Capacity Analysis of Reservation-Based Random Access for Broadband Wireless Access Networks

Alexey Vinel, Qiang Ni, Dirk Staehle, and Andrey Turlikov

***Abstract*—In this paper we propose a novel model for the capacity analysis on the reservation-based random multiple access system, which can be applied to the medium access control protocol of the emerging WiMAX technology. In such a wireless broadband access system, in order to support QoS, the channel time is divided into consecutive frames, where each frame consists of some consequent mini-slots for the transmission of requests, used for the bandwidth reservation, and consequent slots for the actual data packet transmission. Three main outcomes are obtained: first, the upper and lower bounds of the capacity are derived for the considered system. Second, we found through the mathematical analysis that the transmission rate of reservation-based multiple access protocol is maximized, when the ratio between the number of mini-slots and that of the slots per frame is equal to the reciprocal of the random multiple access algorithm's transmission rate. Third, in the case of WiMAX networks with a large number of subscribers, our analysis takes into account both the capacity and the mean packet delay criteria and suggests to keep such a ratio constant and independent of application-level data traffic arrival rate.**

***Index Terms*—random access, capacity, reservation, medium access control, WiMAX.**

I. Introduction

RECENTLY, random multiple access (RMA) technologies have received great attention for broadband wireless access networks (e.g. WiFi and WiMAX). Since 1970s RMA is widely known as an efficient method providing communication between a large number of subscribers with bursty traffic sources in packet-switched data networks. In [1], Rubin is one of the first authors, who considers centralized reservation-based random multiple access which can improve the performance of satellite networks. In Rubin's model [1], time-probabilistic characteristics are computed for different scenarios, particularly considering large propagation delay values, with the emphasis on reservation performed by means of time division multiple access (TDMA). The synchronized subscribers perform reservations, by transferring short requests to the central repeater, and then transmit multiple packet messages. Therefore, the shared broadcast channel is divided into so-called frames. Each frame consists of consequent mini-slots for reservation and slots for actual data packet transmission. In such a reservation-based multiple access system, access to the slots is normally regulated by a central base station using time division technique, each mini-slot can be either assigned periodically (through polling) to a single subscriber or be potentially used by all subscribers in a contention manner. The medium access control (MAC) protocol of contemporary IEEE 802.16 WIMAX broadband wireless technology in point-to-multipoint mode [2] can be treated as an example of reservation-based RMA system.

The most commonly used model for RMA system analysis was described in [5] by Tsybakov. Throughout the rest of the paper we will refer this model [5] as the *basic model*. Later its assumptions were expounded by Gallager in [6]. In contrast to Rubin's model, where a finite number of subscribers is assumed, the *basic model* assumes an infinite number of subscribers. Under this assumption the TDMA system is principally incapable of providing finite mean packet delay, while an RMA algorithm is capable of doing it. Obviously, infinite number of subscribers can not be polled in a TDMA fashion within a finite time. The RMA tree-algorithm, invented 30 years ago by Tsybakov and Mikhailov [3] and independently by Capetanakis [4], is the first-known method to provide a finite mean delay for the infinite number of subscribers model.

Tsybakov and Berkovskii [7] consider the reservation problem in the framework of the basic model. In contrast to [1], in [7] requests are not considered and the subscriber indicates how long it will require the channel in regular packets. Packets from various subscribers compete with each other according to some RMA algorithm. If a packet from some subscriber is received successfully, then all other subscribers in the system stop their transmissions during the specified time interval, thus enabling the subscriber sending the packet to transmit its information without conflict.

In this paper, we propose a novel reservation-based random multiple access system model, which is built upon the combination of the models from [1] and [5]. Using our model, we perform a novel capacity analysis on the considered reservation-based broadband wireless access system. Our model can be utilized to analyze the WiMAX MAC layer. The usage of infinite number of subscribers model is motivated by the vision that the number of subscribers in a WiMAX network is expected to be fairly large. Our main contributions are:

- We first address the problem of the capacity analysis for the reservation-based WiMAX RMA system. Using our

Manuscript received 15 January; revised 15 August 2008. Part of this work has been presented in the XI International Symposium on Problems of Redundancy in Information and Control Systems [15], and 1st International Workshop on Multiple Access Communications [18], both at Saint-Petersburg. Qiang Ni would like to acknowledge the support from BRIEF award on this work.

Alexey Vinel is with Saint-Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, Russia. Part of this work was done when he was at the University of Wuerzburg under the support of German Academic Exchange Service (DAAD) and Alexander von Humboldt Foundation.

Qiang Ni is with the School of Engineering and Design, Brunel University, Uxbridge, London, UB8 3PH, UK (e-mail: Qiang.Ni@brunel.ac.uk).

Dirk Staehle is with University of Wuerzburg, Germany.

Andrey Turlikov is with Saint-Petersburg State University of Aerospace Instrumentation, Russia.

Digital Object Identifier 10.1109/JSAC.2009.090208.

analytical model and simulation analysis, we derive the optimal ratio of contention period and contention-free intervals in each frame, which maximizes the network capacity;
- We introduce a simple practical approach for setting the frame structure in WiMAX based on our analysis and also examine its efficiency.

The rest of the paper is organized as follows. In Section II the basic RMA model is explained and some auxiliary propositions are proved. Our *centralized reservation-based model* as well as the problem statement are presented in Sections III. Upper and lower bounds for the capacity are constructed in Section IV. Mean delay analysis is performed in Section V. Section VII concludes the paper.

II. Basic Random Multiple Access System Model

Here, we briefly explain the basic RMA system model and review some necessary definitions from [5]. Table I lists the notation used within this paper.

A. Review of the Basic Model

Actually the basic model can also be treated as an infinite subscribers model, where each subscriber can have at most one packet requiring transmission. The subscribers are assumed to transmit packets of a fixed length whose duration is taken as a time unit. The system is slotted, so that subscribers can begin packet transmissions only at times $t \in \{0, 1, 2, \ldots\}$. The time interval $[t, t+1)$ will be called a *slot*. The channel is noiseless and it is assumed that each subscriber knows by time $t+1$ which of the following three possible events, *idle slot*, *successful transmission*, or *conflict* (two or more simultaneous transmissions) occurred in the slot $[t, t+1)$. The packet generation times of all subscribers form the overall input traffic, which is assumed to be discrete Poisson. The probability that j new packets are generated at some moment t equals to $e^{-\lambda}\lambda^j/j!$, where λ is the intensity of the overall input traffic.

In the basic model, an *RMA algorithm for the basic system* is defined as a rule that enables any subscriber with a ready-for-transmission packet at any time $t \in \{0, 1, 2, \ldots\}$, to determine whether or not it should transmit this packet in the next slot $[t, t+1)$. Thus we have a function of three arguments. The first argument is the time x of packet generation. The second argument is the sequence $\theta(t) = (\theta_1, \ldots, \theta_t)$ of channel events θ_i, here $\theta_i = 0$ if $[i-1, i)$ was an idle slot, $\theta_i = 1$ if only one subscriber transmitted in this slot, and $\theta_i = 2$ if two or more subscribers transmitted in this slot. The third argument is the sequence $\nu(x, t) = (\nu_1(x), \ldots, \nu_t(x))$ of events at the subscriber where a packet was generated at time x. Here $\nu_i(x) = 0$ if this subscriber has not transmitted a packet in the slot $[i-1, i)$, and $\nu_i(x) = 1$ if it has. Therefore, formally an RMA algorithm is defined as a function $f_0[x, \theta(t), \nu(x, t)]$ with values in the interval $[0, 1]$. Its value is the probability that a packet generated at time x will be transmitted in the slot $[t, t+1)$.

The delay of a packet is the time interval from the moment of its generation till the moment of its successful transmission. The delay $\delta^{(0)}(\lambda, f_0)$ is a random variable. Let a packet be generated at an arbitrary but fixed time t at some subscriber, and let $\delta_t^{(0)}(\lambda, f_0)$, be its delay. The *mean delay* (referred to as *virtual mean delay* in [5]) is defined as $D_0(\lambda, f_0) \triangleq \limsup_{t\to\infty} E[\delta_t^{(0)}(\lambda, f_0)]$.

The *transmission rate*, R_0 (tenacity), of an RMA algorithm (f_0) is the maximum (more precisely, the supremum) intensity of the input traffic that can be transmitted by the algorithm with finite delay: $R_0(f_0) \triangleq \sup_\lambda\{\lambda\colon D_0(\lambda, f_0) < \infty\}$.

The *capacity*[1] of the basic RMA system is defined as $C_0 \triangleq \sup_{f_0\in\mathcal{F}_0} R_0(f_0)$, where $\mathcal{F}_0$ is a set of all RMA algorithms. The exact value of the capacity C_0 is still unknown. As it was mentioned in [8] some researchers conjectured that the optimal value might be 0.5, but this claim was quickly abandoned as baseless. The best known upper bound for the capacity C_0 was found by Likhanov and Tsybakov in [9] and is shown to be $\overline{C}_0 = 0.587$. The fastest known algorithm, a part-and-try one with rate $R_{pt} = 0.487$, was found by Tsybakov and Mikhailov in [10]. Later it was slightly improved, but the core of the algorithm remained the same.

Before presenting our model, we will first prove in the next subsection several auxiliary propositions for the basic RMA systems having some form of *feedback delay*.

B. Several Propositions for the Basic Model

In [11] the feedback information θ_i is assumed to be announced to all subscribers by time $i+N$, where N is the feedback delay. In the basic model the event in slot i is known by the beginning of slot $i+1$, meaning that $N = 1$. In this paper, we assume that all slots are grouped into equal consequent segments of length K. The values of function f_0 do not depend on the values of θ_i related to the current segment. For a given value of K, any RMA algorithm and the set of all RMA algorithms justifying this rule are denoted as $f_0^{(K)}$ and $\mathcal{F}_0^{(K)}$ respectively. Note that $\mathcal{F}_0^{(1)} \triangleq \mathcal{F}_0$. In the following, we will prove several interesting propositions:

Proposition 1: $C_0^{(K)} = \sup_{f_0^{(K)}\in\mathcal{F}_0^{(K)}} R_0(f_0) \le C_0$.

Proof: From the definition of class $\mathcal{F}_0^{(K)}$, it follows directly, that for any K: $\mathcal{F}_0^{(K)} \subset \mathcal{F}_0$ and thus proposition holds. ■

Proposition 2: For any algorithm $f_0 \in \mathcal{F}_0$, having transmission rate R_0, and any value of K an algorithm $f^{(K)} \in \mathcal{F}_0^{(K)}$ exists, which also has the transmission rate R_0.

Proof: Let us show how to construct the desired algorithm. Any algorithm $f_0 \in \mathcal{F}_0$ can be modified in the following way to be in the set $\mathcal{F}_0^{(K)}$. At the moment of a packet generation a subscriber chooses a number r uniformly from $\{1, 2, \ldots, K\}$ once and then "applies" algorithm f_0 only to slots having number r in any segment of K slots. This means, that each subscriber uses feedback from one fixed slot (which has number r in each segment) and can transmit only in such slots. Thus, we "split" our system into K independent basic systems, where each subscriber randomly chooses one system for its operation once and then works independently of those who have chosen a different system according to the

[1]Note that the capacities can be defined *over the class* in the sense that any other class different from $\mathcal{F}_0$ can be used in the above definition.

TABLE I
A SUMMARY OF NOTATION USED IN THIS PAPER.

λ	Intensity of the overall input traffic (per unit of time)
α	Mini-slot duration
K	Number of mini-slots per frame
L	Number of slots per frame
f_0	RMA algorithm for basic system
$\mathcal{F}_0$	Set of all RMA algorithms for basic system
$f_0^{(K)}$	RMA algorithm for the basic system with segmentation into K slots
$\mathcal{F}_0^{(K)}$	Set of RMA algorithms for system with segmentation into K slots
$f^{(K)}$	RMA algorithm for reservation-based system with frame with K mini-slots
$g^{(L)}$	Service discipline for reservation-based system with frame with L slots
$\phi^{(K)}$	RMA algorithm analogous to part-and-try, but for reservation-based system with K mini-slots per frame
$\varphi^{(L)}$	FIFO service discipline (each frame has L slots)
$\delta^{(0)}$	Delay of packet generated at time t in basic system
δ_n	Overall delay of additional packet generated in frame n in reservation-based system
$\delta_n^{(1)}$	Request delay for random access
$\delta_n^{(2)}$	The time from the moment of request successful transmission, to the corresponding packet will be successfully transmitted
D_0	Mean packet delay in basic system
D	Mean overall packet delay in reservation-based system
D_1	Mean request random access delay
$R_0(f_0)$	Transmission rate of RMA algorithm f_0
$R(f^{(K)}, g^{(L)})$	Transmission rate of multiple access protocol $(f^{(K)}, g^{(L)})$
R_{pt}	Transmission rate of part-and-try algorithm
C_0	Capacity of basic RMA system
$\overline{C}_0$	Best known capacity upper bound for basic system
C	Capacity of reservation-based system
$C_0^{(K)}$	Capacity achieved over the class $\mathcal{F}_0^{(K)}$
$\theta_i^{(l)}$	Channel event in mini-slot number l of $(i-1)$-th frame
θ_i	Channel event in slot $[i-1; i)$ for the basic system
$\bar{\theta}_i$	Feedback vector $(\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(K)})$ from $(i-1)$-th frame for a reservation-based system
$\theta(n)$	For basic system: sequence of channel events $(\theta_1, \ldots, \theta_n)$; for reservation-based system: sequence $(\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_n)$
$\nu_i(x)$	Indicator whether a packet generated at time x is transmitted in slot $[i-1; i)$ for basic system
$\nu_i^{(l)}(x)$	Indicator whether a packet generated at time x is transmitted in slot l of $i-1$-th frame for reservation-based system
$\bar{\nu}_i(x)$	Vector $(\nu_i^{(1)}(x), \nu_i^{(2)}(x), \ldots, \nu_i^{(K)}(x))$
$\nu(x, n)$	For basic system: sequence $(\nu_1(x), \ldots, \nu_n(x))$; for reservation-based system: sequence $\nu(x, n) = (\bar{\nu}_1(x), \bar{\nu}_2(x), \ldots, \bar{\nu}_n(x))$
n	Number of stations (for finite-user model)
l	Parameter of BEB algorithm determining minimum contention window, which equals to lK
m	Parameter of BEB algorithm determining maximum contention window, which equals to $2^m lK$

algorithm having transmission rate R_0/K. Thus, the overall transmission rate achieved is R_0. ■

Note that this approach does not necessarily guarantee, that the mean delay of the constructed algorithm will be "good". Moreover, it's easy to give examples when this "splitting" approach leads to unwarrantably high delay values [11].

Proposition 3: For any given K, the capacity $C_0^{(K)}$ achieved over the class $\mathcal{F}_0^{(K)}$ equals to the capacity of the basic system C_0 (achieved over the class $\mathcal{F}_0$).

Proof: On the one hand, from Proposition 1 it follows, that $C_0^{(K)} \leq C_0$. On the other hand, from Proposition 2 follows, that any algorithm from $\mathcal{F}_0$ for any K can be modified in the way that it can be in $\mathcal{F}_0^{(K)}$, without reducing its transmission rate. Thus, $C_0^{(K)} = C_0$. ■

III. Our Novel Reservation-Based Random Access System Model

A. Our System Model

Let us consider a broadband wireless access transmission system (e.g. WiMAX) with one *central base station* and infinite number of *subscribers*. The central station is connected to all subscribers by means of two communication channels, namely uplink and downlink. The *uplink channel* is used for the data transmission from all subscribers to the central station and the *downlink channel* is used for the information transmission from the base station to the subscribers (see Figure 1).

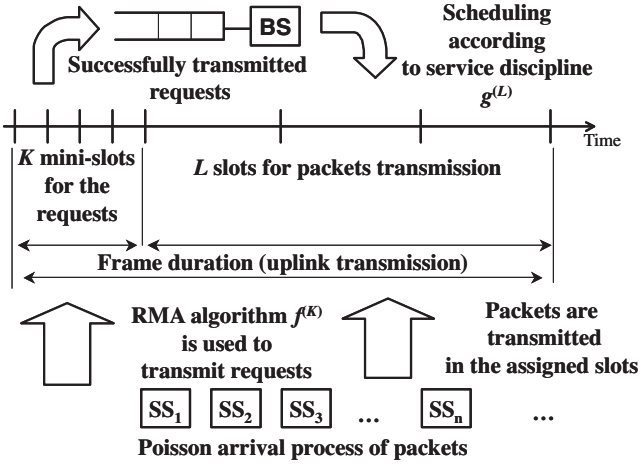


Fig. 1. Illustration for centralized reservation-based random multiple-access system

In our system, the traffic model used is the same as in the basic model - the moments of *packets* arrivals represent

a Poisson process, which provides an arrival rate equal to λ packets per unit of time. However, each subscriber, having a new packet, transmits a special *request* message to the central station in order to reserve uplink channel time. The duration of the request transmission is supposed to be $\alpha < 1$ units of time. In all following considerations we assume, that the *durations of request and packet transmissions are fixed* and the uplink channel usage is organized in the following way. The time axis is slotted into equal intervals of time, which are called *frames*. All frames have a fixed structure. Each frame comprises $K \geq 1$ intervals of time having duration α, which are called *mini-slots*, and $L \geq 1$ intervals of time having a duration equal to one unit of time, which are called *s*lots. Slots are used by the subscribers for transmitting packets, while mini-slots are used for sending requests.

The system is synchronized. The central station and all subscribers know the beginning of each i-th frame $i(\alpha K+L)$, each j-th slot $j+\alpha K\lfloor (j+1)/L \rfloor$ and each k-th mini-slot $k\alpha + L\lfloor k/K \rfloor$, where $i,j,k \in \{0,1,2,\ldots\}$ and transparent numeration of slots and mini-slots is assumed.

Since simultaneous transmissions of subscribers are possible in the mini-slots, three different situations can be distinguished in an arbitrary mini-slot $l \in \{1,2,\ldots,K\}$ of frame number $(i-1)$ (we denote them by $\theta_i^{(l)}$): *successful transmission* of some subscriber ($\theta_i^{(l)}=1$), *empty* mini-slot meaning that there is not any transmission ($\theta_i^{(l)}=0$), and *collision*, when two or more subscribers transmit in the mini-slot ($\theta_i^{(l)}=2$). By the beginning of frame i, the central station transmits information about the situation in the mini-slots of frame $i-1$ to all subscribers. This information is represented by the *feedback vector* $\bar{\theta}_i = (\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(K)})$. In WiMAX this information is implicitly presented in the grants to successfully received requests.

Subscribers transmit requests by means of some reservation-based RMA algorithm $f^{(K)}$, through which each subscriber determines at the beginning of each frame whether or not to transmit a request in a mini-slot of this frame taking into account the situations of previous frames. Analogous to the basic model $f^{(K)}$ is defined as a function of three arguments $f^{(K)}[x,\theta(n),\nu(x,n)]$, $n \in \{0,1,2,\ldots\}$. Here, x is the moment of time, when the packet is generated and $\theta(n) = (\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_n)$ is a sequence of feedback vectors until the beginning of frame n. Finally, $\nu(x,n) = (\bar{\nu}_1(x), \bar{\nu}_2(x), \ldots, \bar{\nu}_n(x))$ is a sequence of vectors for the subscriber x, $\bar{\nu}_i(x) = (\nu_i^{(1)}(x), \nu_i^{(2)}(x), \ldots, \nu_i^{(K)}(x))$. We denote $\nu_i^{(l)}(x)=0$ if the subscriber whose packet has been generated at time x did not transmit a request in the l-th mini-slot of the $(i-1)$-th frame and $\nu_i^{(l)}(x)=1$ otherwise. The possible values of the function f are vectors $\bar{p} = (p^{(1)}, p^{(2)}, \ldots, p^{(K)})$, where each element $p^{(l)}$ represents the probability of the subscriber's transmission in the l-th mini-slot of the n-th frame.

Assume there is an infinite queue buffer for the requests at the central base station. The central station serves the requests from the conducted queue according to some rule, which is referred to as *service discipline* $g^{(L)}$.

At the beginning of frame i the central station transmits grants for successfully received requests in frame $i-1$ indicating the slots for collision free packet transmission in frame i. Throughout this paper we assume, that a subscriber can not make more than one attempt to request a transmission per frame. This leads to the following restriction for considered algorithms. For any $f^{(K)}$: the weight of vector $\bar{\nu}_i(x)$ is either one or zero for any subscriber x and frame i.

In this part, both uplink and downlink channels are assumed to be error-free (noiseless). Neither packets nor requests will be distorted by noise. Error-prone channels are to be analyzed in Section VI. Situations in mini-slots are always correctly distinguished by the central station. Feedback vectors and slot allocation information is always successfully transmitted to all subscribers.

B. Definitions and Problem Statement

In this paper, we call the pair $(f^{(K)}, g^{(L)})$ the *multiple access protocol* for centralized reservation-based systems with parameters (K,L). Here, we introduce definitions analogous to those given previously for the basic RMA model, with extensions corresponding to our system. The time interval from the moment when a packet was generated to the moment it has been successfully transmitted is referred to as packet transmission delay. Then in some arbitrary but fixed frame (having number n) let an additional packet arrive in the system, whose transmission delay is denoted by $\delta_n(\lambda,K,L,f^{(K)},g^{(L)})$. According to the algorithm of the system operation the transmission delay consists of two components. The first one is the request delay for random access $\delta_n^{(1)}(\lambda,K,L,f^{(K)})$. It is the time from the moment of request generation, to the moment of the corresponding successful request transmission. The second one is the time from the moment of successful request transmission, to the time the corresponding packet will be successfully transmitted $\delta_n^{(2)}(\lambda,K,L,g^{(L)})$. We will refer to this value as queuing delay. The value $D(\lambda,K,L,f^{(K)},g^{(L)}) \triangleq \limsup_{n\to\infty} E\delta_n = \limsup_{n\to\infty} E(\delta_n^{(1)}+\delta_n^{(2)})$ for a given arrival rate λ, K mini-slots, L slots and multiple access protocol $(f^{(K)}, g^{(L)})$ will be referred to as the *mean delay of packet transmission*. Further, the mean request delay for the random access is defined as $D_1 \triangleq \limsup_{n\to\infty} E\delta_n^{(1)}$.

The maximal arrival rate (more precisely the supremum of the arrival rate), which can be transmitted by means of some multiple access protocol $(f^{(K)}, g^{(L)})$ for some frame structure (K,L), with finite mean delay $R(K,L,f^{(K)},g^{(L)}) \triangleq \sup_\lambda\{\lambda\colon D(\lambda,K,L,f^{(K)},g^{(L)}) < \infty\}$ will be referred to as *transmission rate (tenacity)* of the multiple access protocol.

If the multiple access protocol is not fixed, using our model, the *capacity* can be calculated as follows:

$$C(K,L,\mathcal{F}^{(K)},\mathcal{G}^{(L)}) \triangleq \sup_{\substack{f^{(K)}\in\mathcal{F}^{(K)} \\ g^{(L)}\in\mathcal{G}^{(L)}}} R(K,L,f^{(K)},g^{(L)}),$$

where $\mathcal{F}^{(K)}$ is the set of all RMA algorithms defined for the system with K mini-slots and $\mathcal{G}^{(L)}$ is the set of all service disciplines, which can be defined for the system with L slots.

Our aim is to compute the upper and lower bounds for the capacity $C(K,L,\mathcal{F}^{(K)},\mathcal{G}^{(L)})$, which will be presented in detail in Section IV.

IV. Capacity Analysis

Let us first consider only one part of the whole system operation, the request transmission during the reservation period, where actual data packet transmission is firstly not considered. This system is referred to as a *reduced* one. Then, transmission rate R_1 and capacity C_1 definitions analogous to those previously mentioned can be introduced for the reduced system, namely $R_1(K,L,f) \triangleq \sup_\lambda\{\lambda\colon D_1(\lambda) < \infty\}$ and $C_1(K,L,\mathcal{F}^{(K)}) \triangleq \sup_{f\in\mathcal{F}^{(K)}} R_1(K,L,f)$.

Then the following proposition is proved.

Proposition 4: If there are K mini-slots per frame then the capacity of the reduced system equals to $(C_0K)/(\alpha K+L)$, where C_0 is the capacity of the basic RMA system $(C_1(K,L,\mathcal{F}^{(K)}) = C_0K/(\alpha K+L))$.

Proof: It is easy to notice that for $K=1$, when each frame consists of only one mini-slot we have exactly the basic RMA system, for which vectors $\bar{\theta}_i$, $\bar{\nu}_i(x)$ and the output of function f turn to scalars. Thus, $\mathcal{F}^{(1)} = \mathcal{F}_0$. Since $\mathcal{F}_0^{(K)} = \mathcal{F}^{(K)}$ for $K \geq 2$, we have the basic RMA system with slots grouped into segments of length K (as it is explained in Section II), whose capacity is proved to be C_0 in Proposition 3. The only difference is that one "slot", which is used in the basic system corresponds to one frame of length $(\alpha K+L)$ in our reduced system, what is taken into account by means of corresponding normalization. ■

Now we are finishing with the analysis of the reduced system and consider the overall reservation model. Below are two necessary conditions for the system stability.

Proposition 5: The mean request delay for the random access D_1 and the mean delay of packet transmission D may be finite if the inequality

$$\lambda(\alpha K+L) < C_0K \tag{1}$$

holds.

Proof: From proposition 4 it directly follows that the request delay for the random access D_1 is infinite if the arrival rate does not satisfy $\lambda < C_0K/(\alpha K+L)$. Obviously, the same is valid for the mean delay D. ■

Proposition 6: Let the arrival rate λ be chosen such that the request delay for the random access D_1 is finite. Then, the mean delay of packet transmission D may be finite if inequality

$$\lambda(\alpha K+L) < L \tag{2}$$

holds.

Proof: Generation and transmission of packets can be described in terms of queueing theory ([12]). We have Poisson packet arrivals with rate $\lambda(\alpha K+L)$ per frame. On the other hand not more than L packets can be transmitted per frame using any service discipline $g^{(L)}$. Thus this queuing system is unstable if (2) does not hold. ■

Now we will construct the upper bound for the system capacity C.

Proposition 7: For a given mini-slot length α, the inequality

$$\max_{K,L} C(K,L,\mathcal{F}^{(K)},\mathcal{G}^{(L)}) \leq \frac{1}{1+\alpha/C_0}, \tag{3}$$

holds for the capacity of centralized reservation-based RMA systems.

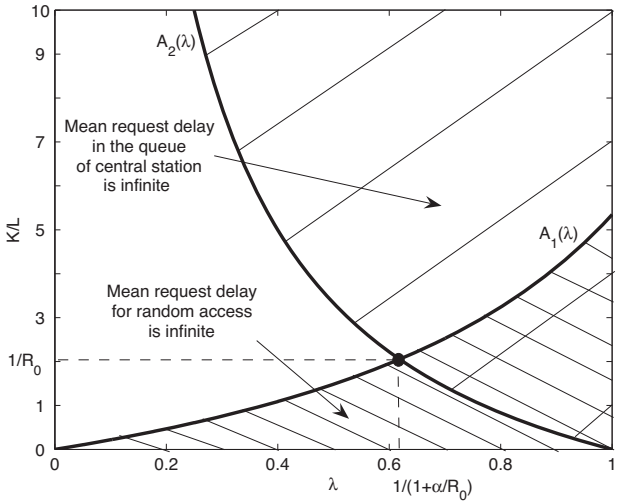


Fig. 2. Areas of instability of random multiple-access protocol $(A_1(\lambda) = \lambda/(R_0-\alpha\lambda)$, $A_2(\lambda) = (1-\lambda)/(\alpha\lambda))$

Proof: Since from Proposition 5, the mean delay of packet transmission may be finite if $\lambda(\alpha K+L) < C_0K$, we easily obtain that it may be finite if arrival rate λ satisfies

$$\lambda < \frac{C_0\frac{K}{L}}{\alpha\frac{K}{L}+1}. \tag{4}$$

On the other hand, from Proposition 6, the mean delay of packet transmission may be finite if $\lambda(\alpha K+L) < L$, hence it may be finite if λ satisfies

$$\lambda < \frac{1}{\alpha\frac{K}{L}+1}. \tag{5}$$

From (4) and (5) we obtain that

$$\lambda < \min\left(\frac{C_0\frac{K}{L}}{\alpha\frac{K}{L}+1}, \frac{1}{\alpha\frac{K}{L}+1}\right),$$

which leads to $\max_{K/L} C(K,L,\mathcal{F}^{(K)},\mathcal{G}^{(L)}) = \frac{1}{\alpha/C_0+1}$ for $K/L = 1/C_0$ and proves (3). Derived areas of instability for RMA protocol are illustrated in Figure 2. ■

Finally, let us construct a lower bound for the system capacity C. For this purpose, we consider the part-and-try RMA algorithm, which, as previously mentioned, is the fastest one known for the basic model. From Proposition 2 it follows that an algorithm exists in class $\mathcal{F}^{(K)}$, which has exactly the same transmission rate. Moreover, an explicit way to construct it is provided in the proof of Proposition 2. Let us denote this RMA algorithm as $\phi^{(K)}$. Then the following proposition can be proven.

Proposition 8: In the centralized reservation-based RMA system, let $\phi^{(K)}$ RMA algorithm and first-input-first-output (FIFO) service discipline (denoted as $\varphi^{(L)}$) be used. Then maximal transmission rate of multiple-access protocol $(\phi^{(K)},\varphi^{(L)})$ for all K and L can be made arbitrary close to $\frac{R_{pt}}{\alpha+R_{pt}}$, where R_{pt} is the transmission rate of the part-and-try-algorithm.

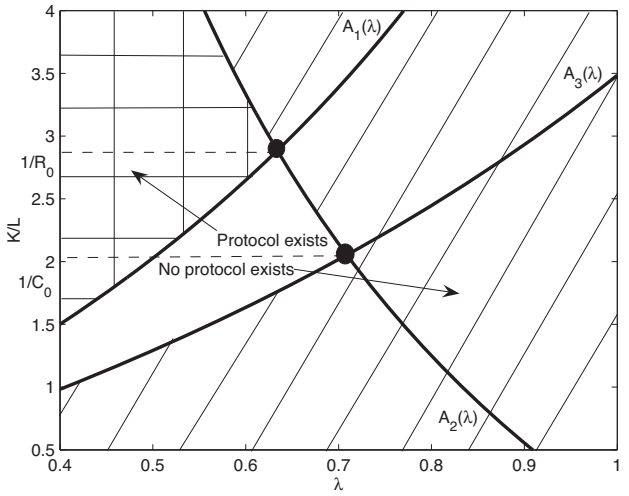


Fig. 3. Capacity bounds of random multiple-access system (A_1 and A_2 are defined in Fig. 2, $A_3(\lambda) = \lambda/(C_0 - \alpha\lambda)$)

Proof: One can show that the necessary and sufficient condition for the mean request delay to be finite, is

$$\lambda(\alpha K + L) < R_{pt} K. \tag{6}$$

Let λ justify Condition (6). Then, the central station queue becomes a $G/D/L$ FIFO queuing system. The input traffic represents the outcome of K basic RMA systems, where subscribers operate independently according to the part-and-try algorithm. One can show that for this queuing system, the Baccelli-Foss conditions [12] are satisfied. Therefore,

$$\lambda(\alpha K + L) < L. \tag{7}$$

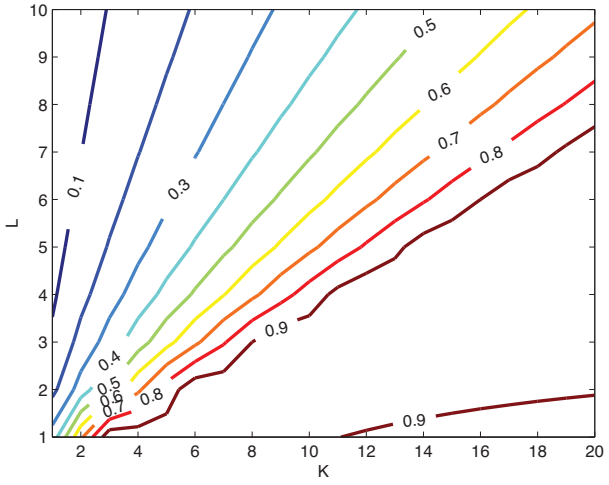
is the necessary and sufficient condition, that mean packet delay in the queue is finite.

From Conditions (6) and (7), and using an approach analogous to the one used in the proof of Proposition 7, we obtain that mean packet delay is finite if and only if both $\lambda < \frac{R_{pt}\frac{K}{L}}{\alpha\frac{K}{L}+1}$ and $\lambda < \frac{1}{\alpha\frac{K}{L}+1}$ hold. Taking into account the fact that for any $\epsilon > 0$, a pair (K, L) exists for which $|K/L - 1/R_{pt}| < \epsilon$, the proposition is proven. ■

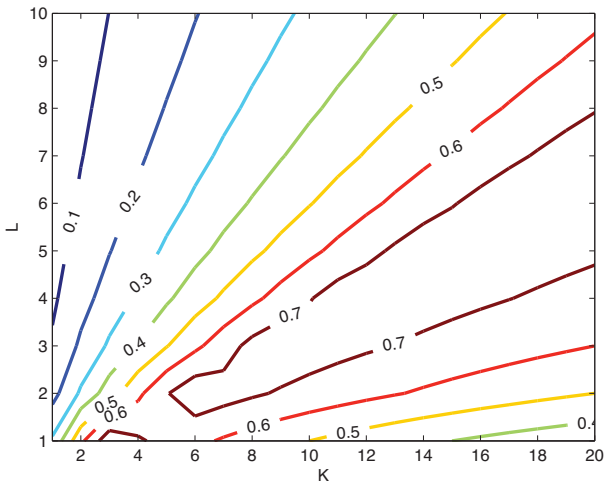
From the proof of this proposition the *corollary* directly follows: the maximal transmission rate of multiple-access protocol $(\phi^{(K)}, \varphi^{(L)})$ is achieved, when $\frac{K}{L} \approx \frac{1}{R_{pt}}$. The capacity bounds derived in Propositions 8 and 9 are illustrated in Figure 3.

We introduced the upper and lower bounds for Tsybakov's capacity of centralized reservation-based RMA system. If some "rational" algorithm $f^{(K)}$ having transmission rate R_0, which is independent of K, and some "simple" service discipline $g^{(L)}$ (like FIFO), are implemented, then the transmission rate of this multiple-access protocol is $R = \min(\frac{R_0 K}{\alpha K+L}, \frac{L}{\alpha K+L})$ and maximized, when $\frac{K}{L} \approx \frac{1}{R_0}$.

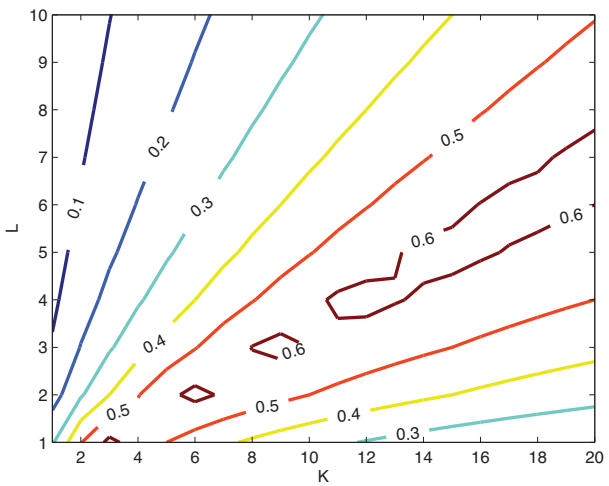
In contemporary IEEE 802.16 WiMAX network a version of the so-called binary exponential back-off (BEB) RMA algorithm is used for bandwidth requests [2]. This algorithm is shown to have zero transmission rate for infinite-users basic RMA model in [13]. For a finite, but fairly large number of users, $\ln(2)/2$ can represent some analog of the transmission



(a) -20pt$\alpha = 0.01$

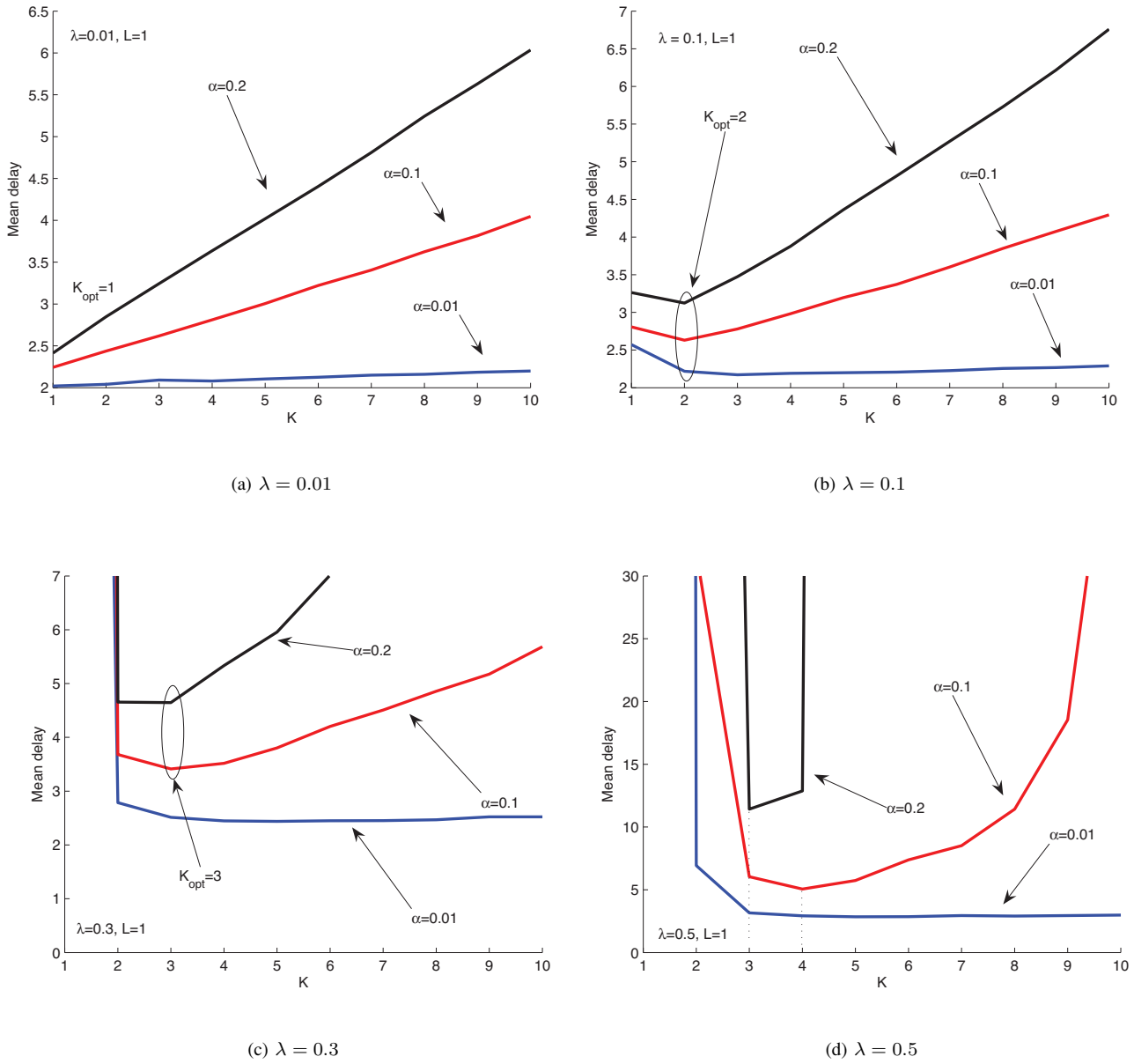


(b) $\alpha = 0.1$



(c) $\alpha = 0.2$

Fig. 4. Theoretical transmission rate bounds for IEEE 802.16 MAC protocol. It is assumed that BEB has finite transmission rate.



(a) $\lambda = 0.01$

(b) $\lambda = 0.1$

(c) $\lambda = 0.3$

(d) $\lambda = 0.5$

Fig. 5. Total mean delay for L=1 and different α values.

rate [14]. With this value, the theoretical transmission rate bounds for the IEEE 802.16 MAC, are depicted in Figure 4. Areas on the plane (K, L) indicate the achievable protocol transmission rates for different α.

V. Mean Delay Analysis

We implemented our simulation model in Matlab (explained in [16] and [17]) to estimate the mean delay of the WiMAX MAC protocol with a finite number of subscribers and using the BEB algorithm. The sercive discipline is FIFO.

We use the following *hypothesis* for estimating the minimal mean delay[2]: the ratio K/L, which minimizes the mean packet delay value $D(\lambda, K, L, f^{(K)}, g^{(L)})$, is a non-decreasing function of arrival rate λ and for any α, values of this function lie in a narrow interval not wider than $[1, 1/R_0]$. Moreover, mean delay itself is minimized, when K and L are minimal among those satisfying optimal ratio K/L. Thus, taking into account our hypothesis, frame structure can be optimally designed and is *almost independent of the ratio between the duration of request and packet transmission*. In the following we validate our hypothesis by means of simulations.

If our hypothesis is valid then the performance of the system is maximized, when $K/L \in \{1, 2, 3\}$ (note that $2/\ln(2) \approx 3$) and never using larger values of this ratio is reasonable. Thus, if we need the simplicity of implementation it may be reasonable to keep $K/L = 3$ always. Now we would like to check the feasibility of this approach. For simplicity we provide the results of the experiments for $L = 1$ (although,

[2]Computation of the mean packet delay in the centralized reservation-based RMA system for the general case is an open question and is out of the scope of this paper.

similar results may be obtained for the $L > 1$ case). The following values of the parameters were used: number of users $n = 50$, BEB parameters $l = 1$ and $m = 10$ [16]. Transmissions during 2×10^4 frames have been simulated (Figure 5, cases a-d). We observe that:

a) For a small arrival rate, e.g. $\lambda = 0.01$, setting $K = 1$ minimizes the mean delay independent of α.

b) For $\lambda = 0.1$, the optimum is $K = 2$ independent of $\alpha \in \{0.01, 0.1, 0.2\}$.

c) For $\lambda = 0.3$, the optimum is $K = 3$ for long mini-slot length $\alpha \in \{0.1, 0.2\}$ and $K = 5$ for the short mini-slot length $\alpha = 0.01$. However, the mean delay for $K = 3$ is not significantly larger.

d) For $\lambda = 0.5$, the optimum is $K = 3$ for $\alpha = 0.2$ and $K = 4$ for $\alpha = 0.1$. For $\alpha = 0.01$ the delay stays almost the same for $3 \leq K \leq 10$. We clearly see two asymptotes of the delay function that correspond to the theoretically derived capacity bounds.

We now depict the relationship between λ and optimal value of K (which minimizes the mean delay - denote K_{opt}) for different α (Figure 6, upper). Also we calculate

$$\Delta = \frac{|D(\lambda;K_{opt};1;BEB;FIFO) - D(\lambda,3,1,BEB,FIFO)|}{D(\lambda,3,1,BEB,FIFO)}$$

which indicates the relative mean delay difference, when K is chosen optimally and when K is set to 3 (Figure 5, cases (c) and (d)). First, we may see that for our scenario hypothesis is not valid, because, for instance, $\alpha = 0.01$ function $K_{opt}(\lambda)$ is not monotone-increasing having values from the interval $[1, 3]$, but has maximum for arrival rate 0.5, with optimal K equals to 7. However, remember that the hypothesis is stated for the infinite subscribers model, but we have simulated a system with 50 subscribers, only. If we increase the number of subscribers to $n = 500$, the function $K_{opt}(\lambda)$ behaves significantly smoother for $\alpha = 0.01$ (Figure 7, a) and is monotone increasing for $\alpha \in \{0.1, 0.2\}$ (Figure 7, b,c). This is a clear indication that the hypothesis is valid in the extreme case of infinite n. The second observation is, that we loose from the mean delay point of view, when K is set to 3 for high λ values if $n = 50$ and $\alpha = 0.01$. However, this degradation decreases as n increases. Here, it should be noticed, that RMA will be used only when arrival rates are small, while for large λ, polling in TDMA fashion should be used. If, for example, $\lambda < 0.5$ the delay lose, when $K = 3$ instead of optimal value K_{opt} is used, will not exceed 10%. For α equals to 0.1 and 0.2 the increased delay occurs for small arrival rates only and does not exceed 25%. The overall conclusion from the $L = 1$ series of experiments is, that if α is rather small (e.g 0.01), like in the IEEE 802.16 protocol, it is reasonable to set $K = 3$ always. If α is larger (e.g. 0.1) it may be reasonable to choose K from $\{1, 2, 3\}$ depending on the arrival rate. Once again, remember, that this conclusion is valid for "typical" RMA usage scenarios namely large number of subscribers and small arrival rates.

VI. Influence of Channel Errors

In the previous sections, an error-free RMA channel is assumed. This assumption of Tsybakov's model is first relaxed by Evseev in [19] as well as by Vvedenskaya in [20], where the so-called false collision model is introduced. In this

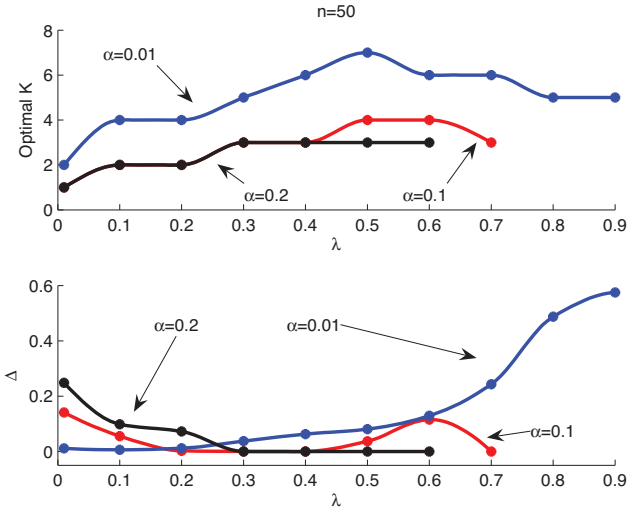


Fig. 6. Values of K, which minimize the mean delay and relative difference between delay value, when K=3 and optimal delay value for different α.

section, we generalize the results obtained in the previous sections for the case of an error-prone channel.

Due to potential noise in the wireless channel, base station makes mistakes when determining the actual channel situations. The false collision probability decision for a mini-slot can be calculated by

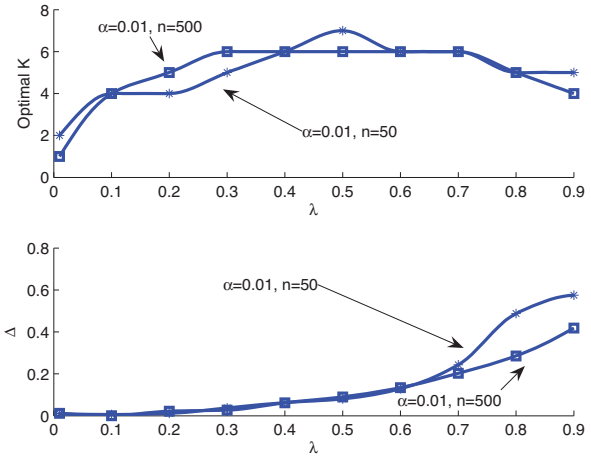
$$q = Pr\{\zeta_i^{(l)} = 2 | \theta_i^{(l)} = 0\} = Pr\{\zeta_i^{(l)} = 2 | \theta_i^{(l)} = 1\} \quad (8)$$

and false collisions in different mini-slots are assumed to be statistically independent. Thus, in order to take into account an error-prone channel in all previous discussions, the feedback vector $\bar{\zeta}_i = (\zeta_i^{(1)}, \zeta_i^{(2)}, \ldots, \zeta_i^{(K)})$ should be used instead of $\bar{\theta}_i$, where the variable $\zeta_i^{(l)} \in \{0, 1, 2\}$ corresponds to the decision of the base station about empty channel, successful transmission or conflict in the l-th mini-slot of the $(i-1)$-th frame.

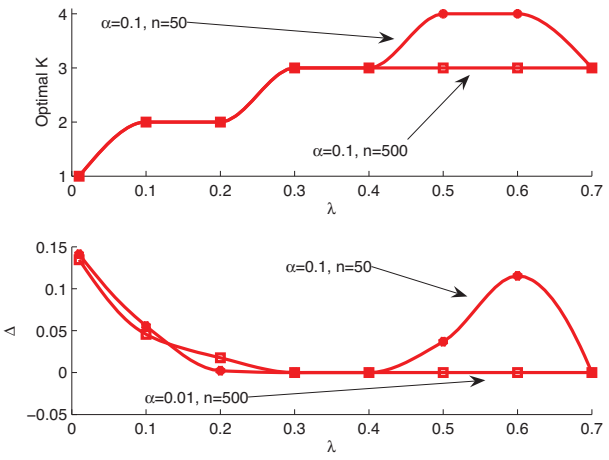
Moreover, we assume that the probability Q for a packet to be distorted by noise is $0 \leq Q < 1$ (in real systems it can be assumed that $Q > q$) and the events corresponding to the packet's distortion are statistically independent. Furthermore, a noiseless downlink channel is assumed. Let the subscribers know about the success/failure result of their transmitted packets in the current frame by the beginning of the next frame. Packets are retransmitted until their successful transmission. Feedback vectors and slot allocation information are always successfully transmitted in the downlink to all the subscribers.

Therefore, we have now two more parameters in our model: (q, Q). All definitions (RMA algorithm, transmission rate, capacity, etc.) can be easily extended for the case of an error-prone channel. Core propositions from Section IV can be modified for the case of an error-prone channel as follows.

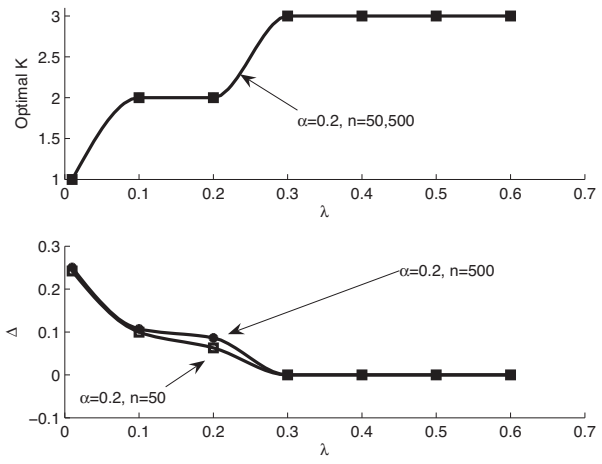
Proposition 9 (error-prone channel case of Proposition 5): The mean request delay D_1 for the random access phase and the mean delay D of the packet transmission may be finite if the inequality $\lambda(\alpha K + L) < C_0(q)K$ holds, where $C_0(q)$ is the basic RMA system's capacity in the error-prone channel. For the case of an error-prone channel with $0 \leq q < 1$, an



(a) $\alpha = 0.01$



(b) $\alpha = 0.1$



(c) $\alpha = 0.2$

Fig. 7. Values of K, which minimize the mean delay and relative difference between delay value, when K=3 and optimal delay value for different n.

upper bound for the capacity was constructed by Tsybakov and Likhanov in [21].

Proposition 10 (error-prone channel case of Proposition 6): Let the arrival rate value λ be chosen such that the request delay for the random access D_1 is finite. Then, the mean delay of the packet transmission in the system D may be finite if the inequality $\lambda(\alpha K + L) < L(1 - Q)$ holds.

Proposition 11 (error-prone channel case of Proposition 7): For a given α value, inequality

$$\max_{K,L} C(K, L, F^{(K)}, G^{(L)}, q, Q) \leq \frac{1}{\frac{\alpha}{C_0(q)} + \frac{1}{1-Q}}$$

holds for the capacity of centralized reservation-based RMA systems in the noisy channel.

Consider the fast tree RMA algorithm from [22], which provides a non-zero transmission rate for any probability $0 \leq q < 1$ (we will refer to this algorithm as "noise-resistant tree algorithm"). It can be shown that an algorithm exists in class $F^{(K)}$, which has exactly the same transmission rate. Let us denote this RMA algorithm as $\Phi^{(K)}$.

Proposition 12 (error-prone channel case of Proposition 8): Let the $\Phi^{(K)}$ algorithm and a first-input-first-output (FIFO) service discipline (denoted as $\phi^{(L)}$ as before) be used. Then, the maximal transmission rate of multiple-access protocol $(\Phi^{(K)}, \phi^{(L)})$ for all K and L can be made arbitrary close to $R(q)/(\alpha + R(q))$, where $R(q)$ is the maximal transmission rate of the noise-resistant tree algorithm for a given q.

Here we omit the detailed proofs due to the page limit. Note that there are no fundamental difficulties in integrating the error-prone channel into our model. Therefore, if we consider the error-prone channel case, from practical point of view it is reasonable to keep the ratio K/L constant and approximately equal to $(1 - Q)/R_0(q)$, where $R_0(q)$ is the rate of the used RMA algorithm in the error-prone channel case.

VII. Conclusion

In this paper, the method to estimate the upper and lower capacity bounds of centralized reservation-based random multiple access systems is developed. It is shown that the *maximal transmission rate* of a reservation-based multiple access protocol is equal to $1/(1+\alpha/R_0)$ and it is achieved when the *ratio* between the number of mini-slots (K) for bandwidth request transmission and the number of slots (L) for data packet transmission equals to the reciprocal of the transmission rate of the used random multiple access algorithm ($1/R_0$). Specifically, in the case of IEEE 802.16 MAC with a large number of subscribers, it is shown that from both capacity and delay points of view, it is reasonable to keep the ratio constant ($K/L = 3$), independently of α and application-level data arrival rate value.

Our future research will include: a) to investigate a reservation-based random multiple access system with TDMA used for the reservation; b) to consider multiple-packets messages transmissions; c) to consider multi-cell situations.

References

[1] I. Rubin, "Access-Control Disciplines for Multi-Access Communication Channels: Reservation and TDMA Schemes," *IEEE Trans. Inform. Theory*, Vol. IT-25, No. 25, pp. 516–538, September 1979.

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

[3] B. S. Tsybakov and V. A. Mikhailov, "Free synchronous packet access in a broadcast channel with feedback," *Problems of Information Transmission*, vol. 14, no. 4, pp. 259–280, October–December 1978.
[4] J. I. Capetanakis, "Tree algorithm for packet broadcasting channel," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 505-515, September 1979.
[5] B. S. Tsybakov, "Survey of USSR Contributions to Random Multiple-Access Communications," *IEEE Trans. Inform. Theory*, Vol. IT-31, No. 2, pp. 143–165, March 1985.
[6] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1st ed., 1987; 2nd ed., 1992.
[7] B. S. Tsybakov and M. A. Berkovskii, "Multiple Access with Reservation," *Problems of Information Transmission*, Vol. 16, No. 1, pp. 35–54, January–March 1980.
[8] A. Ephremides and B. Hajek, "Information Theory and Communication Networks: An Unconsummated Union," *IEEE Trans. Inform. Theory*, Vol. 44, No. 6, pp. 2416–2434, October 1998.
[9] B. S. Tsybakov and N. B. Likhanov, "Upper Bound on the Capacity of a Random Multiple-Access System," *Problems of Information Transmission*, Vol. 23, No. 3, pp. 224–236, July–September 1987.
[10] B. S. Tsybakov and V. A. Mikhailov, "Random Multiple Packet Access: Part-and-Try Algorithm," *Problems of Information Transmission*, Vol. 16, No. 4, pp. 305–317, October–December 1980.
[11] B. Hajek, N. B. Likhanov and S. Tsybakov, "On the Delay in a Multiple-Access System with Large Propagation Delay," *IEEE Trans. Inform. Theory*, Vol. 40, No. 4, pp. 1158–1166, July 1994.
[12] F. Baccelli and S. Foss, "On the Saturation Rule for the Stability of Queues," *Journal of Applied Probability*, 32, 2, pp. 494–507, 1995.
[13] D. Aldous,"Ultimate Instability of Exponential Back-off Protocol for Acknowledgment-based Transmission Control of Random Access Communication Channels," *IEEE Trans. Inform. Theory*, Vol. 33, No. 2, pp. 219–233, March 1987.
[14] N.-O. Song, B.-J. Kwak and L. E. Miller, "On the Stability of Exponential Backoff," *J. Research of the National Institute of Standards and Technology*, Vol. 108, No. 4, pp. 289–297, July-August 2003.
[15] A. Turlikov and A. Vinel, "Capacity Estimation of Centralized Reservation-Based Random Multiple-Access System," *Proc. of the XI International Symposium on Problems of Redundancy in Information and Control Systems*, SUAI, Saint-Petersburg, July 2007, pp. 154–160.
[16] A. Vinel, Y. Zhang., M. Lott, A. Turlikov, "Performance Analysis of the Random Access in IEEE 802.16," *Proc. of the 16th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications - IEEE PIMRC'05*, Berlin, Germany, 2005, pp. 1596–1600.
[17] A. Vinel, Y. Zhang, Q. Ni, A. Lyakhov, "Efficient Request Mechanisms Usage in IEEE 802.16," *Proc. of 49th IEEE Global Telecommunications Conference - GLOBECOM'06*, San Francisco, California, USA, 2006.
[18] A. Vinel and V. Vishnevsky, "Analysis of Contention-Based Reservation in IEEE 802.16 for the Error-Prone Channel," *1st International Workshop on Multiple Access Communications*, Saint-Petersburg, Russia, June 2008.
[19] G. S. Evseev and N. G. Ermolaev, "Performance Evaluation of the Collision Resolution for a Random-Access Noisy Channel," *Problemy Peredachi Informatsii*, Vol. 18, No. 2, pp. 101-105, April-June 1982 (Russian issue).
[20] N. D. Vvedenskaya and B. S. Tsybakov, "Random Multiple Access of Packets to a Channel with Errors," *Problems of Information Transmission*, Vol. 19, No. 2, pp. 131-146, April-June 1983.
[21] B. S. Tsybakov and N. B. Likhanov, "Upper Bound on the Capacity of a Packet Random Multiple Access System with Errors," *Problems of Information Transmission*, Vol. 25, No. 4, pp. 297-308, October-December 1989.
[22] G. S. Evseev and A. M. Turlikov, "Throughput Analysis for a Noise-Resistant Multiple Access Algorithm," *Problemy Peredachi Informatsii*, Vol. 22, No. 2, pp. 104-109, April-June 1986 (Russian issue).

Alexey Vinel (M'07) is a senior researcher of Saint-Petersburg Institute for Informatics and Automation (Russian Academy of Sciences). He received his Bachelor (2003) and Master (2005) degrees in information systems from Saint-Petersburg State University of Aerospace Instrumentation and his Ph.D. (2007) degree in technical sciences from Institute for Information Transmission Problems (Russian Academy of Sciences). He is the fellow of Alexander von Humboldt Foundation and founder of International Workshop on Multiple Access Communications (MACOM). His research interests include random multiple access algorithms and performance evaluation of wireless networks.

Qiang Ni (M'04) is a faculty member in the School of Engineering and Design, Brunel University, West London, United Kingdom, where he heads the Intelligent Wireless Communication Networking Team. Prior to that, he was a Senior Researcher at Hamilton Institute, National University of Ireland Maynooth. His research interests are wireless networking and mobile communications. He has published over 40 refereed papers in the above fields. He worked with INRIA France as a Researcher for 3 years (2001-2004). He received his Ph.D. degree from Huazhong University of Science and Technology (HUST), China. Since 2002 he has been active as an IEEE 802.11 wireless standard working group Voting Member, and a contributor to the IEEE wireless standards.

Dirk Staehle is an assistant professor at the Chair of Distributed systems at the University of Wuerzburg, Germany. He received his doctoral degree (PhD) from the University of Wuerzburg in 2004. He is leading the department's mobile network research group (MNRG). He functions as chairman for the Traffic Engineering working group of the COST 290 action of the European Union entitled Traffic and QoS Management in Wireless Multimedia Networks. His research interests include analytic modeling of WCDMA networks, UMTS radio network planning, source traffic modeling of wireless applications, integration of mobile communication systems with heterogeneous radio access technologies, and capacity evaluation and deployment scenarios of WIMAX networks. He has currently lead multiple industry co-operations in the field of GPRS and UMTS radio network planning with T-Mobile International, France Telecom R&D, and Vodafone Netherlands (former Libertel).

Andrey Turlikov is a professor at Department of Information Systems and Data Protection of Saint-Petersburg State University of Aerospace Instrumentation, Russia. Since 1987 he has been involved in teaching activity. He is the author of about 80 research papers and has been the invited speaker at the number of symposiums and seminars. His research interests cover multi-user telecommunication systems, real-time data transmission protocols, theory of reliability and video compression algorithms.