Energy Efficiency in Latency-Constrained Application Offloading from Mobile Clients to Multiple Virtual Machines

Sandra Lagen, Antonio Pascual-Iserte, *Senior Member, IEEE,* Olga Muñoz, *Member, IEEE,* and Josep Vidal, *Member, IEEE*

***Abstract*—This paper addresses the energy-latency trade-off in distributed application offloading, in which an energy-limited handset offloads totally or partially an application to one or several virtual machines (VMs) located in remote locations or access points (APs) close to the mobile terminal (MT). One of the APs (the serving AP) provides radio access to the MT and is connected to the VMs through non-ideal backhaul (BH) links. In this setting, we optimize the offloading strategy (including the joint optimization of radio and computational resources) to minimize the energy consumption at the MT subject to a maximum latency constraint. In addition, we propose robust designs to cope with imperfect acquisition of the channel state information (CSI) and the BH parameters. Our findings show that, as far as the energy-latency trade-off is concerned, the optimal order of activation of the VMs does not depend on their processing capabilities but the delays of the BH links. However, once a VM is selected to participate in the processing, the optimal amount of processing allocated to such VM depends on its computational capabilities as well as on the features (capacity and delay) of the BH link. Additionally, offloading decisions become more conservative as the uncertainty in CSI and BH parameters increases.**

***Index Terms*—application offloading, battery savings, energy efficiency, energy-latency trade-off, robust design.**

I. Introduction

The almost universal adoption of advanced smartphones has produced a direct impact on the technical requirements of networks. Mobile network operators have been compelled to adopt new standards so as to provide higher bit-rates and widespread coverage to the users [1]. One of the current trends is based on the exploitation of small cell deployments and dense self-organized networks [2]–[4]. In this framework, reducing the coverage area of each access point (AP) allows for higher area spectral efficiencies and lower transmission powers.

In addition to the previous aspects, which have been dealt extensively in the literature over the last years, others are also being contemplated. Currently, users demand not only for high bit-rates, but also high performance in the execution of complex applications on their smartphones. This requires the mobile terminals (MTs) to have high memory and computational resources, which entail a high energy consumption and a reduction of the MT battery lifetime. To solve the problem, the cloud concept for remote computing (known as mobile cloud computing (MCC)) has been proposed [5]–[7]. One possible approach for MCC is based on the use of remote clouds, such as Amazon Elastic Compute Cloud (EC2) [8]. In this regard, the management and architectural aspects of MCC have been analyzed in [9]. References like [10] and [11] perform experimental evaluations of the energy saving associated to application offloading. Works in [12]–[14] focus on the optimization of the offloading strategies by taking into account the energy cost of the radio interface (e.g., 3G or Wi-Fi), but without including the actual channel state of the radio link for data transfer optimization. The radio-cloud interaction is addressed in [15], which includes the energy cost associated to offloading when the MT is transmitting.

The main benefit of the MCC solution stems from the provision of very high storage capacities and computational resources. However, some inconveniences have to be considered. One of the main disadvantages is that these clouds may be situated far away from the end user, which entails high delays that may vary depending on the saturation of the network and the quality of the backhaul (BH) links. A possible solution to cope with this comes from the edge-cloud concept or mobile edge computing (MEC) [16], [17]. Recent surveys on architecture and computation offloading and radio-and-computational resource management for MEC can be found in [18] and [19], respectively.

The key idea of MEC is to enhance the APs, owned by the end-users or by the operators, with some computational and storage capacities. In this way, end users could offload the execution of their applications to closeby servers. In other words, the cloud is moved to the edge of the network, implying shorter distances between the MT and the computing entities and, therefore, shorter delays. According to this concept, in [20] the radio-cloud interaction in MEC is addressed by taking into account the energy cost associated to offloading towards a single serving AP and the current (perfectly acquired) channel conditions of the radio link for uplink (UL) and downlink (DL) transmissions.The radio-cloud interaction for the case in which multiple APs access to a common edge-cloud server is

Manuscript received May 18, 2017; revised September 18, 2017, November 13, 2017; accepted November 22, 2017. S. Lagen was with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona, Spain. She is now with the Mobile Networks Department, Centre Tecnològic de Telecomunicacions de Catalunya, Castelldefels, Spain (email: sandra.lagen@cttc.es). A. Pascual, O. Muñoz, and J. Vidal are with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona, Spain (emails: {antonio.pascual, olga.munoz, josep.vidal}@upc.edu).

addressed in [21]. The main drawback of the MEC approach is, however, that the external computational power and storage, while larger than those of an MT, might be significantly lower than those corresponding to classical remote clouds. In this line, authors in [22] optimize the offloading strategy considering a scenario where one AP, with computational capabilities, and a remote cloud assist the MT in executing applications.

In this paper, we study a hybrid scenario that is composed of virtual machines (VMs) running in close APs and VMs located at remote locations, which are connected through non-ideal BH links that involve higher delays. Our goal is to determine how increasing the pool of heterogeneous computational network resources improves the trade-off between latency and energy consumption at the MT. To that end, we need to answer three key questions:

1) how does having more than one available VM impact on the decision of doing offloading?,
2) is it better to do all the remote processing in the closest AP or, despite the delay of the non-ideal BH, is it better to do the processing in remote VMs if they have better processing capabilities?, and
3) does offloading and the availability of multiple VMs allow reducing effectively the energy spent by MTs?

The answers to these questions will depend on the quality of the BH links that connect the serving AP and the VMs, in terms of round-trip delays and capacities. Furthermore, different sources of imperfection might appear in the acquisition of the system parameters, namely imperfect acquisition of the channel state information (CSI) (both for UL and DL transmissions) and imperfect acquisition of the parameters that characterize the BH links (i.e., capacity and round-trip delay). To cope with the different error sources, worst-case robust designs against parameters estimation errors can be used.

By taking into account the above considerations, this paper optimizes the offloading strategy (including the joint optimization of radio and computational resources) in a hybrid scenario composed of close and remote VMs that are connected to the serving AP through non-ideal and heterogeneous BH links. The optimization aims at minimizing the total energy spent by the MT for UL, DL, and local processing, subject to a maximum latency constraint in the execution of the application. The main contributions of this work are:

- we derive the optimal order of activation of the available VMs,
- we derive the optimal distribution of the offloaded tasks among the multiple available VMs, and
- we propose robust designs for optimizing energy-latency trade-offs under imperfect acquisition of CSI (for UL and DL) and BH link parameters (capacity and delay).

This paper uses some results from [20] that addressed a scenario with a single VM deployed at the serving AP and perfect CSI acquisition. Thus, the fact of having multiple VMs (and therefore the need of dealing with their activation and the distribution of tasks), the impact of non-ideal BH links, and the imperfect acquisition of the system parameters were not considered in [20].

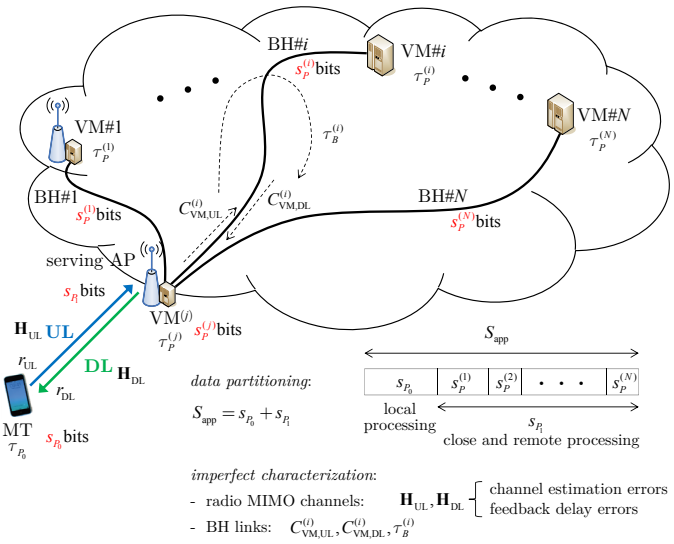

Fig. 1: Scenario for application offloading from MT to multiple VMs. The MT communicates through a radio link with the serving AP, which is connected to N VMs (either in close or remote locations) through non-ideal BH links.

Organization: The paper is organized as follows. Section II presents the system description, the energy consumption models, and the sources of imperfection. Section III proposes robust precoding designs against imperfect CSI acquisition for UL and DL transmissions. Section IV formulates and solves the complete problem for offloading optimization by including multiple VMs with non-ideal BH links. Finally, Sections V and VI present numerical results and conclusions, respectively.

Notation: In this paper, scalars are denoted by italic letters. Boldface lower-case and upper-case letters denote vectors and matrices, respectively. For given scalars a and b, $\min(a,b)$, $\max(a,b)$, $(a)^+$, $|a|$, and $\log_2(a)$, denote the minimum between a and b, the maximum between a and b, the maximum between a and 0, the modulus of a, and the base-2 logarithm, respectively. $J_0(.)$ refers to the zero-order Bessel function of the first kind. $\Pr(a{\leq}b)$ denotes the probability of a being smaller than or equal to b. For a given matrix $\mathbf{A}$, the hermitian matrix is denoted by $\mathbf{A}^H$ and $[\mathbf{A}]_{i,j}$ refers to the (i,j)-th element of the matrix. The operators $|\mathbf{A}|$, $\mathrm{Tr}(\mathbf{A})$, $\mathbb{E}\{\mathbf{A}\}$, $||\mathbf{A}||_{\mathrm{F}}$, $||\mathbf{A}||_2$, refer to the determinant, the trace, the expectation, the Frobenius norm (i.e. $||\mathbf{A}||_{\mathrm{F}}{=}\sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^H)}$), and the spectral norm (i.e. $||\mathbf{A}||_2{=}\sigma_{\max}(\mathbf{A})$), respectively. $\sigma_{\max}(\mathbf{A})$ denotes the maximum singular value of matrix $\mathbf{A}$ and $\lambda_i(\mathbf{A})$ refers to its i-th eigenvalue (when sorting eigenvalues in decreasing order). $\mathbf{A}{\succeq}\mathbf{B}$ means that matrix $\mathbf{A}{-}\mathbf{B}$ is positive semidefinite. Matrix $\mathbf{I}$ denotes the identity matrix. $\mathbb{C}^{m\times n}$ refers to an m by n dimensional complex space.

II. System Model

This section presents the complete system description, the energy consumption models at the MT for UL and DL transmissions, and the sources of imperfections with its associated error modeling.

A. System Description

We consider a general setup where N VMs, placed in neighboring APs or at remote entities in the network, are available to process the MT application either totally or partially. One of the APs (the serving AP) provides radio access for the MT and is connected to the VMs through non-ideal BH links, as shown in Fig. 1. As in [20], the complexity of the application is abstracted in terms of the computation cycles per processed bit ratio, measured in cycles/bit (see [10] for some practical values obtained from measurements), and the data load that is measured by the number of bits to be processed, S_{app}.

We focus on *data partitioned oriented applications*, for which the total amount of data to be processed (S_{app}) is known and the execution can be parallelized into processes (see [20] for a detailed description of the types of applications for which different offloading strategies are needed). Examples of data partitioned oriented applications include virus scanning, file/video compression, and face recognition. Accordingly, the total load can be split into two parts: one to be processed locally at the MT, s_{P_0}, and another to be processed remotely at the VMs, s_{P_1}. When several VMs are available, s_{P_1} can be further distributed among them and the i-th VM will process $s_P^{(i)}$ bits so that $\sum_{i=1}^{N} s_P^{(i)}=s_{P_1}$ (see Fig. 1). No precedence relations of the tasks assigned to the VMs is considered. Including precedence relations of the tasks would require other tools (see [23]), and lies out of the scope of this paper. We assume that the processing time is proportional to the number of bits to be processed (which is reasonable for data partitioned oriented applications).

The computing energy efficiency (i.e. amount of computation that can be performed with a given energy) can be measured in cycles/Joule (see [10]). To simplify the analysis we will use the energy required to process one bit at the MT, denoted by ε_{P_0}. This parameter accounts jointly for the computation to data ratio (in cycles/bit) and the computing energy efficiency (in cycles/Joule). Similarly, we define the time required to process one bit at the MT (τ_{P_0}) and the VMs ($\{\tau_P^{(i)}\}$). These parameters account jointly for the computation to data ratio (in cycles/bit) and the CPU rate (in cycles/second). We assume that the computing model for the VMs is the same as that for the MT, but that the VMs have higher computational power than the MT, i.e., higher memory, higher CPU speed, etc. The computational power impacts on the time required to complete the computation tasks, which is a key parameter in the offloading decision. The numerical values used in the simulation results section illustrate this.

In our model, the amount of bits to be exchanged between the MT and the serving AP is proportional to s_{P_1}, i.e. $s_{\text{UL}}=\beta_{\text{UL}}s_{P_1}$ for the UL and $s_{\text{DL}}=\beta_{\text{DL}}s_{P_1}$ for the DL, where β_{UL} and β_{DL} are proportionality factors that model the overhead in the radio communication. Note that these communication overhead factors could also account for additional information to be offloaded from the MT to the VMs in addition to the data, if needed, such as the program code and/or the execution state. The durations of the UL and DL communication, t_{UL} and t_{DL}, depend on the selected UL and DL rates, r_{UL} and r_{DL}. As higher rates usually imply higher power consumptions, the power consumed by the MT when communicating with the serving AP in UL and DL is, in general, a function of the communication rate, although the precise expressions depend on the consumption model considered for the MT.

For the wireless communication stage, we consider a multiple-input multiple-output (MIMO) system and, thus, we denote the number of antenna elements at the serving AP and the MT by n_{AP} and n_{MT}, respectively.

Note that the selected UL and DL rates, r_{UL} and r_{DL}, are limited by the maximum rates supported by the UL and DL channels, which are denoted by $R_{\text{UL,max}}$ and $R_{\text{DL,max}}$, respectively. $R_{\text{UL,max}}$ and $R_{\text{DL,max}}$ depend on the maximum powers of the transmitters and the specific MIMO channel conditions through Shannon's law [20]. In case of imperfect CSI conditions, the computation of $R_{\text{UL,max}}$ and $R_{\text{DL,max}}$ is derived in Section III by following a robust strategy.

Additionally, the BH link that connects the serving AP and the i-th VM (i.e. the i-th BH link) is abstracted in terms of the BH capacity (including losses due to overhead) when uploading, $\{C_{\text{VM,UL}}^{(i)}\}$, and downloading, $\{C_{\text{VM,DL}}^{(i)}\}$, plus a constant round-trip delay, $\{\tau_B^{(i)}\}$ (see Fig. 1) [24]. In case that the serving AP hosted a VM (e.g., the j-th VM), the parameters associated to such VM would be: $\tau_B^{(j)}=0$, $C_{\text{VM,UL}}^{(j)}\to\infty$, $C_{\text{VM,DL}}^{(j)}\to\infty$.

We assume that the offloading strategy optimization is performed at the AP, which has the knowledge of the CSI in DL and UL as well as the map of the available BH links and their characteristics towards close and remote VMs.

B. Energy Consumption Models

According to [20], [25], from the UL power consumption, p_{UL}, the energy spent by the MT to send s_{UL} information bits during t_{UL} seconds in the UL transmission is modeled as

$$p_{\text{UL}}t_{\text{UL}} = k_{\text{tx},1}t_{\text{UL}} + k_{\text{tx},2}t_{\text{UL}}\text{Tr}(\tilde{\mathbf{Q}}_{\text{UL}}), \tag{1}$$

where $k_{\text{tx},1}$ is a constant related to the extra power consumption for having the radio frequency and baseband transmission circuitries switched on, $k_{\text{tx},2}$ is a constant that measures the linear increase of the transmitted power consumption with the radiated power, and $\tilde{\mathbf{Q}}_{\text{UL}}\in\mathbb{C}^{n_{\text{MT}}\times n_{\text{MT}}}$ denotes the UL power transmit covariance matrix selected at the MT. Notice that, as it will be shown in Section III-B, $\tilde{\mathbf{Q}}_{\text{UL}}$ should be chosen in such a way that the communication link does support the UL rate $r_{\text{UL}}=\frac{s_{\text{UL}}}{t_{\text{UL}}}$.

In the DL, the power consumption at the MT, p_{DL}, increases with the decoding rate. Hence, the energy spent by the MT to receive s_{DL} information bits during t_{DL} seconds in the DL transmission can be modeled by [20], [26]

$$p_{\text{DL}}t_{\text{DL}} = k_{\text{rx},1}t_{\text{DL}} + k_{\text{rx},2}s_{\text{DL}}, \tag{2}$$

where $k_{\text{rx},1}$ is a constant related to the extra power consumption for having the reception circuitry switched on and $k_{\text{rx},2}$ is a constant that measures the linear increase of the power consumption with the decoding rate $r_{\text{DL}}=\frac{s_{\text{DL}}}{t_{\text{DL}}}$.

Even though these energy consumption models are fairly simple, by adjusting properly the constants $k_{\text{tx},1}$, $k_{\text{tx},2}$, $k_{\text{rx},1}$,

and $k_{\text{rx},2}$, they can provide estimations for the spent power that are close to practical measurements, as those presented in [26]. Moreover, because of their simplicity, the models in (1) and (2) make easier to capture the essential trade-offs of the problem considered, as compared to other more complex models available in the literature (e.g., [26]).

These energy consumption models will be used in Sections III and IV-C. However, as we will see in Section IV-B, the optimal distribution of the offloaded tasks among the multiple VMs is independent of the energy consumption models for the MT and, therefore, can be adopted for any energy consumption model that could be used.

Note that the energy consumption required at the MT and the AP for CSI acquisition is not considered because CSI acquisition is performed in any case for other purposes (e.g., channel equalization, precoding/decoding, DL/UL scheduling, etc.) and so the fact of making offloading does not imply having to carry out an additional CSI acquisition.

C. Sources of Imperfection

We consider two sources of imperfections in the system:

- errors in the acquisition of the CSI for wireless UL and DL transmissions and
- errors in the acquisition of the BH parameters that describe the BH capacity and the BH round-trip delay.

In particular, CSI acquisition errors will impact on the energy consumption for the wireless communication betwen the MT and serving AP, and also on the maximum UL and DL rates that can be supported, $R_{\text{UL,max}}$ and $R_{\text{DL,max}}$. On the other hand, BH parameter acquisition errors will affect the BH modeling between the serving AP and the different VMs. In what follows we present the model for every error.

1) CSI Acquisition Errors: Errors in CSI acquisition can arise either due to channel estimation errors or to feedback delay errors [27]. Channel estimation errors take place due to the presence of noise during the training phase to estimate the channel. Thus, its associated error is related to the transmit signal-to-noise ratio (SNR) that is used for channel estimation [28]. On the other hand, feedback delay errors appear due to a likely non-negligible delay between the instant in which the CSI is acquired and the instant in which the CSI is used to design and carry out the communication. In this case, the error depends on the channel coherence time and the aforementioned time delay [29].

Accordingly, instead of taking the estimated channels as perfect channel estimates (naive approach) for UL and DL, $\hat{\mathbf{H}}_{\text{UL}}\in\mathbb{C}^{n_{\text{AP}}\times n_{\text{MT}}}$ and $\hat{\mathbf{H}}_{\text{DL}}\in\mathbb{C}^{n_{\text{MT}}\times n_{\text{AP}}}$, respectively, we consider that the actual channel matrices $\mathbf{H}_{\text{UL}}\in\mathbb{C}^{n_{\text{AP}}\times n_{\text{MT}}}$ and $\mathbf{H}_{\text{DL}}\in\mathbb{C}^{n_{\text{MT}}\times n_{textAP}}$ can be written as

$$\mathbf{H}_{\text{UL}} = \hat{\mathbf{H}}_{\text{UL}} + \mathbf{\Delta}_{\text{UL}}, \quad \mathbf{H}_{\text{DL}} = \hat{\mathbf{H}}_{\text{DL}} + \mathbf{\Delta}_{\text{DL}}, \tag{3}$$

where $\hat{\mathbf{H}}_{\text{UL}}$ and $\hat{\mathbf{H}}_{\text{DL}}$ refer to the imperfect channel estimates, and $\mathbf{\Delta}_{\text{UL}}$ and $\mathbf{\Delta}_{\text{DL}}$ denote the channel estimation errors.

***Lemma** 1:* Assume that the actual channel matrices ($\mathbf{H}_{\text{UL}}$ and $\mathbf{H}_{\text{DL}}$) are composed of independent and identically distributed (*i.i.d.*) complex circularly symmetric Gaussian components with zero mean and variance $\sigma^2_{h,\text{UL}}$ and $\sigma^2_{h,\text{DL}}$, respectively, and that the channel estimates ($\hat{\mathbf{H}}_{\text{UL}}$ and $\hat{\mathbf{H}}_{\text{DL}}$) are obtained by using the minimum mean square error (MMSE) Bayesian approach [30] based on a training phase with *i.i.d.* complex circularly symmetric Gaussian noise. Then, according to the model in (3), the statistics of the actual channels ($\mathbf{H}_{\text{UL}}$ and $\mathbf{H}_{\text{DL}}$) conditioned on the observations in the training phase follow a Gaussian distribution with mean $\hat{\mathbf{H}}_{\text{UL}}$ and $\hat{\mathbf{H}}_{\text{DL}}$ and variances σ^2_{UL} and σ^2_{DL}, respectively. Therefore, the components of the channel estimation errors $\mathbf{\Delta}_{\text{UL}}$=$\mathbf{H}_{\text{UL}}$−$\hat{\mathbf{H}}_{\text{UL}}$ and $\mathbf{\Delta}_{\text{DL}}$=$\mathbf{H}_{\text{DL}}$−$\hat{\mathbf{H}}_{\text{DL}}$ in (3) are also *i.i.d.* complex circularly symmetric Gaussian with zero mean and variances σ^2_{UL} and σ^2_{DL}, respectively. The variances σ^2_{UL} and σ^2_{DL} depend on the quality of the channel estimates and the imperfections that generate the errors during the CSI acquisition process (i.e. channel estimation errors and feedback delay errors). More precisely, assuming that the channel temporal evolution follows Jake's model [29], the variances are given by:

$$\sigma^2_{\text{UL}} = \sigma^2_{h,\text{UL}}\left(\frac{1+\gamma_{\text{UL}}\sigma^2_{h,\text{UL}}(1-J_0^2(2\pi f_{\text{d,UL}}t_{\text{del,UL}}))}{1+\gamma_{\text{UL}}\sigma^2_{h,\text{UL}}}\right), \tag{4}$$

$$\sigma^2_{\text{DL}} = \sigma^2_{h,\text{DL}}\left(\frac{1+\gamma_{\text{DL}}\sigma^2_{h,\text{DL}}(1-J_0^2(2\pi f_{\text{d,DL}}t_{\text{del,DL}}))}{1+\gamma_{\text{DL}}\sigma^2_{h,\text{DL}}}\right), \tag{5}$$

where γ_{UL} and γ_{DL} denote the transmit SNR for channel estimation in UL and DL, respectively, $f_{\text{d,UL}}$ and $f_{\text{d,DL}}$ are the Doppler frequencies in UL and DL, respectively, and $t_{\text{del,UL}}$ and $t_{\text{del,DL}}$ refer to the time delay in UL and DL, respectively.

Proof: See Appendix A. ∎

Motivated by these results, we define uncertainty regions for the channel estimation errors in UL and DL, respectively, through the following spheres:

$$\|\mathbf{\Delta}_{\text{UL}}\|_{\text{F}} \le \epsilon_{\text{UL}}, \quad \|\mathbf{\Delta}_{\text{DL}}\|_{\text{F}} \le \epsilon_{\text{DL}}. \tag{6}$$

As channel estimation errors are *i.i.d.* Gaussian distributed, they will be inside the uncertainty regions with a certain probability p_{in}<1, i.e. $\Pr(\|\mathbf{\Delta}_{\text{UL}}\|_{\text{F}}\le\epsilon_{\text{UL}})$=$p_{\text{in}}$ and $\Pr(\|\mathbf{\Delta}_{\text{DL}}\|_{\text{F}}\le\epsilon_{\text{DL}})$=$p_{\text{in}}$. The probability of belonging to the regions (p_{in}) is related to the size of the uncertainty regions as follows:

$$\epsilon_{\text{UL}} = \sqrt{\tfrac{1}{2}\phi^{-1}(p_{\text{in}})\sigma^2_{\text{UL}}}, \quad \epsilon_{\text{DL}} = \sqrt{\tfrac{1}{2}\phi^{-1}(p_{\text{in}})\sigma^2_{\text{DL}}}, \tag{7}$$

where σ^2_{UL} and σ^2_{DL} are the ones in (4)-(5) and $\phi(.)$ is the cumulative density function of the chi-square distribution with $2n_{\text{AP}}n_{\text{MT}}$ degrees of freedom (since channel estimation errors are composed of $n_{\text{AP}}n_{\text{MT}}$ components, each with independent real and imaginary parts) [31].

Note that the uncertainty regions in (6) take into account the quality of the channel estimate and the imperfections that generate the errors. Furthermore, as shown in (7), the size of these uncertainty regions is larger as the quality of the CSI decreases (e.g., as the transmit SNR for channel estimation diminishes). CSI acquisition errors will be addressed in Section III to derive robust precoding designs for optimizing the energy-latency trade-offs under imperfect CSI conditions.

2) BH Capacity and Delay Acquisition Errors: To properly model BH links, acquisition of the BH capacities and the BH round-trip delays is required. There are many techniques to

estimate different characteristics (bandwidth, capacity, delay) of BH links. In particular, to estimate the end-to-end capacity of a link, the packet pair/train dispersion (PPTD) technique [32] can be used. PPTD estimates the capacity of a link from the dispersion (spacing) experienced by multiple packet pairs. Even though it is simple in principle, PPTD technique produces erroneous estimates mainly due to the presence of cross traffic in the link, which either increases or decreases the capacity estimate [33]. In order to mitigate the effect of cross traffic, a packed-pair delay tracking is introduced in [34], whereby only the dispersion of packet pairs with minimum end-to-end delay are used for capacity estimation. In this way, errors lower than a 10% are obtained. On the other hand, to estimate BH round-trip time delays, time stamps on packet trains can be used and, therefore, the same impairments as for BH capacity estimation arise (i.e. cross traffic) [32].

Based on the results described in the previous papers, instead of taking the BH capacity estimation when uploading and downloading towards the i-th BH link, $\hat{C}^{(i)}_{\text{VM,UL}}$ and $\hat{C}^{(i)}_{\text{VM,DL}}$, respectively, as perfect (naive approach), we consider that the actual BH capacities, $C^{(i)}_{\text{VM,UL}}$ and $C^{(i)}_{\text{VM,DL}}$, are expressed as

$$C^{(i)}_{\text{VM,UL}} = \hat{C}^{(i)}_{\text{VM,UL}} + \Delta^{(i)}_{\text{C,UL}},\ C^{(i)}_{\text{VM,DL}} = \hat{C}^{(i)}_{\text{VM,DL}} + \Delta^{(i)}_{\text{C,DL}}, \quad (8)$$

where $\Delta^{(i)}_{\text{C,UL}}$ and $\Delta^{(i)}_{\text{C,DL}}$ denote the BH capacity estimation errors. We assume that these errors are bounded as

$$|\Delta^{(i)}_{\text{C,UL}}| \leq \epsilon^{(i)}_{\text{C,UL}}, \quad |\Delta^{(i)}_{\text{C,DL}}| \leq \epsilon^{(i)}_{\text{C,DL}}. \quad (9)$$

For example, according to [34], $\epsilon^{(i)}_{\text{C,UL}}$ and $\epsilon^{(i)}_{\text{C,DL}}$ can be considered to be a 10% of the nominal value under PPTD with packet-pair delay tracking.

Similarly, the actual round-trip delay for the i-th BH link ($\tau^{(i)}_{\text{B}}$) can be expressed as a function of the BH round-trip delay estimate ($\hat{\tau}^{(i)}_{\text{B}}$) as

$$\tau^{(i)}_{\text{B}} = \hat{\tau}^{(i)}_{\text{B}} + \Delta^{(i)}_{\tau}, \quad (10)$$

where $\Delta^{(i)}_{\tau}$ denotes the BH round-trip delay estimation error and we assume it to be bounded as

$$|\Delta^{(i)}_{\tau}| \leq \epsilon^{(i)}_{\tau}. \quad (11)$$

BH capacity and BH round-trip delay acquisition errors will be incorporated in Section IV-A within the global problem formulation.

Finally, let us remark that the uncertainty regions defined for CSI in (6), BH capacity in (9), and BH round-trip delay in (11), address errors in the acquisition of the system parameters. Nevertheless, these uncertainty regions could be properly expanded to cover variations of the system parameters according to some temporal evolution. In this case, the robust designs presented in what follows could cover random variations of the system parameters that could come up between the start and the end of the application offloading procedure.

III. Robust Precoder Design for Optimizing Energy-Latency Trade-offs

This section proposes robust precoder designs for optimizing the energy-latency trade-offs in DL and UL transmissions under imperfect CSI conditions. To deal with imperfect CSI acquisition, we use worst-case robust designs, for which the worst channel estimation error that satisfies the bound in (6) is considered [35]. We aim at optimizing the DL and UL transmit covariance matrices for fixed values of s_{DL}, t_{DL}, s_{UL}, and t_{UL}. The design of s_{DL}, t_{DL}, s_{UL}, and t_{UL}, will be later addressed in Section IV through the complete offloading problem statement, which is formulated based on the optimal transmit covariance matrices and the energy functions that we derive in this section.

Before proceeding let us recall that, for any matrix $\mathbf{A}$, the spectral norm $\|\mathbf{A}\|_2{=}\sigma_{\max}(\mathbf{A})$ satisfies $\|\mathbf{A}\|_2{\leq}\|\mathbf{A}\|_{\text{F}}$ [36, Sect. 10.3]. Therefore, the uncertainty regions defined in (6) are included in the following uncertainty regions:

$$\sigma_{\max}(\mathbf{\Delta}_{\text{UL}}) \leq \epsilon_{\text{UL}}, \quad \sigma_{\max}(\mathbf{\Delta}_{\text{DL}}) \leq \epsilon_{\text{DL}}. \quad (12)$$

The uncertainty regions in (12) are larger and contain the ones in (6). However, they would allow us to derive simple hyper robust design solutions in closed-form by applying the framework developed in [37].

Let us define $\mathbf{G}_{\text{UL}}{=}\mathbf{H}^H_{\text{UL}}\mathbf{H}_{\text{UL}}$, $\mathbf{G}_{\text{DL}}{=}\mathbf{H}^H_{\text{DL}}\mathbf{H}_{\text{DL}}$, $\hat{\mathbf{G}}_{\text{UL}}{=}\hat{\mathbf{H}}^H_{\text{UL}}\hat{\mathbf{H}}_{\text{UL}}$ and $\hat{\mathbf{G}}_{\text{DL}}{=}\hat{\mathbf{H}}^H_{\text{DL}}\hat{\mathbf{H}}_{\text{DL}}$, so that

$$\mathbf{G}_{\text{UL}} = \hat{\mathbf{G}}_{\text{UL}} + \mathbf{G}_{\Delta,\text{UL}}, \quad \mathbf{G}_{\text{DL}} = \hat{\mathbf{G}}_{\text{DL}} + \mathbf{G}_{\Delta,\text{DL}}, \quad (13)$$

where $\mathbf{G}_{\Delta,\text{UL}}{=}\hat{\mathbf{H}}^H_{\text{UL}}\mathbf{\Delta}_{\text{UL}}{+}\mathbf{\Delta}^H_{\text{UL}}\hat{\mathbf{H}}_{\text{UL}}{+}\mathbf{\Delta}^H_{\text{UL}}\mathbf{\Delta}_{\text{UL}}$ and $\mathbf{G}_{\Delta,\text{DL}}{=}\hat{\mathbf{H}}^H_{\text{DL}}\mathbf{\Delta}_{\text{DL}}{+}\mathbf{\Delta}^H_{\text{DL}}\hat{\mathbf{H}}_{\text{DL}}{+}\mathbf{\Delta}^H_{\text{DL}}\mathbf{\Delta}_{\text{DL}}$. By using eigenvalue inequalities related to the sum and the product of matrices, it is shown in [37, Sect. 7.3.1] that the singular value bounds in (12) can be transformed into the following bounds for the eigenvalues of the matrices $\mathbf{G}_{\Delta,\text{UL}}$ and $\mathbf{G}_{\Delta,\text{DL}}$ in (13):

$$|\lambda_i(\mathbf{G}_{\Delta,\text{UL}}) - \epsilon^2_{\text{UL}}| \leq \epsilon_{\text{G,UL}}, \quad |\lambda_i(\mathbf{G}_{\Delta,\text{DL}}) - \epsilon^2_{\text{DL}}| \leq \epsilon_{\text{G,DL}}, \quad (14)$$

where

$$\epsilon_{\text{G,UL}} = 2\epsilon_{\text{UL}}\sigma_{\max}(\hat{\mathbf{H}}_{\text{UL}}), \quad \epsilon_{\text{G,DL}} = 2\epsilon_{\text{DL}}\sigma_{\max}(\hat{\mathbf{H}}_{\text{DL}}). \quad (15)$$

By using the relations in (13) and the following eigenvalue decompositions $\hat{\mathbf{G}}_{\text{UL}}{=}\hat{\mathbf{V}}_{\text{UL}}\hat{\mathbf{D}}_{\text{UL}}\hat{\mathbf{V}}^H_{\text{UL}}$ and $\hat{\mathbf{G}}_{\text{DL}}{=}\hat{\mathbf{V}}_{\text{DL}}\hat{\mathbf{D}}_{\text{DL}}\hat{\mathbf{V}}^H_{\text{DL}}$, it can be shown that the bounds in (14) imply that the eigenvalues of $\mathbf{G}_{\text{UL}}$ and $\mathbf{G}_{\text{DL}}$ are lower-bounded by:

$$\lambda_i(\mathbf{G}_{\text{UL}}) \geq (\lambda_i(\hat{\mathbf{G}}_{\text{UL}}) - \epsilon_{\text{G,UL}} - \epsilon^2_{\text{UL}})^+, \quad (16)$$

$$\lambda_i(\mathbf{G}_{\text{DL}}) \geq (\lambda_i(\hat{\mathbf{G}}_{\text{DL}}) - \epsilon_{\text{G,DL}} - \epsilon^2_{\text{DL}})^+. \quad (17)$$

Equivalently, the bounds in (16)-(17) can be compactly expressed as

$$\mathbf{G}_{\text{UL}} \succeq \hat{\mathbf{V}}_{\text{UL}}(\hat{\mathbf{D}}_{\text{UL}} - (\epsilon_{\text{G,UL}}{+}\epsilon^2_{\text{UL}})\mathbf{I})^+\hat{\mathbf{V}}^H_{\text{UL}} = \breve{\mathbf{G}}_{\text{UL}}, \quad (18)$$

$$\mathbf{G}_{\text{DL}} \succeq \hat{\mathbf{V}}_{\text{DL}}(\hat{\mathbf{D}}_{\text{DL}} - (\epsilon_{\text{G,DL}}{+}\epsilon^2_{\text{DL}})\mathbf{I})^+\hat{\mathbf{V}}^H_{\text{DL}} = \breve{\mathbf{G}}_{\text{DL}}. \quad (19)$$

Therefore, worst-case robust designs with imperfect CSI can be obtained in practice simply by using the lower bounds $\breve{\mathbf{G}}_{\text{UL}}$ and $\breve{\mathbf{G}}_{\text{DL}}$ in lieu of $\mathbf{G}_{\text{UL}}$ and $\mathbf{G}_{\text{DL}}$, respectively, as we will see in the forthcoming subsections.

A. Robust Design for DL

In the DL transmission, a pair of values of s_{DL} and t_{DL} is admissible whenever the corresponding rate $r_{\text{DL}}{=}\frac{s_{\text{DL}}}{t_{\text{DL}}}$ is

supported for the worst case-scenario (i.e. the worst channel). In this setup, our objective is to maximize such rate for the worst channel by using a robust design of the DL power transmit covariance matrix, so that the set of admissible values for s_{DL} and t_{DL} is enlarged. Accordingly, for any value of s_{DL} and t_{DL}, the robust design for DL with imperfect CSI can be formulated, including the error modeling in (3)-(6), as:

$$\begin{aligned} \underset{\tilde{\mathbf{Q}}_{\mathrm{DL}}}{\text{maximize}} \ & \min_{\boldsymbol{\Delta}_{\mathrm{DL}}} \ W_{\mathrm{DL}} \log_2 |\mathbf{I} + \mathbf{H}_{\mathrm{DL}} \tilde{\mathbf{Q}}_{\mathrm{DL}} \mathbf{H}_{\mathrm{DL}}^H| \\ \text{subject to} \ & \mathrm{A1}: \ \mathrm{Tr}(\tilde{\mathbf{Q}}_{\mathrm{DL}}) \leq P_{\mathrm{tx,AP}}, \\ & \mathrm{A2}: \ \tilde{\mathbf{Q}}_{\mathrm{DL}} \succeq \mathbf{0}, \\ & \mathrm{A3}: \ \mathbf{H}_{\mathrm{DL}} = \hat{\mathbf{H}}_{\mathrm{DL}} + \boldsymbol{\Delta}_{\mathrm{DL}}, \\ & \mathrm{A4}: \ \|\boldsymbol{\Delta}_{\mathrm{DL}}\|_{\mathrm{F}} \leq \epsilon_{\mathrm{DL}}, \end{aligned} \tag{20}$$

where we have modeled the rate according to the Shannon's law, $\tilde{\mathbf{Q}}_{\mathrm{DL}}$ represents the DL power transmit covariance matrix at the serving AP, $P_{\mathrm{tx,AP}}$ denotes the transmit power available at the serving AP, and W_{DL} refers to the available bandwidth. Equivalently, problem (20) can be written more compactly through the use of the definition in (13) and the bound in (14) for DL, as:

$$\begin{aligned} \underset{\tilde{\mathbf{Q}}_{\mathrm{DL}}}{\text{maximize}} \ & \min_{\mathbf{G}_{\Delta,\mathrm{DL}}} \ W_{\mathrm{DL}} \log_2 |\mathbf{I} + \tilde{\mathbf{Q}}_{\mathrm{DL}} \mathbf{G}_{\mathrm{DL}}| \\ \text{subject to} \ & \mathrm{A1}, \mathrm{A2}, \\ & \mathrm{A3}: \ \mathbf{G}_{\mathrm{DL}} = \hat{\mathbf{G}}_{\mathrm{DL}} + \mathbf{G}_{\Delta,\mathrm{DL}}, \\ & \mathrm{A4}: \ |\lambda_i(\mathbf{G}_{\Delta,\mathrm{DL}}) - \epsilon_{\mathrm{DL}}^2| \leq \epsilon_{\mathrm{G,DL}}, \\ & \mathrm{A5}: \ \mathbf{G}_{\mathrm{DL}} = \mathbf{G}_{\mathrm{DL}}^H \succeq \mathbf{0}. \end{aligned} \tag{21}$$

Now, since the objective function is nondecreasing by definition, we can use the lower bound for $\mathbf{G}_{\mathrm{DL}}$ in (19) to finally write problem (21) as:

$$\begin{aligned} \underset{\tilde{\mathbf{Q}}_{\mathrm{DL}}}{\text{maximize}} \ & W_{\mathrm{DL}} \log_2 |\mathbf{I} + \tilde{\mathbf{Q}}_{\mathrm{DL}} \breve{\mathbf{G}}_{\mathrm{DL}}| \\ \text{subject to} \ & \mathrm{A1}, \mathrm{A2}. \end{aligned} \tag{22}$$

Problem (22) is a convex problem that is equivalent to problem (14) in [20] except that $\mathbf{G}_{\mathrm{DL}}$ has been replaced by $\breve{\mathbf{G}}_{\mathrm{DL}}$. Furthermore, the optimal solution to problem (22) ($\tilde{\mathbf{Q}}_{\mathrm{DL}}^{\star}$) can be obtained through a simple water-filling [38]. Consequently, the maximum DL rate $R_{\mathrm{DL,max}}$ is given by the optimal value of the objective function in (22), i.e. $R_{\mathrm{DL,max}}{=}W_{\mathrm{DL}} \log_2 |\mathbf{I}{+}\tilde{\mathbf{Q}}_{\mathrm{DL}}^{\star}\breve{\mathbf{G}}_{\mathrm{DL}}|$. The obtained $R_{\mathrm{DL,max}}$ will be lower as compared to the perfect CSI case due to the robust design, since the eigenvalues of $\breve{\mathbf{G}}_{\mathrm{DL}}$ are lowered according to the estimation error variance in DL (see (19)).

Summarizing, $e_{\mathrm{DL}}(s_{\mathrm{DL}}, t_{\mathrm{DL}}){=}k_{\mathrm{rx},1}t_{\mathrm{DL}}{+}k_{\mathrm{rx},2}s_{\mathrm{DL}}$ (see (2)) is a valid expression for the DL energy consumption whenever $r_{\mathrm{DL}}{\leq}R_{\mathrm{DL,max}}$, with $R_{\mathrm{DL,max}}$ obtained from (22). In addition, note that the DL energy consumption normalized by the number of bits to be received reduces to $\bar{e}_{\mathrm{DL}}(r_{\mathrm{DL}}){=}\frac{e_{\mathrm{DL}}(s_{\mathrm{DL}},t_{\mathrm{DL}})}{s_{\mathrm{DL}}}{=}\frac{k_{\mathrm{rx},1}}{r_{\mathrm{DL}}}{+}k_{\mathrm{rx},2}$ and depends only on the DL rate, r_{DL}.

B. Robust Design for UL

In the UL transmission, for a fixed value of s_{UL} and t_{UL}, the worst-case design of the UL transmit covariance matrix with imperfect CSI is formulated to minimize the UL energy consumption at the MT subject to a maximum transmit power constraint and the fact that the UL rate should be lower than the rate supported by the worst channel, including the error modeling in (3)-(6):

$$\begin{aligned} \underset{\mathbf{Q}_{\mathrm{UL}}, \tau_{\mathrm{UL}}}{\text{minimize}} \ & k_{\mathrm{tx},1}\tau_{\mathrm{UL}} + k_{\mathrm{tx},2}\mathrm{Tr}(\mathbf{Q}_{\mathrm{UL}}) \\ \text{subject to} \ & \mathrm{B1}: s_{\mathrm{UL}} \leq W_{\mathrm{UL}}\tau_{\mathrm{UL}} \min_{\boldsymbol{\Delta}_{\mathrm{UL}}} \log_2 |\mathbf{I} + \tfrac{1}{\tau_{\mathrm{UL}}}\mathbf{H}_{\mathrm{UL}}\mathbf{Q}_{\mathrm{UL}}\mathbf{H}_{\mathrm{UL}}^H|, \\ & \mathrm{B2}: \tau_{\mathrm{UL}} = t_{\mathrm{UL}}, \\ & \mathrm{B3}: \mathrm{Tr}(\mathbf{Q}_{\mathrm{UL}}) \leq \tau_{\mathrm{UL}} P_{\mathrm{tx,MT}}, \\ & \mathrm{B4}: \mathbf{Q}_{\mathrm{UL}} \succeq \mathbf{0}, \\ & \mathrm{B5}: \mathbf{H}_{\mathrm{UL}} = \hat{\mathbf{H}}_{\mathrm{UL}} + \boldsymbol{\Delta}_{\mathrm{UL}}, \\ & \mathrm{B6}: \|\boldsymbol{\Delta}_{\mathrm{UL}}\|_{\mathrm{F}} \leq \epsilon_{\mathrm{UL}}, \end{aligned} \tag{23}$$

where $\mathbf{Q}_{\mathrm{UL}}$ denotes the UL energy covariance matrix, $P_{\mathrm{tx,MT}}$ denotes the maximum transmit power at the MT, and W_{UL} refers to the available bandwidth. Note that $\mathbf{Q}_{\mathrm{UL}}$ is related to the UL power transmit covariance matrix, $\tilde{\mathbf{Q}}_{\mathrm{UL}}$, through $\mathbf{Q}_{\mathrm{UL}}{=}t_{\mathrm{UL}}\tilde{\mathbf{Q}}_{\mathrm{UL}}$. More compactly, problem (23) can be formulated through the use of the definition in (13) and the bound in (14) for UL as:

$$\begin{aligned} \underset{\mathbf{Q}_{\mathrm{UL}}, \tau_{\mathrm{UL}}}{\text{minimize}} \ & k_{\mathrm{tx},1}\tau_{\mathrm{UL}} + k_{\mathrm{tx},2}\mathrm{Tr}(\mathbf{Q}_{\mathrm{UL}}) \\ \text{subject to} \ & \mathrm{B1}: s_{\mathrm{UL}} \leq W_{\mathrm{UL}}\tau_{\mathrm{UL}} \min_{\mathbf{G}_{\Delta,\mathrm{UL}}} \log_2 |\mathbf{I} + \tfrac{1}{\tau_{\mathrm{UL}}}\mathbf{Q}_{\mathrm{UL}}\mathbf{G}_{\mathrm{UL}}|, \\ & \mathrm{B2}, \mathrm{B3}, \mathrm{B4}, \\ & \mathrm{B5}: \mathbf{G}_{\mathrm{UL}} = \hat{\mathbf{G}}_{\mathrm{UL}} + \mathbf{G}_{\Delta,\mathrm{UL}}, \\ & \mathrm{B6}: |\lambda_i(\mathbf{G}_{\Delta,\mathrm{UL}}) - \epsilon_{\mathrm{UL}}^2| \leq \epsilon_{\mathrm{G,UL}}, \\ & \mathrm{B7}: \mathbf{G}_{\mathrm{UL}} = \mathbf{G}_{\mathrm{UL}}^H \succeq \mathbf{0}. \end{aligned} \tag{24}$$

Now, we can use the lower bound for $\mathbf{G}_{\mathrm{UL}}$ in (19) to finally write problem (24) as:

$$\begin{aligned} \underset{\mathbf{Q}_{\mathrm{UL}}, \tau_{\mathrm{UL}}}{\text{minimize}} \ & k_{\mathrm{tx},1}\tau_{\mathrm{UL}} + k_{\mathrm{tx},2}\mathrm{Tr}(\mathbf{Q}_{\mathrm{UL}}) \\ \text{subject to} \ & \mathrm{B1}: \ s_{\mathrm{UL}} \leq W_{\mathrm{UL}}\tau_{\mathrm{UL}} \log_2 |\mathbf{I} + \tfrac{1}{\tau_{\mathrm{UL}}}\mathbf{Q}_{\mathrm{UL}}\breve{\mathbf{G}}_{\mathrm{UL}}|, \\ & \mathrm{B2}, \mathrm{B3}, \mathrm{B4}. \end{aligned} \tag{25}$$

Problem (25) is similar to problem (11) in [20] except that $\mathbf{G}_{\mathrm{UL}}$ has been replaced by $\breve{\mathbf{G}}_{\mathrm{UL}}$ and that we have included explicitly the transmit power constraint B3. Whenever problem (25) is feasible, the optimal UL energy covariance matrix ($\mathbf{Q}_{\mathrm{UL}}^{\star}$) can be obtained by applying a water-filling among the eigenmodes of $\breve{\mathbf{G}}_{\mathrm{UL}}$ (whose eigenvalues are $(\lambda_i(\breve{\mathbf{G}}_{\mathrm{UL}}) - \epsilon_{\mathrm{G,UL}} - \epsilon_{\mathrm{UL}}^2)^+$, see (17)) so that B1 is fulfilled with equality.

In the context of problem (25), we define the maximum UL rate $R_{\mathrm{UL,max}}$ as the maximum value of $r_{\mathrm{UL}}{=}\frac{s_{\mathrm{UL}}}{t_{\mathrm{UL}}}$ for which problem (25) is feasible. Since more energy is needed in the robust case, the value of $R_{\mathrm{UL,max}}$ will be lower than the one of the perfect CSI case. The value of $R_{\mathrm{UL,max}}$ in the robust case can be calculated based on the expressions (16)-(17) in [20] by substituting λ_i by $(\lambda_i(\hat{\mathbf{G}}_{\mathrm{UL}}) - \epsilon_{\mathrm{G,UL}} - \epsilon_{\mathrm{UL}}^2)^+$ (in other words, $R_{\mathrm{UL,max}}$ is the rate obtained when all the available transmit

power is used).

The cost function of problem in (25) evaluated at $\mathbf{Q}_{\text{UL}}^{\star}$ and $\tau_{\text{UL}}^{\star}$ depends on the parameters t_{UL} and s_{UL} that appear in the constraints, and will be denoted by $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})=k_{\text{tx},1}\tau_{\text{UL}}^{\star}+k_{\text{tx},2}\text{Tr}(\mathbf{Q}_{\text{UL}}^{\star})$ (see (1)). According to [39, Sec. 5.6.1], this function is jointly convex w.r.t. t_{UL} and s_{UL}. Furthermore, the UL energy consumption normalized by the number of bits to be transmitted depends only on the UL rate r_{UL}, i.e. $\bar{e}_{\text{UL}}(r_{\text{UL}})=\frac{e_{\text{UL}}(t_{\text{UL}},s_{\text{UL}})}{s_{\text{UL}}}$, and has a single minimum (see Lemma 2 of [20] for an equivalent proof). In the following, we will use $\check{R}_{\text{UL}}$ to denote the UL rate that minimizes $\bar{e}_{\text{UL}}(r_{\text{UL}})$.

IV. Offloading Optimization

In this section we formulate and solve the complete problem for offloading optimization by including multiple VMs with non-ideal BH links. First, in Section IV-A the complete problem is formulated. Then, in Section IV-B, for a given offloaded load, the optimal load distribution among the VMs is derived, which is independent of the energy consumption model of the MT. Next, in Section IV-C, this result is used to compute the solution of the complete problem. Finally, complexity analysis is included in Section IV-D.

A. Global Problem Statement

We focus on minimizing the total energy spent by the MT subject to a maximum latency in the execution of the application, $L_{\max}$, which is associated to a certain quality of experience perceived by the user. The energy consumed by the MT is the sum of the energy consumed for UL transmission, $e_{\text{UL}}(t_{\text{UL}}, \beta_{\text{UL}}s_{P_1})$, local processing, $\varepsilon_{P_0}s_{P_0}$, and DL transmission, $e_{\text{DL}}(t_{\text{DL}}, \beta_{\text{DL}}s_{P_1})$.

By taking into account the previous definitions, the problem can be formulated as follows[1]:

$$
\begin{aligned}
\underset{\substack{s_{P_0},s_{P_1},\{s_P^{(i)}\},\\ t_{\text{UL}},t_{\text{DL}}}}{\text{minimize}} \quad & e_{\text{UL}}(t_{\text{UL}}, \beta_{\text{UL}}s_{P_1}) + \varepsilon_{P_0}s_{P_0} + e_{\text{DL}}(t_{\text{DL}}, \beta_{\text{DL}}s_{P_1}) \\
\text{subject to} \quad & \text{C1}: s_{P_0} + s_{P_1} = S_{\text{app}}, \\
& \text{C2}: \tau_{P_0}s_{P_0} \le L_{\max}, \\
& \text{C3}: \beta_{\text{UL}}s_{P_1} \le t_{\text{UL}}R_{\text{UL,max}}, \\
& \text{C4}: \beta_{\text{DL}}s_{P_1} \le t_{\text{DL}}R_{\text{DL,max}}, \\
& \text{C5}: \textstyle\sum_{i=1}^{N} s_P^{(i)} = s_{P_1}, \\
& \text{C6}: t_{\text{UL}}+\overline{\tau}_P^{(i)}s_P^{(i)}+\tau_B^{(i)}+t_{\text{DL}} \le L_{\max} \text{ if } s_P^{(i)} > 0, \\
& \text{C7}: s_{P_0}, s_{P_1}, \{s_P^{(i)}\}, t_{\text{UL}}, t_{\text{DL}} \ge 0,
\end{aligned}
\tag{26}
$$

where $\overline{\tau}_P^{(i)} \triangleq \frac{1}{C_{\text{VM,UL}}^{(i)}}+\frac{1}{C_{\text{VM,DL}}^{(i)}}+\tau_P^{(i)}$, N is the number of VMs, and $R_{\text{DL,max}}$ and $R_{\text{UL,max}}$ refer to the DL and UL maximum rates obtained in Sections III-A and III-B, respectively, according to the robust designs for imperfect CSI conditions.

If imperfect acquisition of BH parameters is considered, then $C_{\text{VM,UL}}^{(i)}$, $C_{\text{VM,DL}}^{(i)}$, and $\tau_B^{(i)}$ in C6 should be replaced by its worst-case values. According to the model for BH parameters acquisition errors presented in Section II-C2, the worst-case design is obtained by using $\hat{C}_{\text{VM,UL}}^{(i)}-\epsilon_{\text{C,UL}}^{(i)}$, $\hat{C}_{\text{VM,DL}}^{(i)}-\epsilon_{\text{C,DL}}^{(i)}$, and $\hat{\tau}_B^{(i)}+\epsilon_{\tau}^{(i)}$ in lieu of $C_{\text{VM,UL}}^{(i)}$, $C_{\text{VM,DL}}^{(i)}$, and $\tau_B^{(i)}$, respectively.

Constraint C5 indicates that the offloaded bits, s_{P_1}, are distributed among all the VMs available. However, only those VMs that receive a load $s_P^{(i)}>0$ will be actually active.

Constraints C6 capture the latency constraints associated to the processing in active VMs. In addition to the UL and DL transmission time, the latency of the processing at each active VM depends on the capacity of the i-th BH link in UL and DL. The computation of the latency must include also the time required for the computation in the i-th VM and the fixed round-trip delay $\tau_B^{(i)}$ of the i-th BH link. Note that, according to C6, the serving AP will collect the results from the active VMs and, once all the output bits have been received, it will forward them to the MT through the wireless DL channel. Accordingly, the overall latency experienced by the MT is given by: $L=\max(s_{P_0}\tau_{P_0}, t_{\text{UL}}+\max_i(\overline{\tau}_P^{(i)}s_P^{(i)}+\tau_B^{(i)})+t_{\text{DL}})$.

The previous formulation includes as a particular case the situation in which the serving AP hosts a VM. The parameters associated to such VM would be: $\tau_B^{(j)}=0$, $C_{\text{VM,UL}}^{(j)}\to\infty$ and $C_{\text{VM,DL}}^{(j)}\to\infty$. In case that this VM is the only one available, then $s_P^{(j)}=s_{P_1}$ and the problem (26) could be simplified by removing constraint C5 and rewriting constraints C6 through a single constraint: $t_{\text{UL}}+\tau_{P_1}s_{P_1}+t_{\text{DL}}\le L_{\max}$. This particular case, when perfect acquisitions of the system parameters (CSI in UL, CSI in DL, BH capacity, and BH round-trip delay) is assumed, was considered and solved in [20].

Although problem (26) is not convex due to C6, the optimal solution can be found, as described in the following subsections.

B. Active Set of VMs and Optimal Load Distribution

In this section we present the optimal distribution of the computational load among the VMs for a given value of s_{P_1}. This result, stated in Proposition 1, is valid for any energy functions, $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})$ and $e_{\text{DL}}(t_{\text{DL}}, s_{\text{DL}})$, that model the energy consumption of the MT, and will be used in Section IV-C to compute the global optimal solution of problem (26).

***Proposition** 1:* Consider, without loss of generality, that the N available VMs are ordered increasingly according to the values of $\tau_B^{(i)}$, i.e. $\tau_B^{(1)}\le\ldots\le\tau_B^{(N)}$. Then, for a concrete value of s_{P_1}, the optimal load distribution among the N VMs is given by:

$$
s_P^{(i)}(s_{P_1}) = \frac{s_{P_1}+\sum_{j=1}^{M(s_{P_1})}\frac{\tau_B^{(j)}}{\overline{\tau}_P^{(j)}}}{\overline{\tau}_P^{(i)}\sum_{j=1}^{M(s_{P_1})}\frac{1}{\overline{\tau}_P^{(j)}}} - \frac{\tau_B^{(i)}}{\overline{\tau}_P^{(i)}} \quad \text{for } i=1,\ldots,M(s_{P_1}),
\tag{27}
$$

$$
s_P^{(i)}(s_{P_1}) = 0 \quad \text{for } i=M(s_{P_1})+1,\ldots,N,
\tag{28}
$$

being $M(s_{P_1})$ in (27) and (28) the number of active VMs, which is computed as the value of $\tilde{M}$ for which the following

[1]Note that problem (26) generalizes the problem in [20] to the case of multiple VMs with non-ideal BH links and imperfect acquisition of the system parameters.

conditions hold:

$$\tau_B^{(\tilde{M})} < \frac{s_{P_1} + \sum_{i=1}^{\tilde{M}} \frac{\tau_B^{(i)}}{\tau_P^{(i)}}}{\sum_{i=1}^{\tilde{M}} \frac{1}{\tau_P^{(i)}}} \quad \text{and} \quad \tau_B^{(\tilde{M}+1)} \geq \frac{s_{P_1} + \sum_{i=1}^{\tilde{M}} \frac{\tau_B^{(i)}}{\tau_P^{(i)}}}{\sum_{i=1}^{\tilde{M}} \frac{1}{\tau_P^{(i)}}}. \quad (29)$$

Proof: See Appendix B. ∎

From (29), it can be observed that $M(s_{P_1})$ is an integer and increasing function defined by intervals:

$$M(s_{P_1}) = n, \quad \text{if} \quad S_{P_1,\min}^{(n)} < s_{P_1} \leq S_{P_1,\max}^{(n)}, \quad (30)$$

where

$$S_{P_1,\min}^{(n)} = \tau_B^{(n)} \sum_{i=1}^{n} \frac{1}{\tau_P^{(i)}} - \sum_{i=1}^{n} \frac{\tau_B^{(i)}}{\tau_P^{(i)}} \quad \text{for} \quad n = 1, ..., N, \quad (31)$$

$$S_{P_1,\max}^{(n)} = \tau_B^{(n+1)} \sum_{i=1}^{n} \frac{1}{\tau_P^{(i)}} - \sum_{i=1}^{n} \frac{\tau_B^{(i)}}{\tau_P^{(i)}} \quad \text{for} \quad n = 1, ..., N-1, \quad (32)$$

and $S_{P_1,\max}^{(N)}=\infty$.

Although $M(s_{P_1})$ is a discrete function of s_{P_1}, the values of $s_P^{(i)}$ are continuous w.r.t. s_{P_1}. This can be easily observed in (27). The load of the n-th VM is zero for $s_{P_1}=S_{P_1,\min}^{(n)}$. Then, as s_{P_1} keeps on increasing, the load of the n-th VM increases continuously as a function of the values of $\tau_B^{(i)}$ and $\tau_P^{(i)}$ of all the VMs in the active set. According to (31)-(32) the order of activation of VMs depends only on the values of $\tau_B^{(i)}$.

Remark 1: For any value of s_{P_1}, we can compute in closed form the optimal number of active VMs (from (30)) and the load distribution among them (from (27)-(28)). The optimal set of active VMs is not combinatorial due to the fact that the activation order of VMs depends exclusively on the round-trip delay of the BH connection between the serving AP and each VM (i.e. $\tau_B^{(i)}$).

C. Problem Resolution

The goal of this section is to solve problem (26) and gain insight into the essential trade-offs that appear in it. They, of course, depend on the consumption model of the MT, captured in the UL and DL energy functions, i.e. $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})$ and $e_{\text{DL}}(t_{\text{DL}}, s_{\text{DL}})$. Although the results in Section IV-B are valid for any UL and DL energy functions, we will now consider again the energy functions presented in Section II-B.

Proposition 1 in Section IV-B provided the optimal distribution of $\{s_P^{(i)}(s_{P_1})\}$ for a given value of s_{P_1}. Using this result, we can eliminate variables $\{s_P^{(i)}\}$ and constraint C5 of problem (26). Furthermore, we may decompose problem (26) into N subproblems, one for each of the N intervals that define the function $M(s_{P_1})$ (see (30)) since we know the order of activation of the VMs and we have identified the regions of s_{P_1} for which the different VMs are active.

The n-th subproblem (for which $M(s_{P_1})=n$) will be equal to the original problem (26), except that constraints C5 and C6 will be replaced by the following linear constraints:

$$\text{C5}: \; S_{P_1,\min}^{(n)} \leq s_{P_1} \leq S_{P_1,\max}^{(n)}, \quad (33)$$

$$\text{C6}: \; t_{\text{UL}} + l^\star(s_{P_1}) + t_{\text{DL}} \leq L_{\max}, \quad (34)$$

with $l^\star(s_{P_1})$ given by (57) in Appendix B. Since $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})$ is jointly convex w.r.t. t_{UL} and s_{UL} (see Section III-B), and $e_{\text{DL}}(t_{\text{DL}}, s_{\text{DL}})$ is linear (and jointly convex) w.r.t. t_{DL} and s_{DL}, all subproblems are convex (the cost function in each subproblem is convex w.r.t. to the optimization variables and the constraints C1,...,C7 are linear). Therefore, if they are feasible, the solution can be found in polynomial time. Additionally, each subproblem can be simplified through the following steps:

- As constraint C1 is equivalent to $s_{P_0}=S_{\text{app}}-s_{P_1}$, C2 can be written as $s_{P_1}\geq S_{\text{app}}-\frac{L_{\max}}{\tau_{P_0}}$, and $s_{P_0}\geq 0$ in C7 can be written as $s_{P_1}\leq S_{\text{app}}$. Then, s_{P_0} can be eliminated from the set of optimization variables and C1 is not needed. The meaning of the lower bound on s_{P_1} is straightforward: to fulfill the latency constraint imposed by the application, the amount of local processing s_{P_0} cannot exceed $\frac{L_{\max}}{\tau_{P_0}}$.
- As $e_{\text{DL}}(t_{\text{DL}}, \beta_{\text{DL}} s_{P_1})$ is a non-decreasing function w.r.t. t_{DL}, for a concrete value of s_{P_1}, using the minimum value for t_{DL} allowed by constraint C4 will not increase the cost function but it will make constraints C6 looser. This will enlarge the feasible set for the rest of variables and, consequently, the cost function can be further reduced. Therefore, at the optimum, constraint C4 will be achieved with equality, i.e. $t_{\text{DL}}^\star(s_{P_1})=\frac{\beta_{\text{DL}} s_{P_1}}{R_{\text{DL,max}}}$. Note that this is equivalent to say that the optimal DL rate is $R_{\text{DL,max}}$.
- Replacing the variable t_{UL} by $\frac{\beta_{\text{UL}} s_{P_1}}{r_{\text{UL}}}$, constraint C3 is equivalent to $r_{\text{UL}}\leq R_{\text{UL,max}}$. Additionally, (34) can be re-written as a lower bound: $r_{\text{UL}}\geq r_{\text{UL,min}}(s_{P_1})$, where

$$r_{\text{UL,min}}(s_{P_1}) = \frac{\beta_{\text{UL}} s_{P_1}}{L_{\max} - l^\star(s_{P_1}) - \frac{\beta_{\text{DL}} s_{P_1}}{R_{\text{DL,max}}}}. \quad (35)$$

In order to be able to find a feasible value of r_{UL}, it is required that $r_{\text{UL,min}}(s_{P_1})\leq R_{\text{UL,max}}$. This is equivalent to

$$s_{P_1} \leq \frac{L_{\max} - \sum_{i=1}^{n} \frac{\tau_B^{(i)}}{\tau_P^{(i)}}}{\frac{\beta_{\text{UL}}}{R_{\text{UL,max}}} + \sum_{i=1}^{n} \frac{1}{\tau_P^{(i)}} + \frac{\beta_{\text{DL}}}{R_{\text{DL,max}}}}. \quad (36)$$

Then, re-writing the cost function in terms of the normalized energies per bit $e_{\text{UL}}(t_{\text{UL}}, \beta_{\text{UL}} s_{P_1})=\beta_{\text{UL}} s_{P_1} \overline{e}_{\text{UL}}(r_{\text{UL}})$ and $e_{\text{DL}}(t_{\text{DL}}, \beta_{\text{UL}} s_{P_1})=\beta_{\text{DL}} s_{P_1} \overline{e}_{\text{DL}}(r_{\text{DL}})$, the n-th subproblem (n=1, ..., N) is equivalent to the following problem:

$$\underset{s_{P_1},\, r_{\text{UL}}}{\text{minimize}} \; \beta_{\text{UL}} s_{P_1} \overline{e}_{\text{UL}}(r_{\text{UL}}) - \varepsilon_{P_0} s_{P_1} + \beta_{\text{DL}} s_{P_1} \overline{e}_{\text{DL}}(R_{\text{DL,max}})$$

$$\text{subject to C1}: r_{\text{UL,min}}(s_{P_1}) \leq r_{\text{UL}} \leq R_{\text{UL,max}},$$

$$\text{C2}: \tilde{S}_{P_1,\min}^{(n)} \leq s_{P_1} \leq \tilde{S}_{P_1,\max}^{(n)}, \quad (37)$$

where

$$\tilde{S}_{P_1,\min}^{(n)} = \max\left(0, S_{\text{app}} - \frac{L_{\max}}{\tau_{P_0}}, S_{P_1,\min}^{(n)}\right) \quad (38)$$

and

$$\tilde{S}_{P_1,\max}^{(n)} = \min\left(\frac{L_{\max} - \sum_{i=1}^{n} \frac{\tau_B^{(i)}}{\tau_P^{(i)}}}{\frac{\beta_{\text{UL}}}{R_{\text{UL,max}}} + \sum_{i=1}^{n} \frac{1}{\tau_P^{(i)}} + \frac{\beta_{\text{DL}}}{R_{\text{DL,max}}}}, S_{P_1,\max}^{(n)}, S_{\text{app}}\right). \quad (39)$$

Problem (37) is equal to problem (25) in [20], being the only

TABLE I: COMPUTATION OF THE OFFLOADING STRATEGY

1:	set $e = \infty$
2:	**for** $n = 1, \ldots, N$
3:	compute $\tilde{S}^{(n)}_{P_1,\min}$ and $\tilde{S}^{(n)}_{P_1,\max}$ according to (38)-(39)
4:	**if** $\tilde{S}^{(n)}_{P_1,\max} < \tilde{S}^{(n)}_{P_1,\min}$
5:	subproblem (37) is infeasible: **go to** 9
6:	**otherwise**
7:	compute s_{P_1} in the interval $\tilde{S}^{(n)}_{P_1,\min} \le s_{P_1} \le \tilde{S}^{(n)}_{P_1,\max}$ to minimize $f_o(s_{P_1})$ in (42)
8:	**if** $f_o(s_{P_1}) < e$, **then** $e = f_o(s_{P_1})$ and $s^\star_{P_1} = s_{P_1}$
9:	**end if**
10:	**end for**
11:	**if** $e = \infty$
12:	problem (26) is infeasible
13:	**otherwise**
14:	based on $s^\star_{P_1}$, compute: $s^\star_{P_0}, \{s^{(i)\star}_P\}, r^\star_{\text{UL}}, r^\star_{\text{DL}}, t^\star_{\text{UL}}, t^\star_{\text{DL}}$
15:	**end if**

difference the value of the limits of s_{P_1}, now given by (38) and (39). Then, the optimal value of r_{UL}, i.e. $r^\star_{\text{UL}}(s_{P_1})$, can be obtained[2] as

$$r^\star_{\text{UL}}(s_{P_1}) = \arg\min_{r_{\text{UL}}} \quad \bar{e}_{\text{UL}}(r_{\text{UL}}) \qquad (40)$$
$$\text{subject to} \quad r_{\text{UL,min}}(s_{P_1}) \le r_{\text{UL}} \le R_{\text{UL,max}}.$$

As $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})$ is jointly convex w.r.t. t_{UL} and s_{UL}, $\bar{e}_{\text{UL}}(r_{\text{UL}})$ is a quasi-convex function of r_{UL}, and so $r^\star_{\text{UL}}(s_{P_1})$ can be computed as follows:

$$r^\star_{\text{UL}}(s_{P_1}) = \begin{cases} r_{\text{UL,min}}(s_{P_1}), & \text{if } \check{R}_{\text{UL}} < r_{\text{UL,min}}(s_{P_1}), \\ \check{R}_{\text{UL}}, & \text{if } r_{\text{UL,min}}(s_{P_1}) \le \check{R}_{\text{UL}} \le R_{\text{UL,max}}, \\ R_{\text{UL,max}}, & \text{if } \check{R}_{\text{UL}} > R_{\text{UL,max}}, \end{cases} \qquad (41)$$

where $\check{R}_{\text{UL}}$ is the value of r_{UL} for which $\bar{e}_{\text{UL}}(r_{\text{UL}})$ is lowest.

Finally, each subproblem reduces to a one-dimensional search to find the value of s_{P_1}, with $\tilde{S}^{(n)}_{P_1,\min} \le s_{P_1} \le \tilde{S}^{(n)}_{P_1,\max}$, that minimizes the function $f_o(s_{P_1})$:

$$f_o(s_{P_1}) = \beta_{\text{UL}} s_{P_1} \bar{e}_{\text{UL}}(r^\star_{\text{UL}}(s_{P_1})) - \varepsilon_{P_0} s_{P_1} + \beta_{\text{DL}} s_{P_1} \bar{e}_{\text{DL}}(R_{\text{DL,max}}), \qquad (42)$$

which is a convex function w.r.t. s_{P_1} because $e_{\text{UL}}(t_{\text{UL}}, s_{\text{UL}})$ is jointly convex w.r.t. t_{UL} and s_{UL} and $e_{\text{DL}}(t_{\text{DL}}, s_{\text{DL}})$ is jointly convex w.r.t. t_{DL} and s_{DL}. Note that the actual energy consumption of the MT is $\varepsilon_{P_0} S_{\text{app}} + f_o(s_{P_1})$ and, therefore, $f_o(s_{P_1})$ is the difference between the energy consumption with and without offloading.

Since all the subproblems are convex, simple methods can be applied to calculate the optimal solution to each subproblem, such as for instance the bisection method [39]. Note that the bisection method converges with exponential speed to the value of s_{P_1} minimizing $f_o(s_{P_1})$ in (42), with a resolution better than a given percentage of the length of the interval $\tilde{S}^{(n)}_{P_1,\min} \le s_{P_1} \le \tilde{S}^{(n)}_{P_1,\max}$ (i.e. only 7 iterations are required for a resolution of 1% of the interval length).

Summarizing, the solution reduces to search the best value of s_{P_1} for each of the N possible intervals (n=1, ...,N). The s_{P_1} providing a lower value for $f_o(s_{P_1})$ will be the optimal one to problem (26) (which is denoted by $s^\star_{P_1}$). Once the optimal value of $s^\star_{P_1}$ is obtained, the optimal values for the remaining variables can be directly attained: $s^\star_{P_0} = S_{\text{app}} - s^\star_{P_1}$, $M(s^\star_{P_1})$ as in (30), $s^{(i)\star}_P$ as in (27)-(28), $t^\star_{\text{UL}} = \frac{\beta_{\text{DL}} s^\star_{P_1}}{r^\star_{\text{UL}}}$ with $r^\star_{\text{UL}}$ in (41), and $t^\star_{\text{DL}} = \frac{\beta_{\text{DL}} s^\star_{P_1}}{R_{\text{DL,max}}}$. The procedure is included in Table I.

[2]For a feasible s_{P_1}, condition $r_{\text{UL,min}}(s_{P_1}) \le R_{\text{UL,max}}$ will be fulfilled.

D. Analysis of Complexity

As shown in (30)-(32), the order of activation of the VMs depends only on the values of $\tau^{(i)}_B$. Therefore, the complexity of the proposed solution grows linearly with the number of VMs, because the order of activation of the VMs is known. Note that the exhaustive search solution, which searches among all possible combinations of active VMs, has an exponential complexity with the number of VMs. Notably, the robust designs for application offloading have exactly the same complexity as for the case of perfect CSI, since the robust strategies do only impact on the acquisition of the system parameters.

V. NUMERICAL RESULTS

This section presents some numerical results to illustrate the concepts presented in previous sections when several VMs are available. Each VM and BH link has its own features, captured through the parameters $\overline{\tau}^{(i)}_P$ and $\tau^{(i)}_B$, regarding computational capability and non-ideal BH connection with the serving AP. As an example, we consider up to 5 available VMs with $\overline{\tau}^{\{1,2,3,4,5\}}_P = \{5, 2.5, 2.5, 0.5, 0.5\} \times 10^{-8}$ s/bit and $\tau^{\{1,2,3,4,5\}}_B = \{0, 0.1, 0.2, 0.3, 0.4\}$ s. The rest of the parameters are: ε_{P_0}=8.6 × 10^{-8} J/bit, $\tau_{P_0} = 2\overline{\tau}^{(1)}_P$, $k_{\text{tx},1}$=0.4 W, $k_{\text{tx},2}$=18, $k_{\text{rx},1}$=0.4 W, $k_{\text{rx},2}$=2.86 × 10^{-3} W/Mbps, β_{UL}=1, β_{DL}=0.2, W_{UL}=10 MHz, W_{DL}=10 MHz, $P_{\text{tx,MT}}$=100 mW, $P_{\text{tx,AP}}$=100 mW. The application considered for offloading among multiple VMs is the compression of a set of files with a total size of S_{app}=5 Mbytes.

The parameters related with the computation speed and the computation energy consumption of the MT, τ_{P_0} and ε_{P_0}, are taken from [20], which were derived from the experimental measurements provided in [10, Table 10] for a mobile device running a Gzip compression application. The parameters related to the communication energy consumption of the MT, namely $k_{\text{tx},1}$, $k_{\text{tx},2}$, $k_{\text{rx},1}$, $k_{\text{rx},2}$, are also taken from [20], which were computed through numerical regressions to be aligned with the experimental measurements provided in [26] for an LTE-MT dongle. As for the VMs parameters, we have considered that the VMs can compute the user application between 2 and 20 times faster (note that it is not only a matter of the CPU speed, but also of the computation resources available for this particular MT and the BH capacity when uploading and downloading, i.e., $\{C^{(i)}_{\text{VM,UL}}\}$ and $\{C^{(i)}_{\text{VM,DL}}\}$).

A. Offloading with Perfect Acquisition

In this section we evaluate the performance and offloading decision under perfect acquisition of CSI and BH parameters.

Fig. 2 (top) illustrates the distribution of the total offloaded load among the different VMs, for different sizes of such

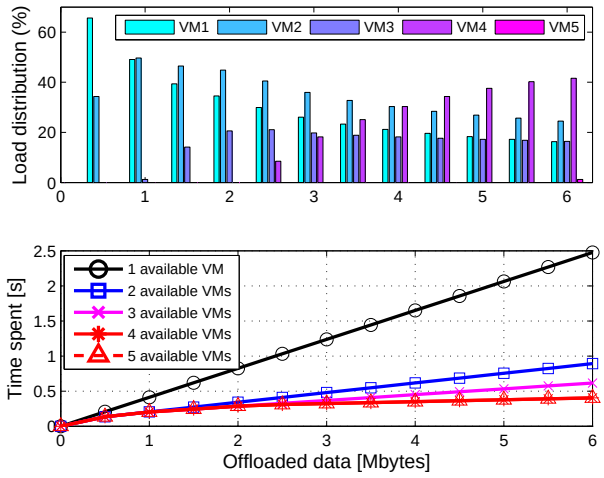

Fig. 2: Load distribution among VMs (top) and time spent for the remote processing and data exchange through the BH (bottom) vs. the size of the offloaded load (s_{P_1}).

load (i.e. s_{P_1}). As proved in previous sections, the order of activation of the VMs depends only on the values of $\tau_B^{(i)}$. Only for high computational loads, all VMs become active. Fig. 2 (bottom) shows the time required for the remote processing, including the time for data exchange among the serving AP and the VMs through the BH, versus the size of the remote computational load s_{P_1}. We present such a result for different numbers of available VMs (although an available VM may not be necessarily active). As expected, having more VMs available allows reducing the time required for the remote processing, particularly for high computational loads, as in this case more VMs become active.

In the following figures, we assess the impact of having several VMs available on the offloading process, for two different values of the channel gain normalized to the noise power, i.e. ρ=σ_h^2/σ_n^2 with σ_h^2=$\sigma_{h,\mathrm{UL}}^2$=$\sigma_{h,\mathrm{DL}}^2$, and n_{AP}=n_{MT}=1. ρ=15 dB and ρ=25 dB are used in Fig. 3 and Fig. 4, respectively. The figures show the percentage of energy saving of the MT w.r.t. doing all the computation locally (top), the number of active VMs (middle), and the total time spent (bottom), for different numbers of available VMs. The horizontal axis corresponds to the maximum latency allowed for the application to be completed, $L_{\max}$.

For ρ=15 dB (see Fig. 3), we observe that as the total latency constraint $L_{\max}$ increases, all the files are processed locally. The reason is that for the parameters considered and from an energy consumption point of view, doing the processing locally is better, since for these channel conditions the energy required for the communication is greater than the saving coming from doing the processing remotely. As a result, only when the time required for doing all the processing locally is greater than $L_{\max}$, part of the processing is shifted to the VMs. This is done at the expenses of increasing the energy consumption at the MT (this is why the energy saving is negative in Fig. 3-top). Restricting the latency constraint increases the amount of files to be compressed remotely and, therefore, the number of active VMs. Note that having more VMs available allows decreasing the value of $L_{\max}$ for which

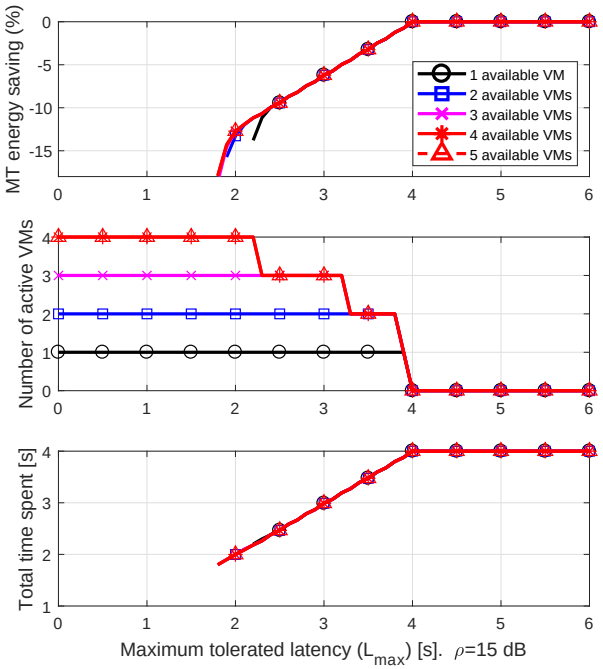

Fig. 3: Energy saving at the MT when offloading (top), number of active VMs (middle), and total time spent (bottom) vs. the total latency constraint ($L_{\max}$) for ρ=15 dB.

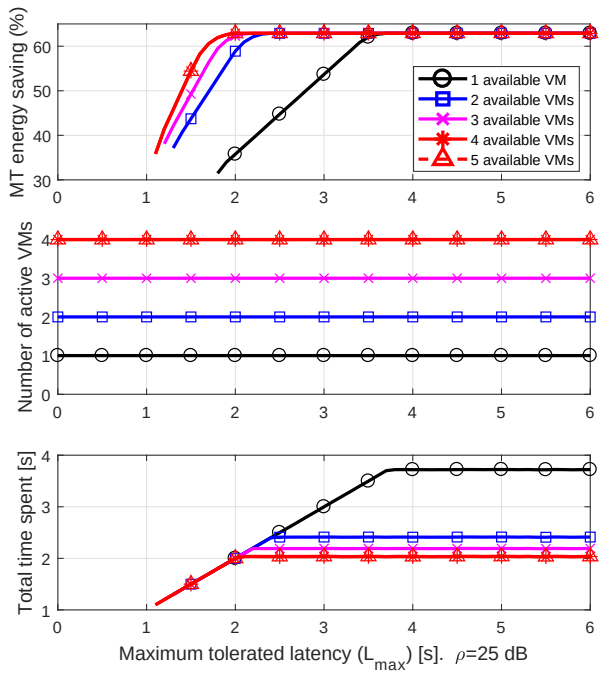

Fig. 4: Energy saving at the MT when offloading (top), number of active VMs (middle), and total time spent (bottom) vs. the total latency constraint ($L_{\max}$) for ρ=25 dB.

problem (26) is feasible. For the given total files size, the 5-th VM is never activated.

When the channel conditions improve (ρ=25 dB, in Fig. 4), the energy consumption in the communication is reduced, making the offloading worthy. In this case, if the total latency constraint allows it, all the files will be compressed remotely.

If $L_{\max}$ is not high enough, some of the files will have to be processed locally to fulfill the total latency constraint. In this case, having more VMs allows processing more data with lower processing time. For such a reason, having more VMs available reduces the energy consumption at the MT and also the total time spent, as it can be observed in Fig. 4. Again, for the given total files size, the 5-th VM is never activated.

Finally, let us note that Fig. 3 and Fig. 4 illustrate the energy-latency trade-off in application offloading: the lower the maximum tolerated latency, the lower is the reduction in the energy consumption (i.e., the higher the energy is). As it is shown in the figures, the trade-off is improved (i.e., for a given tolerated latency, the energy consumption is lower) when the SNR increases and/or when more VMs are available.

B. Offloading with Imperfect Acquisition

In this section we evaluate the impact of imperfect acquisition of BH parameters and CSI on the performance and the offloading decision, assuming that all the 5 VMs are available.

To analyze the robust design against imperfect BH parameters acquisition, we consider different uncertainties in the acquisition of the BH round-trip delay of the 2-nd VM (i.e. $\epsilon_\tau^{(2)}$ in (11)), for ρ=25 dB, $L_{\max}$=4 s, and n_{AP}=n_{MT}=1. Fig. 5 (top) displays the offloaded load distribution among the different VMs versus $\epsilon_\tau^{(2)}$, and Fig. 5 (bottom) shows the total time spent. When the uncertainty in the acquisition of BH parameters related to the 2-nd VM increases (i.e. as $\epsilon_\tau^{(2)}$ increases), then less bytes are sent to be processed at that VM until it becomes inactive. However, as the wireless channel conditions are good, those bytes are distributed among the remaining available VMs, hence resulting in the activation of the 5-th VM and increasing as well the load of the 4-th VM. The total time spent increases because the processing is done in VMs that are farther and/or with lower capabilities, but the increase of the total time spent is shown to be very slight.

To assess the robust design against imperfect CSI acquisition, we consider separately the impact of imperfect CSI due to channel estimation errors in the training phase and due to feedback delay errors (as detailed in Section II-C1). Results are averaged among 100 channel realizations and different MIMO configurations are considered: n_{AP}=n_{MT}=1, n_{AP}=n_{MT}=2, and n_{AP}=n_{MT}=3. p_{in} = 0.9 is used in (7).

First, Fig. 6 depicts the energy saving of the MT w.r.t. doing all the computation locally (top), and the percentage of offloaded data (bottom), versus the transmit SNR used to estimate DL and UL channels ($\bar{\gamma}$=$\frac{P_{\text{train}}T}{\sigma_p^2}$, see (46)) for ρ=25 dB, $L_{\max}$=4 s, and $f_{\text{d}}t_{\text{del}}$=0 (i.e. no feedback delay errors). To fairly compare different MIMO configurations, we use the transmit SNR $\bar{\gamma}$ in the horizontal axis, so that γ_{UL}=$\bar{\gamma}/n_{\text{MT}}$, γ_{DL}=$\bar{\gamma}/n_{\text{AP}}$ in (4)-(5) (see (46)). As it is expected, when the transmit SNR is reduced, the amount of offloaded data is decreased until all data is processed locally at the MT. This is because, when channel estimates are not trusty, the energy consumption in DL/UL is greater and the DL/UL rates are diminished as compared to the perfect CSI case due to the robust designs. It is interesting to note that large MIMO configurations involve large estimation errors because

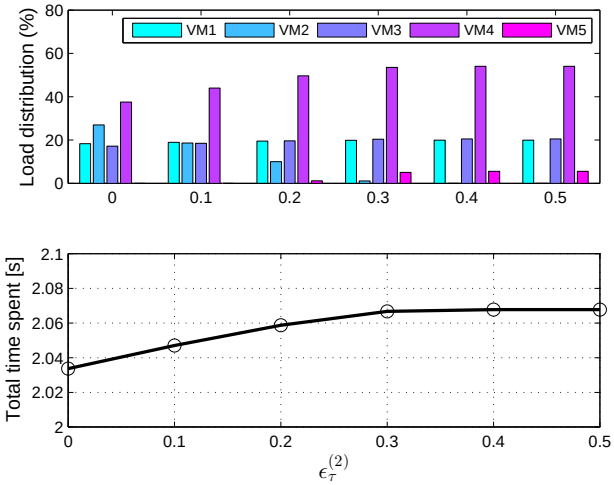

Fig. 5: Load distribution among VMs (top) and total time spent (bottom) vs. the imperfection in the BH delay of the 2-nd VM ($\epsilon_\tau^{(2)}$) for ρ=25 dB and $L_{\max}$=4 s.

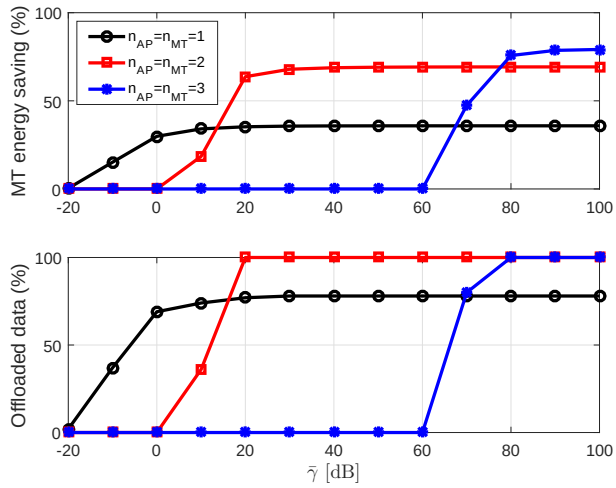

Fig. 6: Energy saving at the MT when offloading (top) and percentage of offloaded data (bottom) vs. the transmit SNR for channel estimation ($\bar{\gamma}$) for ρ=25 dB, $L_{\max}$=4 s, and $f_{\text{d}}t_{\text{del}}$=0.

the total power is distributed among the different antennas and hence less power is available per antenna element for channel estimation. For that reason, large $\bar{\gamma}$ is needed to start offloading with the robust design when larger MIMO setups are considered; however, once offloading is activated and the transmit SNR for channel estimation is good, larger MT energy saving is obtained when the number of antennas increases.

Second, Fig. 7 shows the energy saving of the MT w.r.t. doing all the computation locally (top), and the percentage of offloaded data (bottom), versus the time delay for channel estimation normalized to the inverse of the Doppler frequency (i.e. $f_{\text{d}}t_{\text{del}}$ in (4)-(5)) for ρ=25 dB, $L_{\max}$=4 s, and $\bar{\gamma}\to\infty$ (i.e. no channel estimation errors in the training phase). In this case, when the value of the time delay for CSI acquisition increases then the energy savings are lower as compared to the perfect CSI case. Large MIMO configurations are more sensitive to feedback delay errors because, although the variances of the errors (σ_{UL}^2 and σ_{DL}^2) are the same (see (51)), more coefficients

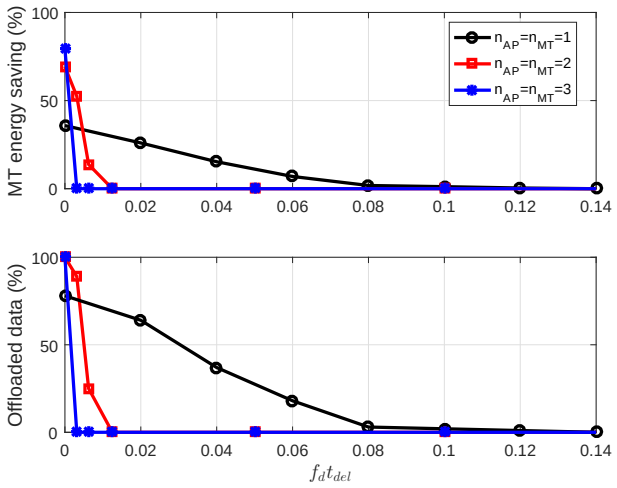

Fig. 7: Energy saving at the MT when offloading (top) and percentage of offloaded data (bottom) vs. the normalized estimation time delay ($f_d t_{del}$) for ρ=25 dB, L_{max}=4 s, and $\bar{\gamma}\to\infty$.

have to be estimated and thus the uncertainty regions in (6) are enlarged (or, equivalently, ϵ_{UL} and ϵ_{DL} in (7) increase).

VI. Conclusion

In this paper, we consider a system in which a given amount of data located at an MT has to be processed by an application. In this system, there are several VMs that could potentially process parts of the data if it is decided to offload such data to the VMs through an AP to which the MT is wirelessly connected. The objective is to minimize the total energy that the MT will spend (accounting for the wireless transmission and the processing of the data that is kept at the MT) while fulfilling a maximum latency constraint in the processing of all the data. In other words, the trade-off between the energy spent by the MT and the experienced delay is exploited. Such minimization of the energy is performed by optimizing the parameters related to the wireless connection between the MT and the AP, and by deciding how much data is kept for processing at the MT and how the rest of the data is distributed among the VMs. Accordingly, this paper extends the results in [20], where only one VM available at the AP was considered. In our case, we assume that the additional close and remote VMs are connected to the serving AP through non-ideal BH links and that the acquisition of the system parameters (such as the CSI of the wireless connection and the characteristics of the BH links) is imperfect.

From the optimization of the system, we have concluded that when several VMs are available, there is an optimal order in the activation of the VMs to which the data is sent to be processed. Such activation order is proved to depend exclusively on the delay of the BH connection between each VM and the serving AP. Furthermore, we have obtained the optimal distribution of the computational load among the set of active VMs. According to the proposed optimization solution, we have evaluated numerically the impact of increasing the number of available VMs on both the energy consumption of the MT and the total time required to complete the data processing. We have observed that, as expected, having more available VMs improves the energy-latency trade-off, although the actual number of active VMs depends on the offloaded load.

To deal with imperfect acquisition of the system parameters (i.e., radio channels for UL and DL transmissions, and BH characterization in terms of capacity and round-trip delay), worst-case robust designs have been proposed. Also, the error uncertainty regions have been related to the physical parameters for channel estimation (i.e., SNR, Doppler frequency, and time delay). It has been observed that offloading decisions become more conservative as the uncertainty in CSI and BH parameters acquisition increases. In the imperfect CSI case, offloading is more affected as the number of antennas increases because more channels have to be estimated. About imperfect acquisition of the BH parameters, it is shown that having imperfectly acquired BH links induces a redistribution of the load among the VMs whose links have been acquired reliably.

Although this paper has dealt only with the case of a single-user scenario, the proposed strategy could be taken as a basis for a multi-user setup. In that case, each user would have its latency constraints, and the resources (in the radio link, in the BH links, and in the VMs) should be shared. This would produce a coupled problem, consisting of both resource allocation and application offloading, that could be significantly more complex depending on the strategy adopted for the resource allocation. In this regard, some initial works with simplified assumptions have been done by the same authors in [40] for a single VM. However, the complete generalization is still to be done and is left for future work.

VII. Acknowledgments

This work has been partially funded by the Spanish Ministerio de Economía y Competitividad and FEDER funds through project TEC2016-77148-C2-1-R (AEI/FEDER, UE): 5G&B RUNNER-UPC, and the Catalan Government AGAUR through grant 2014 SGR 60.

Appendix A
Proof of Lemma 1

Let us focus on the CSI acquisition in the UL transmission. The same result applies to the DL transmission with the appropriate variables. For the sake of simplicity in the notation, let us drop subindexes related to UL and DL. Let $\mathbf{H}\in\mathbb{C}^{n_{AP}\times n_{MT}}$ be the actual channel, for which it is assumed that its components are *i.i.d.* complex circularly symmetric Gaussian distributed[3] with zero mean and variance σ_h^2.

Our objective is to derive the statistics of the actual channel $\mathbf{H}$ conditioned on the observations in the training phase. We will see that the conditional probability density function corresponds to a Gaussian distribution, whose mean (which is, in fact, the MMSE Bayesian channel estimate) will be represented by $\hat{\mathbf{H}}$ and whose variance will be denoted by σ^2. We will also see that this variance is related to the imperfections

[3]For simplicity, we assume that the components of $\mathbf{H}$ are *i.i.d.*, although the analysis could be extended to the correlated case with the proper notation and manipulation.

that generate the errors (namely, channel estimation errors and feedback delay errors) and is given by (4)-(5).

Consider that channel estimation is performed during a training phase, for which orthogonal training sequences $\mathbf{T}\in\mathbb{C}^{n_{\rm MT}\times T}$ are employed to estimate the channel, being T the number of channel uses and $\mathbf{T}\mathbf{T}^H=\alpha\mathbf{I}$. Note that $\alpha=\frac{P_{\rm train}T}{n_{\rm MT}}$, where $P_{\rm train}$ denotes the total transmission power during the training phase. Hence, the received signal $\mathbf{Y}\in\mathbb{C}^{n_{\rm AP}\times T}$ during the training phase can be expressed as [28]

$$\mathbf{Y} = \mathbf{H}_d\mathbf{T} + \mathbf{N}, \tag{43}$$

where $\mathbf{H}_d\in\mathbb{C}^{n_{\rm AP}\times n_{\rm MT}}$ is a delayed version of the actual channel $\mathbf{H}$ (which follows the same distribution as $\mathbf{H}$) and $\mathbf{N}\in\mathbb{C}^{n_{\rm AP}\times T}$ is the noise matrix, which is composed of *i.i.d.* complex circularly symmetric Gaussian components with zero mean, variance σ_n^2, and is assumed to be independent of $\mathbf{H}$. The components of $\mathbf{H}$ and $\mathbf{H}_d$ are correlated through the channel variability model and its correlation depends on the Doppler frequency $f_{\rm d}$ and the time delay $t_{\rm del}$ in the channel estimation. In particular, for Jake's model, we have $\mathbb{E}\{[\mathbf{H}]_{i,j}[\mathbf{H}_d]^*_{m,n}\}=\sigma_h^2 J_0(2\pi f_{\rm d}t_{\rm del})\delta_{i,m}\delta_{j,n}$ (i.e. each component of the channel matrix changes throughout time independently from the other components) [29].

Under this setting, we focus on characterizing the distribution of the actual channel $\mathbf{H}$ conditioned on the received signal $\mathbf{Y}$ during the training phase, i.e. $f(\mathbf{H}|\mathbf{Y})$ [30]. It is well known that $f(\mathbf{H}|\mathbf{Y})$ follows a Gaussian distribution when $\mathbf{H}$ and $\mathbf{N}$ are jointly Gaussian distributed [41], so we need to find its mean and variance. Furthermore, it can be shown that if we define:

$$\bar{\mathbf{H}} = \frac{1}{\alpha}\mathbf{Y}\mathbf{T}^H, \tag{44}$$

then $\bar{\mathbf{H}}$ is a sufficient statistic to estimate the actual channel $\mathbf{H}$ or, equivalently, $f(\mathbf{H}|\mathbf{Y})=f(\mathbf{H}|\bar{\mathbf{H}})$. By including expression (43) into (44), we can express the sufficient statistics $\bar{\mathbf{H}}$ as

$$\bar{\mathbf{H}} = \mathbf{H}_d + \mathbf{E}, \tag{45}$$

where the components of $\mathbf{E}$ are *i.i.d.* complex circularly symmetric Gaussian with zero mean, variance $\sigma_e^2=\frac{\sigma_n^2}{\alpha}$, and independent of the actual channel $\mathbf{H}$. Let us define γ as the transmit SNR for channel estimation, i.e.

$$\gamma = \frac{P_{\rm train}T}{\sigma_n^2 n_{\rm MT}} = \frac{\alpha}{\sigma_n^2}, \tag{46}$$

so that $\sigma_e^2=1/\gamma$.

Therefore, from now on, we focus on computing the distribution $f(\mathbf{H}|\bar{\mathbf{H}})$. The mean of $\mathbf{H}|\bar{\mathbf{H}}$ determines the center of the uncertainty region, which coincides with the MMSE Bayesian channel estimate: $\hat{\mathbf{H}}=\mathbb{E}\{\mathbf{H}|\bar{\mathbf{H}}\}$ [28]. The variance of $\mathbf{H}|\bar{\mathbf{H}}$ determines the size (more precisely, the radius) of the uncertainty region. Note, however, that the MMSE Bayesian channel estimate $\hat{\mathbf{H}}$ might differ from the sufficient statistics $\bar{\mathbf{H}}$.

To compute the mean and the variance of $\mathbf{H}|\bar{\mathbf{H}}$, we use the following statistical result in Lemma 2 that is derived from [42, Prop. 3.13].

***Lemma** 2:* Given two random variables a and b that are jointly Gaussian distributed with means $\mathbb{E}\{a\}=\mu_a$ and $\mathbb{E}\{b\}=\mu_b$, variances σ_a^2 and σ_b^2, respectively, $C_{ab}=\mathbb{E}\{(a-\mu_a)(b-\mu_b)^*\}$, and $C_{ba}=\mathbb{E}\{(b-\mu_b)(a-\mu_a)^*\}=C_{ba}^*$, then:

$$\mathbb{E}_{b|a}\{b|a\} = \mu_b + \frac{C_{ba}}{\sigma_a^2}(a-\mu_a), \tag{47}$$

$$C_{b|a} = \mathbb{E}\{|b-\mathbb{E}_{b|a}\{b|a\}|^2|a\} = \sigma_b^2 - \frac{C_{ba}C_{ab}}{\sigma_a^2}. \tag{48}$$

Under the previous assumptions concerning the channel statistics, channel temporal-variation model, and channel estimation procedure, we can apply the previous Lemma 1 component by component for the matrices involved (since these components are independent) by taking $a_{i,j}=[\bar{\mathbf{H}}]_{i,j}=[\mathbf{H}_d]_{i,j}+[\mathbf{E}]_{i,j}$ and $b_{i,j}=[\mathbf{H}]_{i,j}$ with $\mu_a=0$, $\mu_b=0$, $\sigma_a^2=\sigma_h^2+\sigma_e^2=\sigma_h^2+1/\gamma$, $\sigma_b^2=\sigma_h^2$, $\forall i,j$. For Jake's time-variation model, $C_{ba}=\mathbb{E}\{[\mathbf{H}]_{i,j}[\bar{\mathbf{H}}]^*_{i,j}\}=\mathbb{E}\{[\mathbf{H}]_{i,j}[\mathbf{H}_d]^*_{i,j}\}=\sigma_h^2 J_0(2\pi f_{\rm d}t_{\rm del})$ and $C_{ab}=C_{ba}$ [29]. Accordingly, by using Lemma 2, we obtain the MMSE Bayesian channel estimate as the conditional mean of the actual channel $\mathbf{H}$ given the sufficient statistics $\bar{\mathbf{H}}$:

$$\hat{\mathbf{H}} = \mathbb{E}\{\mathbf{H}|\bar{\mathbf{H}}\} = \frac{\sigma_h^2 J_0(2\pi f_{\rm d}t_{\rm del})}{\sigma_h^2+1/\gamma}\bar{\mathbf{H}} = \frac{J_0(2\pi f_{\rm d}t_{\rm del})}{1+\frac{1}{\sigma_h^2\gamma}}\bar{\mathbf{H}}. \tag{49}$$

Similarly, through Lemma 2, we get the variance of the actual channel $\mathbf{H}$ given the sufficient statistics $\bar{\mathbf{H}}$ as:

$$\begin{aligned}\sigma^2 &= \sigma_h^2 - \frac{(\sigma_h^2 J_0(2\pi f_{\rm d}t_{\rm del}))^2}{\sigma_h^2+1/\gamma} \\ &= \sigma_h^2\left(\frac{1+\sigma_h^2\gamma\left(1-J_0^2(2\pi f_{\rm d}t_{\rm del})\right)}{1+\sigma_h^2\gamma}\right),\end{aligned} \tag{50}$$

and, thus, the characterization is completed.

For the sake of completeness, let us analyze the result in (49)-(50) under extreme cases. If there is no channel estimation error (i.e. $\gamma\to\infty$), then:

$$\hat{\mathbf{H}} = J_0(2\pi f_{\rm d}t_{\rm del})\bar{\mathbf{H}}, \quad \sigma^2 = \sigma_h^2(1-J_0^2(2\pi f_{\rm d}t_{\rm del})). \tag{51}$$

On the other hand, if there is no feedback delay error (i.e. $J_0(0)=1$), then:

$$\hat{\mathbf{H}} = \frac{1}{1+\frac{1}{\sigma_h^2\gamma}}\bar{\mathbf{H}}, \quad \sigma^2 = \frac{\sigma_h^2}{1+\sigma_h^2\gamma}. \tag{52}$$

In addition, for $\gamma\to\infty$ (i.e. no error in the training phase), $\hat{\mathbf{H}}=\bar{\mathbf{H}}$ and $\sigma^2=0$, so that full reliability is given to the observation. However, for $\gamma\to 0$, $\hat{\mathbf{H}}=\mathbf{0}$ and $\sigma^2=\sigma_h^2$, so that only prior information is taken into account and the observation is discarded. This verifies the coherence of the obtained result.

Based on all the previous results, we conclude that the actual channel can be written as

$$\mathbf{H} = \hat{\mathbf{H}} + \mathbf{\Delta}, \tag{53}$$

where $\mathbf{\Delta}$ is a matrix of *i.i.d.* complex circularly symmetric Gaussian entries with zero mean and variance σ^2 given by (50). This model is used to characterize the uncertainty regions in Section II-C1 in the present paper.

APPENDIX B
PROOF OF PROPOSITION 1

Given s_{P_1}, the optimal distribution of $\{s_P^{(i)}\}$ to problem (26) is the one that minimizes the overall latency $L=\max(s_{P_0}\tau_{P_0}, t_{\mathrm{UL}}+\max_i(\overline{\tau}_P^{(i)}s_P^{(i)}+\tau_B^{(i)})+t_{\mathrm{DL}})$ or, equivalently, that minimizes the maximum value of $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}$. This would allow enlarging the sets of feasible values for t_{UL} and t_{DL} (making constraints C3, C4 and C6 looser) and, therefore, the cost function in (26) could be further reduced. This distribution of $\{s_P^{(i)}\}$ can be obtained as the solution to the following optimization problem:

$$\begin{aligned} &\underset{\{s_P^{(i)}>0\}}{\text{minimize}}\ l \\ &\text{subject to } s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)} \le l \quad \text{for } i=1,...,M, \\ &\qquad \textstyle\sum_{i=1}^{M} s_P^{(i)} = s_{P_1}. \end{aligned} \tag{54}$$

In (54), l is minimized when all the terms $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}$ are equal[4], i.e.:

$$s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)} = s_P^{(1)}\overline{\tau}_P^{(1)}+\tau_B^{(1)} \text{ for } i=2,...,M. \tag{55}$$

Combining (55) with the constraint $\sum_{i=1}^{M} s_P^{(i)}=s_{P_1}$, we may compute $\tau_B^{(1)}+s_P^{(1)}\overline{\tau}_P^{(1)}$ and then:

$$s_P^{(i)} = \frac{s_{P_1}+\sum_{j=1}^{M}\frac{\tau_B^{(j)}}{\overline{\tau}_P^{(j)}}}{\overline{\tau}_P^{(i)}\sum_{j=1}^{M}\frac{1}{\overline{\tau}_P^{(j)}}} - \frac{\tau_B^{(i)}}{\overline{\tau}_P^{(i)}} \quad \text{for} \quad i=1,...,M. \tag{56}$$

If M is computed according to (29), the values of of $s_P^{(i)}$ in (56) are positive. Therefore, problem (54) is feasible and its solution is:

$$l^\star(s_{P_1}) = s_P^{(1)}\overline{\tau}_P^{(1)}+\tau_B^{(1)} = \frac{s_{P_1}+\sum_{i=1}^{M}\frac{\tau_B^{(i)}}{\overline{\tau}_P^{(i)}}}{\sum_{i=1}^{M}\frac{1}{\overline{\tau}_P^{(i)}}}. \tag{57}$$

We still need to prove that the optimal set of active VMs is $\{1,...,M\}$, if M is computed according to (29). To that end, we need to prove that no other set can do better. Assume we add the K-th VM, with $K>M$. From (29) and from (57), it follows that $\tau_B^{(K)}\ge l^\star(s_{P_1})$. Therefore, if this VM entered in the active set, as $s_P^{(K)}\overline{\tau}_P^{(K)}+\tau_B^{(K)}>l^\star(s_{P_1})$, the latency of the whole process would be increased. As a conclusion, we cannot do better by adding another VM to the set $\{1,...,M\}$. Additionally, (29) implies that $s_P^{(i)}>0$ for $i=1,...,M$ and we have shown that $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}=l^\star(s_{P_1})$ for $i=1,...,M$. Assume we remove the K-th VM, with $K\le M$. In this case, we should distribute the load of this VM, i.e. $\{s_P^{(K)}\}$, among the remaining VMs. This would increase $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}$ for some (or all) of the remaining VMs, and the latency would be increased. As a conclusion, we cannot do better by removing a VM from the set $\{1,...,M\}$.

[4]This can be justified as follows. Let us assume that not all the terms $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}$ are equal. We define the sets $I_m=\{j:\ s_P^{(j)}\overline{\tau}_P^{(j)}+\tau_B^{(j)} = \min_i(s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)})\}$ (i.e., the set of the indexes for the lowest terms) and $I_M=\{j:\ s_P^{(j)}\overline{\tau}_P^{(j)}+\tau_B^{(j)} = \max_i(s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)})\}$ (i.e., the set of the indexes for the highest terms). In this situation, we could increase the variables $s_P^{(j)}$ for all $j\in I_m$ and, at the same time, decrease the variables $s_P^{(j)}$ for all $j\in I_M$ while keeping the sum $\sum_{i=1}^{M} s_P^{(i)}$ constant. This would allow to find a new configuration for which $\max_i(s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)})$ is reduced, which implies by contradiction that the optimal solution is attained when all the terms $s_P^{(i)}\overline{\tau}_P^{(i)}+\tau_B^{(i)}$ are equal.

REFERENCES

[1] P. Bhat, S. Nagata, L. Campoy, I. Berberana, T. Derham, G. Liu, X. Shen, P. Zong, and J. Yang, "LTE-Advanced: an operator perspective," *IEEE Commun. Mag.*, vol. 50, pp. 104–114, Feb. 2012.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," vol. 52, pp. 82–89, Feb. 2014.

[3] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, pp. 20–27, Jun. 2013.

[4] X. Ge, L. Pan, S. Tu, H.-H. Chen, and C.-X. Wang, "5G ultra-dense cellular networks," *IEEE Wireless Commun.*, vol. 23, pp. 72–76, Feb. 2014.

[5] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," *IEEE Wireless Commun. Mag.*, vol. 20, pp. 34–44, Jun. 2013.

[6] D. Kovachev and R. Klamma, "Framework for computation offloading in mobile cloud computing," *Int. Jorunal of Interactive Multimedia and Artificial Intelligence*, vol. 1, pp. 6–15, Dec. 2012.

[7] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications, Springer Science*, vol. 18, pp. 129–140, Febr. 2013.

[8] Amazon EC2. [Online]. Available: https://aws.amazon.com/ec2/.

[9] L. Gkatzikis and I. Koutsopoulos, "Migrate or not? Exploiting dynamic task migration in mobile cloud computing systems," *IEEE Wireless Commun. Mag.*, vol. 20, pp. 24–32, Jun. 2013.

[10] A. Miettinen and J. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. 2nd USENIX Conference on Hot Topics in Cloud Computing 2010 (HotCloud'10)*, Jun. 2010.

[11] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM'12)*, pp. 945–953, Mar. 2012.

[12] E. Cuervo, A. Balasubramanian, D.-K. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: making smartphones last longer with code offload," in *Proc. Int. Conf. on Mobile Systems, Applications, and Services (MobiSys'10)*, pp. 49–62, Jun. 2010.

[13] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: can offloading computation save energy?," *IEEE Computer*, vol. 43, pp. 51–56, Apr. 2010.

[14] E. Lagerspetz and S. Tarkoma, "Mobile search and the cloud: the benefits of offloading," in *Proc. IEEE Int. Conf. on Pervasive Computing and Commun. (PerCom'11)*, pp. 117–122, Mar. 2011.

[15] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 4569–4581, Sep. 2013.

[16] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogenous networks," *IEEE Signal Process. Mag.*, vol. 31, pp. 45 – 55, Nov. 2014.

[17] X. Chen, L. Jiao, W. Li, and W. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Networking*, vol. 24, pp. 2795 –2808, Oct. 2016.

[18] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys and Tutorials*, vol. PP, Jun. 2017.

[19] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "Mobile edge computing: The communication perpective," *eprint arXiv:1701.01090*, 2017.

[20] O. Muñoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Vehicular Technology*, vol. 64, pp. 4738–4755, Oct. 2015.

[21] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, pp. 89 – 103, Jun. 2015.

[22] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for mobile cloud with computing access point," *Proc. IEEE ICASSP*, pp. 3516 – 3520, Mar. 2016.

[23] R. Hwang, M. Gen, and H. Katayama, "A comparison of multiprocessor task scheduling algorithms with communication costs," *Computers and Operations Research*, vol. 35, pp. 976–993, Mar. 2008.

[24] J. Oueis, E. Calvanese-Strinati, A. Domenico, and S. Barbarossa, "On the impact of backhaul network on distributed cloud computing," in *IEEE Wireless Commun. and Networking Conf. Workshops (WCNCW)*, pp. 12–17, Apr. 2014.

[25] G. Auer *et al.*, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," tech. rep., deliverable report D2.3, ICT-247733 EARTH project, available at: https://www.ict-earth.eu/, Jan. 2012.

[26] A. R. Jensen, M. Lauridsen, P. Mogensen, T. B. Sorensen, and P. Jensen, "LTE UE power consumption model: for system level energy and performance optimization," in *Proc. IEEE Vehicular Technology Conf. Fall (VTC'12)*, pp. 1–5, Sep. 2012.

[27] A. Pascual-Iserte, *Channel state information and joint transmitter-receiver design in multi-antenna systems*. PhD thesis, Universitat Politècnica de Catalunya, Dec. 2004. PhD thesis.

[28] M. Biguesh and A. Gershman, "Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals," *IEEE Trans. Signal Process.*, vol. 54, pp. 884–893, Mar. 2006.

[29] R. Steele and L. Hanzo, *Mobile radio communications*. John Wiley and Sons, 2nd ed. ed., 1999.

[30] A. Pascual-Iserte, D. P. Palomar, A. I. Perez-Neira, and M. A. Lagunas, "A robust maximin approach for MIMO communications with imperfect channel state information based on convex optimization," *IEEE Trans. Signal Process.*, vol. 54, pp. 346–360, Jan. 2006.

[31] A. Papoulis, *Probability, random variables and stochastic processes*. McGraw-Hill Series in Electrical Engineering, third ed., 1991.

[32] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, "Bandwidth estimation: metrics, measurement techniques, and tools," *IEEE Network*, vol. 17, pp. 27–35, Nov. 2003.

[33] C. Dovrolis, P. Ramanathan, and D. Moore, "What do packet dispersion techniques measure?," *Proc. IEEE INFOCOM*, pp. 905–914, Apr. 2001.

[34] R. Kapoor, L.-J. Chen, L. Lao, M. Gerla, and M. Y. Sanadidi, "CapProbe: a simple and accurate capacity estimation technique," *Proc. ACM Special Interest Group Data Comm.*, Aug.-Sep. 2004.

[35] J. Wang and D. P. Palomar, "Worst-case robust MIMO transmission with imperfect channel knowledge," *IEEE Trans. Signal Process.*, vol. 57, pp. 3086–3100, Aug. 2009.

[36] K. B. Petersen and M. S. Pedersen, *The matrix Cookbook*. http://matrixcookbook.com, 2012.

[37] D. P. Palomar, *A unified framework for communications through MIMO channels*. PhD thesis, Universitat Politècnica de Catalunya, May 2003. PhD thesis.

[38] G. Scutari, D. P. Palomar, and S. Barbarossa, "The MIMO iterative waterfilling algorithm," *IEEE Trans. Signal Process.*, vol. 57, pp. 1917–53, May 2009.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[40] M. Molina, O. Muñoz, A. Pascual-Iserte, and J. Vidal, "Joint scheduling of communication and computation resources in multiuser wireless application offloading," in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC'14)*, Sep. 2014.

[41] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice Hall, first ed., 1993.

[42] M. L. Eaton, *Multivariate statistics: a vector space approach*. Institute of Mathematical Statistics, 2007.

Sandra Lagen received the Telecommunication Engineering degree in 2011, the M.S. degree in 2013, and the Ph.D. degree in 2016, from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. From 2012 to 2016 she was a research assistant in the Signal Theory and Communications (TSC) Department at UPC, where she has participated in the EC funded projects FREEDOM and TROPIC, and an industrial collaboration on flexible duplexing for 5G systems. In 2015, she did a research appointment at Nokia Networks in Aalborg, Denmark. Since 2017 she is a researcher in the Mobile Networks Department at Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), where she is participating in an industrial collaboration on 5G New Radio access technology design. Her research interests include wireless communications, signal processing, MIMO, and optimization theory.

Antonio Pascual-Iserte (S'01-M'07-SM'11) was born in Barcelona, Spain, in 1977. He received the M.Sc. degree in electrical engineering and the Ph.D. degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2000 and 2005, respectively. From September 1998 to June 1999, he was a Teaching Assistant in the field of microprocessor programming with the Department of Electronic Engineering, UPC. From June 1999 to December 2000, he was with Retevision R&D, working on the implantation of the DVB-T and T-DAB networks in Spain. In January 2001, he joined the Department of Signal Theory and Communications, UPC, where he worked as a Research Assistant until September 2003. He received a predoctoral grant from the Catalan government for his Ph.D. studies during this period. He became Assistant Professor in September 2003 and Associate Professor in April 2008. Currently, he teaches undergraduate courses on linear systems and signal theory. He also teaches postgraduate courses on advanced signal processing with the Department of Signal Theory and Communications. He has been involved in several research projects funded by the Spanish Government and the European Commission. He has also published several papers in international and national conference proceedings and journals. His current research interests include array processing, robust designs, orthogonal frequency-division multiplexing, multiple-input multiple-output channels, multiuser access, and optimization theory.

Dr. Pascual-Iserte received the 'First National Prize of 2000/2001 University Education' from the Spanish Ministry of Education and Culture and the 'Best 2004/2005 Ph.D. Thesis Prize' from UPC.

Olga Muñoz (M'11) received the M.S. degree in 1993 and the PhD degree in 1998, both in Electrical Engineering from the Universitat Politècnica de Catalunya (UPC), Spain. In 1994 she joined the Department of Signal Theory and Communications at the same University, where she teaches graduate and undergraduate courses related to signal processing and communications. She became an Associate Professor in 2001. She has been a visiting associate professor at Stanford University (September-November 2014 and January-June 2015). She has served as a reviewer for the Spanish Research Council. In addition, she has served also as a reviewer in numerous journal and conferences. She accumulates a substantial experience in European Commission projects. She worked in ROMANTIK (5thFP), FIREWORKS (6thFP), and ROCKET (7thFP), projects devoted to relaying and cooperative upgraded networks. Her experience on heterogeneous and femto-based networks is backed by her work on FREEDOM (7thFP) where she collaborated in several contributions to 3GPP-LTE, and later on by her work in the Spanish Government funded project MOSAIC (call 2010), were she was the Principal Investigator. More recently, she has participated in TROPIC (7thFP) pushing the idea of merging cloud computing with femtocell networking. She also has worked in TUCAN3G (7thFP) project focused on providing connectivity to rural areas through new wireless technologies for the access network as well as WiLD (WiFi for Long Distances)-WiMAX-VSAT heterogeneous backhauling. She has published over 50 papers in books, international conferences and journals in the areas of signal processing and communications.

Josep Vidal (M'91) received the Telecommunication Engineering and the Ph. D. degrees from the Universitat Politècnica de Catalunya (UPC), Barcelona, where he is Professor at the Signal Theory and Communications department. His research interests are in statistical signal processing, information and communication theory, areas in which he has authored +170 journal and conference papers. Since 2002 has coordinated collaborative EC-funded projects ROMANTIK, FIREWORKS, ROCKET, FREEDOM, TROPIC and TUCAN3G, belonging to the FP5, FP6 and FP7 programmes, all in different areas of MIMO relay communications, self-organization, cooperative transmission and heterogeneous networks. He has held research appointments with EPF Lausanne, INP Toulouse and University of Hawaii, and has organized several international workshops. From 2011 through 2014 he served as associate editor of IEEE Transactions on Signal Processing. Since 2016 he is member of the IEEE ComSoc Signal Processing for Communications and Electronics Technical Committee.