Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2010, Article ID 564527, 10 pages
doi:10.1155/2010/564527

Research Article

A QoS-Based Dynamic Queue Length Scheduling Algorithm in Multiantenna Heterogeneous Systems

Nizar Zorba[1] and Christos Verikoukis[2]

[1] Department of Electrical Engineering, University of Jordan, Amman 11942, Jordan
[2] Telecommunications Technological Center of Catalonia (CTTC), Barcelona 08860, Spain

Correspondence should be addressed to Nizar Zorba, n.zorba@ju.edu.jo

Received 30 July 2009; Accepted 25 November 2009

Academic Editor: Yevgeni Koucheryavy

The use of real-time delay-sensitive applications in wireless systems has significantly grown during the last years. Therefore the designers of wireless systems have faced a challenging issue to guarantee the required Quality of Service (QoS). On the other hand, the recent advances and the extensive use of multiple antennas have already been included in several commercial standards, where the multibeam opportunistic transmission beamforming strategies have been proposed to improve the performance of the wireless systems. A cross-layer-based dynamically tuned queue length scheduler is presented in this paper, for the Downlink of multiuser and multiantenna WLAN systems with heterogeneous traffic requirements. To align with modern wireless systems transmission strategies, an opportunistic scheduling algorithm is employed, while a priority to the different traffic classes is applied. A tradeoff between the maximization of the throughput of the system and the guarantee of the maximum allowed delay is obtained. Therefore, the length of the queue is dynamically adjusted to select the appropriate conditions based on the operator requirements.

1. Introduction

The use of real-time delay-sensitive applications such as voice, video streaming, or online-gaming for indoor WLAN applications has been remarkably growing during the last years. Nevertheless, WLAN was designed as a data transmission technology without the considerations of voice and real-time applications, so that commercial IEEE 802.11 WLAN systems do not guarantee strict Quality of Service (QoS) requirements in terms of maximum allowed delay and/or delay jitter. Moreover, the fact that the wireless environments are characterized by a harsh scenario for communications increases the difficulties to guarantee the desired QoS in WLAN-based systems. The specific characteristics of the wireless channel with multiple undesired effects, such as deep fades and multipath, distort the original information. As a consequence, guarantying QoS by using the scarce available resources in an in-home wireless medium is a challenging aspect for future WLAN systems.

Different QoS metrics are defined and used at different layers of the OSI model [1]. The acceptable signal strength level and/or Bit Error Rate at the receiver may represent the QoS at the physical layer, but at the higher layers, the QoS concepts are quite different as they are usually expressed in terms of minimum-guaranteed throughput, and delay either maximum allowed delay or delay jitter. Different procedures are followed at each layer to fulfil QoS requirements. At DLC layer, QoS is guaranteed by appropriate radio resource management algorithms while at the physical layer other mechanisms such as power control, adaptive coding and modulation, or symbol rate are applied to guarantee the quality of the communications.

It has been proved that the vertical coupling among layers, known as Cross-Layer [2], can significantly improve the efficiency of the wireless systems. Both theory and experimental evaluations have demonstrated that cross-layer between the physical and the higher layers seem to be unavoidable in wireless environments in order to exploit the wireless channel instantaneous conditions. Such interchange of information not only helps in increasing the system sum rate performance, but also may be used to guarantee the QoS requirements in systems with heterogeneous type of traffic and applications which need different QoS requirements. In general Cross-Layer further advantages can include improvements in terms of link throughput, reduction of the network latency, energy savings in the mobile nodes, or minimization of transmitted power [2, 3].

One of the resources of the system that can be employed to improve the system performance in terms of both rate and QoS is the spatial diversity. The Multiple-Input-Multiple-Output (MIMO) technology in multiuser scenarios shows very interesting results as several users can be simultaneously serviced within the same frequency, time, and codes. Its employment has already been standardized in IEEE 802.11n and IEEE 802.16e, while it is expected to be part of the forthcoming 4th Generation Long-Term Evolution (LTE) Standard. Among the proposed techniques within MIMO, the Multibeam Opportunistic Beamforming (MOB) strategy that has been suggested in [4] to boost the wireless link capabilities shows the highest performance, lower complexity design, and only partial channel information is required at the transmitter side. MOB can be operated and adopted to fulfil the QoS requirements demanded by the users for their correct operation [5].

An interesting remark concerning the QoS compliance in commercial wireless systems refers to the outage concept [6], where due to the wireless channel characteristics, the 100% satisfaction of the strict QoS demands is impossible, for what is known as outage in the QoS requirements [6]. The notion of outage is widely employed by engineers in the cellular systems where commercial systems (e.g., GSM and WCDMA) allow up to 2–5% outage, depending on the scenario and the application. Therefore, the extension of this concept to WLAN-based systems with delay-sensitive applications seems to be the most tractable approach to asset their efficiency.

Taking into consideration all the previous features, the main contribution of this paper is to propose a Dynamic Queue Length in the Data Link Control Layer, in order to guarantee certain QoS, in the Downlink of multiuser and multiantenna WLAN systems with heterogeneous traffic. As a Cross-Layer philosophy is deployed, then the proposed solution considers both the physical and application layers characteristics of the system. To be more precise, the length of the queue depends on the QoS system requirements, in terms of the system throughput and the maximum allowed delay (and jitter) of the most delay-sensitive applications, where some outage is considered in the QoS requirements of these applications.

As a summary, the contributions of this paper are in the field of Dynamic queues management under QoS demands as follows.

(i) The paper tackles a multiantenna scenario and chooses the MOB scheme for its transmission strategy.

(ii) Through the use of the outage concept, this paper is able to formalize the service distribution characteristics of the MOB scheme, allowing to obtain the minimum rate and maximum scheduling delay in closed form expressions.

(iii) An approach to obtain the opportunistic multiuser gain, while providing the system QoS constraints in terms of minimum-guaranteed rate and maximum allowed delay, is presented.

(iv) This paper presents a Cross-Layer Dynamic queues management strategy, and studies its performance. A Cross-Layer design is required in order to consider the instantaneous channel conditions and QoS demands.

The rest of the paper is organized as follows. Section 2 makes a review of other similar solutions in the literature and underline their innovation. Section 3 presents the system model while the Multibeam Opportunistic Beamforming (MOB) is introduced in Section 4. Section 5 gives an overview of the system QoS performance followed by Section 6 with the Dynamic Queue Length model. Performance evaluation results are depicted and analysed in Section 7, to close the paper with the future research directions and conclusions in Sections 8 and 9, respectively.

2. Related Work

With respect to the aforementioned concepts in a Downlink system with heterogeneous traffic, several proposals in the literature tackle the dynamic queue consideration, but with different objectives and requirements. The authors in [7] propose a Media Access Control (MAC) protocol for a finite-user slotted channel with multipacket reception (MPR) capability. By adaptively changing the size of the contention class (defined as a subset of users who can access the channel at the same time) according to the traffic load and the channel MPR capability, the proposed dynamic queue protocol provides superior channel efficiency at high traffic load and minimum delay at low traffic load. However, this protocol is dynamic in terms of traffic load queue and does not deal with the problem of having different users with different QoS demands.

An admission control problem for a multiclass single-server queue is considered in [8]. The system serves multiple demand streams, each having a rigid due-date lead time. To meet the due-date constraints, a system manager may reject orders when a backlog of work is judged to be excessive, thereby incurring lost revenues. Nevertheless, in this paper, service classes are turned-away based on predefined load (packets in the queue) thresholds and only the average mean delay is guaranteed, while the maximum delay is not.

A dynamically queuing feature for service enhancement is proposed in [9], according to the increment of service subscribers and their mobility. In addition, it presents a dynamic queue manager that handles the queue size to increase call completion rates for service enhancements in wireless intelligent network environments. In spite of this, other QoS demands are not possible and the problem of having different users with different QoS demands is not dealt with.

Various QoS requirements of bursty traffic and a dynamic priority queue with two types of traffic are proposed and analyzed in [10]. The system has two separate buffers to accommodate two types of customers, the capacities of the buffers being assumed to be finite for practical applications. But the service order is only determined by the queue length of the first buffer, so that only average QoS demands can be satisfied.

The scheduler gives some buffers and bandwidth to every priority class at every port in [11]. The scheme adapts to changes in traffic conditions, so that when the load changes the system goes through a transient. Therefore, each queue individually carries out its blocking process, which does not provide any tight control on the QoS demands.

3. System Model

We focus on the single cell Downlink channel where N receivers, each one of them equipped with a single receiving antenna, are being served by a transmitter at the Base Station (BS) provided with nt transmitting antennas, and supposing that N is greater than nt. The considered scenario is actually a multiuser Multiple Input Single Output (MISO) but the results can be easily applied to multiuser MIMO with any receiver processing. This scenario is considered for easiness, as the receiver processing is out of this paper scope, and all main conclusions of the paper are independent of the processing carried out at the receiver. The scenario is identified to be a heterogeneous scenario 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.

It is worth mentioning that the demand of real-time services, such as Voice over IP (VoIP), for strict QoS delay demands, leads to the reconsideration of the ring scattering model [12], which is widely used in the evaluation of WLAN systems with nonreal time (e.g., data traffic) applications. This is because the QoS requirements have to be satisfied in a tighter time scale, which requires for detailed models to account for the instantaneous channel random fluctuations.

A wireless multiantenna channel $\mathbf{h}_{[1\times n_t]}$ is considered between each of the users and the BS, where a quasistatic block fading model is assumed, which keeps constant through 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 circular rings model. Let $\mathbf{x}(t)$ be the $n_t \times 1$ transmitted vector (as we are in a Downlink scenario), while denote $y_i(t)$ as the jth user received signal, given by

$$y_i(t) = \mathbf{h}_i(t)\mathbf{x}(t) + z_i(t), \tag{1}$$

where $z_i(t)$ is an additive Gaussian complex noise component with zero mean and $E\{|z_i|^2\} = \sigma^2$. The transmitted signal $\mathbf{x}(t)$ encloses the independent data symbols $s_i(t)$ to all the selected users with $E\{|s_i|^2\} = 1$. A total transmitted power constraint $P_t = 1$ is considered, and for ease of notation, time index is dropped whenever possible.

4. Multibeam Opportunistic Beamforming (MOB)

One of the main transmission techniques in multiuser multiantenna scenarios is the MOB scheme [4], where random beams are generated at the BS to simultaneously serve more than one user. The beam generation follows an orthogonal manner to decrease the interference among the served users, where n_t beams are generated. Within the acquisition step, a known training sequence is transmitted for all the users in the system. Therefore, each user sequentially calculates the Signal-to-Noise Interference Ratio (SNIR) related to each beam, and feeds back to the BS only the best SNIR value together with an integer number indicating the index of the selected beam. The BS scheduler chooses the user with the highest SNIR value for each one of the beams. So, it gets the multiuser gain from the scenario to increase the system throughput. After that, the BS enters the transmission stage and simultaneously transmits to each one of the n_t selected users, where no user can obtain more than one beam at a time.

Since the users with the best channel conditions are selected for transmission, the scheduler is called *Opportunistic Scheduler*. Therefore, the low complexity MOB strategy achieves high throughput by spatial multiplexing the n_t users with the best channel conditions, making the transmitted signal to enclose the data symbols for the n_t selected users as

$$\mathbf{x} = \sqrt{\frac{1}{n_t}} \sum_{m=1}^{n_t} \mathbf{b}_m s_m \tag{2}$$

with $\mathbf{b}_m$ as the unit-power beam assigned to the mth user, where the square root term is due to a total power constraint of $P_t = 1$.

This scheme is characterized by its SNIR term due to the interference that each beam generates to its nonintended users, and even though the beams are orthogonally generated, some of the orthogonality is lost in the propagation channel [4], stating the SNIR formulation for the ith user through the mth beam as

$$\mathrm{SNIR}_{i,m} = \frac{(1/n_t)|\mathbf{h}_i \mathbf{b}_m|^2}{\sigma^2 + \sum_{u\neq m}^{n_t}(1/n_t)|\mathbf{h}_i \mathbf{b}_u|^2} \tag{3}$$

with $\mathbf{b}_u$ as the unit-power beam assigned to the uth user, and where a uniform power allocation is considered. As the user with the highest SNIR value is selected for each transmitting beam, then the average system throughput of MOB can be written [4] as

$$\mathrm{TH} = E\left\{ \sum_{m=1}^{n_t} \log_2\left(1 + \max_{1\leq i\leq N} \mathrm{SNIR}_{i,m}\right)\right\}, \tag{4}$$

where $E\{\cdot\}$ is the expectation operator to denote the average value. Notice that the value of $\max_{1\leq i\leq N}\mathrm{SNIR}_{i,m}$ reflects the serving SNIR (i.e., the SNIR that the selected user i receives when serviced through the m^{th} beam).

Although it has been shown that MOB improves the average throughput of the system [4], the main target of this work is in providing a precise and guaranteed QoS control for all the users, mainly in terms of the maximum allowed delay, and minimum-guaranteed throughput. As it will be later explained, this is achieved through the optimization of the DLC queue length, where the simulations will show an interesting tradeoff between the QoS satisfaction and the system average throughput. It has to be noted that the minimum allowed rate, the maximum allowed delay and the minimum-guaranteed throughput stand as QoS realistic constraints for both real and nonreal time applications, providing the commercial operator with a wider view than the fairness concept, as the QoS is stated in terms of per user exact requirements.

5. System QoS Performance

For the consideration of any transmission scheme in commercial standards that run real-time applications, the QoS of the users is a very important aspect that can be characterized by several metrics or indicators based on the design objectives. So, QoS can be expressed in terms of rate, reflecting the minimum required rate per user, or in terms of delay, showing the maximum delay that a user can tolerate for its packets. This paper considers both of the aforementioned QoS concepts, where the proposed transmission scheme guarantees a minimum rate R per user, which is presented by a minimum SNIR restriction (snir_{th}), through the classical relation ($R = \log_2(1 + \text{snir}_{\text{th}})$), and delivered to it within a maximum tolerable time delay K.

As this work deals with real-time applications in WLAN systems, then the QoS demands cannot be satisfied for the 100% of cases due to the channel characteristics. Therefore, some outage ξ_{out} in the QoS is accepted [6], where the outage is currently employed in cellular systems design as GSM and UMTS, and expected in WLAN systems when running real-time applications. As an example, VoIP can accept erroneous packets up to 10^{-3} of the total number of packets.

The paper defines two concepts for outage [1]: the scheduling delay outage and the rate outage. The first one is related to the opportunistic access policy and the time instant when the i^{th} user is provided service. Section 5.1 characterizes the user opportunistic access and obtains the expression for its access delay probability. The second outage concept accounts for the received data rate once the ith user is selected for transmission, and whether its rate requirement is satisfied or not. Section 5.2 derives the corresponding SNIR distribution for the selected user and obtains the minimum-guaranteed rate under an outage ξ_{rate}.

5.1. Access Delay Outage. In TDMA systems (e.g., GSM) each user knows, in advance, its exact access slot; but in an opportunistic scheduler, as a continuous monitorization of the users' channel quality is performed to select the best ones in each slot, then the access to the wireless medium is not guaranteed. Therefore, the study of the access to the channel in the MOB scheme offers several challenges that must be solved for the MOB consideration in practical systems.

This section calculates the maximum access delay from the time that a user's packet is available for transmission at the scheduler until the user is serviced through any of the n_t beams of the BS. If an active user is in the system, but it is not scheduled within its maximum allowed delay (e.g., because its channel conditions are not good enough to be selected by the MOB scheduler), then that user is declared as being in access delay, with an outage probability ξ_{access} given by

$$\xi_{\text{access}} = 1 - V(K) \tag{5}$$

with $V(K)$ as the probability that a maximum of K time slots are required to select a user i from a group of N i.i.d. users (along the paper, all the users are assumed to have the same average channel characteristics, and showing the same distribution for the maximum SNIR value, so that each user has the same probability to be selected. If this is not the case (e.g., heterogeneous users 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.), where this probability follows a Geometric Distribution [13] as

$$V(K) = 1 - \left(1 - \overline{P}_{\text{access}}\right)^K. \tag{6}$$

In the MOB scheme, each one of the N independent users attempts to be serviced by one of the n_t generated beams with $\overline{P}_{\text{access}} = n_t/N$ therefore from previous equation, the maximum number of time slots K until the i^{th} user is selected for transmission, with a probability of delay outage ξ_{access}, is given by

$$K = \frac{\log_2(1 - V)}{\log_2\left(1 - \overline{P}_{\text{access}}\right)} = \frac{\log_2(\xi_{\text{access}})}{\log_2(1 - n_t/N)}, \tag{7}$$

showing the effects of the number of active users N and the number of serving beams n_t.

5.2. Minimum Rate Outage. If the BS scheduler selects a user for Downlink transmission, it means that he/she has the maximum SNIR among the users for a specific beam. But the instantaneous channel conditions (i.e., the instantaneous SNIR) may correspond to a transmission rate that does not satisfy its current application rate requirements (e.g., for a predefined Packet Error Rate, the channel can only provide 6 Mbps while the application asks for 24 Mbps). As a consequence, the user is unable to correctly decode the received packets during the current time unit and suffers a rate outage.

Based on the MOB philosophy to deliver service to the users, the serving SNIR value is the maximum SNIR over the active users in the system, corresponding to each generated beam. Using the SNIR equation in (3), note that the numerator follows a Chi-square $\chi^2(2)$ distribution while the interference terms in the denominator are modeled as

$\chi^2(2(n_t - 1))$, which allows to obtain the SNIR probability distribution function (pdf) as [1, 4]

$$f(x) = \frac{e^{-(x \cdot n_t \sigma^2)}}{(1+x)^{n_t}} \left(n_t \sigma^2 (1+x) + n_t - 1\right), \qquad (8)$$

and the cumulative distribution function (cdf) is then formulated as

$$F(x) = 1 - \frac{e^{-(x \cdot n_t \sigma^2)}}{(1+x)^{n_t - 1}}, \qquad (9)$$

and since the serving SNIR is the maximum over all the users' SNIR values (i.e., the opportunistic philosophy), then its cdf is stated as

$$FF(x) = (F(x))^N = \left[1 - \frac{e^{-(x \cdot n_t \sigma^2)}}{(1+x)^{n_t - 1}}\right]^N. \qquad (10)$$

Therefore the minimum required SNIR ($\mathrm{snir}_{\mathrm{th}}$) for each user is achieved with a probability U as

$$U(\mathrm{snir}_{\mathrm{th}}) = 1 - \left[1 - \frac{e^{-(\mathrm{snir}_{\mathrm{th}} \cdot n_t \sigma^2)}}{(1+\mathrm{snir}_{\mathrm{th}})^{n_t - 1}}\right]^N \qquad (11)$$

which relates to the predefined rate outage ξ_{rate} as

$$\xi_{\mathrm{rate}} = \left[1 - \frac{e^{-(\mathrm{snir}_{\mathrm{th}} \cdot n_t \sigma^2)}}{(1+\mathrm{snir}_{\mathrm{th}})^{n_t - 1}}\right]^N, \qquad (12)$$

where the values of $\mathrm{snir}_{\mathrm{th}}$ and ξ_{rate} can be computed on the basis on any system objectives, under the number of users N. With further manipulations, expression (12) can be reformulated as

$$\log_2(1+\mathrm{snir}_{\mathrm{th}}) = \frac{\log_2\left(1/\left(1 - \sqrt[N]{\xi_{\mathrm{rate}}}\right)\right) - \lambda \mathrm{snir}_{\mathrm{th}} \cdot n_t \sigma^2}{n_t - 1}, \qquad (13)$$

obtaining the minimum-guaranteed rate, and where $\lambda = \log_2(e) = 1.4427$ is adopted. Equation (13) shows the rate limits of the system, indicating that high $\mathrm{snir}_{\mathrm{th}}$ requirements induce high outage ξ_{rate} in the system. Negative values in the right hand term indicate infeasibility of the requested rate. We assume in this paper that the minimum SNIR guarantees successful decoding of packets. Therefore, the following unit step function defines the Packet Success Rate (PSR) related to the $\mathrm{snir}_{\mathrm{th}}$ as

$$\mathrm{PSR} = \begin{cases} 1 & \text{if serving SNIR} \geq \mathrm{snir}_{\mathrm{th}}, \\ 0 & \text{if serving SNIR} < \mathrm{snir}_{\mathrm{th}}, \end{cases} \qquad (14)$$

where a direct relation to ξ_{rate} is obtained from (12).

5.3. Outage of the System. As previously explained, the MOB scheme comes controlled by two different outage measures, but the total system performance has to be defined through a single parameter. Notice that the two discussed kinds of outage are totally independent, as the user's access to the channel happens when its SNIR is the maximum over all the other users, with respect to a given beam, but being the user with largest SNIR does not guarantee that this SNIR is larger than an application predefined threshold $\mathrm{snir}_{\mathrm{th}}$. Therefore, the total outage ξ_{out} is defined as

$$\xi_{\mathrm{out}} = 1 - (1 - \xi_{\mathrm{access}}) \cdot (1 - \xi_{\mathrm{rate}}), \qquad (15)$$

standing as the global measure of system outage.

5.4. Maximum Scheduling Delay. In point-to-point scenarios, the queueing delay is the dominant factor in the system delay [14] while in multiuser systems an additional delay factor is introduced, because the system resources are not all the time available to the same user. We name this additional delay factor as the scheduling delay in multiuser systems. In the round robin systems (e.g., TDMA) the user access to the channel is known in advance, so that its scheduling delay can be easily calculated. However, in opportunistic multiuser systems where the users with the best channel conditions are selected for transmission based on their instantaneous SNIR, a user does not have any guarantee for being scheduled in a specific time, which increases its scheduling delay.

In the context of this paper, we define the maximum scheduling delay as the time period from the instant that a user's packet is available for transmission at the scheduler until the packet is correctly received at its destination. The difference with the access delay definition is the requirement of a rate threshold in order to guarantee the decoding without errors, as in (14). Notice that this definition includes both the delay resulting from the scheduling process (i.e., the opportunistic selection) and the delay caused by the requirements to get a rate above to a minimum required threshold to be correctly received. Therefore, the maximum number of time slots to select a user with a total outage ξ_{out} is equal to the K access slots (7), defining the maximum allowed scheduling delay.

In order to avoid misleading conclusions for the reader, a brief numerical example is presented. In a scenario with $N = 30$ total users, $n_t = 3$, a system bandwidth of $B_w = 1$ MHz, $K = 25$, required maximum scheduling delay, $\sigma^2 = 1$ and $R = 580$ Kbps minimum demanded rate for each user, it results that $\xi_{\mathrm{access}} = 7.1\%$ and $\xi_{\mathrm{rate}} = 4.3\%$ are obtained. So that the access delay is 25 slots with an access outage of 7.1%. But even though a user is selected, it may get a rate below its requirement with an outage probability of 4.3%, so that the ξ_{rate} must be introduced. Therefore, a wireless operator can guarantee to each user, the correct reception of its packet within a maximum scheduling delay of 25 slots and with a total outage of $\xi_{\mathrm{out}} = 11.0\%$.

As we consider the scheduling delay, both the buffer management and source statistics for arriving packets are not addressed [15]; the queues stability target [14] is neither considered. Therefore, we assume a saturated system and only consider the delay resulting from the scheduling process. The total delay (scheduling + queueing) will be tackled as a future work.

5.5. Minimum-Guaranteed Throughput per User and per Slot. Obtaining the system throughput formulation is difficult as several processes are included in the communication procedure. The receiver decoding through the unit step function in (14) simplifies the throughput formulation, as the effects of several steps in the communication process (e.g., coding) are avoided.

In opportunistic multiuser scenarios, the user in not always served by the system, so that its throughput is zero for several time units. Therefore, a normalized minimum-guaranteed throughput per user over the time is required. Notice that such definition of throughput per user and per slot accounts for the user's waiting time and hence, for its corresponding scheduling delay expression. Considering that the bandwidth of the system is B_w, then the minimum-guaranteed throughput per user and per slot is denoted as T, in bits, and given as

$$T=\frac{B_w\log_2(1-n_t/N)\left[\log_2\left(1/\left(1-\sqrt[N]{\xi_{\text{rate}}}\right)\right)-\lambda\text{snir}_{\text{th}}\cdot n_t\sigma^2\right]}{(n_t-1)\log_2(\xi_{\text{access}})}, \tag{16}$$

where the expression in (13) is used to provide a closed form solution for the minimum-guaranteed throughput per user, with all the operating variables. Notice that by increasing the number of users N, the minimum-guaranteed rate R goes up and as a consequence higher throughput is obtained. On the other hand, larger N induces larger scheduling delay, increasing in this way the value of K, that drives lower throughput values. This shows a tradeoff on the number of available users in the systems, motivating a control over the N value to achieve the system QoS requirements, as will be shown in the next section.

Note that the minimum-guaranteed throughput is the worst case awarded throughput to the users, but it actually defines the throughput value that an operator can guarantee to its customers, obviously, with a given outage ξ_{out}; where the guaranteed throughput per user is different from the concept of average throughput in (4), previously presented. A very common example in commercial systems for average throughput and the minimum-guaranteed throughput is seen in the ADSL service, where, for example, an operator can provide its costumers 20 Mbps (which is the value that appears in its advertisements), while the minimum-guaranteed value for the user is 2 Mbps (National regulatory telecommunication agencies often ask for a guaranteed value of at least 10% of the average value).

6. Data Link Control with Dynamic Queue Length

Two important aspects to achieve QoS for the serviced users are extracted from the analytical study in the previous section: the impact of the number of available users and their exact QoS demands. To control the different user requirements and their sensitivity to delay and rate, a control on the DLC queue length L is proposed in this paper. The aim of this section is to provide a description of this proposal, performed through a cross-layer scheduling algorithm at the DLC layer of WLAN systems. The main idea of the proposed scheme is depicted in Figure 1. It can be seen that each IP packet is stored at the corresponding priority queue in the IP layer, before moving down to the DLC layer queue. Users from higher priority IP queues are placed at the beginning of the DLC queue following by users with lower priorities traffic.

At the Physical layer, the WLAN systems use different modulation levels, so that variable transmission rates depending on the channel conditions (measured through the received SNIR) are obtained. The MOB scheme is applied to select the users with the best channel conditions per beam in order to maximize the system average throughput.

Regarding the dynamic queue length mechanism, 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 can be increased so that more users can be placed in the DLC layer queue. As a consequence, the MOB scheduler can select the user per beam 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 (4). On the other hand, when the maximum allowed delay requirements are hardly satisfied, then the length of the DLC queue is decreased. Therefore, only packets form users within the higher priority classes can be available in the DLC layer queue, so that the MOB scheduler can only select, for each one of the beams, among these users. Likewise, the same procedure can be applied when the minimum-guaranteed throughput per user is the considered QoS indicator.

Note that the proposed dynamic adjustment in the size of the queue shows the tradeoff between the real-time users' QoS demands and the system average throughput in the network, where the best operating point depends on the network operator requirements. It has to be noted that very delay sensitive applications are in general characterized by short packets lengths, such as VoIP, that do not extract all the benefit from the throughput of the system. To find the best operating point, the dynamic queue length L (i.e., number of available users at the DLC layer) is maximized, subject to some system requirements in terms of the users' QoS demands. Taking into consideration the existence of outage in the QoS satisfaction, a proposed optimization procedure for the system performance can be stated as

$$\begin{aligned}&\max L\\&\quad\text{s.t.}_1\quad \text{Prob}\{\text{SNIR}_i<\text{snir}_{\text{th}}\}\leq\xi_{\text{rate}}\quad \forall i\in L,\\&\quad\text{s.t.}_2\quad \text{Prob}\{D_{\max}<K_i\}\leq\xi_{\text{delay}}\quad \forall i\in L,\\&\quad\text{s.t.}_3\quad \text{Prob}\{T_i\geq T_{\min}\}\geq 1-\xi_{\text{out}}\quad \forall i\in L,\end{aligned} \tag{17}$$

where $D_{\max}$ is the maximum allowed delay and $T_{\min}$ is the minimum required throughput per user and per slot. It is has to be noted that the previous scheme presents the dynamic queue length adjustment together with the QoS concepts (minimum allowed rate, maximum allowed delay, and minimum-guaranteed throughput), where the operator can

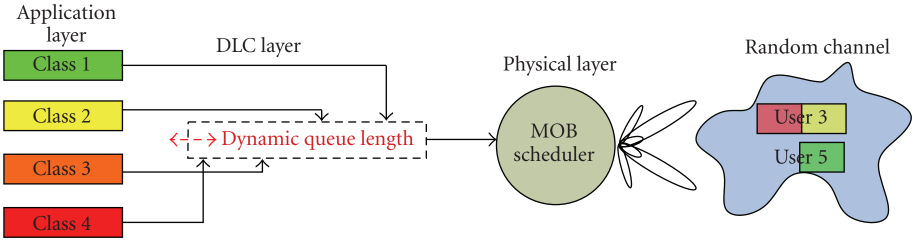

Figure 1: Dynamic queue length scheme.

Table 1: SINR values mapping to rate.

Rate (Mbps)	SINR value
0	< -8
6	-8 to 12.5
9	12.5 to 14
12	14 to 16.5
18	16.5 to 19
24	19 to 22.5
36	22.5 to 26
48	26 to 28
54	>28

choose among the QoS demands for the most appropriate ones for each scenario.

7. Performance Evaluation

To evaluate the performance of the proposed dynamic DLC queue mechanism, a heterogeneous scenario is set up where users with four types of applications coexist in the system. Two transmitting antennas $n_t = 2$ are available, so that two beams are generated and two users in the Downlink can be simultaneously serviced through the same frequency, code, and time. A total of $N = 20$ users are available in the scenario with 5 users for each service traffic class. The length of the packets for the classes 1, 2, 3, and 4 is 100, 512, 1024, and 2312 bytes, respectively. Class 1 has the highest priority, while class 4 is the lowest priority class. A saturated system is considered, where all users have at least one packet available for transmission. A total system bandwidth of 20 MHz and a slot service time of 1 ms are assumed in the simulations. An Indoor complex i.i.d. Gaussian channel with $\sim \mathcal{CN}(0,1)$ entries is considered. A time scale of 10^6 channel visualizations is employed to display the channel continuous variations. Results for an opportunistic scheduler that only transmits to a single user [16] are also shown in the figures to realize the benefits of MIMO from a higher layers perspective. Obviously, the same total power constraint is imposed on both systems in order to have a fair comparison. Table 1 shows how the SNIR values for IEEE 802.11 legacy systems are mapped to the transmission rate per beam, as stated in [17].

The efficiency of our dynamic queue length scheme is compared with a Round Robin-based scheme [18], where the channel conditions are not taken into consideration in the scheduling process, and the users access to the channel are guaranteed at fixed intervals. This technique is implemented in TDMA-based systems (e.g., GSM) and it has been proved to provide the lowest possible scheduling delay, but the obtained throughput is very low as the channel conditions are not regarded in the scheduling process. Moreover, it can not be combined with the MIMO Multiuser capability, since the application of MIMO Multiuser techniques needs for the users' selection principle to choose n_t users that show the least interference among themselves [4].

In Figure 2, the percentage of the outage in the maximum delay satisfaction for Class 1 users versus the length of the queue is presented. A maximum allowed delay of 20 ms is assumed for the class 1 users. It can be seen from Figure 2 that when the length of the queue is $L = 5$ (so that only users of the class 1 exist in the DLC queue), the maximum allowed delay is guaranteed for almost 100% of the cases (with an outage of 0.049%). Notice that increasing the queue length to 20, so that all users are eligible to be selected, the outage reaches a value of 12%. Therefore, the operator can position itself in the most appropriate point based on its requirements and its customers demands. The single user service (indicated as "no-MIMO" in the figures) provides an outage value of 2% for a DLC queue length of 5 and when the DLC queue length is 20, the outage value boosts to 36%, which is an unacceptable value for any communications system. The results show the great benefit of providing QoS delay guarantees with the MOB technique as the users are provided service more frequently (as 2 beams are generated, then the waiting time for the users is decreased, as stated in (7)), thus the probability to violate the maximum delay restriction is lower. Note the exact match between the theory and the simulations results, as approximations were not employed in the equations derivation. From Figure 2 we can also see that in a scenario of 20 users with a maximum of 20 ms maximum allowed delay (remind that the service slot time is 1 ms), then all users are serviced through the Round Robin strategy, delivering a 0% in the outage delay.

From Figure 2 we can see that the outage probability increases with the DLC queue length, which is harmful for the performance of the system. On the other hand, in order to increase the system average throughput a longer length of the DLC queue is required, so that more users are eligible for scheduling selection in the system. This means that class 1 users have lower chance to be serviced by the BS scheduler, which has a direct impact on the time delivery

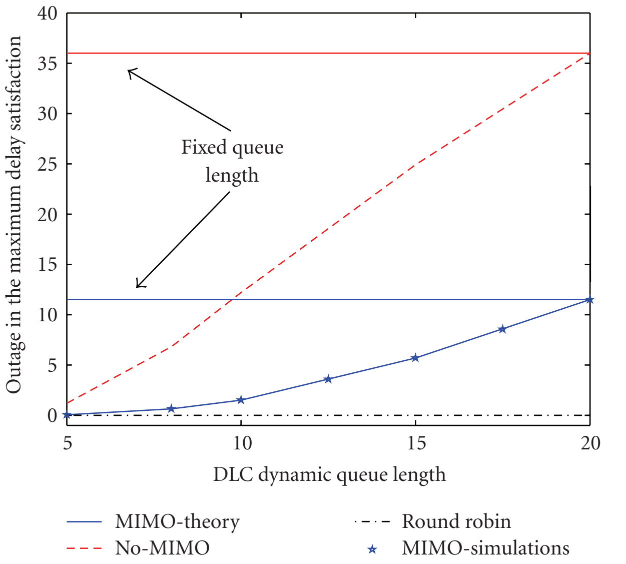

Figure 2: Outage probability (%) in the maximum delay satisfaction for Class 1 users, with a maximum allowed delay threshold = 20 ms.

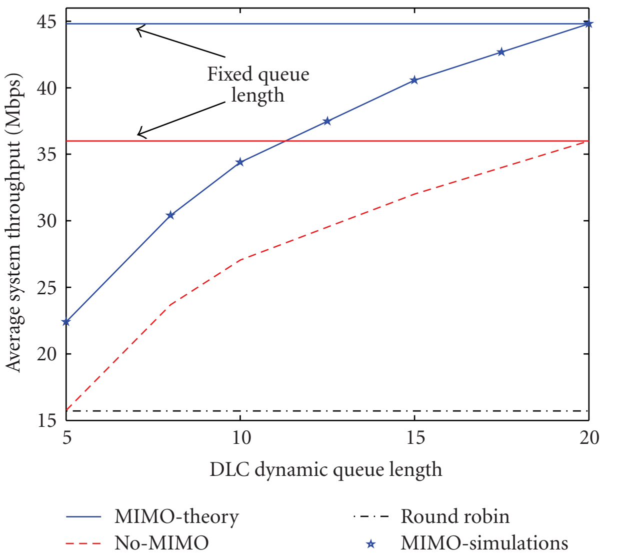

Figure 3: System average throughput for a variable DLC queue length.

of their packets. Figure 3 shows the performance of the average throughput (from (4)) for a variable DLC queue length, where as expected, increasing the queue length (i.e., the number of available users for scheduling), the average throughput values go up due to the opportunistic way of user/s selection in both of MOB and single user selection in [16]. Once again, it can be seen the exact match between simulations and the theoretical analysis. Notice the deficient performance of the round robin strategy, as the scheduler does not tackle the channel conditions, thus delivering very low system throughput performance, which handicaps its implementation in current broadband wireless systems, even of its outstanding delay performance.

Figure 3 shows how the gap between the two schedulers enlarges as the DLC queue length increases, which is motivated by the MOB performance, where a larger number of users enable a better search for a set of users (2 users in our simulations) that do not interfere a lot among them (i.e., better SNIR value). Also realize that the average throughput gain of MOB is not as amazing as the MOB gain in the outage of the QoS satisfaction, as seen in Figure 2. The explanation for this matter is due to the MOB technique where more users can be serviced (2 users in our study case), so that the users have almost twice the probability to be serviced in comparison with the single user scheduling approach in [16]. But on the other hand, the throughput average gain is not twice due to the interference that the users generate between them. Therefore, we can claim that MOB scheme is more suitable for QoS demands than average throughput performance. This conclusion has not been stated previously in the literature (up to the authors' knowledge), where this result is very interesting for the implementation of MOB (and most probably, for any other MIMO multiuser technique).

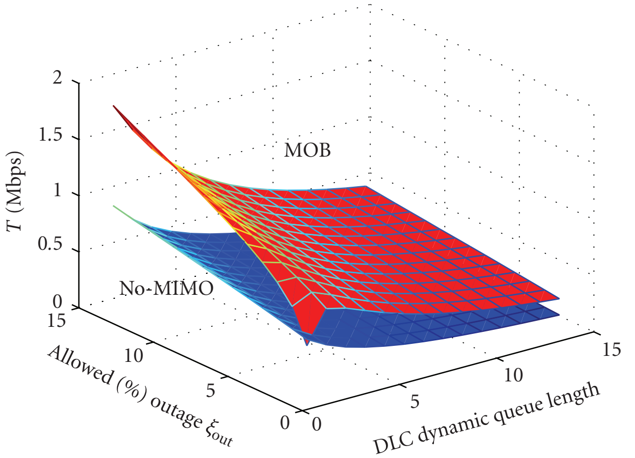

Figure 4: Minimum-guaranteed throughput per user and per slot for a variable DLC queue length and outage.

Regarding the minimum-guaranteed throughput (remind the discussion at the end of Section 5.5 about the difference between guaranteed throughput per user and the average throughput) that the system offers to each user in each slot, Figure 4 depicts it for variable DLC queue length values, as well as for variable allowed outage ξ_{out} values. For the MOB scheduler, it can be seen that there is an optimum DLC queue length L, where the guaranteed throughput has a maximum value for each considered outage. Therefore, the system can be optimized based on specific demands and restrictions, for all the classes of users. For example, if the DLC queue length (i.e., available number of users) is $L = 15$, and each user is guaranteed a minimum throughput value of 10 Mbps within a maximum of 20 time slots (i.e., 20 ms), then the minimum-guaranteed throughput value per user and per slot equals 0.5 Mbps, as shown in Figure 4.

The results for the single user scheduler are also plotted, showing lower performance than the MOB except for the case of short DLC queue length (i.e., small number of users), where the intrabeam interference in MOB limits the multiuser capability [19]. This effect is highlighted for small outage values where all resources should be awarded to a single user to avoid violating its outage constraint.

8. Future Research Directions

This work developed a QoS optimization over the system metrics to guarantee the QoS for the users, but as a future work, a joint optimization over the QoS metrics is also required to avoid any controversial results among them. Moreover, as all current broadband wireless systems are based on the OFDM Access (OFDMA) scheme, a resource management based on the subcarriers allocation is also required to align with current standards.

Another future work is related to the Hour-Aware Resource Management (HA-RMM). As it is defined in the literature, applications running over the different hours have different QoS requirements (e.g., during the night background traffic is the dominant one while during the morning, real-time traffic is needed; where each application has its own QoS demands). Therefore, a smart resource management strategy over the different day hours is desired to achieve a further optimization of the system resources.

9. Conclusions

A dynamic queue length scheduling strategy has been presented in this work for Downlink multiuser and multi-antenna WLAN systems with heterogeneous traffic. Among the users with a packet in their queue, the ones with the best channel conditions are selected for transmission. Through the MOB scheme, the length of the queue defines the maximum achievable average throughput of the system. On the other hand, the QoS requirements of the delay sensitive applications are guaranteed with short DLC queue lengths. A tradeoff appears between the system average throughput and the QoS demands of the users.

The paper proposed a dynamic DLC queue length control, so that the maximum length is allowed to obtain the highest average system throughput, but restricted to the satisfaction of the users QoS. Several alternative QoS measures are presented along the paper and in closed form expressions, so that the wireless operator can choose among them for the most suitable ones for each scenario characteristics and users' QoS requirements.

Besides the dynamic queue proposal, another important outcome of this paper is on how applications and link layers (or in general higher layers) take profit of the advances introduced by multiple antennas and signal processing techniques in the physical layer. A challenge faced by this paper is on how to deal with several aspects from the different layers of the communication process, so that we tried to make the physical layer concepts to be clear for high layers researchers, and vice versa.

Acknowledgment

This work has been partially funded by the research Projects R2D2 (TSI-020400-2009-72), NEWCOM++ (ICT-216715), and JU-DAR 162.

References

[1] N. Zorba and A. I. Pérez-Neira, "CAC for multibeam opportunistic schemes in heterogeneous WiMax systems under QoS constraints," in *Proceedings of the 50th Annual IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 4296–4300, Washington, DC, USA, November 2007.

[2] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, 2003.

[3] E. Kartsakli, A. Cateura, L. Alonso, J. Alonso-Zarate, and C. Verikoukis, "Cross-layer enhancement for WLAN systems with heterogeneous traffic based on DQCA," *IEEE Communications Magazine*, vol. 46, no. 6, pp. 60–66, 2008.

[4] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[5] N. Zorba and A. I. Pérez-Neira, "Opportunistic Grassmannian beamforming for multiuser and multiantenna downlink communications," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1174–1178, 2008.

[6] B. K. Chalise and A. Czylwik, "Robust downlink beamforming based upon outage probability criterion," in *Proceedings of the 60th IEEE Vehicular Technology Conference (VTC '04)*, pp. 334–338, Los Angeles, Calif, USA, September 2004.

[7] Q. Zhao and L. Tong, "A dynamic queue MAC protocol for random access channels with multipacket reception," in *Proceedings of the 34th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1235–1239, Pacific Grove, Calif, USA, October-November 2000.

[8] B. Ata, "Dynamic control of a multiclass queue with thin arrival streams," *Operations Research*, vol. 54, no. 5, pp. 876–892, 2006.

[9] H.-O. Choi, Y.-J. Kim, S. An, and C.-K. Nam, "Dynamic queue management mechanism for service enhancement in wireless intelligent network environment," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '99)*, vol. 1, Rio de Janeiro, Brazil, December 1999.

[10] D. I. Choi and Y. Lee, "Performance analysis of a dynamic priority queue for traffic control of bursty traffic in ATM networks," *IEE Proceedings: Communications*, vol. 148, no. 3, pp. 181–187, 2001.

[11] E. L. Hahne and A. K. Choudhury, "Dynamic queue length thresholds for multiple loss priorities," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 368–380, 2002.

[12] J. Del Prado Pavon and S. Choi, "Link adaptation strategy for IEEE 802.11 WLAN via received signal strength measurement," in *Proceedings of the International Conference on Communications (ICC '03)*, vol. 2, pp. 1108–1113, Anchorage, Alaska, USA, May 2003.

[13] M. R. Spiegel, *Theory and Problems of Probability and Statistics*, McGraw-Hill, New York, NY, USA, 1992.

[14] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[15] 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, no. 4, pp. 796–806, 2006.

[16] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings of the IEEE International Conference on Communications*, vol. 1, pp. 331–335, Seattle, Wash, USA, June 1995.

[17] D. Pubill and A. I. Pérez-Neira, "Handoff optimization with fuzzy logic in 802.11 networks," in *Proceedings of the International Symposium on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '03)*, Paris, France, September 2006.

[18] H. Fu and D. I. Kim, "Analysis of throughput and fairness with downlink scheduling in WCDMA networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2164–2173, 2006.

[19] M. Kountouris and D. Gesbert, "Robust multi-user opportunistic beamforming for sparse networks," in *Proceedings of the 6th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '05)*, pp. 975–979, New York, NY, USA, June 2005.