Application-driven Cross-layer Optimization for Coded OFDMA Systems[1]

S. Khan[1], J. Brehmer[2], W. Kellerer[3], W. Utschick[2], E. Steinbach[1]
[1]Media Technology Group, Technische Universität München, Munich, Germany
[2]Associate Institute for Signal Processing, Technische Universität München, Munich, Germany
[3] Future Networking Laboratory, DoCoMo Eurolab, Munich, Germany

Abstract— **In this paper we describe an application-driven cross-layer optimization approach for a multi-user coded OFDMA system. The physical layer is described as a set of Pareto optimal operating modes, from which the application layer chooses the best mode of operation taking the application-layer performance characteristics into account. The modular design employed by our scheme reduces the complexity problem associated with cross-layer optimization. We use the scheme in a multi-application scenario by defining a common application specific quality metric based on Mean Opinion Score. Our simulation results show that our proposed scheme leads to significant improvements in terms of user perceived quality compared to the traditional throughput maximization approach.**

Key words: Cross-layer optimization, coded OFDMA, Mean Opinion Score.

1. INTRODUCTION

In order to maximize performance of a multi-user wireless communication system an optimization of all system parameters is required. In particular, achieving optimum application layer performance requires looking at the system as a whole, as the performance experienced at the application layer strongly depends on the setup of the lower layers. In a multi-user coded OFDMA system each layer by itself represents a complex sub-system. At the lower layers, signal processing has to allocate resources to users in an optimum manner. At the application layer, multiple users are served with potentially different applications. Each application may be adaptable to the setup of the lower layers.

One approach to cross-layer optimization is to jointly optimize the parameters of all layers in a single optimization problem. A major disadvantage of this approach is that it requires full transparency between layers. In order to solve the joint optimization problem, there has to exist a single instance that knows all details about all layers. Also, a joint optimization requires a large flow of information among the layers and the optimization instance. For an overview of cross-layer optimization approaches, associated challenges, and discussions of inter-layer versus intra-layer optimization, see [10] - [14].

In order to overcome the limitations of the aforementioned brute-force joint optimization approach, we employ a concept for interface-based modular cross-layer optimization as first proposed in [1]. While still providing close to optimum system performance, modular cross-layer optimization achieves a far-reaching decoupling of the design problems at each layer and limits the information flow between layers during operation.

We apply our modular design concept in a multi-application scenario. Optimizing across multiple applications requires a performance metric that is applicable to different applications. In this paper an optimization based on mean opinion score (MOS) is employed [15]. MOS functions are derived for voice, file transfer, and real-time video applications. The MOS functions map a setup of the physical layer into MOS values. The developed MOS maximization aims at maximizing both user satisfaction and fairness among users, given the description provided by the physical layer.

This paper proceeds as follows: Our concept for modular cross-layer optimization is presented in Section 2. Section 3 summarizes the concept concerning the physical layer of coded OFDMA system. Application layer models are presented in Sections 4. Section 5 describes our MOS-based optimization approach. Section 6 contains simulation results.

2. MODULAR CROSS-LAYER OPTIMIZATION

In this section we briefly review the concept for modular cross-layer optimization presented in [1].

We limit our consideration to a simplified two-layer model, consisting of physical (PHY) layer and application (APP) layer only. The PHY layer has a number of parameters, e.g., power allocation and subcarrier allocation. Let a specific parameter setting be collected in a vector $\boldsymbol{x}_1$. These parameters, however, are not of intererst for the APP layer. What counts for the APP layer is the PHY layer performance achieved by a particular parameter setup, where performance is measured in terms of a number of performance metrics. In other words, at the PHY layer we have a so-called layer function f_1 which maps a parameter setup $\boldsymbol{x}_1$ into a performance vector $\boldsymbol{y}_1 = f_1(\boldsymbol{x}_1)$. If we (theoretically) vary $\boldsymbol{x}_1$ over all feasible parameter setups, we obtain a layer output set $\mathcal{Y}_1$, which contains all achievable performance vectors.

The APP layer may have its own parameters, where a specific setting is described by a parameter vector $\boldsymbol{x}_2$. Given a performance $\boldsymbol{y}_1$ of the PHY layer, a certain parameter setup $\boldsymbol{x}_2$ results in an APP layer performance, measured, for

[1] This work was supported by Deutsche Forschungsgemeinschaft (DFG) SPP 1102 and DoCoMo Eurolab, Munich, Germany.

example, in average application layer distortion or average mean opinion score (cf. Section 4). In other words, we have a layer function f_2 which maps $\boldsymbol{y}_1$ and $\boldsymbol{x}_2$ into an APP layer performance $f_2(\boldsymbol{x}_2, \boldsymbol{y}_1(\boldsymbol{x}_1))$. We aim at finding a parameter setup ($\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2$) such that $f_2(\boldsymbol{x}_2, \boldsymbol{y}_1(\boldsymbol{x}_1))$ is maximized.

$$(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2) = \underset{\boldsymbol{x}_1 \in \mathcal{X}_1, \boldsymbol{x}_2 \in \mathcal{X}_2}{\arg\max}\, f_2(\boldsymbol{x}_2, \boldsymbol{y}_1(\boldsymbol{x}_1)), \qquad (1)$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ denote the set of feasible parameter vectors of the PHY and the APP layer, respectively.

In order to achieve a close to optimum parameter setup while preserving modularity, we propose a cross-layer optimization based on information exchange between layers. The key idea is to have the PHY layer offer a set of achievable performance vectors $\mathcal{D}_1 \subseteq \mathcal{Y}_1$ to the APP layer. This set is denoted as layer description. The APP layer then chooses the performance vector from the description $\mathcal{D}_1$ that best fits its needs. The size of description $\mathcal{D}_1$ can be significantly reduced if we assume that the layer function of the APP layer $f_2(\boldsymbol{x}_2, \boldsymbol{y}_1)$ is monotone in $\boldsymbol{y}_1$. Monotonicity basically states that worse performance at the physical layer will not lead to better performance at the application layer. The optimum description of the PHY layer is given by the efficient set of the multiobjective optimization (MOO) problem

$$\mathcal{E} = \max_{\boldsymbol{x}_1 \in \mathcal{X}_1} f_1(\boldsymbol{x}_1)$$

A description in terms of efficient set $\mathcal{E}$ is termed optimum because it achieves the same APP performance as the joint optimization and is the smallest such description [1].

While a description in terms of efficient sets guarantees optimum APP layer performance, its size may still be too large. In [1] we propose to obtain a feasible description by sampling the efficient set. Descriptions obtained in this way are denoted as approximate descriptions.

3. PHYSICAL LAYER MODEL

This Section summarizes the PHY model described in [2]. We employ the metrics outage probability (ε) and data rate R to describe PHY layer performance. We make the idealizing assumption of capacity achieving coding and limit our considerations to the case of only statistical channel state information at the transmitter.

We consider a multi-user OFDMA system using Rayleigh block fading channel model. The system has N subcarriers and K non-cooperating receivers. The available transmit power is limited to E_{tr}. Assuming that the packet period T_p is sufficient to nearly achieve capacity, the appropriate information theoretic performance metric is outage probability.

The outage probability ε_k that the transmission of user k 's data is successful in general depends on the prescribed information rate R_k, the allocated transmit powers δ_i, $k = 1,\dots,K$ and the number N_k of subcarriers allocated to user k. At the physical layer, we aim at maximizing rates while minimizing outage probabilities. The PHY layer function f_1 is given by

$$f_1(\boldsymbol{x}_1) = \left(-\varepsilon_1, R_1, \dots, -\varepsilon_K, R_K\right), \qquad (2)$$

For OFDMA with variable subcarrier allocation the PHY layer parameters are given by $\boldsymbol{x}_1 = (\delta_1, R_1, N_1, \dots, \delta_K, R_K, N_K)$. The set of feasible parameters is given by

$$\mathcal{X}_1 = \left\{ x_1 : \delta_k \geq 0, \sum_{k=1}^{K} \delta_k \leq E_{tr}, R_k \geq 0, \sum_{k=1}^{K} N_k \leq N \right\}$$

s.t. $N_k \in \mathbb{N}_+^0$, where $\mathbb{N}_+^0$ denotes the set of positive integers including 0.

For a resource allocation $\boldsymbol{\delta}=(\delta_1,\dots,\delta_K)$ and $\boldsymbol{N}=(N_1,\dots,N_K)$, the outage probabilities of all users are collected in a tuple $\boldsymbol{\varepsilon}=(\varepsilon_1,\dots,\varepsilon_K)$. Computing the efficient set corresponds to the multiobjective optimization problem [3]:

$$\mathcal{E} = \min_{(\boldsymbol{\delta},\boldsymbol{N}) \in X_1} \boldsymbol{\varepsilon}(\boldsymbol{\delta}, \boldsymbol{N})$$

The efficient set $\mathcal{E}$ contains an infinite number of elements. An approximate description $\mathcal{D}_1$ is created by sampling the efficient set. We apply the sampling algorithm presented in [2]. We seek a good approximation of $\mathcal{E}$ by a finite number of efficient tuples ε using the method of proper equality constraints (PEC) [4]. An approximation of the efficient set for an example system with $K = 3$ users and $(R_1, R_2, R_3) = (100, 300, 300)$ *kbps* is shown in Fig 1.

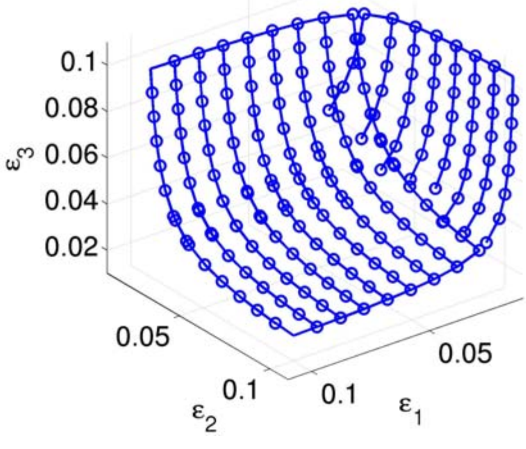

Figure 1: Approximation of outage region for K=3 users for fixed rates R_1, R_2, R_3.

4. MULTI-APPLICATION CROSS-LAYER OPTIMIZATION

The challenge of optimization across multiple applications has been treated in the literature mainly in the form of throughput maximization. Maximizing throughput leads to optimum performance only for applications which are insensitive to delay and packet loss. Multimedia applications such as video and voice are highly sensitive to changes in data rate, delay and packet loss rate. Even the importance of a packet changes dynamically depending on the history of previous packets. Due to these reasons, throughput maximization leads to performance which is usually not optimal with respect to user perceived quality for multimedia applications. Recently, first approaches have appeared that address the issue of optimizing across different applications using utility functions [6]. In [7] utility functions are used to perform resource allocation in wireless OFDM systems.

In our approach we use the well known Mean Opinion Score (MOS) as a common metric for different applications [15]. MOS was originally proposed for voice quality assessment and provides a numerical measure of the quality of human speech signals. The scheme uses subjective tests (opinionated scores) that are averaged to obtain a quantitative indicator of the system performance.

The multi-application CLO approach employed in this paper extends the use of MOS as a user-perceived quality metric to other applications, such as conversational and streaming video, web browsing and file download. This enables us to optimize the resource allocation across different applications using a common optimization metric. The layer function f_2 can be written as the average MOS of all the users competing for the wireless resources

$$f_2(\boldsymbol{x}_2, \boldsymbol{y}_1(\boldsymbol{x}_1)) = \frac{1}{K}\sum_{k=1}^{K} \lambda_k \cdot MOS_k(\boldsymbol{x}_2, \boldsymbol{y}_1(\boldsymbol{x}_1)) \qquad (3)$$

The weighting factor λ_k is the relative importance of the user as for instance determined by the service agreement between the user and the service provider.

The MOS-based optimization approach has several advantages with respect to previous work. First, compared to traditional techniques for multi-user resource allocation it allows us to directly relate network parameters, such as rate (R), packet error probability (p) and delay to a user-perceived application quality metric such as MOS. Second, compared to the application-driven cross-layer optimization described in [5] it allows us to further maximize the optimization gain taking advantage of the diversity not only across multiple users running the same application, but also across users running different applications. Our experiments applied to scenarios including multiple concurrent video, voice and file download applications show that MOS-based optimization significantly outperforms throughput-based optimization. In the following we describe example mappings of the lower layer performance parameters R and p to MOS for three different types of applications.

4.1. Voice communication

Perceptual Evaluation of Speech Quality (PESQ) [8] predicts with high correlation the quality scores that would be given in a typical subjective test. The PESQ algorithm is computationally too expensive to be used in real-time scenarios. To solve this problem we propose a model to estimate MOS as a function R and p. The available rate determines the voice codec that can be used. In Fig. 2 we show experimental curves for MOS estimation as a function of packet loss rate p for different voice codecs. The curves are drawn using an average over a large number of voice samples and channel realizations (packet loss patterns). These curves can be stored in the base station for every codec that is supported. If transcoding from an unsupported codec is required, such curves have to be signaled to the base station as side information. Choosing between different available transmission rates can be addressed by transcoding the incoming speech signal to different voice codecs. Every voice codec leads to a different MOS value in the case of error-free transmission and shows a different sensitivity to packet loss. As an example let us consider two lower layer parameter tuples (R=64 kbps, p=14%) and (R=6.4 kbps, p=0%) and assume these two represent possible operating modes of the lower layers for a particular user. The corresponding MOS values are marked with an arrow in Fig. 2. In this example the second parameter tuple (R=6.4 kbps, p=0%) leads to a gain of 0.3 on the MOS scale and the cross-layer optimizer would select it as its outcome.

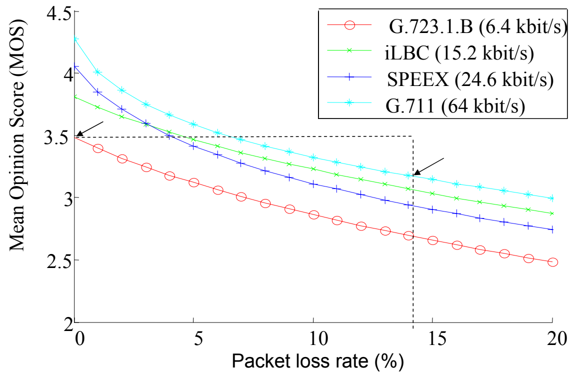

Figure 2: MOS vs. packet error probability p for different voice codecs

4.2. Video communication

In video communication we distinguish between streaming video applications with moderate delay requirements and conversational video applications with very rigid delay requirements. For streaming video the content is typically pre-encoded and we can extract rate-distortion side information during the offline encoding step that can be later used by our cross-layer optimizer for computing the expected distortion in the presence of packet errors.

For both types of video applications the expected reconstruction distortion can be approximated as the sum of source distortion and loss distortion. The source distortion depends only on the encoding rate while the loss distortion depends on the actual packet loss probability. The total distortion is typically measured as the mean square error (MSE) between the original video frames and the reconstructed frames. PSNR which is a logarithmic measure of

MSE is the standard objective way of presenting video quality results. The relationship between PSNR, *R* and *p* are assumed to be available from a table as described in [15].

In this work we assume a simple linear relationship between PSNR and MOS. We assume that maximum user satisfaction is achieved for a PSNR of 45dB and minimum user satisfaction results for PSNR values below 25dB. The upper limit comes from the fact that reconstructed video sequences with 45dB PSNR are almost indistinguishable from the original and below 25dB very severe degradations distort the video.

4.3. File transfer

To estimate user satisfaction for file transfer we use the logarithmic MOS-throughput relationship introduced in [9]. We assume that every user has subscribed for a given data rate and the level of satisfaction is characterized by the real rate received. The MOS is estimated based on the current rate R offered to the user by the system and packet error prob. p

$$MOS(R,p) = a \cdot \log_{10}[b \cdot R(1-p)] \qquad (4)$$

where a and b are determined from the maximum and minimum user perceived quality. If a user has subscribed for a specific rate $R_{service}$ and receives $R_{service}$, then in case of no packet loss the user satisfaction on the MOS scale should be maximum. On the other end, we define a minimum transmission rate (e.g. 10 kbps) and assign to it a MOS value of 1. Using the parameters a and b, we fit a logarithmic curve for the estimated MOS.

5. MOS MAXIMIZATION

The goal of the optimization is to achieve maximum user satisfaction and fairness among the users, given the description provided by the PHY layer.

Each of the K users in the system requests one out of three different types of services. Let us define three sets of users: U - requesting voice service, V - file download and W - video conference. Let $MOS_U(x_U, y_k)$ denote the MOS function of voice service for user k. Let $\boldsymbol{y}_1 = (y_1, ..., y_K)$ be the performance vector of PHY layer and $\boldsymbol{x}_2 = (x_U, x_V, x_W)$ be the parameter tuples from the APP layer (cf. Section 2). The MOS value for voice service depends on the voice service parameters (e.g. the voice codec) and the performance tuple $y_k = (R_k, \varepsilon_k)$ provided by the PHY layer. The APP layer packet error probability p can be calculated from the PHY layer outage probability ε, given the packet sizes at the PHY and APP layers. The MOS functions for file download and video conference are denoted as $MOS_V(x_V, y_k)$, and $MOS_W(x_W, y_k)$, respectively. Moreover, let $\mathcal{X}_U$, $\mathcal{X}_V$, and $\mathcal{X}_W$ denote the set of feasible service parameters for each of the three service classes. As an example, for voice service we may have four voice codecs to choose from, corresponding to

$$\mathcal{X}_U = \{\mathrm{G.711}, \mathrm{G.723.1}, \mathrm{iLBC}, \mathrm{Speex}\}.$$

Our objective function for multi-user multi-application cross-layer optimization is defined in equation (5). A maximization of the sum of the QoS (MOS) perceived by every user in our multimedia wireless network has to be achieved. The parameter λ is used to give higher priority to a given user and it is up to the network operator to choose its value.

$$\max_{\substack{x_U \in \mathcal{X}_U, \\ x_V \in \mathcal{X}_V, \\ x_W \in \mathcal{X}_W, \\ y \in \mathcal{D}_1}} \left(\begin{array}{l} \sum_{k \in U} \lambda_k \mathrm{MOS}_U(x_U, y_k) + \sum_{k \in V} \lambda_k \mathrm{MOS}_V(x_V, y_k) \\ + \sum_{k \in W} \lambda_k \mathrm{MOS}_W(x_W, y_k) \end{array} \right) \qquad (5)$$

Recall that $\mathcal{D}_1$ is the description provided by the PHY layer.

In the simulation described in the next Section we compare the MOS-based application-centric cross-layer optimization approach with maximum throughput optimization. The objective for maximum throughput optimization is to maximize the sum of the goodput available to all the users in the system.

6. SIMULATION

The simulations shown in this paper are performed with the following parameter settings. We assume a total of five users in the wireless network. Two voice users (user1 and user2), one male and one female voice, are used. The voice samples are 30 seconds long. The voice signal comes from the backbone network encoded with G.711 voice codec at 64 kbps. In the base station, following the optimization output, the signal could be transcoded to 6.4 kbps with G.723.1 codec, 15.2 kbps wth iLBC codec, 24.6 kbps with Speex or it can be transmitted without transcoding at 64 kbps.

Two users (user3 and user4) are using video conferencing service. The video sequences used are foreman and grandmother, respectively, encoded with H.264. The frame sequence structure is I-P-P-…-P which is the appropriate format for real-time video.

One user (user5) subscribes for file download using FTP. The user requests for a service with maximum offered transmission rate of 200 kbps. The λ values in equation (4) are all set to 1 in our experiments.

At the PHY layer we use OFDMA with variable subcarrier allocation as described in Section 3. We use a bandwidth of $K \cdot 100$ kHz with 30 subcarriers. The PHY layer provides a description $\mathcal{D}_1$ to the APP layer in terms of (R, ε) tuples. The size of the description depends on the sampling of R and ε. The average transmit SNR per user is assumed to be 15 dB. The users are assumed to be all moving away from the base station from an initial distance of 100 meter, at a rate of 1 meter/sec.

Fig. 3 shows the mean MOS based on 100 simulation runs, each of 30 seconds duration. We observe a gain of mean MOS for our MOS maximization approach, compared to the throughput maximization approach. Fig. 4-6 shows the results for a single simulation run. In Fig. 4 we observe the highest MOS gain for our approach for user3 (foreman video) which is the most demanding of all the users. Fig. 5 and Fig. 6 shows the rate and outage probability of the users for the same simulation run.

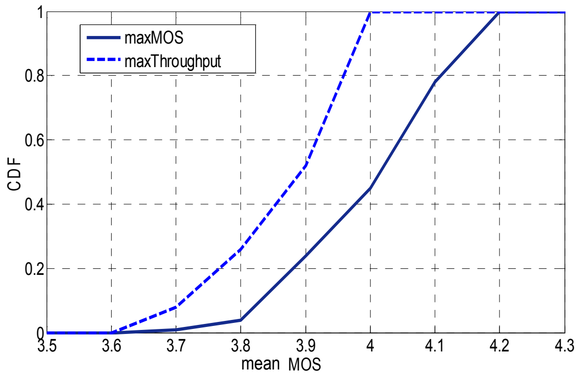

Figure 3: Cumulative Density Function of mean MOS for MOS maximization and throughput maximization.

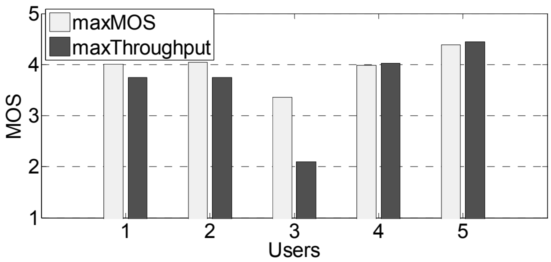

Figure 4: MOS values for a single simulation run

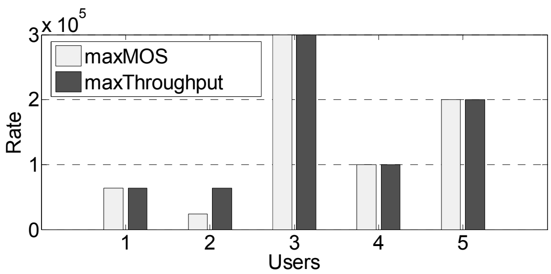

Figure 5: Selected rate in bit/sec for a single simulation run

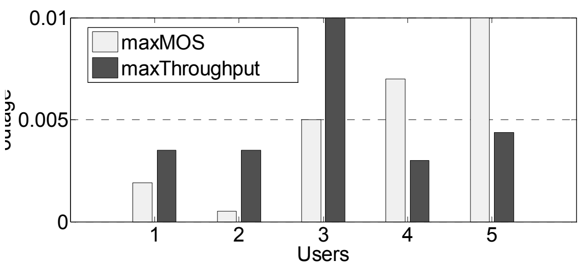

Figure 6: Selected outage probabilities for a single simulation run

7. CONCLUSION

A concept of application-driven modular cross-layer optimization for a coded OFDMA system is presented. The layer parameters are optimized according to an optimality criterion defined at the application layer. The proposed concept is based on a “bottom-up” exchange of layer descriptions between layers. The lower layer capabilities are described by means of an efficient set of Pareto optimal operation modes. We use the scheme in a multi-application scenario by defining a common application specific quality metric called Mean Opinion Score. The simulation results show that our proposed scheme leads to significant improvements in terms of user perceived quality compared to the traditional throughput maximization approach.

REFERENCES

1. J. Brehmer, W. Utschick, “Modular cross-layer optimization based on layer descriptions,” WPMC 2005, Aalborg, Denmark, Sept. 2005.
2. J. Brehmer, C. Guthy, and W. Utschick, “An efficient approximation of the OFDMA outage probability region,” SPAWC 2006, Cannes, France, July 2-5, 2006.
3. Y. Sawaragi, H. Nakayama, and T. Tanino, “Theory of multiobjective optimization,” Academic Press, Orlando, 1985.
4. J. G. Lin, “Multiple-objective problems: Pareto-optimal solutions by method of proper equality constraints,” IEEE Trans. Automat. Contr., vol. 21, pp. 641-650, Oct. 1976.
5. L. Choi, W. Kellerer, and E. Steinbach, “Cross-layer optimization for wireless multi-user video streaming,” IEEE Int. Conf. on Image Processing, ICIP 2004, Singapore, October 2004.
6. G. Song and Y. Li, ”Utility-based Resource Allocation and scheduling in OFDM-based wireless broadband networks,” IEEE Com. Mag., pp. 127-134, Dec. 2005.
7. G. Song and Y. Li, “Cross-layer optimization for OFDM wireless networks – part 1 and part 2,” IEEE Trans. on Wireless Communications, vol. 4, pp. 614-634, Mar. 2005.
8. ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Feb. 2001.
9. F. P. Kelly, “Charging and rate control for elastic traffic,” Eur. Trans. Telecommun., vol. 8, pp. 33-37, Jan. 1997.
10. S. Shakkottai, T. Rappaport, P. Karlsson, “Cross-layer design for wireless networks,” IEEE Com. Mag., vol. 41, no. 10, pp. 74-80, Oct. 2003.
11. V. Kawadia, P. Kumar, “A cautionary perspevtive on cross layer design,” IEEE Wireless Communications, vol. 12, no. 1, pp 3-11, Feb. 2005.
12. S. Khan, M. Sgroi, Y. Peng, E. Steinbach, and W. Kellerer, “Application-driven cross-layer optimization for video streaming over wireless networks,” IEEE Commun. Mag., vol. 44, no. 1, Jan. 2006, pp. 122-130.
13. Lai-U Choi, M. T. Ivrlac, E. Steinbach, and J. A. Nossek, “Bottom-up approach to cross-layer design for video transmission over wireless channels,” IEEE 2005 Spring Semiannular Vehicular Technology Conference, Stockholm, Sweden, 30 May-1 June, 2005.
14. M. van d. Schaar, S. Shankar, “Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms,” IEEE Wireless Communications, vol. 12, no. 4, pp. 50-58, Aug. 2005.
15. S. Khan, S. Duhovnikov, E. Steinbach, M. Sgroi, W. Kellerer, “Application-driven cross-layer optimization for mobile multimedia communication using a common application layer quality metric,” Second Internation Symposium on Multimedia over Wireless, ISMW’06, Vancouver, Canada, July 3-6, 2006.