Scheduling in OFDM Wireless Networks without Tradeoff between Fairness and Throughput

Cédric Gueguen and Sébastien Baey
UPMC Univ. Paris 06
104, avenue du Président Kennedy, 75016 Paris, France
Email: {cedric.gueguen, sebastien.baey}@lip6.fr

Abstract—Providing multimedia services in wireless networks requires maximizing the system throughput without sacrificing fairness. Previous works focus on system capacity optimization but fail to jointly ensure adequate fairness. This paper proposes a new MAC scheduling scheme which dynamically takes into consideration the QoS experienced by the mobiles and the transmission conditions in an extended cross-layer design. Based on a weighted opportunistic algorithm, our resource allocation takes a maximum advantage of multiuser diversity optimizing fairness and system throughput. This provides an efficient support of multimedia services in OFDM wireless networks. Performance evaluation shows that the proposed scheduling widely outperforms the best existing wireless OFDM based scheduling schemes.

Index Terms—Orthogonal Frequency Division Multiplexing, multipath fading, multiuser diversity, opportunistic scheduling, cross-layer design, multimedia, Medium Access Control.

I. INTRODUCTION

In wireless networks, providing an efficient use of the scarce bandwidth is challenging. At the MAC level, many research efforts concentrated on maximizing the cell throughput or ensuring fairness but jointly providing both is not currently well solved yet. In the literature, two classes of multiple access schemes emerge: MaxSNR [1], [2], [3] and Proportional Fair (PF) scheduling [4], [5]. These schedulers are based on opportunistic resource allocation that takes profit of multiuser diversity. At the physical level, Orthogonal Frequency Division Multiplexing (OFDM) efficiently reduces the harmful effects of multipath fading in wireless transmissions [6]. It is now widely recognized as the reference modulation technique for next generation broadband wireless networks (4G systems) and is already implemented in 802.11a/g or 802.16. Operated on top of an OFDM based physical layer, MaxSNR and PF maximize the system capacity.

However, in spite of their high performances in terms of system throughput maximization, both MaxSNR and PF suffer of severe fairness deficiencies owing to unequal spatial positioning of the mobiles. MaxSNR always gives the priority to the mobiles which have the greatest signal-to-noise ratio (SNR) value. With this strategy, the closer mobiles to the access point have a disproportionate priority over mobiles more distant resulting in severe lack of fairness. PF based schemes allocate units of bandwidth to the mobiles which have the best channel state with respect to their respective time average. However, at a short time scale, channel state variations are mainly due to multipath fading, statistically similar for all mobiles. Since all mobiles experience the same channel state variations around their mean, all mobiles obtain an equal number of radio resource units across time. This results in an equal sharing of the total available bandwidth. However, since the farther mobiles have a lower spectral efficiency than the closer ones due to pathloss, the mobiles do not all benefit of an equal average throughput which induces unequal delays [7]. [8], [9] bring solutions for fighting the impact of path loss on fairness but always decreasing the system throughput in exchange.

In this paper we propose the “Weighted Fair Opportunistic” (WFO) scheduling scheme1. The WFO jointly considers both the transmission conditions, the currently measured/experienced QoS and the QoS targets of the mobiles in the bandwidth allocation process. With an original weighted system that introduces dynamic priorities between the flows, the WFO keeps a maximum number of flows active across time but with relatively low traffic backlogs. Thus preserving the multiuser diversity, the WFO takes a maximal benefit of the opportunistic scheduling technique and maximizes the system capacity. Additionally, this also achieves a time uniform fair allocation of the resource units to the flows ensuring short term fairness [10], [11]. This higher layers/MAC/PHY cross-layer approach better conceals the system capacity maximization and fairness objectives.

The paper is organized as follows. Section II gives a description of the system under study. In section III, the scheduling algorithm is presented. Performances are analyzed in section IV. Section V concludes the paper.

II. SYSTEM DESCRIPTION

The physical layer is operated using a structure compatible with the OFDM based transmission mode of the IEEE 802.16-2004 [12]. The total available bandwidth is divided in subfrequency bands or subcarriers. The radio resource is further divided in the time domain in frames and each frame is itself divided in time slots of constant duration. The time slot duration is an integer multiple of the OFDM symbol duration. They are assumed equal in the following. The number of subcarriers is chosen so that the width of each subfrequency band is inferior to the coherence bandwidth of the

---

1This work has been partially supported by the IST European project WIP under contract 27402.
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. In the following, an elementary resource unit (RU) is defined as any (subcarrier, time slot) pair. Each of this RU may be allocated to any mobile with a specific modulation order. The 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 point and the mobile to which it is allocated. This provides the flexible resource allocation framework required for opportunistic scheduling.

In the following, a service flow is defined as a traffic stream and its associated QoS profile, in a given transmission direction. The QoS profile of a service flow is defined as the set of parameters that characterizes its QoS requirements mainly in terms of data integrity but also in terms of delay. The QoS profile is signaled in the connection establishment phase. 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. The 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 resource allocation is considered in a centralized approach. It is performed independently in the uplink and the downlink for the set of mobiles located in the coverage zone of an access point and on a frame by frame basis. The scheduler, located in the access point, grants RUs to each service flow as a function of its required QoS, its currently experienced QoS, its traffic backlog (buffer occupancy), and its channel state. Depending on the related transmission direction, these informations are either signaled by the mobile or directly tracked by the access point.

### III. The WFO Scheduling

The WFO scheduling algorithm relies on weights that set the dynamic priorities for allocating the resource. These weights are built in order to satisfy two major objectives: system throughput maximization and fairness as explained below.

#### A. System Throughput Maximization

The WFO maximizes the system throughput in a MAC/PHY opportunistic approach. Data integrity requirements of the service flows are enforced considering each service flow independently adapting the modulation scheme and the transmit power to the mobile specific channel state. At each scheduling epoch, the scheduler computes the maximum number of bits \( m_{k,n} \) that can be transmitted in a time slot of subcarrier \( n \) if assigned to service flow \( k \), for all \( k \) and all \( n \). This number of bits is limited by two main factors: the data integrity requirement and the supported modulation orders.

Data integrity requirements are specified by a Bit Error Rate (BER) target, denoted by \( \text{BER}_{\text{target},k} \) for service flow \( k \). The required received power \( P_r(q) \) for transmitting \( q \) bits in a RU while keeping below the \( \text{BER}_{\text{target},k} \) of service flow \( k \) is a function of the modulation type, its order and the single-sided power spectral density of noise \( N_0 \). For QAM and a modulation order \( M \) on a flat fading channel [6]:

\[
P_r(q) = \frac{2N_0}{3} \left[ \text{erf}^{-1} \left( \frac{\text{BER}_{\text{target},k}}{2} \right) \right]^2 (M - 1),
\]

where \( M = 2^q \) and \( \text{erf}^{-1} \) is the complementary error function. \( P_r(q) \) may also be determined in practice based on BER history and updated according to information collected on the experienced BER.

The transmit power \( P_{k,n} \) of service flow \( k \) on subcarrier \( n \) is upper bounded to a value \( P_{\text{max}} \) which complies with the transmit Power Spectral Density regulation:

\[
P_{k,n} \leq P_{\text{max}}.
\]

Given the channel attenuation \( a_{k,n} \) experienced by service flow \( k \) on subcarrier \( n \) (including path loss and multipath fading):

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

Perfect knowledge of the channel state is supposed to be available at the receiver. The current channel attenuation on each subcarrier and for each mobile is estimated by the access point based on the SNR of the signal sent by each mobile. Assuming that the channel state is stable on a scale of 50 ms [13], and using a frame duration of 2 ms, the mobiles shall transmit their control information alternatively on each subcarrier so that the access point may refresh the channel state information \( a_{k,n} \) once every 25 frames.

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:

\[
q_{k,n} \leq \left\lfloor \log_2 \left( 1 + \frac{3P_{\text{max}} \times a_{k,n}}{2N_0 \left( \text{erf}^{-1} \left( \frac{\text{BER}_{\text{target},k}}{2} \right) \right)^2} \right) \right\rfloor.
\]

We further assume that the supported QAM modulation orders are limited such as \( q \) 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:

\[
m_{k,n} = \max \{ q \in S, q \leq q_{k,n} \}.
\]

MaxSNR based schemes allocate the resources to the flows which have the greatest \( m_{k,n} \) values. This bandwidth allocation strategy maximizes the bandwidth usage efficiency but suffers of a significant lack of fairness. In order to provide fairness while preserving the system throughput maximization, a new parameter is introduced which modulates this pure opportunistic resource allocation.
The second major objective of the WFO scheme is to provide fairness, i.e. guaranteeing the same level of satisfaction of delay constraints to all service flows. This is achieved by extending the above cross-layer design to higher layers. Delay management is performed considering all service flows jointly and scheduling the packets according to their distance to the delay target. The joint satisfaction of the delay constraints relies on the dynamics of the traffic streams that are multiplexed.

In terms of delay, the meaningful constraint is a limitation of the occurrences of large values. By analogy with the concept of outage used in system coverage planning, the concept of delay outage is defined. 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, \( T_k \) denotes 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 connexion is computed. The result is denoted \( P D O R_k \).

The WFO scheduling objective here is to regulate the experienced PDOR along the lifetime of the service flows such as their values stay below their PDOR targets. Therefore a new “Weighted Fair” (WF) parameter is introduced based on the current estimation of the PDOR of service flow \( k \):

\[
WF_k = f(PDOR_k),
\]

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 service flow \( k \) which has the greatest WFO parameter value \( W F O_{k,n} \) with:

\[
W F O_{k,n} = W F_k \times m_{k,n}.
\]

Based on the PDOR, the WF parameters directly account for the level of satisfaction of the delay constraints for an efficient QoS management. Directly basing the scheduling on the PDOR is more relevant and simpler than considering the service flow throughput, the buffer occupancy or the waiting time of each packet to schedule which would introduce a much greater complexity in the 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 exponent parameter \( \alpha \) allows being more sensitive and reactive to PDOR fluctuations which guarantees fairness at a short time scale. \( \beta \) is a normalization parameter that ensures that \( W F_k \) and \( m_{k,n} \) are in the same order of magnitude. Given that \( PDOR_k \) has an order of magnitude \( 10^{-2} \), \( \beta \) should be set to \( 10^{-2} \). With this choice, \( W F_k \) is always in the same order of magnitude as \( m_{k,n} \) and allows to manage both, fairness and system throughput optimisation.

By extensive simulations, we analyzed the influence of the value of the pair \( (\alpha, \beta) \) on the performances of the WFO scheduling scheme and fine tuned \( f(x) \). Fig. 1 and Fig. 2 illustrate the calibration study\(^2\). Fig. 1 represents the overall PDOR (computed on all transmitted packets) obtained for different values of \( \alpha \) coupled with a \( \beta \) value of \( 10^{2\alpha} \) as defined above. A cubic exponent suits well offering sufficient reactivity to PDOR fluctuations. Hence, in the following \( \alpha \) is assumed to be equal to 3. Fig. 2 shows the WFO performances obtained for each \( \beta \) value when \( \alpha \) is set to 3. It confirms that when \( \beta \) is too small, the weighted parameter has no influence and fairness is lost. On the contrary, if \( \beta \) is too high, \( m_{k,n} \) loses weight in the scheduling and the system throughput maximization decrease. Good values for \( \beta \) range between \( 10^5 \) and \( 10^6 \). In the following, \( \beta \) is taken equal to \( 10^6 \).

Additionally, Fig. 1 and Fig. 2 show the potential of the WFO. Indeed, when \( \alpha \) or \( \beta \) equals zero, the function \( f \) is constant and \( m_{k,n} \) only has influence in the scheduling. With this setting, the WFO behaves as the MaxSNR yielding unfair performances. In contrast, the fine tuning of \( \alpha \) and \( \beta \) brings the wanted fairness.

C. Global WFO Algorithm Description

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

\(^2\)Here, half mobiles are close to the access point and the second half, twice other farther. All mobiles run a same application with same delay and BER requirements as described in Section IV-A.
parameter. 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 for 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\(^3\) 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 \(m_{k,n}\) 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 buffers are empty or all time slots of all subcarriers are allocated and the scheduling ends.

### IV. Performance Evaluation

This section provides a comparison of the proposed Weighted Fair Opportunistic scheduling with the MaxSNR and the PF schemes. Performance evaluation results are obtained using OPNET discrete event simulations. Three essential performance criteria were put under focus: fairness, perceived QoS satisfaction level and offered system capacity.

#### A. Simulation Setup

A frame is assumed to be constituted of 128 subcarriers and 5 time slots. The channel attenuation model on each subcarrier considers free space path loss \(a_k\) and multipath Rayleigh fading \(\alpha_{k,n}\):

\[
a_{k,n} = a_k \times \alpha_{k,n}^2.
\]

\(a_k\) is dependent on the distance to the access point of the mobile owning the service flow \(k\). \(\alpha_{k,n}\) represents the multipath fading experienced by service flow \(k\) if transmitted on subcarrier \(n\). This parameter is Rayleigh distributed with an expectancy equal to unity.

All mobiles run the same videoconference application. This demanding type of application generates a high volume of data with high sporadicity and requires tight delay constraints which substantially complicates the task of the scheduler. Each mobile has only one service flow with a traffic composed of an MPEG-4 video stream [14] multiplexed with an AMR voice stream [15]. The average bit rate of each source is 80 Kbps and its BER target \(k\) value is taken equal to \(10^{-5}\). The PDOR target is set to 5 %.

In order to study the influence of the distance on the scheduling performances, a first half of 6 mobiles are situated close to the access point so that \(P_{\text{max}}/N_0\) is equal to 22.9 dB. The second half of mobile are situated twice over farther so that \(P_{\text{max}}/N_0\) is equal to 16.9 dB. With this setting, the values of \(m_{k,n}\) for the two groups of mobiles are respectively 4 and 2 bits when \(\alpha_{k,n}^2\) equals unity.

#### B. Simulation Results

We first had a look at the QoS level that each mobile perceives. Each mobile runs successive connexions of 5 minutes duration. At the end of each connexion, a final PDOR value is computed. Fig. 4 shows a CDF of these PDOR values, using alternatively the MaxSNR, the PF and the WFO scheme. As shown in Fig. 4(a), close mobiles all satisfy the required QoS while only 54.5 % of the connexions of the far mobiles experience a final PDOR inferior to the PDOR target of 5 %. This first result confirms that MaxSNR is really unfair. Unnecessary priorities are given to close mobiles which easily respects their QoS constraints while more attention should be given to the farther mobiles. The breathing space offered by the easy satisfaction of close mobiles ideally should be exploited for helping the farther mobiles.
PF brings an improvement (cf. Fig. 4(b)) but WFO is the only scheme that ensures absolute fairness. Indeed, in Fig. 4(c)), it appears that all the connexions of close mobiles and more than 98.3% of the connexions of the far mobiles respect their QoS target with a final PDOR smaller than 5%. Making always the most useful allocation, WFO dynamically gives the adequate priority to each connexion. Thereby, it allocates to each mobile the accurate share of bandwidth required for the satisfaction of their QoS constraints, whatever its position.

Additionally, Fig. 5 shows that the WFO keeps more residual bandwidth available for hosting other potential users. This shows that the WFO provides not only the best bandwidth usage but also maximizes the system throughput better than MaxSNR and PF. This improved multiplexing efficiency is obtained by processing all service flows jointly and opportunistically. Keeping more mobiles active but with a relatively lower traffic backlog the WFO preserves the multiuser diversity and takes more advantage of it.

V. Conclusion

Current schedulers presented in the literature offer good system throughput but applying inadequate priorities to the mobiles. The side effect is unfairness in terms of throughput and QoS satisfaction. These schemes serve mobiles in a hierarchical manner preventing to take maximum advantage of the multiuser diversity. Consequently, the good system throughput that they provide is not the optimum either. We think that the solution must be searched in an extended cross-layer approach. This paper proposes a new cross-layer scheduling scheme for OFDM wireless networks, compatible with the existing 802.16 standard. We call this scheme “Weighted Fair Opportunistic (WFO)”. It relies on dynamic weighted priorities that takes into consideration both the physical transmission conditions and higher layer QoS information. Thereby the WFO allocates the adequate priority between mobiles for reaching their QoS satisfaction. All mobiles are served in parallel which allows to exploit the multiuser diversity more efficiently. This results in a well-balanced resource allocation which outperforms the best existing scheduling schemes optimizing the system capacity and providing high QoS satisfaction and fairness.

REFERENCES