas a random access delay of 18 frames occured. The accuracy of the analytic model is again very good though it tends to underestimate large delays. This is due to the fact that the probability for a collided or not detected request is always computed using the number of retransmissions in steady-state. However, the number of retransmitted requests is larger if a request has already been transmitted one or more times. This effect is not taken into account by the analysis.



(a) distribution of retransmissions



(b) distribution of random access delay



(c) distribution of wasted resources

Figure 4: Validation of steady-state performance.

Beside the delay, the other important metric for evaluating the quality of the random access parameter setting is the amount of consumed resources or rather the wasted resources, i.e. the resources spent for BRTOs that are not successfully used. Fig. 4(c) shows the number of wasted BRTOs per frame for the constellation with 128 codes and a detection threshold of 0.5. The analytic results are shown by the solid curve and the simulation results are shown by the

markers. Again, the analytic results have a good accuracy. It is interesting to observe that the number of wasted BRTOs becomes larger than the number of requests, i.e. they exceed the number of successfully used BRTOs when the load is above 20 new requests per frame.

5.3 Setting of Random Access Parameters

In the random access procedure there are three parameters that can be adjusted: the number of ranging subchannels, the number of ranging codes per ranging subchannel, and the detection threshold. Let us first focus on a single ranging subchannel and study the impact of detection threshold and number of ranging codes. The performance metrics will be a) the probability that the random access delay exceeds 5 frames and b) the mean number of wasted slots per frame. The performance evaluation is done for the scenario with idealistic conditions (equal bit powers) and a load of 20 requests per frame. Fig. 5 shows the two performance metrics depending on the number of codes on the x-axis and the detection threshold indicated by the different curves. Some of the points are missing, e.g. for 64 and 96 ranging codes and a detection threshold of 0.75, since the ranging subchannel is not able to carry the load and the iteration diverged. Let us first consider the delay performance in the upper part of the figure. As one would expect, an increased number of codes reduces the collision probability and thus the random access delay. Also, a lower detection threshold reduces the probability that a transmitted code is not detected and thus the random access delay.

Looking at the amount of wasted resources in the lower part of the figure, we observe two trends. First, choosing a higher detection threshold reduces the wasted resources. Second, more ranging codes lead to more wasted resources in particular if the detection threshold is too low. This observation is not in general true as for detection threshold 0.5 the wasted resources increase slightly if the codes are reduced from 96 to 64. This happens since the positive effect of less false detections is annihilated by an increased number of collisions that also lead to wasted resources.

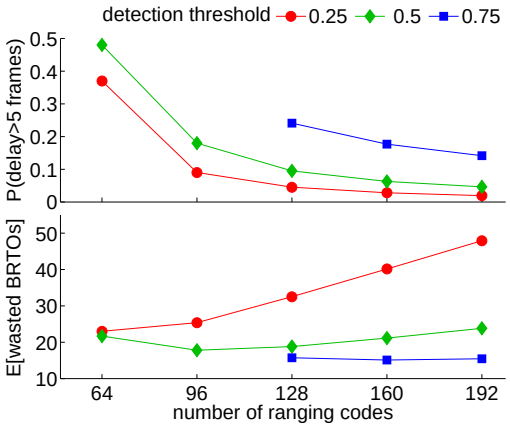


Figure 5: Impact of ranging subchannel parameters on ranging subchannel performance for a load of 20 new requests per frame.

The figure illustrates the trade-off between a low delay and few wasted resources. The lowest delay is obtained for many codes and a low detection threshold but at the price of many wasted resources. The least wasted resources are obtained with a high detection threshold and at least 128 codes which has the negative effect of high delays. A good compromise might be a detection threshold of 0.5 and 160 ranging codes.

Let us now formulate this as an optimization problem and find the optimal ranging subchannel parameters. The target is to minimize the mean number of wasted resources while maintaining a certain performance, e.g. the probability that the random access delay exceeds 5 frames should be less than 5%. In order to obtain the optimal parameters a hill-climbing type of search is used. The initial configuration is with 128 codes and a detection threshold of 0.5. The granularity of the solution space is 4 codes and 3/144 for the detection threshold. The number of codes is limited to a maximum of 196.

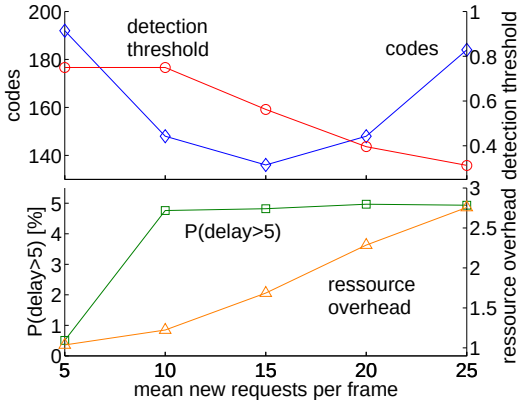


Figure 6: Optimal ranging subchannel parameter setting and resulting performance.

Fig. 6 shows the optimal parameter choice in the upper part and the resulting performance in the lower part. Instead of the wasted resources, we plot the resource overhead defined as the mean number of BRTOs required for successfully transmitting a single requesst. On the x-axis we have a load between 5 and 25 new requests per frame. Let us first discuss the ranging subchannel performance. The probability that the delay exceeds 5 frames increases with the load and approaches its limit of 5%. This shows that with a higher load, the ranging subchannel parameters can be better tuned to obtain the desired performance. The resource overhead is about 0.5 for 5 new requests per frame and increases to 3 for 25 new requests per frame. The latter results means that in average two BRTOs are wasted to transmit one BR.

Let us now study the optimal parameter settings in the upper part of the figure. With an increasing load the detection threshold decreases, i.e. it becomes more and more important to avoid failed detections. The optimal number of codes is at the maximum for a low load. More codes avoid collisions and with only few requests per frame the multiple access interference is small enough to clearly distinguish transmitted and not transmitted codes when choosing an appropriate detection threshold of 0.75. For medium loads, the interference increases and it becomes more difficult to distinguish correct and false detections. Thus, the detection threshold is decreased to ensure the detection of transmit-

ted codes. In order to compensate for the thereby increased probability of false detections the number of codes is reduced. For high loads, the number of codes increases again in order to avoid increasing the number of collisions that occur with more requests per frame.

In order to optimize the number of ranging subchannels per frame, we need to make an approximation since our model is valid only for a single ranging subchannel. The standard defines that a SS retransmits a code requests by randomly selecting a new ranging subchannel and a new ranging code. In the following, we make the assumption that the SS keeps the ranging subchannel and only reselects the ranging code. Thus, a new request chooses one of the S ranging subchannels randomly and the arrival process for a single random subchannel remains Poisson with rate λ/S. The delay distribution of a system with S subchannels and a load of $\mathrm{E}[M] = \lambda\tau$ mean new requests per frame is then identical to the delay distribution of a system with a single ranging subchannel and load $\mathrm{E}[M]/S$. The wasted BRTOs of the S ranging subchannels sum up and the total average is S times the average of a single subchannel. The optimal number of ranging subchannels is found by starting with a high number of ranging subchannels and decrementing them step by step as long as the amount of consumed resources decreases. The consumed resources are (1) 32 bits for the entry in the UL-MAP indicating the location of the ranging subchannels, (2) 144 bit per ranging subchannel, (3) 60 bits per BRTO to announce it in the UL-MAP, and (4) 48 bits for the actual BRTO. Assuming that all bits are transmitted using QPSK 1/2 modulation and coding scheme, we obtain the mean number of consumed resources in terms of subcarriers times OFDMA symbols as

$$\mathrm{E}[B] = 32 + S \cdot 144 + (S \cdot \mathrm{E}[W] + \mathrm{E}[M]) \cdot 108, \qquad (27)$$

where $\mathrm{E}[W]$ is the mean number of wasted BRTO per ranging subchannel. Fig. 7 shows the optimal number of ranging subchannels and the consumed resources for request loads per frame from 2 to 50. The results are shown for the idealistic case and for the case with an ITU Ped. A channel and a standard deviation of the mean received powers of $\mathrm{STD}[\gamma] = 0.25$. The first observation to make is that the consumed resources in terms of bits are a linear function of the load. This means that by adapting the number of ranging subchannels it is possible to avoid an exponential increase of the resource overhead as we have seen in Fig. 6 when increasing the load per ranging subchannel. In numbers, the transmission of a bandwidth request consumes 139 bits with idealistic conditions and 150 bits with imperfections. Of these bits, 108 bits are the minimum requirement for the successful BR. Only, for very low loads of 4 requests per frame or below, the consumed bits per bandwidth request are higher since the load is not high enough to fully utilize the capacity of the ranging subchannel. The optimal number of ranging subchannels is also increasing roughly linearly and as a rule of thumb we have one ranging subchannel per six requests in the idealistic case and one ranging subchannel per five requests with imperfections.

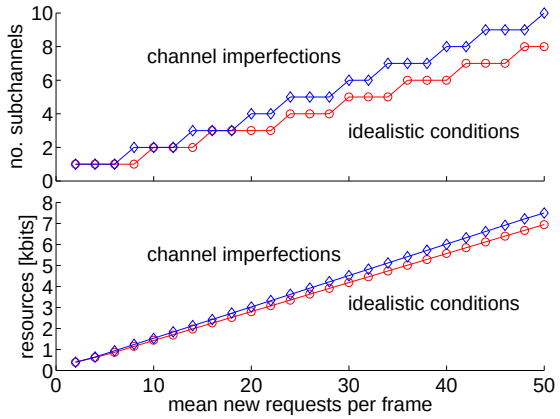


Figure 7: Optimal number of ranging subchannels and resulting consumed resources.

6. CONCLUSION

The IEEE 802.16 standard specifies the CRA mechanism as an alternative to transmit BR for best effort and non-real-time connections. In this paper we present an analysis of this BR mechanism taking into account a detailed radio channel model with both unequal mean powers per SS and unequal received bit powers within one code. As a result, we obtain the distribution of (1) the random access delay and (2) the amount of consumed resources in terms of BRTO. The trade-off between the delay performance and the amount of consumed resources is demonstrated and illustrated at an example. The tunable parameters of a ranging subchannel are the number of codes reserved for BRs and the detection threshold. The optimal subchannel configuration is defined as the configuration minimizing the consumed resources while keeping the probability to exceed a maximum random access delay of 5 frames below 5% and the optimal configuration is shown for loads between 5 and 25 new requests per frame with a single ranging subchannel. Finally, the optimal number of ranging subchannels with optimal configuration is found to be one ranging subchannel per six requests in the idealistic case and one ranging subchannel per five requests with imperfections.

The main and novel contributions of this paper are (1) a realistic and accurate model of the code detection probability, (2) the derivation of the steady-state distribution of back-logged requests using a two dimensional state description, (3) the formulation of the trade-off between delay and wasted resources and (4) the solution to the resulting optimization problem for the number of subchannels, the number of ranging codes per subchannel, and the detection threshold.

APPENDIX

In the following we are going to derive the mean and the variance of the interference term I_j. Therefore, let us first determine mean and variance of $\psi_{j,x}$ for $x \neq j$. We obtain:

$$\begin{aligned}\mathrm{E}[\psi_{i,j}] &= \mathrm{E}\left[\gamma_i \sum\nolimits_{k=1}^{144} \beta_{i,k} \cdot c_{j,k} \cdot c_{i,k}\right] \\ &= \mathrm{E}[\gamma_i] \cdot \sum\nolimits_{k=1}^{144} \mathrm{E}[\beta_{i,k}] \cdot \mathrm{E}[\cdot c_{j,k} \cdot c_{i,k}] = 0 \qquad (28)\end{aligned}$$

$$\mathrm{VAR}\left[\psi_{i,j}\right] = \mathrm{E}\left[\psi_{i,j}^2\right] = \mathrm{E}\left[\gamma_i^2 \cdot \left(\sum\nolimits_{k=1}^{144} \beta_{j,k} \cdot c_{i,k} \cdot c_{j,k}\right)^2\right]$$
$$= \mathrm{E}\left[\gamma_i^2\right] \cdot \mathrm{E}\left[\sum\nolimits_{k=1}^{144} \sum\nolimits_{\ell=1}^{144} \beta_{i,k} \cdot \beta_{i,\ell} \cdot c_{i,k} \cdot c_{i,\ell} \cdot c_{j,k} \cdot c_{j,\ell}\right]$$
$$= \mathrm{E}\left[\gamma_i^2\right] \cdot \mathrm{E}\left[\sum\nolimits_{k=1}^{144} {\beta_{i,k}}^2\right] = \mathrm{E}\left[\gamma_i^2\right] \cdot \Gamma_f. \quad (29)$$

The key assumption in this derivation is that the kth bits $c_{i,k}$ and $c_{j,k}$ of two different codes i and j independently assume the values $+1$ and -1 with equal probabilities. As a consequence, the product $c_{i,k} \cdot c_{j,k}$ also assumes the values $+1$ and -1 with equal probability and the expectation of the product is zero. In Eq. (29), we introduced the variable $\Gamma_f = \mathrm{E}\left[\sum_{k=1}^{144} {\beta_{i,k}}^2\right]$. This variable characterizes the impact of different multi-path propagation profiles. Values for different profiles are obtained by Monte Carlo simulations and shown in Tab. 1. An obvious implication of Eq. (28) is that the mean interference is zero. Let us now consider the variance of I_j. Please note that the interfering codes are all different from j but they are not necessarily different from each other. Consequently, two random variables $\psi_{x,k}$ and $\psi_{y,k}$ are independent only if $x \neq y$ and correlated if $x = y$. As a consequence, the variance of the interference is not simply the sum of the variances derived in Eq. (29). The vector $\bar{z} = (z_1, ..., z_K)$ describes the code usage and z_i is the number of codes that are transmitted by i of the N users. Let us introduce the variable $\psi_{i,x,j}$ for the correlation of code j with the received signal of SS i transmitting code x. Further, let $I_j^{(k)} = \sum_{i=1}^{k} \psi_{i,x,j}$ be a random variable for the interference produced to code j by k independent transmissions of code x. Then, we obtain

$$\mathrm{VAR}\left[I_j^{(k)}\right] = \mathrm{VAR}\left[\sum\nolimits_{i=1}^{k} \psi_{i,x,j}\right] = \mathrm{E}\left[\left(\sum\nolimits_{i=1}^{k} \psi_{i,x,j}\right)^2\right]$$
$$= \mathrm{E}\left[\sum\nolimits_{i=1}^{k} \psi_{i,x,j}^2 + \sum\nolimits_{s=1}^{k} \sum\nolimits_{t=1,t\neq s}^{k} \psi_{s,x,j} \cdot \psi_{t,x,j}\right]$$
$$= k \cdot \mathrm{VAR}\left[\psi_{i,j}\right] + k \cdot (k-1) \cdot \mathrm{E}\left[\psi_{s,x,j} \cdot \psi_{t,x,j}\right]$$
$$\mathrm{E}\left[\psi_{s,x,j} \cdot \psi_{t,x,j}\right]$$
$$= \mathrm{E}\left[\gamma_s \cdot \gamma_t \cdot \sum\nolimits_{k=1}^{144} \sum\nolimits_{\ell=1}^{144} \beta_{s,k} \cdot \beta_{t,\ell} \cdot c_{x,k} \cdot c_{j,k} \cdot c_{x,\ell} \cdot c_{j,\ell}\right]$$
$$= \underbrace{\mathrm{E}\left[\gamma\right]^2}_{=1} \cdot \sum\nolimits_{k=1}^{144} \sum\nolimits_{\ell=1}^{144} \underbrace{\mathrm{E}\left[\beta_{s,k}\beta_{t,\ell}\right]}_{=1/144^2} \underbrace{\mathrm{E}\left[c_{x,k}c_{j,k}c_{x,\ell}c_{j,\ell}\right]}_{=\begin{cases} 0 & \text{if } k \neq \ell \\ 1 & \text{if } k = \ell \end{cases}} = 1$$

Summarizing the results, we obtain the variance of I_j as a function of the code usage vector $\bar{z}$:

$$\mathrm{VAR}\left[I_j|\bar{z}\right] = \sum\nolimits_{k=1}^{K} z_k \cdot \mathrm{VAR}\left[I_j^{(k)}\right]$$
$$= \sum\nolimits_{k=1}^{K} k \cdot z_k \cdot \left(\mathrm{E}\left[\gamma^2\right] \cdot \Gamma_f + (k-1)\right). \quad (30)$$

1. REFERENCES

[1] L.-W. Chen and Y.-C. Tseng. Design and analysis of contention-based request schemes for best-effort traffics in IEEE 802.16 networks. *IEEE Communications Letters*, 12(8):602–604, Aug. 2008.

[2] Y. P. Fallah, F. Agharebparast, M. R. Minhas, H. M. Alnuweiri, and V. C. M. Leung. Analytical modeling of contention-based bandwidth request mechanism in IEEE 802.16 wireless networks. *IEEE Transactions on Vehicular Technology*, 57(5):3094–3107, Sept. 2008.

[3] J. He, K. Guild, K. Yang, and H.-H. Chen. Modeling contention based bandwidth request scheme for IEEE 802.16 networks. *IEEE Communications Letters*, 11(8):689–700, August 2007.

[4] IEEE LAN/MAN Standards Committee. DRAFT Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems, P802.16Rev2/D9 January 2009.

[5] D. H. Lee and H. Morikawa. Performance analysis of ranging process in IEEE 802.16e OFDMA systems. In *Proc. IEEE WiMob*, Oct 2007.

[6] Q. Ni, A. Vinel, Y. Xiao, A. Turlikov, and T. Jiang. Investigation of bandwidth request mechanisms under point-to-multipoint mode of WiMAX networks. *IEEE Communications Magazine*, 45(5):132–138, May 2007.

[7] R. Pries, D. Staehle, and D. Marsico. Performance evaluation of piggyback requests in IEEE 802.16. In *Proc. IEEE VTC Fall*, Sep 2007.

[8] H.-H. Seo, B.-H. Ryu, Choong-Ho, and H.-W. Lee. *Proc. of Wired/Wireless Internet Communications*, chapter Traffic Characteristics Based Perfromance Analysis Model for Efficient Random Access in OFDMA-PHY System, pages 213–222. Springer, 2005.

[9] J.-B. Seo, H.-W. Lee, and C.-H. Cho. Performance of ieee802.16 random access protocol - steady state queuing analysis. In *Proc. IEEE GLOBECOM*, Nov 2006.

[10] D. Staehle and R. Pries. Comparative study of the IEEE 802.16 random access mechanisms. In *Proc. NGMAST 2007*, Sep 2007.

[11] A. Vinel, Y. Zhang, M. Lott, and A.Tiurlikov. Performance analysis of the random access in IEEE 802.16. In *Proc. of IEEE PIMRC*, 2005.

[12] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov. Efficient request mechanism usage in IEEE 802.16. In *Proc. IEEE GLOBECOM*, Nov 2006.

[13] A. Viterbi and A. Viterbi. Erlang capacity of a power controlled CDMA system. *Journal on Selected Areas in Communication*, 11(6):892–900, Aug. 1993.

[14] J. You, K. Kim, and K. Kim. Capacity evaluation of the OFDMA-CDMA ranging subsystem in IEEE 802.16-2004. In *Proc. IEEE WiMob*, Aug, 2005.

[15] J. Zhou, Y. Yang, X. Jin, J. Shi, and Z. Li. Contention region allocation optimization in IEEE 802.16 OFDMA systems. In *Proc. ACM MSWiM*, Oct 2007.

[16] Y. Zhou, Z. Zhang, and X. Zhou. OFDMA Initial Ranging for IEEE 802.16e Based on Time-domain and Frequency-domain Approaches. In *Proc. ICCT*, Nov 2006.