Congestion-Based Pricing Resource Management in Broadband Wireless Networks

Najah AbuAli, Mohammad Hayajneh, and Hossam Hassanein

***Abstract*—Supporting the diverse QoS requirements of multimedia applications is an essential requirement for broadband wireless access (BWA) networks. Due to the projected dynamics in traffic patterns, more capable resource management functionalities are needed. We propose a game-theoretic, congestion-based pricing scheduler that incorporates two sub-schemes: a bandwidth provisioning sub-scheme to address the bandwidth scarcity to provision in fourth generation (4G) BWA technologies and an efficient packet scheduler sub-scheme. To the best of our knowledge, the proposed scheduler is the first one to simultaneously control congestion and fairness while providing differentiated QoS guarantees in BWA networks. Simulation results show that the proposed scheme realizes our objectives of controlling congestion, providing differentiated QoS guarantees, and catering to proportional fairness among the different network classes and among connections within the same class.**

***Index Terms*—Congestion control; scheduling; QoS; game theory; radio resource management (RRM).**

I. Introduction

FOURTH Generation (4G) Wireless Networks are developed to further improve the user's experience of access services through higher data access rates and improved quality of service (QoS) provisions. Emerging Broadband Wireless Access (BWA) such as IEEE 802.16m and 3GPP's Long Term Evolution-Advanced (LTE-Advanced) are the leading candidate technologies to be utilized for Fourth Generation (4G) BWA. These technologies are expected to serve a wide range of applications including data, voice, gaming, and multimedia. 4G BWA networks are challenged by the need to guarantee the diverse QoS requirements of such applications over a limited radio spectrum. Such a challenge dictates a need for capable network planning and resource management and requires sound novel approaches to ensure QoS. To address this need, the IEEE 802.16 standard [1] defines a service flow framework that supports multiple classes: Unsolicited Grant Service (*UGS*), real time Polling Service (*rtPS*), non-real-time Polling Service (*nrtPS*), Best Effort (*BE*) and Extended real time Polling Service (*ErtPS*). Likewise, LTE [2] defines two main QoS classes: the guaranteed rate class and the non-guaranteed rate class. Each traffic flow in LTE belongs to one of these two classes, and is associated with QoS parameters that precisely specify the values of packet delay and loss that can be tolerated by the flow application.

Packet scheduling is the function that enforces the QoS parameters in WiMAX and LTE. WiMAX and LTE standards do not define a specific packet scheduler. Several WiMAX [3], [4], [5] and LTE [6], [7], [8] scheduling schemes have been recently proposed in the literature. However, these schemes are mainly designed to maximize network throughput or operator revenue, in addition to enforcing users' flows QoS requirements. Congestion control in WiMAX and LTE has received little attention in literature. Few proposals address the issue of designing QoS schedulers at the MAC layer with explicit congestion control mechanisms. Song and Li [9] propose a scheduler for an OFDM wireless network for a mixture of delay sensitive and best–effort traffic. The proposed scheme utilizes two utility functions for each traffic type. The delay sensitive traffic utility function is employed to meet the delay requirements of the traffic, while the best–effort utility function is used to control the traffic greediness. Hence, the scheduling scheme is effectively a congestion management tool, as opposed to a congestion prevention tool. In addition, the scheduler is only triggered by the greediness of best–effort traffic. The work presented in reference [10] establishes a QoS-driven adaptive congestion control framework that provides QoS guarantees for a single service flow, namely VoIP. The framework is composed of three radio resource management algorithms: admission control, packet scheduling and load control. The guarantees for VoIP QoS requirements are realized by the adaptive load control algorithm, which monitors the VoIP quality and dynamically adapts the parameters of the admission control and the packet scheduler algorithms when congestion is projected. The adaptation is based on a trade-off between VoIP QoS guarantees and efficient network resource usage in the mixed services scenario.

Other proposals address the congestion control problem following one of two approaches: either designing a call admission control (CAC) scheme [11] or enhancing (or re-designing) the congestion control mechanism at the transport layer [12]. The main goal of designing CAC schemes is to avoid congestion and maintain the delivered QoS to different connections at a target level by means of blocking new connections. CAC schemes can avoid long–term network congestion but cannot adapt to short–term congestion. On the other hand, addressing congestion control at the transport layer is more involved, especially in wireless networks. This is because packets may be lost due to bad channel quality rather than congestion, and the transport layer mistakenly

Manuscript received August 26, 2009; revised February 1, 2010; accepted March 13, 2010. The associate editor coordinating the review of this paper and approving it for publication was N. Kato.

N. AbuAli and M. Hayajneh are with the College of IT, United Arab Emirates University, Al-Ain, United Arab Emirates (e-mail: {najah, mhayajneh)@uaeu.ac.ae).

H. Hassanein is with the School of Computing, Queen's University, Kingston, Canada (e-mail: hossam@cs.queensu.ca).

This work was supported by grant no. 2009/053 from the Emirates Foundation, United Arab Emirates, and in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

The authors would like to thank the anonymous reviewers whose feedback helped improve the quality of this paper.

Digital Object Identifier 10.1109/TWC.2010.061810.091283

reacts to this type of packet loss as if it is due to congestion. The Base Station (BS) is the main entity that is capable of detecting the network status in wireless networks and not the end nodes. Effective MAC layer scheduling scheme at the BS in congestion with an end-to-end scheme can better control congestion. Additionally, most congestion control mechanisms at the transport layer require changes in the operating systems of users' devices, which is not practically a favorable solution.

While dynamic pricing is a powerful tool for controlling congestion in BWA networks, it has received little attention and is still an open research issue. As well, and to the best of our knowledge, there is no study that addresses *dynamic pricing* as a tool for controlling congestion and differentiating services. The current proposals utilize pricing models to address other problems in BWA networks. For example, the authors of [13] proposed a pricing model for bandwidth sharing in heterogeneous networks of WiMAX and WiFi technologies. Likewise, in reference [14] a pricing model for using WiMAX as a network failure backup for a WiFi network is proposed. The proposal in [15] presented a market-based cognitive radio network pricing model to admit new users to an access network by service providers that minimizes the effect on QoS of existing users and increases the service provider revenue. The closest work to our proposal is presented in [16]. The author proposes a pricing model for WiMAX UGS, rtPS and BE classes. The proposed model statically differentiates pricing based on the class type, and is hence not dynamic. The model is threshold based, and may not be an efficient tool for reactive short–term congestion control. Also, the proposed model supports only a single QoS metric and left other metrics such as the delay for future work.

In this paper, we address the void for a congestion control scheme in BWA networks to better utilize the limited radio spectrum. The main contribution of this paper is designing a congestion-based pricing scheme that consists of two non-cooperative game-theoretic pricing algorithms at two levels: class level and connection level. The scheme is designed to cater to fairness and congestion control while efficiently differentiating services in BWA networks. Using game theory as a robust tool is motivated by the success of its application in many fields where it has been shown to provide capable solutions [17], [18], especially in situations where rational agents share a common resource while exercising potentially conflicting objectives.

Designing a fair scheduling scheme is achieved through exercising proportional fairness among the different classes at one level and connections within same class at other level. Effective congestion control is achieved through introducing a pricing parameter in a market-based scheme. The BS varies a pricing factor based on the traffic load in the network at both the class level and the connection level. Simulation results show that varying the pricing factor results in controlling congestion by reducing the average throughput of the served connections. This reduction is achieved without compromising the fairness among the classes or the QoS guarantees of the connections, since the average throughput achieved is larger than or equal to the minimum data rate requested by this connections.

We proceed by providing an overview of the proposed scheme in Section II. The system model is presented in Section III. The problem is formulated and solved at two levels: the class level and the connection level, which are respectively presented in Section IV and Section V. In Section VI, we evaluate the performance of our proposed approach and discuss the results. Finally, in Section VII we conclude the paper.

II. Scheme Overview

The scheme consists of two congestion-based pricing sub-schemes, one for bandwidth provisioning at the class level and the other for scheduling at the connection level. The two–level design efficiently guarantees the QoS requirements of connections within each class. The class-level scheme isolates the classes with diverse QoS requirements. For example, for real-time traffic, isolating VoIP traffic from video streaming is required since VoIP traffic has a stricter delay requirements than video streaming. Hence, if a VoIP packet and a video streaming packet both have the same waiting time in a queue, the VoIP packet is given higher priority. On the other hand, differentiating connections within the same class enhances the QoS guarantees, especially for the traffic flows that have more than one QoS metric. For example, within the same class (e.g VoIP class), packets approaching their delay bounds are given higher priority than others. QoS guarantees is hence achieved by overseeing the guarantees at the two levels.

Designing an efficient and robust bandwidth provisioning algorithm requires taking different issues into account. The objectives of our bandwidth provisioning algorithm are as follows.

1) Enforcing the requirements of various QoS classes;
2) Capturing and making use of the technology's physical and functional specifications;
3) Efficiently providing for a wide range of QoS requirements;
4) Satisfying fairness among classes; and
5) Adapting to network status, especially in congestion instances.

These objectives form the basis of our design and mathematical modeling. For our purposes, we identify game theory as a powerful tool to design efficient resource management schemes with the above characteristics in a congested network environment. Game–theoretic approaches explore how strategic interactions between rational players (or agents) generate outcomes according to the players' conflicting preferences [19], [20]. For bandwidth provisioning, we consider the players to be the different classes supported by the BWA network, the players' conflicting preferences as the classes' QoS requirements, and the outcomes as the resources allocated within a predefined epoch. An increased demand of a specific class for bandwidth results in decreasing the available bandwidth resources of the other classes in the network. Game theory can hence establish a suitable framework to capture the interactions of self-interested players with potentially conflicting interests.

At the class level, we adopt a one stage non-cooperative congestion-based pricing game at the BS to calculate allocation of time slots of a given time frame for the different classes' services in a BWA network. We call the decision epoch within which this calculation is performed the *interclass*

resource allocation phase. We define four players corresponding to four service flow classes. Each player is associated with a utility function to address fairness among the different service classes. A pricing factor is introduced to penalize greedy service classes and control network congestion.

At the connection level, we design a single stage, non-cooperative congestion-based game to be executed at the BS for Down-Link (DL) scheduling and at the Subscriber Station (SS) for UP-Link (UL) scheduling. Players correspond to connections within a same class. Similar to the interclass resource allocation phase, each player is associated with a utility function to address fairness among connections of the same class. This phase is called the *intra–class resource allocation phase*.

We remark that our scheme is generic and can be applied in any class-based wireless network such as WiFi, HSDPA or LTE. In this work, and partially for illustration purposes, the performance of our algorithm is investigated in a WiMAX setting.

III. System Model

We propose a scheme to be implemented at the BS in a single cell consisting of one BS with several SSs. This is applicable in a Point to Multi-Point (PMP) mode or a multi-hop relay network with a tree-based infrastructure. Specifically it can be implemented in the relay non-transparent mode, where the non-transparent node plays the role of the base station for the mobile station. It is assumed that the BS (or the non-transparent relay node) has complete channel state information of all connections through a robust feedback channel. An SS conveys its received signal to noise power ratio (SNR) from itself to the BS. This information does not vary within a single frame as the wireless channel between each SS and BS on each subcarrier is assumed to undergo flat fading that is fixed over a single frame period. The Adaptive Modulation Coding (AMC) technique described in IEEE 802.16 standard [1] divides the range of the received SNR into seven non-overlapping regions as shown in Table I. According to the received SNR at a specific SS, the BS decides the suitable transmission mode and consequently the data rate for each SS. We consider a real-time environment with four classes, rtPS, ertPS, nrtPS and BE. We do not consider the UGS class in our design, since the UGS class is designed to serve periodic constant bit rate traffic. Hence, it requires fixed number of time slots during the UGS connection life, which can be easily scheduled by allocating these slots once the UGS connection is admitted. The scheme is designed to guarantee a minimum transmission rate for each connection belonging to each class. The transmission rate can be increased up to a maximum value based on the network load.

TABLE I
Modulation and Coding Schemes for IEEE 802.16.

Modulation (coding)	Info bits/symbol	Required SNR
BPSK(1/2)	0.5	6.4
QPSK(1/2)	1	9.4
QPSK(3/4)	1.5	11.2
16QAM(1/2)	2	16.4
16QAM(3/4)	3	18.2
64QAM(2/3)	4	22.7
64QAM(3/4)	4.5	24.4

IV. Bandwidth Provisioning: Class Level

The bandwidth provisioning algorithm is designed to calculate the number of time slots per class. The scheduling algorithm proposed in Section V allocates time slots amongst connections within the same class. We remark here that the bandwidth provisioning algorithm can be used as a stand alone algorithm that need not be associated with the scheduling algorithm proposed in Section V.

The bandwidth provisioning algorithm isolates the service classes by allocating K_i time-slots for each class i out of K total time-slots in a given time frame. In IEEE 802.16, time frames are divided into constant number of time slots with fixed time-slot duration measured in microseconds. The slot duration measured by number of bits transferred within the fixed time-slot duration may change on per-frame bases depending on the channel conditions and consequently the modulation and coding mode adapted at the current time-frame.

The total K time-slots are completely partitioned among the classes such that the difference between a utility function and a pricing function is maximized. The utility function is defined as a logarithmic normalized function of the ratio of each class's allocated slots to the total number of allocated slots for other classes. The utility function facilitates optimally allocating the frame time-slots to the different service classes in order to guarantee proportional fairness among the service classes. The pricing function is defined as a weighted function of class i allocated slots. In addition to controlling congestion, the role of the pricing function is to counter greedy demand for time slots. We define P1 as our first optimization problem with the objective function $L(K_i, \mathbf{K}^{-i})$, where $\mathbf{K}^{-i}$ is an array of the allocated time-slots of all classes except the i^{th} class:

$$\text{P1:} \quad \max_{K_i \in \mathcal{K}_i} \left\{ L(K_i, \mathbf{K}^{-i}) = U_i \log\left(1 + \frac{K_i}{\Upsilon_i}\right) \right. \tag{1}$$

$$+ \quad \psi_i \log(K_i - \rho_i K_{i,min}) \tag{2}$$

$$- \quad \left. \lambda K_i \right\}, \tag{3}$$

The utility function is defined as $U_i \log(1+\frac{K_i}{\Upsilon_i})$ over the set of class i's time slots denoted by $\mathcal{K}_i \in [K_i^{min} \; K_i^{max}]$. U_i is introduced as a utility factor to measure how much bandwidth service class i is requesting. Υ_i, defined as $\Upsilon_i \triangleq v_i + \sum_{m \neq i}^{4} K_m$, is a normalizing factor of the fairness measure. The parameter v_i is used as a tuning factor.

The utility function in (1) ensures proportional fairness among the different service classes. The barrier function in (2) guarantees that the new allocated time slots are enough to support the payload of all active connections of class i. ψ_i is a prioritizing parameter with $U_i \gg \psi_i$. The parameter ρ_i is a tuning factor introduced to give the network operator flexibility in deciding how much resources to allocate for classes above their minimum requested time-slots, $K_{i,min}$. If desired, the effect of ρ_i can be disabled by choosing $\rho_i = 1$. A linear pricing function in (3) with pricing factor λ is introduced to prevent greedy use of network resources and to control congestion.

TABLE II
OFDM FRAME DURATION AND NUMBER OF FRAMES PER SECOND.

Code	Frame Duration (ms)	Frames per second
0	2.5	400
1	4	250
2	5	200
3	8	125
4	10	100
5	12.5	80
6	20	50

Mapping the requested bandwidth to a maximum and a minimum number of time-slots is based on the fact that IEEE 802.16 standard defines a fixed size of time frames and their corresponding number of frames sent per second as shown in Table II. Given a requested bandwidth requirement per second for each connection n, the *requested bandwidth requirement per frame is equal to the requested bandwidth per second divided by the number of frames per second.* Thus, given minimum reserved bandwidth per frame, ($\tilde{B}_n^{min}$), and maximum sustainable bandwidth per frame ($\tilde{B}_n^{max}$), the number of time-slots for each connection n can be represented as

$$\hat{K}_n^{min} = \frac{\tilde{B}_n^{min}}{r_n} \quad (4)$$

$$\hat{K}_n^{max} = \frac{\tilde{B}_n^{max}}{r_n} \quad (5)$$

where r_n is an average value that reflects the channel quality over a predefined window size using a proper prediction algorithm. The minimum and maximum number of time slots of service class i are: $K_i^{min} = \sum_{k=1}^{|\mathcal{C}_i(t)|} \hat{K}_k^{min}$ and $K_i^{max} = \sum_{k=1}^{|\mathcal{C}_i(t)|} \hat{K}_k^{max}$, respectively. $\mid \mathcal{C}_i(t) \mid$ is the number of connections of class i.

To represent problem P1 as a formal game theoretic problem, G1, we define $x_i \triangleq \frac{K_i}{\Upsilon_i}$, $x_i \in \mathcal{X}_i$, with $\mathcal{X}_i \in [x_i^{min} \quad x_i^{max}]$ being the action profile or the strategy space of the i_{th} player (the i^{th} service class). Consequently, the objective function in problem P1 is modified as follows

$$\text{G1}(x_i, \mathbf{X}^{-i}, \lambda_i): \quad \max_{x_i \in \mathcal{X}_i} \left\{ L_i(x_i, \mathbf{X}^{-i}, \lambda_i) \right\}, \forall\, i \in \mathcal{I}, \quad (6)$$

with $L_i(x_i, \mathbf{X}^{-i}, \lambda_i)$ given by

$$\begin{aligned} L_i(x_i, \mathbf{X}^{-i}, \lambda_i) &= U_i \log(1 + x_i) \\ &+ \psi_i \log(x_i - \rho_i\, x_i^{min}) - \lambda_i\, x_i \end{aligned} \quad (7)$$

$\text{G1}(x_i, \mathbf{X}^{-i}, \lambda_i)$ is a Noncooperative Game with Pricing (NGP) in which player i decides its next decision x_i based on local information $(\mathbf{X}^{-i}, \lambda_i)$ that maximizes its objective function $L_i(x_i, \mathbf{X}^{-i}, \lambda_i)$. The vector of players' actions (decisions) is denoted by $\mathbf{X} = (x_1, x_2, x_3, x_4)$, while $\mathbf{X}^{-i}$ is the vector $\mathbf{X}$ without the ith element. The pricing factor, to prevent greedy usage of the time slots resource, λ_i is equal to $\lambda\, \Upsilon_i$. Finally, $\mathcal{I} = \{1, 2, 3, 4\}$ is the indexing set of the players. The limits of the action profile $\mathcal{X}_i$, namely x_i^{min} and x_i^{max} are set such that $\Upsilon_i\, x_i^{min} = K_i^{min}$ and $x_i^{max}\, \Upsilon_i = K_i^{max}$, respectively.

To simplify representation, L_i will be used as a simplified form of $L_i(x_i, \mathbf{X}^{-i}, \lambda_i)$. The optimal value x_i^* that maximizes L_i in (6) is the feasible solution of the following equation

$$\frac{\partial L_i}{\partial x_i} = \frac{U_i}{1 + x_i} + \frac{\psi_i}{x_i - \rho_i\, x_i^{min}} - \lambda_i = 0, \forall\, i \in \mathcal{I} \quad (8)$$

or

$$x_i^2 - \Xi_i\, x_i + \phi_i = 0 \quad (9)$$

Thus,

$$x_i^* = \frac{\Xi_i}{2} \left(1 + \sqrt{1 - \frac{4\,\phi_i}{\Xi_i^2}} \right), \forall\, i \in \mathcal{I} \quad (10)$$

where $\Xi_i \triangleq -1 + \rho_i\, x_i^{min} + \frac{U_i + \psi_i}{\lambda_i}$ and $\phi_i \triangleq -\rho_i\, x_i^{min} - \frac{\psi_i - U_i\, \rho_i\, x_i^{min}}{\lambda_i}$. Then the optimal number of time slots of service class i, K_i^* can be expressed in terms of x_i^* as follows

$$K_i^* = x_i^* \Upsilon_i, \forall i \in \mathcal{I} \quad (11)$$

In order to guarantee that $x_i^* > 0$, the utility factor U_i should be lower bounded by

$$U_i > \lambda_i\, (1 - \rho_i\, x_i^{min}) - \psi_i \quad (12)$$

A. Existence of a Nash equilibrium operating point

In this subsection we demonstrate the existence of Nash Equilibrium (NE) operating point in the game G1. The existence of a NE point is guaranteed if the objective function is quasiconcave and optimized on a convex strategy space [20]. We denote the NE operating point by $\mathbf{K}^*$ (allocated time-slots vector of the four different service classes) where $\mathbf{K}^* = (K_1^*, K_2^*, K_3^*, K_4^*)$, and at equilibrium K_i^* is given by

$$K_i^* = x_i^*\, \Upsilon_i^* = x_i^* \left(\sum_{m \neq i}^{4} K_m^* + \upsilon_i \right), \forall i \in \mathcal{I} \quad (13)$$

To prove the existence of NE operating point K_i^* of G1, it is enough to prove that L_i in (6) is quasiconcave function in x_i given $\mathbf{X}^{-i}$ on the convex set $\mathcal{X}_i = [x_i^{min} \quad x_i^{max}]$. Hence, we calculate the second derivative of G1 and proof that it is always negative on the compact convex set $\mathcal{X}_i$. The second order derivatives of L_i in (6) with respect to x_i is

$$\begin{aligned} \frac{\partial^2 L_i}{\partial x_i^2} &= -\frac{U_i}{(1 + x_i)^2} - \frac{\psi_i}{(x_i - \rho_i\, x_i^{min})^2} \\ &< 0, \forall x_i \in \mathcal{X}_i, \forall\, i \in \mathcal{I} \end{aligned} \quad (14)$$

Therefore, L_i is a strictly concave function on the compact convex set $\mathcal{X}_i$, which implies that L_i is a quasiconcave function on $\mathcal{X}_i$. Therefore, an NE point $\mathbf{K}^*$ exists. Also, as a result of strict concaveness, the unconstrained maximizer x_i^* is unique.

B. Uniqueness of the NE operating point of G1

Before proving the uniqueness of the NE point $\mathbf{K}^*$ of the game G1, we present the following proposition:

Proposition 1: Given the actions of the other players (QoS classes) $\mathbf{X}^{-i}$, the best response that the player i can decide is

$$b_i(\mathbf{X}^{-i}) = \min \{x_i^*, x_i^{max}\}. \quad (15)$$

Proof: It is obvious that x_i^* is the unconstrained maximizer of the objective function L_i given the other players decisions $\mathbf{X}^{-i}$. However, if this maximizer is not feasible, i.e. $x_i^* > x_i^{max}$, then the best decision that maximizes L_i is x_i^{max}

as $\partial L_i/\partial x_i > 0$ in the interval $\{x_i : x_i^{min} < x_i \leq x_i^*\}$, which means that the objective function is increasing in this interval. Thus, given $\mathbf{X}^{-i}$, $b_i(\mathbf{X}^{-i}) = x_i^{max}$ maximizes L_i if x_i^* is not feasible and this concludes the proof. ■

Recall that $K_i = x_i \Upsilon_i$ which is a one-to-one mapping relationship, consequently the best response function in (15) can be written in terms of the number of time slots K_i as follows

$$b_i(\mathbf{K}^{-i}) = \min\{K_i^*, K_i^{max}\}, \tag{16}$$

where $\mathbf{K} = (K_1, K_2, K_3, K_4)$ and $\mathbf{K}^{-i}$ is the vector $\mathbf{K}$ without the i^{th} element.

Define the best response vector function $b(\mathbf{K})$ as follows:

$$b(\mathbf{K}) \triangleq (b_1(\mathbf{K}^{-1}), b_2(\mathbf{K}^{-2}), b_3(\mathbf{K}^{-3}), b_4(\mathbf{K}^{-4})) \tag{17}$$

In light of the following theorem based on the observations in [21], we can prove the uniqueness of the NE operating point $\mathbf{K}^*$ of game G1.

Theorem 1: A non-cooperative game with a standard best response function adopts a unique equilibrium point if it exists.

At this stage and for completeness, we provide the definition of a generic standard function $\Theta(\mathbf{K})$ as follows [21]:

Definition 1: A vector function $\Theta(\mathbf{K})$ is called a standard vector function if it satisfies the following: 1) Positivity: $\Theta(\mathbf{K}) > 0$, i.e. each element is positive, 2) Monotonicity: if $\mathbf{K} > \hat{\mathbf{K}}$ then $\Theta(\mathbf{K}) \geq \Theta(\hat{\mathbf{K}})$ (entry wise), and 3) Scalability: $\forall \delta > 1,\ \delta\Theta(\mathbf{K}) > \Theta(\delta\,\mathbf{K})$ (entry wise).

Thus, to prove the uniqueness of $\mathbf{K}^*$ we only need to prove that $b(\mathbf{K})$ is a standard function.

Positivity of the best response $b(\mathbf{K})$ is clear from the fact that the strategy space of each player $\mathcal{X}_i$ is a convex and compact subset of the positive real numbers $\mathbb{R}^+$. To prove that $b(\mathbf{K})$ is a monotone function in $\mathbf{K}$ it is enough to prove that each of its components is a monotone function in $\mathbf{K}$. To do so, let $\mathbf{K} > \hat{\mathbf{K}}$. Then

$$b_i(\mathbf{K}) = b_i(\mathbf{K}^{-i}) = x_i^*\,\Upsilon_i = x_i^*\,(v_i + \sum_{m\neq i}^{4} K_m), \forall\, i \in \mathcal{I} \tag{18}$$

while

$$b_i(\hat{\mathbf{K}}) = b_i(\hat{\mathbf{K}}^{-i}) = x_i^*\,\hat{\Upsilon}_i = x_i^*\,(v_i + \sum_{m\neq i}^{4} \hat{K}_m), \forall\, i \in \mathcal{I} \tag{19}$$

By examining the equations (18) and (19), one can see that $b_i(\mathbf{K}) > b_i(\hat{\mathbf{K}})$, i.e. it is a monotone function in $\mathbf{K}$, then the monotonicity of the best response function $b(\mathbf{K})$ is implied.

Consider $\delta > 1$. Then

$$\delta\, b_i(\mathbf{K}) = \delta\, x_i^*\,(v_i + \sum_{m\neq i}^{4} K_m), \forall\, i \in \mathcal{I}. \tag{20}$$

On the other hand,

$$b_i(\delta\,\mathbf{K}) = x_i^*\,(v_i + \delta \sum_{m\neq i}^{4} K_m), \forall\, i \in \mathcal{I}. \tag{21}$$

Equations (20) and (21) show that $\delta\, b(\mathbf{K}) > b(\delta\,\mathbf{K})$, i.e. $b(\mathbf{K})$ is a scalable standard function. Then the uniqueness of the NE point is guaranteed based on Theorem 1.

C. Pareto optimality of the NE point for G1

In this Subsection, we investigate the efficiency of the NE operating point $\mathbf{K}^*$. We prove that it is a Pareto optimal operating point, i.e. it is the best operating point that the players (service classes) can reach in a non-cooperative manner at equilibrium. Similar to [19], a Pareto optimal operating point is defined as follows.

Definition 2: The vector $\mathbf{K}^*$ is Pareto optimal if there exists no other vector $\tilde{\mathbf{K}} > \mathbf{K}^*$ such that $L_i(\tilde{\mathbf{K}}) \geq L_i(\mathbf{K}^*)$ for all $i \in \mathcal{I}$ and $L_i(\tilde{\mathbf{K}}) > L_i(\mathbf{K}^*)$ for some $i \in \mathcal{I}$.

The following Lemma demonstrate the Pareto optimality of $\mathbf{K}^*$.

Lemma 1: The NE operating point $\mathbf{K}^*$ of the game G1 is a Pareto optimal operating point.

Proof: Without loss of generality, let $\tilde{K}_i = \pi_i\, K_i^*, \forall\, i \in \mathcal{I}$, where $\pi_i > 1$, then

$$\begin{aligned} L_i(\tilde{K}_i, \tilde{\mathbf{K}}^{-i}) &= U_i \log\left(1 + \frac{\pi_i\, K_i^*}{v_i + \sum_{m\neq i} \pi_m\, K_m^*}\right) \\ &+ \psi_i\, \log(\pi_i\, K_i^* - \rho_i\, K_{i,min}) \\ &- \lambda\, \pi_i K_i^* \end{aligned} \tag{22}$$

The behavior of L_i in (22) with respect to π_i can be found by differentiating L_i with respect to π_i as follows

$$\begin{aligned} \frac{\partial\, L_i(\tilde{K}_i, \tilde{\mathbf{K}}^{-i})}{\partial\, \pi_i} &= \left(\frac{U_i}{v_i + \sum_{m\neq i}^{4} \pi_m\, K_m^* + \pi_i\, K_i^*}\right. \\ &+ \left.\frac{\psi_i}{\pi_i\, K_i^* - \rho_i\, K_i^{min}} - \lambda\right) K_i^* \\ &\triangleq K_i^*\, f(D_\pi\, \mathbf{K}^*), \forall i \in \mathcal{I}, \end{aligned} \tag{23}$$

where

$$D_\pi \triangleq \begin{pmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \pi_3 & 0 \\ 0 & 0 & 0 & \pi_4 \end{pmatrix}$$

Recall that K_i^* is the unconstrained maximizer of L_i in (1), therefore at equilibrium

$$\begin{aligned} \left.\frac{\partial\, L_i(K_i, \mathbf{K}^{-i})}{\partial\, K_i}\right|_{\mathbf{K}=\mathbf{K}^*} &= \left(\frac{U_i}{v_i + \sum_{m\neq i}^{4} K_m^* + K_i^*}\right. \\ &+ \left.\frac{\psi_i}{K_i^* - \rho_i\, K_{i,min}} - \lambda\right) = 0 \\ &\triangleq f(\mathbf{K}^*), \forall i \in \mathcal{I}. \end{aligned} \tag{24}$$

Since $K_i^* > 0$ and by observing the two equations (23) and (24), one can conclude that

$$f(D_\pi\, \mathbf{K}^*) < f(\mathbf{K}^*) = 0, \forall \pi_i > 1,$$

which implies that

$$\frac{\partial\, L_i(\tilde{K}_i, \tilde{\mathbf{K}}^{-i})}{\partial\, \pi_i} < 0.$$

That is, when L_i in (22) decreases, π_i increases. Thus, according to Definition 2, $\mathbf{K}^*$ is a Pareto optimal operating point. ■

D. Class time-slots allocation algorithm

In reference [21], it was proven that both synchronous and asynchronous algorithms with standard best response functions converge to the same point. Therefore, we consider asynchronous time slots allocation control algorithms that converge to the unique Nash equilibrium point $\mathbf{K}^*$ of game G1. In this algorithm, the players (service classes) update their numbers of time slots as follows

Assume class m updates its required number of time slots at time tics [1], where these time tics refer to the iteration indices in the set $T_m = \{t_{m_1}, t_{m_2}, ...\}$, with $t_{m_n} < t_{m_{n+1}}$ and $t_{m_0} = 0$ for all $m \in \mathcal{I}$. Let $T = \{t_1, t_2, \cdots\}$ where $T = T_1 \bigcup \cdots \bigcup T_4$ with $t_n < t_{n+1}$.

Define $\underline{\mathbf{K}}$ to be the vector of numbers of time slots picked randomly from the total strategy space $\mathcal{K} = \mathcal{K}_1 \bigcup \cdots \bigcup \mathcal{K}_4$.

Algorithm 1: Consider the game G1 given in (6) and generate a sequence of time slots vectors as follows:

- **(a)** Set the time slots vector at time $t_0 = 0$: $\mathbf{K}(0) = \underline{\mathbf{K}}$, let $k = 1$
- **(b)** For all $m \in \mathcal{I}$, such that $t_k \in T_m$: Given $\mathbf{K}(t_{k-1})$, calculate $U_m(t_k) > \lambda_m(t_{k-1})\,(1 - \rho_i\, x_m^{min}) - \psi_m(t_{k-1})$, and $\psi_m(t_k) = 0.01 U_m(t_k)$, then $x_m^*(t_k) = \underset{x_m \in \mathcal{X}_m}{\operatorname{argmax}} L_m(x_m, \mathbf{X}^{-m}(t_{k-1}), \lambda_m(\mathbf{K}(t_{k-1})))$, then set $K_m^*(t_k) = x_m^*(t_k)\,(v_m + \sum_{j \neq m}^{4} K_j(t_{k-1}))$ and let the number of time slots $K_m(t_k) = b_m(t_k) = \min(K_m^*(t_k), K_m^{max})$
- **(c)** If $\mathbf{K}(t_k) = \mathbf{K}(t_{k-1})$ stop and declare the Nash equilibrium number of time slots vector as $\mathbf{K}(t_k)$, else let $k := k + 1$ and go to **(b)**.

The output of Algorithm 1 is the optimal allocated time slots for the different service classes, which can be computed in polynomial time [22]. The work presented in [22] proved that a multi-player game can be computed in polynomial time if it is guaranteed that the game is a pure Nash equilibrium problem. Our proposed algorithm is mathematically formulated as a pure Nash equilibrium problem (as opposed to a randomized Nash equilibrium), hence based on the study conducted in [22], the algorithm computation can run in a polynomial time.

In the following section, we calculate the optimal allocated time slots per connection within the same class.

V. Downlink/Uplink Game-Theoretic Scheduling Algorithms: Connection Level

After calculating the optimal number of time-slots for each QoS service class (interclass bandwidth allocation phase), we introduce our game theoretic scheduling algorithm based on a formulation for the optimization problem P2 below. The objective function of P2 is to optimally allocate time slots among connections of same class (intraclass bandwidth allocation phase) using the results of the interclass phase.

$$\text{P2:} \quad \max_{B_i \in \mathcal{B}_i} \left\{ L(B_i, \mathbf{B}^{-i}) = \hat{U}_i \log\left(1 + \frac{B_i}{\Theta_i}\right) \right. \tag{25}$$

$$\left. + \quad \hat{\psi}_i \, \log(B_i - \hat{\rho}_i \, B_{i,min}) \right\} \tag{26}$$

The utility function in (25), $\hat{U}_i \log\left(1 + \frac{B_i}{\Theta_i}\right)$, is introduced to provide proportional fairness among the different connections of a given service class through the ratio $\frac{B_i}{\Theta_i}$, where $\Theta_i \triangleq c_i + \sum_{m \neq i}^{M} B_m$ and the set of connection i's allocated bandwidth is denoted by $\mathcal{B}_i = [B_i^{min}, B_i^{max}]$. $\hat{U}_i$ is a weighting factor to the utility function, consider it as the utility factor. This factor is introduced to provide the network operator the flexibility to enforce users' service level agreements based on how much bandwidth they are requesting or/and what delay bounds their applications can tolerate. The network operators can decide on the granularity of this factor based on their policies. For example, $\hat{U}_i$ can be a function of $\frac{1}{\Delta t_m^1}$ to cater for the flows' required bound, where Δt_m^1 is the residual time to reach the latency bound of connection m. Hence, in this example, the weighting factor, $\hat{U}_i$, can be used to prioritize users who spend longer time in the system waiting for the service.

The function $\hat{\psi}_i \, \log(B_i - \hat{\rho}_i \, B_{i,min})$ in (26) is the connection's minimum requirement constraint with $\hat{U}_i \gg \hat{\psi}_i$. Given that WiMAX and LTE define a maximum data rate per traffic flow, we use $\hat{\rho}_i$ as a tuning factor to provide the network operator with the flexibility to decide how much resources to allocate for connections above their minimum requested bandwidth, $B_{i,min}$. If the network operator's policy is to allocate the minimum required bandwidth of the connection, then $\hat{\rho}_i$ is set to 1.

The problem in P2 is generically formulated to be equally applied for all defined service classes in the network. The parameters $\hat{U}_i$, $\hat{\rho}_i$, c_i, and $\hat{\psi}_i$ in P2 can be utilized when needed as prioritization variables among connections within a same class based on the connections QoS bounds. For example, $\hat{U}_i$ can be a function of the delay bound of a connection (for example; the ertPS or rtPS class connections in WiMAX or GBR class connections in LTE) to prioritize connections approaching there delay bound within the same class. Another example is utilizing $\hat{\rho}_i$ to serve network operators' policies towards the content of traffic connections of the same class. Thus, small values of $\hat{\rho}_i$ indicate that the network operator is not concerned whether a traffic class, say HTTP is starved.

Problem P2 is not a formal game theoretic problem. To reformulate it, we define $y_i \triangleq \frac{B_i}{\Theta_i}$, $\forall\, y_i \in \mathcal{Y}_i$, with $\mathcal{Y}_i = [y_i^{min}, y_i^{max}]$ being the strategy space of the i^{th} connection. Consequently, the objective function in problem P2 is modified as follows

$$\text{G2}(y_i, \mathbf{Y}^{-i}): \quad \max_{y_i \in \mathcal{Y}_i} \left\{ J_i(y_i, \mathbf{Y}^{-i}) \right\}, \forall\, i \in \mathcal{M}, \tag{27}$$

with $J_i(y_i, \mathbf{Y}^{-i})$ given by

$$\begin{aligned} J_i(y_i, \mathbf{Y}^{-i}) \quad &= \quad \hat{U}_i \log(1 + y_i) \\ &+ \quad \hat{\psi}_i \log(y_i - \hat{\rho}_i \, y_i^{min}) \end{aligned} \tag{28}$$

$\text{G2}(y_i, \mathbf{Y}^{-i})$ is a non-cooperative bandwidth control game (NBG), where connection i decides its next decision, y_i based on the values given by the vector decisions, $(\mathbf{Y}^{-i})$ such that the objective function $J_i(y_i, \mathbf{Y}^{-i})$ is maximized. The vector $\mathbf{Y}^{-i}$ is the vector $\mathbf{Y} = (y_1, y_2, \cdots, y_M)$ without the ith element, where M is the total number of connections of a given class, and the boundary of connection i's decision $(\mathcal{Y}_i)$ is lower bounded by y_i^{min} and upper bounded by y_i^{max}. y_i^{min} is related to B_i^{min} by $\Theta_i \, y_i^{min} = B_i^{min}$ and y_i^{max} by $y_i^{max} \, \Theta_i = B_i^{max}$, respectively.

To find the optimal value y_i^* that maximizes the objective function $J_i(y_i, \mathbf{Y}^{-i})$, we calculate the first derivative of J_i as

[1]We distinguish the time tics from the time slots of the time frame as the algorithm iteration indices.

follows

$$\frac{\partial J_i}{\partial y_i} = \frac{\hat{U}_i}{1+y_i} + \frac{\hat{\psi}_i}{y_i - \hat{\rho}_i\, y_i^{min}} = 0, \forall\, i \in \mathcal{M} \qquad (29)$$

Hence,

$$y_i^* = \frac{\hat{U}_i\, \hat{\rho}_i\, y_{i,min} - \hat{\psi}_i}{\hat{\psi}_i + \hat{U}_i}, \forall\, i \in \mathcal{M} \qquad (30)$$

and the optimal bandwidth allocated to a connection i is given by

$$B_i^* = y_i^* \left(c_i + \sum_{m \neq i}^{M} B_m \right). \qquad (31)$$

We omit the proofs of existence, uniqueness and Pareto optimality of NE operating point of game G2 as they are almost similar to the proofs presented above for game G1.

A. Connections' bandwidth allocation control algorithm

In this subsection, we present the algorithm for allocating the control bandwidth that converges to the unique Nash equilibrium point $\mathbf{B}^*$ of game G2. At each iteration of the algorithm, the connections update their allocated bandwidth as follows until convergence:

Assume class m updates its required number of time slots at time tics in the set $\hat{T}_m = \{\hat{t}_{m_1}, \hat{t}_{m_2}, ...\}$, with $\hat{t}_{m_n} < \hat{t}_{m_{n+1}}$ and $\hat{t}_{m_0} = 0$ for all $m \in \mathcal{M}$. Let $\hat{T} = \{\hat{t}_1, \hat{t}_2, \cdots\}$ where $\hat{T} = \hat{T}_1 \bigcup \cdots \bigcup \hat{T}_M$ with $\hat{t}_n < \hat{t}_{n+1}$ and define $\underline{\mathbf{B}}$ to be the values of the bandwidths' vector randomly chosen from the total strategy space $\mathcal{B} = \mathcal{B}_1 \bigcup \cdots \bigcup \mathcal{B}_M$.

Algorithm 2: Consider the game G2 given in (27) and generate a sequence of bandwidth vectors as follows:

- **(a)** Set the bandwidth vector at time $\hat{t}_0 = 0$: $\mathbf{B}(0) = \underline{\mathbf{B}}$, let $k = 1$
- **(b)** For all $m \in \mathcal{M}$, such that $\hat{t}_k \in \hat{T}_m$: Given $\mathbf{B}(\hat{t}_{k-1})$, calculate $\psi_m(t_k) = 0.01 U_m(t_k)$, then $y_m^*(\hat{t}_k) = \underset{y_m \in \mathcal{Y}_m}{\operatorname{argmax}} J_m(y_m, \mathbf{Y}^{-m}(\hat{t}_{k-1}))$, then set $B_m^*(\hat{t}_k) = y_m^*(\hat{t}_k)\,(c_m + \sum_{j \neq m}^{M} B_j(\hat{t}_{k-1})$ and let the bandwidth $B_m(\hat{t}_k) = b_m(\hat{t}_k) = \min(B_m^*(\hat{t}_k), B_{m,max})$
- **(c)** If $\mathbf{B}(\hat{t}_k) = \mathbf{B}(\hat{t}_{k-1})$ stop and declare the Nash equilibrium number of time slots vector as $\mathbf{B}(\hat{t}_k)$, else let $k := k + 1$ and go to **(b)**.

The output of Algorithm 2 is the optimal number of time slots per connection, which can be computed in a polynomial time [22], since the game is a pure Nash equilibrium problem. In the following section, we evaluate the performance of the interclass and intra-class schemes through simulation.

VI. Performance Evaluation

A. Simulation setup

We evaluate the performance of our proposed congestion-based pricing scheme in an IEEE 802.16 setting. We implemented a simulator comprising a TDD cell with one BS and several SSs. The wireless channel is modeled based on the Nakagami-m channel model which is adopted to accurately describe the statistical variation of the channel gains between the BS and the SSs based on OFDM channel multiplexing. The AMC module defined by the IEEE 802.16 standard divides the range of the received SNR into seven non-overlapping regions where the dividing thresholds are evaluated based on a target prescribed bit error rate (BER). According to the received SNR at a specific SS, the module in the PHY layer of BS decides the suitable transmission mode for each SS as shown in Table I. Each transmission mode consists of a modulation and coding pair aimed at efficiently using the bandwidth while satisfying a prescribed BER.

Nodes are randomly placed over a simulation grid of 5000m×5000m. Number of subscriber stations simulated are 1 to 30. Each subscriber station can simultaneously have different types of connections including voice, video, FTP or background traffic. In loading the network, the ratio between the different traffic types is maintained for each experiment. The ratio is read Voice:Video:FTP:Background. For example, the ratio 1:2:1:3 means that Video and Background are respectively 2 and 3 times either voice or FTP traffic. We shall refer to each loading instance by *loading ratio*, e.g. loading ratio 1:2:1:3. The loading ratio is used for studying the fairness of the algorithm in supporting different types of applications with different data rate requirements when the network is congested. For example, the experiment scenario with loading ratio of 1:1:1:1 and 3:1:1:1 studies the effect of voice traffic (which has the highest priority in the system) on the fairness of the algorithm when the number of voice flows is equal to the other flow types number, and when the number of voice flows is the largest number of flows in the system. The effect of the loading ratio is discussed in Subsection VI-D.

The frame size is fixed at a value of 10 ms equally divided between UL and DL traffic, the OFDM symbol duration is 12.5 μs, and the rate of frames is 100 frames/second. Time–slots are allocated for active flows at the beginning of each frame. The DL bandwidth is simulated as 20 MHz. The channel quality of each SS remains constant per frame, but is allowed to vary from frame to frame.

B. Traffic model

For connections' traffic model, we adopt the traffic model presented in [23], (a model specifically designed and tested for WiMAX simulation). This model implements VoIP traffic for the *ErtPS* class, video streaming traffic for the *rtPS* class, FTP traffic for the *nrtPS* class and background traffic for the BE class. Table III shows the traffic model used in simulation. Each simulation scenario is repeated 20 times. MRTR in table is the Minimum Reserved Traffic Rate, MSTR is the Maximum Sustained Traffic Rate and ML is the Maximum Latency. Since we are interested in studying the performance of the scheme in congestion status, connections are admitted to the network based on a generic CAC algorithm based on their MRTR value (the design of a CAC algorithm is out of the scope of this work). However, connections can send traffic at a rate equal to or higher than MRTR and up to their MSTR, which then results in congestion.

C. Performance metrics

We utilize the following metrics in our performance evaluation:

Average Throughput: The rate of packets successfully transmitted during the simulation time. This metric is used to understand the effect of load, congestion and relative priority of the classes and connections.

TABLE III
TRAFFIC MODEL

Service	Traffic Type	MRTR (kbps)	MSTR (kbps)	ML (ms)	Packet (bytes)
ErtPS	Voice	12.5	64	80	60
rtPS	Stream video	64	500	150	170-320
nrtPS	FTP	45	500	–	250
BE	Background	1	64	–	250

Average queuing delay: The time between the arrival of a packet to the departure of the packet from the queue. The value is reported in milliseconds (ms) and is averaged over the number of packets.
Packet loss: The percentage of packets dropped from the queue out of all the packets that arrived into the queue. The metric indicates the percentage of packets that missed their delay bounds. Both average delay and packet loss will allow us to determine how effectively the scheme satisfies the QoS requirements of real-time connections.
Fairness index: Fairness is measured among all connections (interclass fairness). We use Jain's fairness index to calculate interclass fairness, which is defined as

$$\text{Jain's Fairness Index} = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n * \sum_{i=1}^{n} x_i^2}, \quad (32)$$

where n is the total number of connections in the network. To calculate interclass fairness using Jain's index, we use normalized average throughput for x_i. The average throughput of a connection is normalized with respect to the MRTR of the connection.

D. Results

We evaluate the performance of our scheme based on the desired characteristics discussed in Subsection II. The performance of the algorithm is evaluated under light and heavy loading conditions. Both fairness and satisfaction of QoS requirements are studied for the different classes and for connections within the same class. For each simulation scenario, experiments are repeated 20 times and results are obtained with 95% confidence interval lower than 0.0028.

Figure 1 shows the performance of the scheme under light load. Given that the network is not congested, it is expected to have all traffic classes achieving a throughput higher than their minimum required bandwidth. The x axis in this and the following figures indicates the loading ratio associated with each point of evaluation.

We note in Fig. 1 that for loading ratios 1:3:1:1, 1:1:3:1, 1:1:1:3, the average throughput of the more loaded classes (respectively video, FTP and background) is the lowest. This is due to the contention among same class connections for the fixed amount of bandwidth allocated for these classes. However, the voice class does not undergo the same trend, achieving almost a constant average throughput of 40kbps. This is because voice is assigned the highest priority in the network, and also due to the voice class's low traffic requirements

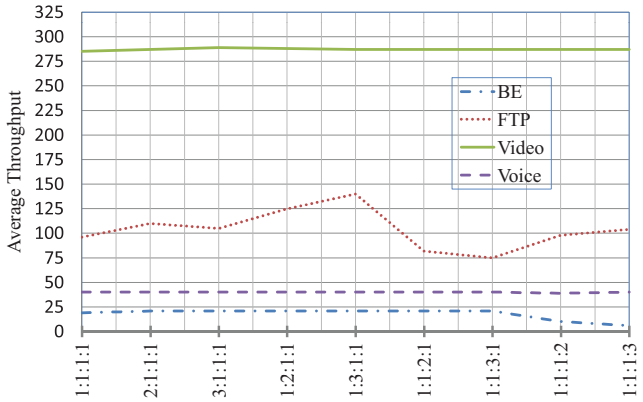


Fig. 1. Average throughput in (Kbps) of BE, FTP, video and voice under light loaded network.

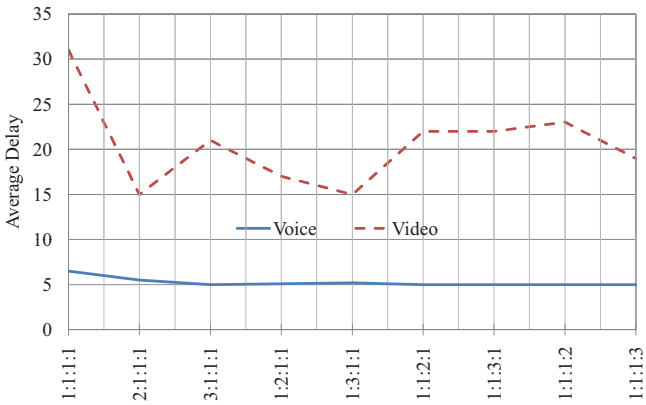


Fig. 2. Average delay in (ms) of video and voice traffic under light loaded network.

(MRTR=12.5kbps and MSTR=64kbps). The latter justification is also emphasized by observing the video traffic results, which show the highest throughput under the loading ratio 3:1:1:1. Since video is prioritized second after voice, and because the requirements of the voice connection are small, the bandwidth available for video traffic is higher in this case than that under other loading ratios. Video traffic, consequently, exploits this availability to achieve the higher throughput.

Since the network is lightly loaded, the average delay in Fig. 2 is much lower than classes' delay bounds. Similarly, the packet loss of the audio and video traffic in this experiment is zero.

Figure 3 shows the results when the network is heavily loaded. This is an important experiment as it evaluates our scheme in a congestion setting. The figure shows that the scheme was able to adapt to the network status as the number

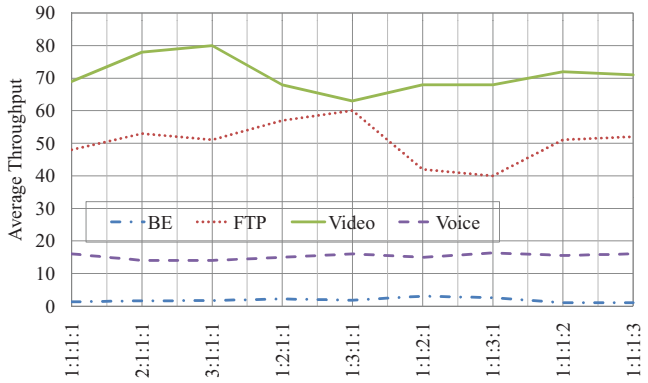


Fig. 3. Average throughput in (Kbps) of BE, FTP, video and voice under heavy loaded network.

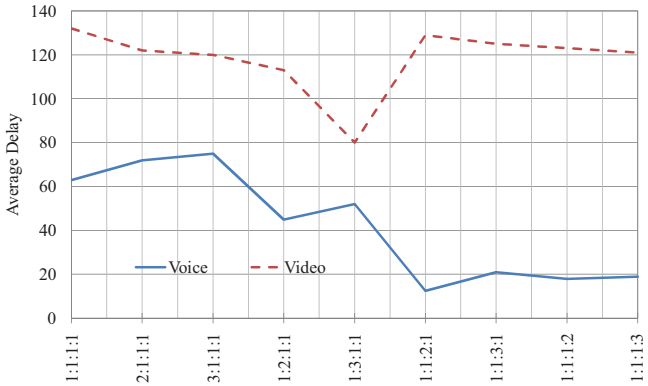


Fig. 4. Average delay in (ms) of video and voice traffic under heavy loaded network.

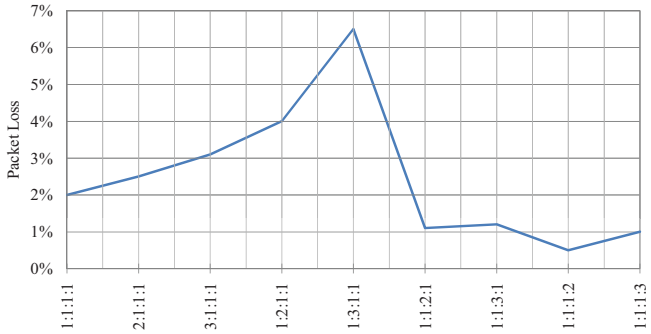


Fig. 5. Average packet loss of video traffic under heavy loaded network.

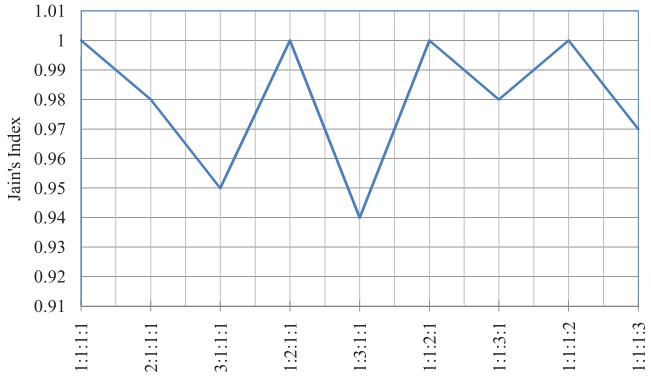


Fig. 6. Jains index fairness among voice, video, FTP and BE connection classes.

of connections changed for the different classes. We observe that the lowest voice throughput occurs at loading ratios 2:1:1:1 and 3:1:1:1. Understanding this outcome is intuitive, because in these loading ratios, more voice traffic connections are competing for the allocated bandwidth. The same behavior can be observed for video, FTP and background traffic. However, the average throughput of each class (1kbps for BE, 45kpbs for FTP, 64kbps for video and 14kbps for voice) shows that the minimum bandwidth required by each class is achieved, despite the fact that the connections are receiving lower allocations than the lightly loaded experiment.

Figure 4 also shows the effect of increasing the number of FTP connections on the average video throughput. The value of the average throughput of video traffic is 68kbps and 67kbps when the loading ratio is respectively 1:1:2:1 and 1:1:3:1. These throughput values are almost the required minimum bandwidth of the video traffic. Since the allocated bandwidth for the FTP class is larger, the allocated bandwidth for the video traffic is smaller. Recall that both classes have high bandwidth requirements (45kbps for FTP and 64kbps for video). More importantly, the scheme's response indicates its robustness in handling congestion instances.

In Fig. 4, the delay of voice and video traffic is higher than that of the light load setup. In the figure, the measured delay does not exceed the delay bounds for VoIP and video traffic. Note that packets queued beyond their delay bounds are dropped and their delay is not included in calculating the average delay. The figure shows that the delay of the voice is less when the network is more loaded with FTP and background traffic (1:1:2:1, 1:1:3:1, 1:1:1:2 and 1:1:1:3) than when more loaded with video traffic (loading ratios 1:2:1:1 and 1:3:1:1). This is due to the fact that FTP and BE traffic are delay insensitive, while video traffic is a delay sensitive traffic. Accordingly, voice and video packets approaching their delay bound will compete for the current time slot of transmission at the scheduler. The results in the figure also reveal that the delay increases as the number of connections increases. This is due to the contention among connections of the same class. We remark here that the major influence on connections' delay performance is due to the scheduler. However, the bandwidth provisioning algorithm controls the delay through controlling the amount of time-slots allocated to each real time class. It is also worth observing that the difference between the average delay and the delay bound for voice traffic is smaller than that of the video traffic, and that the average delay for video traffic is almost at its required delay bound at all loading ratios. These results illustrate the effectiveness of the proposed algorithm as it consistently prioritizes voice over video.

For packet loss, the results in Fig. 5 should be coupled with those in Fig. 4. Figure 5 shows that the packet loss has the largest value at a loading ratio of 1:3:1:1, while Fig. 4 shows minimum delay at this loading ratio. Given that the delay of dropped packets is not included in the delay calculations, and that the packets dropped from the head of the queue allow other queued packets to be sent with lower delay, the delay of the video class is lower at this loading ratio.

Figure 6 presents the results of fairness among the different classes with the network heavily loaded. The results show that maximum fairness (i.e. fairness index = 1) is achieved when

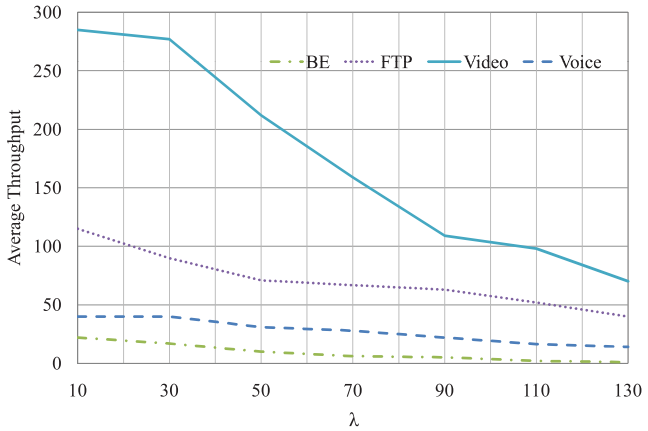


Fig. 7. Average throughput in (Kbps) of BE, FTP, video and voice for loading ratio 1:1:1:1 vs. the pricing factor λ.

all traffic types are equally loaded (i.e. loading ratio 1:1:1:1). The fairness index decreases when any traffic type is loaded 3 times the minimum ratio (i.e. at loading ratios 3:1:1:1, 1:3:1:1, 1:1:3:1 and 1:1:1:3). This decrease is especially pronounced with voice or video (i.e. 3:1:1:1 and 1:3:1:1). Because these two classes have highest priorities, scheduling and bandwidth allocation of these classes reduce resources left for the FTP and BE class, directly decreasing the fairness index. Despite this reduction, however, the index is maintained at a relatively large value (0.94). The figure hence reveals that our objective of designing a fair congestion control resource management scheme is fulfilled. The fairness results under light load are not shown as the experiments always yield a fairness index of 1 for all loading ratios.

Finally, Fig. 7 shows the average throughput against different values of the pricing factor λ. As the figure shows, an increase of the pricing factor results in decreasing the average throughput. Hence, the pricing factor λ is effective in preventing greedy use of network resources and in controlling congestion as desired. Nevertheless, the required MRTR of connections is maintained even for large values of λ.

VII. Conclusions and Future Work

In this paper, we addressed the void for a congestion control scheme in BWA networks by presenting a game-theoretic model that controls congestion through dynamic pricing while providing for fairness and QoS guarantees. The scheme consists of two congestion-based pricing sub-schemes, each over a different operational phase: one for bandwidth provisioning at the class level operating in an inter-class phase and the other for scheduling at the connection level operating in an intra-class phase. We defined characteristics and requirements to design an efficient and robust interclass bandwidth provisioning algorithm. These characteristics are used as the guidelines for the design and the performance evaluation of the interclass algorithm. The interclass algorithm is designed as a general algorithm irrespective of the underlying technology as long as it is time frame based. Simulation results show that the proposed scheme at the class level is able to meet these characteristics. Similarly, the scheme was able to meet the QoS requirements at the connection level for all connections in each class while maintaining fairness objectives.

The proposed scheme is implemented in a municipality WiFi solar-powered test-bed. The proposed scheme is coded in C- language and integrated into an open-source Linux based operating system of the access points. As for future work, the authors will study the performance of the scheme in mitigating congestion and meeting the QoS requirements of different types of real-time and non real-time applications carried by the municipality WiFi test-bed.

References

[1] IEEE Standard 802.16 Working Group, IEEE 802.16-2005e Standard for Local and Metropolitan Area Networks: Air interface for fixed broadband wireless access systems - amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands, Dec. 2005.

[2] 3GPP TS 36.300 Version 9, Technical Specification 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall description; Stage 2 (Release 9), June 2009.

[3] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance evaluation of the IEEE 802.16 MAC for QoS support," *IEEE Trans. Mobile Computing*, vol. 6, pp. 26–38, 2007.

[4] K. Vinay, N. Sreenivasulu, D. Jayaram, and D. Das, "Performance evaluation of end-to-end delay by hybrid scheduling algorithm for QoS in IEEE 802.16 network," *International Conference on Wireless and Optical Communications Networks 2006 (IFIP)*, Aug. 2006.

[5] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 839–847, May 2006.

[6] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proc. International Conference on Wireless Communications, Networking and Mobile Computing 2007 (WiCom 2007)*, pp. 2861–2864, Sep. 2007.

[7] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "Mac scheduling scheme for VoIP traffic service in 3g LTE," in *Proc. IEEE 66th Vehicular Technology Conference (VTC-Fall 2007)*, pp. 1441–1445, Oct. 2007.

[8] L. Ruiz de Temino, G. Berardinelli, S. Frattasi, and P. Mogensen, "Channel-aware scheduling algorithms for SC-FDMA in LTE uplink," in *Proc. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2008)*, pp. 1–6, Sep. 2008.

[9] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[10] E. B. Rodrigues and F. R. P. Cavalcanti, "QoS-driven adaptive congestion control for voice over IP in multiservice wireless cellular networks," *IEEE Commun. Mag.*, vol. 46, no. 1, pp. 100–107, Jan. 2008.

[11] B. Rong, Y. Qian, and H.-H. Chen, "Adaptive power allocation and call admission control in multiservice WiMAX access networks: radio resource management and protocol engineering for IEEE 802.16," *IEEE Wireless Commun.*, vol. 14, no. 1, pp. 14–19, Feb. 2007.

[12] J. A. Perez, B. Donnet, and O. Bonaventure, "Preliminary analysis of the TCP behavior in 802.16 networks," in *Proc. WEIRD Workshop on WiMAX, Wireless and Mobility*, 2006.

[13] D. Niyato and E. Hossain, "Integration of WiMAX and WiFi: optimal pricing for bandwidth sharing," *IEEE Commun. Mag.*, vol. 45, no. 5, pp. 140–146, May 2007.

[14] V. Rakovic, O. Ognenoski, L. Gavrilovska, M. Bogatinovski, and V. Atanasovski, "User perception of QoS and economics for a WiMAX network in a backup scenario," in *Proc. Wireless VITAE 2009*.

[15] M. Chatterjee, S. Sengupta, and R. Chandramouli, "Dynamic pricing for service provisioning and network selection in heterogeneous networks," *Physical Commun.*, vol. 2, no. 1, pp. 138–150, Mar. 2009.

[16] P. Maille, A. Belghith, and L. Nuaymi, "Pricing of differentiated-QoS services WiMAX networks," in *Proc. Global Telecommunications Conference (GLOBECOM 2008)*.

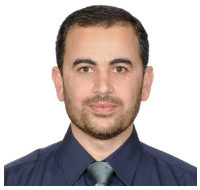
[17] J. Zhang, L. Zhao, and H. Zhang, "Using incompletely cooperative game theory in wireless mesh networks," *IEEE Network*, vol. 22, no. 1, pp. 39–44, Jan. 2008.

[18] R. Jayaparvathy and S. Geetha, "Resource allocation and game theoretic scheduling in IEEE 802.16 fixed broadband wireless access systems," in *Proc. 2008 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'08)*.

[19] D. Ross, *What People Want: The Concept of Utility from Bentham to Game Theory*. University of Cape Town Press, South Africa, 1st edition, 1999.

[20] D. Fudenberg and J. Tirole, *Game Theory*. The MIT Press, 1991.
[21] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.
[22] C. Papadimitriou, A. Fabrikant, and K. Talwar, "The complexity of pure Nash equilibria," in *Proc. Thirty-Sixth Annual ACM Symposium on theory of Computing*, 2004.
[23] B. Kim and Y Hur, *Application Traffic Model for WiMAX Simulation*. POSDATA, Ltd, 2007.

Najah AbuAli received her B.Sc. and M.Sc. degrees in Electrical Engineering in 1989 and 1995 respectively from University of Jordan, Amman, Jordan and her PhD degree in 2006 in Computer Networks in Electrical Engineering department at Queen's University, Kingston, Canada. She joined the College of Information Technology, United Arab Emirates University (Al Ain, UAE), as an Assistant Professor with the Computer Networks Engineering track. She had a postdoctoral fellowship at the School of Computing, Queen's University from January 2006 to August 2006. She worked as an instructor and the head of the Engineering Department at Queen Noor College in Jordan from 1995 to 2003. Her research interests comprise wired and wireless communication networks. Specifically, analytical and measurement based network performance management and Quality of Service and resource management of single and multi-hop wireless networks.



Mohammad Hayajneh (SM'01, M'04) received his B.Sc and M.Sc in Electrical Engineering majoring in Electronics and Communications from Jordan University of Science and Technology in 1995, 1998, respectively. He received his PhD in Electrical Engineering from University of New Mexico (Albuquerque, USA) in 2004. In September 2004 he joined the College of Information Technology, United Arab Emirates University (Al Ain, UAE), as an Assistant Professor with the Computer System Engineering track. Before joining the PhD program at University of New Mexico, Dr. Hayajneh held several positions in the academia: he was lecturer and head of the Telecom. Department in the Institute of Science for Telecom and Technology in Riyadh, KSA from 2000-2001. And from 1998-2000 he was an instructor at Princess Sumaya University of Technology, (PSUT) Amman, Jordan. Dr. Hayajneh research interests include: power control for wireless data networks, WiMAX cross-layer scheduling, resource allocation and pricing in wireless data networks, MIMO systems, performance analysis and modeling of OFDMA based wireless networks, Cooperative networks.

Hossam Hassanein is with the School of Computing at Queen's University working in the areas of broadband, wireless and variable topology networks architecture, protocols, control and performance evaluation. Dr. Hassanein obtained his Ph.D. in Computing Science from the University of Alberta in 1990. He is the founder and director of the Telecommunication Research (TR) Lab http://www.cs.queensu.ca/ trl in the School of Computing at Queen's. Dr. Hassanein has more than 350 publications in reputable journals, conferences and workshops in the areas of computer networks and performance evaluation. He has delivered several plenary talks and tutorials at key international venues, including Unconventional Computing 2007, IEEE ICC 2008, IEEE CCNC 2009, IEEE GCC 2009, IEEE GIIS 2009, ASM MSWIM 2009 and IEEE Globecom 2009. Dr. Hassanein has organized and served on the program committee of numerous international conferences and workshops. He also serves on the editorial board of a number of International Journals. He is a senior member of the IEEE, and is currently chair of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). Dr. Hassanein is the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research award in 2003. He received several best paper awards, including at IEEE Wireless Communications and Network (2007), IEEE Global Communication Conference (2007), IEEE International Symposium on Computers and Communications (2009), IEEE Local Computer Networks Conference (2009) and ACM Wireless Communication and Mobile Computing (2010). Dr. Hassanein is an IEEE Communications Society Distinguished Lecturer.