QoS-based User Scheduling for Multiuser MIMO Systems

Marios Kountouris [(1)], Ashish Pandharipande [(2)], Hojin Kim [(2)], David Gesbert [(3)]

[(1)] France Telecom
38-40 rue du Général Leclerc
92794 Issy-les-Moulineaux, France
marios.kountouris@francetelecom.com

[(2)] Communication & Network Lab
Samsung Advanced Institute of Technology
P.O. Box 111, Suwon, Korea
pashish@ieee.org, hkim73@samsung.com

[(3)] Mobile Communication Group
Eurecom Institute
Sophia-Antipolis, France
david.gesbert@eurecom.fr

Abstract – **Next generation wireless networks are being developed to support a wide variety of data services with a broad range of Quality of Service (QoS) requirements. We consider the problem of downlink scheduling wherein a base station (BS) with multiple antennas serves a multiuser system. Assuming that different users can be served by the transmitter antennas at each slot, we allocate each BS antenna to users based on certain priority functions. We study a QoS-based allocation scheme, wherein the priority functions capture the user QoS demands quantified in terms of throughput and delay. We attempt to identify scheduling schemes that provide a good trade-off between (i) higher system capacity by exploiting inherent multiuser diversity, (ii) fairness among users based on their instantaneous channel conditions relative to average channel conditions, and (iii) tolerable latency requirements specified by the user applications. Simulation results evaluate the performance of the above allocation scheme and show that multiple antennas and delay-aware scheduling rules can be efficiently used to provide QoS at little expense of throughput.**

Keywords – **Multiuser MIMO systems, packet scheduling, broadcast channels, Quality of Service.**

I. Introduction

Recently, there has been an increasing demand for wireless connectivity for both delay-tolerant and delay-sensitive applications, such as voice and video conferencing. Additionally, for applications such as wireless internet and video, a high data rate downlink is needed. Multiple Input Multiple Output (MIMO) systems have been recognized lately as a promising solution to provide high data rates and highly reliable data link. We consider the application of MIMO techniques on the downlink of a multiuser point-to-multipoint system, and we examine user scheduling for wireless packet transmission in order to maximize system throughput, providing also fairness in the allocation of system resources.

Traditionally, the fundamental problem of resource allocation among users of a system has been treated separately, both practically and conceptually. Providing diverse Quality of Service (QoS) guarantees to users is a challenging issue, especially considering the dichotomy between the physical and the medium access control (MAC) layer. At the physical layer, QoS is synonymous to an acceptable signal-to-noise ratio (SNR) level, minimum rate or bit error rate (BER) at the receiver, while at the MAC or higher layers, QoS is usually expressed in terms of maximum delay guarantees or delay jitter for a certain minimum rate. Although that dichotomy has the advantage of offering a better understanding of the functionality of each layer, there is a growing awareness that significant performance gains can be achieved by a physical layer that operates in synergy with the MAC layer. This cross-layer approach can address efficiently the capacity/delay and capacity/fairness tradeoffs.

A challenging task in designing such systems is to meet the QoS demands of multiple users while maintaining high system throughput. In wireless communication systems where a common medium is shared, a good scheduling policy should provide a satisfactory tradeoff between (i) maximizing capacity, (ii) achieving fairness, and (iii) satisfying delay constraints of real-time application users. In this paper, we consider a multiuser system serviced by a base station (BS) with multiple antennas. Our goal is to endow the BS with the capability to service users such that the aforementioned objectives are met by the scheduling algorithm. We examine the throughput and delay performance of a multiple antenna allocation scheme applied on the downlink of a heterogeneous MIMO system, where users have different QoS constraints. Instead of using just throughput or bit error rare as QoS metric, the maximum tolerable delay and maximum dropping probability are also taken into account.

Most of the prior work on scheduling has focused either on maximizing system throughput or achieving fairness for users with delay tolerant applications. In this context, scheduling algorithms can exploit channel variations across users and attempt to transmit to users with the “best” channel conditions. This effect is called multiuser diversity [1], [2], and has given the rationale to a group of scheduling rules referred often as ‘opportunistic’. Opportunistic scheduling (OS) attempts to exploit the time varying propagation channels between the BS and the mobile stations (MSs) when they reach their peak rate capability and defer using channels when in bad state. However, when the fading is slow in comparison with an acceptable packet delay, or weak, OS is not very useful. Proportional fair scheduling [3], [4] is another approach seeking to maximize long-term average throughputs, thus maintaining long-term fairness among users. To mitigate the problem of delay guarantees several algorithm have been proposed. In [5], users in a Code Division Multiple Access (CDMA) cell are scheduled taking into account the channel variations as well as their QoS requirements in terms of probabilistic packet delay bounds. In [6], an algorithm which strikes a balance between throughput and delay constraints was proposed but its complexity limited its usage. Another approach is to use dynamic programming [7] to design schedulers that can increase capacity while maintaining QoS guarantees, but in such an approach the program's state space grows exponentially with the number of users and the delay requirement.

The rest of the paper is organized as follows. In Section II, we present the system model and the MIMO transmission scheme that is employed. In Section III, we present the scheduling rules that are used as priority functions in our QoS allocation scheme. In Section IV, we evaluate the performance of the scheme in terms of throughput and delay through simulations. Finally, Section V summarizes the paper.

II. System Model

We consider the downlink channel of a system with a single BS having $M > 1$ transmit antennas, and K users each having $N \geq 1$ receiver antennas. This model is a multiple antenna broadcast channel. The received signal at user k at time slot t is given by

$$y_k(t) = \sqrt{\zeta_k} H_k(t) s(t) + n_k(t)$$

where $s(t)$ is the M×1 transmitted signal at time slot t, H_k the M×N complex channel matrix, ζ_k the received SNR, and n_k is a N×1 additive noise. We assume that the channel matrix H_k is perfectly known to the receiver, and that H_k and n_k have a zero mean and unit variance complex Gaussian distribution. The transmitter is subject to an average power constraint and we assume that the total power is equally distributed to the transmit antennas. The total transmit power is assumed to be $E\{s^*s\} = M$, thus the transmit power per antenna is one. Our allocation scheme uses multiple transmit antennas for downlink transmission assuming that it has the possibility of joint, or independent, scheduling across antennas. In order to support simultaneous transmission to multiple users, transmit beamforming is used. Although sub-optimal, beamforming is equivalent to dirty paper coding (DPC) at high SNR [8]. For demonstration reasons, our method is built on recent advances realized in [9] in the area of multi-user downlink precoding and scheduling based on partial transmitter channel state information (CSIT). The use of a transmission scheme similar to [9] does not restrict the generality and validity of our method. Under minor modifications, our QoS-based scheduling scheme can be implemented for several other existing beamforming and channel inversion schemes [10, 11]. The key point is to be able to express the rate of each user if served from any of the transmit antennas. Depending on the scheme, the rate can be calculated either from the signal-to-interference-plus-noise ratio (SINR) or from the effective channel gains. A brut force approach is to try calculating the achievable rates exhaustively for all possible user permutations. Let,

$$S \subset \{1,...,K\},\ |S| \leq M$$

be a subset of users that the BS intends to transmit to. The search space of S is given by:

$$\sum_{i=1}^{M} \binom{K}{i}$$

and becomes very large for large K. Nevertheless, in practical systems the number of active users K is not so prohibitively large. If the number of users is small, the achievable rates of each user can be calculated via exhaustive search on all the possible combinations of user grouping for scheduling.

The transmission scheme we adopt here generates M random beams independently from one time slot to the other. Assuming $N=1$, a MxM unitary matrix Q is generated according to an isotropic distribution. At time slot t the transmitted signal is

$$s(t) = \sum_{m=1}^{M} q_m(t) s_m(t)$$

where $s_m(t)$ is the m-th transmit symbol at time slot t and q_m are M×1 random orthonormal vectors for $m=1,...,M$.

Therefore, the received signal at the k-th receiver is,

$$y_k(t) = \sum_{m=1}^{M} H_k(t) q_m(t) s_m(t) + n_k(t)$$

Assuming that the k-th receiver knows $H_k q_m$ for $m=1,..,M$, it can calculate its SINRs on each one of the M random beams using:

$$SINR_{k,m} = \frac{\left|H_k q_m\right|^2}{1/\zeta_k + \sum_{j \neq m} \left|H_k q_j\right|^2}$$

Each user feeds back its SINR for each one of the M beams, and the transmitter assigns s_m to the user indicated by the priority functions that we would present in the following section. That is the information our scheme needs under any feasible transmission scheme. Once we have calculated the rates each antenna or beam can assign to each one of the active users of the system, either by knowing the other users to be scheduled, thus the interference, or by checking all possible combinations, our QoS-based scheme remains unchanged.

III. QoS-Based Scheduling

We have assumed that the transmitter can send packets to different users at each timeslot. In a given time slot, each transmitter antenna independently chooses the user with highest priority, where certain user priority functions model their respective QoS requirements. In order to decide the best user set to be allocated on each transmit antenna, we consider the following scheduling rules: the *Maximum Delay* rule (MAX-DELAY), the *Maximum Rate* rule (MAX-RATE) [2], the *Modified Largest Weighted Delay First* rule (M-LWDF) [5], the *Proportionally Fair Scheduler* (PFS) [4], and the *Exponential rule* (EXP) [5]. These different rules result in different priority functions.

Our goal is to gain an understanding of the QoS performance of these schemes when applied to multiuser MIMO systems. We define $R_k(t)$ as the actual rate supported by the channel of user k. This rate is constant over one slot. $T_k(t)$ is the mean data rate observed by user k over a long sliding window, and $W_k(t)$ the head-of-the-line (HOL) latency, i.e. the amount of time the HOL packet of user k has spent at the base station. Each user has its own probabilistic QoS requirement [5] of the form

$$\Pr\{W_k > D_k\} \le \delta_k$$

where parameters D_k and δ_k are the latency threshold and the maximum probability of exceeding it, respectively.

The MAX-DELAY discipline schedules the user whose HOL packet has spent the longest time at the base station, i.e.,

$$j = \arg\max_k (W_k(t))$$

The MAX-RATE rule schedules the user whose channel can support the largest rate over the next slot.

$$j = \arg\max_k (R_k(t))$$

The PFS selects the best user such as:

$$j = \arg\max_k \left(\frac{R_k(t)}{T_k(t)} \right)$$

where $T_k(t)$ is the mean rate actually given to the user k, and measured over a relative long "sliding window" of size t_c. The average throughputs $T_k(t)$ can be updated using the following exponential filter [4]:

$$T_k(t+1) = \begin{cases} \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} R_k(t), & k = k^* \\ \left(1 - \frac{1}{t_c}\right) T_k(t), & k \neq k^* \end{cases}$$

The M-LWDF scheduler aims to balance the weighted latencies of individual user all by trying to utilize the wireless channel characteristics efficiently. Explicitly, user

$$j = \arg\max_k \left(\gamma_k R_k(t) W_k(t) \right)$$

is scheduled at each time slot. In [5], it was found that a good choice for γ_k is $\gamma_k = \alpha_k / T_k(t)$, where $\alpha_k = -(\log\delta_k)/D_k$ characterizes the desired QoS levels for the individual users.

Finally, the Exponential scheduler tries to equalize the weighted latencies of all the queues when their differences are large [5], [12]. At the slot t, user

$$j = \arg\max_k \left(\gamma_k R_k(t) \exp\left(\frac{a_k W_k(t) - \overline{aW}}{1 + \sqrt{\overline{aW}}} \right) \right)$$

is scheduled. Here, $\overline{aW} = (1/N) \sum_k a_k W_k$ and γ_k is the same as in the M-LWDF rule.

Under QoS-based scheduling in a multiuser system, we assign on each beam the user that is selected from the above five scheduling rules which have been mainly used up to now for the single user case. Thus, on each antenna, we take independent scheduling decisions and user allocations, without neglecting the possibility that all transmit antennas could be assigned to the same user. In fact, through simulations we saw that except the Max-Delay rule that assigns all antennas to the user with the longest delay, the rest of the rules transmit simultaneously to different users all the time.

IV. Simulation Results

To assess the relative performance of the MIMO scheduling schemes, we consider as metrics the latency, fairness and average rate. We assume that the fading process $\{h_k(t)\}$ evolve in time according to Jakes' autocorrelation model, where $E\{h_k(t)h_k(t-\tau)^H\} = J_0(2\pi f_D T\tau)I$, with f_D and T denoting the one-sided Doppler bandwidth (in Hz) and the sampling duration respectively. We have assumed a heterogeneous network in the sense that the received SNR for all users are not identical. Specifically, we suppose that users have SNRs uniformly distributed from 3dB to 15dB, therefore the users corresponding to the SNR of 3dB and 15dB are the worst and the best users, respectively. In our simulations we have considered mobile speeds 3km/h. The service rates change in time randomly for different users and all active users have infinite backlog. The actual rate of user k is calculated using $R_k = log_2(1+SINR_k)$. The scheduling decisions are made once every slot, which is defined as T = 1 ms, and the results are averaged over Monte Carlo simulations. The QoS requirements are chosen as follows: the first quarter of users has delay threshold 4sec with probability 0.21, and the second quarter has delay threshold 3sec with probability 0.1 of exceeding it. The third and the second quarters of users have delay threshold 2 sec and 1sec with dropping probability equals to 0.1 and 0.05 respectively. The length of the window in PFS is t_c = *500 slots.*

In Fig. 1 and 2, we plot the sum capacity (in total number of bits per second per hertz (b/s/Hz)) and the average delay of the downlink channel as a function of the number of antennas, respectively. We assume that there are 200 single-antenna MSs in the system. It should be noted that the use of multiple antennas has significant impact on the delay performance of Max Rate scheduler. From Fig. 2, we can see that the average delay of Max Rate is significantly decreasing while the number of antennas is increasing. The small degradation in throughput of the PFS and EXP compared with Max Rate rule is the cost we have to pay in order to take into account fairness in the PFS definition or QoS satisfaction as that that EXP tries to guarantee. However, both rules have the same scaling laws as Max-Rate rule, which is particularly desirable.

In Fig. 3 and 4 we plot the sum capacity and the average delay of the downlink channel as a function of active users, respectively.

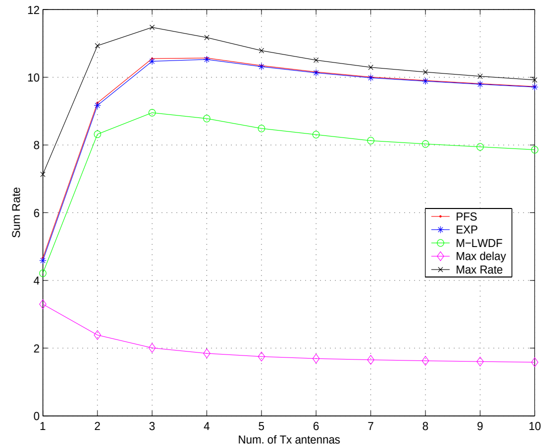

Figure 1. Sum Capacity versus the number of transmit antennas for a heterogeneous network with various QoS constraints.

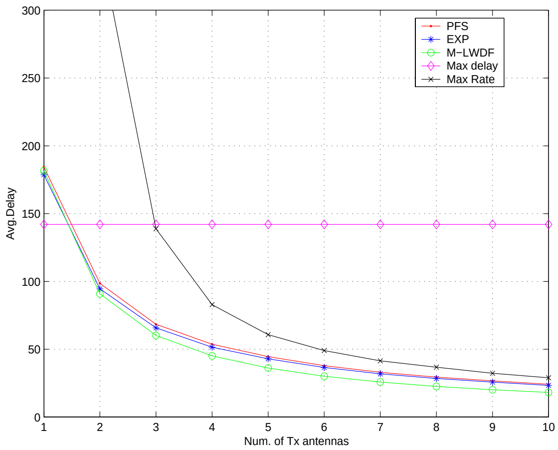

Figure 2. Average user delay (in ms) versus the number of antennas for a heterogeneous network with various QoS constraints

The BS has six antennas and serves single-antenna users uniformly distributed throughout the system. The fact that the receivers have one antenna does not affect the validity of our model; having more receiver antennas results in better interference cancellation or stream multiplexing. As it was expected the sum capacity of the Max Rate rule increases with the number of users in the system exploiting the multiuser diversity. Similar performance is seen also by PFS and EXP rule, while M-LWDF is sub-optimal in terms of sum capacity. Max-Delay is harshly sub-optimal in throughput as it does not take into account the channel variations. In a single-user case, Max-Delay is the optimum scheduler in terms of average delay, but this is different in the multiuser case. Assigning all antennas to the same user at each slot, it decreases the served user's delay at the expense of other users' delays. All the other rules showed better average delay performance as they can exploit the freedom of simultaneous transmission to different users.

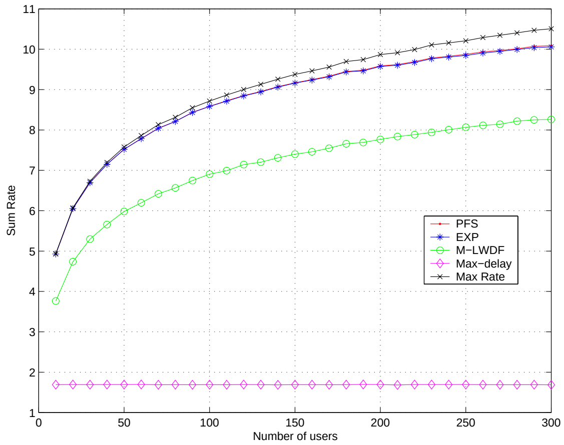

Figure 3. Sum Capacity as a function of the number of active users

The delay distribution of the worst and the best user is given in Figure 5. With dashed lines we plot the distribution of the best user and with solid lines that of the worst user. We should note that the delay distributions depend among others on the SNR distribution of users and the cell loading. Thus, this figure can be used to evaluate the behavior of each rule and so the values have only relative significance. One of the main claims is that EXP rule shows to have excellent delay distribution performance, approaching that of Max-Rate for the case of the best user. The delay distribution of the worst user under Max Rate rule can often explode and violate the QoS constraints. In addition to that, QoS-based schedulers (M-LWDF, EXP) improve the delay performance of users with relatively bad channel conditions. PFS rule in the multiuser case has delay performance close to that of EXP rule, something which is not the case in the single-user case. This is due to the fact of the use of multiple antennas for transmission to different users at the same time. Something that is also remarkable in the multiuser case is that PFS approaches EXP rule delay performance, and EXP rule approaches the throughput of PFS, being only around 0.02 b/s/Hz lower than PFS.

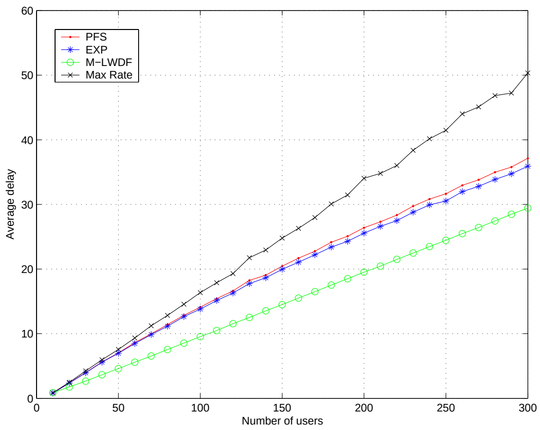

Figure 4. Average user delay (in ms) as a function of active users in the system. (Max delay performance is not shown for illustration reasons as its delay explodes and it is higher than all the others)

Figure 6 shows the number of times that each user with the corresponding SNR is chosen out of 20.000 channel realizations. It is obvious that EXP and PFS allocate at the same way most of the times the system resources, whereas Max Rate offers more resources to users with strong received SNR values. As M-LWDF is the rule that tries the most to guarantee the statistical QoS of the users, it allocates resources mainly based on QoS requirements and not based on channel conditions. This can be seen by the four distinct allocations regions (steps) in Fig. 6 which are defined by the difference to the QoS requirements.

Finally, we kept track of how many times PFS, EXP and Max Rate take the same decisions on user allocation. Although PFS and EXP seem to have almost identical performance in terms of sum rate and average delay, this does not necessarily implies that they allocate transmit antennas in the same way all the time. This can be seen by the fact that the per user capacity and delay under PFS and EXP was shown to be quite different in our simulations. The percentage that PFS and EXP decide to transmit to the same users decreases when the number of active users increase for fixed *M*, and increases when the number of antennas is increasing for fixed number of users. For small number of users, PFS and EXP rules schedule the same users from the same antennas around 85% of the times, falling to 25% as *K* is approaching 300. In a system with 300 users, PFS works the same as EXP in 8% of the times for small number of antennas and reaches up to 25% for number of transmit antennas bigger than 4.

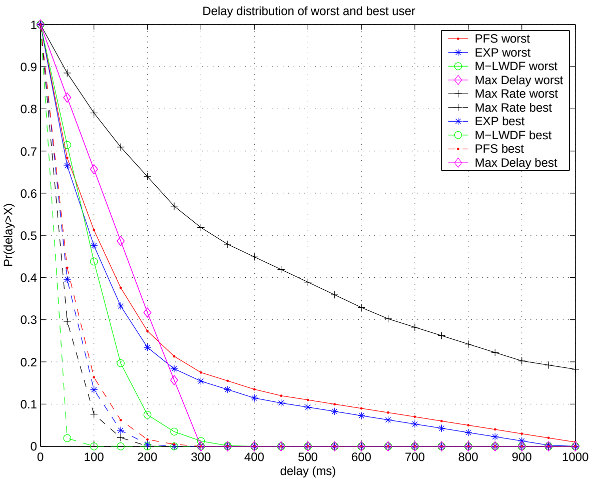

Figure 5. Delay distribution of the strongest and the weakest users of the heterogeneous network under different scheduling rules

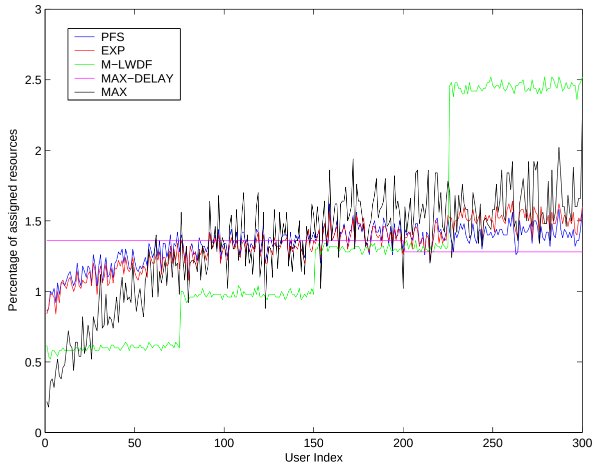

Figure 6. Frequency that each user with the corresponding SNR is chosen under different scheduling rules

Finally, through simulations we remarked that the EXP rule supplies larger throughput to the good channel users than M-LWDF, which gives more throughput to the bad channel users resulting to poor throughput performance. Comparing PFS and EXP rule, one might see that EXP rule assigns slightly more resources to weakest users than PFS, while high SNR users are given more throughput from PFS rule than EXP rule. This is evidently due to the fact that EXP rule tries to satisfy the QoS requirements of all users, even if they are in bad channel conditions, especially when their delay is increased more than the average delay of all users.

V. CONCLUSION

In this paper, we provided a framework for scheduling and antenna assignment in a multiuser MIMO system based on QoS priorities. We proposed a simple scheme how to use existing delay-sensitive schedulers in MIMO downlink channel and we evaluated its performance with different QoS-based priority functions. Through simulation study, we showed that exploiting the degrees of freedom offered by MIMO systems and transmitting simultaneously to different users according to delay-aware schedulers, QoS can be provided at little expense of throughput.

REFERENCES

[1] R. Knopp and P. A. Humblet, "Information Capacity and power Control in Single-Cell Multiuser Communications", *Proc. IEEE ICC*, Seattle, WA, June 1995.

[2] R. W. Heath, Jr., M. Airy, and A. J. Paulraj, "Multiuser Diversity for MIMO Wireless Systems with Linear Receivers", *Proc. of IEEE Asilomar Conf. on Signals, Systems, and Computers*, pp. 1194 -1199, vol.2, Pacific Grove, California, Nov. 4 - 7, 2001.

[3] F. Kelly, "Charging and Rate Control for Elastic Traffic", *European Transactions of Telecommunications*, vol. 8, pp. 33-37, 1998.

[4] P. Viswanath, D. Tse and R. Laroia, "Opportunistic Beamforming using Dumb Antennas", IEEE Transactions on Information Theory, vol. 48(6), June, 2002.

[5] M. Andrews, K. Kumaran, K. Ramanan, S. Stolyar, R. Vijayakumar, P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions", *Bell Labs Technical Memorandum*, April 2000.

[6] I. Bettesh and S. Shamai, "A low delay algorithm for the multiple access channels with Rayleigh fading," in *Proc. IEEE Personal, Indoor and Mobile Communications (PIMRC'98)*, 1998.

[7] I. Bettesh and S. Shamai, "Optimal power and rate control for fading channels," in *Proc. IEEE Vehicular Technology Conference*, Spring 2001.

[8] N. Jindal and A.J. Goldsmith, "Dirty Paper Coding vs. TDMA for MIMO Broadcast Channels," *to appear in IEEE Trans. on Information Theory*, 2005

[9] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channel with Partial Side Information," *to appear in IEEE Trans. on Information Theory*, 2004.

[10] B.M. Hochwald, C.B. Peel, and A.L. Swindlehurst, "A Vector-Pertubation Technique for Near-Capacity Multi-Antenna Multi-User Communication – Part I & II," in *Proc. of the 41st Allerton Conference on Comm., Control , and Computing*, October 2003.

[11] T. Yoo, and A.J. Goldsmith, "Optimality of Zero-Forcing Beamforming with Multiuser Diversity", *submitted toIEEE International Conf. on Communications (ICC)* 2005

[12] S. Shakkottai and A. Stolyar, "Scheduling for Multiple flows Sharing a Time-Varying Channel: The Exponential Rule", American Mathematical Society Translations, Vol. 207, 2002.