ENABLING EFFECTIVE MOBILE EDGE COMPUTING USING MILLIMETER WAVE LINKS

Sergio Barbarossa[1], *Elena Ceci*[1], *Mattia Merluzzi*[1], *Emilio Calvanese-Strinati*[2]

[1] Sapienza University of Rome, DIET, Via Eudossiana 18, 00184 Rome, Italy
[2] CEA, LETI, Minatec Campus, 17 rue des Martyrs, 38054 Grenoble, France

ABSTRACT

Mobile Edge Computing (MEC) plays a key role in the 5G roadmap, as a way to bring information technology (IT) services closer to the mobile users by empowering radio access points with additional functionalities, like caching or computation offloading capabilities. At the physical layer, some of the key technologies enabling very low latency mobile services are dense deployment of radio access points, massive MIMO and millimeter-wave (mmW) communications for radio access as well as for radio fronthaul/backhaul. In this paper, we merge computation offloading techniques for mobile edge computing with mmW communications and show how the joint optimization of computation/communication resources is crucial to design an energy efficient mobile edge computing system. In particular, we tackle the intermittency of mmW links by considering multi-link communications and show how to take advantage of preliminary estimation of blocking probabilities.

Index Terms— Mobile edge computing, millimeter-wave communications, computation offloading

1. INTRODUCTION

Even though the fifth generation (5G) of mobile technology is yet to be specified, there is a wide acceptance of the fact that 5G will not be just a further generation of radio access standard, but rather an efficient integration of cross-domain networks serving diverse sectors, or *verticals*, such as industry 4.0, automotive, energy, multimedia, e-health, etc. 5G will enable operators to provide networks on an as-a-service basis, to be able to meet the wide range of use cases through the design of *network slices* [1]. To meet the requirements of such a broad vision, 5G will need to be able to deliver services with very low latency and high reliability. Within this framework, Mobile Edge Computing (MEC) has been identified as a key architecture to enable proximity access to IT services from mobile users [2]. The goal of MEC is to provide an IT service environment and cloud-computing capability at the edge of the mobile network, within the Radio Access Network (RAN) and in close proximity to mobile subscribers. This is obtained by enabling radio access points (RAP) to get fast access to distributed storage and computation resources in order to meet the mobile users requests as close to the mobile user as possible, using a fully scalable architecture where users' requests may be handled at the mobile device, or in the nearest cloud-enhanced RAP, or cluster of RAPs, properly interconnected through high capacity links [3], or in a (possibly far away) cloud server. MEC will be particularly effective in delivering localized or context-aware services. Some of the potential applications are augmented reality, localized extraction of analytics from live video streaming for distributed security, RAN-aware optimization of video download, caching, and computation offloading, and so on.

At the physical layer, the use of millimeter wave (mmW) links makes possible the large increase of data rate with respect to current standards (three orders of magnitude) through the significant increase of bandwidth, dense deployment of RAP's, and massive multi-input/multi-output (MIMO). The idea pushed forward in this paper, reflecting also the focus of the H2020 project 5G-MiEdge [4], is to merge MEC and mmW technologies within a common holistic view that sees communication, computation, and caching as three primary components of a single system. An effective design of such a system requires a *joint optimization* of these three primary resources, as also recently suggested in [5], [6]. In particular, in this paper we will focus on computation offloading strategies. Mobile cloud computing (MCC) has received great attention as a way to provide cloud services (either storage or computation) to mobile users [7]. Alternative algorithms for the joint optimization of computation and communication resources for computation offloading were proposed in [8]. The coordinated management of interference and CPU time allocation in the multi-user case was then considered in [9] and, more thoroughly, in [10], assuming perfect channel knowledge. These approaches were recently generalized to incorporate the limitations of the backhaul link and considering both uplink and downlink channel in [11]. The previous techniques are inherently static. Dynamic approaches have been considered as well, incorporating the channel randomness in [12], [13], [14], and [15].

Clearly, MEC in general, and computation offloading in particular, may greatly benefit from the introduction of mmW links for possibly both radio access and backhaul, because

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

of the data rate increase. However, the strict latency constraints foreseen in 5G require taking care of the *intermittent* nature of mmW channels: mmW links may in fact fail either because of an obstacle appearing between transmit and receive sides, or because of an interferer transmitting within the mainlobe of the receive array. In this paper, we extend the joint optimization approach proposed in [9], [10], by incorporating knowledge (estimation) of the blocking probabilities into the optimization problem. In this way, we counteract the losses due to channel intermittency, by allocating additional radio resources to guarantee that the latency constraints are satisfied, in average sense, even in the presence of unforeseen blocking events. Furthermore, we consider multi-link communications between the mobile user and nearby access points to overcome blocking. We start with the single user case and show how to optimize the resource allocation taking into account preliminary estimations of the blocking probability. Then, we consider the multi-user case and show how a joint allocation of communication and computation resources can provide considerable advantages with respect to disjoint strategies.

2. SCENARIO

In this section, we briefly recall some of the basic concepts about mmW communication models and computation offloading, as they will be used later on. We assume MIMO communications and use beamforming at both transmit ad receive sides to counteract the attenuation of the received power as a function of the carrier frequency. Using planar arrays and beamforming at both transmit and receive sides, the received power P_R, from the Friis' formula in function of equivalent areas, varies as a function of the distance R between transmit and receive sides as [16]

$$P_R = \frac{P_T A_T A_R \eta}{(\lambda R)^2}\, e^{-\alpha R} \tag{1}$$

where P_T is the transmit power, A_T is the area of the transmit antenna, A_R is the area of the receive antenna, λ is the wavelength associated to the carrier frequency, η is an efficiency coefficient (less than one) incorporating losses at both transmit and receive sides and α is the attenuation factor due to absorption in the propagation through the atmosphere.

2.1. Statistical interference model

In [16], the authors proposed an analytic model for the probability of having an interference on mmW MIMO links in a mesh network. In this work, we consider a small cell scenario, where the radio access points placed over lamppost, building corners, i.e with a very small elevation angle between access points and mobile users. In such a case, we may proceed similarly to [16] and assume that a transmission is successfully decoded by the access point receiver if the total signal-to-interference-plus-noise ratio (SINR) is above a given threshold, say β. Otherwise, the transmission is considered to be lost. For simplicity, thermal noise is assumed to be negligible with respect to the interference. Let us denote by N_{Tx} the number of transmit nodes over a deployment area A, assuming a Poisson random variable with density ρ. We consider first the probability of high interference due to a single interferer randomly located at a distance R and angle ϕ_1 relative to the receiver. The angle ϕ_2 represents the pointing direction of the interferers beam relative to the receiver. We use the functions $g(\phi_1), g(\phi_2)$ as the normalized power gain pattern for an N-element linear array in which each individual flat-top element has beam angle (or sector size) $\Delta\phi_{ft}$. The probability p_c of having an interference is $p_c = A_c/A$, with [16]

$$A_c := \iint\limits_{(r,\phi_1 \in A)} \int_{-\pi}^{\pi} \mathbf{1}\left(g(\phi_1) g(\phi_2) \frac{R_0^2}{r^2} e^{-\alpha(r-R_0)} \geq \frac{1}{\beta} \right) \frac{r}{2\pi} dr d\phi_1 d\phi_2 \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function that takes the value 1 when its argument is true, and 0 otherwise. In the above expression, the antenna gains along the elevation angle are assumed to be the same for both the useful user and the interferer. If we consider now $N_{Tx} - 1$ interferers placed randomly over a surface of area A, the probability of interference is [16]:

$$P_I = 1-(1-p_c)^{N_{Tx}-1} = 1- \lim_{A\to\infty} \left(1-\frac{A_c}{A}\right)^{\rho A} = 1-e^{-\rho A_c} \tag{3}$$

In our set-up the interferers are the users associated to a cell different from the cell pertaining to the user of interest. In words, this probability depends on the antenna azimuth beamwidth, the density of potential interfering devices, the area illuminated from the antenna beam and on the distance between interferers and receiver.

2.2. Computation offloading

Computation offloading is a key strategy to move the computation burden from resource-hungry mobile devices to more powerful fixed servers. The purpose may be either to save energy at the mobile side, thus overcoming the limited battery resources of mobile devices, or to make possible for mobile devices to run computationally intensive applications. Computation offloading is convenient when: i) the energy necessary to offload computation remotely is less than the energy necessary to run the application locally, under a given latency constraint; or ii) the latency resulting from offloading is less than the latency experienced locally, for a given energy consumption; iii) the mobile device (maybe a sensor or IoT device) is unable to run sophisticated applications. We use the term *latency* here to mean the time necessary to run an application and denote by L the maximum latency limit. We

quantify the computational load associated to offloading with the following parameters: w measures the amount of CPU cycles necessary to run the application; n_b is the number of bits to be transmitted from the mobile user to the fixed server to enable the transfer of the execution; f_S is the number of CPU cycles/sec of the server (virtual machine) dedicated to run the mobile application. From the communication side, we use MIMO transceivers and denote by n_T and n_R the number of antennas at the transmit and receive sides, respectively. $\mathcal{E}_{loc}$ is the energy spent to run the application at the mobile side. $\boldsymbol{F}$ is the precoding (beamforming) matrix $\mathbf{Q}$ is the covariance matrix of the transmitted symbols ($\mathbf{Q} = \sigma_s^2 \boldsymbol{F}\boldsymbol{F}^H$ when the symbols before precoding are uncorrelated, with variance σ_s^2), with $\mathrm{tr}(\mathbf{Q}) = P_T$, where P_T is the maximum transmit power of the mobile device; $\boldsymbol{H}$ is the $n_R \times n_T$ channel matrix between mobile user and base station; $\boldsymbol{R}_n$ is the disturbance (interference plus noise) covariance matrix. Offloading takes place if the energy spent for offloading is less than the energy necessary to run the application locally, under a given latency constraint. The decision is taken by a MEC server coordinating possibly multiple RAPs, based on the users' requests and channel status. The strategy is to find the optimal precoding matrix (equivalently, the covariance matrix of the transmitted symbols) in order to minimize power consumption, subject to (s.t.) the following constraints: i) latency constraint; ii) energy for offloading less than the energy to be spent for local processing; iii) transmit power less than available power budget. For the single user case, the problem can be formulated as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{Q}\succeq\mathbf{0}} & \mathrm{trace}(\mathbf{Q}), \ \text{ subject to (s.t.)} \\ \mathcal{X} \triangleq & \text{i)} \dfrac{n_b}{R(\mathbf{Q})} + \dfrac{w}{f_S} + \Delta_R \le L \\ & \text{ii) } \mathrm{tr}(\mathbf{Q}) \le P_T, \quad \mathbf{Q} \succeq \mathbf{0} \end{array} \tag{4}$$

where $R(\mathbf{Q}) = B \log_2 \det\left(\mathbf{I} + \mathbf{H}\mathbf{Q}\mathbf{H}^H \mathbf{R}_n^{-1}\right)$ and Δ_R is the time necessary to send the result back to the mobile user. The constraints have the following meaning: $n_b/R(\mathbf{Q})$ is the time necessary to transmit n_b bits over a bandwidth B using optimal coding. The symbol $\mathcal{X}$ denotes the feasible set: If $\mathcal{X}$ is empty, offloading is not convenient or impossible to carry out within the user's requirements and processing is performed at the mobile device; if $\mathcal{X}$ is non-empty, the previous problem is convex and the solution can be found with effective numerical tools. The latency constraint in (4) can be expressed, equivalently, as

$$R(\mathbf{Q}) \ge R_{min} := n_b/(L - w/f_S - \Delta_R). \tag{5}$$

3. MULTI-LINK COMMUNICATIONS

As mentioned in Section 2, one of the major impairments of mmW communications is that the channel is intermittent, depending on the occurrence of blocking or large interference. We model interference as additive noise, for the sake of simplicity. Better performance could be achieved by applying successive decoding, but this of course requires additional complexity and it goes beyond the scope of this paper. The most effective strategy to counteract blocking is to enable multi-link communications, so that the Mobile User Equipment (MUE) can transmit towards multiple radio access points, depending on their status. We focus on two access points, for simplicity of explanation, but the approach can be generalized to multiple base stations in a straightforward way. Let us denote by $\mathbb{P}_{s,r}$ the joint probability of having channels s and/or r available. More specifically, $\mathbb{P}_{1,0}$ is the probability that channel 1 is on, while channel 2 is off, $\mathbb{P}_{0,1}$ is the probability that channel 2 is on, while channel 1 is off, and $\mathbb{P}_{1,1}$ is the probability that both channels are on. In case of statistically independent blocking events, we simply have $\mathbb{P}_{1,0} = (1 - P_{I_1})P_{I_2}$, $\mathbb{P}_{0,1} = (1 - P_{I_2})P_{I_1}$ and $\mathbb{P}_{1,1} = (1 - P_{I_1})(1 - P_{I_2})$, where P_{I_i} is the probability that the single channel i is blocked. In case of interference, P_{I_i} is given by Eqn. (3). Suppose now that at the beginning of each frame, there are control signals used to check if a channel is on or off. Then, depending on channel availability, the MUE will use the precoding matrix $\boldsymbol{F}_1$ if only channel 1 is available, matrix $\boldsymbol{F}_2$ if only channel 2 is available, and matrix $\boldsymbol{F}_3$ if both channels are available. Denoting with $\mathbf{Q}_i$ the covariance matrices of the transmitted symbols corresponding to using precoding matrices $\boldsymbol{F}_i$, under ergodicity assumptions, the average transmit power coincides with the expected value of the transmit power, equal to

$$\mathbb{E}\{p\} = \mathbb{P}_{1,0}\mathrm{tr}(\mathbf{Q}_1) + \mathbb{P}_{0,1}\mathrm{tr}(\mathbf{Q}_2) + \mathbb{P}_{1,1}\mathrm{tr}(\mathbf{Q}_3). \tag{6}$$

Similarly, the average rate is

$$\mathbb{E}\{R\} = \mathbb{P}_{1,0}R_1(\mathbf{Q}_1) + \mathbb{P}_{0,1}R_2(\mathbf{Q}_2) + \mathbb{P}_{1,1}R_3(\mathbf{Q}_3) \tag{7}$$

where

$$R_i(\mathbf{Q}_i) = B \log_2 \left| \boldsymbol{I} + \boldsymbol{H}_i \mathbf{Q}_i \boldsymbol{H}_i^H \boldsymbol{R}_{n_i}^{-1} \right|. \tag{8}$$

where $R_i = R_i(\mathbf{Q}_i) = \log_2 |\boldsymbol{I} + \sigma^{-2}\boldsymbol{H}_i\mathbf{Q}_i\boldsymbol{H}_i^H|$, for $i = 1, 2$; when $i = 3$, i.e. the MUE is transmitting towards both RAP's, the channel matrix is the $(n_{R_1} + n_{R_2}) \times n_T$ matrix obtained by stacking matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ on top of each other, i.e., $\boldsymbol{H}_3 := (\boldsymbol{H}_1^H, \boldsymbol{H}_2^H)^H$. If we consider the eigendecomposition $\boldsymbol{H}_i^H \boldsymbol{R}_{ni}^{-1} \boldsymbol{H}_i^H := \boldsymbol{U}_i \boldsymbol{\Lambda}_i \boldsymbol{U}_i^H$, we can set, without loss of generality, $\mathbf{Q}_i = \boldsymbol{U}_i \boldsymbol{\Gamma}_i \boldsymbol{U}_i^H$, so that the rates can be rewritten as

$$R_i(\mathbf{Q}_i) = B \sum_{k=1}^{r_i} \log_2 \left(1 + \gamma_i(k)\lambda_i(k)\right) \tag{9}$$

where r_i denotes the rank of $\boldsymbol{H}_i$.

Extending the approach of Section 2 in order to incorporate blocking event, we can formulate the optimization problem as the minimization of the average power (energy), under the constraint of respecting a given latency constraint, besides

the usual power budget constraint. In formulas, we can write

$$\min_{\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3} \mathbb{E}\{p\} \text{s.t. } \mathbb{E}\{R\} \geq \overline{R}_{min} \tag{10}$$

where $\overline{R}_{min} = (n_b \log 2)/(L - w/f_S - \Delta_R)$. If the feasible set is non empty, the above problem is convex and the unique solution can be obtained with any efficient convex optimization problem solver. To gain insight into the solution is now useful to study what happens in the typical case where there are only line-of-sight (LOS) links between the transmit and receive arrays. In such a case, the matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are both rank-1 matrices, whereas $\boldsymbol{H}_3$ is rank-2. More specifically, we have

$$\boldsymbol{H}_i = c_i \, \boldsymbol{a}_{R_i}(\theta_{R_i}, \phi_{R_i}) \boldsymbol{a}_T^H(\theta_{T_i}, \phi_{T_i}), \tag{11}$$

where $\boldsymbol{a}_{R_i}(\theta_{R_i}, \phi_{R_i})$ is the steering vector at the receiver i along the direction identified by the azimuth angle θ_{R_i} and the elevation angle ϕ_{R_i}, $\boldsymbol{a}_T(\theta_{T_i}, \phi_{T_i})$ is the steering vector at the transmit side along the direction identified by the azimuth/elevation pair $(\theta_{T_i}, \phi_{T_i})$, and c_i is a scalar coefficients that incorporates all attenuation terms, i.e., $c_i = \sqrt{\eta e^{-\alpha R_i} (\lambda/4\pi R_i)^2}$.

To use a more compact notation, let us introduce the symbols $\mathbb{P}_1 := \mathbb{P}_{1,0}, \mathbb{P}_2 := \mathbb{P}_{0,1}, \mathbb{P}_3 = \mathbb{P}_4 := \mathbb{P}_{1,1}$. The optimization problem can then be reformulated in terms of the scalar powers p_i as

$$\begin{aligned} &\min_{\boldsymbol{p}} \ \textstyle\sum_{i=1}^4 \mathbb{P}_i \, p_i, \ \text{s.}t. \\ &\quad i) \ \textstyle\sum_{i=1}^4 \mathbb{P}_i \log(1 + a_i p_i) \geq \overline{R}_{min} \\ &\quad ii) \ p_i \geq 0, i = 1, \ldots, 4; \\ &\quad iii) \ p_i \leq P_T, i = 1, 2; p_3 + p_4 \leq P_T, \end{aligned} \tag{12}$$

where $\boldsymbol{p}$ denotes the vector of powers, and a_i is the i-th channel response. The Lagrangian associated to this constrained problem is

$$\begin{aligned} \mathcal{L} = &\ \mathbb{P}_{1,0} \, p_1 + \mathbb{P}_{0,1} \, p_2 + \mathbb{P}_{1,1}(p_3 + p_4) + \\ &- \beta(\mathbb{P}_{1,0} \log(1 + a_1 p_1) + \mathbb{P}_{0,1} \log(1 + a_2 p_2) \\ &+ \mathbb{P}_{1,1} \log(1 + a_3 p_3) + \mathbb{P}_{1,1} \log(1 + a_4 p_4) - \overline{R}_{min}) \\ &- \sum_{i=1}^4 \alpha_i p_i + \sum_{i=1}^2 \nu_i (p_i - P_T) + \nu_3 (p_3 + p_4 - P_T) \end{aligned}$$

where β, $\alpha_i, i = 1, \ldots, 4$, and $\nu_i, i = 1, \ldots, 3$ are the Lagrange multipliers associated to the constraints $i)$, $ii)$, and $iii)$, respectively, in (12). Applying the Karush-Kuhn-Tucker conditions, we can express all powers p_i in terms of the Lagrange multiplier β in the case of a single available link as

$$p_i = \left[\beta - \frac{1}{a_i} \right]_0^{P_T}, i = 1, 2 \tag{13}$$

where $[x]_a^b$ denotes the projection of the real variable x in the interval $[a, b]$. In the case of two available links, we can express the two powers as a function of the Lagrange multiplier ν_3 as follows

$$p_i = \left[\frac{\beta \mathbb{P}_i}{\mathbb{P}_i + \nu_3} - \frac{1}{a_i} \right]_+, \ i = 3, 4. \tag{14}$$

where $[x]_+$ denotes $max(0, x)$. When the constraints $iii)$ of (12) hold with the inequality sign, we can set $\nu_i = 0, \forall i$, so that (13) and (14) can be written as

$$p_i = \beta - \frac{1}{a_i}, i = 1, \ldots, 4. \tag{15}$$

Otherwise, we can find ν_3 imposing the equivalence $p_3 + p_4 = P_T$, while $p_i = P_T, i = 1, 2$. Note that β is necessarily different from zero, or otherwise the powers would not respect the condition to be non-negative, and it can be found by imposing that condition $i)$ in (12) holds with the equality sign. Interestingly, whenever the parameters of the optimization problem are such that no powers reach the maximum power budget P_T and the blocking events on the two channels are statistically independent, β can be expressed in closed form as

$$\beta = \exp\left(\frac{c}{2 - P_{I_1} - P_{I_2}} \right) \tag{16}$$

where $c := \overline{R}_{min} - \sum_{i=1}^4 \mathbb{P}_i \log(a_i)$. In their simplicity, expressions (15) and (16) show the advantage of establishing multi-link communications. In fact, from (16) it is clear that the transmit power can increase dramatically only if both channels are usually blocked, i.e. both P_{I_1} and P_{I_2} tend to 1. However, as far as at least one channel is not blocked most of the time, i.e. at least P_{I_1} or P_{I_2} are small, then multi-link communication facilitates an energy-efficient communication.

As a numerical example, in Fig. 1, we compare the average transmitted energy obtained with multi-link as opposed to single-link communication, under the same latency constraint, assuming a $f_c = 60$ GHz carrier frequency. We consider interference as the blocking event. We can notice the large gain obtained obtained through a 2-link communication. Furthermore, we see that the gain increases with distance because, as distance increases the probability of having an interference increases as well [16]. Of course, computation offloading performs differently as a function of the kind of application being offloaded. As an example, in Fig. 2, we show the energy consumption as a function of the computational load w to be offloaded, for a given number of bits to be transmitted to enable the transfer of the execution. We can see how the improvement resulting from a double-link connection increases for more computationally demanding applications.

4. JOINT ALLOCATION OF COMMUNICATION/COMPUTATION RESOURCES IN A MULTIUSER SCENARIO

Let us consider now the more challenging case where there are K users who can access the network through 2 RAP's.

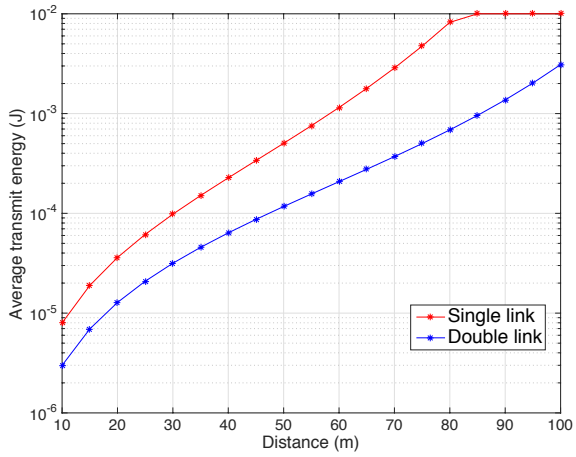

Fig. 1. Average transmit energy: single-link vs. multi-link.

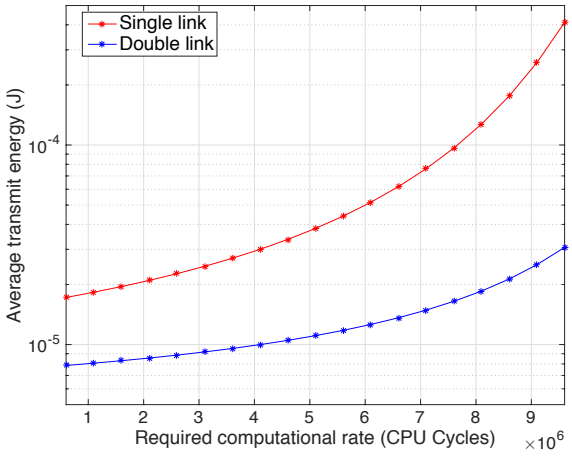

Fig. 2. Average transmit energy vs. computational load.

Each of them can transmit to either one RAP or to both, depending on the blocking event, exactly as in previous section. We assume that the base stations assign orthogonal radio resources, either time or frequency slots, to different users in order to avoid interference. However, the receive antennas can receive interference from mobile users associated to other base stations. We assume that there is a single computing system serving the K users. Each user wishes to offload computations to this server, under the constraint that the latency incorporating round-trip communications plus execution time be less than a prescribed threshold. This entails a joint allocation of communication and computation resources. Let us assume, for simplicity, as in the previous section, that there are LOS links between each MUE and the RAP's or, in other words, that each channel matrix is rank-1, except the channel matrix from each user and the ensemble of the two base stations, that has rank-2. Let us formulate now the allocation problem as the minimization of the total transmit power, subject to latency constraints. The problem can be formulated as follows:

$$\begin{aligned}
&\min_{\boldsymbol{p},\boldsymbol{f}} \sum_{k=1}^{K}\sum_{i=1}^{4} \mathbb{P}_{k,i}\, p_{k,i} \\
&\text{s.t. } \frac{c_k}{\sum_{i=1}^{4} \mathbb{P}_{k,i} \log(1+a_{k,i}p_{k,i})} + \frac{w_k}{f_k} \le \Delta_k, \quad \forall k; \\
&p_{k,i} \ge 0, i = 1,\dots,4, \quad \forall k; \\
&p_{k,i} \le P_T, i = 1,2; p_{k,3} + p_{k,4} \le P_T, \quad \forall k; \\
&\sum_{k=1}^{K} f_k \le f_S
\end{aligned} \tag{17}$$

where $c_k := n_k \log 2/B$ and $\boldsymbol{f}$ denotes the vector of computational rates. This problem can be equivalently reformulated in terms of the rates $R_{k,i} := \log(1 + a_{k,i}p_{k,i})$, considering that

$$p_{k,i} = \frac{1}{a_{k,i}} \left(e^{R_{k,i}} - 1\right). \tag{18}$$

Denoting by $\boldsymbol{R}$ the vector of all rates $R_{k,i}$, we have then

$$\begin{aligned}
&\min_{\boldsymbol{R},\boldsymbol{f}} \sum_{k=1}^{K}\sum_{i=1}^{4} \frac{\mathbb{P}_{k,i}}{a_{k,i}} \left(e^{R_{k,i}} - 1\right) \\
&\text{s.t. } \frac{c_k}{\sum_{i=1}^{4} \mathbb{P}_{k,i} R_{k,i}} + \frac{w_k}{f_k} \le \Delta_k, \quad \forall k; \\
&R_{k,i} \ge 0, i = 1,\dots,4, \quad \forall k; \\
&R_{k,i} \le \log(1 + a_{k,i}P_T), i = 1,2, \quad \forall k; \\
&\frac{1}{a_{k,3}} \left(e^{R_{k,3}} - 1\right) + \frac{1}{a_{k,4}} \left(e^{R_{k,4}} - 1\right) \le P_T, \quad \forall k; \\
&\sum_{k=1}^{K} f_k \le f_S.
\end{aligned} \tag{19}$$

This problem is convex and then it can be solved very efficiently. To gain insight into the solution, it is useful to consider the single link case, as in such a case we can find closed form expressions for the relation between communication and computational rate allocation.

As a numerical example, in Fig. 3 we compare the power spent for ensuring a minimum average rate in two cases: i) the disjoint optimization case, where the computational resources are assigned in proportion to the users' requests and then the rates are optimized over each access channel, in order to minimize the transmit power under a latency constraints; ii) the joint optimization case, which passes through the solution of the optimization problem given in (19). We can see from Fig. 3 that the joint optimization yields a non negligible gain.

5. CONCLUSION

In this paper we have shown some of the benefits in terms of computation offloading and have proposed strategies to

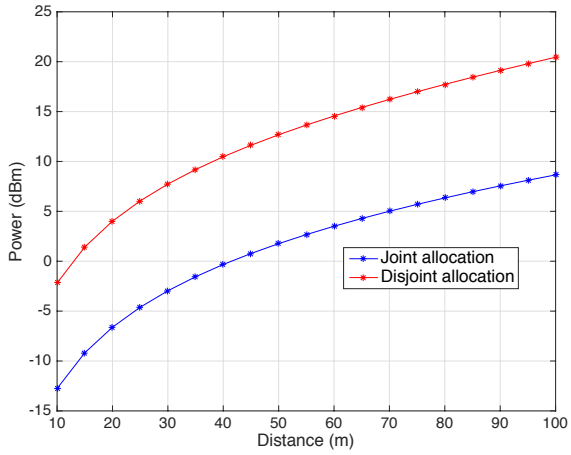

Fig. 3. Transmit power vs. distance: Joint vs. disjoint allocation.

tackle the blocking effect in mmW using multi-link with the goal of minimizing the power consumption at the terminal side necessary to perform computation offloading. In particular, we have shown that the joint optimization of communication/computation resources provides non negligible gain with respect to the typical disjoint strategies. In this work, the computation load was described in very simple manner. A deeper look into the computation mechanisms can provide further gains, even if at the cost of increased complexity. Furthermore, interference has been treated as additive noise, but coordinating multiple base stations in a cloud-RAN architecture may enable joint decoding with consequent performance benefits.

6. REFERENCES

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

[2] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," *ETSI White Paper No. 11*, September 2015.

[3] J. Oueis, E. Calvanese Strinati, A. De Domenico, and S. Barbarossa, "On the impact of backhaul network on distributed cloud computing," in *WCNC 2014, CLEEN Workshop*, 2014, pp. 12–17.

[4] "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*.

[5] S. Andreev, O. Galinina, A. Pyattaev, J. Hosek, P. Masek, H. Yanikomeroglu, and Y. Koucheryavy, "Exploring synergy between communications, caching, and computing in 5G-grade deployments," *IEEE Comm. Mag.*, vol. 54, no. 8, pp. 60–69, 2016.

[6] H. Liu, Z. Chen, and L. Qian, "The three primary colors of mobile systems," *IEEE Comm. Mag.*, pp. 15–21, Sep. 2016.

[7] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

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

[9] ——, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *IEEE Workshp SPAWC 2013*, pp. 26–30.

[10] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. on Signal and Inform. Proc. over Networks*, vol. 1, no. 2, pp. 89–103, 2015.

[11] A. N. Al-Shuwaili, O. Simeone, A. Bagheri, and G. Scutari, "Joint uplink/downlink optimization for backhaul-limited mobile cloud computing with user scheduling," *arXiv preprint arXiv:1607.06521*, 2016.

[12] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. on Wireless Comm.*, vol. 11, no. 6, pp. 1991–1995, 2012.

[13] Y. Zhang, D. Niyato, P. Wang, and C.-K. Tham, "Dynamic offloading algorithm in intermittently connected mobile cloudlet systems," in *IEEE int. Conf. on Communications (ICC)*, 2014, pp. 4190–4195.

[14] W. Labidi, M. Sarkiss, and M. Kamoun, "Joint multi-user resource scheduling and computation offloading in small cell networks," in *WiMob 2015*, pp. 794–801.

[15] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," *preprint arXiv:1604.07525*, 2016.

[16] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE/ACM Trans. on Networking*, pp. 1513–1527, 2011.