Joint Optimization of a Dynamic Queue Scheduling Scheme for the Provision of QoS in WLAN

Nizar Zorba
University of Jordan
Amman, Jordan
n.zorba@ju.edu.jo

Jesús Alonso-Zárate
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
Barcelona, Spain
jesus.alonso@cttc.es

ABSTRACT
We present in this paper the joint optimization of a dynamic queue scheduling scheme for the provision of Quality of Service (QoS) over Wireless Local Area Networks (WLANs). Previous works tune the size of the queue of the scheduler according to the requirements of the users either in terms of expected delay or required throughput. In this paper, we present a joint optimization of the scheduler wherein both delay and throughput constraints are simultaneously considered.

General Terms
Opportunistic Scheduling, WLAN, QoS Guarantees

1. INTRODUCTION
The increasing popularity of in-home Wireless Local Area Networks (WLAN) pushes towards a continuous redesign and update of wireless communications technology. One of the most interesting challenges to face consists in satisfying the stringent Quality of Service (QoS) requirements of real-time traffic in networks that are becoming more and more heterogeneous in terms of variety of applications and types of users. The fact is that current standards for WLAN connectivity, such as the IEEE 802.11 [1], fall short to manage in an efficient manner such demanding scenarios, and thus it is necessary to devise new methods that can provide users with some degree of QoS.

Traditionally, the concept of QoS has been applied separately to the different layers of the protocol stack. However, the latest years have witnessed how the exchange of information between the different layers of the protocol stack can have a dramatic impact in the performance of wireless communications. The main idea is to make decisions at one layer taking into consideration information from other layers. This is known as cross-layer (CL) design [2].

The independent design and optimization of the different layers of the protocol stack is an appropriate approach for wired networks where the channel conditions do not vary with time and the topology of the network is fixed. However, the time-varying characteristics of the wireless channel, which suffers from propagation losses, shadowing, and fading, as well as the capability of the users to freely move and modify the topology of the network in an unpredictable manner, impose a dynamic operation of the communication protocols. In this sense, the use of CL information traveling from the physical (PHY) layer to the upper layers of the protocol stack is of great interest for the design of efficient wireless communications.

In this paper, we focus on the use of channel state information for the scheduling tasks as a way to fulfill the QoS requirements demanded by the users. More specifically, we are interested in the use of the adaptive opportunistic scheduling (OT) scheme presented in [3] where the user with the best channel conditions at each time instant is served in order to enhance the system average throughput and, in general, the QoS satisfaction of the final user. A key characteristic of this approach is that only partial channel information is required at the transmitter side. In fact, the use of OT has been already standardized in UMTS-HSDPA and it is expected to be part of the standard for Very High Throughput (VHT) for WLAN [4].

In addition, we consider the use of a dynamic queue-length for the OT scheduler. With this approach, the number of eligible users for the scheduler can be tuned at will in order to optimize any performance metric of interest. Previous work on dynamic queues has been presented in [5] [6] but without the consideration of the different QoS metrics among the users. Taking into account the recent developments in commercial standards, offering a wide range of applications to the network, the previous work has become somehow obsolete.
More specifically, we extend in this paper the work presented in [7], where a dynamic queue length mechanism for the Data Link Control (DLC) layer is designed to guarantee certain QoS in the downlink of multuser WLAN systems. The proposed solution considers both the PHY and application layers characteristics of the system. The general idea is that data packets arriving from the application layer enter the scheduler queue at the DLC layer. Then, an opportunistic selection is carried out to determine which user is selected for transmission. The size of this queue plays an important role in the performance of the system. An infinite queue may provide no packet losses, but on the other hand, may turn unfeasible the provision of QoS. Therefore, the length of the queue is tuned according to the QoS requirements of each application, taking into account the constraints of the system in terms of throughput and the maximum allowed delay, and tolerating a certain permissible outage in the satisfaction of the QoS requirements. In the work presented in [7], either the throughput or the delay requirements are taken into account to optimize the length of the queue. In this paper, we consider a joint optimization of the problem assuming heterogeneous traffic conditions where both delay and throughput constraints must be taken into account to define the length of the queue.

The reminder of the paper is organized as follows: The system model is presented in Section II. The description of the opportunistic scheduler is given in Section III and its QoS performance metrics are reviewed in Section IV. The proposed dynamic queue length strategy with a joint of optimization considering both delay and throughput requirements is presented in Section V. Some results are presented in Section VI. Finally, Section VII concludes the paper and gives some final remarks.

2. SYSTEM MODEL

We focus on a single-cell downlink channel where \( N \) receivers, all of them equipped with a single receiving antenna, are being served by a transmitter at the Base Station (BS) also provided with a single antenna. A heterogeneous scenario is considered where users run any of the four different classes of applications. Class 1 represents voice users (the most delay-sensitive application) and has the highest priority, while Class 4 is the lowest priority best-effort class (background traffic).

A symmetric wireless channel with response \( h(t) \) is considered between the BS and each of the receivers. As in [3] and [8], a quasi-static block fading model is considered where the channel remains constant along the coherence time and independently changes between consecutive time intervals with independent and identically distributed (i.i.d.) complex Gaussian entries \( \sim \mathcal{CN}(0, 1) \). Therefore, the channel for each user is assumed to be fixed within each fading block (i.e., scenario coherence time) and i.i.d from block to block, so that for the QoS objective, this model captures the instantaneous channel fluctuations in a better approach than the traditional circular ring model [9].

Let \( s_i(t) \) denote the uncorrelated data symbol transmitted from the BS to the \( i^{th} \) user with \( E[|s_i|^2] = 1 \), being \( E[\cdot] \) the expectation operator. The received signal at the \( i^{th} \) user is denoted by \( y_i(t) \) and given by

\[
y_i(t) = h_i(t) s_i(t) + z_i(t),
\]

where \( z_i(t) \) is an additive i.i.d. complex noise component with zero mean and \( E[|z_i|^2] = \sigma^2 \). A total transmission power of \( P_i = 1 \) is assumed and, for ease of notation, time index is dropped whenever possible.

3. OPPORTUNISTIC TRANSMISSION

As mentioned before, an OT selects the user with the best channel conditions (among those in the queue of the scheduler) at each time in order to increase the system aggregate throughput.

During the acquisition step, a predetermined known training sequence is transmitted from the BS to all the users in the system. Upon reception, all the users calculate the received Signal-to-Noise-Ratio (SNR) and feed it back to the BS (this is a partial channel information, as the whole channel information (SNR + phase) is not fed back to the BS). With this information, the BS scheduler is capable of choosing the user with the highest SNR value for each time block and the so-called multiuser gain can be exploited. After this acquisition step, the BS enters to the transmission stage and starts transmitting to the selected user.

This opportunistic strategy has low complexity and is demonstrated that it is optimal in the sense that it obtains the maximum average throughput (TH) computed as [3]

\[
TH = E \left[ \log_2 (1 + \max_{1 \leq i \leq N} SNR_i) \right].
\]

Note that the value of \( \max_{1 \leq i \leq N} SNR_i \) corresponds to the SNR of the user \( i \) selected for transmission.

Although it has been demonstrated in [3] that this OT scheme improves the system average throughput, it cannot provide by itself any precise and guaranteed QoS control for all the users neither in terms of maximum allowed delay nor minimum guaranteed throughput. As it will be explained later, these guarantees can be achieved through an optimization of the DLC scheduler queue length. Before diving into the details of the proposed mechanism, we first review in the next section the different metrics that determine the degree of QoS for a system.

4. SYSTEM PERFORMANCE METRICS

The QoS perceived by a user can be characterized by several metrics or indicators that can be used for the design of different scheduling strategies. Among other possibilities, we will consider in this paper that QoS can be expressed in terms of:

1. Maximum Scheduling Delay, reflecting the maximum delay that a user can tolerate for its transmission.
2. Minimum Guaranteed Rate, reflecting the minimum required throughput per user.

However, the QoS demands cannot be satisfied for 100% of the cases due to the wireless channel impairments. Therefore, some outage in the QoS must be tolerated in the system [10]. Only as an example, current cellular systems such as...
as GSM or UMTS use the same approach and it is expected that future WLAN systems will employ the outage concept when running real-time applications. Another example are Voice over IP (VoIP) applications which can accept erroneous and delayed packets up to $10^{-3}$ of the total number of packets.

For the interest of this paper, we define two concepts for outage [11]:

1. The scheduling delay outage, denoted by $\xi_{\text{access}}$ and defined as the probability that the $i^{th}$ user is provided service after the maximum time that its application can tolerate.

2. The minimum rate outage, denoted by $\xi_{\text{rate}}$ and defined as the probability that once the $i^{th}$ user is selected for transmission, it cannot satisfy its requirements in terms of rate.

We further elaborate on these concepts in the following subsections.

### 4.1 Maximum Scheduling Delay Outage

Contrary to point to point scenarios where the queueing delay is the dominant factor in the system delay [12], in multiuser systems an additional delay factor is introduced due to the fact that the system resources are not always available for the same user. We call this additional delay factor the scheduling delay in multiuser systems.

In round robin based systems (e.g., deterministic Time Division Multiple Access, TDMA) the user channel allocation is known in advance and thus the scheduling delay can be easily calculated. However, in opportunistic multiuser systems, where the user with the best channel conditions is selected for transmission based on its instantaneous SNR, a user does not have any guarantee for being scheduled in a specific time and thus its scheduling delay takes a random value that changes along time.

In the context of this paper, we define the maximum scheduling delay as the maximum allowed period of time from the instant that a user’s packet is available for transmission at the scheduler until the packet is correctly received at its destination. This delay is denoted by $K$ and is measured in access slots. In order to focus on the scheduling problem we will assume hereafter a saturated system. Therefore, we will not consider either the buffer management and the source statistics for arriving packets [13], nor the queues stability target [12]. We leave the integration of all the delays for future work.

As mentioned before, the maximum scheduling delay may not be guaranteed for 100% of the time, and a certain outage probability must be tolerated by the system. If an active user is not scheduled within its maximum allowed delay, e.g., because its channel conditions are not good enough to be selected by the OT scheduler, then this user is said to be in access outage. The probability of outage is denoted by $\xi_{\text{access}}$ and given by

$$\xi_{\text{access}} = 1 - V(K),$$

with $V(K)$ the probability that a maximum of $K$ access slots are required to select a user $i$ from a group of $N$ i.i.d. users. We will assume hereafter that all the users have the same average channel characteristics and, as they show the same distribution for the maximum SNR value, then they have the same probability to be selected for transmission in each access slot. If this is not the case (e.g., heterogeneous user distribution in the cell with some users far from the BS), then a channel normalization (e.g., division by the path loss) can be accomplished for such a scenario.

The value $V(K)$ follows a Geometric Distribution [14] as

$$V(K) = 1 - (1 - P_{\text{access}})^K,$$

with $P_{\text{access}}$ as the average probability for being served in an access slot. Since all users have the same probability to be selected then we obtain $P_{\text{access}} = 1/N$. Therefore, the maximum number of time slots $K$ until the $i^{th}$ user is selected for transmission, with a probability of delay outage $\xi_{\text{access}}$, is given by

$$K = \frac{\log_2(1 - V)}{\log_2(1 - P_{\text{access}})} = \frac{\log_2(\xi_{\text{access}})}{\log_2(1 - 1/N)}.$$  

Note that this expression depends directly on the number of users $N$.

### 4.2 Minimum Rate Outage

The BS scheduler selects the user with the maximum SNR among all the users in the DLC queue at each access slot. However, it may happen that the instantaneous channel conditions (i.e., the instantaneous SNR) with the selected user correspond to a transmission rate that does not satisfy the current requirements of that user. As a consequence, the user is unable to correctly decode the received packets and suffers from a rate outage. We compute this outage probability as follows.

First, and according to [8], the cumulative distribution function (cdf) of the SNR of the i.i.d. complex Gaussian channel associated to each user can be written as

$$F(x) = 1 - e^{-x^2/2\sigma^2},$$

where recall that $\sigma$ is the noise spectral density.

Then, based on the OT philosophy, the serving SNR value is the maximum SNR over the active users in the system. Therefore, the cdf of the serving SNR can be expressed as [11]

$$FF(x) = (F(x))^N = \left[1 - e^{-(x^2/\sigma^2)}\right]^N.$$  

Taking into account the cdf of the serving SNR, the minimum required rate $snr_{th}$ is not achieved for each user with a predefined rate outage probability

$$\xi_{\text{rate}} = \left[1 - e^{-(snr_{th}^2/\sigma^2)}\right]^N.$$  

The values of $snr_{th}$ and $\xi_{\text{rate}}$ can be computed to meet any system objectives given any number of users $N$. By properly selecting the value of $snr_{th}$ it is possible to assume that the minimum SNR value guarantees that the user’s decoding process will be successful. In that case, a unit step function
can be used for the detection procedure, making the Packet Success Rate (PSR) to relate to $\text{SNR}_{th}$ as

$$\text{PSR} = \begin{cases} 1 & \text{if } \text{SNR} \geq \text{SNR}_{th} \\ 0 & \text{if } \text{SNR} < \text{SNR}_{th} \end{cases}$$

With further manipulations, the $\text{SNR}_{th}$ from (8) can be expressed as

$$\text{SNR}_{th} = \frac{1}{\lambda} \log_{2} \left( \frac{1 - 1 - \sqrt{\xi_{rate}}} {1 - \sqrt{\xi_{rate}}} \right)$$

(10)

where the effect of all the involved parameters is shown, with $\lambda = \log_{2}(e) = 1.4427$. This equation shows the rate limits of the system, indicating that high $\text{SNR}_{th}$ requirements induce high outage $\xi_{rate}$ in the system. Negative values for $\text{SNR}_{th}$ indicate infeasibility of the requested rate.

5. DYNAMIC QUEUE LENGTH

The analysis presented in the previous section shows that the number of users in the system has a direct impact on the QoS perceived by each user. Therefore, by controlling the number of eligible users for the scheduler, it is possible to control the QoS perceived by the users. We describe in the next subsection the concept of using a dynamic queue length at the DLC scheduler. Then, we jointly optimize this length taking into account both delay and throughput constraints for heterogeneous traffic conditions.

5.1 Description

The use of a dynamic queue length for the DLC scheduler was first proposed in [7]. When the maximum allowed delay (or minimum allowed rate) in the delivery of the most delay sensitive application is smoothly satisfied, then the length of the queue, denoted by $N$, can be increased so that more users can be accommodated in the DLC layer queue. As a consequence, the OT scheduler can select the user with the best channel conditions in a bigger pool of choices, increasing in this way the performance of the system in terms of the average throughput in (2). On the other hand, when the maximum allowed delay requirements are hardly satisfied, then the length of the DLC queue is shortened. Therefore, only packets form users within the higher priority classes can be available in the DLC layer queue, so that the OT scheduler can only select among these users. Likewise, the same procedure can be applied when the minimum guaranteed throughput per user is the considered QoS indicator.

The proposed dynamic adjustment in the size of the queue poses a tradeoff between the real-time users’ QoS demands and the system average throughput in the network. The best operating point depends on the network operator requirements, but in general, the operator will attempt to accept as much users as possible to increase its income.

5.2 Joint Optimization

Taking into consideration the acceptance of the system to a certain outage in the QoS satisfaction, a proposed optimization procedure for the system performance can be stated as

$$\text{max } N$$

s.t. $1. \text{Prob}(\text{SNR}_i < \text{SNR}_{th}) \leq \xi_{rate} \forall \ i \in N$

s.t. $2. \text{Prob}(D_{max} < K_i) \leq \xi_{access} \forall \ i \in N$

(11)

where $D_{max}$ is the maximum allowed delay. The previous formulation stands a linear optimization problem [15], where the objective is to get the maximum value of $N$ to satisfy the two requirements.

Thanks to the obtained closed form expressions for the scheduling delay and minimum rate in the previous sections, the first restriction is directly obtained from (10), and by mathematical manipulations, the value of $N$ can be obtained as

$$N \geq \frac{\log_{2}(\xi_{rate})}{\log_{2}(1 - \frac{1}{\text{SNR}_{th} \cdot \sigma^2})}$$

(12)

Similarly for the maximum delay demands and employing the obtained equations for the scheduling delay, the second restriction reformulates as

$$N \leq \frac{1}{1 - 2^{\left( \frac{\log_{2}(\xi_{access})}{D_{max}} \right)}}$$

(13)

The result for the optimization problem in (11) is very interesting as the obtained restrictions for the value of $N$ expressed in (12) and (13) show an opposite trend. For the minimum rate demands, the $N$ is desired to be the largest possible, i.e., $N \geq c_1$, with $c_1$ as the boundary value in (12). On the other hand, for the maximum delay demands, the $N$ value is desired to be the smallest possible, i.e. $N \leq c_2$, with $c_2$ as the boundary value in (13). Therefore, the value of $N$ must stand in between the two boundary values, $c_1$ and $c_2$. According to [15], we will refer to this region as the feasibility region in the optimization formulation.

Since the optimization problem is restricted by the feasibility region, it is necessary to to check whether the restrictions can be jointly satisfied or not before solving the problem. Once both restrictions are guaranteed to be within the feasibility region, the solution of the optimization problem yields a maximization of the number of users that the system can support, thus increasing the system throughput and ensuring that the QoS demands are satisfied.

In the next section, we present some numerical results supported by computer simulations.

6. PERFORMANCE EVALUATION

To evaluate the performance of the proposed dynamic DLC queue mechanism, a heterogeneous scenario has been set up where users with four types of applications coexist in a network. The results are presented through both theoretical analysis and with Monte Carlo simulations carried out with Mathworks Matlab.

We consider a scenario formed by a total of 20 users. These users are separated in 4 groups of 5 users each. Each group corresponds to each of the 4 different traffic classes. In the simulations we use different lengths for the queue ($N$). For example, if $N = 12$, it means that we take the 5 users from class 1, the 5 users from class 2, and 2 users from class 3.

The length of the packets for the classes 1, 2, 3 and 4 are 100, 512, 1024 and 2312 bytes, respectively. Class 1 has the highest priority, while class 4 is the lowest priority class. Class 1 can be VoIP and/or on-line gaming applications while Class 4 may represent an FTP download application.
worth noting the exact match between the theory and the requirements increases with the length of the queue. It is assumed for the class 1 users. It can be seen from Fig. 1 that the class 1 exist in the DLC queue), the maximum allowed delay is guaranteed for the 98% of the cases (with an outage of 2%). Notice that increasing the queue length to 20, so that all users are eligible for scheduling selection in the system. This means that class 1 users (lowest priority) have lower chances to be served by the BS scheduler. This has a direct impact on the time delivery of their packets. Fig. 2 shows the performance of the average throughput (from (2)) for a variable DLC queue length scheme where, as expected, the average throughput increases with the value of N due to the opportunistic user selection of the scheduler. Note that when N = 16, the result of the optimization formulation goes out of the feasibility region as the maximum delay demands can not be satisfied, as expressed in (13). On the other hand, when N <= 6, the problem also leads to the other infeasibility region where the minimum rate demands cannot be satisfied, as predicted by (12).

7. CONCLUSIONS

We have presented in this paper the joint optimization of a dynamic DLC queue length for an opportunistic scheduler. We have focused on the downlink of a WLAN scenario, where a number of users have to be served by the base station. By dynamically adjusting the size of the queue of the OT scheduler it is possible to provide the users with certain QoS, given that some outage is accepted in the system.

The main contribution of this paper has been the joint opti-
mization of the queue size taking into account both through-
put and delay constraints in a heterogeneous scenario. Both 
theoretical analysis and computer simulation results have 
been presented in the paper. The main conclusion is that the 
requirements of throughput and delay have opposite trends 
(with respect to the number of users that can be served in 
the system), and thus there are some infeasibility regions 
where both requirements cannot be met.

Future work aims at including user behavior patterns in the 
optimization of the queue length as the kind of traffic de-
manded by users is different depending on the time of the 
day.

Acknowledgements
This work has been partially funded by the research Projects 
R2D2 (CP6-013), PERSEO (TEC2006-10459/TCM) and NE-
WCOM++ (ICT-216715).

8. REFERENCES
[1] Wireless LAN Medium Access Control (MAC) and 
Physical Layer (PHY) Specifications, IEEE Standard 
802.11, Revision 2007.
"Cross-Layer Design for Wireless Networks," IEEE 
[3] P. Viswanath, D. N. Tse, and R. Laroia, 
"Opportunistic beamforming using dumb antennas," 
[5] B. Ata, "Dynamic Control of a Multiclass Queue with 
Thin Arrival Streams," Operations Research, vol.54, 
September 2006.
Length Thresholds for Multiple Loss Priorities," 
IEEE/ACM Transactions on Networking, vol.10, June 
2002.
Verikoukis "Cross Layer QoS Guarantees in Multiuser 
WLAN Systems," Special Issue on “Cross-Layer Design 
for Future Generation Wireless Networks" for Wireless 
Personal Communications, accepted for publication, 
2009.
[8] L. Yang and M.S. Alouini, "Performance Analysis of 
Multiuser Selection Diversity," IEEE-ICC, Paris-France, 
for IEEE 802.11 WLAN Via Received Signal Strength 
[10] B. K. Chalise and A. Czywilk, "Robust Downlink 
Beamforming based upon Outage Probability Criterion," 
IEEE-VTC Fall, Los Angeles-USA, September 2004.
Opportunistic Schemes in Heterogeneous WiMax 
Systems under QoS Constraints," IEEE Globecom, 
Power Allocation and Routing for Time Varying 
[13] T. Issariyakul and E. Hossain, 
"Channel-Quality-Based Opportunistic Scheduling with 
ARQ in Multi-Rate Wireless Networks: Modeling and 
Analysis," IEEE Transactions on Wireless 
Communications, vol.5, April 2006.
[16] H. Fu and D.I. Kim, "Analysis of Throughput and 
Fairness with Downlink Scheduling in WCDMA 
Networks," IEEE Transactions on Wireless 
Communications, vol.5, August 2006.

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7925
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7925