Overbooking Radio and Computation Resources in mmW-Mobile Edge Computing to Reduce Vulnerability to Channel Intermittency

Sergio Barbarossa, Elena Ceci, Mattia Merluzzi
Sapienza University of Rome, Dept. of Information Engineering, Electronics, and Telecommunications
Via Eudossiana 18, 00184, Rome, Italy
E-mail: {sergio.barbarossa, elena.ceci}@uniroma1.it; mattia.merluzzi@gmail.com

***Abstract*—One of the key features of 5G roadmap is mobile edge computing (MEC), as an effective way to bring information technology (IT) services close to the mobile user. Moving computation and caching resources at the edge of the access network enables low latency and high reliability services, as required in many of the verticals associated to 5G, such as Industry 4.0 or automated driving. Merging MEC with millimeter wave (mmW) communications provides a further thrust to enable low latency and high reliability services, thanks to the high data rate of mmW links and the ability to handle interference through massive beamforming. However, mmW links are prone to blocking events, which could limit the effectiveness of the mmW-MEC deployment. To overcome blocking events and robustify mmW-MEC, in this paper we propose and analyze the performance of two strategies to overcome the effect of blocking: i) overbooking of computation and communication resources, based on the statistics of blocking events, and ii) adopting multi-link communications[1].**

I. Introduction

The main goal of 5G roadmap is to design a common communication infrastructure enabling new business opportunities in diverse sectors, or *verticals*, such as connected vehicles, automotive industry, augmented reality, videogames and IoT systems. The way to enable such diverse services, with different requirements in terms of latency, reliability and data rate, passes through *network slicing*, where a single physical network is partitioned into multiple *virtual* networks, each matched to its specific requirements and constraints, thus enabling operators to provide networks on an as-a-service basis, while meeting a wide range of use cases in parallel [1]. This new reality, sometimes called fourth industrial revolution, can be realized by a new architecture able to meet advanced requirements, especially in terms of latency (below 5 ms), reliability, coverage (with density up to 100 devices$/m^2$), and bandwidth. At the physical layer, 5G builds on a significant increase of system capacity by incorporating massive MIMO techniques, dense deployment of radio access points and wider bandwidth. All these strategies are facilitated by the use of mmW communications. However, the significant improvement achievable at the physical layer could be still insufficient to meet the challenging and diverse requirements of very low latency and ultra reliability. A further improvement comes from a paradigm shift that puts applications at the center of the system design. Virtualization of network functionalities and MEC are the key tools of this application-centric networking. In particular, MEC plays the key role of bringing cloud-computing resources at the edge of the network, within the Radio Access Network (RAN), in close proximity to mobile subscribers [2]. MEC is particularly effective to deliver context-aware services or to enable computation offloading from resource-poor mobile devices to fixed servers or to perform intelligent cache pre-fetching based on local learning of the most popular contents across space and time.

Merging MEC with an underlying mmW physical layer creates a unique opportunity to bring IT services at the mobile user with very low latency and high reliability. This merge is the objective of the joint Europe/Japan Project called 5G-MiEdge (Millimeter-wave Edge Cloud as an Enabler for 5G Ecosystem) [3]. In this paper, we will address some of the issues related to this merge. In particular, we will concentrate on computation offloading, as one of the key services offered by MEC, and we will propose some strategies to counteract one of the shortcomings of mmW communications, namely vulnerability to blocking [4]. It is in fact well known that mmW links are prone to blocking caused by the sudden appearance of an obstacle between transmitter and receiver or because of an interference coming through the receive mainlobe. Blocking events may jeopardize the benefits of ultra low latency resulting from proximity mmW access to MEC services. In this work, we extend the jointly optimal allocation of communication and computation resources for computation offloading proposed in, e.g., [5], [6], [7], to mmW-MEC, incorporating channel blocking events. In particular, we propose two ways to counteract the undesired effect of blocking: i) overbooking of computation and communication resources, based on knowledge (estimation) of the blocking event statistics; and ii) multi-link communications. More specifically, this work extends our recent contribution to the analysis of mmw-MEC systems [8], by enabling parallel access to more than two access points, considering statistically dependent blocking events, and deriving closed form expressions for the allocation of communication (capacity) and computation (virtual machines) resources. These closed expressions help to better understand the effect of blocking and show how to counteract it through over-provisioning of computation/communication resources.

II. Computation offloading

Computation offloading is advantageous for a mobile device when any of the following conditions is satisfied: i) the energy to be spent by the device to run an application locally is greater than the energy spent to offload it, under a latency constraint; ii) the latency resulting from running an

[1]This work has been supported by H2020 EU-Japan Project 5G-MiEdge, Nr. 723171.

application locally is larger than the latency obtained from offloading; iii) the mobile device (e.g., a sensor) is unable to run the application. Here the term latency is used to indicate the overall delay experienced by the end user to get the result of its application. In a real context, where multiple users raise requests to offload their computations elsewhere, the decision on where to run the applications is taken by a MEC server, which is aware of the radio access network as well as of the available computing resources. The MEC server has in fact access to both radio resources, expressed in terms of capacity of the radio links, and computation resources, expressed in terms of CPU cycles/sec assigned by an hypervisor to the virtual machines serving each user's computational requests. The assignment of these two types of resources is coupled by the latency constraint, which includes both communication time and computation time.

Let us introduce some basic notations. From the computation side, w is the number 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 assigned to the virtual machine running the mobile application; the maximum latency is denoted by L. 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. $\boldsymbol{F}$ 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; $\boldsymbol{R}_n$ is the disturbance (interference plus noise) covariance matrix; $\boldsymbol{H}$ is the channel matrix between mobile user and base station. One possibility to cast the resource optimization problem is to find the optimal precoding matrix (equivalently, the covariance matrix of the transmitted symbols) that minimizes the energy consumption at the mobile side, 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. This problem is non-convex, but in [6] it was proved that it can be equivalently cast as a convex problem, associated to the minimization of the power consumption, whose mathematical formulation is:

$$\begin{array}{ll} \min\limits_{\mathbf{Q}\succeq\mathbf{0}} & \mathrm{trace}(\mathbf{Q}), \ \text{subject to (s.t.)} \\ \mathcal{X} \triangleq & \text{i)} \dfrac{n_b}{R(\mathbf{Q})} + \dfrac{w}{f_S} + \Delta_R \leq L \\ & \text{ii)} \ \mathrm{tr}(\mathbf{Q}) \leq P_T, \quad \mathbf{Q} \succeq \mathbf{0} \end{array} \tag{1}$$

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 latency constraint includes the time $n_b/R(\mathbf{Q})$ necessary to transmit n_b bits over a bandwidth B and the computation time w/f_S. 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 (1) can be expressed, equivalently, as

$$R(\mathbf{Q}) \geq R_{min} := n_b/(L - w/f_S - \Delta_R). \tag{2}$$

III. Single-User Multi-link scenario

In this section we start studying the simple case of a single user accessing MEC services through multiple radio access points (RAPs). In such a case, the channel matrix $\boldsymbol{H}$ is built by stacking on top of each other the single channel matrices $\boldsymbol{H}_i$ from the user to each of the $i-th$ available radio access points, with dimension $s\, n_R \times n_T$, where s is the number of available RAPs. Only for the sake of finding closed form expressions, in the following we make the simplifying assumption that each channel is line of sight (LOS) case. In such a case, every channel matrix $\boldsymbol{H}_i$ can be written as

$$\boldsymbol{H}_i = c_i \, \boldsymbol{a}_{R_i}(\theta_{R_i}, \phi_{R_i}) \boldsymbol{a}_T^H(\theta_{T_i}, \phi_{T_i}), \tag{3}$$

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 for every array element, i.e., $c_i = \sqrt{\eta \left(\frac{\lambda}{4\pi r_i}\right)^2 e^{-\alpha r_i}}$.

Let us consider now the situation where the k-th uplink channel from the user to the k-th RAP is blocked with probability P_{I_k}. We start with the situation where the blocking events are statistically independent.

1) Independent events: In [8], we studied the case of statistically independent events modeled as in [9], and a scenario with two base stations. In such a scenario, $\mathbb{P}_1 = (1 - P_{I_1})P_{I_2}$ is the probability that channel 1 is on and channel 2 is off, $\mathbb{P}_2 = (1 - P_{I_2})P_{I_1}$ vice versa, and $\mathbb{P}_3 = (1 - P_{I_1})(1 - P_{I_2})$ is the probability that both channels are on. Proceding as in [8], the problem can be reformulated in terms of the scalar powers p_i as

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

where $\boldsymbol{p}$ denotes the vector of powers, a_i the channel response coefficients and $\overline{R}_{min} = R_{min} \cdot \log(2)$. Writing the Lagrangian and using the Karush-Kuhn-Tucker (KKT) conditions, the scalar powers allocated on each channel can be expressed in closed form. In particular, whenever the parameters of the optimization problem are such that no powers reach the maximum power budget P_T, the powers can be expressed as follows

$$p_i = \beta - \frac{1}{a_i}, i = 1, \ldots, 4 \tag{5}$$

where β is the Lagrange multiplier associated to the first constraint. Moreover, in this case, β can be expressed in closed form as

$$\beta = \exp\left(\frac{c(P_{I_1}, P_{I_2})}{2 - P_{I_1} - P_{I_2}}\right) \tag{6}$$

where $c(P_{I_1}, P_{I_2}) := \overline{R}_{min} - \sum_{i=1}^4 \mathbb{P}_i \log(a_i)$. The power expressions (5) show the advantages of multi-link communications: in case of two links, we see from (6) that when both channels are off (i.e. P_{I_1} and P_{I_2} tend to 1), as expected, the transmit power increases dramatically (the numerator of exponent of β in (6) tends to a constant and denominator

tends to infinity). However, considering statistical independent blocking events, as far as at least one channel is not completely off, multi-link communications enable a large energy saving.

2) Dependent events: Let us consider now statistically dependent blocking events. This situation arises, for example, when the blocking event is due to a large obstacle that obscures more than one LOS. For sake of simplicity, we assume a scenario with one single mobile user and two RAPs, so the user can be served by either one or both RAPs. The problem is formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{p}} \ & \mathbb{P}_{off}\left(p_1+p_2\right)+\left(1-\mathbb{P}_b\right)^{2-tri_{\Delta\phi}(\delta\phi)}\left(p_3+p_4\right) \ s.t. \\ & i)\mathbb{P}_{off}[\log(1+a_1p_1)+\log(1+a_2p_2)]+ \\ & (1-\mathbb{P}_b)^{2-tri_{\Delta\phi}(\delta\phi)}[\log(1+a_3p_3)+ \\ & \log(1+a_4p_4)] \geq \overline{R}_{min}; \\ & ii)p_i \geq 0, i=1,\ldots,4; \\ & iii)p_i \leq P_T, i=1,2; p_3+p_4 \leq P_T, \end{aligned} \tag{7}$$

where $\mathbb{P}_b$ is the blocking probability, $\Delta\phi$ is the angle shadowed by the obstacle, $\delta\phi$ is the angle under which the two RAPs are seen from the mobile user and $tri_{\Delta\phi}(\delta\phi) := (1-|\delta\phi|/\Delta\phi)$ for $|\delta\phi| < \Delta\phi$ and 0 elsewhere. Consider, for example, a bus passing between the mobile user and the RAPs. If the bus is moving perpendicular to the LOS, its length is D and it passes at a distance d from the mobile user, $\Delta\phi \approx D/d$. $\mathbb{P}_{off}(\delta\phi) = \frac{1-\mathbb{P}_b^{2-tri_{\Delta\phi}(\delta\phi)}-(1-\mathbb{P}_b)^{2-tri_{\Delta\phi}(\delta\phi)}}{2}$ is the probability of having one link on and the other one off, depending on the angle $\delta\phi$. Furthermore, $R_i, i=1,\ldots,3$ are the rates achievable when only the first, only the second or both RAPs are available, respectively. The KKT conditions lead to the same solution as in (5) for the power allocation. However, the statistical dependence of the blocking events leads to a different expression for the Lagrange multiplier β. In such a scenario, β is given as follows

$$\beta = \exp\left(\frac{c'(P_b)}{1-P_b^{2-tri_{\Delta\phi}(\delta\phi)}+(1-P_b)^{2-tri_{\Delta\phi}(\delta\phi)}}\right) \tag{8}$$

where

$$c'(P_b) = \mathbb{P}_{off}\sum_{i=1}^{2}\log(a_i)+(1-\mathbb{P}_b)^{2-tri_{\Delta\phi}(\delta\phi)}\sum_{i=3}^{4}\log(a_i)$$

As a numerical example, in Fig. 1 we report the minimum energy consumption for offloading as a function of the distance between mobile user and RAPs (for simplicity, all RAPs are assumed to be at the same distance from the mobile user). The different curves correspond to different angles $\delta\phi$. We can see from Fig. 1 that the smaller is the angle with which the user sees the RAPs, the larger is the energy spent by the mobile user, since the statistical dependence between the two blocking events increases. When the optimization set is no more feasible or the energy spent for offloading is greater than the one necessary to run the application at the mobile side, the offloading is no longer possible or convenient for the MU, and this explains the ceiling (see the top right side of Fig.1) equal to the energy spent locally.

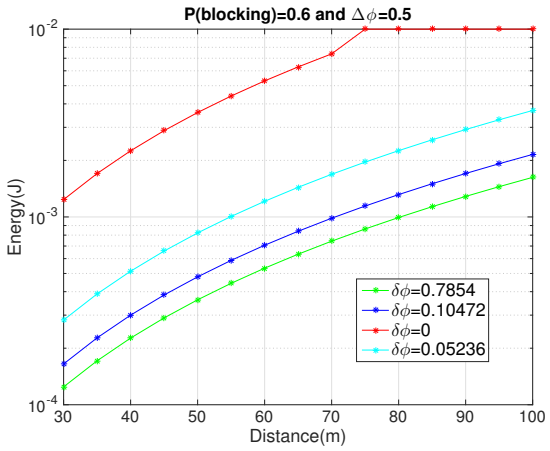

Figure 1. Energy spent in case of statistically dependent blocking events.

IV. Multi-user single link

Now we analyze the more interesting multi-user scenario, where computation and communication resources are jointly optimized. We start with single link communications and we assume that different users are accommodated either in different time or frequency slots, in order to avoid mutual interference. However, every link can be affected by blocking events. The effect of these blocking events is to generate an intermittent channel, so that the average rate experimented by the k-th user decreases as a function of the blocking probability P_{I_k}. We assume a single computing system serving K users in parallel. The aim is to minimize the average transmit power under a latency and a total power budget constraints, similarly to Section III-1, but with the important difference that now the computing resources of the single server are optimally split across the virtual machines running each user applications. We assume LOS conditions for every user, so that all channel matrices are rank-1, and they are defined as in (3). The result of the optimization provides a joint allocation of computation and communication resources, that are linked to each other by the latency constraint that incorporates both sources of delay. In particular, the latency incorporates the time needed to offload the program, the time to run the application and to send the result back to the user. Although we want to minimize the total average transmit power, noting that $p_k = \frac{1}{a_k}(e^{\overline{R}_k}-1)$, where $\overline{R}_k = R_k \cdot \log(2)$ the problem can be formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{R},\boldsymbol{f}} \ & \sum_{k=1}^{K}(1-P_{I_k})\frac{1}{a_k}\left(e^{\overline{R}_k}-1\right) \\ \text{s.t.} \ & \frac{c_k}{(1-P_{I_k})\,\overline{R}_k}+\frac{w_k}{f_k} \leq \Delta_k, \forall k \\ & \overline{R}_k \geq 0, \forall k \\ & \overline{R}_k \leq \log(1+a_kP_T), \forall k \\ & \sum_{k=1}^{K} f_k \leq f_S; \\ & 0 \leq f_k \leq f_S, \forall k. \end{aligned} \tag{9}$$

In (9), f_k is the number of CPU cycles/sec assigned to user k. The Lagrangian associated to this constrained problem is

$$\mathcal{L} = \sum_{k=1}^{K}(1-P_{I_k})\,\frac{1}{a_k}\left(e^{\overline{R}_k}-1\right) - \sum_{k=1}^{K}\alpha_k\,\overline{R}_k + \sum_{k=1}^{K}\beta_k\,(\overline{R}_k - R_{max_k}) + \sum_{k=1}^{K}\gamma_k\Big(\frac{c_k}{(1-P_{I_k})\overline{R}_k} + \frac{w_k}{f_k} - \Delta_k\Big) - \sum_{k=1}^{K}\mu_k f_k + \sum_{k=1}^{K}\theta_k(f_k - f_S) + \nu\left(\sum_{k=1}^{K} f_k - f_S\right).$$

where $\alpha_k, \beta_k, \gamma_k, \mu_k, \theta_k$ and ν are the Lagrange multipliers associated to the constraints of the problem in (9). To find closed form solutions, we write the KKT:

$$\begin{aligned}
&a)\quad \nabla_{\overline{R}_k}\mathcal{L} = \frac{(1-P_{I_k})}{a_k}\,e^{\overline{R}_k} - \alpha_k + \beta_k + \\
&\qquad\qquad - \gamma_k\,\frac{c_k}{(1-P_{I_k})\overline{R}_k^2} = 0, \quad \forall k;\\
&b)\quad \nabla_{f_k}\mathcal{L} = -\gamma_k\,\frac{w_k}{f_k^2} - \mu_k + \theta_k + \nu = 0, \quad \forall k;\\
&c)\quad \alpha_k\,\overline{R}_k = 0, \quad \alpha_k \geq 0, \quad \overline{R}_k \geq 0, \quad \forall k;\\
&d)\quad \beta_k(\overline{R}_k - R_{max_k}) = 0, \quad \beta_k \geq 0, \quad \overline{R}_k \leq R_{max_k}, \quad \forall k;\\
&e)\quad \gamma_k\left(\frac{c_k}{(1-P_{I_k})\overline{R}_k} + \frac{w_k}{f_k} - \Delta_k\right) = 0, \quad \gamma_k \geq 0,\\
&\qquad \frac{c_k}{(1-P_{I_k})\,\overline{R}_k} + \frac{w_k}{f_k} \leq \Delta_k, \quad \forall k;\\
&f)\quad \mu_k f_k = 0, \quad \mu_k \geq 0, \quad f_k \geq 0, \quad \forall k;\\
&g)\quad \theta_k(f_k - f_S) = 0, \quad \theta_k \geq 0, \quad f_k \leq f_S, \quad \forall k;\\
&h)\quad \nu\left(\sum_{k=0}^{K} f_k - f_S\right) = 0, \quad \nu \geq 0, \quad \sum_{k=0}^{K} f_k \leq f_S. \qquad (10)
\end{aligned}$$

Considering the case the user is served, from $c)$ we set $\overline{R}_k \neq 0$ and $\alpha_k = 0$ and from $f)$, $g)$ $f_k \neq 0$ and $f_k \neq f_S$, so $\mu_k = 0$ and $\theta_k = 0$. Using b), we obtain:

$$f_k = \sqrt{\frac{\gamma_k\,w_k}{\nu}}, \quad \forall k. \qquad (11)$$

We can then see that the first term of $a)$ is different from zero and since $\beta_k \geq 0$, then $\gamma_k \neq 0$. Using the previous equation in $e)$ we get:

$$\overline{R}_k = \frac{c_k}{(1-P_{I_k})\left(\Delta_k - \sqrt{\frac{w_k\nu}{\gamma_k}}\right)}, \quad \forall k \qquad (12)$$

In $b)$, the first term is different from zero, so $\nu \neq 0$ and from h):

$$\sum_{k=1}^{K}\sqrt{\frac{\gamma_k\,w_k}{\nu}} = f_S, \qquad (13)$$

from which, we obtain:

$$f_k = \frac{\sqrt{w_k\,\gamma_k}}{\sum_{k=1}^{K}\sqrt{w_k\gamma_k}}f_S, \quad \forall k \qquad (14)$$

and (12) becomes:

$$\overline{R}_k = \frac{c_k}{(1-P_{I_k})\left(\Delta_k - \sqrt{\frac{w_k}{\gamma_k}}\sum_{k=1}^{K}\sqrt{w_k\gamma_k}/f_S\right)}, \quad \forall k. \qquad (15)$$

In general, from $a)$ we have:

$$\gamma_k = \frac{(1-P_{I_k})^2\overline{R}_k^2}{c_k a_k}\,e^{\overline{R}_k} + \beta_k\frac{(1-P_{I_k})\,\overline{R}_k^2}{c_k}, \quad \forall k. \qquad (16)$$

By replacing (16) into (15), we get a closed form expression for the rate of every user, in case of single link communications. In particular, if we consider a case in which no rate reaches the maximum value R_{max_k}, from $d)$ we can set $\beta_k = 0\ \forall k$, and from (15) we have the value for the optimal rate. Instead, in the case where one or more users need the maximum value for the rate (equivalently the power), we can set $\overline{R}_k = R_{max_k}$ in (15) and find the value of β_k. Interestingly, the expression for f_k in (14), resulting from the joint allocation of computation/communication resources, shows the difference between the joint allocation and the disjoint fair allocation, where the computation resources are allocated in proportion to the users' requests, so that

$$f_k = \frac{\omega_k}{\sum_{k=1}^{K}\omega_k}f_S.$$

V. Multi-users Multi-link communications

In this section we extend the analysis to the case of multi-link communications, where multiple users can access MEC services through multiple RAPs, labeled with numbers from 1 to N. From the computation side, only one base station is equipped with computation and storage capabilities, but all base stations are linked to each other by a high capacity backhaul, so that the small cloud server can always receive the bits necessary to run the application, with negligible latency. In such a case, considering the interference events, we must take into account all the 2^N possible configurations of base stations used during the communication, since every channel can be available or not, according to blocking events. For every possible configuration, we have a certain number of available RAPs for the offloading and the total transmit power is the sum of the powers allocated on each channel. Let us introduce the index $z_i, i = 1, \ldots, N$, such that $z_i = 1$ if the channel to the $i-th$ RAP is available, and 0 otherwise. To gain insight into the solution, it 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 channel matrix $\boldsymbol{H}$(see Eq.(3)) is obtained by stacking on top of each other the channel matrices $\boldsymbol{H}_{\bar{i}}$ with $\bar{i} \in \{i : z_i = 1\}$, since the other channels are off (see section II). It should be noted that, for every possible configuration, the powers allocated on each channel change according to the availability of the single RAP. Let us introduce another index $w_{z_1,\ldots,z_N}$ such that, the subscript of w indicates which base stations are available ($z_i = 1$). As an example, in case of three RAPs, if the first and third ones are available, we have $w_{z_1,\ldots,z_N} = w_{1,0,1}$. The problem, with K users served by N possible base stations, whose aim is to minimize the total transmit power under a latency constraint and a total power budget constraint, can be

formulated as follows:

$$\min_{\boldsymbol{p},\boldsymbol{f}} \sum_{k=1}^{K} \sum_{z_{1,k}=0}^{1} \cdots \sum_{z_{N,k}=0}^{1} P_{I_{1,k}}^{(1-z_{1,k})} \cdots P_{I_{N,k}}^{(1-z_{N,k})} (1-P_{I_{1,k}})^{z_{1,k}} \cdot \ldots \cdot (1-P_{I_{N,k}})^{z_{N,k}} \cdot \left(\sum_{i=1}^{N} p_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}}^{z_{i,k}} \cdot z_{i,k} \right) \tag{17}$$

$$\begin{aligned} \text{s.t.} \quad & \frac{c_k}{\overline{R}_k} + \frac{\omega_k}{f_k} \leq \Delta_k, \quad \forall k \\ & p_{i,k} \geq 0, \quad i = 1, \ldots, N, \quad \forall k \\ & \sum_{i=1}^{N} p_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}}^{z_{i,k}} \cdot z_{i,k} \leq P_T, \quad \forall w_{z_{1,k},\ldots,z_{N,k}}, \quad \forall k \\ & 0 \leq f_k \leq f_S, \quad \forall k \\ & \sum_{k=1}^{K} f_k \leq f_S, \end{aligned} \tag{18}$$

where $P_{I_{i,k}}$ is the blockage probability on the link between the $k-th$ user and the $i-th$ RAP, $p_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}}$ is the power allocated by the $k-th$ user to the beam pointing the $i-th$ RAP according to the index $w_{z_{1,k},\ldots,z_{N,k}}$, and $\overline{R}_k$ is the average rate given by

$$\overline{R}_k = \sum_{z_{1,k}=0}^{1} \cdots \sum_{z_{N,k}=0}^{1} P_{I_{1,k}}^{(1-z_{1,k})} \cdots P_{I_{N,k}}^{(1-z_{N,k})} \cdot (1-P_{I_{1,k}})^{z_{1,k}} \cdot \ldots \cdot (1-P_{I_{N,k}})^{z_{N,k}} \cdot \left(\sum_{i=1}^{N} \log^{z_{i,k}} \left(1 + a_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}} \cdot p_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}} \right) \cdot z_{i,k} \right) \tag{19}$$

where $a_{i,k_{w_{z_{1,k},\ldots,z_{N,k}}}}$ is the channel experienced by the $k-th$ user towards the $i-th$ RAP according to the index $w_{z_{1,k},\ldots,z_{N,k}}$. Although this problem is not easy to solve in closed form, it is convex and then it admits a unique global solution that can be found with convex optimization numerical tools. As a numerical example, in Fig.2, we compare the average transmitted power obtained with multi-link and multi-user communication system, for the same latency constraint. We consider interference as the blocking event. Firstly, we can notice the large gain obtained through a 2-link communication compared with single-link and how this gain is larger than the one between 2-link and 4-link. Indeed, as the number of RAPs increases, the gain of having more link decreases. Fig. 2 shows also the difference between the joint optimization of communication/computation resources and the disjoint allocation of computational resources (dotted curves), even if the latter ones benefit from multi-link communications. Nevertheless, the power gap with the respective joint solution is indicative of the gain resulting from the joint optimization. Moreover the disjoint allocation in case of single link (dotted red curve in Fig.2) becomes even unfeasible (there are no solutions), while the joint allocation allows us to offload at longer distances, as discussed in section IV.

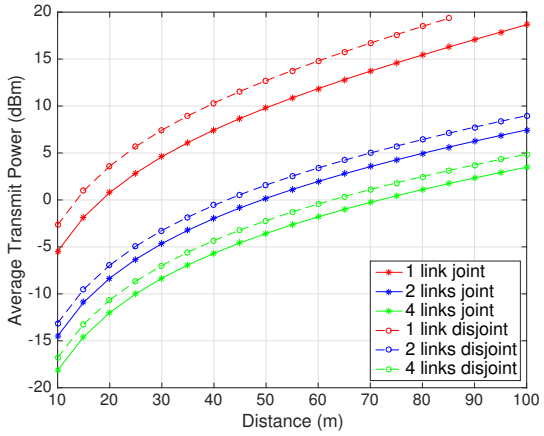

Figure 2. Transmit power vs. distance in a multi-user scenario with multi-link communications.

VI. Conclusion and Further Developments

We have proved the beneficial effects of multi-link mmW communication systems for computation offloading, starting from a single user case and then moving to the multi-user scenario. Typically, computation offloading is more advantageous for applications requiring intensive computations (e.g., large ω) and limited transfer of data (e.g., small n_b), as shown also in [8]. We have shown the improvement achievable through the joint optimization of communication/computational resources, for a multi-link multi-user system. The multi-link system helps the disjoint allocation as well, but major gains are achieved by the joint allocation in the multi-link communication. As a further development, we plan to extend our approach to the multi-server case and to optimize admission control and the choice of which are the best access points and the best server for each user, under latency constraints.

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] "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*.

[4] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, 2017.

[5] 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.

[6] 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.

[7] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," 2017, http://arxiv.org/abs/1702.00892.

[8] S. Barbarossa, E. Ceci, M. Merluzzi, and E. Calvanese-Strinati, "Enabling effective mobile edge computing using millimeter wave links," in *IEEE International Communication Conference*, May 2017.

[9] 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.