# QoE-Driven Cross-Layer Optimization for High Speed Downlink Packet Access

Srisakul Thakolsri[1], Shoaib Khan[2], Eckehard Steinbach[2], Wolfgang Kellerer[1]
[1]DOCOMO Euro-Labs, Munich, Germany
{thakolsri, kellerer}@docomolab-euro.com
[2]Institute for Media Technology,
Technische Universität München (TUM), Munich, Germany
khan@mytum.de, Eckehard.Steinbach@tum.de

***Abstract*— This paper proposes a Quality of Experience (QoE) based cross-layer design (CLD) framework for High Speed Downlink Packet Access (HSDPA). The proposed scheme aims at maximizing the user satisfaction by taking advantage of the link adaptation mechanism of HSDPA and the rate adaptation capability of multimedia applications. The main contributions of the paper are as follows. First, we describe the multiuser rate region of HSDPA by constructing a long-term radio link layer model. Next, we formulate multimedia QoE by constructing long-term utility functions, describe the multiuser utility space and derive its properties. We show analytically that the maximization of the sum of utility (max-MOS) can be efficiently solved by a fast greedy algorithm which searches only through the boundary of the utility space. We investigate two alternatives to the max-MOS approach, which introduce additional fairness in the system. We compare our proposed QoE-based cross layer optimization schemes to a system that is configured to maximize the overall throughput. For the sake of completeness, we also compare our approaches to a non-optimized HSDPA system. The performance comparison is made by simulation using a software implementation of an actually deployed HSDPA system. Results show that our QoE-based approach leads to significantly improved user perceived quality compared to the other approaches.**

***Index Terms*— Quality of Experience (QoE), High Speed Downlink Packet Access (HSDPA), Application-driven cross layer optimization.**

## I. INTRODUCTION

The increased usage of a wide variety of cellular multimedia services is putting an ever increasing demand for high data rates on the wireless interface. As the downlink of the cellular system often acts as the bottleneck link, an efficient usage of downlink wireless resources becomes essential in order to provide high quality of services to the largest possible number of users. The time varying transmission conditions of the wireless channel and the dynamic changes of application requirements of multimedia services make the optimization of downlink resources a challenging task. Cross-layer design (CLD) has been proposed to address this issue [1]. By exchanging key parameters across the layers [2], a CLD scheme can enable a highly efficient wireless resource allocation [3], [4], [5].

High Speed Downlink Packet Access (HSDPA) provides a shared packet-switched service with variable bitrate [6]. Thus, HSDPA allows for the concurrent usage of diverse, resource demanding applications such as video streaming, video conferencing, network gaming and 3D navigation, which all require variable data rates. However, an efficient distribution of the HSDPA resources to the concurrent applications is still a challenging task. When traditional Quality of Service (QoS) measures, such as Guaranteed Bit Rate (GBR) are used for these services, it either results in congestion due to the increase of data rates from the initial rates, or underutilization of resources due to rate decrease. In both cases, the users suffer from unsatisfactory services.

This paper proposes a Quality of Experience (QoE) based CLD approach for resource allocation in HSDPA. The CLD framework is integrated into an OPNET-based HSDPA simulator, although it would be equally applicable to other future packet based services as well. We use an application-driven approach which considers the QoE accross multiple different applications. This framework takes into account the total resource constraint of the system and periodically reassigns the resources with the aim of maximizing the user satisfaction. In this way our CLD framework adds a software-based optimizer component to an existing system but does not violate the protocol layering.

*Remark*: QoE-based resource allocation for future cellular multimedia networks is important for several reasons. Firstly, current throughput-based optimization only makes sense in case of packet-based charging. High cellular bitrates, e.g. 3GPP Long Term Evolution (LTE), which is expected to approach a peak bit rate of 100 Mbps [7], would further push a flat-rate billing model or models based on quality guarantees. In that scenario, the operators would find a clear motivation to maximize the satisfaction of their customers, irrespective of the requirements of their services. Secondly, user satisfaction is gaining importance to the operators who realize that unsatisfied users would usually quit the network without ever complaining to the operator, and would possibly

Manuscript received March 15, 2009; revised April 15, 2009; accepted May 15, 2009.


doi:10.4304/jcm.4.9.669-680

share their experience with other potential customers, resulting in severe loss of revenues. Thirdly, QoE-based optimization allows potentially more customers to be served simultaneously without a loss of user perceived quality.

QoE-based resource allocation as presented in this paper is performed at intervals of seconds. It fills a gap between physical layer transmission intervals (2ms in HSDPA) and long term application layer mechanisms such as adaptive streaming or TCP congestion control (10s of seconds).

*Literature Review*: The challenge of downlink resource optimization across multiple applications has been treated mainly in the form of throughput maximization [8], [9]. Multimedia applications such as video streaming and voice telephony are highly sensitive to changes in data rate, delay, delay jitter and packet losses. Even the importance of a packet changes dynamically depending on the transmission 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.

HSDPA throughput performance results are presented and compared with the WCDMA Release '99 dedicated channel in [10], [11], [12] and [13]. The packet multiplexing of HSDPA provides a significant multiuser diversity gain when combined with a channel aware scheduling scheme. Performance of several channel aware scheduling schemes, e.g., max C/I, proportional fair (PF) and modified largest weighted delay first (M-LWDF) have been shown for HSDPA [14], [15], [16]. [14] gives a comparison of several scheduling schemes for streaming and non-realtime applications. [14] and [15] propose to use the M-LWDF scheduler to ensure uninterrupted media playout. [16] proposes a scheduling scheme to improve wireless TCP performance over HSDPA. [17] and [18] propose using queue length information in scheduling in order to ensure fairness.

Specific scheduling schemes for different applications, e.g., video streaming and voice over IP (VoIP) over HSDPA, are proposed in [19] and [20], respectively. [21] and [22] propose quality-driven access control and the use of the 3GPP QoS framework [23] in HSDPA.

A QoE-based application-layer adaptation for HSDPA has not been proposed so far in the literature. A utility-based optimization framework was first proposed in [24], where the utility function is assumed to capture the user satisfaction with respect to data rate. For a comprehensive overview of the Network Utility Maximization (NUM) framework please refer to [25] and the references therein. Contrary to most of the NUM literature where only concave, continuously differentiable utility functions and theoretical link models are assumed, our scheme proposes a framework considering realistic utility functions and applies the framework to a standardized system.

*Main Results*: In this paper we present details of our QoE-based CLD framework for HSDPA and its evaluations. First, we propose long term link layer and application layer models by extracting key parameters from the respective layers. The parameters are communicated to a cross-layer optimizer which acts as a downlink resource allocator. The optimizer periodically reviews the total system resources and makes an estimate of the time-share needed for each user for each possible application-layer rate. If necessary, the optimizer suggests re-adaptation of the application rates. This adaptation of the application-layer rate is motivated by the following arguments:

1) Admission control policies are traditionally used to detect violation of QoS. With the advent of new demanding applications it becomes challenging to maintain reasonable user satisfaction without compromising too much on efficiency. The proposed CLD framework goes beyond admission control by continuously adapting the network and thereby ensuring both good user experience and efficient utilization of spectrum.
2) Unlike the second and third generation wirless standards, HSDPA is inherently a variable bit rate channel. Dynamic adaptation of the application-layer streams is thus attractive to use in this context, which can take advantage of the high bit rate provided by the standard. A media stream can be adapted by using transcoding and packet dropping. Transcoding is flexible but computationally intensive, whereas packet dropping leads to lower complexity at the cost of reduced quality. Adaptation in our work is performed by transcoding, which is performed on a node close to the base station. The transcoding is applied to the voice and video streams and its impact on the utility function is also taken into account. (see Fig. 6)

The rest of the paper is organized as follows. Section II introduces our long term link layer and application layer models. Section III shows the problem formulation of our QoE-based CLD framework. Section IV describes the greedy algorithm in detail. Section V gives simulation results and our conclusions are drawn in Section VI.

## II. PRELIMINARIES

### A. HSDPA Overview

The key concept of HSDPA is to increase the packet data throughput using link adaptation and fast retransmission from the base station (Node B). Link adaptation of HSDPA uses Adaptive Modulation and Coding (AMC) with two modulation schemes, QPSK and 16-QAM, and a rate $1/3$ turbo code with variable amount of puncturing. AMC adapts to the radio condition based on the Channel Quality Indicator (CQI) report from the receiver every Transmission Time Interval (TTI) which is fixed at 2ms.

Figure 1 shows the scenario considered in this paper, with the three main network elements involved in HSDPA: Radio Network Controller (RNC), Base Station or Node B, and the User Equipment (UE). The RNC is responsible for the control of the radio resources. The Node B schedules the packets to the UEs, taking

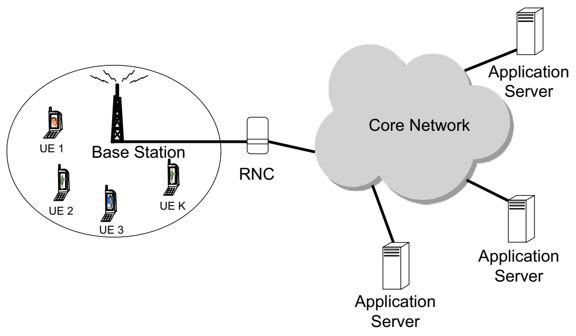


Figure 1. Scenario considered in this paper

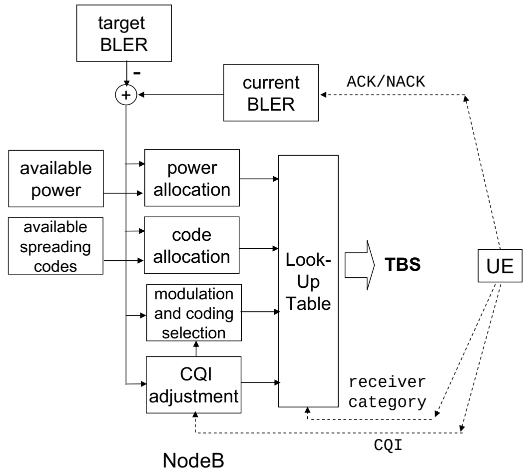


Figure 2. TBS estimation process adapted from the standard [26] and extended for the deployed HSDPA system.

advantage of AMC. At the RNC, IP packets are received from the core network and each of them is encapsulated into one Radio Link Control (RLC) Service Data Unit (SDU). The RLC-SDU is then segmented into fixed-size RLC-PDUs. At the Node B, one transport block (TB) is sent over the air each TTI. The number of information bits that can be sent in each TB is denoted as the Transport Block Size (TBS) which depends on the CQI of the user.

The process of estimating the TBS is shown in Fig. 2. Each TTI one or more users are selected to be scheduled. When multiple users are allowed to be scheduled within the same TTI, the available power and code resources are calculated using a resource-allocation algorithm. When user multiplexing is not used, all the available power and code resources can be allocated to a single user during the TTI. Look-Up Tables (LUT) are used to get the TBS, given the available power, code and CQI values. The UE also sends ACK/NACK messages associated to the previous TB. This helps to estimate the actual Block Error Rate (BLER) of the user. An appropriate TBS is chosen for a target BLER of 10%. The difference between the target and the current BLER is used to update the power, code and CQI values to be used in the LUT.

Let $\mathcal{K}$ be the set of users, $\mathcal{K} = \{1, 2, \cdots, K\}$. Let $k^i$ be the user who is given access to the channel at time $i$, where $i$ is the index of TTI, $k \in \mathcal{K}$, $i \in \mathcal{Z}_+$ with $\mathcal{Z}_+$ being the set of positive integers. Let $\mathcal{Q}$ be the set of possible CQI, $\mathcal{Q} = \{1, 2, \cdots, 30\}$, and $Q_k^i$ be the CQI of user $k$ at time $i$, $Q_k^1 \in \mathcal{Q}$.

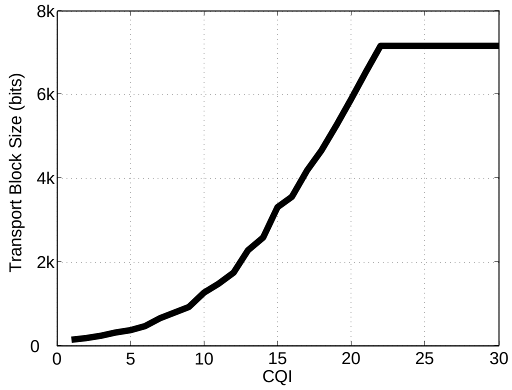


Figure 3. TBS vs. CQI for a category 6 receiver [26]. A variable TBS is attained by using adaptive modulation and coding, combined with a variable number of spreading codes. The category 6 receiver can use either QPSK or 16QAM and a maximum of five spreading codes.

Using AMC the Node B chooses a transmission format for a fixed target BLER resulting into a TBS of $B_k^i$ which depends on $Q_k^{i-d}$:

$$B_k^i = g(Q_k^{i-d}) \tag{1}$$

where $d$ is the link adaptation delay. The relationship in (1) is standardized by 3GPP [26].

### *B. Radio Link Layer Model*

In this paper we adopt the long term link layer model originally proposed in [27]. We estimate the long term average rate of each user $k$, denoted by $R_{max,k}$, $k \in \mathcal{K}$, that the user can support when all the resources are allocated to the user. Let $R_k$ be the long term data rate provided to user $k$, given the time share $\alpha_k$. Then the radio-link layer is described as:

$$R_k = \alpha_k \cdot R_{max,k}, 0 \leq \alpha_k \leq 1, \forall k \tag{2}$$

(2) defines the HSDPA rate region. In the following the estimation of $R_{max,k}$ is performed for HSDPA. For the analysis we consider an individual user at some time instant. Hence, we drop the user and the time index. Let $r$ be the instantaneous data rate of the user. Assuming that the scheduler selects only the users who have packets to send, and $Q$ is slowly varying, $r = B$, and from (1) follows:

$$r = g(Q) \tag{3}$$

Taking expected values on both sides of (3):

$$E\{r\} = E\{g(Q)\} \tag{4}$$

Assuming only one user is scheduled at a TTI, all the resources are allocated to the user, so that

$$E\{r\} = R_{max} \tag{5}$$

Within the time interval of interest, it can be assumed that the mean of $CQI$ does not change considerably.

Therefore, although (3) is a non-linear function when the whole domain of the function is taken into account, it can be approximated by a piecewise linear curve, as shown in Figure 3. Hence,

$$E\{g(Q)\} = g(E\{Q\}) = g(\bar{Q}) \tag{6}$$

where $E\{Q\} = \bar{Q}$. From (4), (5) and (6):

$$R_{max} = g(\bar{Q}). \tag{7}$$

Hence, $R_{max}$ can be estimated by observing the mean CQI values over a period of time.

### C. Application Layer Model

We use utility functions to describe the Quality of Experience (QoE) for different applications as a function of *lower* or *radio link* layer parameters, e.g., rate or throughput, time share, power, spreading code, bandwidth, etc. As a measure of utility, we use the Mean Opinion Score (MOS), as proposed in [28]. The utility functions in [28] are described as a function of transmission data rate and packet loss rate. Due to the HSDPA MAC-layer retransmission mechanism, we assume that all packets are transmitted successfully and therefore, the utility function can be simplified as a function of transmission data rate as given below:

$$U = f(R), f : \mathcal{R} \rightarrow MOS \tag{8}$$

where $\mathcal{R}$ is the set of possible rates, and $MOS = [1 : 4.5]$. MOS 4.5 means that the user would rate the service with an excellent quality, while MOS 1 means the service is expected to be rated by all users with a very poor quality. Below we describe the derivation of the utility functions of different applications and the multiuser utility space in details.

*1) Voice call application:* Assessment of voice quality can be done by performing subjective tests with panels of human listeners. Such tests are not suitable for online system optimization. Alternatively, objective measures predicting the one-way voice quality scores (MOS) given by the user such as the ITU-T Perceptual Evaluation of Speech Quality (PESQ) [29] can be used. However, the algorithms are computationally expensive and require the original speech signal, and hence are also not suitable for dynamic online system optimization. To solve this we precompute voice utility functions that estimate MOS by using the PESQ algorithm. Previous work on measuring voice quality has shown that the QoE for voice applications depend on the encoding rate, packet loss rate and delay [30]. For simplicity and considering the retransmission at the MAC-layer, we define MOS as a function of the transmission rate R as depicted in Fig.4. Each point represents a different codec (G.723, iLBC, SPEEX and G.711) and the MOS is measured from a set of speech files with different contents for the case of error-free transmission. Due to distortion imposed by the source codec, every voice codec leads to a different MOS value. This utility curve can be stored at the base station for information when performing QoE-based optimization for resource allocation.

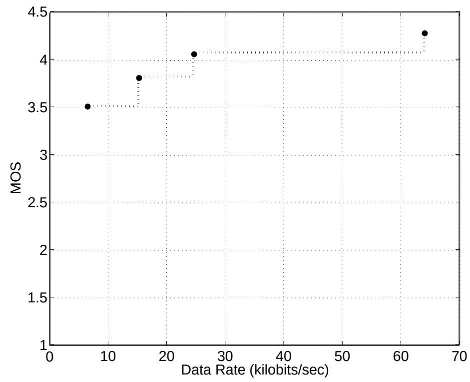


Figure 4. PESQ-based MOS as a function of the available data rate for different voice codecs [28]. The utility curve consists of 4 discrete points since the 4 codecs operate at fixed bit rates of 6.4kbps, 15.2kbps, 24.6kbps, and 64kbps, respectively.

*2) File download application:* File transfer or web browsing applications are considered to be elastic services, for which the utility function is an increasing, strictly concave, and continuously differentiable function of throughput [24]. Based on this assumption, the data transfer utility function is assumed to be logarithmic with respect to rate as following.

$$MOS = a \cdot log_{10}(b \cdot R) \tag{9}$$

where a and b are determined from the maximum and minimum user perceived quality. If a user has subscribed for a specific rate service R and receives this service rate R when downloading the file, then in case of no packet loss user satisfaction on the MOS scale should be maximum, i.e, 4.5. On the other hand, we define a minimum transmission rate (e.g., 10kbps in Fig.5) and assign to it a MOS value of 1. Using the parameters a and b, we fit the logarithmic curve in (9) for the estimated MOS. Fig.5 presents the MOS curve by varying the actual transmission rate R. In fact, data transfer uses TCP as its transport protocol, which has its own end-to-end mechanisms between the sender and the receiver such as the flow control and the congestion control. The flow control adapts the sending data rate in order to prevent a fast sender from over running a slow receiver, while the congestion control keeps the data flow below a rate that would trigger a network congestion, which makes the network performance fall. In order to adapt the transmission rate to optimize the transmission in a base station, the cross-layer optimizer might contact the sender or simply, e.g., slow down the TCP flow. If this happens on a small time scale (e.g., seconds) TCP will not notice. If this situation pertains then TCP will react accordingly by adapting its sending rate. We assume that the TCP rate adaptation process, which is for example modeled by the TCP Friendly Rate Control (TFRC) equation, has no significant impact on the user perceived throughput (as shown in (9)).

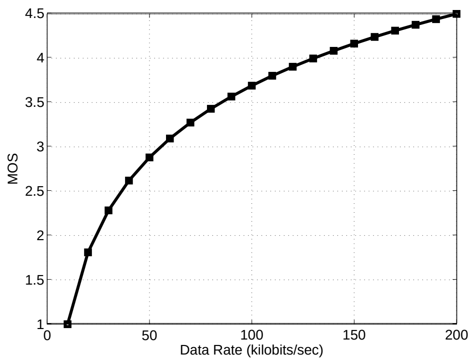


Figure 5. MOS as a function of transmission rate for file download applications [28].

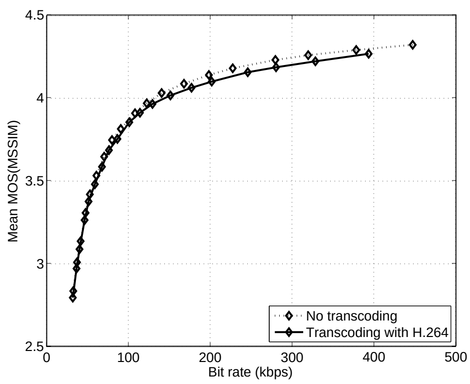


Figure 6. MOS for a video streaming application with the 'Foreman' video sequence.

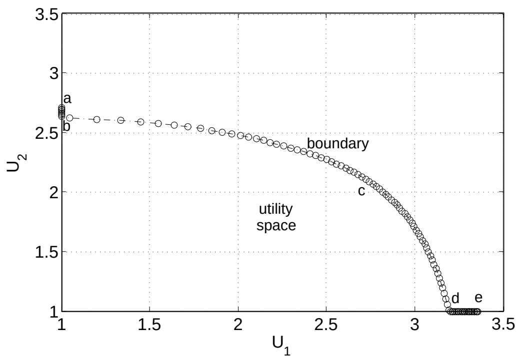


Figure 7. Utility space, $\mathcal{U}$ and its boundary, $\mathcal{B}_U$ for a two user case with $\bar{Q}_1 = 15dB$ and $\bar{Q}_2 = 5dB$. User 1 and user 2 are using video streaming and FTP services, respectively.

*3) Video streaming application:* Peak Signal to Noise Ratio (PSNR) has been used widely to measure the full-reference video quality due to its simplicity of calculation. However, many studies [31], [32], [33] show that PSNR does not match well to perceived visual quality. Due to the fact that the human visual system is highly adapted to extract structural information, we use the Structural SIMilarity (SSIM) index [33] to assess perceptual video quality, which measures the structural information change. In principle, SSIM measures the similarity of the two signals (the original signal and the distorted signal) by comparing the luminance, the contrast and the structure. The luminance is the mean intensity from the signal. The contrast is the standard deviation of the signal. The structure is the signal after luminance subtraction and variance normalization. These two signals are taken from a local window, which is just a part of the whole image. To evaluate the overall image quality, we calculate a mean SSIM (MSSIM).

$$MSSIM(\mathbf{X}, \mathbf{Y}) = \frac{1}{M} \sum_{j=1}^{M} SSIM(\mathbf{x}_j, \mathbf{y}_j) \tag{10}$$

where **X** and **Y** are the reference and distorted images, respectively; $\mathbf{x}_j$ and $\mathbf{y}_j$ are the image contents at the *j*-th local window; and *M* is the number of local windows of the image. To obtain the utility functions for video streaming, we vary the quantization steps for encoding the raw video and measure the average data rate of the video and the average MSSIM of all images in the video. For simplicity, the video utility functions assume a linear mapping from MSSIM of the entire video to MOS. Fig. 6 depicts an example of a video utility curve for the 'Foreman' video sequence. The dotted line and the solid line are the utility curves considering the distortion caused by the source encoding and the distortion caused by using a simple transcoding with the same codec (re-encoding) in the core network, respectively. In this example, we transcode the video with a high data rate (450kbps) to a lower data rate. Obviously, the transcoding causes an additional video quality degradation due to the re-encoding process. We assume that the utility functions are precomputed at the streaming server and signalled as side information along with the video bitstreams.

*4) Multiuser Utility Space:* The multiuser utility space, $\mathcal{U}$, defines a set of feasible utility vectors constrained by the total system resources:

$$\mathcal{U} \subseteq R^K, \sum_k \alpha_k \leq 1, \tag{11}$$

where $R^K$ is the $K$ dimensional Euclidean Space and $\alpha_k$ is some normalized resource share to user $k$. Since HSDPA is a time division multiplexed (TDM) channel, in the rest of the paper we assume $\alpha_k$ to be the time share given to user $k$.

A multiuser utility space can be formed by combining the transmission policies of every user. First, for simplicity, we show an example for a two user case, where the users are using video and file download services. If we assume that the mean receiver-side Signal-to-Noise-Ratio (SNR) of the video and file download user are 15dB and 5dB, respectively, the utility space ($\mathcal{U}$) and its boundary ($\mathcal{B}_U$) for this scenario can be illustrated as depicted in Fig. 7. $a$–$b$ and $d$–$e$ correspond to user rates of $(0, R_{max,1})$ and $(0, R_{max,2})$. $c$ is the optimum point with respect to the objective function described in the next section.

## III. QoE-BASED CROSS-LAYER OPTIMIZATION

### A. Utility-based optimization

The utility functions introduced in Section II-C provide the information about the required transmission rate at the application-layer in order to achieve a certain level of QoE. The representation of the lower-layers depends on the channel quality of each user. Information about the channel quality is obtained by CQI feedback from the UE as described in Section II-B. Depending on the selected objective function, the optimizer allocates the wireless system resources differently. Below we discuss two QoE-based objective functions applied in our work.

*1) Utility maximization:* The optimizer maximizes the objective function which is defined as the average utility of all users:

$$\mathcal{F}(\tilde{\mathbf{x}}) = \frac{1}{K} \cdot \sum_{k=1}^{K} U_k(\tilde{\mathbf{x}}) \tag{12}$$

where $\mathcal{F}(\tilde{\mathbf{x}})$ is the objective function with the cross-layer parameter tuple $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$. $K$ is the total number of users in the system, $\tilde{\mathbf{X}}$ is the set of possible parameter tuples abstracted from the protocol layers representing the set of candidate operation modes. The decision of the optimizer can be expressed as:

$$\tilde{\mathbf{x}}_{opt} = \arg\max_{\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}} \mathcal{F}(\tilde{\mathbf{x}}) \tag{13}$$

where $\tilde{\mathbf{x}}_{opt}$ is the parameter tuple which maximizes the objective function. After selection of the optimal values of the parameters, those parameters are sent back to the individual layers, which are responsible for translating them back into actual layer-specific modes of operation. Further details of parameter abstraction can be found in [2], [34] and [35].

Depending on the type of appplication, we create different sets of transmission policies, which specify possible transmission data rates. We denote the set of transmission policies for a user $k$ by $T_k$. With utility-based optimization, the optimizer chooses a combination of resource allocation that maximizes the following objective function:

$$\mathcal{F}(\tilde{\mathbf{x}}) = \sum_{k=1}^{K} \sum_{j=1}^{|T_k|} E\{\mathcal{I}_{kj} \cdot U_{kj}(\tilde{\mathbf{x}})\} \tag{14}$$

where $k$ denotes the user index, $j$ refers to the index of the transmission policy. $\mathcal{I}_{kj}$ is the indicator function. Its value is 1 when the transmission policy $j$ is chosen for user $k$, and 0 otherwise.

*2) Max-min utility:* The max-min fairness concept [36] applied to our QoE-based cross layer optimization means that the optimizer allocates the resources such that all users experience the maximally possible same level of quality. The max-min objective function is defined as:

$$\tilde{\mathbf{x}}_{opt} = \arg\max_{\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}} \left\{ \min_{k \in K} U_k(\tilde{\mathbf{x}}) \right\} \tag{15}$$

A drawback of using max-min fairness is the unequal quality loss. For instance, when a single user runs a very demanding application or has a very poor channel quality, the optimizer tries to give this user more resources and therefore forces all other users to share this poor experience. A modified max-min technique [37] has been proposed to allow for setting a minimum guarantee of service quality. It first checks whether there is enough resources to provide all users with that guaranteed quality. If not, the system will drop the user with the highest resource consumption, meaning that no resources are given to this user until the next optimization loop. After checking the constraint, it performs a usual max-min utility based optimization as described in (15). In the following, we discuss the objective function in (13). Later in our experimental results, we then compare the performance using (13) and (15).

### B. Rate adaptation

At each TTI, a number of data blocks or RLC PDUs are passed from the higher layers to the radio link layer. The size of a data block to be transmitted in one TTI depends on the Channel Quality Indicator (CQI), which is carried via the uplink High Speed-Dedicated Physical Control Channel (HS-DPCCH). The TTI is set to 2ms, meaning that there are 500 TTI slots available in one second period to be shared among users.

The optimizer decides the best combination of all user's operation modes, which maximizes the selected objective function. To assure the data rate of each user, the number of TTI slots must be assigned correctly. Estimation of the required number of TTI is done by using the following equation:

$$S_k = \lceil \frac{A_{app,k} + OH_k}{\bar{B}_k} \rceil \tag{16}$$

where $S_k$ is the number of transmission opportunities to be allocated to user $k$. $A_{app,k}$ is the number of bits to be sent in one second. We assume that the application is sending with a constant bit rate (CBR) during the time interval of interest. $\bar{B}_k$ is the mean size of a transport block. $OH_k$ is the amount of overhead due to transport and network layer headers.

The use of the proposed framework does not exclude the possibility of setting Guaranteed Bit Rate (GBR), Scheduling and Priority Indicator (SPI) and Discard Timer (DT) for quality control, as proposed in [21]. GBR can be set at once as the values out of the optimization or periodically reconfigured during optimization. Setting SPI would be essential in order to ensure delay guarantees. In this paper we assume that the streaming and realtime traffic are prioritized with respect to file download traffic. The exact priority indices would largely depend on the scheduler used. The approach taken in this paper does not rely on any particular scheduling scheme, and hence can be used with any scheduler.

## IV. GREEDY OPTIMIZATION

A greedy algorithm [38] makes a locally optimal choice at each step with the hope of finding the global optimum.

Greedy algorithms are in general not guaranteed to find the optimal solution since they usually do not operate exhaustively through the whole constraint space. Because of this, greedy algorithms are usually much faster than the full search. In this paper we propose a greedy algorithm to solve the utility maximization problem. First, we derive some properties of the constraint space which we call the *utility space*. Next we describe the algorithm in detail. Then we derive the worst case properties of the algorithm and compare it with that of the full search approach.

### A. Properties of the utility space

**Theorem 1.** *Let $P$ be a set of points in the utility space corresponding to $\sum_k \alpha_k(p) = 1$, $P = \{p \in \mathcal{U}\ s.t.\ \sum \alpha_k(p) = 1\}$. Let $\tilde{x}^*$ be the optimum mode of operation: $\tilde{x}^* = \arg\max \sum_k U_k$. Then, $\tilde{x}^* \in P$.*

**Theorem 2.** *The optimum of the objective function, $\tilde{x}^*$ lies on the boundary of the utility space, i.e., $\tilde{x}^* \in B_{\mathcal{U}}$.*

*Proof:* Let $p$ be an interior point of the utility space $\mathcal{U}$, $p \in \mathcal{U}$, $p \notin B_{\mathcal{U}}$ and let $d(x, y)$ denote the Euclidean distance between points $x$ and $y$. Then there exists another point $q \in \mathcal{U}$, $d(q, 0) - d(p, 0) > 0$ such that $\mathcal{F}(p) < \mathcal{F}(q)$. The existence of $q$ is guaranteed until $q$ lies on the boundary of $\mathcal{U}$, i.e., $q \in B_{\mathcal{U}}$. But $\sum_k U_k(p) < \sum_k U_k(q)$, so that an interior point of $\mathcal{U}$ cannot be an optimum. In other words, the optimum must lie on the boundary: $\tilde{x}^* \in B_U$.

**Theorem 3.** *Assume monotonically increasing utility functions, $U_k(\alpha)$ for $\forall k$. Let $P$ be a set of points in the utility space corresponding to $\sum_k \alpha_k = 1$, $P = \{p\ s.t.\ \sum_k \alpha_k = 1\}$. Then $P = B_{\mathcal{U}}$.*

*Proof:* First we show that $P \subseteq B_{\mathcal{U}}$. Let $q \in \mathcal{U}$, $q \in P$ and $q \notin B_{\mathcal{U}}$. Then there exists another point $r \in B_{\mathcal{U}}$ such that $d(r, 0) - d(q, 0) > 0$. Hence, $\sum U(q) < \sum U(r)$ and $U_k(q) < U_k(r)$ for some $k$. Since $U_k(\alpha_i) > U_k(\alpha_j)$ only if $\alpha_i > \alpha_j$ (non-decreasing utility functions), $\sum \alpha(q) < \sum \alpha(r)$ which implies $\sum \alpha(r) > 1$. But then $r \notin \mathcal{U}$ and hence, $r \notin B_{\mathcal{U}}$. Therefore, $q \in B_{\mathcal{U}}$ which implies $P \subseteq B_{\mathcal{U}}$. Similarly, $B_{\mathcal{U}} \subseteq P$ can be proved by using the fact that $\alpha_i > \alpha_j$ only if $U(\alpha_i) > U(\alpha_j)$ (strictly increasing utility functions). $P \subseteq B_{\mathcal{U}}$ and $B_{\mathcal{U}} \subseteq P$ implies that $P = B_{\mathcal{U}}$.

The proof of Theorem 1 follows from results of Theorem 2 and Theorem 3.

*Discussion*: Theorem 1 implies that the optimum of the utility maximization problem lies on the boundary of the utility space, so that a search through the whole utility space is not required. Hence, any algorithm that performs an exhaustive search over the set $B_{\mathcal{U}}$ would eventually find the global optimum.

### B. Algorithm description

We consider a time window of $S_o$ TTI. Let $S_k$ be the number of TTI assigned to user $k$. Then we have, $\sum_{k=1}^{K} S_k \leq S_o$.

The greedy algorithm for the utility maximization is described below. Throughput maximization is performed in a similar fashion. The algorithm is initialized by assigning an amount of resource for every user such that $\sum_{k=1}^{K} S_k = S_o$. At each subsequent iteration a small amount of resources is taken from the user with the lowest sensitivity with respect to decrease of utility and assigned to the user which receives the maximum benefit. This process is repeated until there is no further improvement in the objective function.

Let $U_k$ denote the utility function and $\alpha_k$ the fraction of total TTI assigned to user $k$: $\alpha_k = \frac{S_k}{S_o}$, $\sum_{k=1}^{K} \alpha_k = 1$. We consider a discrete set of $\alpha_k$:

$$\alpha_k \in \{n \cdot \Delta\alpha\ s.t.\ n \in \mathcal{Z}_o, 0 \leq \alpha_k \leq 1\}, \forall k \tag{17}$$

where $\mathcal{Z}_o$ denotes the set of non-negative integers.

Let $\Delta U_k$ denote the change of utility for user $k$ due to a change of its resource share, $\Delta\alpha$. The greedy algorithm can be expressed as an iterative maximization of the incremental utility values of two users $k^+$ and $k^-$, $k^+ \neq k^-$ such that

$$k^+ = \arg\max_k \{\Delta U_k | \alpha_k \leftarrow \alpha_k + \Delta\alpha\} \tag{18}$$

$$k^- = \arg\min_k \{\Delta U_k | \alpha_k \leftarrow \alpha_k - \Delta\alpha\} \tag{19}$$

The greedy algorithm is summarized in Algorithm 1.

### C. Complexity

The worst case complexity of the greedy algorithm described in the previous section depends on the number of users and the granularity of the sampling of $\alpha$. It can be shown that the cardinality of the constraint set, and hence the number of points that have to be searched in the worst case increases with both the number of users and the granularity of sampling. Specifically, it is shown that the cardinality of the constraint set stays constant when the number of users and the number of samples are interchanged.

Let $h$ be the number of possible modes for each user, $h \in \{1, 2, \cdots\}$. We assume the modes to be equally spaced, so that $\Delta\alpha = 1/(h-1)$. Let $P$ be a set of vectors such that $P = \{(p_1 \cdots p_K)'\ s.t. \sum_{k=1}^{K} p_k = h, p_k \in \{0, 1, \cdots, h\}\}$. Then $P$ is the set of points corresponding to $\sum_{k=1}^{K} \alpha_k = 1$. Hence the cardinality of the set $P$, $|P|$ is the worst case number of iterations for the greedy algorithm.

Let $|P| = N_G(h, K)$. Then, $N_G(h, K) =$

$$\begin{pmatrix} h+K-1 \\ K-1 \end{pmatrix} = \begin{pmatrix} h+K-2 \\ K-1 \end{pmatrix} + \begin{pmatrix} h+K-2 \\ K-2 \end{pmatrix}$$

$= N_G(h-1, K) + N_G(h, K-1)$.

This results in a 2D symmetric matrix of $N_G(h, K)$ which implies that we can interchange the number of users with the granularity of the sampling and yet the worst case number of iterations for the algorithm stays constant. This fact can be taken advantage of by using less

**Algorithm 1** Greedy Algorithm

**Input**: Utility function $U$, Transmission policies $T$, number of user $K$, resource budget $S_o$, step size $\Delta\alpha$, increase of step size $\Delta\alpha_{inc}$, minimum expected utility change $\Delta U_{min}$, maximum number of iterations $I_{max}$.
**Output**: Optimal operating mode $\tilde{x}_{opt}$;
**Initialization**: initial resource share: $\alpha = [1, 0, 0, \cdot, 0]$, set $\Delta U_{max,inc}$ to a value greater than $\Delta U_{min}$. Iteration index, $I = 0$.
**for** $k = 1$ to $K$ **do**
  get operating mode $\tilde{x}_k$ from $\alpha_k$, $\tilde{x}_k \in T_k$
  Compute $U_k$
**end for**
**loop**
  **for** $k = 1$ to $K$ **do**
    get operating mode $\tilde{x}_{inc,k}$ from $\alpha_k + \Delta\alpha$, where $\tilde{x}_{inc,k} \in T_k$;
    get operating mode $\tilde{x}_{dec,k}$ from $\alpha_k - \Delta\alpha$, where $\tilde{x}_{dec,k} \in T_k$;
    compute $\Delta U_k(\tilde{x}_{inc,k})$ and $\Delta U_k(\tilde{x}_{dec,k})$;
  **end for**
  **if** $\Delta U_{max,inc} < \Delta U_{min}$ **then**
    set $\Delta\alpha$ to $\Delta\alpha + \Delta\alpha_{inc}$
  **else**
    find $k^+$, $k^-$ using equations 18 and 19
    $\Delta U_{max,inc} = \Delta U_k(\tilde{x}_{inc,k}) - \Delta U_k(\tilde{x}_{dec,k})$
    set $\Delta\alpha$ to $\Delta\alpha_{inc}$
  **end if**
  $I + +$;
  **if** $I > I_{max}$ **then**
    break;
  **end if**
**end loop**
**output**: $\tilde{x}_{opt}$

granularity of sampling as the number of users grow, such that the real-time computation of the optimum remains feasible. In comparison, the number of iterations for a full search is $h^K$ which becomes infeasible when $K \gg 1$.

## V. Evaluation Results

In our simulation, we consider a single cell scenario. We allocate all the resources to HSDPA users. We simulate a 10 user scenario: three voice users, four video streaming users, two FTP users and one video conferencing user. In our simulation we compare five schemes as follows:

1) *No-adaptation*: this is the default HSDPA mode which uses GBR, SPI and DT as QoS parameters. Application-layer rate adaptation is not performed and the system is left to run into overload.
2) *Max-Rate*: System overload is avoided by adapting the multimedia bit rate. Adaptation is done so as to maximize the total cell throughput.
3) *Max-MOS*: Adaptation is done so as to maximize the mean MOS (Quality of Experience) over all users, using both the full-search and the greedy algorithm.
4) *MaxMin-MOS*: With the max-min fairness, the total resources are allocated such that all users experience the same perceived quality (MOS).
5) *MaxMin-MinMOSX.Y-MOS*: Similar to the max-min fairness approach, this scheme first sets a minimum guarantee of MOS *X.Y* for all users and then adapts the resource allocation so as to achieve the same MOS that is equal or higher than the guarantee MOS. If the system cannot provide all users with the guaranteed MOS, a user or more requiring the highest amount of resources is dropped.

TABLE I.
Simulation Parameters

| | |
|---|---|
| Total transmit power | 15.8W |
| Power allocated to HS-DSCH | 11W |
| Carrier Frequency | 2GHz |
| User speed | 3km/h |
| Distance from Node B | 500m – 1.8km |
| UE category | 6 |
| Target BLER | 10% |
| CQI averaging cycle | 1sec |
| RLC PDU size | 40byte |
| Scheduler | Proportional Fair |
| $\mathcal{R}_{vs}$ | $\{0, 30, ..., < 500\}$kbps |
| $\mathcal{R}_{vc}$ | $\{0, 96\}$kbps |
| $\mathcal{R}_{voice}$ | $\{0, 6.4, 15.2, 24.6, 64\}$kbps |
| $\mathcal{R}_{FTP}$ | $\{0, 50, 100, \cdots, 250\}$kbps |
| Video codec used | H.264 |
| Voice codec used | G.723, iLBC, SPEEX, G.711 |
| Loss concealment | Copy previous frame |
| Video/Voice rate shaping | Transcoding |
| $DT_{vs}$, $DT_{FTP}$ | 2sec |
| $DT_{vc}$, $DT_{voice}$ | 150ms |
| Simulator | OPNET 9.1 with NTT DoCoMo HSDPA plugin |

It should be noted that schemes 2) to 5) are application-aware. Optimization is performed every five seconds. The utility function of scheme 2 is:

$$U_k = \bar{R}_k, \forall k \in \mathcal{K} \quad (20)$$

whereas the utility function of scheme 3) to 5) is:

$$U_k = MOS_k(\bar{R}_k), \forall k \in \mathcal{K} \quad (21)$$

where $MOS_k$ is the MOS-based utility function of user $k$.

The parameters used in our simulations are given in Table I. A proportional fair scheduler is used. At the scheduler, we assign lower priority to FTP with respect to other services. A set of possible rates, $\mathcal{R}_{vs}$, $\mathcal{R}_{vc}$, $\mathcal{R}_{voice}$, and $\mathcal{R}_{FTP}$ for video streaming, video conferencing, voice and FTP services, respectively, are chosen as shown in Table I. Discard timer, $DT$ are set as shown in the table.

The evaluation methodology is of particular importance to the quality-aware optimization framework, as we are

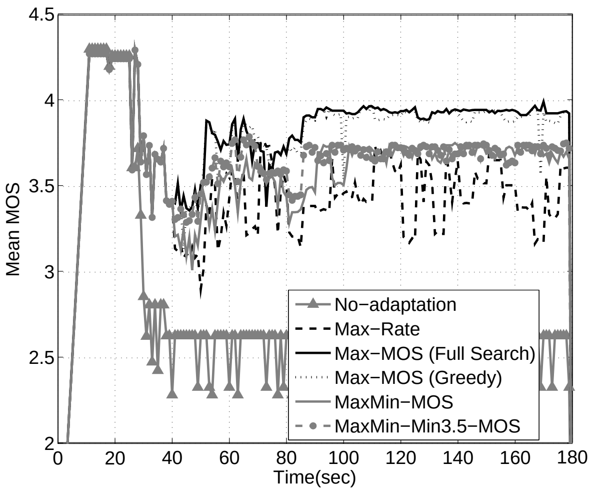


Figure 8. Mean utility for the 10 user case as a function of simulation time.

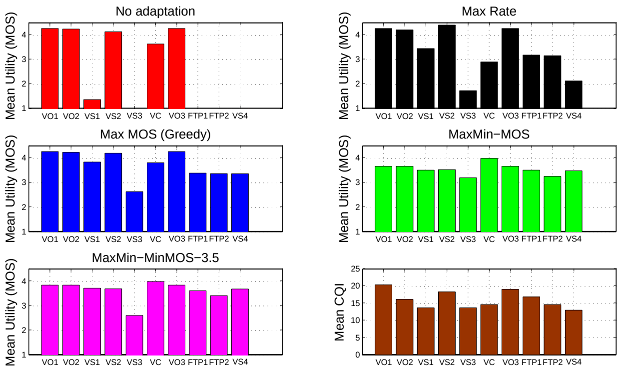


Figure 9. Mean utility and the corresponding mean CQI values for 10 users.

interested in characterizing the system performance in terms of user perceived quality instead of only network-related parameters.

The simulation of a particular scenario produces packet traces which contain the time of generation and arrival of each packet and the chosen rate/operating mode corresponding to the packet. From this information, an offline evaluation is performed. Each media type is encoded into a set of possible rates. The packet trace file is used to infer the rates chosen for each user. Errors introduced to the bit-stream due to late arrival of the packets are simulated using the packet arrival times. This distorted bitstream is then decoded by the audio/video decoder with error-concealment enabled. The distortion between the original input stream and the output distorted stream is measured and converted to MOS following the approach outlined in Sec. II-C. For more details on the evaluation methodology please refer to [28].

Fig. 8 shows the mean utility of all the users over the simulation period of 3 min. From time 10sec to 35 sec, users join the system one by one. The rate-based scheme and all utility-based schemes start at 40 sec. We see a significant performance gain between the no-

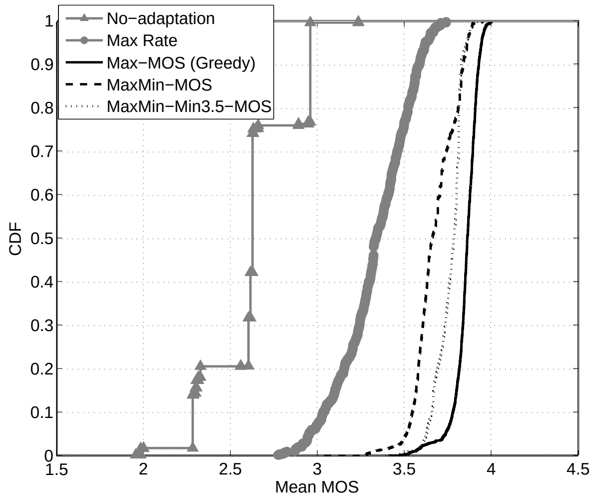


Figure 10. CDF of mean utility for the 10 user case using MSSIM as a video quality assessment

adaptation scheme and the other schemes. It can also be observed from Fig. 8 that the MOS-based utility optimization scheme leads to an additional gain compared to the rate-based scheme. From Figure 9 we see that most of the gain for the *Max-MOS* scheme comes from the users experiencing relatively bad channel conditions and demanding applications (e.g. VS3). For the *MaxMin-MOS* approach, all users experience a similar service quality (around MOS 3.4). When setting a minimum guarantee of service quality with MOS 3.5, the VS3 user suffers the most, since VS3 requires the largest amount of resources to reach a higher service quality and therefore the optimizer does not allocate any resources to this user. Nevertheless, the overall quality of the other users increases due to the resources taken from the user VS3. The average perceived service quality for the two FTP users is slightly lower than for the other users due to the lower priority setting at the scheduler and the TCP slow-start behaviour.

Fig. 10 shows the Cumulative Distribution Function (CDF) of mean MOS over all the users over 300 simulation runs, each consisting of three minutes of simulation time. Out of the three minutes, we take the results of only the last two minutes to avoid the effects of startup. For clarity of the picture, we have left out the results of MaxMOS with full search, since the MaxMOS with greedy optimization performs as good as MaxMOS with full search. We see an average increase of 0.6 MOS for the Max-Rate optimization scheme when compared to the no-adaptation case. Using the MOS-based utility optimization scheme gives a further gain of 0.4 MOS on average. The CDF curves also show that the MOS-based schemes have the lowest dispersion around the mean value, which results in more stable user perceived quality compared to the other schemes. Although using Max-Rate approach would lead to the best result in terms of throughput, it does not guarantee that the quality

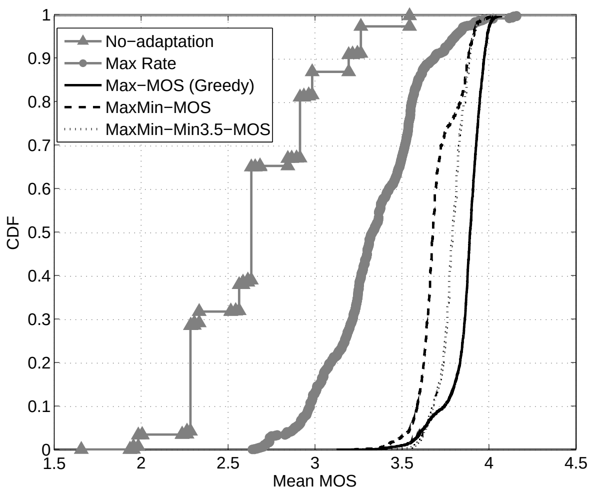


Figure 11. CDF of mean utility for the 10 user case using PSNR as a video quality assessment

perceived by the end-user will be the best. With the MOS-based scheme, the resources are allocated by considering the cost and the gain when giving more or less resources to the users running different multimedia applications.

By having more voice call users in the cell and less users for other applications, the gain of MOS-based scheme is expected to be less due to the smaller number of steps in the voice utility function. This gives us fewer operating points to adjust the data rate in the network, and therefore, less possibility to find an operating point that improves the quality.

All the results that we have discussed so far are based on the MSSIM-based video quality assessment as described in Sec. II-C.3. We have also run simulations for the PSNR-based video quality measurement and the CDF curve depicted in Fig. 11 shows a similar result as for the case of the MSSIM-based CDF. We conclude from the similarity of these two results that whichever video quality assessment type we use, the MOS-based utility optimization schemes outperform the no-adaptation and the rate-based optimization scheme.

## VI. Conclusion

We propose a QoE-driven optimization framework for HSDPA in situations when the total resource of the system is unable to support the system load. Conventionally, this situation is avoided by a strict admission control policy. But by doing this, users would suffer from high blocking probability and operators would loose revenue. We propose that the applications be re-adapted, taking into account the utility functions of the applications. This policy results in better mean quality of experience for given system resources and a fixed number of users, and the admission of more users for a given target quality. The results from both video quality assessment types (structural similarity and error sensitivity (e.g., PSNR)) show that all QoE-based utility optimization schemes (maximization mean MOS, max-min MOS fairness and max-min MOS with minimum MOS guarantee) outperform the no adaptation and the rate-based adaptation. The selection of the particular objective function for QoE-based utility optimization depends on the operator policy. Although the proposed framework is applied to HSDPA due to its current relevance, it can be also integrated into future packet-based systems, e.g. in LTE, as well.

## References

[1] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, October 2003.

[2] M. T. Ivrlac and J. A. Nossek, "Cross layer optimization-an equivalence class approach," in *ITG Workshop Smart Antennas*, Munich, Germany, Mar. 18-19, 2004.

[3] H. Jiang, W. Zhuang, and X. Shen, "Cross-layer design for resource allocation in 3G wireless networks and beyond," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 120–126, December 2005.

[4] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[5] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, September 2004.

[6] Third Generation Partnership Project, "Physical layer aspects of utra high speed downlink packet access," 3GPP, Sophia Antipolis, France, Technical Specification TS 25.848, March 2001, v4.0.0.

[7] H. Ekstrom *et al.*, "Technical solutions for the 3G Long-Term Evolution," *IEEE Communications Magazine*, vol. 44, no. 3, pp. 38–45, March 2006.

[8] X. Liu, E. K. P. Chong, and N. B. Schroff, "Transmission scheduling for efficient wireless utilization," in *IEEE Annual Joint Conference of Computer and Communications Societies (INFOCOM '01)*, Anchorage, Alaska, USA, April 2001, pp. 776–785.

[9] V. Tsibonis, L. Georgiadis, and L. Tassiulas, "Exploiting wireless channel state information for throughput maximization," in *IEEE Annual Joint Conference of Computer and Communications Societies (INFOCOM '03)*, San Francisco, California, USA, 22-26 April 2003, pp. 301–310.

[10] R. Love, A. Ghosh, W. Xiao, and R. Ratasuk, "Performance of 3GPP High Speed Downlink Packet Access (HSDPA)," in *IEEE Vehicular Technology Conference, VTC2004-Fall*, Los Angeles, CA, USA, 26-29 September 2004, pp. 3359–3363.

[11] F. Frederiksen and T. E. Kolding, "Performance and Modeling of WCDMA/HSDPA Transmission/H-ARQ Schemes," in *IEEE Vehicular Technology Conference, VTC 2002-Fall*, Vancouver, Canada, 24-28 September 2002, pp. 472–476.

[12] K. I. Pedersen *et al.*, "Network Performance of Mixed Traffic on High Speed Downlink Packet Access and Dedicated Channels in WCDMA," in *IEEE Vehicular Technology Conference, VTC2004-Fall*, Los Angeles, CA, USA, 26-29 September 2004, pp. 4496–4500.

[13] M. Lundevali *et al.*, "Streaming applications over HSDPA in mixed service scenarios," in *IEEE Vehicular Technology Conference, VTC2004-Fall*, Los Angeles, CA, USA, 26-29 September 2004, pp. 841–845.

[14] P. Ameigeiras, "Packet scheduling and Quality of Service in HSDPA," Ph.D. dissertation, Department of Communication Technology, Aalborg University, October 2003.

[15] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming multimedia users in high-speed downlink packet access (HSDPA)," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 844–855, August 2006.

[16] W. S. Jeon, D. G. Jeong, and B. Kim, "Packet scheduler for mobile internet services using high speed downlink packet access," *IEEE Transaction on Wireless Communications*, vol. 3, no. 5, pp. 1789–1801, September 2004.

[17] G. Manfredi, P. Annese, and U. Spagnolini, "A channel aware scheduling algorithm for hsdpa system," in *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, Germany, 11-14 September 2005, pp. 2136–2140.

[18] N. Nasser, Al-Manthari, and H. Hassanein, "A performance comparison of class-based scheduling algorithms in future umts access," in *IEEE International Performance, Computing and Communications Conference*, Phoenix, Arizona, USA, 7-9 April 2005, pp. 437–441.

[19] G. Liebl, H. Jenkac, T. Stockhammer, C. Buchner, and A. Klein, "Radio Link Buffer Management and Scheduling for Video Streaming over Wireless Shared Channels," in *IEEE Packet Video Workshop*, Irvine, California, USA, December 2004.

[20] B. Wang, K. I. Pedersen, T. E. Kolding, and P. E. Mogensen, "Performance of VoIP on HSDPA," in *IEEE Vehicular Technology Conference, VTC 2005-Spring*, 30 May - 1 June 2005, pp. 2335–2339, vol. 4.

[21] K. I. Pedersen, P. E. Mogensen, and T. E. Kolding, "Overview of QoS options for HSDPA," *IEEE Communications Magazine*, vol. 44, no. 7, pp. 100 – 105, July 2006.

[22] K. I. Pedersen, "Quality based HSDPA access algorithms," in *IEEE Vehicular Technology Conference, VTC-2005-Fall*, Dallas, USA, 25-28 September 2005, pp. 2498–2502, vol. 4.

[23] S. Dixit, Y. Guo, and Z. Antoniou, "Resource management and quality of service in third-generation wireless networks," *IEEE Commun. Mag.*, vol. 39, pp. 125–133, Feb. 2001.

[24] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transaction of Telecommunication*, vol. 8, pp. 33–37, Jan. 1997.

[25] M. Chiang, S. H. Low, R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: a mathematical theory on network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[26] Third Generation Partnership Project, "Physical layer procedures (FDD)," 3GPP, Sophia Antipolis, France, Technical Specification TS 25.214 V7.1.0, June 2006.

[27] A. Saul, S. Khan, G. Auer, W. Kellerer, and E. Steinbach, "Cross-layer optimization using model-based parameter exchange," in *IEEE Int. Conf. on Communications, ICC '07*, Glasgow, Scotland, 24-28 June 2007.

[28] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, "MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication," *Advances in Multimedia*, 2007, article ID 94918.

[29] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," in *ITU-T Recommendation P.862*, February 2001.

[30] J. Janssen, D. D. Vleeschauwer, M. Buchli, and G. H. Petit, "Assessing voice quality in packet-based telephony," *IEEE Internet Computing*, vol. 6, no. 3, pp. 48–56, May/June 2002.

[31] B. Girod, "What's wrong with mean-squared error," *Digital Images and Human Vision*, pp. 207–220, May 1993, the MIT press.

[32] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Orlando, Florida, USA, 13-17 May 2002.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[34] L.-U. Choi, W. Kellerer, and E. Steinbach, "On cross-layer design for streaming video delivery in multi-user wireless environments," *Eurasip Journal on Wireless Communications and Networking*, 2006, vol. 2006, Special Issue on Radio Resource Management in 3G+ Systems, Article ID 60349.

[35] S. Khan, M. Sgroi, Y. Peng, E. Steinbach, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Communications Magazine*, pp. 122–130, Jan. 2006.

[36] B. Radunovic and J.-Y. L. Boudec, "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Trans. on Networking*, vol. 15, no. 5, pp. 1073–1083, Oct. 2007.

[37] A. Saul, "Wireless resource allocation with perceived quality fairness," in *IEEE Annual Asilomar Conference on Signals, Systems, and Computers*, PACIFIC GROVE, CA, USA, 1-4 November 2008.

[38] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to algorithms*. McGraw Hill, 2001.

**Srisakul Thakolsri** is currently a researcher of the ubiquitous networking research group at NTT DOCOMO's European research labs in Munich, Germany, and a Ph.D. candidate at the Technische Universität München, Munich, Germany. He received a Bachelor degree and a Masters degree in electrical engineering from Thammasat University in 1999, and from Technische Universität Kaiserslautern in 2001, respectively. His current research interests are mobile multimedia communication, quality of experience for multimedia applications, signaling for mobile networks.

**Shoaib Khan** received his Bachelor degree from Bangladesh University of Engineering and Technology in 2001 and Masters degree from Technische Universität München in 2003, both in electrical engineering. From December 2003 to December 2008 he was a Member of the Research Staff at the Institute for Media Technology of Technische Universität München, Munich, Germany. His research interests include cross-layer design, multimedia user experience, multimedia streaming and networking. Since January 2009 he is with Keynote SIGOS GmbH based in Nuremberg, Germany.

**Eckehard Steinbach** studied electrical engineering at the University of Karlsruhe, Karlsruhe, Germany, the University of Essex, Colchester, U.K., and ESIEE, Paris, France. He received the Engineering Doctorate from the University of Erlangen-Nuremberg, Germany, in 1999. From 1994 to 2000, he was a Member of the Research Staff of the Image Communication Group, University of Erlangen-Nuremberg. From February 2000 to December 2001, he was a Postdoctoral Fellow with the Information Systems Lab, Stanford University, Stanford, CA. In February 2002, he joined the Department of Electrical Engineering and Information Technology, Technische Universität München, Munich, Germany, as a Professor for Media Technology. His current research interests are in the area of audio-visual-haptic information processing, image and video compression,

error-resilient video communication, and networked multimedia systems.

**Wolfgang Kellerer** is senior manager of the ubiquitous networking research group at NTT DOCOMO's European research labs in Munich, Germany. Before he joined DOCOMO he has worked at the Institute of Communication Networks at Munich University of Technology (TUM). In 2001 he was a visiting researcher at the Information System Laboratory of Stanford University, California. He received a PhD and a Dipl.-Ing. degree in electrical engineering and information technology from the Munich University of Technology in 2002 and 1995, respectively. His current research interests are next generation mobile Internet networks, peer-to-peer overlay networks, mobile multimedia communication, signaling networks and service platforms.