

User Allocation in Multi-System, Multi-Service Scenarios: Upper and Lower Performance Bound of Polynomial Time Assignment Algorithms

Ingmar Blau and Gerhard Wunder¹

Fraunhofer German-Sino Mobile Communications Lab, Heinrich-Hertz-Institut
Einstein-Ufer 37, D-10587 Berlin, Germany
{blau,wunder}@hhi.fhg.de

Abstract—In this paper we cover the problem of how users of different service classes should be assigned to a set of radio access technologies (RAT). All RAT have overlapping coverage and the aim is to maximize a weighted sum of assignable users. Under the constraint that users cannot be split between multiple air-interfaces the problem is identified as NP-complete. In the first part of the paper we derive upper and lower bounds of polynomial assignment algorithms. Using Lagrangian theory and continuous relaxation we show for polynomial assignments that in scenarios with M air-interfaces there are at most M users less assigned than in the optimum solution. In the second part we present an algorithm and compare its performance to standard load-balancing strategies.

I. INTRODUCTION

We address the problem of assigning users of different service classes to a set of orthogonal resources to maximize a certain utility function. This relates to a scenario when an operator offers a set of services on multiple radio access technologies (RAT) with overlapping coverage and mobiles can cope with these standards. In this case operators have the freedom to apply assignment algorithms that select a suitable RAT for each user [1]. Of great interest are assignment strategies that maximize a weighted sum of assigned users under the constraint, that users cannot be split between multiple air-interfaces or assigned only partially. Under these constraints the problem can be identified as the general assignment problem (GAP) which is NP-complete and no algorithms are known, that solve it in polynomial time [2]. Due to the exponentially growing computational effort, applying suboptimal algorithms is therefore often inevitable, however, little is known about the performance loss compared to the optimal solution.

In this paper we present an algorithm based on continuous relaxation. The relaxed problem is a Linear Program for which convergence to the optimum can be achieved in polynomial time. Further upper bounds on the optimal solution and a lower bound (upper bounds on the performance loss of the suboptimal algorithm) are derived using Lagrangian theory. The presented algorithm assigns at most M users less than the optimum one, making the lower bound with M air-interfaces, I users and $M \ll I$ astonishingly tight. The performance of the proposed algorithm is compared to a standard multi-

RAT load-balancing algorithm and shows considerable improvement of the utility.

II. SYSTEM MODEL

Notations: In the following small and large bold fonts denote vectors and matrices respectively. Calligraphic letters are used for sets and $|\mathcal{S}|$ is the cardinality of set \mathcal{S} . Matrix transpose relates to $(\cdot)^T$ and $\lceil x \rceil$ ($\lfloor x \rfloor$) denote the closest greater (smaller) integer value to x .

We consider a scenario where a set of users $\mathcal{I} = \{1, \dots, I\}$ is in the coverage area of a set of different radio access technologies $\mathcal{M} = \{1, \dots, M\}$ and mobiles are able to support all technologies. Interference between different systems is precluded since each air-interface is assumed to use an individual non overlapping frequency band. In general, resources like bandwidth and transmission power of air-interfaces and/or users are restricted. Depending on the access technology operators can distribute them in different granularity between users, adapted to users channel gains and service requests. In CDMA systems with fixed spreading gain, a maximum sum power can be split continuously between users in downlink direction. In orthogonal transmission schemes like GSM, there is only a finite number of time-slots that can be assigned to meet users service requests. Systems, which allow more than one degree of freedom in assigning resources, like OFDM, where a set of subcarriers and transmission power can be tuned, are beyond the scope of this paper. In our setting, each user requests a service, which corresponds to a guaranteed transmission rate. If user i is assigned to air-interface m a certain percentage $c_{i,m}$, $i, m \in \mathcal{I}, \mathcal{M}$ of the available resource of this RAT is needed to fulfill its service request. In general this relative resource cost depends not only on the service class and channel of the user but also on the characteristics of the air-interface. Different granularity of distributable resources, modulation and coding schemes, robustness to interference and the system load make costs of each user dependent on the transmission technology. In air-interfaces with non-orthogonal resources the costs usually depend on the interference of other users in the system. In this class of systems we restrict our

¹The authors are supported in part by the *Bundesministerium für Bildung und Forschung (BMBF)* under grant FK 01 BU 566

model to air-interfaces where the costs depend only on the sum of interference and not on its compilation. Otherwise, finding a feasible subset in a single-system scenario that maximizes the number of assigned users is a combinatorial problem, which is general expensive to solve. Interference tolerating single-cell SISO scenarios with e.g. a sum power constraint and a SIR of the form

$$SIR_{i,m} = \frac{h_{i,m}P_{i,m}}{\sum_{i \neq j} h_{i,m}P_{j,m} + \sigma^2}. \quad (1)$$

apply to our model. Here $h_{i,m}$ denotes the channel gain, $P_{i,m}$ the power of user i in air-interface m and $\sum_{i \in \mathcal{I}} P_{i,m} \leq P_{max,m}$. The noise variance is represented by σ^2 . For such a model, given a minimum SIR $\gamma_{i,m}$ that corresponds to the requested rate r_i of user i , (1) can be solved for $P_{i,m}$ depending on the totally assigned power $P_{tot,m}$. Then, assuming a fully loaded system with $P_{tot,m} = P_{max,m}$ we obtain the resource cost

$$c_{i,m} = \frac{P_{i,m}}{P_{max,m}} = \frac{1 + \sigma^2/(h_{i,m}P_{max,m})}{1/\gamma_{i,m} + 1}, \quad (2)$$

For this class of systems $c_{i,m}$ represents the needed percentile of available transmission power in a fully loaded system.

In TDMA systems with a maximum number of time-slots S_{max} with fixed length and under the constraint that users cannot share time-slots the costs can be calculated by

$$c_{i,m} = \frac{1}{S_{max}} \lceil r_i/R(P_{max,m}h_{i,m}/\sigma^2) \rceil. \quad (3)$$

Here $R(\cdot)$ gives the rate corresponding to the SIR. After calculating the costs of all users in all RAT we define an $M \times I$ cost matrix \mathbf{C} and $[\mathbf{C}]_{i,m} = c_{i,m}$. The association of each user with the set of air-interfaces is represented in the $M \times I$ assignment matrix \mathbf{V} , with elements

$$v_{i,m} = \begin{cases} 1 & \text{if user } i \text{ is assigned to air-interface } m \\ 0 & \text{else.} \end{cases} \quad (4)$$

III. OPTIMIZATION PROBLEM AND RELAXATION

In our setup an operator receives service requests from a set of users and can assign them to a set of orthogonal air-interfaces. The assignment should be performed to maximize an operator's utility function. The utility function is restricted to a weighted sum of assigned users. This relates to maximizing the total number of assigned users, when all weights are equal, or allows prioritization of a certain service or user class. From a cross-layer perspective the weights could also represent the coupling between the physical and higher layers. In multi-system scenarios, due to the separated architecture and enhanced signaling effort, it is in general not an option to split service requests and assign users to multiple air-interfaces at the same time. Also, due to minimum QoS requirements, assigning users only partially is not desired. Under these

premises the optimization problem can be written as follows:

$$\begin{aligned} y_{opt} &= \max_{\mathbf{V}} \sum_{i,m \in \mathcal{I}, \mathcal{M}} w_i v_{i,m} \\ \text{subj. to} & \sum_{i \in \mathcal{I}} v_{i,m} c_{i,m} \leq 1 \quad \forall m \in \mathcal{M} \\ & \sum_{m \in \mathcal{M}} v_{i,m} \leq 1 \quad \forall i \in \mathcal{I} \\ & v_{i,m} \in \{0, 1\} \quad \forall i, m \in \mathcal{I}, \mathcal{M} \end{aligned} \quad (5)$$

Here w_i denotes the weight of user i . The first set of constraints in (5) assures that no more resources than available are assigned to each air-interface, the second one prevents multiple assignment of each user. In order to avoid splitting of users the third constraint is used. The latter constraint however leads to the combinatorial nature of the problem for which complexity grows exponentially with the degrees of freedom. Problem (5) can be identified as the Generalized Assignment Problem (GAP), a generalization of the Multiple Knapsack Problem (MKP), which is NP-complete and APX-hard¹ [2]. Thus, using suboptimum algorithms is often inevitable.

In our case problem (5) can be transformed into a convex optimization if we relax the third constraint and allow fractional assignment of users [3]. We then obtain the following representation:

$$\begin{aligned} y^* &= \max_{\mathbf{V}} \sum_{i,m \in \mathcal{I}, \mathcal{M}} w_i v_{i,m} \\ \text{subj. to} & \sum_{i \in \mathcal{I}} v_{i,m} c_{i,m} \leq 1 \quad \forall m \in \mathcal{M} \\ & \sum_{m \in \mathcal{M}} v_{i,m} \leq 1 \quad \forall i \in \mathcal{I} \\ & v_{i,m} \geq 0 \quad \forall i, m \in \mathcal{I}, \mathcal{M}, \end{aligned} \quad (6)$$

with \mathbf{V}^* the optimum relaxed assignment. An efficient algorithm to solve the relaxed problem above is presented in section VI.

Due to the relaxation the optimal solution will probably contain users which are assigned only partially or split between multiple air-interfaces. A feasible solution of the integer problem (5) can now be constructed by not assigning any user with $0 < v_{i,m} < 1$

$$\tilde{v}_{i,m} = \begin{cases} 1, & \text{if } v_{i,m}^* = 1 \\ 0, & \text{else} \end{cases} \quad (7)$$

resulting in

$$\tilde{y} = \sum_{i,m \in \mathcal{I}, \mathcal{M}} w_i \tilde{v}_{i,m}. \quad (8)$$

¹APX-hard means that there does not even exist an approximation scheme that comes arbitrarily close to the optimum value in polynomial time. For the GAP only a 2-approximation exists. This means the tightest lower bound of a polynomial approximation is half of the optimum solution.

IV. BOUNDS

A. Upper Bound

The following statements can be easily derived:

- 1) The relaxed problem (6) optimizes over an extended space \mathbf{V} and this space includes all possible solutions of (5). Therefore its solution is an upper bound of the combinatorial problem.
- 2) If $v_{i,m}^* \in \{1, 0\} \forall i, m \in \mathcal{I}, \mathcal{M}$, then the maximum solution to (6) is also the optimum solution to (5).
- 3) Assume $w_i = 1 \forall i \in \mathcal{I}$. Since \mathbf{V}^* is an upper bound and the solution of (5) is integer, no better solution than the largest integer inside the feasible solutions of (6) can exist. Then (8) is the optimum solution to (5) if

$$\tilde{y} = \lfloor y^* \rfloor. \quad (9)$$

The upper bound of the problem gives valuable information on the quality of a suboptimal solution. Nevertheless, it offers little insight into the general performance of a polynomial time algorithm. Therefore a lower bound on the performance of a suboptimal algorithm is of interest. It can be obtained if an upper bound on the number of users with $0 < v_{i,m}^* < 1$ is found.

B. Lower Bound

In order to derive a lower bound we exploit the Lagrange function and the Karush-Kuhn-Tucker (KKT) conditions [4]. The Lagrange function to problem (6) is

$$\begin{aligned} L(\mathbf{V}, \boldsymbol{\lambda}) = & \sum_{i,m \in \mathcal{I}, \mathcal{M}} w_i v_{i,m} - \sum_{m \in \mathcal{M}} \lambda'_m \left(\sum_{i \in \mathcal{I}} c_{i,m} v_{i,m} - 1 \right) \\ & - \sum_{i \in \mathcal{I}} \lambda''_i \left(\sum_{m \in \mathcal{M}} v_{i,m} - 1 \right) + \sum_{i,m \in \mathcal{I}, \mathcal{M}} \lambda'''_{i,m} v_{i,m}, \end{aligned} \quad (10)$$

where λ are the non-negative dual variables. Since (6) is a convex problem and strong duality holds the KKT conditions are necessary and sufficient for the optimum solution:

$$\frac{\partial L(\mathbf{V}, \boldsymbol{\lambda})}{\partial v_{i,m}} = w_i - \lambda'_m c_{i,m} - \lambda''_i + \lambda'''_{i,m} = 0 \quad \forall i, m \in \mathcal{I}, \mathcal{M} \quad (11)$$

$$\lambda'_m \left(\sum_{i \in \mathcal{I}} c_{i,m} v_{i,m} - 1 \right) = 0 \quad \forall m \in \mathcal{M} \quad (12)$$

$$\lambda''_i \left(\sum_{m \in \mathcal{M}} v_{i,m} - 1 \right) = 0 \quad \forall i \in \mathcal{I} \quad (13)$$

$$\lambda'''_{i,m} v_{i,m} = 0 \quad \forall i, m \in \mathcal{I}, \mathcal{M} \quad (14)$$

Now the following is true at the optimum:

Proposition 1: Assume $c_{i,m}/w_i \neq c_{j,m}/w_j \forall \{i, m\} | i \in \mathcal{I}, i \neq j, m \in \mathcal{M}$, then there are at most M partially assigned users with $0 < \sum_{m \in \mathcal{M}} v_{i,m} < 1$.

Proof: Assume user i is only partially assigned to the set of all air-interfaces. For (13) and (14) to be true $\lambda''_i = 0$ and $\lambda'''_{i,m} = 0$ for at least one air-interface. Substituting this into (11) leads to

$$1 = \lambda'_m c_{i,m}/w_i. \quad (15)$$

Due to the assumption of non equal weighted costs the condition above can only be true for at most one user per air-interface and thus M in total \blacksquare

If multiple users have the same costs in a RAT it is easy to see that multiple assignments, resulting in the same optimum, exist and there is always one with at most M partially assigned users.

In many air-interfaces the channel gain of a user directly influences its resource costs. Then, due to uncorrelated fading and unequal pathloss exponents in different air-interfaces it is a valid assumption that all entries of \mathbf{C} are independently drawn from a set of random distributions. This assumption is helpful if the rank of a matrix with $c_{i,m}$ as entries is needed. We refer to this model as randomly afflicted costs.

Proposition 2: If all entries of \mathbf{C} are randomly afflicted, then there are at most $M - 1$ users with resources assigned to more than one air-interface.

Proof: Assume user u is split between two air-interfaces m_u and n_u . In this case $v_{u,m_u}, v_{u,n_u} > 0$ and $\lambda'''_{u,m_u} = \lambda'''_{u,n_u} = 0$. Substituting this into (11) leads to

$$w_i - \lambda''_u = \lambda'_{m_u} c_{u,m_u} = \lambda'_{n_u} c_{u,n_u}. \quad (16)$$

Thus, a user can only be split if its costs weighted by λ' are equal in multiple air-interfaces. Now assume there is a subset $\mathcal{U} \subseteq \mathcal{I}$ of split users (to two RAT). We define $\mathbf{C}_{\mathcal{U}}$ to be a $|\mathcal{U}| \times M$ matrix where each row has two non-zero entries, c_{u,m_u} in the m_u^{th} and $-c_{u,n_u}$ in the n_u^{th} column. Then (16) can be extended to matrix form

$$\mathbf{C}_{\mathcal{U}} \boldsymbol{\lambda}' = \mathbf{0}, \quad (17)$$

with $\boldsymbol{\lambda}' = \{\lambda'_1, \dots, \lambda'_M\}^T$. Equation (17) has only a non-trivial solution if $\text{rank}(\mathbf{C}_{\mathcal{U}}) \leq M - 1$. Since we assumed that all resource costs are randomly afflicted matrix $\mathbf{C}_{\mathcal{U}}$ has full rank and there can be at most $M - 1$ split users. In the example above users were split between two air-interfaces only. It can be seen, however, that the rank argumentation is also valid if users are split between more than two air-interfaces. \blacksquare

From Proposition 1 and 2 we can conclude that there are at most $2M - 1$ split or partially assigned users, if the matrix $\mathbf{C}_{\mathcal{U}}$ has full rank.

The assumption of randomly afflicted costs is not suitable for all air-interfaces. In TDMA systems with fixed time-slot length like GSM resource costs are bound to a finite set of discrete values. In this case the independence of all rows in $\mathbf{C}_{\mathcal{U}}$ can not be guaranteed anymore and there can exist an arbitrary number of split users in the optimum solution. Then, however, an equivalent assignment with at most $M - 1$ split users, that results in the same optimum, can always be found. This is shown next:

Proposition 3: Assume \mathbf{V}^* maximizes (6) and \mathcal{U}^* is the corresponding set of split users with $|\mathcal{U}^*| > M - 1$. Then there always exists a feasible assignment $\mathbf{V}^\#$ resulting in the same optimum with $\mathcal{U}^\#$ and $|\mathcal{U}^\#| \leq M - 1$.

Proof: First assume an optimal solution of (6) has $|\mathcal{U}| > M - 1$ split users, each assigned to two air-interfaces.

Equation (11) and (13) are satisfied regardless of the proportions users are split. In addition a feasible solution for (12) exists. Following the proof of Proposition 2 we construct $\mathbf{C}_{\mathcal{U}^*}$. Now we can rewrite (12) to

$$r_m = 1 - \sum_{i \notin \mathcal{U}^*} c_{i,m} v_{i,m}^* = \sum_{i \in \mathcal{U}^*} c_{i,m} v_{i,m}^* \quad \forall m \in \mathcal{M} \quad (18)$$

and reformulate the equation above to matrix form:

$$\mathbf{r}' = \mathbf{C}_{\mathcal{U}^*}^T \mathbf{v} \quad (19)$$

Here \mathbf{r}' is a non-negative $M \times 1$ vector with $r'_m = r_m + k_m$, k_m a constant. The $\mathcal{U}^* \times 1$ vector \mathbf{v} contains one $v_{i,m}$ of each user in \mathcal{U}^* . Using the fact that $\text{rank}(\mathbf{C}_{\mathcal{U}^*}^*) \leq M-1$, an infinite number of solutions to the equation above and therefore to (12) exists with at least $|\mathcal{U}^*| - M + 1$ degrees of freedom that result in the optimum utility. Since (12) is linear all solutions lie on a hyperplane in an $|\mathcal{U}^*|$ dimensional space, each dimension representing the assignment $v_{i,m}$ of the split users. This is illustrated for $|\mathcal{U}^*| = M = 3$ in Figure 1 for one degree of freedom on the left hand side and two degrees of freedom on the right one. Any piercing point of the hyperplane with the plotted cube, edge length one, in the figure is a feasible solution where a user is assigned or erased completely from a RAT (and therefore erased or assigned completely to the complementary RAT). Piercing points of the hyperplane with an edge of the cube reduce the number of split users to one. Each hyperplane has to cut the cube because otherwise all split users in the air-interface could be assigned completely and resources would be left unused, which cannot be optimum. In case some users in \mathcal{U}^* are split to more than two air-interfaces we can extend \mathcal{U}^* and $\mathbf{C}_{\mathcal{U}^*}$ by pseudo users. Without loss of generality assume user u is split between 3 air-interfaces m, n, l . Then we can represent user u in \mathcal{U}' by u_1, u_2 with $v_{u_1,m} = t, 0 < t < 1$ and $v_{u_2,n} + v_{u_2,l} + t = 1$. In $\mathbf{C}_{\mathcal{U}'}$ user u appears by the constraints $\lambda'_m c_{u_1,m} = \lambda'_n c_{u_1,n}$ and $\lambda'_n c_{u_2,n} = \lambda'_l c_{u_2,l}$ respectively. Using the argumentation from above equivalent to (19) one can find a representation

$$\mathbf{r}'' = \mathbf{C}_{\mathcal{U}'}^T \mathbf{v}', \quad (20)$$

where \mathbf{v}' is a $\mathcal{U}' \times 1$ vector and each element is a linear function of the elements of \mathbf{v} and t . The remainder of the proof is equivalent to the case, when users are only split between two RAT. ■

Based on the precedent observations, we know that there always exists an optimum solution to the relaxed problem with at most $2M - 1$ users that are partially assigned or split. Next we show how we can further reduce this number.

Proposition 4: If a user is split between n air-interfaces, then there is always an optimum solution where only one of them has a partially assigned user.

Proof: We proof by contradiction: Assume there is an optimum solution where user u is split between two RAT m, n and both RAT have an additional partially assigned user p, q . Then we can reformulate (12)

$$\begin{aligned} r_m &= c_{u,m} v_{u,m} + c_{p,m} v_{p,m} \\ r_n &= c_{u,m} (1 - v_{u,m}) + c_{q,n} v_{q,n}, \end{aligned} \quad (21)$$

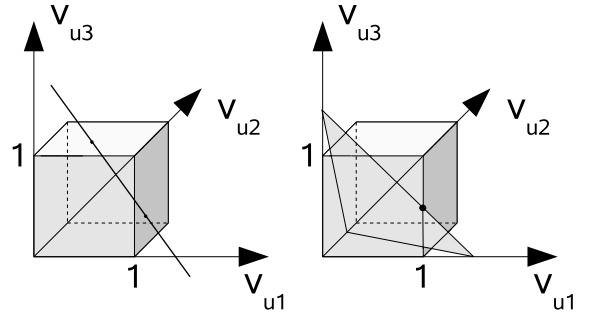


Fig. 1. All optimum solutions lie on the line segment (left) or hyperplane(right) inside the cube. Any piercing point of the cube is a feasible solution with less split users. Left: 1 degree of freedom. Right: 2 degrees of freedom

where $r_m = 1 - \sum_{i:v_{i,m}=1} c_{i,m}$ represents the left resource for the split and partially assigned user in RAT m . The share of utility corresponding to the observed users

$$y_{u,p,q} = w_u v_{u,m} + w_u (1 - v_{u,m}) + w_p v_{p,m} + w_p v_{q,n} \quad (22)$$

is maximum at $v_{u,m}$. Substituting (21) into (22) we get

$$y_{u,p,q} = k + \left(\frac{c_{u,n}}{c_{q,n}} - \frac{c_{u,m}}{c_{p,m}} \right) v_{u,m}. \quad (23)$$

The equation above is either independent of $v_{u,m}$ (then it can be set to one) or $v_{u,m} = \{0, 1\}$ which is a contradiction to the assumption that user u is split. Using the same arguments the proof can be extended if users are split between more than two RAT. ■

In general Proposition 4 states that for each split user at least one partially assigned disappears (or can be erased) from the optimum solution. We can now formulate the lower bound of our polynomial algorithm by summarizing the precedent Propositions:

Theorem 1: There is always a solution to the relaxed optimization problem (6) with at most M partially assigned or split users, that can be achieved in polynomial time. Therefore, the suboptimal solution (8) can be bounded below by

$$\tilde{y} = \lceil y^* - M \rceil, \quad (24)$$

if $w_i = 1 \forall i \in \mathcal{I}$ and else

$$\tilde{y} = y^* - M \max_i w_i. \quad (25)$$

V. INTERPRETATION OF THE LAGRANGE MULTIPLIERS

The following interpretations intend to shed light on the optimization problem from an intuitive point of view. First we take a look at which air-interface a user is assigned in general. Without loss of generality we restrict our analysis to a scenario with two air-interfaces and $c_{i,m} \lambda'_m > c_{i,n} \lambda'_n$. Using condition (11) gives

$$w_i - \lambda''_i = c_{i,m} \lambda'_m - \lambda'''_{i,m} = c_{i,n} \lambda'_n - \lambda'''_{i,m}. \quad (26)$$

Since all λ are non-negative it follows that only $\lambda'''_{i,n}$ can be zero and the user is assigned to air-interface n . This example

makes clear that a user is always assigned to the air-interface with the minimum weighted cost, where λ' represents the weight. If the average cost of users in one air-interface is low its λ' increases so that the resources of other air-interfaces are also exploited. Thus λ' is a measure of the capacity of an air-interface. Next we examine λ'' . It gives information on the merit of a user on a scale between zero and w_i independent of the air-interface. As seen earlier $\lambda_i'' = 0$ if the user is assigned only partly, if $\lambda_i'' = w_i$ the cost of the user has to be zero. The interpretations of the Lagrange multipliers can help to design further simplified assignment algorithms. For example, if averaged values of λ' are known for all air-interfaces the assignment of users can be performed based on the weighted costs without solving the optimization problem for each set of user requests.

VI. ALGORITHM

Our suboptimal algorithm to optimization problem (5) can be divided into two parts. First we solve the relaxed problem (6), which is a linear program (LP). Then we use heuristics to move from the relaxed optimum to a feasible solution of the combinatorial problem.

A variety of algorithms and ready to use tools are known in literature to solve LP with different complexity and convergence characteristics. In this paper we apply the ellipsoid method [5] and optimize over the dual variables λ'_m . For this combination we observe very fast convergence to the optimum in simulations. The dual function is defined by the supremum of the Lagrangian (10) over the primal variables

$$g(\lambda) = \sup_{\mathbf{V}} L(\mathbf{V}, \lambda) \geq y^*. \quad (27)$$

By making constraint (13) explicit and due to compactness of \mathbf{V} we obtain, which we refer to as the inner problem

$$g(\lambda') = \max_{\mathbf{v}, \sum_{m \in \mathcal{M}} v_{i,m} \leq 1} \sum_{i,m \in \mathcal{I}, \mathcal{M}} w_i v_{i,m} - \sum_{m \in \mathcal{M}} \lambda'_m \left(\sum_{i \in \mathcal{I}} c_{i,m} v_{i,m} - 1 \right). \quad (28)$$

Since the inner problem overestimates y^* and Slater's condition applies we can solve (6) by

$$y^* = \min_{\lambda'} g(\lambda'). \quad (29)$$

For a given λ' (28) can be solved by dividing users into subsets

$$\mathcal{I}_m = \{i | i \in \mathcal{I}, i \notin \mathcal{I}_{n \neq m}, m = \arg \min_{m \in \mathcal{M}} c_{i,m} \lambda'_m, w_i > \lambda'_m c_{i,m}\}. \quad (30)$$

Then

$$g(\lambda') = \sum_{m \in \mathcal{M}} \sum_{i \in \mathcal{I}_m} (w_i - \lambda'_m c_{i,m}) + \sum_{m \in \mathcal{M}} \lambda'_m. \quad (31)$$

The optimum λ' is found using the ellipsoid method, a multi-dimensional extension to the bisection method. It generates a sequence of ellipsoids with shrinking volume. In each iteration a half-space of the actual ellipsoid can be ruled out by using

Algorithm 1 Ellipsoid Method

initialize $\lambda_m^{(0)} = \frac{1}{2} \lambda_{m,max}$ with $\lambda_{m,max} = \max_{i \in \mathcal{I}} \frac{w_i}{c_{i,m}}$
and $\mathbf{E} = \frac{1}{\|\lambda_{max}\|^2} \mathbf{I}_M$

while $\max(\text{eig}(\mathbf{E}^{(n)})) > \text{tol}$ **do**

(1) calculate $\mathcal{I}_m \forall m \in \mathcal{M}$ and set

$$v_{i,m}^{(n)} = \begin{cases} 1, & \text{if } i \in \mathcal{I}_m \\ 0, & \text{else} \end{cases}$$

(2) calculate $[\nu(\hat{\lambda}')_m]^{(n)} = \sum_{i \in \mathcal{I}} c_{i,m} v_{i,m}^{(n)} - 1$

(3) update ellipsoid

$$\mathbf{E}^{(n+1)} = \frac{|\mathcal{M}|^2 - 1}{|\mathcal{M}|^2} \left(\mathbf{E}^{(n)} + \frac{2}{|\mathcal{M}| - 1} \frac{\nu^{(n)} \nu^{(n)T}}{\nu^{(n)T} \mathbf{E}^{(n)-1} \nu^{(n)}} \right)$$

with new centroid

$$\lambda^{(n+1)} = \lambda^{(n)} + \frac{1}{|\mathcal{M}| + 1} \frac{\mathbf{E}^{(n)-1} \nu^{(n)}}{\sqrt{\nu^{(n)T} \mathbf{E}^{(n)-1} \nu^{(n)}}}$$

(5) assure $\nu^{(n)} \in \mathbb{R}^+$

end while

a subgradient and a new ellipsoid, containing the half-space with the optimum solution, is build.

To calculate a subgradient define

$$\hat{v}_{i,m} = \arg \max_{v_{i,m}} L(\mathbf{V}, \lambda') \quad \text{subject to } \sum_m c_{i,m} v_{i,m} \leq 1. \quad (32)$$

Then

$$g(\lambda') \geq L(\hat{\mathbf{V}}, \lambda') = g(\hat{\lambda}') + \sum_{m \in \mathcal{M}} \left(\sum_{i \in \mathcal{I}} c_{i,m} \hat{v}_{i,m} - 1 \right) (\lambda'_m - \hat{\lambda}'_m) \quad (33)$$

and a valid subgradient $\nu(\hat{\lambda}')$ with elements

$$[\nu(\hat{\lambda}')_m] = \sum_{i \in \mathcal{I}} c_{i,m} \hat{v}_{i,m} - 1. \quad (34)$$

can be extracted. Given an ellipsoid with center $\hat{\lambda}'$ from (33) it is clear that the half-space corresponding to

$$\lambda' : (\lambda' - \hat{\lambda}') \nu > 0 \quad (35)$$

can be ruled out. The ellipsoid method is summarized in Algorithm 1.

To extract a feasible solution to (5) we first divide users into subsets using the optimal λ'^*

$$\mathcal{I}_m^* = \{i | i \in \mathcal{I}, m = \arg \min_{m \in \mathcal{M}} c_{i,m} \lambda_m^*\}. \quad (36)$$

If all \mathcal{I}_m^* are disjoint no split users exist and the optimization problem decouples in M independent Knapsack problems. A good approximation is achieved by first ordering the elements in the subsets with decreasing weight-cost ratios $w_i/c_{i,m} \forall i \in \mathcal{I}_m^*$. Then, starting with the element with the highest weight-cost ratio one assigns users until no complete one can be

assigned without violating (11) [6]. If the subsets are not disjoint we use the following procedure: For all RAT without split users we proceed like above. From the remaining air-interfaces we pick those with only one split user and without partially assigned ones ($c_{i,m}\lambda_m^* \neq 1, i \in \mathcal{I}_m$). Here we assign all users with better weight-cost ratios than the split one. Next we choose all air-interfaces where the subsets overlap with the previously picked ones and assign users like in the decoupled case. We repeat this until only RAT with partially assigned users remain. Those are assigned last using the same procedure. We refer to this strategy as polynomial assignment in the simulations which has at most M users less than the optimum solution. Often it is possible to further improve the assignment. In case weights of users differ one can try to fill the unused resources with previously not assigned users with low costs. In a second step remaining resources can sometimes be rearranged by interchanging users between RAT that additional users fit into the system. However, nothing can be said about the performance of this additional steps in general. Simulations, where we apply the interchanging will be denoted as heuristically improved results.

VII. SIMULATION RESULTS

In this section the performance of the proposed algorithm and its bounds are evaluated using simulations and compared to a standard multi-RAT load-balancing algorithm. We consider a single-cell downlink scenario with a UMTS and GSM RAT with overlapping coverage. A number of 20 voice users, requesting a minimum data rate of 12.2kbit/s and 20 streaming users with requests of 128kbit/s are equally distributed on a circular playground with radius 1200m. All users have equal weights. The UMTS and the GSM base-stations are both positioned in the middle of the playground. Continuous SIR-rate mapping curves and pathloss exponents from measurements are used to calculate the cost matrix using (2),(3). Extra-cell interference, fading and arrival processes are neglected. In simulations the performance of the following 3 algorithms and the bounds are compared:

- Polynomial Assignment: The proposed algorithm from Section VI.
- Heuristically Improved Algorithm (HIA): After applying the algorithm above, the sum of unused resources of both RAT is sometimes greater than the cost of a not assigned user. Then there is a chance that shifting of users rearranges the left resources that additional users can be assigned. Due to complexity reasons we check rearranging only for a limited number of users. For each assigned user we check if it would fit into the alternative RAT and if the freed resource plus the unused resource allows the assignment of an additional user.
- Load-balancing Assignment: The set of users \mathcal{I} is randomly split into the subsets \mathcal{I}_{GSM} and \mathcal{I}_{UMTS} with equal size. In each RAT the assignment is performed like in the polynomial assignment based on the disjoint sets \mathcal{I}_{UMTS} and \mathcal{I}_{GSM} . If one air-interface is not fully loaded after the assignment procedure, it attempts to assign as many

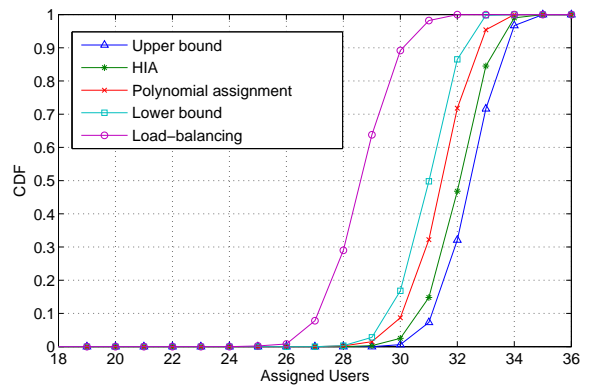


Fig. 2. Cumulative distribution function of maximum assignable users

additional users from the set of non assigned users from the alternative RAT.

Simulation results are presented in Figure 2. Here the cumulative distribution functions of the maximum assignable users are shown. Based on random user positions 1000 cost matrices have been generated and used as input for the 3 algorithms. It can be seen that the presented suboptimal algorithm and the bounds top the load-balancing considerably at relatively low computational expenses. Using heuristic improvements the simulations indicate further increase of the utility.

VIII. CONCLUSION

In this paper we studied the problem how users should be assigned in a heterogeneous multi-service scenario. Based on the concept of resource costs, we formulated an optimization problem that maximizes the weighted number of users assignable to the set of air-interfaces. Due to the NP-completeness exactly solving this problem is in general prohibitive because of exponentially growing complexity. We therefore derived an upper and lower bound of polynomial assignment algorithms using continuous relaxation and Lagrangian theory. Based on the interpretation of the dual variables we presented a suboptimal algorithm performing at least as good as the lower bound. In simulations we compared our algorithm with standard load-balancing procedures and observed a significantly increased utility.

REFERENCES

- [1] Anders Furuskär and Jens Zander, "Multiservice Allocation for Multiaccess Wireless Systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, January 2005.
- [2] L. Fleischer, M.X. Goemans, V.S. Mirrokni, and M. Sviridenko, "Tight Approximation Algorithms for Maximum General Assignment Problems," in *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms*, 2006.
- [3] Joakim Jalden, Cristoff Martin, and Björn Ottersten, "Semidefinite Programming for Detection in Linear Systems - Optimality Conditions and Space-Time Decoding," in *Acoustics, Speech, and Signal Processing*, 2003.
- [4] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [5] R. M Freund and C. Roos, "The ellipsoid method," www.isa.ewi.tudelft.nl/~roos/courses/wi485/ellips.pdf.
- [6] Hans Kellerer, Ulrich Pferschy, and David Pisinger, *Knapsack Problems*, Springer, Berlin, Germany, 2004.