Optimal association of mobile users to multi-access edge computing resources

Stefania Sardellitti, Mattia Merluzzi, and Sergio Barbarossa
Sapienza Univ. of Rome, DIET Dept, Via Eudossiana 18, 00184, Rome, Italy
e-mail: {stefania.sardellitti, mattia.merluzzi, sergio.barbarossa}@uniroma1.it

***Abstract*—Multi-access edge computing (MEC) plays a key role in fifth-generation (5G) networks in bringing cloud functionalities at the edge of the radio access network, in close proximity to mobile users. In this paper we focus on mobile-edge computation offloading, a way to transfer heavy demanding, and latency-critical applications from mobile handsets to close-located MEC servers, in order to reduce latency and/or energy consumption. Our goal is to provide an optimal strategy to associate mobile users to access points (AP) and MEC hosts, while contextually optimizing the allocation of radio and computational resources to each user, with the objective of minimizing the overall user transmit power under latency constraints incorporating both communication and computation times. The overall problem is a mixed-binary problem. To overcome its inherent computational complexity, we propose two alternative strategies: i) a method based on successive convex approximation (SCA) techniques, proven to converge to local optimal solutions; ii) an approach hinging on matching theory, based on formulating the assignment problem as a matching game.**

***Index Terms*—Multi-access edge computing, computation offloading, resources allocation, cloud assignment.**

I. Introduction

The main goal of 5G communication networks is to design a communication infrastructure that will enable an efficient integration of cross-domain networks serving multiple sectors, or verticals, such as Industry 4.0, automotive, multimedia, energy, etc [1], [2]. A plethora of new mobile applications and services for 5G are envisioned, such as interactive gaming, virtual reality and natural language processing, to name a few. Most of these applications are rather demanding in terms of computation needs and energy consumption. This trend raises a conflict between resource-hungry applications and the limited battery lives of mobile devices. This conflict poses a significant challenge to the implementation of the novel mobile applications in an energy efficient manner. One promising solution is to leverage Multi-access Edge Computing (MEC), a new architecture that provides information technologies (IT) and cloud-computing services within the Radio Access Network (RAN), in close proximity to mobile subscribes [3]. Exploiting MEC, a mobile user equipment (UE) can offload computation-intensive and latency-critical applications to the MEC servers at the edge of the network, rather than utilizing the servers in the core network. Thanks to proximity, offloading to MEC servers is convenient for reducing latency, saving mobile users energy consumption and enabling simple devices, like inexpensive sensors, to run sophisticated applications [4]. From a user perspective, one of the parameters mostly affecting the quality of experience is the end-to-end (E2E) latency, i.e. the time necessary to get the result of running an application. Using computation offloading, the E2E latency includes the transmission time to send bits from the UE to the MEC server to transfer the program execution plus the execution time needed to run the application remotely. Therefore the overall latency couples communication and computation resources. This motivates the joint allocation of these resources, as proposed in [4], [5]. A further substantial improvement to computation offloading comes from the introduction of millimeter wave (mmWave) links. Merging MEC with an underlying mmWave physical layer creates indeed a unique opportunity to bring IT services to the mobile user with very high data rate. This merge is indeed one of the main objectives of the joint Europe/Japan H2020 Project called 5G-MiEdge (Millimeter-wave Edge Cloud as an Enabler for 5G Ecosystem) [6]. Since mmWave links are prone to blocking events, which may jeopardize the benefits of computation offloading, a possible way to counteract blocking events in mmWave links was proposed in [7], [8].

Several works investigated computation offloading optimization strategies in MEC systems in the multi-user case [9]–[12]. In [9], a joint optimization of radio and computation resources was investigated, in a multi-user MIMO scenario, taking into account inter-cell interference. In [10], the authors minimize the overall energy consumption at the mobile side, in case of TDMA and OFDMA systems, while the authors of [11] proposed a joint optimization of offloading decision and allocation of computation and communication resources. In [12], MEC computation offloading decision was formulated as a computation offloading game. Only few works focus on the association of users to APs and MEC servers. In [13], we proposed a sub-optimal association strategy minimizing the users energy consumption, taking into account radio and computation parameters jointly. The server selection problem was studied in [14] for a multiuser system to decide whether to offload computation either to the edge server or to the central cloud. In [15], the server selection over multiple MEC servers is formulated as a congestion game. Another approach for the Cloud Radio Access Networks (C-RAN) is presented in [16], based on matching theory.

In this paper we consider a mmWave edge cloud scenario, composed of multiple APs and multiple MEC servers concur-

This work has been supported by H2020 EUJ Project 5G-MiEdge, Nr. 723171.

ring to serve multiple UEs. The association of a UE to a pair of AP and MEC server depends not only on radio channel parameters, but also on the availability of computational resources at the MEC server and the state of the backhaul network. A UE can get radio access from a certain AP, but its application can run on a MEC server located elsewhere, exploiting wired or wireless backhaul. We formulate the offloading problem as the jointly optimal association between UEs, APs and MEC servers, and allocation of mobile radio and computational resources. To solve the resulting mixed-binary problem with affordable complexity, we propose two alternative sub-optimal strategies: i) a method based on SCA techniques, as developed in [17], which extends our previous approach [13] by incorporating the penalty method recently proposed in [18]; ii) a method based on matching theory [19], extending the approach of [20] to deal with computation offloading.

II. System model

Let us consider a mmWave based cloud access network where multiple users may get radio access through multiple APs and multiple MEC servers. In particular, we consider a system composed of N_b small cell access points, N_c MEC servers and K mobile users. Denote with $\mathcal{I} \triangleq \{k : k = 1, \dots K\}$ the set of users asking for computation offloading of their applications to a set of MEC servers. From the offloading point of view, we simplify the classification of applications by assuming that each of them is characterized through the following parameters: i) the number b_k of bits to be transmitted from the mobile user to the MEC server to transfer the program execution; ii) the number of CPU cycles ω_k needed to run the application. We denote by L_k the E2E latency requested by UE k to run its application. In case of offloading, the overall latency experienced by the k-th UE for accessing the network through the access point n when served by cloud m, is given by

$$T_{knm} = T^{\mathtt{exe}}_{mk} + T^{\mathtt{tx}}_{kn} + T^{\mathtt{rx}}_{knm} + T_{Bnm}. \tag{1}$$

The first term in (1) is the server execution time:

$$T^{\mathtt{exe}}_{mk} = \omega_k / f_{mk}, \tag{2}$$

where ω_k is the number of CPU cycles to be executed and f_{mk} is the number of CPU cycles/second allocated by the m-th server to the k-th UE; $T^{\mathtt{rx}}_{knm}$ is the time necessary for the server to send the result back to the k-th UE; T_{Bnm} is the backhaul delay between access point n and MEC server m; $T^{\mathtt{tx}}_{kn}$ is the time spent to send the program state and input (encoded with b_k bits) from the k-th UE to the n-th AP. This time enables the transfer of the program execution from the UE to the MEC server. More specifically, the time $T^{\mathtt{tx}}_{kn}$ necessary for UE k to transmit b_k bits over a channel of bandwidth B to the n-th AP is

$$T^{\mathtt{tx}}_{kn}(p_{kn}) = \frac{c_k}{r_{kn}(p_{kn})} \tag{3}$$

where $c_k = b_k/B$ and $r_{kn}(p_{kn})$ is the spectral efficiency, which, in the interference-free regime, assumes the form

$$r_{kn}(p_{kn}) = \log_2 (1 + \alpha_{kn} p_{kn}) \tag{4}$$

where p_{kn} is the transmit power of user k and α_{kn} is an equivalent channel coefficient. We assume mmWave communications for the radio access and, under Line Of Sight (LOS) conditions, we use Friis formula to model the path loss. Each pair of UE and AP is supposed to be equipped with, respectively, n_T transmit antennas and n_R receive antennas. We also denote with d_{kn} the distance between UE k and AP n. In a LOS condition with a single path with isotropic array elements, the channel matrix $\boldsymbol{H}_{kn} \in \mathbb{C}^{n_R \times n_T}$ between UE k and AP n is rank one. In this case, the channel coefficient α_{kn} in (4) is $\alpha_{kn} \triangleq v_{kn}^2 \xi_{kn} / \sigma_n^2$ with ξ_{kn} the eigenvalue of the rank one matrix $\boldsymbol{H}_{kn} \boldsymbol{H}_{kn}^H$; σ_n^2 is the noise variance; and the coefficient v_{kn} is defined as $v_{kn} \triangleq \frac{\lambda \zeta}{4\pi\,(d_{kn}/d_0)}\, e^{-\beta\, d_{kn}/2}$ where ζ incorporates some efficiency terms, λ is the wavelength associated to the carrier frequency, d_0 is the far field reference distance, β is the atmospheric absorption coefficient.

Within this edge-cloud scenario, the association of a UE to a pair of AP and MEC server depends not only on the radio channel parameters, but also by the computation resources availability of the MEC servers. Therefore, by extending our previous approach in [13], in the next section we propose an optimization strategy to jointly find the optimal computation and communication resources allocation and the optimal association between UEs, APs and MEC servers.

III. MEC Offloading over multi-server networks

Our goal is now to devise an optimal strategy to assign each user to an access point and to a MEC server, while jointly optimizing the radio and computation resources allocation. The objective is to minimize the transmit power consumption of all users, under power budget and latency constraints. The assignment is performed by properly selecting the binary values $a_{knm} \in \{0, 1\}$ for $k = 1, \dots, K$, $n = 1, \dots, N_b$, $m = 1, \dots, N_c$, where the subscripts k, n, and m denote, respectively, UE, AP, and MEC server indexes. For the sake of simplicity, we assume that each user is served by a single AP and a single MEC server. Therefore, for each k, $a_{knm} = 1$ if user k accesses the network through AP n and it is served by the m-th MEC server, while $a_{knm} = 0$ otherwise.

The objective function we wish to minimize is the sum of the powers spent by all UE's:

$$f(\mathbf{p}, \mathbf{a}) \triangleq \sum_{k=1}^{K} \sum_{n=1}^{N_b} \sum_{m=1}^{N_c} p_{kn} a_{knm}$$

where $\mathbf{p} \triangleq (\mathbf{p}_k)_{k \in \mathcal{I}}$, $\mathbf{p}_k \triangleq (p_{kn})_{\forall n}$, $\mathbf{a} \triangleq (\mathbf{a}_k)_{k \in \mathcal{I}}$, $\mathbf{a}_k \triangleq$

$(a_{knm})_{\forall n,m}$. The resulting optimization problem is:

$$\begin{aligned}
\min_{\mathbf{p},\mathbf{f},\mathbf{a}} \quad & f(\mathbf{p},\mathbf{a}) \triangleq \sum_{k=1}^{K}\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} p_{kn}a_{knm} \qquad (\mathcal{P}) \\
\text{s.t.} \quad & \text{i) } g_{knm}(p_{kn}, f_{mk}, a_{knm}) \leq L_k, \forall\, k,n,m \\
& \text{ii) } p_{kn} \leq P_k, \quad p_{kn} \geq 0, \forall\, k,n \\
& \text{iii) } h_m(\mathbf{f},\mathbf{a}) \triangleq \sum_{k=1}^{K}\sum_{n=1}^{N_b} a_{knm}f_{mk} \leq F_m,\ \forall\, m,\ \mathbf{f} \geq \mathbf{0} \\
& \text{iv) } \sum_{n=1}^{N_b}\sum_{m=1}^{N_c} a_{knm} = 1,\ a_{knm} \in \{0,1\},\ \ \forall\, k,n,m
\end{aligned} \tag{5}$$

where we define the function $g_{knm}(p_{kn}, f_{mk}, a_{knm}) \triangleq a_{knm}\left(\frac{c_k}{r_{kn}(p_{kn})} + \frac{w_k}{f_{mk}} + T_{Bnm}\right)$. The above constraints have the following meaning: i) the overall latency of each user k must be lower than the maximum value L_k; ii) the total power spent by each user must be lower than a fixed total power budget P_k; iii) the sum of the computational rates f_{mk} assigned by each server cannot exceed the server computational capability F_m; iv) each mobile user should be served by one AP-MEC server pair, so that we enforce the constraint $\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} a_{knm} = 1$, for each k. For simplicity we have incorporated the term T^{rx}_{knm} in the latency limit L_k. It can be noted from the latency expression in (1) the interplay between radio access and computational aspects and this calls for a *joint* optimization of the radio resources, the transmit power $\mathbf{p}$ of the UEs and the computational rates $\mathbf{f} \triangleq (f_{mk})_{\forall m,k}$. Unfortunately, problem $\mathcal{P}$ is a mixed-binary problem and, in general, NP-hard. To handle its computational cost with affordable complexity, in the following we propose two alternative suboptimal strategies.

IV. SCA-BASED OPTIMIZATION STRATEGY

In this section we propose a suboptimal optimization strategy to solve problem $\mathcal{P}$, combining our previous approach in [13] with the successive convex approximation strategy proposed in [17] and incorporating an efficient penalty term, recently proposed in [18], to relax the binary variables to be real while driving the solution towards the situation where each UE is served by a single AP and a single MEC cloud. More specifically, the penalty method in [18] is based on the fact that, given the following problem,

$$\begin{aligned}
\min_{\mathbf{a}_k} \quad & \| \mathbf{a}_k + \epsilon\mathbf{1} \|_q^q \triangleq \sum_{n=1}^{N_b}\sum_{m=1}^{N_c} (a_{knm} + \epsilon)^q \\
\text{s.t.} \quad & \| \mathbf{a}_k \|_1 = 1, \\
& a_{knm} \in [0,1], \quad \forall\, n,m
\end{aligned} \tag{6}$$

where $q \in (0,1)$, $\epsilon > 0$, the optimal solution of (6) is binary, i.e. only one element is one and all the others are zero. The optimal solution is $c_{\epsilon,k} = (1+\epsilon)^q + (N_bN_c - 1)\epsilon^q$. Therefore, we relax our binary variables a_{knm} to be real and belonging to the following convex set

$$\mathcal{A} = \{(\mathbf{a}_k)_{k\in\mathcal{I}} \ :\ a_{knm} \in [0,1], \sum_{n=1}^{N_b}\sum_{m=1}^{N_c} a_{knm} = 1, \forall\, k,n,m\},$$

and we add a penalty to the objective function so that our relaxed optimization problem becomes

$$\begin{aligned}
\min_{\mathbf{p},\mathbf{f},\mathbf{a}} \quad & f_{P_\sigma}(\mathbf{p},\mathbf{a}) \triangleq f(\mathbf{p},\mathbf{a}) + \sigma P_\epsilon(\mathbf{a}) \qquad (\mathcal{P}_\sigma) \\
\text{s.t.} \quad & \text{i) } g_{knm}(p_{kn}, f_{mk}, a_{knm}) \leq L_k, \forall\, k,n,m \\
& \text{ii) } h_m(\mathbf{f},\mathbf{a}) \triangleq \sum_{k=1}^{K}\sum_{n=1}^{N_b} a_{knm}f_{mk} \leq F_m,\ \forall\, m,\ \mathbf{f} \geq \mathbf{0} \\
& \text{iii) } p_{kn} \leq P_k, \quad p_{kn} \geq 0, \forall\, k,n,\ \mathbf{a} \in \mathcal{A}
\end{aligned} \tag{7}$$

where $\sigma > 0$ is the penalty parameter and

$$P_\epsilon(\mathbf{a}) \triangleq \sum_{k=1}^{K} \| \mathbf{a}_k + \epsilon\mathbf{1} \|_q^q - c_{\epsilon,k}. \tag{8}$$

Even by relaxing the binary variables $\mathbf{a}$, problem $\mathcal{P}_\sigma$ is still non-convex, since the objective function and the constraints i), ii) are non convex. In what follows, we exploit the structure of problem $\mathcal{P}_\sigma$ and building on some recent advances on SCA techniques [17], we devise an efficient iterative penalty SCA approximation algorithm (PSCA) converging to a local optimal solution. To solve the non-convex problem $\mathcal{P}_\sigma$ efficiently, we adopt an SCA-based algorithm where the original problem is replaced by a sequence of strongly convex problems. To do this, we start by finding a suitable convex approximation of the nonconvex objective function that is the sum of the non-convex term $f(\mathbf{p},\mathbf{a})$ and the concave function $P_\epsilon(\mathbf{a})$.

Let $\mathbf{x} \triangleq (\mathbf{p},\mathbf{f},\mathbf{a})$ and $\mathbf{x}_k \triangleq (\mathbf{p}_k,\mathbf{f}_k,\mathbf{a}_k)$ with $\mathbf{f}_k \triangleq (f_{mk})_{\forall m}$. We denote by $\mathcal{X}$ the feasible set of problem $\mathcal{P}_\sigma$ and we denote by $\mathbf{x}^\nu \triangleq (\mathbf{p}^\nu,\mathbf{f}^\nu,\mathbf{a}^\nu)$ the set of variables at iteration ν of SCA. Following [17], the main idea is to approximate around the current iterate $\mathbf{x}^\nu \in \mathcal{X}$, the original nonconvex nonseparable term with a strongly convex function, say $\tilde{f}_{P_\sigma}(\mathbf{x};\mathbf{x}^\nu)$, that has the same first order behaviour of the original objective function at $\mathbf{x}^\nu$. To find a convex approximant of the objective function observe that $f(\mathbf{p},\mathbf{a})$ has a bilinear structure, since it is the sum of the terms $s_{knm}(p_{kn},a_{knm}) \triangleq p_{kn}a_{knm}$. Therefore, as suggested in [17], s_{knm} can be written as a difference of convex (DC) functions, i.e.

$$s_{knm}(p_{kn},a_{knm}) = \frac{1}{2}(p_{kn}+a_{knm})^2 - \frac{1}{2}(p_{kn}^2 + a_{knm}^2). \tag{9}$$

A valid convex upper approximation of s_{knm}, for any given $(p^\nu_{kn}, a^\nu_{knm}) \in \mathbb{R}^2$, is then

$$\tilde{s}_{knm}(p_{kn},a_{knm};p^\nu_{kn},a^\nu_{knm}) \triangleq \frac{1}{2}(p_{kn}+a_{knm})^2 - \frac{1}{2}\left(p^{\nu\,2}_{kn} + a^{\nu\,2}_{knm}\right) - p^\nu_{kn}(p_{kn}-p^\nu_{kn}) - a^\nu_{knm}(a_{knm}-a^\nu_{knm}).$$

Finally, the concave function $P_\epsilon(\mathbf{a})$ can be approximated by its first order approximation at the iterate $\mathbf{a}^\nu$, i.e.

$$P_\epsilon(\mathbf{a}) \approx P_\epsilon(\mathbf{a}^\nu) + \nabla P_\epsilon(\mathbf{a}^\nu)^T(\mathbf{a}-\mathbf{a}^\nu). \tag{10}$$

Then, a convex approximation of $f_{P_\sigma}(\mathbf{p},\mathbf{a})$ can be defined as:

$$\tilde{f}_{P_{\sigma^\nu}}(\mathbf{x};\mathbf{x}^\nu) \triangleq \sum_{k=1}^{K}\Bigg[\sum_{n=1}^{N_b}\sum_{m=1}^{N_c}\tilde{s}_{knm}(p_{kn},a_{knm};p_{kn}^\nu,a_{knm}^\nu)+ \sigma_\nu \nabla P_\epsilon(\mathbf{a}_k^\nu)^T(\mathbf{a}_k-\mathbf{a}_k^\nu)+\tau_p\|\mathbf{p}_k-\mathbf{p}_k^\nu\|^2+\tau_f\parallel \mathbf{f}_k-\mathbf{f}_k^\nu\parallel^2+ \tau_a\|\mathbf{a}_k-\mathbf{a}_k^\nu\|^2\Big], \tag{11}$$

where we added quadratic regularization terms to make $\tilde{f}_{P_\sigma}$ strongly convex with respect to $\mathbf{x}$. Note that in (11) we use a monotonically increasing penalty sequence $\{\sigma_\nu\}_\nu$ to guarantee that the obtained solution $\mathbf{a}$ is binary [18].

We show now how to reduce the non-convex constraint $g_{knm}(p_{kn},f_{mk},a_{knm})$ to a convex form. To do so, observe that at any feasible point $(\mathbf{p},\mathbf{f},\mathbf{a})$, it must be $r_{kn}(p_{kn})>0$, $f_{mk}>0$ and $L_k > T_{Bnm}a_{knm}-\omega_k a_{knm}f_{mk}$, for all k,n,m, then the constraints $g_{knm}(p_{kn},f_{mk},a_{knm})$ in (7) can be rewritten as

$$g_{knm}(p_{kn},f_{mk},a_{knm}) = -r_{kn}(p_{kn})+q_{knm}(f_{mk},a_{knm})$$

that is the sum of the convex term $-r_{kn}(p_{kn})$ and the convex function $q_{knm}(f_{mk},a_{knm}) \triangleq \dfrac{c_k a_{knm} f_{mk}}{f_{mk}(L_k-T_{Bnm}a_{knm})-\omega_k a_{knm}}$.

Finally, the non-convex bilinear constraint $h_m(\mathbf{f},\mathbf{a})$ may be replaced by the following convex approximation

$$\tilde{h}_m(\mathbf{f},\mathbf{a};\mathbf{x}^\nu) \triangleq \sum_{k=1}^{K}\sum_{n=1}^{N_b}\frac{1}{2}\left(a_{knm}+f_{mk}\right)^2-\frac{1}{2}(a_{knm}^{\nu\,2}+f_{mk}^{\nu\,2}) - a_{knm}^\nu(a_{knm}-a_{knm}^\nu)-f_{mk}^\nu(f_{mk}-f_{mk}^\nu).$$

We can now introduce the proposed convex approximation of the nonconvex problem $\mathcal{P}_\sigma$. Given the feasible point $\mathbf{x}^\nu \in \mathcal{X}$, we have

$$\begin{aligned}\hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq \underset{\mathbf{x}=(\mathbf{p},\mathbf{f},\mathbf{a})}{\text{argmin}} \quad & \tilde{f}_{P_{\sigma^\nu}}(\mathbf{x};\mathbf{x}^\nu) && (\mathcal{P}^\nu)\\ \text{s.t.} \quad & g_{knm}(p_{kn},f_{mk},a_{knm})\le 0,\ \forall k,n,m\\ & \tilde{h}_m(\mathbf{f},\mathbf{a};\mathbf{x}^\nu)\le F_m,\ \forall m\\ & f_{mk}\ge 0,\ \forall k,m\\ & p_{kn}\le P_k,\quad p_{kn}\ge 0, \forall\, k,n,\ \mathbf{a}\in\mathcal{A}\end{aligned}$$

where we denoted by $\hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq (\hat{\mathbf{p}}(\mathbf{x}^\nu),\hat{\mathbf{f}}(\mathbf{x}^\nu),\hat{\mathbf{a}}(\mathbf{x}^\nu))$ the unique solution of the strongly convex optimization problem $\mathcal{P}^\nu$. The proposed solution method consists in solving iteratively problem $\mathcal{P}^\nu$, starting from a feasible point $\mathbf{x}^0$. First we find out an optimal solution $\hat{\mathbf{x}}$ of $\mathcal{P}^\nu$ by setting the penalty coefficient σ to zero. Hence, taking this optimal solution as initial point, we iteratively solve $\mathcal{P}^\nu$ with an increasing penalty coefficient σ_ν. In Algorithm 1 below we provide a formal description of the algorithm. The convergence proof of the proposed algorithm is given in [21] and is omitted here because of space limitation.

Note that in Step 2 of the algorithm we allow a memory in the update of the iterate $\mathbf{x}^\nu \triangleq (\mathbf{p}^\nu,\mathbf{f}^\nu,\mathbf{a}^\nu)$.

Algorithm 1 : PSCA Algorithm for $\mathcal{P}$

Data: $\mathbf{x}^0 \triangleq (\mathbf{p}^0,\mathbf{f}^0,\mathbf{a}^0)\in\mathcal{X}$, $\{\gamma^\nu\}_\nu \in (0,1]$, $\tau_p,\tau_f>0$; $\tau_a>0$, $0<\eta<1<\beta$, $\epsilon_0>0$, $\sigma_0>0$. Set $\nu=0$;
(S.1): If $\mathbf{x}^\nu$ satisfies a suitable termination criterion, go to (S.5);
(S.2): Compute $\hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq (\hat{\mathbf{p}}(\mathbf{x}^\nu),\hat{\mathbf{f}}(\mathbf{x}^\nu),\hat{\mathbf{a}}(\mathbf{x}^\nu))$ by solving $\mathcal{P}^\nu$ with $\sigma=0$;
(S.3): Set $\mathbf{x}^{\nu+1}=\mathbf{x}^\nu+\gamma^\nu(\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu)$;
(S.4): $\nu\leftarrow\nu+1$ and go to (S.1);
(S.5): If $\hat{\mathbf{a}}^\nu$ is binary, stop;
(S.6): Else initialize $\mathbf{x}^0\triangleq\mathbf{x}^\nu$, N_{max}, $T_{max}<N_{max}$, $\nu=0$;
(S.7): While $\nu<N_{max}$ do
(S.8): Compute $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ by solving $\mathcal{P}^\nu$;
(S.9): Set $\mathbf{x}^{\nu+1}=\mathbf{x}^\nu+\gamma^\nu(\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu)$;
(S.10): If $\hat{\mathbf{a}}^\nu$ is binary, stop;
(S.11): Otherwise set $\nu=\nu+1$;
(S.12): If $\nu\le T_{max}$, then $\sigma_\nu=\beta\sigma_{\nu-1}$, $\epsilon_\nu=\eta\epsilon_{\nu-1}$;
(S.13): End

V. Matching theory based optimization

In this section, we propose an alternative approach to overcome the combinatorial complexity of the assignment problem by devising an optimization strategy based on matching theory [19]. Inspired by [20], which used matching theory for the uplink selection of AP, we generalize the approach of [20] to computation offloading. The assignment problem is formulated as a matching game in which users and AP-MEC pairs rank one another using suitable preference functions associated to the transmit power used by each user to implement computation offloading under latency constraints. Matching theory is a powerful and simple tool to associate agents of two different sets using suitable preference lists. A typical matching problem is the college admission problem [22], where students apply to colleges based on their preference lists and are accepted based on colleges' preference lists. Each college cannot accept more students than a certain number, defined as its quota q. The aim of matching theory algorithms is to find a stable assignment. An assignment of applicants to colleges is called *unstable* if there are two applicants α and β who are assigned to colleges A and B, respectively, although β prefers A to B and A prefers β to α. Matching theory has been extensively used in economics, and recently introduced in wireless networks [23]. In the context of C-RAN, the authors in [16] find an assignment of UE's to Radio Remote Head (RRH), Base Band Unit (BBU) and computing resources to minimize the refusal ratio, i.e. the portion of requests that cannot meet their deadlines. The preference function is based on the *expected* latency that a user would experience choosing a certain triple RRH, BBU, and computing resource. In [22], the Deferred Acceptance (DA) algorithm is presented and proved to converge to a stable matching. In the DA algorithm, students apply to their preferred college, which subsequently accept students based on their preference lists, rejecting the least preferred ones. Applying matching theory, and in particular the DA algorithm to problem $\mathcal{P}$ is not straightforward due to inter-dependencies of utility functions necessary to build preference lists. In fact, while users get accepted by a pair AP-MEC server, the convenience of being assigned to that pair changes due to the need for resource sharing. As pointed

out in [20], in case of interdependent preferences, the general college admission game becomes complex. In [20], matching is used only for the uplink selection of AP, and the R-factor, a parameter that incorporates both the delay and the packet success rate, is used as utility function. To overcome the problem of interdependent preferences, the authors propose to divide the problem into two interdependent subgames: a matching game, where users build their preferences based on the potential R-factor guarantees (supposing that each access point n fills up its quota q_n), and a second subgame, where users can request to be transferred to another AP to improve their R-factors.

Generalizing this approach to our assignment problem, we first need to define a utility function to build the users' preference lists. In our joint allocation of communication and computation resources, we incorporate both communication and computation parameters in the preference function. As in Section III, for the sake of simplicity we assume perfect beamforming and interference-free channels. In particular, every UE is supposed to be served with the same frequency band at the same time. We focus instead on the delay caused by computation resource sharing. To define the utility function, we consider constraint $i)$ of problem $\mathcal{P}$. Even though we do not have any a priori information on allocated resources, we can get an approximate estimation of the minimum transmit power that a user would experience choosing a certain pair AP-MEC server using the delay constraint. To do this, we compute an *expected* minimum transmit power in case of a disjoint allocation. In particular, given a certain allocation of computation resources, the minimum transmit power necessary to meet the latency constraint can be easily found. As we do not know a priori the assignment of users to each pair AP-MEC server, initially we assume that each MEC server m serves all users as far it does not exceed its quota q_m, in order to consider the maximum computation delay. Thus, for the first assignment, we compute f_{mk}, for each user k and server m, with a proportional rule as follows

$$f_{mk} = \frac{w_k}{\sum_{i=1}^{K} w_i} F_m. \tag{12}$$

Replacing (12) in the execution delay expression given in (2), the minimum rate to meet the latency constraint L_k can be written as

$$r_{knm}^{\min} = \frac{c_k}{L_k - L_{Bnm} - \frac{w_k}{f_{mk}}}. \tag{13}$$

Inverting (4), the associated minimum transmit power is then

$$p_{knm}^{min} = \frac{2^{r_{knm}^{\min}} - 1}{\alpha_{kn}}. \tag{14}$$

We define the utility function for user k accessing AP n and MEC server m as

$$U_{knm} = -p_{knm}^{min}. \tag{15}$$

Based on this utility function, each user builds its preference list. Similarly, each AP build its preference list based on the best SNR. For simplicity, we assume that all MEC servers can accept an unlimited number of users. However, this condition can lead to a solution very far from the optimum, since a single MEC server has limited resources. Indeed, a first stage for the assignment is not sufficient due to interdependency of the preference functions of all users. For this reason, as in [20], we perform a second stage with a coalitional game to transfer users given the new conditions. A coalition $\mathcal{C}_{nm}$ is the set of all users associated to AP n and MEC server m. Once users are assigned with the DA algorithm, one can compute the new proportional disjoint allocation of computation resources as follows:

$$f_{mk} = \begin{cases} \frac{w_k}{\sum_{j=1}^{N_b} \sum_{i \in \mathcal{C}_{jm}} w_i} F_m, & \text{if } \exists\, j : k \in \mathcal{C}_{jm} \\ \frac{w_k}{\sum_{j=1}^{N_b} \sum_{i \in \mathcal{C}_{jm}} w_i + w_k} F_m, & \text{if } \nexists\, j : k \in \mathcal{C}_{jm}. \end{cases} \tag{16}$$

Computing the new approximate computation delays, we can compute the *expected* minimum transmit powers from (14) and build the new preference lists. Now, users can request to be transferred, based on the new utility functions. In particular, as in [20], user k requests to be transferred to coalition $\mathcal{C}_{n'm'}$ from coalition $\mathcal{C}_{nm}$ if $U_{kn'm'} > U_{knm}$. If more users request to be transferred to a certain coalition, only the most preferred user is considered. Each transfer is accepted if the following two conditions hold [20]:

1) MEC server m' does not exceed its quota $q_{m'}$;
2) The social welfare, represented by the sum of users' utility functions.

Formally, the second condition can be written as follows:

$$W(\mathcal{C}_{nm} \setminus \{k\}) + W(\mathcal{C}_{n'm'} \cup \{k\}) > W(\mathcal{C}_{nm}) + W(\mathcal{C}_{n'm'})$$

where $W(\mathcal{C}_{nm}) = \sum_{k \in \mathcal{C}_{nm}} U_{knm}$, and $\mathcal{C}_{nm} \setminus \{k\}$ is defined as the set obtained by removing user k from $\mathcal{C}_{nm}$. This stage stops if there are no more transfer requests or the social welfare is not improved by any transfer. In [20] it is proved that, given any initial assignment, this second game will converge to a Nash-stable partition, where no user has any incentive to execute a transfer. Once the assignment has been performed, for every MEC server, we optimize the radio and computation resources jointly as in $\mathcal{P}$. This second stage is solved in closed form using our previous work [8].

VI. Numerical results and Conclusions

To test the effectiveness of the proposed offloading strategy, in Fig. 1 we report the optimal total transmit power consumption vs. the maximum latency L_k, assumed equal for all users. To test the effectiveness of the proposed algorithms, we compare their performance with the optimal results achieved with the exhaustive search. We consider a network composed of $K = 4$ users, a number of base stations equal to the number of clouds, i.e. $N_b = N_c = 2$. The other parameters are set as follows: $F_1 = 2.7 \cdot 10^9$, $F_2 = 3 \cdot 10^8$, $P_k = 1.35 \cdot 10^{-1}$, $q = 0.7$. We may observe that both the PSCA and the matching game algorithms provide results very close to the exhaustive search algorithm whose complexity is exponential. Additionally, we consider as comparison term the SNR-based association method, in both cases where the

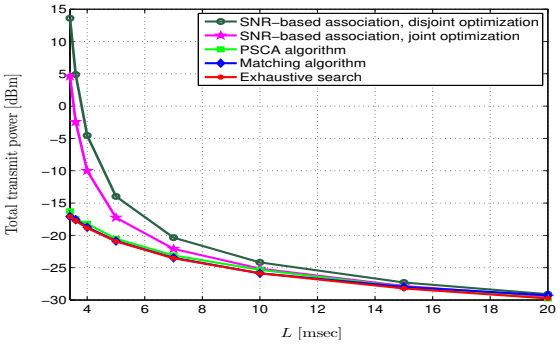

Fig. 1: Overall UE transmit power consumption vs. L.

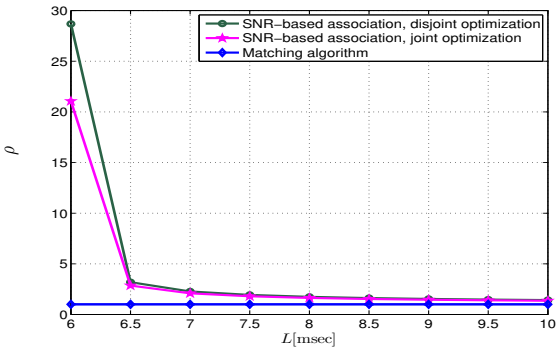

Fig. 2: Averaged ratio ρ vs. L.

radio and computational resources are jointly and disjointly optimized. It can be noted that both proposed approaches yield considerable power savings with respect to SNR-based methods, taking advantage of the optimal assignment of each user to a cloud through the most convenient base station. It has to be remarked that the complexity of the matching-based algorithm has a polynomial growth with the number of players (users and AP-MEC pairs), although the reached final solution could be suboptimal, as the preference lists are built based on an approximate a priori knowledge. To further test the effectiveness of the matching algorithm, in Fig. 2 we show the ratio ρ between the overall power consumptions achieved with two different association rules and the global optimal solution, averaged over the channel realizations. It is interesting to note from Fig. 2 that ρ keeps quite close to 1 for the proposed matching algorithm.

In summary, in this paper we formulated the offloading problem in MEC systems as a joint optimization of the association between users and AP-MEC server pairs and the allocation of radio and computation resources. To solve the resulting mixed-binary nonconvex problem, we proposed two alternative strategies that, albeit suboptimal, converge to solutions very close to the results achieved with an exhaustive search.

REFERENCES

[1] *5G empowering vertical industries*, 5G PPP White paper, Feb. 2016.

[2] J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C.K. Soong, and J.C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] ETSI, "Mobile-edge computing introductory technical white paper," *White Paper, Mobile-edge Comput. Indus. Init.*, [Online]. Available: https://portal.etsi.org/portals/0/tbpages/mec/docs/ mobile-edge computing - introductory technical white paper v1.

[4] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

[5] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *Proc. of IEEE Work. Signal Process. Adv. Wir. Commun. (SPAWC 2013)*, 16–19 Jun. 2013.

[6] "5G-MiEdge millimeter-wave edge cloud as an enabler for 5G ecosystem," *Europe/Japan project co-funded by the European Commission's Horizon 2020 and Japanese Ministry of Internal Affairs and Communications; website: http://5g-miedge.eu*.

[7] S. Barbarossa, E. Ceci, M. Merluzzi, and E. Calvanese-Strinati, "Enabling effective mobile edge computing using millimeterwave links," in *Proc. of IEEE Int. Conf. Commun. Work. (ICC Wkshps)*, May 2017, pp. 367–372.

[8] S. Barbarossa, E. Ceci, and M. Merluzzi, "Overbooking radio and computation resources in mmw-mobile edge computing to reduce vulnerability to channel intermittency," in *Proc. of 2017 Eur. Conf. Net. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[9] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Net.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[10] C. You, K. Huang, H. Chae, and B. H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wir. Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[11] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. 5, pp. 11255–11268, 2017.

[12] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Net.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[13] S. Sardellitti, S. Barbarossa, and G. Scutari, "Distributed mobile cloud computing: Joint optimization of radio and computational resources," in *Proc. of 2014 IEEE Globecom Work. (GC Wkshps)*, Dec. 2014, pp. 1505–1510.

[14] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "A cooperative scheduling scheme of local cloud and Internet cloud for delay-aware mobile cloud computing," in *Proc. of 2015 IEEE Globecom Work. (GC Wkshps)*, Dec. 2015, pp. 1–6.

[15] Y. Ge, Y. Zhang, Q. Qiu, and Y.-H. Lu, "A game theoretic resource allocation for overall energy minimization in mobile cloud computing system," in *Proc. of ACM/IEEE Int. Symp. Low Pow. Elec. Des.*, Jul.-Aug. 2012, pp. 279–284.

[16] T. Li, C. S. Magurawalage, K. Wang, K. Xu, K. Yang, and H. Wang, "On efficient offloading control in cloud radio access network with mobile edge computing," in *Proc. of IEEE 37th Int. Conf. Dist. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 2258–2263.

[17] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization - Part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.

[18] N. Zhang, Y. F. Liu, H. Farmanbar, T. H. Chang, M. Hong, and Z. Q. Luo, "Network slicing for service-oriented networks under resource constraints," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2512–2521, Nov. 2017.

[19] A. E. Roth and M. Sotomayor, "Two-sided matching," *Handbook of game theory with economic applications*, vol. 1, pp. 485–541, 1992.

[20] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. of IEEE Conf. Comput. Commun. (INFOCOM 2014)*, Apr. 2014, pp. 1096–1104.

[21] S. Sardellitti, S. Barbarossa, and M. Merluzzi, *Optimal association of mobile users to multi-access edge computing resources*, submitted to IEEE Trans. Signal Inform. Process. Net., 2017.

[22] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The Amer. Math. Month.*, vol. 69, no. 1, pp. 9–15, 1962.

[23] Zhu Han, Yunan Gu, and Walid Saad, *Matching Theory for Wireless Networks*, Springer Publish. Comp., Inc., 1st edition, 2017.