Throughput and Fairness in IEEE 802.16 Broadband Wireless Networks

Radostin A. Pachamanov[1] and Boris P. Tsankov[2]

***Abstract* – The IEEE 802.16 standard defines the air interface specifications for broadband access in wireless metropolitan area networks. These networks support various services that require specific data throughput. In order to maintain appropriate quality of service (QoS), the network should perform an adequate resource allocation. Optimizing the network throughput as a primary objective may lead to unfairness from the users' point of view. In this paper we investigate the "fair" resource allocation problem. We propose an optimization-based approach for resource allocation, which takes into consideration fairness issues.**

***Keywords* – IEEE 802.16, resource allocation, quality of service, throughput, fairness.**

I. INTRODUCTION

The newest generation of Internet connectivity is the broadband wireless access (BWA) technology, based on the IEEE 802.16 standard, also known as WiMAX (Worldwide Interoperability for Microwave Access) [1]. The main advantages of IEEE 802.16 wireless access are high transmission rate, flexibility for the connection, and the ability for pre-defined quality of service (QoS) framework. WiMAX, as a wireless solution, is a promising alternative for "last mile" access in crowded urban or suburban areas where installation of cable-based infrastructure is economically or technically infeasible.

The WiMAX network is able to provide a wide range of services that are varied in their nature and demand different performance levels in order to maintain an appropriate QoS. The mechanisms for providing and granting QoS in WiMAX are through connection admission control (CAC) and resource allocation. CAC is used to limit the number of connections in the network in order to maintain the pre-defined QoS parameters of the applications that are already accepted for service. If the available resources could be allocated among the ongoing and the incoming connections so that the QoS requirements of both types of connections could be fulfilled and maintained at the target level, the new connections are accepted. Through a resource allocation schemes the applications that require better throughput are granted more bandwidth. The resource allocation scheme should take into consideration the fact that in wireless networks the transmission rate is affected by channel conditions and also depends on allocated power [2]. It needs to take into consideration not only the different rates and QoS requirements of the applications, but the SSs' positioning in the cell (the distance to the BS, SINR, modulation used, etc.). Although the IEEE 802.16 standard comprises the signaling for the multiple access, the algorithms for CAC and resource allocation remain open issues.

There are several articles published on the resource allocation [3-5] and CAC problems [6-7]. They investigate different architectures and scheduling algorithms to guarantee QoS. A very interesting approach is the formulation of an optimization problem having as an objective function QoS parameters such as delay, loss, throughput, etc. [5-7].

In this paper, we investigate the resource allocation problem in order to obtain maximum network throughput. But maximizing the network throughput as a primary objective may lead to "unfair" resource allocation. Within a wireless network, two different aspects of fairness can be distinguished. First (and this is typical for every network serving simultaneously traffic flows with different traffic characteristics), the applications with less stringent QoS requirements may be neglected compared to high priority traffic. This problem has been broadly studied in recent years with focus on IP networks, and different schemes for providing fairness have been introduced [8-9]. A new and specific problem with fair resource allocation appears in IEEE 802.16 wireless networks. Since the SSs distant from the BS use different modulation schemes, the throughput achieved for the same amount of bandwidth may differ. Therefore, the base station may refuse service to remote subscribers in order to maximize network throughput. To the best of our knowledge, this problem has not yet been discussed in the literature. Considering these two aspects, we propose an optimization-based joint connection admission control and resource allocation scheme for IEEE 802.16-based wireless networks where a fairness element is implemented. Our scheme takes into account traffic characteristics of the ongoing and incoming connections such as mean data queue length, maximum allowed mean data delay, and mean packet arrival rate. We formulate a corresponding optimization problem where we assume that each user is assigned a throughput-dependent utility function. This function, which may be different for different SSs, represents the utility of the corresponding throughput achieved and implements the fairness concept. Our goal is to optimize the total utility over all users.

Optimization of such a network is a complicated task since it is a complex problem consisting of joint throughput and power allocation optimization. One possible approach is to

[1]Radostin A. Pachamanov is with the Dept. of Telecommunications, Technical University of Sofia, "Kl.Ohridski" Blvd. No8, Sofia 1000, Bulgaria, e-mail: radostin_ap@tu-sofia.bg

[2]Boris P. Tsankov is with the Dept. of Telecommunications, Technical University of Sofia, "Kl.Ohridski" Blvd. No8, Sofia 1000, Bulgaria, e-mail: bpt@tu-sofia.bg

split the problem in two parts [9]. According to [10], the gain obtained from throughput optimization is bigger than the gain from power optimization. Therefore, we assume that the power is equally spread over all slots and solve the slot allocation problem, considering it as a part of the global optimization problem.

The rest of the paper is organized as follows. Section II gives a brief overview of IEEE 802.16 specifics taken in consideration. Section III presents the fairness problem with resource allocation in a cell structure. Section IV presents the system model and the formulated optimization problem. Section V concludes with final remarks.

II. An Overview of the Considered IEEE 802.16 Specifics

The IEEE 802.16 standard, or so called WirelessMAN, is an air interface standard that defines the first two network layers (physical-PHY and data-MAC) [1] of the OSI model. The IEEE 802.16 system architecture consists of two logical entities – a base station (BS) and a subscriber station (SS).

The medium access control layer (MAC) is structured to support multiple PHY specifications, depending on the particular operational environment. The physical layer operates at 10-66 GHz, or 2-11 GHz frequency band, where for the latter propagation without direct line-of-sight is possible. For frequencies below 11 GHz, three alternatives for PHY specifications are provided: Orthogonal Frequency Division Multiplexing (OFDM), Orthogonal Frequency Division Multiple Access (OFDMA) and Single-Carrier (SC). In our investigations we consider OFDMA, since it is proposed for the mobile version IEEE 802.16e of the standard. To maximize the spectral efficiency of the air link, each specification uses multilevel modulation scheme. The modulation is optimized for each SS based on the quality of the radiofrequency channel. If link conditions permit, a more efficient modulation is used to maximize the tradeoff between bandwidth and robustness. The supported modulations are QPSK, 16-QAM and 64-QAM.

The MAC protocol is performed in a way that provides high data rates in both uplink (UL) and downlink (DL) directions, and comprises medium access admission and bandwidth allocation algorithms. The protocol is connection-oriented and all data communications for both transport and control are in the context of a unidirectional connection. BS is responsible for coordinating the access of SSs to the medium. Data transmission is organized in framed format. Separate subframes are used for uplink and downlink directions. There are two ways supported by IEEE 802.16 for separating uplink from downlink transmissions: frequency division duplexing (FDD) and time division duplexing (TDD). In FDD, downlink and uplink are performed at different frequencies, and therefore may overlap in time. TDD divides time into uplink and downlink transmission periods. In both schemes the frame has fixed duration. We consider TDD since it has the advantage that the network can adjust the size of the UL and DL subframes within the frame depending on the traffic load in the corresponding direction.

In order to provide a QoS-based network operation, the IEEE 802.16 standard groups the applications with similar QoS requirements and traffic characteristics into a small number of classes, named *scheduling services*. Each scheduling service is tailored to support specific class of application. Four types of services are defined.

Unsolicited grant service (UGS): Supports constant-bit-rate (CBR) real time traffic with stringent QoS requirements, such as voice over IP. The medium is granted on a periodic basis.

Real-time polling service (rtPS): Supports various-bit-rate real time traffic. The amount of bandwidth required for this type of service is determined based on the required QoS parameters, the channel quality and the traffic arrival rates of the sources. An example for such kind of applications is VoIP with silence suppression.

Non real-time polling service (nrtPS): Supports traffic with less stringent QoS requirements. This is suitable for applications such as file transfer. The bandwidth allocation is also adaptive as in the case of rtPS.

Best-effort service (BE): Supports best-effort traffic with no QoS guarantee.

In our model we focus the latter three types of services. We assume that a certain amount of bandwidth is granted for the UGS, and the BS has to allocate the rest among the other three types.

III. Fairness in IEEE 802.16-Based Networks

In telecommunications the term “fairness” is used to represent a criterion for distributing the available network resources among competing traffic flows – the resources should be “fairly” distributed among the subscribers in such a way that everybody is satisfied with the achieved QoS. Generally, policies for resource sharing that are characterized by low level of fairness provide high average throughput, but low stability in the service quality, meaning that the achieved service quality is varying in time, depending on the behavior of other users. If this instability is a frequent event, it may result in displeased customers that may choose another, more stable communication network.

In a simple vision, fairness may be interpreted as allocating the same share of bandwidth to all. However, in a case of a network supporting a wide range of applications with various QoS requirements, such a simple view does not make sense.

As we mentioned above, two different aspects of fairness following the specifics of the wireless networks can be distinguished. The network allocates the resources depending on the type of *scheduling service* to which the application belongs. The applications from classes with lower priority may suffer. Therefore, a special scheme for resource distribution is needed.

The second important and specific for IEEE 802.16 aspect concerns the throughput that can be achieved with a certain amount of bandwidth (we consider OFDMA where a number of subcarriers are dedicated to a subscriber for a certain time period). The SSs are randomly situated within the cell and the distance between a particular SS and the BS may vary. Depending on the wireless link condition, SINR, the distance, etc., a different modulation scheme may be used. Some

modulation schemes are more efficient than others, and the same amount of bandwidth can deliver better throughput. An example is given in Fig.1.

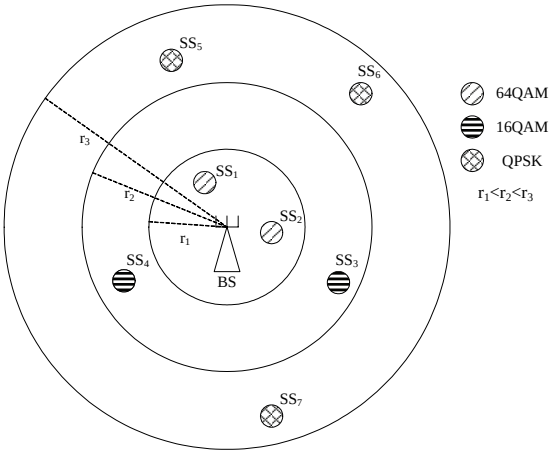

Fig.1. SSs working with different modulation schemes

It is natural for a network operator to focus on maximizing the network throughput, since in that way the system utilization is increased. However, maximizing the throughput without taking into consideration the fairness issues described above may worsen the network operation. A simple way to evaluate the level of fairness achieved is through the Fairness Index (FI), proposed in [12]:

$$FI = \left(\sum_{i=1}^{N} r_i \right)^2 / \left(N \sum_{i=1}^{N} r_i^2 \right), \qquad (1)$$

where N denotes the number of active SSs, and r_i represents the resource portion allocated to, or throughput achieved by, user i. Since the applications have different traffic characteristics, they require different resource portions to achieve certain QoS level. Therefore, we define r_i as a "surplus" resource portion/throughput of user i, and $r_i = a_i - a_i^{\min}$, where $a_i^{\min}$ denotes the minimum required data throughput for user i. If all users get the same surplus throughput, then FI will be 1, and the system will be 100% fair.

Three fairness criteria are presented in the literature – max-min fairness, proportional fairness, and utility fairness [8-9]. Max-min fairness puts emphasis on maintaining high values for the smallest rates. Proportional fairness is an alternative definition and provides fairness on a proportion basis. In the "utility" fairness approach, every user has a utility function that indicates the value to that source to have a certain throughput/rate. In our model, we consider utility fairness, since it is a more general concept, and comprises both max-min and proportional fairness.

IV. THROUGHPUT OPTIMIZATION

Our system model consists of a BS that serves incoming and outgoing traffic from and to a certain number of SSs. We assume that every SS has a corresponding queue in the BS. Let us denote the number of active SSs in the system by *N*. We assume that the BS is responsible for both UL and DL and follows the state of *2N* queues, since each user/application has separate queues for UL and DL directions. Henceforth, by user/application we will mean the traffic flow that enters the corresponding queue. Let $\vec{\lambda} = (\lambda_1,...,\lambda_N)$ represent the vector of arrival rates for data that are entering the queues. Depending on the QoS parameters of the connections, the BS allocates resources to a particular queue, and the data are transmitted.

We consider WirelessMAN-OFDMA as a PHY specification. In OFDMA, a set of transmitter's carriers is divided into subsets, each of which can address a different receiver at any given time depending on the specific throughput requirements of the user. In this paper we assume that the available bandwidth is divided into a number of subchannels (*m*), each of which is made of multiple subcarriers according to OFDMA. The duration of one frame is divided into a given number of time-slots (*n*) (Fig.2).

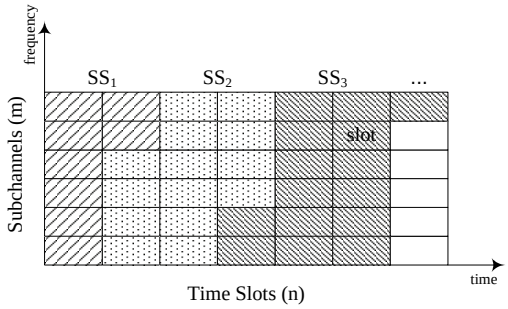

Fig.2. Bandwidth allocation

Each subchannel/time-slot pair will be referred to as a *slot*. A slot can be assigned to at most one user/application and one transmission direction. Let us denote the transmission rate that can be achieved within a single slot with s_k ($k=1,\ldots,nm$). It varies depending on the used modulation and coding scheme, the condition of the wireless link, the allocated power, etc. The number of slots assigned to a user x_i ($i=1,\ldots,nm$) and the specific transmission rates depending on the chosen type of modulation of these slots determine its overall throughput which we will denote by $a(x_i)$.

In our optimization problem for allocating the total available bandwidth to the ongoing and the incoming connections, we associate with each user a utility function of its throughput $u_i(a(x_i))$ We assume that these functions are increasing, strictly concave, continuously differentiable over the range $x_i \geq 0$, and $u_i(0) = 0$. These functions represent the utility of that user/application to the corresponding throughput achieved. The objective is to maximize the overall utility of all users/applications. The main advantage in using such an approach is that these utility functions could be sophisticated and able to provide fair resource allocation. All the fairness issues concerning the type of scheduling service to which the application belongs may be taken into consideration through the specification of the utility functions. Let $u_i(a(x_i))$ indicate the "value" to source i of having throughput $a(x_i)$. We can assume that each "slot" has a cost function $g(x_k), k = 1,..,nm,$

which indicates the cost to the network of supporting x_k slots for connection i, and may represent the second aspect of the fairness problem – the type of modulation used. Then, a "utility-fair" allocation of resources in an allocation which maximizes $U(\vec{x})$, defined by:

$$U(\vec{x}) = \sum_{i=1}^{2N+2} u_i\left(a(x_i)\right) - \sum_{k=1}^{nm} g(x_k), \quad (2)$$

where $\vec{x}$ presents the optimal solution vector. The BS has to estimate if the available resources may be allocated among the ongoing and the new incoming connection in such a way that QoS for each application can be met. In other words, if the optimization problem is feasible, the new request should be accepted for service, and the optimal solution itself gives the optimal resource allocation.

We add a number of constraints to the optimization problem. First, the total number of slots that are allocated should be less than or equal to the total number of available slots (bandwidth) in the system. The second constraint is related to the mean queue length for each user. When an application has a certain packet arrival rate, it is expected that in the corresponding queue there would be a certain number of packets, as that number is related to the arrival rate. Slots should be allocated to this queue only if we expect that there will be enough data to fill them. We denote by $\overline{L_i}$ the mean number of packets in the i^{th} queue and by $l(x_i)$ the amount of data that can be transmitted with x_i allocated slots.

Data delay as an important QoS parameter determines the third constraint. We assume that each user demands a service that has stringent requirements about the maximum allowed average data delay $\hat{d}_i$, and the next constraint in our optimization problem will ensure that the rate achieved through resource allocation should be enough to guarantee it. Let $\overline{w}(\lambda_i, x_i)$ denote the average delay for packet arrival rate λ_i and x_i allocated slots. We assume a Poisson request arrival process.

The optimization problem can be formulated as follows:

$$\begin{aligned} \max_{\mathbf{x}} \quad & U(\vec{x}) \\ s.t. \quad & \sum_{i=1}^{2N+2} x_i \le nm \\ & l(x_i) \le \overline{L_i}, \quad i = 1,\ldots,2N+2 \\ & \overline{w}(\lambda_i, x_i) \le \hat{d}_i, \quad i = 1,\ldots,2N+2 \\ & x_i \ge 0, \quad i = 1,\ldots,2N+2 \end{aligned} \quad (3)$$

If the problem is feasible, the optimal solution will contain the amount of bandwidth allocated to each connection, including the new one. If the solution is infeasible, there is no bandwidth allocation scheme such that the throughput and delay requirements for all users can be satisfied, and the incoming connection is blocked. The optimization problem may be used as a stand-alone scheme for optimal resource allocation that can be performed at the beginning of each frame. The fairness is provided through the specification of the utility functions.

V. Conclusions

In this paper we presented the specifics of the IEEE 802.16 standard, the methods for QoS provisioning, and the corresponding problems with "fair" resource allocation. We formulated an optimization problem where we focused on maximizing a utility-based objective function related to users' throughput, which comprises fairness. In future studies we intend to develop further the concept of utility functions.

Acknowledgements

This research was done under contract BУ-TH-105/2005 between Technical University of Sofia and the Bulgarian Ministry of Education and Science.

References

[1] IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems. June, 2004,

[2] Eklund C., R. Marks, S. Ponnuswamy, K. Stanwood, N. van Waes, *WirelessMAN: Inside the IEEE 802.16TM Standard for Wireless Metropolitan Area Networks*, IEEE press, 2006,

[3] Cicconetti C., C. Eklund, L.Lenzini, E. Mingozzi, "Quality of Service Support in IEEE 802.16 networks", IEEE Network, vol. 20, no. 2, March 2006

[4] Sayenko A., O. Alanen, J. Karhula, T. Hamalainen, "Ensuring the QoS Requirements in 802.16 Scheduling", Proc. of the 9th ACM symposium on Modeling analysis and simulation of wireless and mobile systems, 2006, Pages: 108 – 117,

[5] Iyengar R., K. Kar, B. Sikdar, "Scheduling Algorithms for Point-to-Multipoint Operation in IEEE 802.16 Networks", Proc. Of 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006, pp1-7,

[6] Niyato D., E. Hossain, "Joint Bandwidth Allocation and Connection Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks", IEEE ICC Proceedings, 2006,

[7] Hosein P. "On the Optimal Allocation of Downlink Resources in OFDM-Based Wireless Networks". Proc. of WWIC 2006, Bern, Switzerland,

[8] Bertsekas D., R. Gallager, *Data Networks*, Englewood Cliffs, NJ: Prentice Hall, 1987,

[9] Kelly F., A. Mauloo, D. Tan "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability", J. Operational Research Society, 1998, pp. 237-252,

[10] Bohge M., J. Gross, A. Wolisz, M. Meyer, "Dynamic Resource Allocation in OFDM Systems: An Overview of Cross-Layer Optimization Principles and Techniques", IEEE Network, Jan/Feb 2007,

[11] Song G., Y. Li, L. J. Cimini Jr., H. Zheng, "Joint Channel-Aware Data Scheduling in Multiple Shared Wireless Channels", IEEE WCNC 2004 Proceedings, Vol.3, pp. 1939-1944,

[12] Kim K., Y. Han, "A Proportional Fair Scheduling for Multicarrier Transmission Systems", IEEE Commun. Lett. vol.9, no.3, March 2005, pp. 210-212.