Abstract—Efficient support of multimedia services in next generation wireless networks requires advanced scheduling schemes that achieve both system capacity maximization and full QoS differentiation. This paper proposes a new MAC scheduling scheme for multiuser OFDM wireless networks. Designed in an opportunistic cross-layer approach, this scheme uses a system of weights that dynamically adjusts the priority of the flows considering the transmission conditions and the QoS that they currently experience. This results in a very efficient management of the bandwidth. Performance evaluation shows that the proposed scheduling outperforms existing wireless OFDM based scheduling schemes providing full QoS differentiation, high fairness and system throughput maximization.

Index Terms—Multiple access, multimedia, QoS, Orthogonal Frequency Division Multiplexing, multiuser diversity, opportunistic scheduling, cross-layer design.

I. INTRODUCTION

Considerable research effort has been given to the design of efficient modulation techniques for next generation wireless networks (4G systems). Orthogonal Frequency Division Multiplexing (OFDM) is now widely acknowledged as the most promising physical layer technique for broadband transmission. Supporting multimedia services in these networks now requires the design of fair and efficient multiuser OFDM scheduling schemes with adequate QoS support. The most advanced scheduling algorithms focus on maximizing the system capacity taking advantage of multiuser diversity. Mainly two classes of schemes emerge in the literature: MaxSNR based schemes [1], [2], [3] and Proportional Fair (PF) based schemes [4], [5], [6], [7]. In MaxSNR, priority is given at every scheduling event to the mobiles which have the greatest signal-to-noise ratio (SNR). In PF based schemes, the bandwidth is allocated to the mobiles with the best channel state with respect to their time average.

MaxSNR and PF are both very good at maximizing the system capacity. However these schemes were designed for scheduling service flows with same traffic profile and identical QoS requirements like voice; multimedia services and QoS differentiation are not supported. Recently, [8] proposed the Multimedia Adaptive OFDM Proportional Fair (MAOPF) algorithm, an evolution of the classical PF for considering multimedia services. The principle of the MAOPF is to share the total available bandwidth among users proportionally to their bit rate requirement. Although this enables the coexistence of applications with unequal bit rates, heterogeneous QoS requirements are still not well supported. Moreover, the MAOPF allocates all OFDM subcarriers to the same mobile. This does not fully take advantage of the multiuser diversity and has a negative impact on the system capacity.

In this paper we proposed a new MAC scheduling scheme called “Weighted Fair Opportunistic” (WFO)1 for full support of multimedia services with adequate QoS. The proposed scheme was designed for best profiting of the multi-user diversity taking advantage of the dynamics of the multiplexed traffics. The WFO jointly considers the required QoS, the experienced QoS and the transmission conditions in an extended cross-layer approach. Based on weights, the scheduler dynamically favors the flows that go through a critical period and always attributes the adequate priorities for improved QoS support.

The paper is organized as follows. Section II provides a detailed description of the system under study. Section III introduces a generic approach for QoS management and Section IV defines the proposed WFO scheduling algorithm. In Section V, a detailed performance evaluation through simulation is presented, before concluding.

II. SYSTEM DESCRIPTION

The paper focuses on the proper allocation of radio resources among the set of mobiles situated in the coverage zone of an access node. In a centralized approach, the packets originating from the backhaul network are buffered in the access node which schedules the downlink transmissions. In the uplink, the mobiles signal their traffic backlog to the access node which builds the uplink resource mapping. We assume that the physical layer is operated using the structure described in Fig. 1 which ensures a good compatibility with the OFDM based transmission mode of the IEEE 802.16-2004 [9], [10]. The total available bandwidth is divided in sub-frequency bands or subcarriers. The radio resource is further divided in the time domain in frames. Each frame is itself made of time slots of constant duration, integer multiple of the

1This work has been partially supported by the IST European project WIP under contract 27402.
OFDM symbol duration. The number of subcarriers is chosen so that the width of each sub-frequency band is inferior to the coherence bandwidth of the channel. Moreover, the frame duration is fixed to a value much smaller than the coherence time of the channel. With these assumptions, the transmission on each subcarrier is subject to flat fading with a channel state that can be considered static during each frame.

The elementary resource unit (RU) is defined as any (subcarrier, time slot) pair. Each RU may be allocated to any mobile with a specific modulation order. Transmissions performed on different RUs by different mobiles are assumed to have independent channel state variations. On each RU, the modulation scheme is QAM with a modulation order adapted to the channel state between the access node and the mobile to which it is allocated. This provides the flexible resource allocation framework required for opportunistic scheduling.

The system is operated using time division duplexing with four subframes: the downlink feedback subframe, the downlink data subframe, the uplink contention subframe and the uplink data subframe. The uplink and downlink data subframes are used for the transmissions of user data. In the feedback and contention subframes, control information is communicated between the mobiles and the access node. This frame structure supposes time and frequency synchronization between the mobiles and the access node as described in [11]. Therefore, each frame starts with a long preamble used for synchronization purposes.

III. QoS MANAGEMENT

A crucial objective for modern multiple access schemes is to fully support multimedia transmission services, including the widest range of services: VoIP, videoconference, email, file transfer... This requires the coexistence of delay sensitive flows as well as non real time traffic with looser delay constraints but with tight data integrity targets. In order to deal with the various and heterogeneous QoS requirements of multimedia services, the proposed scheme relies on a generic approach of QoS management.

We define a service flow as a traffic stream and its QoS profile, in a given transmission direction. A mobile may have multiple service flows both in the uplink and the downlink. An application may also use several service flows enabling for instance the implementation of Unequal Error Protection schemes in the physical layer. Each service flow possesses its own transmission buffer. In the following, index $k$ is used to designate a given service flow among the set of service flows to be scheduled in a given transmission direction.

The QoS profile is defined as the set of parameters that characterizes the QoS requirements of a service flow mainly in terms of data integrity and delay. In the following, data integrity requirements are specified by a Bit Error Rate (BER) target, denoted $BER_{\text{target},k}$ for service flow $k$. Regarding delay requirements, the meaningful constraint is a limitation of the occurrences of large delay values. By analogy with the concept of outage used in system coverage planning, we define the concept of delay outage. A service flow is in delay outage when its packets experience a delay greater than a given threshold. The Packet Delay Outage Ratio (PDOR) target is defined as the maximum ratio of packets that may be delivered after this fixed delay threshold. In the following, $T_k$ represents the delay threshold of service flow $k$. This characterizes the delay requirements of any service flow in a generic approach.

The PDOR experienced by each service flow is tracked all along its lifetime. At each transmission of a packet of service flow $k$, the total number of packets whose delay exceeded the threshold divided by the total number of packets transmitted since the beginning of the connection is computed. The result is denoted $PDOR_k$. Fig. 2 illustrates an example cumulative distribution of the packet delay of service flow $k$ at a given time instant. The objective of the WFO is to regulate the experienced PDOR along the lifetime of the service flow so that its value stays below the PDOR target. This ensures the
satisfaction of the delay requirements at a short time scale.

In the WFO scheduling, QoS management comprises two parts: data integrity management and delay management. Data integrity is guaranteed by the physical layer mainly by adapting the modulation scheme and the transmit power to the mobile specific channel state. This is achieved considering each service flow independently. Delay management is performed considering all service flows jointly and scheduling the packets according to their distance to their PDOR target. The joint consideration of the delay constraints relies on the dynamics of the traffic streams that are multiplexed. Fairness consists in guaranteeing the same PDOR to all service flows.

IV. THE WFO SCHEDULING ALGORITHM

The WFO scheduling is performed during the uplink data transmission phase. The scheduler, located in the access node, grants RUs to each service flow as a function of:

- its QoS profile (BER target, delay threshold and PDOR target);
- its currently experienced QoS (BER and PDOR);
- its traffic backlog;
- its channel state.

The QoS profile is signaled in the connection establishment phase. In the uplink, the currently experienced PDOR and the traffic backlog (buffer occupancy) are signaled by the mobile in the contention subframe. The experienced BER is tracked directly by the access node. Reciprocally, in the downlink, the currently experienced PDOR and the traffic backlog are calculated by the access node and the experienced BER is signaled.

Additionally, knowledge of the channel state is supposed to be available at the receiver. The current channel attenuation $a_{k,n}$ experienced by service flow $k$ on subcarrier $n$ (including path loss and multipath fading):

$$P_r(q,k) \leq a_{k,n} P_{\text{max}}.$$  

Hence, the maximum number of bits $q_{k,n}$ of service flow $k$ which can be transmitted on a time slot of subcarrier $n$ while keeping below its BER target is:

$$q_{k,n} \leq \left\lfloor \log_2 \left( 1 + \frac{3P_{\text{max}} \times T_k \times a_{k,n}}{2N_0 \left( \text{erfc}^{-1} \left( \frac{BER_{\text{target},k}}{2} \right) \right)^2} \right) \right\rfloor. \quad (4)$$

We further assume that the supported QAM modulation orders are limited so that $q$ belongs to the set $S = \{0, 2, 4, \ldots, q_{\text{max}}\}$. Hence, the maximum number of bits $m_{k,n}$ that will be transmitted on a time slot of subcarrier $n$ if this RU is allocated to the service flow $k$ is:

$$m_{k,n} = \max \{s \in S, q \leq q_{k,n}\}. \quad (5)$$

MaxSNR based schemes allocate the resources to the flows which have the greatest $m_{k,n}$ values. We introduce a new parameter which modulates this pure opportunistic resource allocation in order to provide full and fair QoS support while preserving the system throughput maximization. This parameter called "Weighted Fair" (WF) parameter is based on...
the current estimation of the PDOR of service flow $k$ and defined by:
\[ WF_k = f(PDOR_k), \]
where $f$ is a strictly positive and monotonically increasing function. The WFO scheduling principle is then to allocate a time slot of subcarrier $n$ to the mobile $k$ which has the greatest WFO parameter value $WFO_{k,n}$ with:
\[ WFO_{k,n} = WF_k \times m_{k,n}. \]
Based on the PDOR, the WF parameters directly account for the level of satisfaction of the delay constraints for an efficient QoS management. The PDOR is more relevant and simpler to use than the service flow throughput, the buffer occupancy or the waiting time of each packet to schedule which would introduce a great complexity in the scheduling algorithm. The WFO parameters introduce dynamic priorities that delay the flows which currently easily respect their delay threshold to the benefit of others which go through a critical period.

Our studies on the algorithm performance have shown that a polynomial function $f$ suits well:
\[ f(x) = 1 + \beta x^\alpha. \]
The influence of the value of the $(\alpha, \beta)$ pair on the performances of the WFO scheduling scheme was analyzed by extensive simulations in order to optimally calibrate $f(x)$. The exponent parameter $\alpha$ allows being more sensitive and reactive to PDOR fluctuations which guarantees fairness at a short time scale. Several values of $\alpha$ suit well but a cubic exponent offers the best reactivity to PDOR fluctuations. $\beta$ is a normalization parameter that ensures that $WF_k$ and $m_{k,n}$ are in the same order of magnitude. If $\beta$ is too low, the WF parameter has no influence and fairness decreases. On the contrary, if $\beta$ is too high, $m_{k,n}$ looses weight in the scheduling and the system throughput is not maximized. Given that $PDOR_k$ has an order of magnitude $10^{-2}$, $\beta$ should roughly be set to $10^{0.5}$. Simulations have shown that the good values of $\beta$ range between $10^5$ and $10^6$ with $\alpha$ fixed equal to 3. In the following, $\alpha$ is assumed to be equal to 3 and $\beta$ is taken equal to $10^6$. With this choice, $WF_k$ is always in the same order of magnitude as $m_{k,n}$ and allows to manage both fairness and system throughput optimisation.

The dynamic priorities introduced by the WFO algorithm evolve as a function of the specific channel conditions and currently experienced QoS of each service flow in a cross-layer higher layers/MAC/PHY approach. This results in a well-balanced resource allocation which keeps a maximum number of service flows active across time but with continuously relatively low traffic backlogs. Preserving this multiuser diversity allows to continuously take a maximal benefit of opportunistic scheduling maximizing the spectral efficiency. Additionally, this also achieves a time uniform fair allocation of the RUs to the service flows ensuring the required short term fairness [14], [15].

The WFO scheduling algorithm is detailed in Fig. 3. The scheduling is run subcarrier by subcarrier and on a time slot basis for improved granularity. In the allocation process of a given time slot, the priority of a service flow with respect to another is determined by the magnitude of its WFO parameter. All service flows are scheduled simultaneously in a single run of the algorithm, whatever their QoS profile. QoS differentiation is achieved by means of the WFO parameters. Service flows with low delay constraints like best effort traffic are qualified with a quite high delay threshold. As a result, their PDOR are always very small compared to other low latency traffics whose priority increases dramatically as soon their smaller delay threshold are not respected. In the following, the proposed scheduling algorithm is described step by step.

- **Step 0:** The scheduler refreshes the current $PDOR_k$ and buffer occupancy $BO_k$ values of each service flow $k$ and computes the $m_{k,n}$, $WF_k$, and $WFO_{k,n}$ parameters for each service flow and each subcarrier. Then, $n$ and $t$ are initialized to 1.
- **Step 1:** For subcarrier $n$, the scheduler selects the service
flow $k$ with the greatest $WFO_{k,n}$ value.

- Sub-step 1-1: If the virtual buffer occupancy$^2$ of service flow $k$ is positive, the scheduler goes to Sub-step 1-2. Else, if all virtual buffers are null or negative, the scheduler goes to Step 2. Otherwise, the scheduler selects the next service flow $k$ with the greatest $WFO_{k,n}$ value and restarts Sub-step 1-1.

- Sub-step 1-2: The scheduler allocates time slot $t$ of subcarrier $n$ to service flow $k$ with a capacity of $m_{k,n}$ bits, removes $m_{k,n}$ bits of its virtual buffer and increments the value of $t$. If $t$ is smaller than the maximum number $t_{\text{max}}$ of time slots by subcarrier, go to Sub-step 1-1 for allocating the next time slot. Else, go to next sub-step.

- Sub-step 1-3: Increment the value of $n$. If $n$ is smaller than the maximum number $n_{\text{max}}$ of subcarriers, go to Step 1 for allocating the time slots of the next subcarrier. Otherwise, go to Step 2.

- Step 2: All virtual buffers are empty or all time slots of all subcarriers are allocated and the scheduling ends.

V. PERFORMANCE EVALUATION

In this section we compare the proposed Weighted Fair Opportunistic scheduling with the MaxSNR, the PF and the MAOPF implemented with subcarrier by subcarrier allocation. Performance evaluation results are obtained using OPNET discrete event simulations.

$^2$We define the virtual buffer occupancy as the current buffer occupancy of service flow $k$ minus the number of bits already allocated to this service flow.

A. Simulation Setup

In the simulations, a frame is composed of 5 time slots and 128 subcarriers. We consider a multipath Rayleigh fading channel model, i.e. the attenuation experienced by service flow $k$ if transmitted on subcarrier $n$ is:

$$a_{k,n} = \alpha_{k,n}^2,$$

where $\alpha_{k,n}$ represents a Rayleigh distributed random variable with an expectancy equal to unity.

We chose to test our scheme with a very demanding type of application. In the simulations presented below, all mobiles run a videoconference application. Each mobile has only one service flow with a traffic composed of an MPEG-4 video stream [16] and an AMR voice stream [17]. This generates a high volume of data with high sporadity and requires tight delay constraints which substantially complicates the task of the scheduler. The BER target was set to $10^{-3}$ for all mobiles.

B. Simulation Results

1) Performance evaluation with heterogeneous bit rate sources: We first study the scheduling performances with heterogeneous bit rate traffics (Fig. 4). Mobiles are divided in two groups. Group 1 is constituted of 9 mobiles each with an average bit rate of 80 Kbps while Group 2 is composed of 3 mobiles with an average bit rate of 240 Kbps. For all mobiles, the delay threshold $T_k$ is fixed to 80 ms. The maximum value of the SNR ($P_{\text{max}}T_s/N_0$) is set to 22 dB. The four scheduling strategies provide the same bandwidth usage ratio of 82%. However delay management considerably differs.

Fig. 4(a) shows the overall ratio of packets delivered after the...
threshold time respectively in Group 1, Group 2 and globally. The results show that the MaxSNR and the PF easily respect the delay constraints of low bit rate mobiles but fail for the second group of mobiles. In contrast, the MAOPF and the WFO schemes provide fairness with an equal and moderate ratio of packets in delay outage whatever the source bit rate. The overall PDOR obtained with the MAOPF and the WFO is smaller than with the two other schemes. Here, the two multimedia oriented schedulers provide fair QoS management and better QoS support.

Another comparison criterion was the QoS satisfaction level that each mobile perceives across the lifetime of a connection. The connection of each mobile was divided in cycles of five minutes representative of the duration of a session and we checked at the end of each cycle if the delay constraint was met or not. Then the number of times that the mobile is not satisfied (experienced PDOR \( \geq P D O R_{\text{target}} \)) divided by the total number of cycles was computed. Considering the usual \( P D O R_{\text{target}} \) of 5%, Fig 4(b) shows that the WFO outperforms the other schedulers including the MAOPF which do not directly manage the PDOR fluctuations. Fig 4(c) confirms this observation for a \( P D O R_{\text{target}} \) of 10%. 100% satisfaction is obtained with the WFO in both groups while the MaxSNR and the PF provide this satisfaction level only for Group 1 mobiles.

2) Performance evaluation with heterogeneous delay constraints: The scheduling performances in a context of heterogeneous delay requirements was then studied. Mobiles with identical traffic sources, all generating an average bit rate of 80 Kbps were considered. The maximum value of the SNR is set to 15 dB. Two groups of 7 mobiles were constituted and the delay threshold \( T_k \) was fixed to 80 ms in the first group and 250 ms in the second group. In this context where all mobiles have an equal source bit rate, the MAOPF and the PF perform the same scheduling. Fig. 5 clearly shows that the WFO outperforms the other schemes ensuring fair QoS support and provides the largest QoS satisfaction level. Concretely, the MaxSNR, the PF and the MAOPF manage all mobiles identically (without QoS differentiation) while some of them need more attention. These approaches do not optimize the global mobile satisfaction. Indeed, it could be more judicious to take benefit of the breathing space offered by the easy satisfaction of mobiles with loose delay requirements to help the others with tighter delay constraints. With its weighted system, the WFO dynamically adjusts the relative priority of the flows according to their experienced delay and efficiently regulates the PDOR, fully supporting QoS differentiation.

VI. CONCLUSION

In this paper, we proposed a new scheduling scheme, called "Weighted Fair Opportunistic". It operates on top of an OFDM-based physical layer and shows a good compatibility with the existing 802.16 standard. Designed for both the uplink and the downlink, the full support of advanced multimedia services with QoS differentiation is enabled by the introduction of generic QoS attributes. Data integrity and delay requirements are managed in a Higher Layers/MAC/PHY cross-layer approach. Using dynamic priorities, the WFO keeps the traffic backlog of each service flow at a continuously low level. A maximum number of service flows are preserved active across time which takes the best benefit of the multiuser diversity. This results in a fair and efficient resource management scheme fully adapted to advanced multimedia traffics with heterogeneous bit rates and QoS constraints. A great and equal satisfaction among mobiles whatever their respective traffic profile and QoS requirements is provided. The performance analysis shows that the WFO outperforms the best existing wireless OFDM based scheduling schemes in terms of QoS satisfaction and fairness.

REFERENCES