2004 ITG Workshop on Smart Antennas

Cross Layer Optimization – An Equivalence Class Approach

Michel T. Ivrlač Josef A. Nossek

Institute for Circuit Theory and Signal Processing
Munich University of Technology
80290 Munich, Germany
phone: +49 89 289285 {11,02}
email: {ivrlac, nossek}@nws.ei.tum.de

Abstract— **A cross layer design method is presented which is built upon the idea of equivalence classes of key-parameters of several layers in the protocol stack. An equivalence class is composed of all key-parameter tuples which fulfill a desired quality of service. Since different parameter tuples are usually associated with different costs (e.g. transmit power), a cross-layer design can select the most cost-efficient parameter tuple.**

I. Introduction

The concept of cross layer design in mobile communication networks introduces inter-layer coupling in the protocol stack and allows for a joint optimization of key parameters of several layers. This is especially important in mobile wireless networks due to the highly time-varying and fading nature of the wireless link. Combined with demanding quality of service (QoS) requirements of multi-media applications this challenge is hard to meet with a strictly layered protocol stack. It may therefore be promising to introduce a cross-layer design, in which (limited) exchange of information takes place between the involved layers. In this paper such a design method is presented, which is build upon the idea of equivalence classes of key-parameters of several layers in the protocol stack [1]. An equivalence class is composed of all key-parameter tuples which fulfill a desired quality of service. Since different parameter tuples are usually associated with different costs (e.g. transmit power), a cross-layer design can select the most cost-efficient parameter tuple.

II. Cross-Layer Optimization

Let us first develop a formal model of the cross-layer optimization we are aiming at. The formalism essentially consists of a few sets of tuples, a relation between the sets and a cost function. Let us pack all parameters of all layers that shall be optimized (or tuned) by cross-layer optimization into tuples p_i and collect all possible tuples into the set

$$\mathcal{P} = \{p_1, p_2, \ldots\}.$$

Elements of the tuples p_i can be for instance type of modulation, type of signal processing and its parameters, type of channel-code, code-rate, frame-error probability, packet-size, interleaver-length, buffer-size, air-time, SDMA-grouping, type of packet scheduling and its parameters, application-specific parameters (e.g. source data rate for a streaming media application). It will prove convenient to write the set $\mathcal{P}$ as the Cartesian product of two sets

$$\mathcal{P} = \mathcal{P}^{\mathrm{OP}} \times \mathcal{P}^{\mathrm{OM}}, \qquad (1)$$

where $\mathcal{P}^{\mathrm{OP}}$ contains parameter tuples which are influencing QoS, while the set $\mathcal{P}^{\mathrm{OM}}$ contains the remaining parameter tuples which do *not* influence QoS. For reasons, which will become clearer later, the sets $\mathcal{P}^{\mathrm{OP}}$ and $\mathcal{P}^{\mathrm{OM}}$ are called the set of operating *points* and the set of operating *modes*, respectively. Frequently, data-rate and error-ratio will be part of tuples contained in the set $\mathcal{P}^{\mathrm{OP}}$ and will specify an operating *point*. On the other hand, type of modulation, air-time, packet-size, interleaver-length, type of signal-processing will often be part of tuples contained in the set $\mathcal{P}^{\mathrm{OM}}$, and specify an operating *mode*. Let furthermore the set

$$\mathcal{Q} = \{q_1, q_2, \ldots\}$$

contain all possible tuples q_i, which specify the quality of service. At the application layer this may be for instance the image distortion and flow continuity in a video stream. At the transport layer it may be the minimum required data throughput and the maximum allowed delay time. The mapping from $\mathcal{Q}$ onto $\mathcal{P}^{\mathrm{OP}}$ is defined as an injective[1] relation

$$\mathrm{Q}_f \subset \mathcal{Q} \times \mathcal{P}^{\mathrm{OP}}. \qquad (2)$$

The subscript f in (2) indicates that the relation Q depends on an additional parameter tuple $f \in \mathcal{F}$, where $\mathcal{F}$ is the set of all parameter tuples of a communication system, which can (or should) *not* be altered by cross-layer optimization, but have influence on the mapping (2). These parameter tuples are called *side-effects*.

For a specific quality of service tuple $q \in \mathcal{Q}$ and a given side-effect $f \in \mathcal{F}$ the set

$$\Gamma_{q,f}^{\mathrm{OP}} = \{p \mid \forall p \in \mathcal{P}^{\mathrm{OP}} : (q,p) \in \mathrm{Q}_f\} \qquad (3)$$

contains all operating points, which lead to the same quality of service q given a side-effect f. Hence, $\Gamma_{q,f}^{\mathrm{OP}} \subseteq \mathcal{P}^{\mathrm{OP}}$ is an *equivalence class* in $\mathcal{P}^{\mathrm{OP}}$. For $|\Gamma_{q,f}^{\mathrm{OP}}| > 1$, we obtain some degrees of freedom for the cross-layer optimization, since we are *free to choose* any particular element of $\Gamma_{q,f}^{\mathrm{OP}}$ as our operating point. To determine this choice we can associate each element of $\mathcal{P}^{\mathrm{OP}}$ with a cost[2] by the cost function

$$\mathrm{cost} : \mathcal{P}^{\mathrm{OP}} \times \mathcal{P}^{\mathrm{OM}} \times \mathcal{F} \to \mathcal{R} \cup \{\infty\}, \qquad (4)$$

[1] Different $q \in \mathcal{Q}$ are mapped onto different $p \in \mathcal{P}^{\mathrm{OP}}$.

[2] In a cellular communication system, a cost may be expressed e.g. in terms of transmit power required to realize a given operating point with a given operating mode and side-effect.

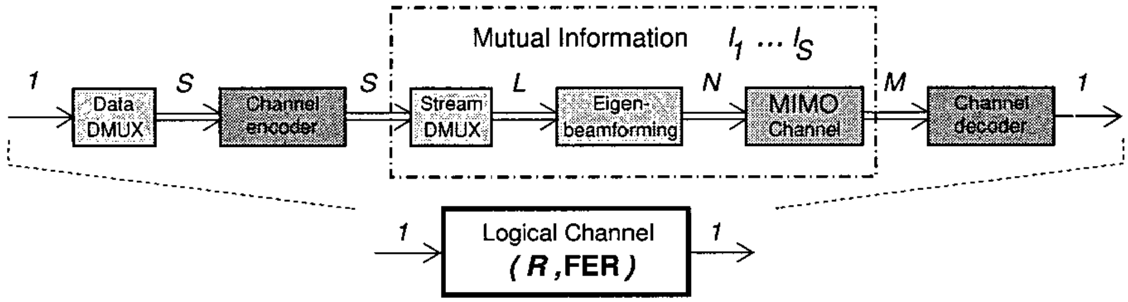

Fig. 1. Air-interface of a single-link MIMO system using Eigenbeamforming (long-term transmitter channel state information).

which depends also on operating mode and side-effects. Here, $\mathcal{R}$ represents the set of real numbers. We assign infinite costs to parameter combinations which cannot be implemented (e.g. a data-rate which is larger than permitted by the modulation scheme).

For a given operating point $p \in \mathcal{P}^{\mathrm{OP}}$ and a side-effect $f \in \mathcal{F}$ we define the set of *optimum operating modes* as

$$\mathcal{P}_{p,f}^{\mathrm{opt-OM}} = \{\, \tilde{p} \,|\, \forall\, \tilde{p} \in \mathcal{P}^{\mathrm{OM}}, \forall \tilde{p}\,' \in \mathcal{P}^{\mathrm{OM}} :\ \mathrm{cost}(p,\tilde{p},f) \leq \mathrm{cost}(p,\tilde{p}\,',f)\} \quad (5)$$

Hence, $\mathcal{P}_{p,f}^{\mathrm{opt-OM}}$ contains those operating modes, which implement the given operating point with minimum cost. The process of cross-layer optimization is then reduced to the selection of the optimum operating point. The set of *optimum operating points* for a given quality of service $q \in \mathcal{Q}$ and side-effect $f \in \mathcal{F}$ is then given by

$$\mathcal{P}_{q,f}^{\mathrm{opt-OP}} = \{\, p \,|\, \forall p \in \Gamma_{q,f}^{\mathrm{OP}},\ \tilde{p} \in \mathcal{P}_{p,f}^{\mathrm{opt-OM}}, \forall p' \in \Gamma_{q,f}^{\mathrm{OP}}, \tilde{p}' \in \mathcal{P}_{p',f}^{\mathrm{opt-MO}} : \mathrm{cost}(p,\tilde{p},f) \leq \mathrm{cost}(p',\tilde{p}',f)\} \quad . \quad (6)$$

In case $|\mathcal{P}_{q,f}^{\mathrm{opt-OP}}| > 1$, the solution is not unique, and the set $\mathcal{P}_{q,f}^{\mathrm{opt-OP}} \subseteq \Gamma_{q,f}^{\mathrm{OP}}$ is another equivalence class in $\mathcal{P}^{\mathrm{OP}}$, which elements all lead to minimum costs for implementing quality of service $q \in \mathcal{Q}$ with side-effect $f \in \mathcal{F}$.

The crucial part in the cross-layer optimization specified in (6), (5) and (3) is actually not the problem of optimization itself, but rather the problem of *modelling*, which purpose is the development of the relations Q_f between quality of service and the adjustable parameters of the system on the one hand, and the development of the relationship between costs and operating points, operating modes and side-effects, on the other hand.

In the following sections we will give detailed examples on how this modelling can be applied for different applications and air-interfaces. Explicitly we will consider two different air-interfaces: a single-user MIMO system, and a multi-user MIMO system, both with long-term transmitter channel state information. The applications in mind will be video-streaming and file-transfer.

III. Single-Link Optimization

Here we aim at optimization of a single link only. Such a situation occurs in a multi-user communication system when the users are separated by other means, e.g. time (TDMA) or frequency (FDMA) with fixed time- or frequncy slot assignments. Fig. 1 shows an example air-interface of a single-link spanned over a noisy, frequency-flat fading MIMO channel operating with N transmit and M receive antennas. The sequence of information bits to be transmitted is first demultiplexed into $S \geq 1$ data streams which are then encoded independently for error protection, and then demultiplexed again onto $L \geq S$ streams which are transmitted over the L dominant long-term eigenmodes of the MIMO channel (MIMO-eigenbeamforming). The transmitter is aware of the channel on average only. Instantaneous channel state information or feedback is not available. For simplicity we ommit modulation schemes here, and assume a Gaussian codebook at the output of the channel encoders.

A. Operating modes and operating points

From a higher layer point of view, the whole air-interface can be described by a so-called *logical channel*, which is characterized by two parameters, namely the net data-rate R and the residual frame error probability FER at the output of the channel decoder at the receiver. These two parameters will clearly influence quality of service in a video-streaming or file-transfer application. Hence, we define the set of operating points

$$\mathcal{P}^{\mathrm{OP}} = \{(R, \mathrm{FER}) \mid R > 0,\ 0 < \mathrm{FER} < 1\}. \quad (7)$$

The number of data streams $S \geq 1$ and powered up eigenmodes $L \geq S$ on the other hand do not influence QoS directly, since any combination of R and FER can be implemented for any valid combination of S and L. For simplicity, let us restrict our self to $S \leq 2$ and $L \leq 2$. Hence, the set of operating modes is

$$\mathcal{P}^{\mathrm{OM}} = \{(S,L)\} = \{(1,1),(1,2),(2,2)\}. \quad (8)$$

In the mode $(1,1)$ one data-stream is transmitted over the dominant eigenmode. We will refer to this as the "antenna-gain"-

mode (ANT), since this mode of operation is aiming at maximizing signal to noise ratio at the receiver by utilizing antenna gain through beamforming. In the mode $(1,2)$ a single data-stream is transmitted over two eigenmodes in a round-robin fashion. This mode is called "diversity"-mode (DIV), as the time-switching between the two eigenmodes increases link-reliability by introducing diversity. One can think of it as interleaving the encoded data stream over the two dominant eigenmodes. Finally, the mode $(2,2)$ is called "multiplexing"-mode (MUX), since the two eigenmodes are used to transmit independent data streams, and therefore aiming at multiplexing gain. Both antenna-gain and diversity are sacrificed for having indepedent multi-stream transmission. Which of this operating modes is the best? To answer this question we have to establish a cost function.

B. The cost-function

In all cases we assume the receiver has complete and correct instantaneous channel knowledge (coherent signalling). In order to take into account the error control capabilities of the channel codes we will make the following abstraction: the channel codes are assumed to be capacity achieving, which means, that a decoding error occurs only, if the current data-rate is above the current mutual information of the channel including the eigenbeamforming (see Fig. 1). Hence, we have

$$\mathrm{FER} = \left\{ \begin{array}{ll} 1 - \Pr\{I_1 > R\} & \text{for } S = 1 \\ 1 - \Pr\{I_1 > R_1,\, I_2 > R_2\} & \text{for } S = 2 \end{array} \right\}, \tag{9}$$

where I_1 and I_2 are the mutual information for the first and the second data stream, respectively, and R_1 and R_2 are the net data rates for each of the data streams in the multiplexing mode. They have to be chosen in an optimum way, i.e.

$$R_1, R_2 = \arg\min_{R_1,R_2} \mathrm{FER} \;\; \text{s.t.} \; \left\{ \begin{array}{l} R_1 \geq 0 \\ R_2 \geq 0 \\ R_1 + R_2 = R \end{array} \right. . \tag{10}$$

In the operating mode "ANT" ($S = 1, L = 1$), the mutual information is given by

$$I_1^{(\mathrm{ANT})} = \log_2\left(1 + \frac{P_\mathrm{T}}{\sigma_n^2}\, \boldsymbol{t}_1^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{H} \boldsymbol{t}_1\right), \tag{11}$$

where $\boldsymbol{H} \in \mathcal{C}^{M\times N}$ is the MIMO channel matrix, $\boldsymbol{t}_1$ is the unity norm, dominant eigenvector of $\mathrm{E}[\boldsymbol{H}^\mathrm{H}\boldsymbol{H}]$, while P_T and σ_n^2 are the transmit power, and the receiver noise[3] power per antenna element, respectively. In the "DIV" mode ($S = 1, L = 2$), we obtain

$$I_1^{(\mathrm{DIV})} = \frac{1}{2} \sum_{i=1}^{2} \log_2\left(1 + \frac{P_\mathrm{T}}{\sigma_n^2}\, \boldsymbol{t}_i^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{H} \boldsymbol{t}_i\right) \tag{12}$$

since we switch between the two dominant eigenbeams. The "MUX"-mode ($S = 2, L = 2$) is the most complicated one.

[3] Spatially white Gaussian noise is assumed for simplicity.

We assume interference cancellation after decoding at the receiver. If the first stream is detected first we have

$$\begin{aligned} I_{1,1} &= \log_2\left(1 + \frac{P_1}{\sigma_n^2} \boldsymbol{t}_1^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{R}_1^{-1} \boldsymbol{H} \boldsymbol{t}_1\right) \\ I_{2,1} &= \log_2\left(1 + \frac{P_2}{\sigma_n^2} \boldsymbol{t}_2^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{H} \boldsymbol{t}_2\right) \end{aligned},$$

where $P_1 > 0$ and $P_2 > 0$ are the transmit powers assigned to each eigenmode ($P_1 + P_2 = P_T$), and

$$\boldsymbol{R}_1 = \frac{P_2}{\sigma_n^2} \boldsymbol{H} \boldsymbol{t}_2 \boldsymbol{t}_2^\mathrm{H} \boldsymbol{H}^\mathrm{H} + \mathbf{I}_M$$

is the normalized noise and interference correlation matrix for the first stream before interference cancellation. Here $\mathbf{I}_M$ is the $M \times M$ identity matrix. If the second stream is detected first we have

$$\begin{aligned} I_{2,2} &= \log_2\left(1 + \frac{P_2}{\sigma_n^2} \boldsymbol{t}_2^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{R}_2^{-1} \boldsymbol{H} \boldsymbol{t}_2\right) \\ I_{1,2} &= \log_2\left(1 + \frac{P_1}{\sigma_n^2} \boldsymbol{t}_1^\mathrm{H} \boldsymbol{H}^\mathrm{H} \boldsymbol{H} \boldsymbol{t}_1\right) \end{aligned},$$

where

$$\boldsymbol{R}_2 = \frac{P_1}{\sigma_n^2} \boldsymbol{H} \boldsymbol{t}_1 \boldsymbol{t}_1^\mathrm{H} \boldsymbol{H}^\mathrm{H} + \mathbf{I}_M$$

is the normalized noise and interference correlation matrix for the second stream before interference cancellation. Even though, since $I_{1,1} + I_{2,1} = I_{1,2} + I_{2,2}$ the order of detection is not important for the *sum* of mutual information, it is of importance in a fading channel, since the probability distributions are different. We assume the order is chosen such that

$$I_i^{(\mathrm{MUX})} = \left\{ \begin{array}{ll} I_{i,1} & \text{for } \min(I_{1,1}, I_{2,1}) \geq \min(I_{1,2}, I_{2,2}) \\ I_{i,2} & \text{else} \end{array} \right., \tag{13}$$

where $i = 1, 2$. The power distribution between the two eigenmodes (P_1 and P_2) is determined by the waterfilling algorithm based on the eigenvalues of $\mathrm{E}[\boldsymbol{H}^\mathrm{H}\boldsymbol{H}]$. In case the waterfilling solution is switching off one eigenmode, the powers are shared equally between the two eigenbeams instead.

The channel matrix $\boldsymbol{H}$ is treated as a random variable which makes the mutual information random variables, too. The cost function is now defined as the transmit power P_T which is necessary to achieve a given frame error ratio FER at a given net data-rate R,

$$\mathrm{cost}\left((R, \mathrm{FER}), \mathrm{mode}, \text{side-effect}\right) = P_\mathrm{T}/\sigma_n^2 \tag{14}$$

where mode can be "ANT" ($S = 1, L = 1$), "DIV" ($S = 1, L = 2$), or "MUX" ($S = 2, L = 2$), and the side-effect is the probability distribution of the channel matrix $\boldsymbol{H}$. The cost function can be computed numerically from (9), (10), (11), (12) and (13).

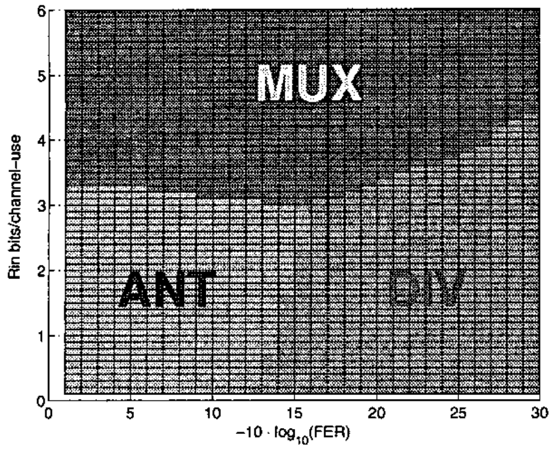

Fig. 2. COM-Chart: Optimum operating modes ("ANT", "DIV", "MUX") as functions of the operating point (R, FER). The FER is shown logarithmically and is decreasing from left to right.

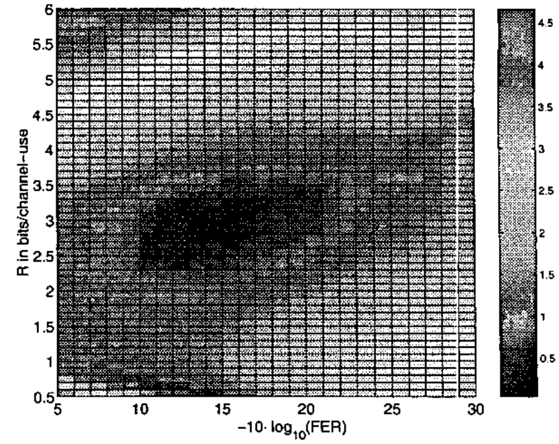

Fig. 3. Worst-case excess power ratio in dB. The chart shows what increase in transmit power is necessary when the worst operating mode is selected instead of the best one.

C. Optimum Operating Mode

For a given operating point (R, FER) and side-effect (distribution of $\boldsymbol{H}$) the optimum operating mode can now be computed by applying (5). For a MIMO system with $N = 8$ transmit and $M = 4$ receive antennas[4], with semi-correlated Rayleigh-fading channel with one departing path with 20° angle-spread departing from bore-side, the result is shown in Fig. 2. The qualitative result is what we would expect intuitively. The diversity mode becomes optimal if high link reliability, i.e. low values of FER are required, while the multiplexing mode is favourable for high data-rates. For moderate data-rates and error ratios one should go for antenna-gain. The quantitative results show the border-lines, i.e. the optimum trade-off between those three modes of operation. It should be emphasized, that the quantitative results are depending on the distribution of $\boldsymbol{H}$, which is dependent on the correlation properties of the channel. The chart in Fig. 2 is called "COM-Chart", as it essentially displays *c*apacity, *o*utage and optimal operating *m*ode. If the operating mode is *not* selected in the optimum way, we have to live with an increase in transmit power. The amount of this increase is shown in Fig. 3 in the worst-case, i.e. when selecting the worst instead of the best operating mode. The range of increase of power is substantial.

IV. Multi-Link Optimization

Let us have a look at the case of optimizing several links simultaniously. This happens in a multi-user communication system when the user separation (e.g. in time or frequency) is adaptive, i.e. the number of time- or frequency slots can be set differently for each link. It also happens in case the transmission for both users is done at the same time in the same frequency band, and the separation (with or without interference) is done in the space domain (SDMA). Fig. 4 shows such a case for two links. The two receivers are equipped with $M_1 \geq 1$ and $M_2 \geq 1$ antennas, the transmitter can use N antennas. The transmitter may serve the users with differernt multiple access schemes: either by TDMA (separation in time) or by SDMA (separation in space). In the former case the relative air-time α_i can be set for each link. Furthermore, the number of eigenmodes L_1 and L_2 can be set individually for each link.

[4]A uniform linear antanna array with half-wavelength spacing is assumed.

A. Operating modes and operating points

Similar to the single-link case discussed in Section III-A the net data-rate R and the frame error ratio FER are forming an operating point. Since we now have a multi-link situation, we need a pair (R, FER) for each link:

$$\begin{aligned}\mathcal{P}^{\text{OP}} &= \{(R_1, R_2, \text{FER}_1, \text{FER}_2) \mid \\ &R_1, R_2 > 0,\ 0 < \text{FER}_1, \text{FER}_2 < 1\}\end{aligned}, \qquad (15)$$

The operating modes contain the choice about the multiple access scheme (TDMA/SDMA), the relative air-time α_i for each link, the number of eigenmodes (L_1, L_2) for each link and may in addition contain modulation schemes, the number of data-streams which leads to "ANT", "DIV" and "MUX" mode as in the single-link case. Also it may contain the choice of signal processing algorithm used for spatial separation in SDMA mode. For simplicity, we will here however assume that the algorithm for SDMA is fixed and furthermore only one data-stream and eigenmode is used ($L_1 = L_2 = 1$). We will also ommit modulation schemes (Gaussian codebook). In this way, we have a simple set of operating modes

$$\mathcal{P}^{\text{OM}} = \{\text{"TDMA"}, \text{"SDMA"}\} \times \{(\alpha_1, \alpha_2)|, 0 < \alpha_1, \alpha_2 \leq 1\} \qquad (16)$$

which offers only the choice between the two multiple access schemes and the relative air-time.

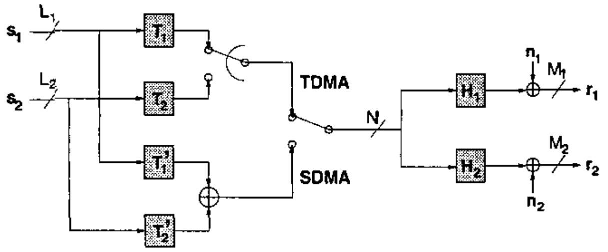

Fig. 4. Air-interface of a multi-link (here two links) MISO or MIMO system using TDMA and SDMA.

B. The cost-function

Since the TDMA mode is adaptive by allowing different air-time for each link, the channel data-rate and the net data-rate are different and related by

$$R_i = R_{c,i} \cdot \alpha_i, \text{ or } R_{c,i} = R_i/\alpha_i, \tag{17}$$

where $R_{c,i}$ is the data-rate when the $i-$th link is activated. For TDMA all α_i must sum up to unity. To keep the same notation also when using SDMA, we let all α_i be equal to unity in this case.

We assume the receiver has complete and correct instantaneous channel knowledge (coherent signalling). As in the single-link case we take the effects of channel coding into account by using the abstraction of capacity achieving channel codes. The frame error ratios are therefore given as

$$\text{FER}_i = \Pr\{I_i \leq R_{c,i}\} = \Pr\{I_i \leq R_i/\alpha_i\} \text{ for } i = 1,2. \tag{18}$$

As we are using only one eigenmode, the mutual information for the two links in the system from Fig. 4 is given for the TDMA mode as

$$\begin{aligned} I_1^{(\text{TDMA})} &= \log_2\left(1 + \frac{P_1}{\sigma_{n,1}^2}\, \boldsymbol{t}_1^{\text{H}} \boldsymbol{H}_1^{\text{H}} \boldsymbol{H}_1 \boldsymbol{t}_1\right) \\ I_2^{(\text{TDMA})} &= \log_2\left(1 + \frac{P_2}{\sigma_{n,2}^2}\, \boldsymbol{t}_2^{\text{H}} \boldsymbol{H}_2^{\text{H}} \boldsymbol{H}_2 \boldsymbol{t}_2\right) \end{aligned}, \tag{19}$$

where P_1 and P_2 are the transmit power for each user and $\sigma_{n,1}^2$ and $\sigma_{n,2}^2$ represent the receiver noise[5] power per antenna. The vectors $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ are the eigenvectors of $\text{E}[\boldsymbol{H}_1^{\text{H}}\boldsymbol{H}_1]$ and $\text{E}[\boldsymbol{H}_2^{\text{H}}\boldsymbol{H}_2]$, respectively, which correspond to the largest eigenvalue. Note, that the users are separated interference-free. In SDMA mode, on the other hand the two users are not separated free of interference, since the transmitter is aware of the channel only on average. The mutual information therefore becomes

$$\begin{aligned} I_1^{(\text{SDMA})} &= \log_2\left(1 + \frac{P_1}{\sigma_{n_1}^2} \boldsymbol{t}_1'^{\text{H}} \boldsymbol{H}_1^{\text{H}} \boldsymbol{R}_1^{-1} \boldsymbol{H}_1 \boldsymbol{t}_1'\right) \\ I_2^{(\text{SDMA})} &= \log_2\left(1 + \frac{P_2}{\sigma_{n_2}^2} \boldsymbol{t}_2'^{\text{H}} \boldsymbol{H}_2^{\text{H}} \boldsymbol{R}_2^{-1} \boldsymbol{H}_2 \boldsymbol{t}_2'\right) \end{aligned}, \tag{20}$$

[5] For simplicity spatially white Gaussian noise is assumed-

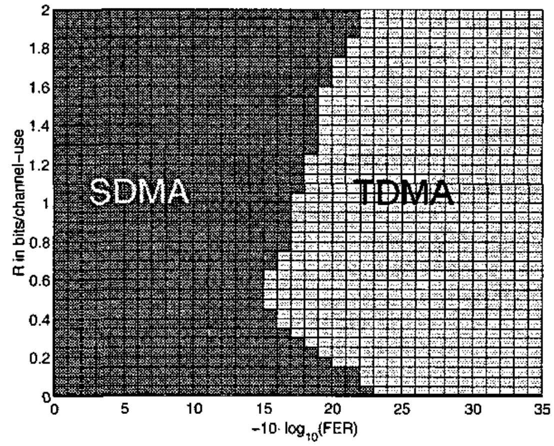

Fig. 5. Optimum operating modes ("SDMA", "TDMA") as functions of the operating point (R, FER). The FER is shown logarithmically and is decreasing from left to right.

with

$$\begin{aligned} \boldsymbol{R}_1 &= \frac{P_2}{\sigma_{n,1}^2} \boldsymbol{H}_1 \boldsymbol{t}_2' \boldsymbol{t}_2'^{\text{H}} \boldsymbol{H}_1^{\text{H}} + \mathbf{I}_{M_1} \\ \boldsymbol{R}_2 &= \frac{P_1}{\sigma_{n,2}^2} \boldsymbol{H}_2 \boldsymbol{t}_1' \boldsymbol{t}_1'^{\text{H}} \boldsymbol{H}_2^{\text{H}} + \mathbf{I}_{M_2} \end{aligned}. \tag{21}$$

The vectors $\boldsymbol{t}_1'$ and $\boldsymbol{t}_2'$ are determined by a suitable signal-processing algorithm for SDMA with long-term channel state information (CSI). Here we use the long-term CSI version of the block-zero-forcing algorithm [2], which in the special case of $L_1 = L_2 = 1$ becomes

$$\begin{aligned} \boldsymbol{t}_1' &= \text{dom.EV}\{\,\text{E}[\,\text{Proj}\{\,\text{null}\,\boldsymbol{H}_2\,\}\,]\,\} \\ \boldsymbol{t}_2' &= \text{dom.EV}\{\,\text{E}[\,\text{Proj}\{\,\text{null}\,\boldsymbol{H}_1\,\}\,]\,\} \end{aligned}. \tag{22}$$

Here dom.EV{.}, Proj{.} and null refer to the dominant eigenvector, the orthogonal projection matrix and the null-space, respectively.

The channel matrix $\boldsymbol{H}$ is treated as a random variable which makes the mutual information random variables, too. The cost function is now defined as the transmit power

$$P_{\text{T}} = \alpha_1 \cdot P_1 + \alpha_2 \cdot P_2 \tag{23}$$

which is neccessary to achieve a given combination of frame error ratio and net data-rate for the two users:

$$\text{cost}\left((R_1, R_2, \text{FER}_1, \text{FER}_2), \text{mode}, \text{side-effect}\right) = P_{\text{T}}. \tag{24}$$

Here "mode" is a combination of the multiple access scheme (TDMA/SDMA) and the relative air-times. The side-effect is the probability distribution of the channel matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ and the receiver noise powers[6] and $\sigma_{n,2}^2$. The cost function can be computed numerically from (18), (19) and (20).

[6] Regarding the optimization one acutally needs only the *ratio* of the receiver

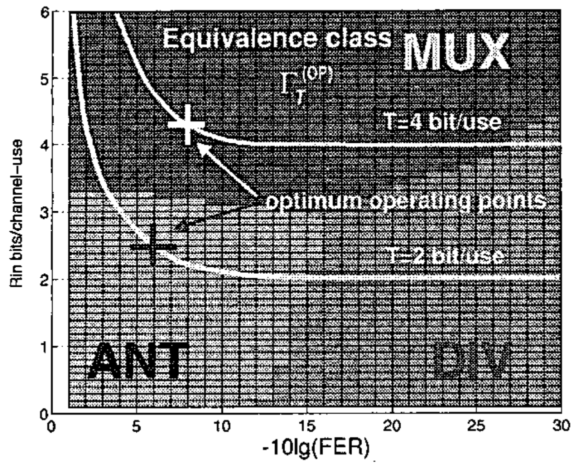

Fig. 6. COM-Chart with Γ_T^{OP} equivalence classes. The optimum operating points are determined from the COP chart in Fig. 7

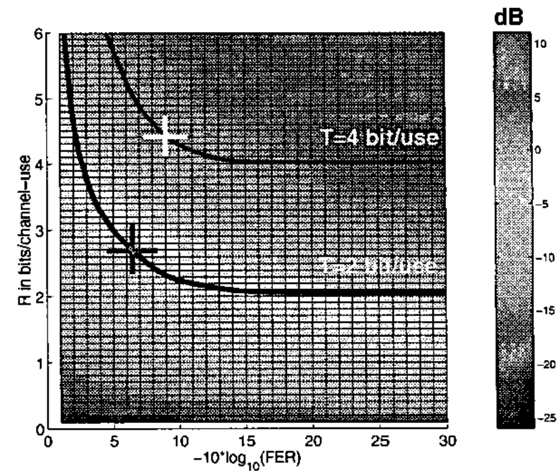

Fig. 7. COP-Chart with Γ_T^{OP} equivalence classes. This is a visualization of the cost-function. The colors represent transmit power neccessary for different operating points. The optimum points are determined by the minimum costs within Γ_T^{OP}.

C. Optimum Operating Mode

For a given operating point $(R_1, R_2, \mathrm{FER}_1, \mathrm{FER}_1)$ and side-effect (distribution of $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$) the optimum operating mode can now be computed by applying (5). Since the tuples forming an operating point are now four-dimensional, we cannot display the result in a two dimensional chart as in the previous example of the single-link case. However, we can draw a cross-cut through the four-dimensional operating point space, by looking at the special case $R_1 = R_2 = R$, $\mathrm{FER}_1 = \mathrm{FER}_2 = \mathrm{FER}$. For a 2-user MIMO system with $N = 6$ transmit and $M_1 = M_2 = 2$ receive antennas[7], the result is shown in Fig. 5. The MIMO channels are semi-correlated Rayleigh-fading with one departing path with 28° angle-spread departing from an azimuthal angle of 50° for the first user and −50° for the second. The TDMA mode is preferable when high link reliability is requiered, while for moderate link reliability SDMA takes the lead. For very high data-rates TDMA also becomes the favourite, because the long-term SDMA mode is interference limited. In this example, the two channels are however well separable, so that the interference limitation occurs at rather large data-rate values which are not displayed in Fig. 5.

V. Mapping QoS on Operating Points

In the previous two section we have seen two examples for the definitions of the parameter sets of operating points and operating modes, the cost-function and the optimization of the operating mode. In this section we will now look at the problem of mapping QoS requirements onto the set of operating points. We will consider two applications: file transfer and streaming video.

A. File transfer

Let us start with this simple application which is aiming at transfering large data files. There are no delay or continuity restrictions. However the file has to be received absolutely error-free. To ensure this requirement, an automatic repeat request protocol (ARQ) is set up. Frames that cannot be decoded correctly are retransmitted until they are decoded correctly. The quality of service is measured in the average data-rate, which includes the effect of the retransmissions due to the ARQ protocol. This average data-rate, is called throughput and can be computed as

$$T = R \cdot (1 - \mathrm{FER}), \tag{25}$$

since the instantaneous data-rate is equal to R with probability $(1 - \mathrm{FER})$ and zero with probability FER. In the single-link case we can therefore write the quality of service relation (2)

$$\begin{aligned} \mathrm{Q}_f = \mathrm{Q} \; &= \{(T, (R, \mathrm{FER})) \mid \\ & R \cdot (1 - \mathrm{FER}) = T,\ R > 0,\ 0 < \mathrm{FER} < 1\} \end{aligned} \tag{26}$$

In this case the relation does not depend on side-effects. The equivalence class from (3) is then given by

$$\Gamma_T^{\mathrm{OP}} = \{(R, \mathrm{FER}) \in \mathcal{P}^{\mathrm{OP}} \mid R \cdot (1 - \mathrm{FER}) = T\}, \tag{27}$$

which is shown in graphical form in the COM-Chart in Fig. 6. The final optimization step now simply consists of picking the operating point out of Γ_T^{OP} which minimizes the cost-function. The so called COP-Chart[8] in Fig. 7 shows the cost-function

noise powers, i.e. $\sigma_{n,1}^2/\sigma_{n,2}^2$, since the cost-function can be scaled by any real, non-zero number. By using $1/\sigma_{n,1}^2$ as the scaling factor, the cost-function depends on the *ratio* $\sigma_{n,1}^2/\sigma_{n,2}^2$ instead of their values.

[7]A Uniform linear antenna arrayw with half-wavelength spacing is assumed.

[8]COP stands for: Capacity-outage-power

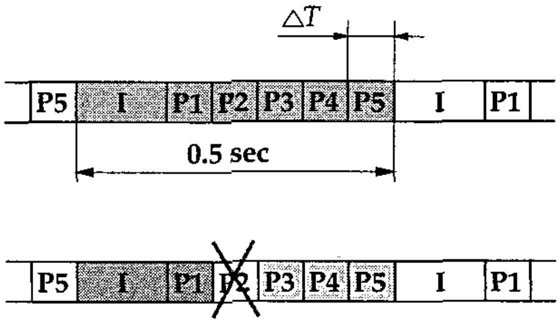

Fig. 8. *Upper part:* Considered frame structure of streaming video. One I-frame and $(F-1)$ P-frames (here $F = 6$) form a group of pictures (GOP). The duration of a GOP is 0.5 sec, independent of F. *Lower part:* error-event S_2: a first time frame loss at the third frame

and indicates the optimum operating points. By transfering these operating points into the COM-Chart of Fig. 6 we can also see which operating mode is optimum.

B. Streaming Video

In a streaming video application, a commonly used measure of quality is the mean squared error (MSE) between the original frame and the decoded frame. This MSE is influenced by two issues, which are 1) the source-distortion due to lossy compression at the encoder, and 2) distortion due to transmission errors [3]. The source-distortion $\mathrm{MSE}_{\mathrm{SD}}(R_s)$ depends on the source-encoder and is essentially a monotonic decreasing function of the source rate R_s, which is the rate at which the compressed video is emitted from the source-encoder. The source-distortion expressed in dB as the peak signal to noise ratio (PSNR) can be modelled with

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{255^2}{\mathrm{MSE}_{\mathrm{SD}}} = a - b\sqrt{\frac{R_s}{c}}\left(\frac{c}{R_s} - 1\right), \quad (28)$$

where the constants a, b and c are obtained by fitting to measurement data[9]. The distortion due to transmission errors $\mathrm{MSE}_{\mathrm{TE}}$ depends on the error characteristics, e.g. their probability and burstiness, and the error-concealing properties of the decoder. Fig. 8 shows a typical frame structure of a video stream. A so called group of pictures (GOP) consists of an anchor frame (I-frame) followed by $(F-1)$ P-frames. The I-frame uses space-only compression (JPEG), while the P-frames are result of a space-time compression by motion compensated predictive coding. In this example, the I-frame is twice as large as the P-frames.[10] Successful decoding of a P-frame requires that all previous P-frames and the I-frame have been decoded successfully. This error-propagation property makes the video stream sensitive to frame losses. To mitigate the effect of frame losses, the receiver usually performs some

[9] In the example we have used $a = 25$, $b = 3.7$ and $c = 1$bit/channel-use.

[10] For simplification we also assume that all P-frames have the same size. The case of variable size frames is also tractable, however is a little bit more involved.

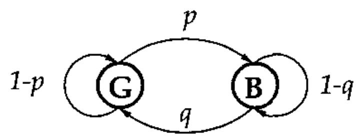

Fig. 9. Gilber-Elliot model

kind of error concealment. In this example we will assume a very simple type of error coinceilment, which just copies the last correctly decoded frame into the remaining frames, i.e. "freezes" the video stream until a new I-frame is received correctly. In order to quantify the distortion $\mathrm{MSE}_{\mathrm{TE}}$ we define the error event S_j when the first transmission error occurs at the j-th frame (numbering from 0). Because of the "freezing" error concealment, the receiver does not care about transmission errors following the first one in a GOP. The $\mathrm{MSE}_{\mathrm{TE}}$ can then be written as

$$\mathrm{MSE}_{\mathrm{TE}} = \sum_{j=0}^{F-1} \mathrm{MSE}_j \cdot \Pr\{S_j\}, \quad (29)$$

where MSE_j is the mean square error that remains after "freezing" the video stream at the j-th frame in the GOP. The set of values $\mathcal{D} = \{\mathrm{MSE}_0, \mathrm{MSE}_1, \ldots, \mathrm{MSE}_{F-1}\}$ is called a *distortion profile* [4]. Distortion profiles depend on the current scene in the video stream and have to be measured. In this example we use the following distortion profile:

$$\mathcal{D} = \{250, 197, 184, 171, 158, \ldots, 15\} \quad (30)$$

for a GOP of $F = 16$ frames. It is common to model the burstiness of transmission errors by a two-state Markov model (Gilbert-Elliot model) shown in Fig. 9. Transmission errors only occur, when the channel is in "bad" state. The transition probabilities p and q control the length and frequency of error bursts. The probability of the error event S_j from (29) can now be written as

$$\Pr\{S_j\} = \frac{p}{p+q}\begin{cases} 1+q & \text{for } j = 0 \\ q \cdot (1-p)^j & \text{for } j \geq 1 \end{cases}. \quad (31)$$

In a block-fading wirless channel, the transition probabilities p and q can be related to frame-error ratio FER and channel coherence time T_{coh} [5]:

$$\begin{aligned} p &= \mathrm{FER} \cdot \Delta T / T_{\mathrm{coh}} \\ q &= (1-\mathrm{FER}) \cdot \Delta T / T_{\mathrm{coh}} \end{aligned}, \quad (32)$$

where ΔT is the time needed to transmit one P-frame. For a given distortion profile we have $\mathrm{MSE}_{\mathrm{TE}}(\mathrm{FER}, T_{\mathrm{coh}})$ as function of frame-error probability and coherence time of the wireless channel. The total distortion

$$\mathrm{MSE} = \mathrm{MSE}_{\mathrm{SD}} + \beta \cdot \mathrm{MSE}_{\mathrm{TE}} \quad (33)$$

is modelled [3] as the weigthed sum of the two individual distortions, where β is the weighting factor. For fixed β and given coherence time T_{coh} the total distortion

$$\mathrm{MSE} = \mathrm{MSE}(R_s, \mathrm{FER}), \quad (34)$$

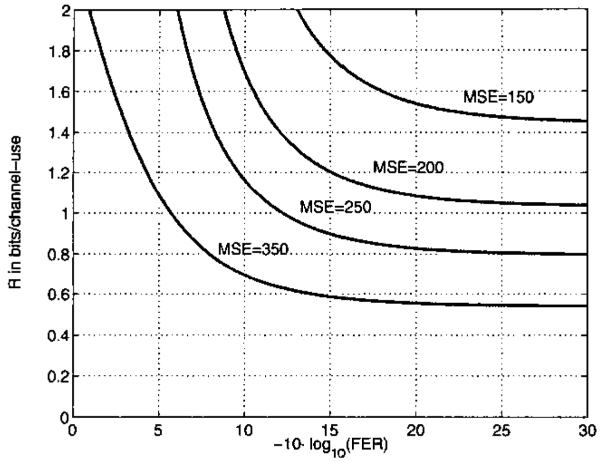

Fig. 10. Equivalence class $\Gamma^{\mathrm{OP}}_{(\mathrm{MSE}_1,\mathrm{MSE}_2),f}$ displayed for the first link, i.e. tuples $(R_1 = R, \mathrm{FER}_1 = \mathrm{FER})$ are shown as function of $\mathrm{MSE}_1 = \mathrm{MSE}$

is a function of source-rate and frame-error probability and can be computed by application of (28), (29), (31) and (32). If the source-rate R_s is equal to the net data rate R, i.e.

$$R_s = R, \tag{35}$$

we obtain the total distortion MSE as function of R and FER. In a two-link scenario we can therefore write the quality of service relation (2) as

$$\begin{aligned} Q_f = \ & \{((\mathrm{MSE}_1, \mathrm{MSE}_2), (R_1, R_2, \mathrm{FER}_1, \mathrm{FER}_2)) \mid \\ & \mathrm{MSE}(R_i, \mathrm{FER}_i, f) = \mathrm{MSE}_i, \ i = 1, 2\} \end{aligned}, \tag{36}$$

The side-effect $f \in \mathcal{F}$ consists of the coherence time T_{coh}, the distortion profile $\mathcal{D}$ and the probability distributions of the MIMO channels. The equivalence class from (3) then becomes

$$\begin{aligned} \Gamma^{\mathrm{OP}}_{(\mathrm{MSE}_1,\mathrm{MSE}_2),f} & = \{(R_1, R_2, \mathrm{FER}_1, \mathrm{FER}_2) \in \mathcal{P}^{\mathrm{OP}} \mid \\ & \mathrm{MSE}(R_i, \mathrm{FER}_i, f) = \mathrm{MSE}_i, \ i = 1, 2\} \end{aligned}, \tag{37}$$

which is shown for the first user in Fig. 10.

To obtain the optimum operating point we have to pick the element from $\Gamma^{\mathrm{OP}}_{(\mathrm{MSE}_1,\mathrm{MSE}_2),f}$ which has the lowest associated costs. For a coherence time $T_{\mathrm{coh}} = 0.1$ sec a sample result is shown in Fig 11. The TDMA-mode is favourable in case of either very high requested quality (i.e. low MSE) or large quality inbalance (large difference in MSE of the users). For very high requested quality the link has to have high reliability, i.e. low FER. By looking at the COM-Chart in Fig. 5 it becomes clear that TDMA will get the lead in this situation. In case of high quality inbalance it does not pay off to separate the users in space. Assigning the user with the higher quality requirements a longer air-time needs less transmit power. On the other hand, SDMA is favourable to TDMA in situations of similar quality requirements, which are not too high.

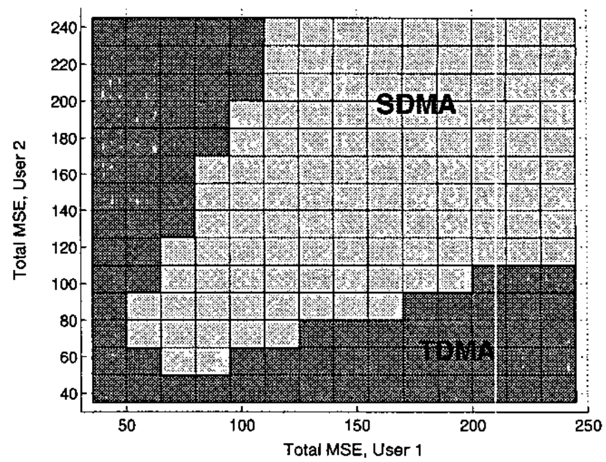

Fig. 11. Optimum operating modes (long-term TDMA, long-term SDMA) as functions of total distortion for two users.

VI. Conclusion

A cross layer design method is presented which is built upon the idea of equivalence classes of key-parameters of several layers in the protocol stack. An equivalence class is composed of all key-parameter tuples which fulfill a desired quality of service. Since different parameter tuples are usually associated with different costs (e.g. transmit power), a cross-layer design can select the most cost-efficient parameter tuple. The crucial part of this design approach is the development of a relation between quality of service and adjustable parameters on the one hand, and the association of costs with these parameters on the other hand. Two detailed examples are provided which show how this method can be applied for different air-interfaces and service applications.

References

[1] M. T. Ivrlac and F. Antreich, "Cross osi-layer optimization – an equivalence class approach," *Technical Report, Munich University of Technology*, vol. TUM-LNS-TR-03-09, Mai 2003.

[2] M. T. Ivrlac, L. U. Choi, J. A. Nossek, and R. D. Murch, "Effective use of long-term transmit channel state information in multi-user mimo communication systems," in *IEEE Vehicular Technology Conference, VTC Fall*, October 2003.

[3] B. Girod, E. Steinbach, and N. Faerber, "Performance of the h.263 video compression standard," *Journal of VLSI Signal Processing: Systems for Signal, Image, and Video Technology. Special Issue on Recent Development in Video: Algorithms, Implementation and Applications*, vol. 17, pp. 101–111, 1997.

[4] Lai U Choi, E. Steinbach, and W. Kellerer, "A joint optimization of application and radio link layers in wireless multimedia communication," in *ICC*, 2004, p. (submitted).

[5] M. T. Ivrlac, "Parameter selection for the gilbert-elliot channel model," *Technical Report, Munich University of Technology*, vol. TUM-LNS-TR-03-05, Mai 2003.