QoS-constrained Energy Minimization in Multiuser Multicarrier systems

Qing Bai, Michel T. Ivrlač, and Josef A. Nossek

Institute for Circuit Theory and Signal Processing,
Technische Universität München, Munich, Germany
bai.qing@nws.ei.tum.de,
Home page: http://www.nws.ei.tum.de/

Summary. In this paper the QoS-constrained resource allocation problem in multicarrier systems is considered. Within the established cross-layer framework, parameters for subchannel assignment, adaptive modulation and coding, and ARQ/HARQ protocols are jointly optimized. Instead of the conventional transmit power minimization, the total energy consumption for the successful transmissions of all information bits is set as the optimization goal. The nonconvex primal problem is solved by using Lagrange dual decomposition and the ellipsoid method. Numerical results indicate that the recovered primal solution is well acceptable in performance, and efficient in terms of computational effort.

Key words: resource allocation, multicarrier systems, cross-layer optimization, energy minimization

1 Introduction

Resource allocation in wireless communication networks is both important and challenging not only because of the scarcity of radio resources and time-variant channel conditions, but also due to the increasing demand to support heterogeneous *quality of service* (QoS) requirements of various applications. From a mathematical point of view, one specific resource allocation corresponds to a mapping from the available radio resources to a set of QoS values. When parameters from different protocol layers are jointly taken into account in the mapping, the optimizations we do, either optimizing QoS with limited resources or minimizing the amount of resources required to achieve certain QoS, are referred to as *cross-layer optimizations*, and the resource allocation itself is termed as *cross-layer assisted resource allocation*. In this paper, a QoS provisioning resource minimization problem at the downlink of a multicarrier system is investigated, where the cross-layer framework adopted integrates PHY and MAC layer functionalities such as subchannel assignment, adaptive modulation and coding, and retransmission protocols.

In most studies on resource allocation for wireless communication systems, the objective for the QoS-constrained resource minimization is to minimize the sum transmit power, *e.g.*, [1][2]. Since retransmission protocols are taken into

account in this work, it is of interest and necessity to consider the transmit power spent over time, *i.e.*, *energy*, instead of to merely consider the power consumption for the first transmission, because on the long run, what is consumed at the transmitter is energy. Based on this analysis, we formulate the minimization goal as the sum energy consumption required to transmit a certain number of information bits within respective latency times for a group of end users.

Though having different physical interpretations, structurally similar optimizations can be found in the literature such as in [3], [4] and [1]. However, due to the discontinuity and nonconvexity of our objective, the methods therein to solve the optimizations and the optimality conditions derived can not be directly applied. Exploiting the discontinuity, we set up a look-up table to lessen the computational burden for the dual methods employed, and a primal recovery scheme is developed to give primarily feasible resource allocations from the obtained dual optimal solutions.

2 System Model

We consider the downlink scenario of an isolated single-cell multicarrier system with K users, each having one data stream to be served. The resource allocation is done on a per *slot* basis, where a *slot* is a short time period of length T during which the wireless channel is assumed to stay constant. As information bits loaded onto consecutive slots are independently modulated and coded, a slot can formally be referred to as a *Transmission Time Interval* (TTI), and the bit-loading procedure inherently includes *packetization* of the information bits. For every TTI, each data stream has a number of information bits to be transmitted, depending on its *throughput* requirement. The other relevant QoS parameter characterizing the data streams, the *latency*, is defined as:
Definition: The latency τ_k of a packet from user k is the delay it experiences until received correctly with an outage probability of no more than the predefined value $\pi^{(\text{out})}$. Let $f_k[m]$ be the probability that it takes exactly m TTI's to transmit a packet error-free, then $\tau_k = (M_k - 1)(\text{RTD} + T) + T$ where RTD represents *round trip delay*, and

$$M_k = \min_{M} M \quad \text{s.t.} \quad \sum_{m=1}^{M} f_k[m] \geq 1 - \pi^{(\text{out})}.$$

In the following subsections, the mathematical descriptions of the regarded system components are derived which lay the basis for cross-layer optimization.

2.1 Channel Model

The downlink broadcast channel is modeled as frequency-selective fading over the total system bandwidth and frequency-flat fading over each *subchannel*, which is consist of N_{c} adjacent subcarriers. FDMA is employed meaning the assignment

of every subchannel is exclusive to one user, and *intercarrier interference* (ICI) is not taken into account. Moreover, we restrict ourselves here to the single-antenna case both at the base station (BS) and at the mobile stations (MS).

Let $H_{k,n}$ and $\sigma^2_{k,n}$ be the channel coefficient and Gaussian noise variance of user k on the nth subchannel, and p_n be the amount of power allocated on subchannel n. When assigned to user k, the *signal-to-noise-ratio* (SNR) on subchannel n can be computed as

$$\gamma_{k,n} = \frac{|H_{k,n}|^2}{\sigma^2_{k,n}} \cdot p_n. \tag{1}$$

Note that throughout this work the index k refers to users and index n refers to subchannels. And as in the remaining part of this chapter, the focus is on any one of the subchannels which is assigned to one user, we drop the subscripts k and n for simplicity.

We choose the TTI to be of length $T = 2$ ms. The WiMAX standard suggests a symbol duration of 102.9 μs in a system with 10 MHz bandwidth and an FFT size of 1024. Based on this number we assume that one TTI contains $N_\mathrm{s} = 16$ symbols for data transmission.

2.2 FEC coding and modulation

We assume that modulation and coding across the subchannels are done independently, and with reference to the WiMAX standard 8 modulation and coding schemes (MCS) are chosen as candidates, which are listed in Table 1.

Table 1. Modulation and Coding Schemes (MCS)

Index	Modulation Type	Alphabet Size A	Code Rate R	$R \log_2 A$
1	BPSK	2	1/2	0.5
2	QPSK	4	1/2	1
3	QPSK	4	3/4	1.5
4	16-QAM	16	1/2	2
5	16-QAM	16	3/4	3
6	64-QAM	64	2/3	4
7	64-QAM	64	3/4	4.5
8	64-QAM	64	5/6	5

Since with the help of cyclic prefix or an equalizer, intersymbol interference is not present in the system, each subchannel can be modeled as a *discrete memoryless channel* (DMC) over which the *noisy channel coding theorem* [5] can be applied. Let the modulation alphabet and the coding rate on the subchannel under consideration be $\mathcal{A} = \{a_1, \ldots, a_A\}$ and R respectively. The *cutoff rate* of the subchannel with SNR γ can be expressed as

$$R_0(\gamma, A) = \log_2 A - \log_2 \left[1 + \frac{2}{A} \sum_{m=1}^{A-1} \sum_{l=m+1}^{A} e^{-\frac{1}{4}|a_l - a_m|^2 \gamma} \right]. \tag{2}$$

The noisy channel coding theorem states that there always exists a block code with block length l and binary code rate $R \log_2 A \leq R_0(\gamma, A)$ in bits per subchannel use, such that with maximum likelihood decoding the error probability $\tilde{\pi}$ of a code word satisfies

$$\tilde{\pi} \leq 2^{-l(R_0(\gamma,A) - R \log_2 A)}. \tag{3}$$

In order to apply this upper bound on code word error probability to the extensively used turbo decoded convolutional code, quantitative investigations have been done in [2] and an expression for the *equivalent block length* is derived based on link level simulations. The result from [2] shows that the performance of a turbo decoded convolutional code applied to a coded packet of length L in a very good approximation equals the performance of a block code with block length

$$n_{\text{eq}} = \beta \ln L, \tag{4}$$

where parameter β is used to adapt this model to the specifics of the employed turbo code, and $L = N_{\text{c}} N_{\text{s}} \log_2 A$. Consequently, the transmission of L bits is equivalent to the sequential transmission of L/n_{eq} blocks of length n_{eq} and has an error probability of

$$\pi = 1 - (1 - \tilde{\pi})^{\frac{L}{n_{\text{eq}}}} \leq 1 - \left(1 - 2^{-n_{\text{eq}}(R_0(\gamma,A) - R \log_2 A)}\right)^{\frac{L}{n_{\text{eq}}}}. \tag{5}$$

2.3 Protocol

At the MAC layer both *automatic repeat request* (ARQ) and *incremental redundancy hybrid ARQ* (IR HARQ) protocols are studied. The data sequence transmitted in one TTI on one subchannel, *i.e.*, a *packet*, is used as the retransmission unit.

ARQ: The corrupted packets at the receiver are discarded. Hence we assume that the *packet error probability* (PEP) of a retransmitted packet is the same as that of its original transmission, *i.e.*,

$$f[m] = \pi^{m-1}(1 - \pi), \quad m \in \mathbb{Z}^+.$$

When the number of transmissions M is given, the maximum allowable PEP can be obtained as

$$\sum_{m=1}^{M} f[m] = 1 - \pi^M \geq 1 - \pi^{(\text{out})} \quad \Rightarrow \quad \pi \leq \sqrt[M]{\pi^{(\text{out})}}.$$

HARQ: The corrupted packets at the receiver are combined and jointly decoded using rate-compatible punctured convolutional codes. For the particular

IR scheme where the retransmissions contain pure parity bits of the same length as the first transmission, the code rate for the mth transmission can be expressed as

$$R[m] = \frac{B}{m \cdot L} = \frac{1}{m} R[1] = \frac{1}{m} R. \tag{6}$$

Let $\tilde{m}$ denote the maximum number of transmissions determined by the mother code. The equivalent block length n_{eq} is then given by

$$n_{\text{eq}} = \beta \ln(\tilde{m} L). \tag{7}$$

Plugging (6)(7) into (5) gives the PEP expression for the mth transmission as

$$\pi[m] = 1 - \left(1 - 2^{-\beta \ln(\tilde{m}L)(R_0(\gamma) - \frac{1}{m} R \log_2 A)}\right)^{\frac{mL}{\beta \ln(\tilde{m}L)}}. \tag{8}$$

When $R_0(\gamma) \leq \frac{1}{m} R \log_2 A$, (8) suggests that $\pi[m] = 1$. And when $R_0(\gamma)$ increases from $\frac{1}{m} R \log_2 A$, $\pi[m]$ approaches 0 very fast. As a result, given the number of transmissions M, $f[m]$ can be approximated by

$$f[M] = 1 - \pi^{(\text{out})}, \quad f[m] = 0, m = 1, \ldots, M-1, \tag{9}$$

where $R_0(\gamma)$ satisfies $\frac{1}{M} R \log_2 A < R_0(\gamma) \leq \frac{1}{M-1} R \log_2 A$.

The quantities mentioned in this section, their notations, as well as their simulation values are summarized in Table 2.

Table 2. System Parameters

Total bandwidth		10 MHz
Center frequency	f_{c}	2.5 GHz
FFT size		1024
Number of data subcarriers		720
Number of subchannels	N	30
Number of subcarriers per subchannel	N_{c}	$720/30 = 24$
Transmission Time Interval (TTI)	T	2 ms
Number of data symbols per TTI	N_{s}	16
Round Trip Delay (RTD)	RTD	10 ms
Maximum number of transmissions allowed	$\tilde{m}$	5
Turbo code dependent parameter	β	32
Outage probability	$\pi^{(\text{out})}$	0.01

3 Problem Formulation

Suppose for the current TTI, the number of information bits intended for user k is b_k, and the maximum latency time for the transmission is $\tau_k^{(\text{rq})}$. The energy minimization problem can be formulated as

$$\begin{aligned} \min_{\boldsymbol{B}} \quad & \sum_{k=1}^{K}\sum_{n=1}^{N} \eta_{k,n}(B_{k,n}, \tau_k^{(\mathrm{rq})}) \\ \text{s.t.} \quad & \sum_{n=1}^{N} B_{k,n} = b_k,\ k = 1, \ldots, K, \\ & \boldsymbol{B} \in \mathcal{B}, \end{aligned} \tag{10}$$

where $\boldsymbol{B} \in \mathbb{Z}_{+,0}^{K\times N}$ represents the bit-loading matrix with its entry $B_{k,n}$ as the number of information bits for the kth user loaded onto the nth subchannel, and $\eta_{k,n}(B_{k,n}, \tau_k^{(\mathrm{rq})})$ is the minimum energy consumption needed for the successful transmission of $B_{k,n}$ bits within the latency time $\tau_k^{(\mathrm{rq})}$. The first constraint in (10) is the completeness of bit-loading for the K users, and the second constraint comes from FDMA in which $\mathcal{B} \subset \mathbb{Z}_{+,0}^{K\times N}$ represents the set of matrices that have only one nonzero entry in each of their columns.

3.1 The η function

We define a tuple (A, R, M) which is a modulation type, FEC code rate, and number of transmissions combination as one *mode of operation*. With 5 as the maximum number of transmissions for each packet and 8 available MCS, we have in all 40 different modes of operation, denoted by set $\mathcal{M}$. For a fixed B, each mode of operation (A, R, M) leads to a (latency, expected energy consumption) pair (τ, E) with

$$\begin{aligned} \tau &= (M-1)(\mathrm{RTD} + T) + T, \\ E &= \left\lceil \frac{B}{R\log_2 A} \right\rceil \cdot T_{\mathrm{s}} \cdot \gamma(A,R,M) \cdot \sum_{m=1}^{M} f[m] \left(\frac{\sigma^2}{|H|^2} + \frac{(m-1)\sigma^2}{|H^{(\mathrm{avg})}|^2} \right) \\ &\overset{!}{=} \phi \cdot \sum_{m=1}^{M} f[m] \left(\frac{\sigma^2}{|H|^2} + \frac{(m-1)\sigma^2}{|H^{(\mathrm{avg})}|^2} \right), \end{aligned}$$

where $|H|^2$ and $|H^{(\mathrm{avg})}|^2$ are the instantaneous and average channel gains, and σ^2 is the noise power on one subcarrier. $\gamma(A, R, M)$ is the SNR required to convey the packet within M transmissions when MCS (A, R) is employed, which can be obtained from a binary search on the cutoff rate curve. Note that ϕ as defined is independent of the channel realizations. $\eta(B, \tau^{(\mathrm{rq})})$ is then given by

$$\eta(B, \tau^{(\mathrm{rq})}) = \min_{(A,R,M)\in\mathcal{M}} E(A,R,M) \quad \text{s.t.} \quad \tau(A,R,M) \le \tau^{(\mathrm{rq})}. \tag{11}$$

Limited by the highest MCS, the number of information bits that can be loaded onto one subchannel in one TTI is upper bounded by $B^{(\mathrm{u})} = 5 \cdot N_{\mathrm{s}} N_{\mathrm{c}}$. Let (τ_1, E_1) and (τ_2, E_2) be two (latency, energy) pairs. Analytical derivations show that if $\tau_1 < \tau_2$ and $\phi_1 < \phi_2$, then $E_1 < E_2$. That means, only those modes of operation that are Pareto efficient in (τ, ϕ)[1] can lead to the solution

[1] A mode of operation $(\tilde{A}, \tilde{R}, \tilde{M})$ is called Pareto efficient in (τ, ϕ) if the pair $(\tilde{\tau}, \tilde{\phi})$ it leads to is Pareto efficient, *i.e.*, there $\nexists (\tau, \phi)$ resulting from other modes of operations in $\mathcal{M}$ such that $\tau \le \tilde{\tau}$ and $\phi \le \tilde{\phi}$.

of (11). Therefore, for each $B \in [1, B^{(\mathrm{u})}]$, finding and storing the (τ, ϕ) Pareto efficient points via an enumeration of all modes of operation are sufficient to solve (11) given the instantaneous channel realizations, which is to say, an offline computable look-up table can be established beforehand. At run time, only some simple calculations are needed to compute $\eta(B, \tau^{(\mathrm{rq})})$. An exemplary η function is shown in Fig. 1, where $\tau^{(\mathrm{rq})}$ is set to infinity.

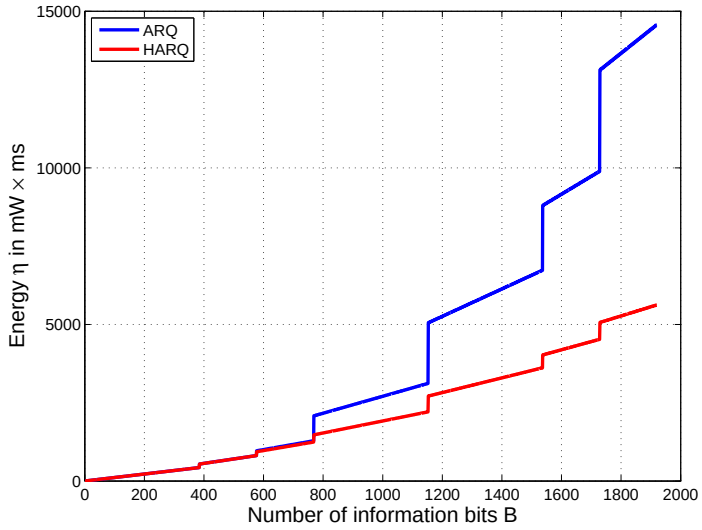


Fig. 1. An exemplary η function for ARQ and HARQ protocols

From the visualization some of the expectations of the η function are verified: it is monotonically increasing with the number of information bits B, the energy increments for the same increment in B become larger with increasing B, and HARQ consumes less energy than ARQ for fairly large B. However, the η function is not convex due to the discrete inputs and changes of the optimum mode of operation at some B. As a result, the optimization (10) is not convex in both objective and constraints. Therefore when dual methods are applied, the solution is bound to suffer from the duality gap.

4 The Resource Allocation Algorithm

4.1 Dual Methods

The Lagrange dual decomposition method and the ellipsoid method are employed to solve the optimization problem (10), following a similar procedure as proposed in [4]. Introducing Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{K \times 1}$ to the K bit-loading constraints in (10) gives the Lagrangian

$$L(\boldsymbol{B},\boldsymbol{\lambda}) = \sum_{k=1}^{K}\sum_{n=1}^{N}\eta_{k,n}(B_{k,n},\tau_k^{(\mathrm{rq})}) + \sum_{k=1}^{K}\lambda_k(\sum_{n=1}^{N}B_{k,n} - b_k), \tag{12}$$

and the dual function $g(\boldsymbol{\lambda})$ follows as

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\boldsymbol{B}\in\mathcal{B}} L(\boldsymbol{B},\boldsymbol{\lambda}) \\ &= \inf_{\boldsymbol{B}\in\mathcal{B}} \sum_{n=1}^{N}\left(\sum_{k=1}^{K}\eta_{k,n}(B_{k,n},\tau_k^{(\mathrm{rq})}) + \sum_{k=1}^{K}\lambda_k B_{k,n}\right) - \sum_{k=1}^{K}\lambda_k b_k \\ &= \sum_{n=1}^{N}\inf_{\boldsymbol{B}\in\mathcal{B}}\sum_{k=1}^{K}\left(\eta_{k,n}(B_{k,n},\tau_k^{(\mathrm{rq})}) + \lambda_k B_{k,n}\right) - \sum_{k=1}^{K}\lambda_k b_k \\ &\overset{!}{=} \sum_{n=1}^{N} g_n(\boldsymbol{\lambda}) - \sum_{k=1}^{K}\lambda_k b_k, \end{aligned}$$

where $g_n(\boldsymbol{\lambda}), n = 1,\ldots,N$ are N independent optimization problems resulting from the decomposition of minimizing $L(\boldsymbol{B},\boldsymbol{\lambda})$. In solving the dual problem, *i.e.*, $\max g(\boldsymbol{\lambda})$, the update of the dual variable $\boldsymbol{\lambda}$ is done efficiently using the ellipsoid method. We denote the optimal value and solution to the dual problem as d^* and $\boldsymbol{\lambda}^*$ respectively, and the bit-loading matrix obtained with $\boldsymbol{\lambda}^*$ as $\tilde{\boldsymbol{B}}$. By weak duality, d^* gives a lower bound on the primal optimal value. However, $\tilde{\boldsymbol{B}}$ is not necessarily primal-feasible, which makes primal recovery necessary.

4.2 Primal Recovery Scheme

Due to the nonconvexity of the objective function of (10), the conclusion drawn in [3] that the duality gap vanishes when the number of subchannels approaches infinity is not valid anymore. Consequently, the subchannel assignment (SA) implicitly given by $\tilde{\boldsymbol{B}}$ ($\tilde{B}_{k,n} > 0$ indicates that the nth subchannel is assigned to the kth user) can not be assumed optimum. In fact, as $B_{k,n}$ is limited by $B^{(\mathrm{u})}$ from above, the dual optimum SA can be infeasible, especially when the total number of information bits to be loaded is large. Therefore, in order to perform primal recovery based on the dual optimum SA, we have to assure its feasibility first.

The minimum number of subchannels needed by user k can be computed as $N_k^{(\mathrm{l})} = \left\lceil \frac{b_k}{B^{(\mathrm{u})}} \right\rceil$. Let the set of subchannels assigned to user k by the dual optimum SA be $\mathcal{S}_k$, *i.e.*, $\mathcal{S}_k = \{n : \tilde{B}_{k,n} > 0\}$. If $\exists k$ with $|\mathcal{S}_k| < N_k^{(\mathrm{l})}$, then the dual optimum SA is infeasible. Denote the set of users with $|\mathcal{S}_k| > N_k^{(\mathrm{l})}$ as $\mathcal{K}_{\mathrm{o}}$[2]. One adjustment scheme can be to solve

[2] An empty set $\mathcal{K}_{\mathrm{o}}$ indicates the infeasibility of (10), the case of which should be tested and excluded at the beginning of the whole program. In order to provide the resource allocation entity with appropriate traffic loads, a scheduling component on its top is necessary.

$$(k^*, n^*) = \underset{k' \in \mathcal{K}_o, n \in \mathcal{S}_{k'}}{\operatorname{argmin}} \Big(\eta_{k,n}(B^{(\mathrm{u})}, \tau_k^{(\mathrm{rq})}) - \eta_{k',n}(B^{(\mathrm{u})}, \tau_{k'}^{(\mathrm{rq})}) \Big),$$

and reassign subchannel n^* to user k instead of its former possessor k^* by updating $\mathcal{S}_k$ and $\mathcal{S}_{k^*}$ accordingly.

Fixing the feasible SA, we have K decoupled minimization problems, one for each user, as

$$\begin{aligned} &\min_{\{B_{k,n}: n \in \mathcal{S}_k\}} \sum_{n \in \mathcal{S}_k} \eta_{k,n}(B_{k,n}, \tau_k^{(\mathrm{rq})}) \\ &\text{s.t.} \quad \sum_{n \in \mathcal{S}_k} B_{k,n} = b_k, \end{aligned} \tag{13}$$

which can again be solved in the dual domain. Let the dual optimal bit-loading be $\{B_{k,n}^* : n \in \mathcal{S}_k\}$. If $\sum_{n \in \mathcal{S}_k} B_{k,n}^* \neq b_k$, we can load or unload the extra bits one by one on the subchannel that leads to the minimum energy increment or the maximum energy decrement. Mathematically, we iteratively find

$$n^* = \begin{cases} \underset{n \in \mathcal{S}_k}{\operatorname{argmin}} \Big(\eta_{k,n}(B_{k,n}^* + 1, \tau_k^{(\mathrm{rq})}) - \eta_{k,n}(B_{k,n}^*, \tau_k^{(\mathrm{rq})}) \Big), & \sum_{n \in \mathcal{S}_k} B_{k,n}^* < b_k, \\ \underset{n \in \mathcal{S}_k}{\operatorname{argmax}} \Big(\eta_{k,n}(B_{k,n}^*, \tau_k^{(\mathrm{rq})}) - \eta_{k,n}(B_{k,n}^* - 1, \tau_k^{(\mathrm{rq})}) \Big), & \sum_{n \in \mathcal{S}_k} B_{k,n}^* > b_k, \end{cases} \tag{14}$$

and update B_{k,n^*}^*, until $\sum_{n \in \mathcal{S}_k} B_{k,n}^* = b_k$ is satisfied. Such a recovery scheme is simple, but greedy and performance-degrading.

5 Simulation Results

For simulations, $K = 10$ users uniformly located in a cell of radius 2 km are assumed. The wireless channel is modeled as a frequency-selective fading channel consisting of six independent Rayleigh multipaths with an exponentially decaying power profile. The delay spreads are uniformly distributed within 1 μs, resulting in a rms delay spread of about 0.3 μs which is consistent with the assumed channel coherence bandwidth. The path loss in dB is computed as $PL(d) = 140.6 + 35.0 \log_{10} d$ following the COST-Hata model, where d is the distance between MS and BS in km, and the receiver noise level is assumed to be -174 dBm/Hz.

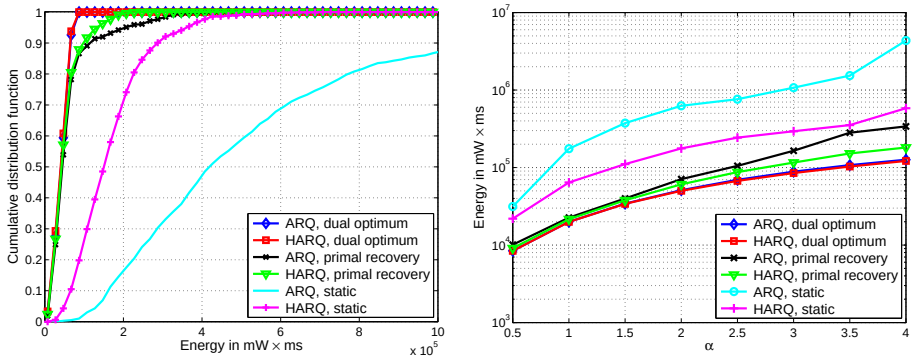
Each user's information bits to be served and latency requirements are listed in Table 3, where the unit for b_k is bit and the unit for τ_k is ms, and α is a scalar that takes values from $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$. Besides the test results of the algorithm discussed in the previous sections, a static resource allocation scheme is simulated for comparison purpose. The static scheme first assigns each user with a fixed set of subchannels and then performs the greedy bit-loading, in the same way as used for primal recovery. Each test scenario has been simulated under 1000 independent channel realizations.

In Fig. 2 the statistics of energy consumptions are shown, where Fig. 2(a) shows the cumulative distributions of the energy spent under different retransmission protocols and resource allocation schemes, and Fig. 2(b) illustrates the

Table 3. QoS requirements of 10 users for simulation

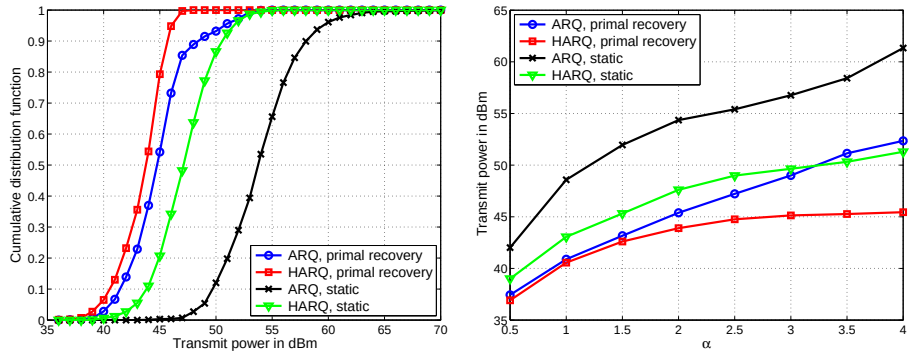
User	b_k	τ_k	User	b_k	τ_k	User	b_k	τ_k
1-4	$512{\cdot}\alpha$	20	5-7	$800{\cdot}\alpha$	40	8-10	$1600{\cdot}\alpha$	80

average energy consumption over 1000 simulations at each α value. Note that the actual optimal energy curves lie between the dual optimum and the primal recovery curves. The corresponding statistics for the transmit power spent for the first transmission are drawn in Fig. 3.



(a) CDF of energy at $\alpha = 2$ (b) Average energy at different α's

Fig. 2. Energy Comparisons



(a) CDF of transmit power at $\alpha = 2$ (b) Average Tx power at different α's

Fig. 3. Transmit Power Comparisons

It is clear from the figures that the algorithm developed in this paper greatly outperforms the static resource allocation scheme, and HARQ protocol outperforms ARQ, in reducing energy consumption as well as transmit power consumption. The higher density the traffic has, the more obvious the advantages. However, with increasing traffic density, the deviation from the primal recovered objective to the dual optimum also gets larger, *e.g.*, the ratio of the deviation to the dual optimum increases from 8% at $\alpha = 0.5$ to 50% at $\alpha = 4$ on average for the HARQ case. On the one hand, this could be caused by possibly larger optimal duality gaps at higher traffic densities, while on the other hand, the more frequent situation at higher traffic densities that infeasible SA is obtained from solving the dual problem, which has to be heuristically adjusted, may deteriorate the performance of primal recovery and in turn, deteriorate the performance of the whole algorithm.

6 Conclusions

A novel energy minimization problem for QoS provisions in multicarrier systems has been formulated and solved, within a cross-layer framework that involves adaptive modulation and coding and retransmission protocols. By using the cutoff rate theorem, the channel and user independent parameters are connected to the time-varying resource allocation parameters with the SNR, which provides the means to reducing computations by setting up offline-computable look-up tables. Though the algorithm has been proved efficient at low to medium traffic densities, there are more issues to be studied: first, the optimal duality gap of the optimization should be estimated; second, more delicate primal recovery schemes are necessary to further improve system performance; last but not the least, the transmit power constraint at the BS should be integrated into the optimization. The three-fold work will be left to our future research.

References

1. C. Y. Wong, R. S. Cheng, K. B. Letaief and R. D. Murch, *Multiuser OFDM with adaptive subcarrier, bit and power allocation*, in IEEE Journal on Selected Areas of Communications, pp. 1747-1758, vol. 17, 1999.
2. B. Zerlin, M. T. Ivrlač, W. Utschick, J. A. Nossek, I. Viering and A. Klein, *Joint optimization of radio parameters in HSDPA*, in IEEE 61st Vehicular Technology Conference VTC 2005-Spring, vol. 1, pp. 295-299.
3. W. Yu and R. Lui, *Dual Methods for Nonconvex Spectrum Optimization of Multicarrier Systems*, in IEEE Transactions on Communications, July 2006, vol. 54, no. 7, pp. 1310-1322.
4. K. Seong, M. Mohseni and J. M. Cioffi, *Optimal Resource Allocation for OFDMA Downlink Systems*, in IEEE Int. Symposium on Information Theory, July 2006, pp. 1394-1398.
5. R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Son, 1968.