QoS Scheduling in Heterogeneous Traffic Multiuser Multiantenna WLAN Systems

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

Christos Verikoukis
Centre Tecnològic de Telecomunicacions de Catalunya
Barcelona, Spain
cveri@cttc.es

Ana I. Pérez-Neira
Technical University of Catalonia
Barcelona, Spain
anuska@gps.tsc.upc.edu

***Abstract*—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. An opportunistic scheduling algorithm is applied, while real time classes are prioritized. A trade-off between the throughput maximization of the system and the guarantee of the QoS requirements is obtained. Therefore the length of the queue is dynamically tuned to select the appropriate conditions based on the operator requirements.**

I. Introduction

The demand for using in-home Wireless Local Area Networks (WLANs) to support real-time delay-sensitive applications such as voice, video streaming or online-gaming has been remarkably growing during the last years. However, current IEEE 802.11 WLAN systems fail to fulfill the strict Quality of Service (QoS) requirements in terms of maximum allowed delay and/or delay jitter for such applications. Therefore, providing QoS by using the scarce resources in the wireless medium is a challenging aspect for such system objective.

Different notions of QoS are available at different communication layers [1]. At the physical layer, QoS means an acceptable signal strength level and/or Bit Error Rate at the receiver, while at the Data Link Control (DLC) or higher layers, QoS is usually expressed in terms of minimum guaranteed throughput, maximum allowed delay and/or delay jitter. The fulfillment of QoS requirements depends on procedures that follow each layer. At the DLC layer, QoS guarantees can be provided by appropriate scheduling and resource allocation algorithms, while at the physical layer, adaptation of transmission power, modulation level or symbol rate are employed to maintain the link quality.

One of the system resources that can be employed to improve the system performance in terms of both rate and QoS is the spatial resource. 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 code. Its employment has been already standardized in IEEE 802.11n and IEEE 802.16e, while it is expected to be part of LTE Standard. Among all the techniques within the MIMO technology, the Multibeam Opportunistic Beamforming (MOB) strategy has been suggested in [2] to boost the wireless link capabilities, showing high performance, low complexity design and only partial channel information is required at the transmitter side. MOB can be operated and adopted to fulfill the QoS requirements demanded by the users for their correct operation [1].

This paper proposes 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. Therefore, the proposed solution considers both the physical and application layers characteristics of the system. 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 delay-sensitive applications, where some outage [3] is considered in the QoS requirements of these applications.

II. 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 n_t transmitting antennas, and supposing that N is greater than n_t. A heterogeneous scenario[1] has been set up where users run one of 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.

A wireless multiantenna channel $\mathbf{h}_{[1\times n_t]}$ is considered between each of the users and the BS, where a quasi-static 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 i^{th} user received signal, given by

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

[1]The considered scenario is actually a multiuser Multiple Input Single Output (MISO) but its results can be immediately 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 the paper's objectives and conclusions are independent of the processing carried out at the receiver.

978-1-4244-5213-4/09/ $26.00 ©2009 IEEE

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.

III. Multibeam Opportunistic Beamforming (MOB)

One of the main transmission techniques in multiuser multiantenna scenarios is the MOB scheme [2], 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 m^{th} 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 non-intended users, representing a major drawback of this system, and stating the SNIR formulation for the i^{th} user through the m^{th} beam as

$$SNIR_{i,m} = \frac{\frac{1}{n_t} |\mathbf{h}_i \mathbf{b}_m|^2}{\sigma^2 + \sum\limits_{u \neq m}^{n_t} \frac{1}{n_t} |\mathbf{h}_i \mathbf{b}_u|^2} \tag{3}$$

where a uniform power allocation is considered. As the user with highest SNIR value is selected for each transmitting beam, then the average system throughput of MOB can be written [2] as

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

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

The MOB scheme is shown to improve the system average throughput [2], but 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 rate. 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 and the maximum allowed delay stand as QoS realistic constraints for both real and non-real 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.

IV. 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 ($R = log_2(1+snir_{th})$), and delivered to it within a maximum tolerable time delay K.

As this work deals with real-time applications in wireless systems, then the QoS demands can not be satisfied for the 100% of cases due to the channel characteristics. Therefore, some outage ξ_{out} in the QoS is accepted [3].

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. Subsection (IV-A) 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 i^{th} user is selected for transmission, and whether its rate requirement is satisfied or not. Subsection (IV-B) derives the corresponding SNIR distribution for the selected user, and obtains the minimum guaranteed rate under an outage ξ_{rate}.

A. 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

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

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[2], where this probability follows a Geometric Distribution [4] as

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

In the MOB scheme, each one of the N independent users attempts to be serviced by one of the n_t generated beams with $\overline{P}_{access} = \frac{n_t}{N}$, 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

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

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

In point to point scenarios, the queueing delay is the dominant factor in the system delay [5] 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. Therefore, the maximum number of time slots to select a user is equal to the K access slots (7), defining the maximum allowed scheduling delay.

As we consider the scheduling delay, both the buffer management and source statistics for arriving packets are not addressed [6]; and the queues stability target [5] 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.

[2]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.

B. 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] [2]

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

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

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

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

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

Therefore the minimum required SNIR ($snir_{th}$) for each user is achieved with a predefined rate outage ξ_{rate} as

$$\xi_{rate} = \left[1 - \frac{e^{-(snir_{th} \cdot n_t \sigma^2)}}{(1+snir_{th})^{n_t - 1}} \right]^N \tag{11}$$

where the values of $snir_{th}$ and ξ_{rate} can be computed on the basis on any system objectives, under the number of users N. With further manipulations, the expression (11) can be reformulated as

$$\log_2(1+snir_{th}) = \frac{\log_2\left(\frac{1}{1-\sqrt[N]{\xi_{rate}}}\right) - \lambda snir_{th} \cdot n_t \sigma^2}{n_t - 1} \tag{12}$$

obtaining the minimum guaranteed-rate, and where $\lambda = \log_2(e) = 1.4427$ is adopted. Eqn.(12) shows the rate limits of the system, indicating that high $snir_{th}$ 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 $snir_{th}$ as

$$PSR = \begin{cases} 1 & \text{if} \quad serving\ SNIR \geq snir_{th} \\ 0 & \text{if} \quad serving\ SNIR < snir_{th} \end{cases} \tag{13}$$

where a direct relation to ξ_{rate} is obtained from Eqn.(11).

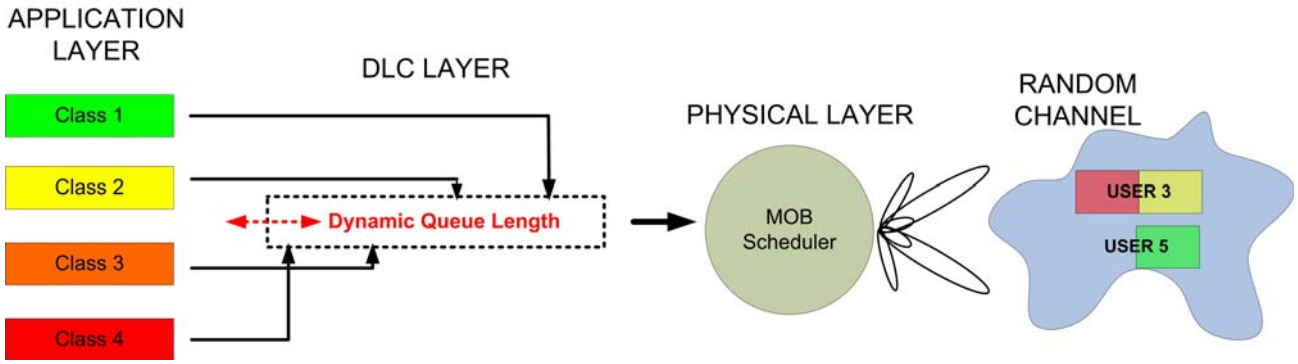

Fig. 1. *Dynamic queue length scheme.*

V. 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 Fig. 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 Eqn.(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 among these users. Likewise, the same procedure can be applied when the minimum guaranteed rate 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

$$\begin{array}{lll} \max & L & \\ s.t._1 & Prob\{SNIR_i < snir_{th}\} \leq \xi_{rate} & \forall\ i \in L \\ s.t._2 & Prob\{D_{max} < K_i\} \leq \xi_{delay} & \forall\ i \in L \end{array} \quad (14)$$

where D_{max} is the maximum allowed delay. It is has to be noted that the previous scheme presents the dynamic queue length adjustment together with the QoS concepts (minimum allowed rate and maximum allowed delay), where the operator can choose among the QoS demands for the most appropriate ones for each scenario.

VI. Performance Evaluation

To evaluate the performance of the proposed dynamic DLC queue mechanism, an 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

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

Table I: SINR values mapping to rate

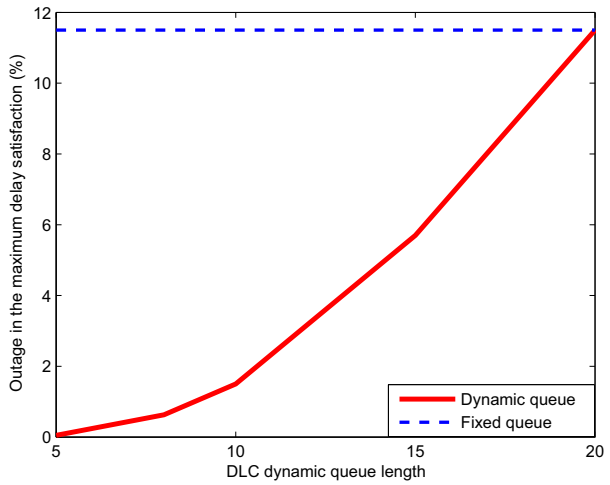

Fig. 2. *Outage probability (%) in the maximum delay satisfaction for Class 1 users, with a maximum allowed delay threshold=20msec.*

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. 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 msec are assumed in the simulations. Table 1 shows how the SNIR values for IEEE 802.11 legacy systems are mapped to the transmission rate per beam, as stated in [7].

In Fig. 2, the percentage of the outage in the maximum delay satisfaction for Class 1 users is presented versus the length of the queue. A maximum allowed delay of 20msec is assumed for the class 1 users. It can be seen from Fig. 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

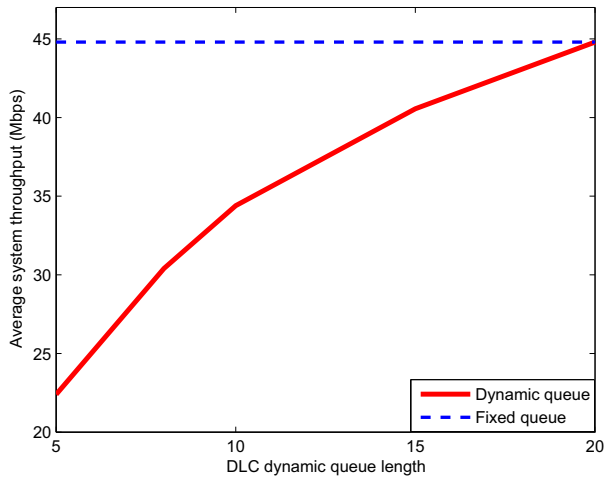

Fig. 3. *System average throughput for a variable DLC queue length.*

demands. 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 Eqn.(7)), thus the probability to violate the maximum delay restriction is lower.

From Fig. 2 we saw that increasing the DLC queue length increases the outage probability 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 of their packets. Fig. 3 shows the performance of the average throughput (from Eqn.(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 MOB.

VII. Conclusions

A dynamic queue length scheduling strategy has been presented in this work for Downlink multiuser and multiantenna 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 users' QoS demands.

Acknowledgements

This work has been partially funded by the research Projects R2D2 (CP6-013), PERSEO (TEC2006-10459/TCM) and JUDAR 162.

References

[1] N. Zorba and A.I. Pérez-Neira, "Robust Power Allocation Schemes for Multibeam Opportunistic Transmission Strategies Under Quality of Service Constraints," *IEEE JSAC special issue on MIMO for Next-Generation Wireless Networks*, no.8, August 2008.

[2] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channel with Partial Side Information," *IEEE Transactions on Information Theory*, vol.51, February 2005.

[3] B.K. Chalise and A. Czylwik, "Robust Downlink Beamforming based upon Outage Probability Criterion," *IEEE-VTC Fall*, Los Angeles-USA, September 2004.

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

[5] M.J. Neely, E. Modiano, and C.E. Rohrs, "Dynamic Power Allocation and Routing for Time Varying Wireless Networks," *IEEE JSAC*, vol.23, January 2005.

[6] 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.

[7] D. Pubill, A.I. Pérez-Neira, "Handoff Optimization with Fuzzy Logic in 802.11 Networks," *IPMU Conference*, Paris-France, September 2006.