REAL-TIME IMPLEMENTATION OF CROSS-LAYER OPTIMIZATION: MULTI-ANTENNA HIGH SPEED UPLINK PACKET ACCESS

T. Haustein[⋆]*, M. Wiczanowski*[†]*, H. Boche*[†] *and E. Schulz*[††]

[⋆]Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, 10587 Berlin, Germany, haustein@fhg.hhi.de

[†] Technical University of Berlin, Heinrich Hertz Chair for Mobile Communications, Einsteinufer 25, 10587 Berlin, Germany

marcin.wiczanowski, holger.boche@TU-Berlin.de, [††] Siemens AG ICM N PG NT RC FR, Information and Communication Mobile Networks R&C, Sankt-Martin Str.76, 81541 Munich, Germany, egon.schulz@siemens.com

ABSTRACT

In this paper we present a multi-antenna real-time implementation of a cross-layer uplink scheduler based on a recently developed framework. We outline the theory behind the stability-optimal scheduler, which achieves the entire stability region of the multiple access channel. We discuss the trade-off between theoretical optimality and real-time operation under real-world constraints, which lead to the development of an efficient *fair scheduler*. Finally, we evaluate the performance of the *fair scheduler* in experimental measurements (on state-of-the-art hardware) in a typical office scenario, which show an increase in spectral efficiency and the possibility to guarantee a certain link quality of service thresholds.

1. INTRODUCTION

The current evolution of mobile communication technology is driven by rising customer requirements in terms of provided service art, service quality and availability. This tendency is mirrored in the increasing importance of data-based services (so called *new services*), which are expected to outnumber traditional voice connections in the future. The tendency is already recognizable on the design of UMTS networks, currently rolling out all over the globe. In this context the transmission schemes High Speed Uplink Packet Access (HSUPA) and High Speed Downlink Packet Access (HSDPA) are of fundamental importance for the realization of broadband data services with up to 50 Mbit/s. Those rate requirements are challenging and can only be satisfied under the use of multiple antennas at the base station (BS) and (possibly) at the mobile terminal side (MT), both together corresponding to a multiple input multiple output (MIMO) transmission.

One crucial requirement for efficient exploitation of the available frequency spectrum is a scheduled adaptive transmission strategy which considers the channel qualities of each link, their individual quality of service (QoS) requirements. The spectral efficiency promised for multi-antenna systems [1, 2] is considered to be achievable in practical applications only under the use of such scheduled transmission. Recent designs of scheduling policies have leaned towards joint treatment of the physical and data link layer (cross-layer design) and hence towards joint optimization of issues like buffer occupancies, delays, data rates etc. Among cross-layer policies, the policy achieving the entire stability region of the uplink (a so called stability-optimal policy) might be of special interest to a network operator. Such policy allows for the service of the densest traffic under system stability, i.e. under finiteness of all buffer occupations, see [3], [4] and references therein. The stability-optimal policy provides balanced throughput maximization, which is also interesting in terms of pure physical layer issues.

In this work we shortly outline the theory behind the stability-optimal policy and discuss the trade-off of optimality against complexity and real-world constraints. The analysis gives rise to a *fair scheduler*. We evaluate the performance of the fair scheduler and compare it to other schedulers by means of experimental measurements in a typical office scenario. The results show high efficiency of fair scheduling and its flexibility in terms of assuring link QoS thresholds.

2. SYSTEM MODEL AND THEORY OUTLINE

We assume a cellular uplink with K MTs (mobile terminals) indexed by $k = 1, 2, \ldots, K$, each equipped with multiple transmit antennas. The base station (BS) is assumed to have full channel state information (CSI) and to utilize the MMSE (minimum mean square error) detector with successive interference cancellation (SIC). In the physical layer we denote the vector of data rates in n-th time slot as $\mathbf{R}(n) = (R_1(n), R_2(n), \ldots, R_K(n))$ and group the MIMO channel states in the matrix set $\mathcal{H}(n) = \{\mathbf{H}_1(n), \mathbf{H}_2(n), \ldots, \mathbf{H}_K(n)\}$. Similarly, we group the instantaneous transmit covariance matrices in the matrix set $\mathcal{Q}(n) = \{\mathbf{Q}_1(n), \mathbf{Q}_2(n), \ldots, \mathbf{Q}_K(n)\}$. The SIC-orders are denoted by permutation symbols $\pi = \pi(1) \leftarrow \pi(2), \ldots, \leftarrow \pi(K)$, where $\pi(1)$ is the last decoded link, ... and $\pi(K)$ the first decoded link. In the data link layer we assume the K processes of bit arrivals into the buffers to be bulk Poisson processes (i.e. independent arrival times of bursts of variable size in [bit]). The bit arrival rates in [bit/s] are grouped in the vector $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_K)$. Similarly, we denote the vector of instantaneous buffer occupancies (bit queue lengths) as $\mathbf{q}(n) = (q_1(n), q_2(n), \ldots, q_K(n))$.

We regard scheduling policies as slot-wise mappings associating a transmit strategy with the cross-layer system state, i.e.

$$\{\mathcal{H}(n), \mathbf{q}(n), n\} \longrightarrow \phi(\{\mathcal{H}(n), \mathbf{q}(n), n\}) = \{\mathcal{Q}(n), \pi_k(n)\}. \tag{1}$$

The principle of policy computation is depicted in Fig. 1. With the above policy notion and iid-property of the fading processes over time the queue system evolves like a *Discrete Time Markov Chain* (DTMC) and for the evolution each queue $1 \leq k \leq K$ we have

$$q_k(n+1) = [q_k(n) - R_k(\phi, \mathcal{H}(n))T]_+ + a_k(n), \tag{2}$$

with $a_k(n)$ as the number of bits arrived at k-th queue in the n-th time slot. The objective of our desired policy is associated with

the notion of stability of the queue system. Stability can be characterized by several different definition, like e.g. weak stability, strong stability, non-evanescence [3] etc. A special role is played by so called observation-based stability notion, which gives rise to the definition of the stability region.

Definition*The system of K queues is called stable (in the observation-based sense), if for all $k = 1, 2, \ldots, K$ holds*

$$\lim_{M\to\infty} g_k(M) = 0, \tag{3}$$

with

$$g_k(M) = \limsup_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}_{\{q_k(\tau)\geq M\}}\, d\tau, \tag{4}$$

and $\mathbf{1}_{\{q_k(\tau)\geq M\}}$ as the indicator function of the event $q_k(\tau) \geq M$.

Definition *The stability region $\mathcal{D}$ of the system of K queues is the set of all arrival rate vectors $\boldsymbol{\rho}$, such that there exists a policy achieving stability in the observation-based sense for all $\boldsymbol{\rho}$ from the interior of $\mathcal{D}$.*

The scheduling policy achieving the entire stability region of the MIMO uplink can be referred to as stability-optimal and is of interest to the network operator. According to both definitions, in broad terms stability-optimal policy allows the operator for handling the densest traffic under finiteness of bit queues at all times. In other words it minimizes the set of arrival rates, which lead to infinite blow-up of bit queues and force the operator to drop servicing of some links, resulting in a decreased revenue. It can be shown that the stability region in the MIMO uplink corresponds to the ergodic capacity region in the uplink [3], [4].

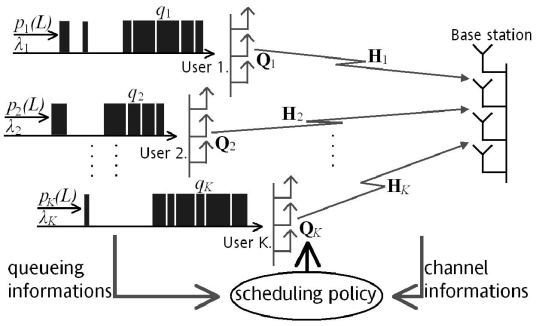

Fig. 1. Routine to compute a cross-layer scheduling policy in the MIMO uplink. λ_k and $p_k(L)$ are the burst arrival rate and burst size distribution, the bit arrival rate is derivable according to $\rho_k = \lambda_k \int Lp_k(L)dL$.

It is not surprising for readers familiar with computer network control and switch theory, that the stability-optimal policy for the MIMO uplink bases on the maximization of a weighted sum of rates, with weights corresponding to queue lengths. The weighted sum is associated with the stability issue throughout different application areas. However, for the MIMO uplink such characterization does not provide direct information on the optimal transmit covariance matrices and the optimal SIC order. The desired characterization is provided in [3] by the following statement.

Proposition *The largest stability region in the MIMO uplink with K links is achieved by the scheduling policy $\hat{\phi}$ using in each time slot $n \in \mathbb{N}$ the transmit covariance matrices of form*

$$\mathbb{Q}_s = \arg\max_{\mathbb{Q}_s\in\mathcal{M}} \sum_{k=1}^{K} q_k(n) R_k\left(\mathbb{Q}_s, \mathcal{H}(n)\right) \tag{5}$$

and the SIC order π satisfying

$$q_{\pi(1)}(n) \geq q_{\pi(2)}(n) \geq \ldots, \geq q_{\pi(K)}(n) \geq 0. \tag{6}$$

Using the notion of *S-rate region* $\mathcal{S}_\pi$ as regions of rates achievable under fixed SIC-order π, the above proposition is easily interpretable in terms of optimization over rates. The stability optimal rate vector in each time slot solves $\max_{\mathbf{R}\in\mathcal{S}_\pi(\mathcal{H}(n))} \mathbf{q}^T(n)\mathbf{R}$, with π characterized by (6). Since the optimization objective in (5) represents a hyperplane with normal vector $\mathbf{q}(n)$, the stability-optimal rate vector corresponds to the point at the boundary of the S-rate region $\mathcal{S}_\pi$ (with π from (6)) which is supported by the hyperplane with normal vector $\mathbf{q}(n)$. Note, that the stability-optimal policy is of pure spatial nature, i.e. does not need to utilize time-sharing techniques. Furthermore, the stability-optimal SIC order is independent of physical layer issues (fading, noise) and is determined solely by the queue system states. The theory outlined above is addressed in detail in [3], [4] and references therein.

We observe that the stability-optimal policy can be also regarded as the policy of balanced throughput maximization. This allows for its flexible use and application to pure physical layer goals. For instance, replacing the part of the weights in the objective by rate-priority factors for the links can result in throughput optimization with good stability behavior. Furthermore, under symmetric arrival rates we could set all weights in (5) equal without introducing significant stability performance loss. This would clearly result in a classic throughput maximization. Such flexibility features of the stability-optimal policy give rise to the design of efficient real-time schedulers under acceptable stability-optimality loss, discussed below.

3. FROM THEORY TO PRACTICE

The complexity of the cross-layer optimization discussed in section 2 might exceed the real time computational capabilities of today's hardware. Still, we can profit from the theory results if we take relevant parameters carefully into account which might simplify the computational complexity significantly. For instance, in the case of a multi-user scenario with only one antenna at each mobile (a very likely scenario as a first step to enhance the system throughput by MIMO signal processing at the BS) the calculation of transmit covariance matrices reduces to simple power allocation per active user. The sum power constraint which is a very common theoretical assumption will in reality not be applicable with reasonable effort since transmit power amplifiers have a limited dynamic range and occasional power peaks for certain Tx antennas will hardly justify amplifiers at much higher costs. Therefore an individual per antenna power constraint plus an additional sum power constraint will be more realistic in a real application. Furthermore in some scenarios a suboptimum but fast scheduler might be preferred to an optimum but slow scheduler. For instance, if individual power constraint is valid and a MMSE-SIC receiver is used at the BS then the problem reduces to finding the right user set. The right choice of a user set supported at a certain instance plus the right detection order is most important for the system stability (in reality the data buffer at each terminal will be limited to some kbyte or Mbytes). So, in practice we have to match the complexity of the scheduling task (MFlops) at a given computational capability of the scheduling unit (MFlops/s) with the timing constraints of the real time scheduling application. If this is done in a smart way the sub-optimality of the applied scheduling policy can be adjusted automatically with higher processing power at the BS.

For the initial MU-scheduling experiments which already showed obvious advantages of even sub-optimum scheduling in a cross-layer manner, we used 4 users with one Tx antenna each and a BS with 3 Rx antennas, which means that a maximum number of three users can be supported per time slot. For the cross-layer scheduler which will be denoted as *Fair Scheduler* in the following we chose always the user with the longest queue state and then two more users were selected which maximized the sum throughput with the first selected user. Now, SIC and individual power constraint require that each user transmits at maximum power and the optimum SIC detection order is found when the user with longest queue is detected last and the user with the shortest queue is detected first. In this way a cross-layer scheduling policy could be implemented in real time for a typical office scenario.

4. EXPERIMENTAL SETUP

Several scheduling policies were implemented on a reconfigurable multi-antenna test-bed [5] using a hybrid FPGA and DSP architecture for the base band signal processing. The configuration used for the measurements is depicted in Fig. 2 The MTs are distributed

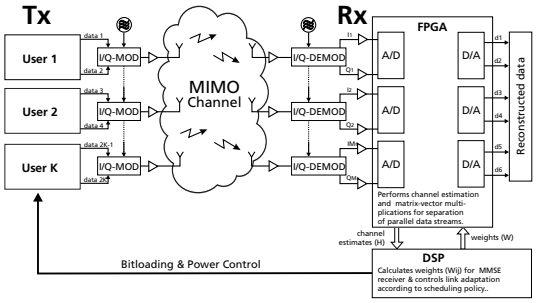

Fig. 2. Setup of the test-bed for the MU SIMO MAC scenario.

in the room such that the averaged single user channel gains are similar. To introduce sufficient channel statistics we move the BS across the room on a railway like construction which allows high reproducibility of the channel realizations which is necessary for a comparison of different scheduling policies.

The BS acquires channel knowledge by a correlation based measurement of all user channels. Furthermore the average rate requirements of all users are assumed to be the same and known to the BS for convenience. The packet arrival rates are assumed to be the same and fixed for simplicity. After initialization the BS can now easily keep track of all user queues and consider them together with the actual channel realization for the scheduling policy. The actual computation of the MMSE or MMSE-SIC receive filter and the belonging bit and power allocation is performed by a standard DSP. Finally, the BS is signalling the power and bit allocation to each user over a feed-back link. Since the allocated power and modulation can be quantized with only a few information bits this feed-back does not require much bandwidth and can be transmitted together with the general MAC signalling which is needed in any coordinated multi-user access scheme.

5. MEASURED EXPERIMENTAL RESULTS

The performance of several scheduling policies with adaptive bit-loading was measured and evaluated with regard to sum throughput, delay (queueing state) wih certain QoS (BER and average rate) targets of the individual users. The real-time data transmission was performed with up to 5 MSym/s and up to 64-QAM modulation.

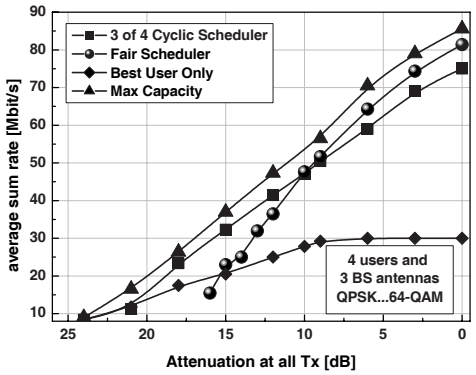

Fig. 3. Measure average sum rate with MU scheduling.

Fig. 3 shows the achievable average sum rate in a typical office scenario. The 3 of 4 cyclic scheduler is outperformed by the fair scheduler and the max. capacity approach in the high SNR region. With decreasing SNR the fair scheduler degrades below the cyclic scheduler since the sum rate is here dominated by the user which has the worst average channel. We clearly see that spatial multiplexing is mandatory with a multi-antenna BS otherwise 60% of the achievable throughput are lost, see lower cut-off rate for the best user only scheme at about 30 Mbits/s.

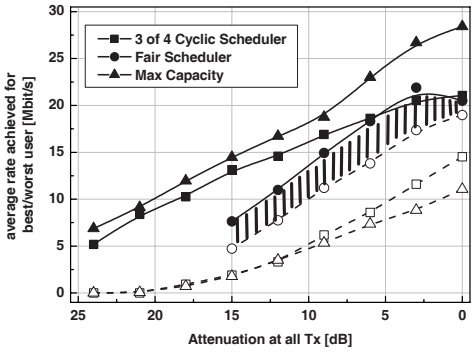

Fig. 4. Average individual user rates with different schedulers. filled symbols: best rate user, open symbols: lowest rate user.

Fig. 4 shows the possible average throughput per user. Here, over the whole SNR range the fair scheduler achieves the highest minimum average rate per user. This minimum rate, at least can be assured (open circles, lower bound of the shaded area) to all users.

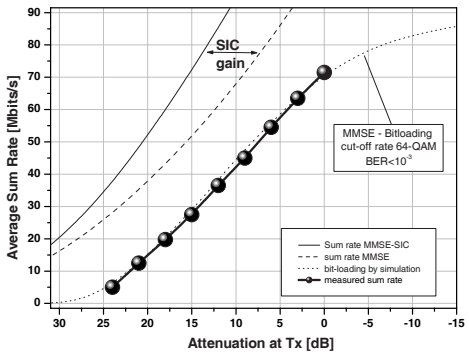

Fig. 5. Simulated and measured (•) sum throughput with and without bit-loading and linear/non-linear MMSE detector.

Fig. 5 depicts the comparison of the sum throughput achieved in the experiment (circles) with the expected throughput on the measured channel along the 5m trek in the lab. We clearly see the rate potential obtained with SIC due to the entanglement of the channel vectors. The measured throughput in Fig. 6 with a recently implemented MMSE-VBLAST detector [6] confirm that a substantial rate and SNR gain can be obtained from SIC detection.

The spatial multiplexing gain [7] which is expected to be 3 with the 3 BS antennas is not found in the experiment which is due to the fact that before full spatial multiplexing can be exploited the sum rate is cut-off due to the limited level of the QAM modulation. This limitation is to be seen in all sum rate plots. Therefore the measurement results coincide very well with what theory predicts to be achievable with real application constraints.

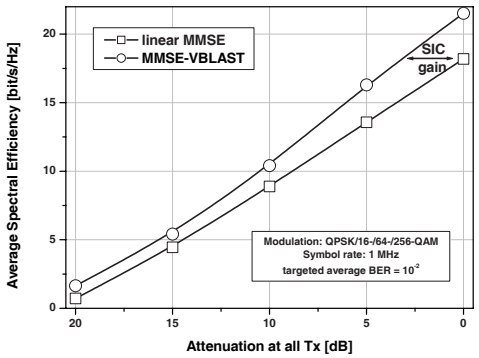

Fig. 6. Measured average spectral efficiency achieved with 4 users and a BS with 5 antennas and linear / non-linear MMSE detection.

Furthermore SIC stabilizes the transmission reliability to be seen in the empirical cdfs depicted in Fig. 7. The cdf curves become steeper and the tails at low rates reduce dramatically which would be reflected in higher outage sum rates e.g. 8 Mbits/s with linear MMSE and 16 Mbits/s with MMSE-VBLAST for 1% outage. In Fig. 7 no multi user diversity was exploited yet, therefore a further improvement towards narrower sum rate variations can be expected when many users are included within a scheduling policy.

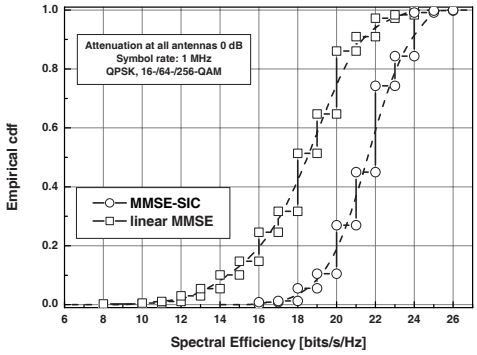

Fig. 7. CDF of the measured spectral efficiency.

6. CONCLUSIONS

We showed that by exploiting appropriate tools for the cross layer optimization the computational complexity can be reduced dramatically and real-time implementations of rate and QoS aware multi-user scheduling policies are feasible on state of the art hardware. Our measurements showed a significant increase of the spectral efficiency with a multi-antenna BS and a first implementation of a fair scheduling scheme showed that it allows the operator to selectively control the data queues at the users such that higher individual rates can be guaranteed for all users as a QoS, which is a significant improvement compared to many best effort schemes used today. A further spectral efficiency increase is expected from the usage of non-linear detection schemes at the BS as indicated by recent measurements with MMSE-VBLAST.

7. REFERENCES

[1] G.J. Foschini and M.J. Gans, "On the Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas," *Wireless Pers. Commun.*, p. pp. 315ff, 1998.

[2] E. Telatar, "Capacity of multi-antenna Gaussian channels," Tech. Rep., AT & T Bell Labs Internal Technical Memorandum, June 1995.

[3] H. Boche and M. Wiczanowski, "Optimal Scheduling for High Speed Uplink Packet Access - A Cross-Layer Approach," in *IEEE VTC Spring, Milano, Italy*, May 2004.

[4] H. Boche and M. Wiczanowski, "Stability Region of Arrival Rates and Optimal Scheduling for MIMO-MAC - A Cross-Layer Approach," in *Proc. IEEE International Zurich Seminar, Zurich*, Feb. 2002.

[5] T. Haustein, A. Forck, H. Gäbler, C.v. Helmolt, V. Jungnickel, and U. Krueger, "Real-Time MIMO Transmission Experiments with Adaptive Bitloading," in *IASTED WOC, Banff, Canada*, July 8-10th 2004.

[6] T. Haustein, A. Forck, H. Gäbler, and S. Schiffermüller, "From Theory to Practice: MIMO Real-Time Experiments of Adaptive Bit-loading with linear and non-linear Transmission and Detection Schemes.," in *IEEE VTC Spring, Stockholm, Sweden.*, May. 2005.

[7] L. Zheng and D.N.C. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple Antenna Channels," *IEEE Trans. on Information Theory*, vol. 49, no. 5, May 2003.

[8] E. Jorswieck H. Boche and T. Haustein, "Channel Aware Scheduling for Multiple Antenna Multiple Access Channels," in *Asilomar 2003, Stanford, USA.*, Nov 2003.