2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC) | 978-1-6654-9455-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/SPAWC51304.2022.9833958

Reconfigurable Intelligent Surface Aided Mobile Edge Computing over Intermittent mmWave Links

Fatima Ezzahra Airod[1], Mattia Merluzzi[1], Paolo Di Lorenzo[2,3], Emilio Calvanese Strinati[1]
[1]CEA-Leti, Université Grenoble Alpes, F-38000 Grenoble, France
[2]DIET department, Sapienza University of Rome, Italy, [3]CNIT, Parma, Italy
email:{fatima-ezzahra.airod, mattia.merluzzi, emilio.calvanese-strinati}@cea.fr,{paolo.dilorenzo}@uniroma1.it

***Abstract*—The advent of Reconfigurable Intelligent Surfaces (RISs) in wireless communication networks unlocks the way to support high frequency radio access (e.g. in millimeter wave) while overcoming their sensitivity to the presence of deep fading and blockages. In support of this vision, this work exhibits the forward-looking perception of using RIS to enhance the connectivity of the communication links in edge computing scenarios, to support computation offloading services. We consider a multi-user MIMO system, and we formulate a long-term optimization problem aiming to ensure a bounded end-to-end delay with the minimum users' average transmit power, by jointly selecting uplink user precoding, RIS reflectivity parameters, and computation resources at a mobile edge host. Thanks to the marriage of Lyapunov stochastic optimization, projected gradient techniques and convex optimization, the problem is efficiently solved in a per-slot basis, requiring only the observation of instantaneous realizations of time-varying radio channels and task arrivals, and that of communication and computing buffers. Numerical simulations show the effectiveness of our method and the benefits of the RIS, in striking the best trade-off between power consumption and delay for different blocking conditions, also when different levels of channel knowledge are assumed.**

I. Introduction

The advent of the sixth generation of mobile communication systems (6G) unveils several ambitions that span from the support of new services, to completely new key performance indicators. Indeed, we are facing an unprecedented revolution of applications, such as immersive virtual reality, connected autonomous systems, and the industrial Internet of Things, all verticals that require real time data transmission and processing [1]. Of course, to be effective, these services require to be enabled with new levels of dependability, reliability and sustainability. From a radio access perspective, the adoption of higher frequency such as millimeter wave (mmWave) and TeraHertz (THz) bands certainly enhances radio access network capacity, although at the price of a higher sensitivity to the presence of spatial blockages and, in general, to deep fading events that may hinder the aforementioned vision on performance [2], [3]. To this end, Reconfigurable Intelligent Surfaces (RISs) have recently emerged as a promising candidate to counteract the above mentioned issue, thanks to their ability to opportunistically shape the wireless propagation environment. More precisely, RISs are composed of scattering elements that can be adaptively configured to shape the incident wave through adjustable phase shifts, with the aim of improving system performance in specific locations in space and time [4], [5]. Therefore, owing to their abilities of customizing time-varying wireless propagation environments, RISs mark undeniably the dawn of the 6G era. Another key technological enabler of the 6G vision, already introduced in 5G, is Multi-access Edge Computing (MEC), which brings storage and computing resources close to end users, aiming to enable a new class of connect-compute services [6]. As such, the interplay between RISs and MEC plays a key role in improving network performance, thanks to the double benefit of computation and communication aspects, to be tackled and optimized jointly. In this paper, we focus on computation offloading services, whose goal is to move the execution of computation demanding applications from resource-poor end devices to nearby Mobile Edge Hosts (MEHs), to enable energy efficient, low latency, and reliable processing [7]. In particular, we investigate on the promising convergence of RISs and MEC, mainly focusing on a joint optimization of radio and computing resources, down to the wireless propagation environment properties.

Related works. Most of investigated works in the literature have been focused on addressing computation offloading upon appropriate wireless environments, tackling the joint optimization of communication and computation resources in MEC-enabled wireless networks [6], [9]. However, inevitably, moving towards higher frequency bands to cope with large data volumes is no more suitable for MEC systems due to the unpredictable and intermittent nature of wireless links. Indeed, blocking events may deteriorate the overall network performance. In line with this, the recent literature has involved the prominence of RISs to boost the performance of MEC systems in terms reliability [7], [10], [11]. Nevertheless, to the best of our knowledge, a dynamic joint optimization of computing resources and RIS-aided MIMO radio parameters is lacking.

Contribution. In this work, we propose an algorithm aiming to dynamically configure RIS parameters, users' uplink precoding, and computation resources, with the goal of minimizing users' transmit power, with guaranteed finite E2E delay of the offloading service. Thanks to the theory of Lyapunov stochastic optimization, we are able to split a long-term problem into consecutive deterministic optimization problems, based on instantaneous observations of context parameters. The solution of the latter, from a radio perspective, builds on an alternating optimization strategy that couples a projected

The work of Airod, Di Lorenzo and Calvanese Strinati has been partially funded by the H2020 project RISE-6G no. 101017011.

gradient step for the RIS parameters [13], and a water-filling solution for the users' precoding [14]. The computation resource allocation problem is also solved with low complexity.

II. System Model

In this work, we consider a dynamic system, in which a set $\mathcal{U}$ of N users continuously generate data locally and offload them to an MEH through the wireless connection with an Access Point (AP). In such a scenario, context parameters such as wireless channels and data arrivals vary over time, thus calling for a dynamic cross-layer optimization involving users' precoding for uplink transmission, RIS reflectivity matrix, and computation resources at the MEH. Then, we consider time as organized in time slots $t = 1, 2, \dots$ of equal duration τ. Given a random variable X, we denote by $\overline{X}$ its long-term average:

$$\overline{X} = \lim_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathbb{E}\{X(t)\} \tag{1}$$

A. Communication model

1) RIS reflectivity model: An RIS can be modeled as an array of M nearly passive elements, whose phases can be opportunistically tuned. Therefore, the RIS can be characterized by its reflectivity matrix that is represented, at time t, as $\mathbf{\Theta}(t) = \text{diag}\{r_1(t), ..., r_M(t)\}$, where r_i, $i = 1..., M$, are the complex reflection coefficients of each element, characterized by a fixed amplitude, which we assume to be unitary, and an adjustable phase shift θ_i, therefore we have $r_i = \alpha_i e^{j\theta_i}$. In the sequel, we will denote by $\mathbf{r}(t)$ the vector $\mathbf{r}(t) = \{r_i(t)\}_{i=1}^{M}$.

2) Channel model: We consider a MIMO system, in which the AP is equipped with N_a antennas, while each user is equipped with K antennas. For each user k, the E2E channel matrix $\mathbf{H}_k(t)$ at time t is composed of: i) a direct channel $\mathbf{H}_{k,d}(t) \in \mathbb{C}^{N_a \times K}$ between the user and the AP; ii) an indirect link, comprising the channel $\mathbf{H}_{k,r}(t) \in \mathbb{C}^{M\times K}$ between the user and the RIS, and the channel $\mathbf{H}_{r,a}(t) \in \mathbb{C}^{N_a\times M}$ between the RIS and the AP. Also, since we consider blocking, we define $\beta_{k,a}(t) \in \{0, 1\}$ ($\beta_{k,r}(t) \in \{0, 1\}$), which equals 1 if the direct (indirect) link experiences a blockage event. Then, the overall channel matrix of user k can be written as follows [15] (we omit the index t for ease of notation):

$$\mathbf{H}_k = (1 - \beta_{k,a})\mathbf{H}_{k,d} + (1 - \beta_{k,r})\mathbf{H}_{r,a}\Theta\mathbf{H}_{k,r}. \tag{2}$$

Note that, in the sequel, we will denote by $p_{k,a}$ ($p_{k,r}$) the blocking probability of the direct (indirect) link, i.e. the probability that $\beta_{k,a}$ ($\beta_{k,r}$) equals 1, which can be computed as the expectation of $\beta_{k,a}$ ($\beta_{k,r}$). Uplink transmission is a fundamental phase of computation offloading services, foreseen to also increase future uplink traffic [16]. Thus, let us now formalize the uplink parameters, whose optimization will be presented in the sequel. Letting $\mathbf{Q}_k(t) \in \mathbb{C}^{K\times K}$ be the transmit covariance matrix of user k at time t, the experienced data rate reads as follows

$$R_k(t) = W_k \log_2 \left|\mathbf{I} + \sigma_k^{-2}\mathbf{H}_k(t)\mathbf{Q}_k(t)\mathbf{H}_k^H(t)\right| \tag{3}$$

where W_k represents the bandwidth assigned to user k, and the noise power is $\sigma_k^2 = N_0 W_k$, with N_0 the noise power spectral density. Here, we assume that users are served thorough orthogonal channels with a frequency division multiplexing. Obviously, besides the current channel state conditions, the data rate depends on the user transmit covariance and the RIS parameters, which we will jointly optimize in the sequel.

B. Queuing Model and delay

Computation offloading services generally entail three phases, along with their respective delays: i) uplink communication buffering and transmission of input data; ii) computation buffering and computation; iii) downlink communication buffering and transmission of results. In this work, we consider the first two delays, although considering the last one would not substantially change the system model, as presented in [7]. Therefore, we model the E2E delay thorough a queueing system, comprising an uplink communication buffer $B_{l,k}(t)$, and a computation buffer $B_{r,k}(t)$ for each user k.

Communication buffer: The uplink buffer of each user k is fed by new arrivals $A_k(t)$ at time t, and drained by transmitting bits over the wireless interface at rate $R_k(t)$ (cf. (3)). Given a slot of duration τ, the queue evolves as:

$$B_{l,k}(t+1) = \max\left(0, B_{l,k}(t) - \tau R_k(t)\right) + A_k(t) \tag{4}$$

Computation buffer: Assuming that all computation tasks are offloaded to the MEH, we consider a computation queue for each UE, which is fed by the its arriving data in uplink, and drained by the computation performed at the MEH. We assume a linear relation between the number of transmitted bits and the CPU cycles. Then, denoting by J_k the number of CPU cycles per bit, the remote computation queue evolves as:

$$\begin{aligned} B_{r,k}(t+1) = &\max\left(0, B_{r,k}(t) - \tau f_k(t)/J_k\right) \\ &+ \min(B_{l,k}(t), \tau R_k(t)), \end{aligned} \tag{5}$$

where $f_k(t)$ represents the amount of resources (in CPU cycles/s) allocated to user k during time slot t. Due to Little's law, the average E2E delay experienced by each device is proportional to the sum queue length [6]: $\overline{D_k} = \tau \frac{\overline{B_{l,k}} + \overline{B_{r,k}}}{\overline{A_k}}$.

III. Problem formulation

In this paper, we jointly optimize users' uplink covariance matrix, RIS parameters, and computation resources at the MEH, to minimize the users' transmit power under queue stability constraints. The problem can be formulated as follows:

$$\begin{aligned} \min_{\{\mathbf{Q}_k(t)\}_k, \mathbf{r}(t), \{f_k(t)\}_k} \quad & \sum\nolimits_{k\in\mathcal{U}} \overline{\text{Tr}(\mathbf{Q}_k)} \\ \text{subject to } & (a)\ \overline{B_{l,k}} < \infty, \quad \forall k \ \ (b)\ \overline{B_{r,k}} < \infty, \quad \forall k \\ & (c)\ \mathbf{Q}_k(t) \succcurlyeq 0, \quad \forall k \ \ (d)\ \text{Tr}(\mathbf{Q}_k(t)) \le P_k^{\max}, \quad \forall k \\ & (e)\ |r_i(t)| = 1,\ \forall i \quad (f)\ f_k(t) \ge 0, \quad \forall k \\ & (g)\ \sum\nolimits_{k\in\mathcal{U}} f_k(t) \le f_{\max}. \end{aligned} \tag{6}$$

The constraints of (6) have the following meaning: (a)-(b) the local and remote queues of each user are stable; (c) the transmit covariance matrix of each user is semidefinite positive; (d) the uplink transmit power of each user is lower than

a maximum value $P_k^{\max}$; (e) the RIS reflectivity entries are complex exponential; $f)$ the CPU cycle frequency allocated to each user by the MEH is non-negative; (g) The sum all CPU cycle frequencies assigned to each user is at most equal to the MEH CPU maximum frequency $f_{\max}$. Problem (6) is a priori very complex to solve, as it involves time averages performed on variables whose statistics are supposed to be unknown. To solve it in an efficient way, we leverage on Lyapunov stochastic optimization [17], which allows us to define a sequence of deterministic problems, based on instantaneous observations of context parameters. In particular, following [17], and defining the vector $\mathbf{b}(t) = [\{B_{l,k}(t)\}_k, \{B_{r,k}(t)\}_k]$, we can write the *Lyapunov function* as $L(\mathbf{b}(t)) = \frac{1}{2}\sum_{k\in\mathcal{U}}[B_{l,k}^2(t) + B_{r,k}^2(t)]$ [17], which is a measure of the overall congestion state of the system. Our aim is to drive the network towards stability, with the minimum transmit power. To this end, as in [17], let us define first the *drift-plus-penalty* (DPP) function $\Delta_p(t) = \mathbb{E}\{L(\mathbf{b}(t+1)) - L(\mathbf{b}(t)) + V\sum_{k\in\mathcal{U}}\text{tr}(\mathbf{Q}_k(t))|\mathbf{b}(t)\}$, which is the one slot conditional expected change of the Lyapunov function, with a penalty factor, weighted by a parameter V, used to trade-off users' transmit powers and queue backlogs, thus shaping the desired trade-off between power consumption and E2E delay. Interestingly, queues' stability ((a)-(b) in (6)) is guaranteed if the DPP is bounded by a finite constant for all t [17]. As in [17], we now proceed by minimizing a suitable upper bound of the DPP. The upper bound, whose derivations are omitted due to the lack of space (see, e.g., [17]) reads as

$$\Delta_p(t) \le C + \mathbb{E}\Big\{\sum\nolimits_{k\in\mathcal{U}}[(B_{r,k}(t) - B_{l,k}(t))\,\tau R_k(t) + A_k(t)B_{l,k}(t) - \tau B_{r,k}(t)f_k(t)/J_k + V\text{tr}(\mathbf{Q}_k(t))]|\mathbf{b}(t)\Big\},$$

where C is a positive constant, omitted due to the lack of space. By greedily minimizing this upper bound in each time slot (i.e. removing the expectation), queues' stability is guaranteed, as well as the asymptotic optimality of the solution as V increases, with the cost of increased queue backlogs (i.e. higher E2E delay) [17]. It is easy to show that the resulting problem can be split, in each time slot t, into a radio resource allocation sub-problem, including the optimization of user covariance matrices and RIS parameters, and a computation resource allocation sub-problem, to optimize the MEH CPU scheduling. The overall proposed dynamic resource allocation procedure is described in Algorithm 1, whose steps are described in the following. In particular, Sec. III-A describes the implementation of step 1, to optimize radio resources. The implementation of step 2 follows in Sec. III-B.

A. Radio resource allocation sub-problem

The radio resource allocation sub-problem (step 1 of Algorithm 1) involves $\{\mathbf{Q}_k(t)\}_k$ and $\mathbf{r}(t)$, and is formulated as

$$\min_{\{\mathbf{Q}_k(t)\}_k,\mathbf{r}(t)} \sum_{k\in\mathcal{U}} (V\text{Tr}(\mathbf{Q}_k(t)) - \tau(B_{l,k}(t) - B_{r,k}(t))R_k(t)) \quad \text{subject to} \quad (c)\text{-}(e) \text{ of (6)} \tag{7}$$

Problem (7) is non convex, due to the non linear equality constraint (e). However, given the RIS parameters, the problem is convex and enjoys a simple water-filling solution [14]. Then, the solution of (7) is built on an iterative optimization algorithm that alternatively optimizes (7) with respect to the RIS phase shift using the projected gradient descent method (PGM), as in [13], and optimally updates the uplink covariance matrices of all users through the water-filling method [14]. Steps 1.1 and 1.2 are implemented as follows.

1) RIS optimization step (Step 1.1 *of Algorithm 1):* The projected gradient step (step 1.1) with respect to $\mathbf{r}(t)$ comes with low complexity, as both the gradient and the projection can be written in closed form [13, Eq. (17a)]. However, differently from [13], we deal with a multi-user case. Nevertheless, thanks to the decoupling obtained through the Lyapunov optimization framework, in this case, the gradient is a weighted sum of different terms (corresponding to different users), where the weights include both communication and computation queues. This naturally introduces a scheduling of the RIS, which is therefore optimized to prioritize users with worse queueing states. The gradient with respect to $\mathbf{r}$ reads as follows [13]:

$$\nabla_{\mathbf{r}} f(\mathbf{r}, \{\mathbf{Q}_k\}_k) = -\tau \sum\nolimits_{k\in\mathcal{U}} W_k(B_{l,k} - B_{r,k}) \times \text{diag}\left(\mathbf{H}_{r,a}^H \left(I + \mathbf{Z}_k\mathbf{Q}_k\mathbf{Z}_k^H\right)^{-1}\mathbf{Z}_k\mathbf{Q}_k\overline{\mathbf{H}}_{k,R}^H\right), \tag{8}$$

where $\mathbf{Z}_k = \mathbf{H}_k/\sigma$, and $\overline{\mathbf{H}}_{k,R} = \mathbf{H}_{k,r}/\sigma$, and the operator $\text{diag}(L)$ saves the diagonal elements of an $N_L \times N_L$ matrix L into a vector. Finally, since $|r_i| = 1$ must hold for all $i = 1, \dots M$, the projection onto the unit circle reads as [13]:

$$P_\Theta(r_i) = r_i/|r_i|, \qquad \forall i = 1, \dots, M. \tag{9}$$

Step 1.1 of Algorithm 1 is implemented through (8) and (9).

2) Uplink covariances optimization (Algorithm 2): From (7), it can be easily observed that, once the RIS configuration is fixed, the problem with respect to $\{\mathbf{Q}_k(t)\}_k$ admits a low complexity solution. First of all, the problem is separable among the N users. Moreover, for a generic user k, if $B_{l,k}(t) \le B_{r,k}(t)$, both terms in (7) are monotone non-decreasing functions of the user transmit power. Therefore, in this case, the optimal solution is $\mathbf{Q}_k^*(t) = \mathbf{0}_K$, i.e. user k does not transmit (step $1.2.a$ of Algorithm 1). This holds true also in the case in which all links are blocked. Instead, for a generic user k for which $B_{l,k}(t) > B_{r,k}(t)$ holds, the problem is convex and is similar to the one presented in [14], thus admitting the water-filling procedure in Algorithm 2.

B. Computation resource allocation sub-problem

The second sub-problem, necessary to implement step 2 of Algorithm 1, is formulated as follows:

$$\begin{aligned} \max_{f_k(t)} \quad & \sum\nolimits_{k\in\mathcal{U}} B_{r,k}(t)f_k(t)/J_k \\ \text{subject to} \quad & a)\ 0 \le f_k \le \min\left(f_{\max}, B_{r,k}(t)J_k/\tau\right), \quad \forall k \\ & b)\ \sum\nolimits_{k\in\mathcal{U}} f_k(t) \le f_{\max}, \end{aligned} \tag{10}$$

where, for efficiency purposes, we added the constraint in (a) that prevents each user to be allocated more frequency than the one needed to empty the remote queue. Problem (10) is linear, and the optimal frequencies can be iteratively found by

assigning the whole available frequency to the user with the highest ratio $B_{r,k}(t)/J_k$. If this leaves available frequency, the left part is assigned to the subsequent users until draining the whole CPU power of the server $f_{\max}$, or serving all users [7].

Algorithm 1 Dynamic optimization of RIS-assisted MEC

Require: V, $\mathcal{U} = \{1, \ldots, N\}$ N_{slots}, τ, P_k^{max}, $B_{l,k}(0)$, $B_{r,k}(0)$, J_k, $\forall k \in \mathcal{U}$, $f_{\max}$,
for $t = 1 : N_{slots}$ **do**
step 1: Optimize $\{\mathbf{Q}_k(t)\}_k$ and $\mathbf{r}(t)$.
  for $n = 1 : I_{\max}$ **do**
step 1.1: $\mathbf{r}^{n+1} = P_\Theta(r_n - \rho \nabla_{\mathbf{r}} f(\mathbf{r}^n, \{\mathbf{Q}_k^{(n)}\}_k))$
step 1.2:
    for $k = 1, \ldots, N$ **do**
      a: If $B_{l,k} \leq B_{r,k}$, $\mathbf{Q}_k^{(n+1)} = \mathbf{0}_K$, else
      b: Update optimal $\{\mathbf{Q}_k^{(n+1)}\}_k$ with Algorithm 2
    end for
  end for
step 2: Optimize $\{f_k(t)\}_k$ as in Section III-B
step 3: Compute $R_k(t), \forall k \in \mathcal{U}$ as in (3);
step 4: Update $B_{l,k}$ and $B_{r,k}$ as in (4) and (5), respectively.
end for

Algorithm 2 Uplink covariance optimization for user k [14]

step 1: Compute $\mathbf{H}_k^H \mathbf{H}_k = \mathbf{U}^H \mathbf{\Sigma} \mathbf{U}$, with $\mathbf{\Sigma}$ a diagonal matrix with non-negative elements $\sigma_i, i = 1, ..., K$
step 2: Check if $\sum_{i=1}^{K} \max\left(0, \frac{\tau W_k (B_{l,k} - B_{r,k})}{V} - \frac{1}{\sigma_i}\right) \leq P_k^{\max}$ holds. If yes, then let $\mu^* = 0$ and $\lambda_i^* = \max\left(0, \frac{\tau W_k (B_{l,k} - B_{r,k})}{V} - \frac{1}{\sigma_i}\right)$, else,
step 3: Take all $\sigma_i, i = 1, \ldots, K$ in a decreasing order, i.e. as, $\sigma_{d(1)} > \sigma_{d(2)} > \ldots > \sigma_{d(K)}$.
step 4:
Start with $S_0 = 0$.
for $i = 1 : K$ **do**
Let $S_i = S_{i-1} + \frac{1}{\sigma_{d(i)}}$ and $\mu^* = \frac{i}{S_i + P} - \frac{V}{\tau W_k (B_{l,k} - B_{r,k})}$. If $\mu^* \geq 0$, $\frac{1}{\mu^* + \frac{V}{\tau W_k (B_{l,k} - B_{r,k})}} - \frac{1}{\sigma_{d(i)}} \geq 0$, and $\frac{1}{\mu^* + \frac{V}{\tau W_k (B_{l,k} - B_{r,k})}} - \frac{1}{\sigma_{d(i+1)}} \leq 0$, then stop the loop, otherwise, move to the next iteration.
end for
step 5: Let $\lambda_i^* = \max\left[0, \frac{1}{\mu^* + \frac{V}{\tau W_k (B_{l,k} - B_{r,k})}} - \frac{1}{\sigma_i}\right]$.
step 6: $\mathbf{Q}_k^* = \mathbf{U}^H \mathbf{\Lambda}^* \mathbf{U}$, with $\mathbf{\Lambda}^*$ diagonal matrix with entries λ_i^*

IV. Numerical results

In this section, numerical results are provided to assess the performance of our strategy. We consider a scenario with $N = 6$ users aiming to offload their tasks to a MEH collocated at the AP serving the users. Each user is assigned an equal portion of the total bandwidth $B = 1$ MHz, while the noise power spectral density is set to $N_0 = -174$ dBm/Hz. The slot duration is set to $\tau = 10$ ms. The arrival rate is 1 Mbps with Poisson distribution, for all users. The maximum available CPU cycle frequency is $f_{\max} = 4.5$ GHz and $J_k = 500$ $\forall k$ (cf. (5)). All channels (cf. (2)) are generated for a typical mmWave operating frequency, $f = 28$ GHz, as in [15], with: $K = 4$, $N_a = 4$, and $M = 64$. The maximum transmit power for a single user k is set to $P_k^{\max} = 100$ mW. In the sequel, for the sake of comparison, we consider both scenarios with and without the RIS. Also, we assume two different degrees of channel knowledge: i) **Alg. 1:** instantaneous knowledge of $\beta_{k,r}(t)$ and $\beta_{k,a}(t)$ (cf. 2) is assumed, and Alg. 1 is used for radio resource allocation; ii) **Alg. 1, statistical**: only a statistical knowledge of the blockage, i.e. the blocking probabilities $p_{k,a}$ and $p_{k,r}$ is assumed. In this case, Algorithm 1 is used, but $\beta_{k,r}(t)$ and $\beta_{k,a}(t)$ are replaced by $p_{k,a}$ and $p_{k,r}$ in (2), for the optimization. Obviously, the data rate experienced by each user is computed with the true channel in (2). Furthermore, for all cases, we consider also the case in which the RIS phase shifts are randomly selected; whereas, for **Alg. 1**, we also consider the case in which, after step 1.1 of Algorithm 1, the RIS phase shifts are quantized with 2 bits, which is a practical constraint of RIS implementation [7].

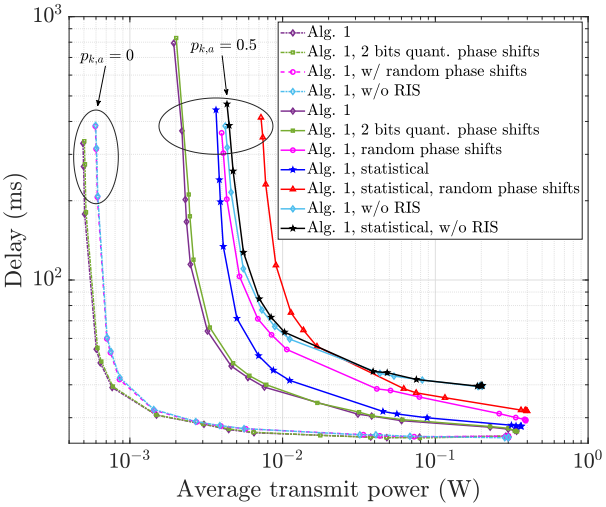

Fig. 1: Delay-power trade-off

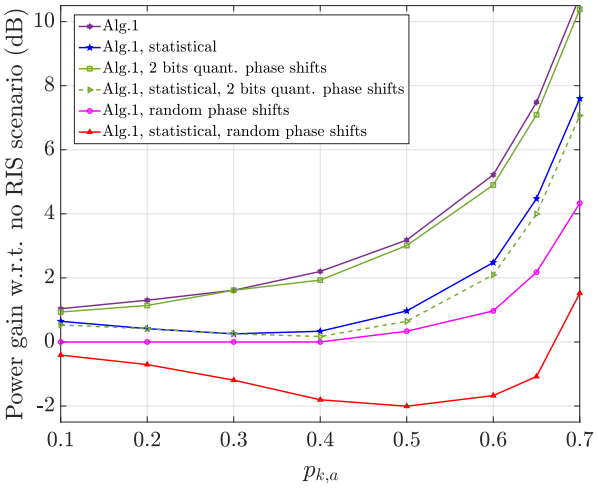

Fig. 2: Avg. transmit power vs. AP blocking probability

As a first result, in Fig. 1, we show the trade-off between

average E2E delay and transmit power, for different blocking probabilities $p_{k,a}$ on the direct link. We specifically plot the results for $p_{k,a} = 0$ and $p_{k,a} = 0.5$[1], obtained by increasing the Lyapunov trade-off parameter V from right to left. For all curves we can notice how, by increasing V, the system average transmit power decreases while the average service delay increases. Also, all scenarios with the optimized RIS outperform the scenario without the RIS, although with negligible gain for the case without blocking. This suggests that the benefits of the RIS are more significant in case of high blocking probability of the direct link. Also, the **Alg. 1, statistical** strategy performs better than the non RIS case, except for the random phase shifts case. Finally, for **Alg. 1**, it can be noticed that a 2-bit quantization of RIS phase shifts yields parallel performance as the ideal case (i.e., continuous) with a very negligible gap, thus suggesting that our method can be exploited also for practical RIS optimization.

To further highlight the previously mentioned remarks, we illustrate, through Fig. 2, the gain in terms of average transmit power of each strategy with respect to the non RIS case, as a function of the direct link blocking[2], for a fixed E2E delay bound of 150 ms, obtained by tuning the trade-off parameter V. As expected, as the blocking probability increases, the gain notably increases with the **Alg. 1** strategy, (up to 10 dB for $p_{k,a} = 0.7$), also with quantized phases. Conversely, the gain of **Alg. 1, statistical** is visible only for higher blocking probabilities, due to the fact that, in this case, the channel knowledge is well-matched to the real channel states. Eventually, this implies that unreliable blocking knowledge is critical for the performance. Optimizing the RIS through step 1.1 of Algorithm 1 leads to a better exploitation of the indirect path. However, for the random phase case, it can be noticed that no gain is achieved (less than 2 dB in the best case), which is obvious since we have no control on the RIS. More specifically, it can be noticed that for $p_{k,a}$ around 0.5, the channel knowledge is completely mismatched. Instead, at high blocking probabilities $p_{k,a} \geq 0.6$, the mismatch is reduced and higher gain is achieved. Overall, we can conclude that the use of an RIS is prominent to satisfy a reliable MEC-based task offloading in case of bad conditions of the direct link, i.e., higher $p_{k,a}$, for all strategies. More specifically, the benefit of the RIS starts to be noticed with a lower $p_{k,a}$ for the best strategy, while it becomes more visible with higher $p_{k,a}$ for the worst strategy.

V. Conclusion

In this work, we have explored the effectiveness of using an RIS to counteract the blocking (e.g. unreliability) when increasing communication frequency, in the case of computation offloading services. To this end, we considered a blocking aware framework through which we investigated the dynamic joint optimization of computing resources and RIS-aided multi-user MIMO communication parameters. Then, for dynamic configuration, we applied Lyapunov optimization tools to transform a complex long-term optimization problem into a per-slot deterministic problem that requires only instantaneous observations of the context parameters and properly defined state variables. Numerical results show the inherent gain of empowering MEC with RISs, while assuming different degrees of channel knowledge along with different scenarios and blocking conditions.

[1]For $p_{k,a} = 0$, $\beta_{k,a}(t) = 1, \forall k, t$, so that only **Alg. 1** is shown

[2]Problem (6) is not feasible for $p_{k,a} > 0.7$ without the RIS

References

[1] E. Calvanese Strinati et al., "Wireless Environment as a Service Enabled by Reconfigurable Intelligent Surfaces: The RISE-6G Perspective," Proc. of EUCNC 6G Summit, Porto, Portugal, Jun. 2021.

[2] C. K. Anjinappa, F. Erden and I. Güvenç, "Base Station and Passive Reflectors Placement for Urban mmWave Networks," in IEEE Trans. Veh. Technol., vol. 70, no. 4, pp. 3525-3539, Apr. 2021.

[3] G. Zhou, C. Pan, H. Ren, K. Wang, M. Elkashlan, and M. Di Renzo, "Stochastic learning-based robust beamforming design for RIS-aided millimeter-wave systems in the presence of random blockages," IEEE Trans. Veh. Technol., vol. 70, no. 1, pp. 1057–1061, Jan. 2021.

[4] M. Di Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and The Road Ahead," IEEE J. Sel. Areas Commun., vol. 38, no. 11, pp. 2450–2525, 2020.

[5] Q. Wu and R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," IEEE Commun. Mag., vol. 58, no. 1, pp. 106–112, 2019.

[6] M. Merluzzi, P. D. Lorenzo, S. Barbarossa and V. Frascolla, "Dynamic Computation Offloading in Multi-Access Edge Computing via Ultra-Reliable and Low-Latency Communications", IEEE Trans. Signal Inf. Process. Netw., vol. 6, pp. 342-356, Mar. 2020.

[7] P. Di Lorenzo, M. Merluzzi, E. C. Strinati, and S. Barbarossa, "Dynamic Edge Computing Empowered by Reconfigurable Intelligent Surfaces", arXiv preprint arXiv:2112.11269, 2021.

[8] Pengtao Zhao, Hui Tian, Cheng Qin, and Gaofeng Nie, "Energy-Saving Offloading by Jointly Allocating Radio and Computational Resources for Mobile Edge Computing," IEEE Access, vol. 5, pp. 11255–11268, 2017.

[9] D. Huang, P. Wang, and D. Niyato, "A Dynamic Offloading Algorithm for Mobile Computing," IEEE Trans. Wirel. Commun. vol. 11, no. 6, pp. 1991–1995, 2012.

[10] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency Minimization for Intelligent Reflecting Surface Aided Mobile Edge Computing," IEEE J. Sel. Areas Commun., vol. 38, no. 11, pp. 2666–2682, 2020.

[11] S. Huang, S. Wang, R. Wang, M. Wen, and K. Huang, "Reconfigurable Intelligent Surface Assisted Mobile Edge Computing with Heterogeneous Learning tasks", IEEE Trans. Cogn. Commun. Netw., 2021.

[12] S. Mao, S. Leng, S. Maharjan and Y. Zhang, "Energy Efficiency and Delay Tradeoff for Wireless Powered Mobile-Edge Computing Systems With Multi-Access Schemes", IEEE Trans. on Wirel. Commun, vol. 19, no. 3, pp. 1855-1867, Mar. 2020.

[13] N. S. Perović, L. -N. Tran, M. Di Renzo, and M. F. Flanagan, "Achievable Rate Optimization for MIMO Systems With Reconfigurable Intelligent Surfaces", IEEE Trans. on Wirel. Commun, vol. 20, no. 6, pp. 3865-3882, Jun. 2021.

[14] H. Yu and M. J. Neely, "Dynamic Transmit Covariance Design in MIMO Fading Systems With Unknown Channel Distributions and Inaccurate Channel State Information", IEEE Trans. on Wirel. Commun, vol. 16, no. 6, pp. 3996-4008, Jun. 2017.

[15] E. Basar, I. Yildirim, "Reconfigurable Intelligent Surfaces for Future Wireless Networks: A Channel Modeling Perspective", IEEE Wireless Commun., vol. 28, no. 3, pp. 108–114, Jun. 2021.

[16] J. Oueis and E. C. Strinati, "Uplink traffic in future mobile networks: Pulling the alarm", in Int. Conf. CROWN, pp. 583-593, Springer, Cham, May 2016.

[17] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems", M. & C. Publishers, 2010