Full and Partial Offloading in a Mobile Edge Computing(MEC) scenario with Distributed Uplink Non Orthogonal Multiple Access(DU-NOMA)

Draft of Master's Thesis
Supervisor: George Karagiannidis

Askitopoulos Athanasios

1 Introduction

1.1 Next generation wireless networks (5G and beyond)

With the commercial development and deployment of 5^{th} generation networks, evolving globally, intense interest is observed from both the academic community and the industry for the design of upcoming generation networks (5G and beyond, and 6G), as well as for determining the specifications, application fields, and appropriate techniques to meet the increasing demands. Each new generation so far has been characterized by an enhancement of communication capabilities with support for voice (1G and 2G), text messages (3G), internet connectivity (4G+), along with the accompanying increase in demands for user numbers, coverage capacity, communication rates, and spectral efficiency. The transition to the 5^{th} and 6^{th} generations not only requires an exponential increase in the above requirements but also the addition of radically different functionalities to the wireless network. An initial estimate of the growth compared to 4G predicts the ability to support 100 times higher mobile data volume per area, 10 to 100 times faster data rates, a fivefold reduction in communication latency, and support for up to 100 times more devices without increasing costs or energy consumption [1], [2].

Additionally, at a global level, average monthly data usage is expected to rise from 21 GB in 2023 to 56 GB in 2029, with total telecommunications

traffic tripling over the same period and the share of 5G reaching 76% [3], [4], [5].

Managing this high volume of telecommunications traffic as well as the upcoming needs for massive and reliable communication of heterogeneous devices led to the introduction of three basic service categories of 5G: enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine Type Communications (mMTC). Specifically, enhanced Mobile Broadband (eMBB) supports stable connections with high data rates for wide coverage and mobility scenarios and can be seen as a direct extension of existing networks, while Ultra-Reliable Low-Latency Communications (URLLC) supports highly reliable and ultra-low-latency connections from a limited set of active terminals for automation in industrial and healthcare sectors. Finally, mMTC supports short-duration, low-power connections from a large number of occasionally active devices in IoT network applications. These connectivity categories make it evident that the 5G network is designed as a "platform" of technologies that will enable the coverage of heterogeneous specifications in a spectrum of possible use cases.

The capabilities provided by these technologies can be summarized as follows: massive connectivity, increased spectral efficiency, reduced end-to-end latency, increased transmission rates, and low energy consumption.

Although the transition to the 5^{th} generation has not been completed, the vision for the characteristics and potential applications of the 6G network is rapidly unfolding within the academic community and the industry. Building on the existing capabilities of 5G, 6G will continue the effort to realize the Internet of Everything (IoE), expanding human-to-machine (as eMBB) and machine-to-machine type communications (as mMTC and URLLC), and deeply integrating Artificial Intelligence (AI). Vision-wise, 6G is purported as a distributed, omnipresent, "smart" network that will provide communication links between the physical and digital worlds, between humans, things, and machines [6]. This vision relies on the utilization of untapped spectrum areas such as sub-6GHz, THz, and optical bands in merging computation, communication, sensing, control, and "smart" decision-making capabilities and providing immersive, sensory experiences to users through holographic communication and extended reality (XR) applications [7], [8].

1.2 Non-Orthogonal Multiple Access (NOMA)

Non-Orthogonal Multiple Access (NOMA) represents a significant advancement in the field of 5G and is expected to serve as an exemplary method for designing radio spectrum access techniques for future generations of networks (Next generation Multiple access) [9], [10], [11].

In contrast to traditional access techniques such as TDMA, FDMA, CDMA, and OFDM, which utilize orthogonal, non-overlapping blocks of resources (time, frequency, code), thus limiting the number of served users to the number of available resources, in NOMA systems, users can simultaneously use the same block to access the network. Overloading of existing resources is achieved through some form of multiplexing at the transmitter and demultiplexing at some or all receivers, resulting in better utilization of telecommunication resources and increased efficiency.

In its primary formulation [11], during downlink communication between a base station (BS) and a set of users, the base station employs superposition coding (SC) to create a linear combination of users' messages, weighted by a percentage of the transmitted power P for each user, and transmits it over the channel. The power percentage is determined by the gain each user receives on its channel, defining strong and weak users accordingly. On the receiver side, each user decodes its message using successive interference canceling (SIC). Signals from weaker users relative to the receiver are progressively decoded and subtracted from the received signal to decode the message, using signals from stronger users as interference [12]. It is important to note that the base station assigns more power to weak users based on the channel coefficient, and less to strong users, allowing successful SIC operation by removing the strongest signal first at the receivers.

In a system using Orthogonal Multiple Access (OMA), each user would occupy one resource, even those with very poor channel conditions, where spectrum allocation would lead to reduced spectral efficiency and system throughput. With NOMA, strong and weak users can be adequately served with the same spectrum, achieving better user fairness and improvements in throughput.

1.3 Cloud - Radio Access Network (C-RAN)

Beyond multiple access techniques, the stringent requirements of next-generation networks in terms of coverage capacity for high user density and telecommunications traffic volume are expected to be met by the radical restructuring of the radio access network (RAN). The trend towards redefining the architecture of the radio access network is already evident in the transition from

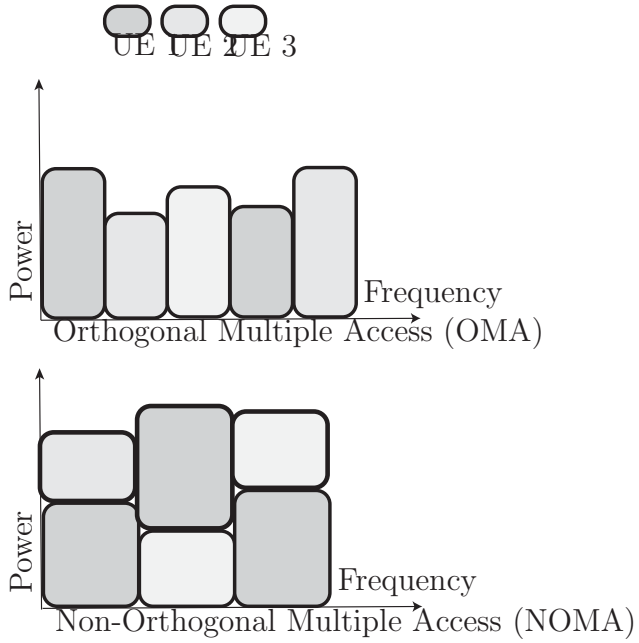

Figure 1: Representation of OMA and NOMA

the 2G RAN, with the common signal processing of the baseband and radio emission in Base Stations (BS), to the various architectures of 3G and 4G networks based on D-RAN, where this processing takes place in separate nodes: Remote Radio Heads (RRH) and Baseband Units (BBU). The necessity to cover heterogeneous Quality-of-Service requirements, along with the rapid increase in user data, has led network operators to adopt centralization, cloudification and virtualization techniques for BBUs and their corresponding RRHs.

The efforts led to the emergence of the Centralized Radio Access Network, an innovative architectural solution initially proposed by (IBM) under the name Wireless Network Cloud[13] and further elaborated in a white paper by the Chinese Mobile Research Institute[14]. As presented in , in the basic scenario of the application of the C-RAN architecture, the Basic Broadcasting Units (BBUs) are separated from their Remote Radio Heads (RRHs) and consolidated in a central, virtualized pool based on cloud computing. Each RRH is connected to the BBU pool via Fronthaul links, while the pool, which can support multiple RRHs, is connected to the network core through a Backhaul link. The RRHs are released from a dedicated BBU and are now served by virtualized BBUs which are created according to the service needs. By using optical fibers for these links, high capacity and reduced end-to-end latency (latency) are ensured for data transmission between the core and the

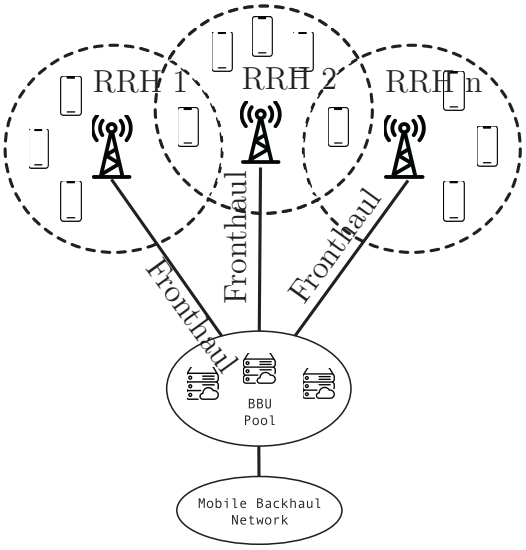

Figure 2: Representation of CRAN architecture

radio stations[15],[16].

The C-RAN architecture provides various advantages to 5G networks such as: the ability to expand network coverage or split existing cells for capacity enhancement, due to the ease of adding new RRHs connected to the BBU pool, support for intra-cell cooperation for network resource management and signal processing, improvement of spectral efficiency, ease of mobility management and load balancing, and reduction of network energy consumption[17],[18].

The energy efficiency of C-RAN has been noted as a crucial characteristic to support Green Communications[19], which require careful network expansion to support a greater number of devices, with a focus on energy savings without losses in communication quality. The majority of the energy used in previous-generation networks was consumed by base stations for the achievement of RAN processes as well as for the operation and cooling of equipment. Cloud computing techniques combined with centralized radio processing in the C-RAN architecture allow for a reduction in the number of BSs without a corresponding drop in network performance and capacity, as virtualized BBUs can dynamically cover telecommunication traffic according to user demand conditions[17].

Finally, centralized resource management and the use of a feedback link for communication among RRHs enable the use of cooperative communication techniques (CoMP)[15], non-orthogonal multiple access (NOMA), and distributed resource optimization for even greater energy savings[16].

1.4 Mobile Edge Computing (MEC)

The primary objective of future network generations beyond 5G is to reliably serve a vast number of devices running applications with heterogeneous latency requirements and connectivity capabilities. Emerging applications in healthcare (smart medicine) [20, 21], transportation and manufacturing (smart manufacturing), agriculture, and electrification (smart grids) [22] highlight the need for real-time processing of significant computational loads, alongside parallel secure and reliable data transfer and storage. One of the emerging solutions for this technological development framework is edge-centric computing [23], which combines the principles of cloud computing development with the capabilities provided by modern communication networks to leverage edge network nodes for computation.

To understand the need for transitioning from cloud computing to edge computing, it is worth mentioning the capabilities and limitations of the cloud. Specifically, cloud computing constitutes a set of technologies and a development standard that enables end-users to store and process data volumes at much higher speeds than the limited computational power and space of local devices allow. Central systems and servers with enhanced computational power and parallel processing capabilities, managed by cloud service providers, handle user tasks and transmit the final results via the network infrastructure. This dependency on the Internet infrastructure is also the main limitation of the technology, as the long distances between cloud servers and users introduce time delays and increase the likelihood of errors, rendering the system unreliable for delay-sensitive applications [24].

Edge computing was developed to exploit the significant amount of idle processing power and storage space of edge devices, mediating between end-users and cloud servers, bringing cloud computing services closer to the user. The application of this idea to the telecommunications network led to the development of Mobile Edge Computing (MEC), a network architecture concept proposed by the European Telecommunications Standards Institute (ETSI) in 2014. According to the ETSI definition, "Mobile Edge Computing provides Cloud Computing capabilities within the Radio Access Network (RAN) and in close proximity to mobile devices. Their goal is to reduce latency, ensure efficient network operation, provide services, and improve user experience" [25].

A key feature of MEC is the ability to offload computational tasks from edge devices to edge servers, which are installed at base stations (BS) in close proximity to users [24, 26, 27]. Computational tasks can be offloaded and executed either entirely (full offloading) or partially (partial offloading), offering flexibility in task processing and increasing energy efficiency. Fi-

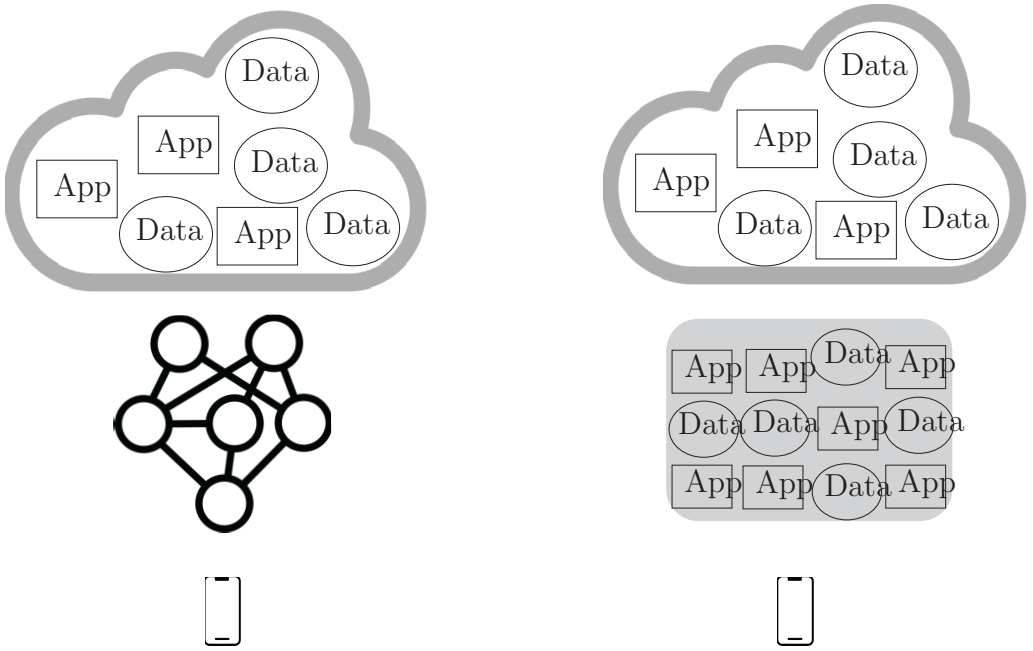

Figure 3: cloud computing vs MEC

nally, edge servers can execute resource management optimization tasks for both RAN and users, achieving resource sharing of computational and communication resources and providing the necessary infrastructure to support emerging real-time applications (AR, streaming, self-driving cars, IoT) and artificial intelligence in 5G/6G networks [28–31].

1.5 Related Research

In recent years, several studies have been conducted on NOMA and the possibilities it offers, in combination with other techniques, for multi-access networks. Specifically, many different versions of NOMA have been proposed, such as power domain NOMA, code domain NOMA, multi-carrier NOMA, sparse code multiple access NOMA, and distributed NOMA, all based on the same fundamental principle, whereby more than one user can be served by a single block of resources (e.g., time slot, frequency channel, spreading code) [9], [10], [11], [12]. The advantages of the method for multi-user access to the channel are attributed to resource savings, avoiding the exclusive allocation of blocks to users with poor channel conditions [11].

Power Domain NOMA

Power domain NOMA has received significant attention among these different versions for its ability to achieve the capacity of the downlink channel and the multiple access channel in the uplink, using Successive Interference Cancellation (SIC) by the receiver [11], [32] [33]. A key factor for the successful interference elimination is the selection of a decoding order for users at the receiver, based on precise knowledge of the gains the channel brings to each transmission. In the typical version of NOMA, a fixed and decreasing order is chosen based on how "strong" each user's channel is.

During uplink transmission, power domain NOMA uses time sharing among the potential transmission orders of users-transmitters, which can lead to improving the "fairness" in terms of the allocation of transmission rates and energy and achieving any point in the capacity region of the multiple access channel [34], [16]. The choice of time sharing allows users with reduced rate gains in a given decoding order to experience better reception reliability in another, due to the switching of orders within the transmission period T. The increasing complexity of SIC with the increase in the number of users and thus the potential decoding orders can be addressed by dividing users into pairs and applying power-domain NOMA between them [35]. It has been shown that with this method, the overall sum rate of all users can be optimized.

Distributed NOMA

Given the dominance of uplink links in emerging IoT applications as well as the C-RAN architecture, a distributed version of NOMA, the Distributed Uplink NOMA (DU-NOMA), is considered particularly important. Specifically, in [16], the problem of maximizing the capacity region in a DU-NOMA usage scenario in the uplink of a C-RAN architecture is investigated, assuming non-fixed transmission rates of users and the ability of communication among the RRHs of the system via high-capacity fiber links for error-free SIC execution. The case of SIC execution via limited-capacity links and imperfect communication among RRHs has been explored in [9].

The application of NOMA in MEC systems has been recognized in [34],[35],[36], as a combination of techniques that can lead to significant reductions in latency and energy consumption. Existing literature has focused on the advantages that NOMA can provide over traditional OMA techniques and the possibility that these techniques may need to be combined within the MEC framework. Specifically, the differences in latency tolerance of different users may render the use of pure NOMA inadequate, leading, in [37], [38], to the

adoption of hybrid NOMA-OMA architectures. In more detail, in a scenario of data offloading of 2 users, keeping one user's tolerated latency and transmit power fixed and assuming the other user has a higher latency tolerance, the problem of minimizing latency and energy consumption is solved. Each transmission frame is divided into a phase where users simultaneously transmit NOMA and a phase where only the user with higher latency tolerance transmits, using OMA, to complete the offloading process. During the SIC process, a fixed order of decoding is chosen, with the message of the user with fixed latency always decoded last, during the NOMA phase of the process, and the offloading is fully executed on the edge servers.

Partial Offloading

In research works [40], [41], [42], [43], models with the capability of partial offloading to edge servers were analyzed.Specifically, [40], [41] focused on minimizing energy consumption, while [42], [43] focused on minimizing the completion time of users' tasks using partial offloading in MEC systems.

The relationship between the parameters of local processing on edge devices and the offloading capability is another important direction of the literature in MEC systems. Thus, in [39], the ability to control the CPU clock speed of the device for computing part of a task is introduced through dynamic voltage scaling technique, followed by solving a joint optimization problem of computational speed, device transmit power, and offloading ratio with objective functions of energy consumption and total task execution time. It was observed that these two metrics exhibit a trade-off, and controlling the offloading ratio and CPU speed can effectively balance them.

A common metric of Quality of Experience, based on task execution time and device energy consumption, is maximized in [44] through a heuristic optimization algorithm that iteratively finds the appropriate offloading decision timing and optimal management of computational and communication resources of the proximal cloud. Managing the computational load of the edge server of an MEC system using NOMA for offloading can be aided by device-to-device (D2D) communications, as analyzed in [45]. In this publication, the weighted sum of energy consumption and total delay of all system users is selected as the optimization objective, while a heuristic algorithm based on particle swarm optimization (PSO) is chosen as the optimization method for the non-convex problem of power allocation to D2D pairs.

In contrast to the aforementioned works, in [46], the optimization of common computational and communication resources takes into account both the impact of the decoding order of users during the SIC process and the ability to vary the CPU clock speed with controlled clock speed. A similar inves-

tigation is conducted in [27], in a hybrid NOMA-OMA protocol, by adding time-sharing between different decoding order during the interference elimination phase in NOMA and using variable clock speed for local processing. Specifically, as in publications [37] and [38], where two phases were used, NOMA for simultaneous user offloading and OMA for remaining offloading of the weakest user, similarly in [27], a phase with time-sharing NOMA is used to achieve any channel rate of the multiple access channel, and an OMA phase to complete the offloadings. In the case of full offloading, the delay minimization problem is examined, limiting the energy consumption (delay-constrained energy minimization), and the energy minimization problem, limiting the acceptable delay (energy-constrained delay minimization). In the case of partial offloading, the problem of minimizing the weighted sum of delays for completing users' tasks is examined, with controlled CPU clock speed.

2 Contributions

The main contributions are summarized below:

- Describing a MEC scheme with the application of the DU-NOMA protocol, during which time sharing is used in the SIC process, through communication between the RRH, and presenting the rates achieved by the users for each possible decoding sequence of users.
- Formulating the problems of minimizing the energy consumption with limited acceptable delay (delay-constrained energy minimization), and minimizing the delay with limited acceptable energy consumption (energy-constrained delay minimization) for the case of full offloading, and the problem of minimizing the offloading delay, taking into account the constraints of local processing and the ability to dynamically control the speed of the CPU.
- Solving the optimization problems by transforming them into convex formulations through a combined methodology of geometric programming and successive convex optimization via Difference-of-Convex (DC) programming.
- Analyzing the impact of distance from the base stations on the optimal delay and energy consumption of the system and comparing the effectiveness of the methodology with a simpler methodology based on exhaustive search of possible power assignments during time sharing.

3 System Model and Proposed Protocol

We consider a MEC scenario with DU-NOMA consisting of 2 users (User Equipment): UE_i,UE_j offloading their computations on 2 RRHs: RRH_a,RRH_b, as shown in Figure 1.It is assumed that each remote radio head has perfect channel information and can perform the necessary optimization scheme.Let T denote the total duration of the time-sharing phase and let $g_{lm} = |h_l m|^2 * w_l m$ denote the channel gain of the l-th user,$\forall l \in i, j$ to RRH_m,$\forall m \in a, b$.Each channel gain includes both the small scale fading h_{lm} and the path loss factor w_{lm} for each choice of user and remote head.As we will see more thoroughly in the description of the protocol below, in each of the 8 communication scenarios between user l and RRH_m, the power expended by each user ,during the k-th scenario, is denoted by $P_l^{(k)}$ and the respective SNR as $p_l^{(k)} = \frac{P_l^{(k)}}{B*N_o}$,$\forall l \in i, j$ and $\forall k = 1, 2.., 8$.The power spectral density of the additive white Gaussian noise(AWGN) is denoted as N_o and B is the available bandwidth.

Each user is equipped with a CPU ,as depicted in Figure 1, and as such two optimization schemes are considered: full offloading, in which all the computation bits are sent to the RRHs, processed and send back[Ref for offloading] as well as partial offloading, in which some of the bits are processed on the local CPU of each UE.

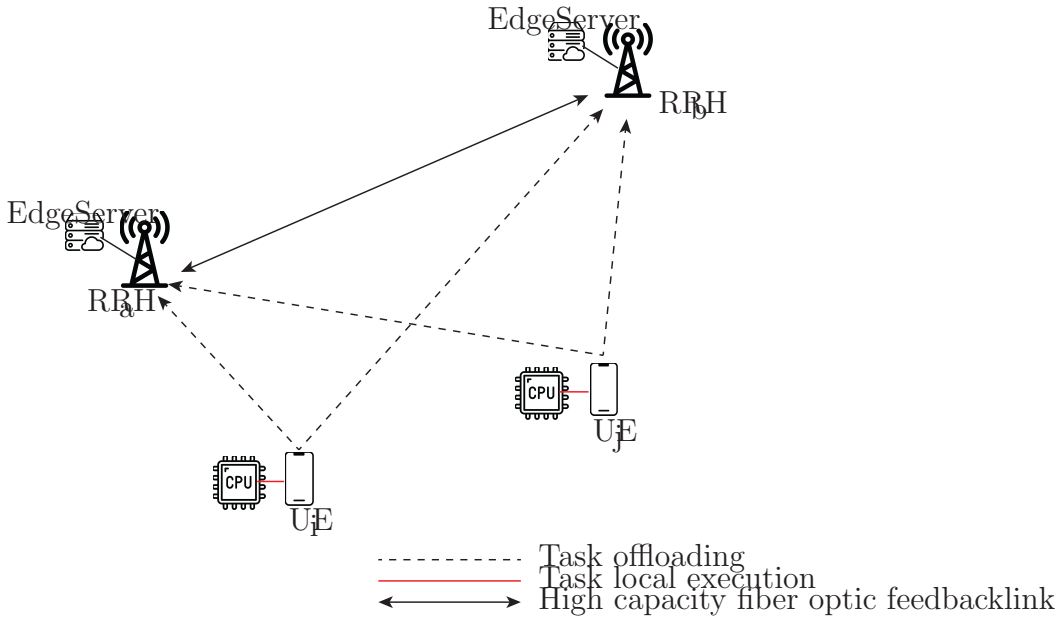

Figure 4: ή ά DU-NOMA MEC

It is important to note that the cost, in time and energy, for sending and receiving the bits are omitted and the energy cost for the computation in each RRH is considered minimal[33].

3.1 Protocol Implementation

The usage of the DU-NOMA protocol in this MEC scenario is facilitated by utilizing time-sharing between the possible communication environments that both users experience when they transmit simultaneously.More specifically, for each up-link transmission, the RRHs use all possible decoding orders of the two users performed by either RRH during the time-sharing phase duration T,while assuming an ideal feedback link between the remote heads and thus,perfect SIC in each sub-phase of the time-sharing duration[16].Each message is received by each RRH and it is either considered the first to be decoded and then used ,in SIC, to decode the other message by the same RRH or the other.By having 2 RRHs,the possible decoding orders ,considering the first user to be decoded as fixed, is 4 and by having 2 users and thus, 2 possible choices for the first message, all the possible scenarios amount to 8. For a given decoding order $z = (u, w), (v, x)$ such that message u is decoded first,while treating the message v as interference, from RRH_w and v is decoded second from RRH_x by performing SIC, the achievable rates for users i,j, for the k-th phase of the time sharing, are given as:

$$R_u^{(k)} = \log\left(1 + \frac{g_{uw} * p_u^{(k)}}{1 + g_{vw} * p_v^{(k)}}\right) \tag{1}$$

$$R_v^{(k)} = \log\left(1 + g_{vx} * p_v^{(k)}\right) \tag{2}$$

$$\forall (u, v) \in i, j, \forall (w, x) \in a, b$$

It is important to note that,for half of the duration T of the time-sharing scheme,one of the users' messages is always decoded first, regardless of the channel conditions or the constraints in energy or rate of the users.

Every communication scenario ,with its respective rates, is present for a percentage s_k of the duration of time-sharing,T and thus, the total achievable rate during T is given by:

$$R_{i,total} = \sum_{k=1}^{8} s_k * R_i^{(k)} \tag{3}$$

$$R_{j,total} = \sum_{k=1}^{8} s_k * R_j^{(k)} \tag{4}$$

When NOMA with time-sharing is employed, any point in the capacity region of the multiple access channel can be reached and the number of bits that can be offloaded by users i, j is constrained by the capacity region[25],[44].This is expressed as inequality constraints for each user as such:

$$T * B * R_{i,total} \geq \tilde{N}_i \tag{5}$$

$$T * B * R_{i,total} \geq \tilde{N}_j \tag{6}$$

where $\tilde{N}_i$,$\tilde{N}_j$ is the number of bits that each user can offload to the remote heads.The expressions for the energy consumption and delay for the cases of full and partial offloading are discussed below.

3.1.1 *Full Offloading*

In this scenario, it is assumed that the computation load of each user is processed solely by the edge server, at each RRH and as such, the individual delay incurred is solely imposed by the offloading delay and is given by:

$$T_{FO,i} = T_{o,i} = T \tag{7}$$

$$T_{FO,j} = T_{o,j} = T \tag{8}$$

Moreover,the energy consumed by each user id denoted by $E_{o,i}$,$E_{o,j}$ and is given by:

$$E_{o,i} = E_{FO,i} = T \cdot p_i = T \cdot \sum_{k=1}^{8} s_k \cdot p_i^{(k)} \tag{9}$$

$$E_{o,j} = E_{FO,j} = T \cdot p_j = T \cdot \sum_{k=1}^{8} s_k \cdot p_j^{(k)} \tag{10}$$

The energy consumption of each user must be limited and thus, the following inequalities hold:

$$E_{o,i} \leq E_i \tag{11}$$

$$E_{o,j} \leq E_j \tag{12}$$

Accordingly, the amount of bits that can be offloaded must be at least equal to the amount of bits that are needed for the computational task and thus:

$$\tilde{N}_i \geq N_i \tag{13}$$

$$\tilde{N}_j \geq N_j \tag{14}$$

3.1.2 *Partial offloading*

In this scenario,part of the computation load is offloaded to the edge server while the other is processed locally.As the local computation delay for the execution of the i-th user's sub-task is given by :

$$T_{loc,i} = \frac{L_i \cdot X_i}{f_i}$$

where L_i is the sub-task size in bits,X_i is the computation workload in CPU cycles per bit and f_j is the clock speed of the local CPU.It is assumed that clock speed can be adjusted dynamically such that $f_j \in (0, fmax]$, by using dynamic voltage scaling[ref] and that the transmission of signals between each edge server and user and the local processing can be done in parallel.Accordingly, the delay for completing the j-th user's task is given by :

$$T_{PO,j} = \max\left(T_{o,j}, T_{loc,j}\right)$$

since the total delay is determined by either the local processing or the offloading of said task.By setting $L_j = 0$, the full offloading scenario is covered since there are no bits to process locally.

Each CPU cycle consumes energy proportional to f_j^2 and can be expressed as $k_j \cdot f_j^2$, where k_j is a constant parameter related to the hardware architecture.For a sub-task of size L_j and workload X_j the local energy consumption can be derived by[8]

$$E_{loc,j} = k_j L_j X_j f_j^2$$

and subsequently, the total consumption is given by

$$E_{PO,u} = E_{o,j} + E_{loc,j}$$

.Finally, the number of bits that are either locally processed L_j or are successfully offloaded $\tilde{N_j}$ at the edge are bound by the capacity region of the multiple access channel and as such, the following inequality holds:

$$\tilde{N_j} + L_j \geq N_j$$

3.2 Full offloading Optimization

In this section, the case of full offloading using the proposed protocol is considered, taking into account the constraints in the achievable rate during the time-sharing phase, given by (5),(6) as well as those in energy consumption, given by (11),(12).Two different optimization objectives, the energy constrained minimization and the delay constrained minimization are discussed below.

3.2.1 Energy constrained delay minimization

The performance metric to be minimized is the total duration of the time-sharing phase for the full offloading of each users' bits, $T_{FO,i} = T_{FO,j} = T$.Note that both users must offload all the computation workload during the same period T.To that end, we formulate the optimization problem below:

$$\begin{aligned} \underset{\mathbf{T},\mathbf{p},\tilde{\mathbf{N}},\mathbf{s}}{\mathbf{min}} \quad & T \\ \mathbf{s.t.} \quad & \mathrm{C}_1 : TBR_{i,total} \geq \tilde{N}_i, \\ & \mathrm{C}_2 : TBR_{i,total} \geq \tilde{N}_j, \\ & \mathrm{C}_3 : E_{o,i} \leq E_i, \\ & \mathrm{C}_4 : E_{o,j} \leq E_j, \\ & \mathrm{C}_5 : s_k, p_j^{(k)}, p_i^{(k)} \geq 0, \\ & \mathrm{C}_6 : s_k \in [0,1], \forall k = 1,..,8 \end{aligned} \tag{15}$$

where $\mathbf{T}, \mathbf{p}, \tilde{\mathbf{N}}, \mathbf{s}$ denote the vectors containing the variables $T_{FO,i}, T_{FO,j}, p_i, p_j, \tilde{N}_i, \tilde{N}_j$ and s_K respectively.Constraints C_1, C_2 represent the bound on the number of bits that must be offloaded while C_3, C_4 are associated with the maximum available energy for the process.By setting $z_k = s_k \cdot T$ as a new variable that represents the time interval that a specific decoding order is used, the problem can be formulated as:

$$\begin{aligned} \underset{\mathbf{p},\mathbf{z}}{\mathbf{min}} \quad & \textstyle\sum_{k=1}^{8} z_k \\ \mathbf{s.t.} \quad & \mathrm{C}_1 : z_1 BR_i^{(1)} + z_2 BR_i^{(2)} + \ldots + z_7 BR_i^{(7)} + z_8 BR_i^{(8)} \geq \tilde{N}_i, \\ & \mathrm{C}_2 : z_1 BR_j^{(1)} + z_2 BR_j^{(2)} + \ldots + z_7 BR_j^{(7)} + z_8 BR_j^{(8)} \geq \tilde{N}_j, \\ & \mathrm{C}_3 : z_1 p_i^{(1)} + \ldots + z_8 p_i^{(8)} \leq \tfrac{E_i}{BN_0}, \\ & \mathrm{C}_4 : z_1 p_j^{(1)} + \ldots + z_8 p_j^{(8)} \leq \tfrac{E_i}{BN_0}, \\ & \mathrm{C}_5 : z_k, p_j^{(k)}, p_i^{(k)} \geq 0, \forall k = 1,..,8 \end{aligned} \tag{16}$$

where constraints C_1, C_2 where expanded using (3),(4) and C_3, C_4 using (9) and (10) and the objective is rewritten as $T = \sum_{k=1}^{8} z_k$.Due to the coupling of the time intervals z_k and the normalized powers of each decoding order, $p_i^{(k)}, p_j^{(k)}$ in C_1, C_2 and the appearance of interference terms $p_v^{(k)}, \forall v \in \{i, j\}$, in the expressions of R_i or R_j, the problem, as formulated, is non-convex.We introduce slack variables $r_i^{(k)}, r_j^{(k)}$ such that:

$$r_i^{(k)} \leq R_i^{(k)} \tag{17}$$

$$r_j^{(k)} \leq R_j^{(k)} \tag{18}$$

where $R_i^{(k)}, R_J^{(k)}$ are given by (1),(2) for the k-th decoding order.For illustrative purposes, we consider the decoding order for $k = 1$ through 4, in which the message of user i is always decoded first by RRH_w, treating the message of user j as interference ,at the same RRH_w, while the message of user j is decoded using SIC ,interference-free at RRH_v for every choice of $w, v \in \{a, b\}$.Therefore, the inequalities (17),(18) considering the above and relations (1),(2) can be written as:

$$r_i^{(k)} \leq \log\left(1 + \frac{g_{iw} * p_i^{(k)}}{1 + g_{jw} * p_j^{(k)}}\right) \tag{19}$$

$$r_j^{(k)} \leq \log\left(1 + g_{jx} * p_j^{(k)}\right) \tag{20}$$

$$(w, x) \in \{a, b\}, \forall k = 1, .., 4$$

After some mathematical manipulations the inequalities are written as:

$$\log\left(\frac{1}{g_{iw} p_i^{(k)}} + \frac{g_{jw} p_j^{(k)}}{g_{iw} p_i^{(k)}}\right) + \log\left(2^{r_i^{(k)}} - 1\right) \leq 0 \tag{21}$$

$$\log\left(2^{r_j^{(k)}} - 1\right) - p_j^{(k)} - \log\left(g_{jx}\right) \leq 0 \tag{22}$$

The same inequalities hold for $k = 5, .., 8$ with indices i, j reversed.The coupled terms $z_k p^{(k)}$ in constraints C3,C4 as well as the terms $\frac{p_j^{(k)}}{p_i^{(k)}}$ are monomials of the optimization variables[Boyd,geometric ref] and as such, motivate the use of geometric transformations of the variables, of the form $x = e^{\tilde{x}}$ on the elements of $\mathbf{p}, \mathbf{z}$ and the vector of slack variables $\mathbf{r}$.After taking the logarithm of the objective and the constraints, the problem is formulated as follows:

$$\begin{aligned}
\underset{\mathbf{p},\mathbf{z},\mathbf{r}}{\mathbf{min}} \quad & \log \textstyle\sum_{k=1}^{8} e^{\tilde{z_k}} \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \log\left(\tfrac{\tilde{N}_i}{B}\right) - \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{r_i}^{(k)}}\right) \leq 0 \\
& \mathrm{C}_2 : \log\left(\tfrac{\tilde{N}_j}{B}\right) - \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{r_j}^{(k)}}\right) \leq 0 \\
& \mathrm{C}_3 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_i}^{(k)}}\right) - \log\left(\tfrac{E_i}{BN_0}\right) \leq 0, \\
& \mathrm{C}_4 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_j}^{(k)}}\right) - \log\left(\tfrac{E_j}{BN_0}\right) \leq 0, \\
& \mathrm{C}_{(5+k)} : \log\left(\tfrac{1}{g_{ik}} e^{-\tilde{p_i}^{(k)}} + \tfrac{g_{jk}}{g_{ik}} e^{\tilde{p_j}^{(k)} - \tilde{p_i}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_i}^{(k)}}} - 1\right) \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(2^{e^{\tilde{p_j}^{(k)}}} - 1\right) - \tilde{p_j}^{(k)} - \log g_{jk} \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(\tfrac{1}{g_{jk}} e^{-\tilde{p_j}^{(k)}} + \tfrac{g_{ik}}{g_{jk}} e^{\tilde{p_i}^{(k)} - \tilde{p_j}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_j}^{(k)}}} - 1\right) \leq 0, \forall k = 5, .., 8 \\
& \mathrm{C}_{(12+k)} : \log\left(2^{e^{\tilde{p_i}^{(k)}}} - 1\right) - \tilde{p_i}^{(k)} - \log g_{ik} \leq 0, \forall k = 5, .., 8
\end{aligned} \tag{23}$$

It is observed that constraints $C3$ and $C4$ are now convex as the sum of a log-sum-exp function of an affine function of the optimization variables[Boyd] and a constant.Moreover, the first terms in $C_{(5+k)}, \forall k = 1, ..4$ and in $C_{(9+k)}, \forall k = 5, ..8$ are log-sum-exp of affine functions (the gain terms can be transformed as $x = e^{\log(x)}$, contributing a constant) of the variables and the second terms i.e $f(r) = \ln(2^{e^r} - 1)$ are convex.This is easily proved, considering that $\frac{\partial^2 f}{\partial r^2} = \frac{2^z z \ln(2)(2^z - z - 1)}{(2^z - 1)^2}$ and $\frac{\partial^2 f}{\partial r^2} \geq 0$,with $z = e^r$.The term $2^{e^r} e^r \ln(2)$ is always positive,$\forall r \in \mathbb{R}$,and as $e^r > 0$,it follows that $2^{e^r} - 1 > 0, \forall r \in \mathbb{R}$.Also, the term $w = (2^z - z - 1)$ is an increasing function of z and as $z \to 0, w \to 0$.Accordingly,constraints $C_{(9+k)}, \forall k = 1, ..4$ and $C_{(12+k)}, \forall k = 5, ..8$ are convex as the sum of a convex term $f(p) = \ln(2^{e^p} - 1)$, for $p \in \{p_i^{(k)}, p_i^{(k)}\}$ and an affine function of $p_i^{(k)}$ or $p_i^{(k)}$. Constraints C_1, C_2 are recognized as difference of convex(DC), since the terms $\log(N/B), \forall N \in \{N_i, N_j\}$ are constants and thus, convex functions and the terms $\log(\sum_{k=1}^{8}(e^{\tilde{z}_k + \tilde{r}})), \forall \tilde{r} \in \tilde{r}_i^{(k)}, \tilde{r}_j^{(k)}$ are convex as log-sum-exp functions of affine functions of the variables[Boyd].By exploiting the DC structure of the constraints, we can use successive convex approximation procedure[DC papers] by replacing the second term in each expression with its first order multi-variable Taylor approximation.Thus, we define $\mathbf{x_l} = [\mathbf{z}, \mathbf{r_l}], \forall l \in \{i, j\}$ as the concatenated vector of vectors $\mathbf{z}$ and $\mathbf{r_l}$ or $\mathbf{r_l}$, respectively.The second terms can then be written as:

$$h_l(\mathbf{x_l}) = \log(\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{r_l}^{(k)}}), \forall l \in \{i, j\}$$

and the approximation ,around the point $\mathbf{x_{l_0}}$ is given by:

$$h_l(\mathbf{x}) \approx h_l(\mathbf{x_{l_0}}) + \nabla h_l(\mathbf{x_{l_0}})(\mathbf{x_l} - \mathbf{x_{l_0}}), \forall l \in \{i, j\} \tag{24}$$

where:

$$\begin{aligned} \nabla h_l(\mathbf{x_{l_o}}) &= [\mathbf{e_z}, \mathbf{e_{rl}}], \forall l \in \{i.j\} \\ \mathbf{s.t}\ \mathbf{e_z}(k) &= \frac{\partial h}{\partial z_k}, \\ \mathbf{e_{rl}}(k) &= \frac{\partial h_l}{\partial r_{l_k}}, \forall k = 1, ...8 \\ \frac{\partial h_l}{\partial zk} = \frac{\partial h_l}{\partial r_{lk}} &= \frac{e^{\tilde{z_k} + \tilde{r_l}^{(k)}}}{\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{r_l}^{(k)}}} \end{aligned}$$

The vectors $\mathbf{e_z}, \mathbf{e_{rl}}$ contain the partial derivatives with respect to z_k and $r_{lk}, \forall l \in \{i, j\}$.Replacing the approximants in constraints C_1 and C_2, the initial non-convex problem is replaced with a convex approximation of the form:

$$
\begin{aligned}
\underset{\mathbf{p},\mathbf{x_i},\mathbf{x_j}}{\mathbf{min}} \quad & \log \textstyle\sum_{k=1}^{8} e^{\tilde{z_k}} \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \log\left(\tfrac{\tilde{N_i}}{B}\right) - h_i(\mathbf{x_{i_0}}) - \nabla h_i(\mathbf{x_{i_0}})(\mathbf{x_i} - \mathbf{x_{i_0}}) \leq 0 \\
& \mathrm{C}_2 : \log\left(\tfrac{\tilde{N_j}}{B}\right) - h_i(\mathbf{x_{j_0}}) - \nabla h_i(\mathbf{x_{j_0}})(\mathbf{x_j} - \mathbf{x_{j_0}}) \leq 0 \\
& \mathrm{C}_3 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_i}^{(k)}}\right) - \log\left(\tfrac{E_i}{BN_0}\right) \leq 0, \\
& \mathrm{C}_4 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_j}^{(k)}}\right) - \log\left(\tfrac{E_j}{BN_0}\right) \leq 0, \\
& \mathrm{C}_{(5+k)} : \log\left(\tfrac{1}{g_{ik}} e^{-\tilde{p_i}^{(k)}} + \tfrac{g_{jk}}{g_{ik}} e^{\tilde{p_j}^{(k)} - \tilde{p_i}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_i}^{(k)}}} - 1\right) \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(2^{e^{\tilde{p_j}^{(k)}}} - 1\right) - \tilde{p_j}^{(k)} - \log g_{jk} \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(\tfrac{1}{g_{jk}} e^{-\tilde{p_j}^{(k)}} + \tfrac{g_{ik}}{g_{jk}} e^{\tilde{p_i}^{(k)} - \tilde{p_j}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_j}^{(k)}}} - 1\right) \leq 0, \forall k = 5, .., 8 \\
& \mathrm{C}_{(12+k)} : \log\left(2^{e^{\tilde{p_i}^{(k)}}} - 1\right) - \tilde{p_i}^{(k)} - \log g_{ik} \leq 0, \forall k = 5, .., 8
\end{aligned}
\tag{25}
$$

In this case, a simple algorithm is developed that approximates the solution of the initial non-convex problem by a sequence of convex optimization problems, that can be solved , at each iteration, by the interior point method,in polynomial time[46,47].Moreover,the interior of the "while" loop has a linear convergence rate[46] and the algorithm can be realistically implemented in a practical MEC scenario, on an edge server of adequate computational capabilities.

Algorithm 1 Solution of DC optimization problem (25)

Initialize: $A, e, \tilde{\mathbf{z}}, \tilde{\mathbf{r}}$
  while $A > e$ **do**
Solve: optimization problem (9) for $\mathbf{z}^*, \mathbf{r}^*$
  $A = \|[\tilde{\mathbf{z}}, \tilde{\mathbf{r}}] - [\mathbf{z}^*, \mathbf{r}^*]\|_2^2$
    $\tilde{\mathbf{z}} \leftarrow \mathbf{z}^*$
    $\tilde{\mathbf{r}} \leftarrow \mathbf{r}^*$
  end while
Output: $\tilde{\mathbf{T}}^*, \tilde{\mathbf{z}}^*, \tilde{\mathbf{p}}^*$

3.2.2 Delay constrained energy minimization

The above analysis can be easily extended to the case of energy minimization, taking into account the rate constraints in (5), (6) and adding a maximum

delay constraint such that $T \leq T_{max}$.The metric to be used is the weighted sum of users' energy consumption, defined as

$$E_{FO} = \sum_{l \in \{i,j\}} w_l \cdot E_{FO,l}$$

while $w_l \in [0,1]$ is a positive constant that is used for the assignment of different priorities to each user and ensure certain notions of "fairness"[fairness paper].By setting $z_k = s_k T$ and $E_{FO,i} = \sum_{k=1}^{8} p_i^{(k)} z_k$,$E_{FO,j} = \sum_{k=1}^{8} p_j^{(k)} z_k$, the minimization problem can be formulated as:

$$\begin{aligned}
\underset{\mathbf{p},\mathbf{z}}{\mathbf{min}} \quad & w_i \textstyle\sum_{k=1}^{8} p_i^{(k)} z_k + w_j \sum_{k=1}^{8} p_j^{(k)} z_k \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : z_1 BR_i^{(1)} + z_2 BR_i^{(2)} + \ldots + z_7 BR_i^{(7)} + z_8 BR_i^{(8)} \geq \tilde{N}_i, \\
& \mathrm{C}_2 : z_1 BR_j^{(1)} + z_2 BR_j^{(2)} + \ldots + z_7 BR_j^{(7)} + z_8 BR_j^{(8)} \geq \tilde{N}_j, \\
& \mathrm{C}_3 : \\
& \mathrm{C}_5 : z_k, p_j^{(k)}, p_i^{(k)} \geq 0,
\end{aligned} \tag{26}$$

Similarly with the delay minimization solution, to remove the coupling between variables we introduce slack variables

$$r_i^{(k)} \leq R_i^{(k)}, r_j^{(k)} \leq R_j^{(k)}$$

, we apply the geometric programming method[Boyd paper] and because C_1 and C_2 are DC functions , we replace them with their first order Taylor approximations.Note that constraint C_3 is relaxed,so that after the geometric transformation,the constraint remains convex[Boyd].The problem is thus, transformed to:

$$\begin{aligned}
\underset{\mathbf{p},\mathbf{x_i},\mathbf{x_j}}{\mathbf{min}} \quad & \log\left(w_i \textstyle\sum_{k=1}^{8} e^{\tilde{p}_i^{(k)} + \tilde{z}_k} + w_j \sum_{k=1}^{8} e^{\tilde{p}_j^{(k)} + \tilde{z}_k}\right) \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \log\left(\tfrac{\tilde{N}_i}{B}\right) - h_i(\mathbf{x_{i_0}}) - \nabla h_i(\mathbf{x_{i_0}})(\mathbf{x_i} - \mathbf{x_{i_0}}) \leq 0 \\
& \mathrm{C}_2 : \log\left(\tfrac{\tilde{N}_j}{B}\right) - h_i(\mathbf{x_{j_0}}) - \nabla h_i(\mathbf{x_{j_0}})(\mathbf{x_j} - \mathbf{x_{j_0}}) \leq 0 \\
& \mathrm{C}_3 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_i}^{(k)}}\right) - \log\left(\tfrac{E_i}{BN_0}\right) \leq 0, \\
& \mathrm{C}_4 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k} + \tilde{p_j}^{(k)}}\right) - \log\left(\tfrac{E_j}{BN_0}\right) \leq 0, \\
& \mathrm{C}_{(5+k)} : \log\left(\tfrac{1}{g_{ik}} e^{-\tilde{p_i}^{(k)}} + \tfrac{g_{jk}}{g_{ik}} e^{\tilde{p_j}^{(k)} - \tilde{p_i}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_i}^{(k)}}} - 1\right) \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(2^{e^{\tilde{p_j}^{(k)}}} - 1\right) - \tilde{p_j}^{(k)} - \log g_{jk} \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(9+k)} : \log\left(\tfrac{1}{g_{jk}} e^{-\tilde{p_j}^{(k)}} + \tfrac{g_{ik}}{g_{jk}} e^{\tilde{p_i}^{(k)} - \tilde{p_j}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_j}^{(k)}}} - 1\right) \leq 0, \forall k = 5, .., 8 \\
& \mathrm{C}_{(12+k)} : \log\left(2^{e^{\tilde{p_i}^{(k)}}} - 1\right) - \tilde{p_i}^{(k)} - \log g_{ik} \leq 0, \forall k = 5, .., 8
\end{aligned} \tag{27}$$

The above formulation can be used in an iterative algorithm similar to Algorithm 1,to converge to the global solution of the non-convex problem (26).

3.3 Partial Offloading

In this section, the partial offloading case is examined while taking into consideration tha latency and energy constraints outlined in the description of the system model.This case proposes several benefits due to the joint optimization of computation and communication resources.The controllabillity of the CPU clock speed can ensure that latency is reduced in the local computation while reducing the time that interference is experienced by the users, but it also increases the energy consumption of the edge devices.The metric to be minimized is the total delay of both users and the optimization problem formulated below can easily be extended to the energy minimization of similar schemes:

$$
\begin{aligned}
\underset{\mathbf{T},\mathbf{p},\tilde{\mathbf{N}},\mathbf{s}}{\mathbf{min}} \quad & T_{PO} \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : T_{PO,i} BR_{i,total} \geq \tilde{N}_i, \\
& \mathrm{C}_2 : T_{PO,j} BR_{j,total} \geq \tilde{N}_j, \\
& \mathrm{C}_3 : \tilde{N}_i + L_i \geq N_i, \\
& \mathrm{C}_4 : \tilde{N}_j + L_j \geq N_j, \\
& \mathrm{C}_5 : E_{PO,i} \leq E_i, \\
& \mathrm{C}_6 : E_{PO,j} \leq E_j, \\
& \mathrm{C}_7 : f_l \leq f_{max}, \forall l \in \{i, j\} \\
& \mathrm{C}_5 : T_{PO,i}, T_{PO,j}, \tilde{N}_i, \tilde{N}_j, p_j^{(k)}, p_i^{(k)} \geq 0,
\end{aligned}
\tag{28}
$$

To handle the non-convex objective function we can write the optimization problem in epigraph form , while also writing the constraints in terms of the time interval of each decoding order $z_k = s_k T_{FO,i} = s_k T$ and $T = \sum_{k=1}^{8} z_k$.Note that inequalities C_1 and C_3 ,as well as C_2 and C_4 , can be merged as they preserve the order between variables.Thus,

$$
\begin{aligned}
\underset{\mathbf{z},\mathbf{p},\mathbf{L},\mathbf{f},y}{\mathbf{min}} \quad & y \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \textstyle\sum_{k=1}^{8} z_k R_i^{(k)} + \frac{L_i}{B} \geq \frac{N_i}{B}, \\
& \mathrm{C}_2 : \textstyle\sum_{k=1}^{8} z_k R_j^{(k)} + \frac{L_j}{B} \geq \frac{N_j}{B}, \\
& \mathrm{C}_3 : \textstyle\sum_{k=1}^{8} z_k p_i^{(k)} + k_i X_i L_i f_i^2 \leq E_i, \\
& \mathrm{C}_4 : \textstyle\sum_{k=1}^{8} z_k p_j^{(k)} + k_j X_j L_j f_j^2 \leq E_j, \\
& \mathrm{C}_5 : \textstyle\sum_{k=1}^{8} z_k \leq y, \\
& \mathrm{C}_6 : \frac{L_i X_i}{f_i} \leq y, \\
& \mathrm{C}_7 : \frac{L_j X_j}{f_j} \leq y, \\
& \mathrm{C}_8 : f_i \leq f_{max}, \\
& \mathrm{C}_9 : f_j \leq f_{max}, \\
& \mathrm{C}_{10} : L_i, L_j, f_i, f_j, p_j^{(k)}, p_i^{(k)}, z_k \geq 0, \forall k = 1, \ldots, 8
\end{aligned}
\tag{29}
$$

where $\mathbf{L}, \mathbf{f}$ denote the vectors corresponding to variables L_l, $f_l \, \forall l \in \{i, j\}$.Subsequently, we follow the same steps as the previous minimization problems , by introducing slack variables $r_l^{(k)} \leq R_l^{(k)} \forall l \in \{i.j\}$ and using the geometric transform $x = e^{\tilde{x}}$ on every element of the optimization vectors $\mathbf{z}, \mathbf{r}, , \mathbf{p}, \mathbf{L}, \mathbf{f}$ and the epigraph variable y.By taking the logarithm of the objective and the constraints,the problem ,after some mathematical manipulations is given by:

$$
\begin{aligned}
\underset{\tilde{\mathbf{z}},\tilde{\mathbf{p}},\tilde{\mathbf{L}},\tilde{\mathbf{f}},\tilde{y}}{\mathbf{min}} \quad & \tilde{y} \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \log\left(\tfrac{N_i}{B}\right) - \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z}_k + \tilde{r}_i^{(k)}} + \tfrac{1}{B} e^{\tilde{L}_i}\right) \leq 0, \\
& \mathrm{C}_2 : \log\left(\tfrac{N_j}{B}\right) - \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z}_k + \tilde{r}_j^{(k)}} + \tfrac{1}{B} e^{\tilde{L}_j}\right) \leq 0, \\
& \mathrm{C}_3 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z}_k + \tilde{p}_i^{(k)}} + k_i X_i e^{\tilde{L}_i + 2\tilde{f}_i}\right) - \log\left(E_i\right) \leq 0, \\
& \mathrm{C}_4 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z}_k + \tilde{p}_j^{(k)}} + k_i X_i e^{\tilde{L}_j + 2\tilde{f}_j}\right) - \log\left(E_j\right) \leq 0, \\
& \mathrm{C}_5 : \log\left(\textstyle\sum_{k=1}^{8} e^{z_k}\right) - \tilde{y} \leq 0, \\
& \mathrm{C}_6 : \tilde{L}_i - \tilde{f}_i + \log X_i - \tilde{y} \leq 0, \\
& \mathrm{C}_7 : \tilde{L}_i - \tilde{f}_i + \log X_i - \tilde{y} \leq 0, \\
& \mathrm{C}_8 : \tilde{f}_i - \log\left(f_{max}\right) \leq 0, \\
& \mathrm{C}_9 : \tilde{f}_j - \log\left(f_{max}\right) \leq 0, \\
& \mathrm{C}_{(10+k)} : \log\left(\tfrac{1}{g_{ik}} e^{-\tilde{p_i}^{(k)}} + \tfrac{g_{jk}}{g_{ik}} e^{\tilde{p_j}^{(k)} - \tilde{p_i}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_i}^{(k)}}} - 1\right) \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(14+k)} : \log\left(2^{e^{\tilde{p_j}^{(k)}}} - 1\right) - \tilde{p_j}^{(k)} - \log g_{jk} \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(14+k)} : \log\left(\tfrac{1}{g_{jk}} e^{-\tilde{p_j}^{(k)}} + \tfrac{g_{ik}}{g_{jk}} e^{\tilde{p_i}^{(k)} - \tilde{p_j}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_j}^{(k)}}} - 1\right) \leq 0, \forall k = 5, .., 8 \\
& \mathrm{C}_{(17+k)} : \log\left(2^{e^{\tilde{p_i}^{(k)}}} - 1\right) - \tilde{p_i}^{(k)} - \log g_{ik} \leq 0, \forall k = 5, .., 8
\end{aligned}
\tag{30}
$$

In this formulation,C_3, C_4, C_5 are convex, as the sum of a constant and the composition of log-sum-exp functions with affine combinations of the optimization variables[Boyd],C_7, C_7, C_8, C_9 are linear with respect to the variables and thus, convex and the rest of the constraints are identical to those in problems (25),(27) of the above sections, and are proven to be convex.Although the objective function $\tilde{y}$ is convex, constraints C_1, C_2 are concave and thus, the problem is still non-convex.By observing that those constraints are Difference of Convex(DC) functions,as the difference between a constant function and a composition of a log-sum-exp function with affine combinations of the optimization variables, the full problem can be replaced with a convex approximation by changing the log-sum-exp function to its Taylor approxi-

mation as is shown below:

$$g_l(\mathbf{x_l}) = \log\left(\sum_{k=1}^{8} e^{\tilde{z_k}+\tilde{r_l}^{(k)}}\right), \mathbf{x_l} = [\tilde{\mathbf{z}}, \tilde{\mathbf{r}}_\mathbf{l}, \tilde{L}_l] \forall l \in \{i, j\}$$

and the approximation ,around the point $\mathbf{x_{l_0}}$ is given by:

$$g_l(\mathbf{x}) \approx h_l(\mathbf{x_{l_0}}) + \nabla g_l(\mathbf{x_{l_0}})(\mathbf{x_l} - \mathbf{x_{l_0}}), \forall l \in \{i, j\} \tag{31}$$

where:

$$\begin{aligned}
\nabla g_l(\mathbf{x_{l_o}}) &= [\mathbf{e_z}, \mathbf{e_{rl}}, \frac{\partial g}{\partial L_{lk}}], \forall l \in \{i.j\} \\
\mathbf{s.t}\ \mathbf{e_z}(k) &= \frac{\partial g}{\partial z_k}, \\
\mathbf{e_{rl}}(k) &= \frac{\partial g}{\partial r_{l_k}}, \forall k = 1, ...8 \\
\frac{\partial g}{\partial zk} = \frac{\partial g}{\partial r_{lk}} &= \frac{e^{\tilde{z_k}+\tilde{r_l}^{(k)}}}{\sum_{k=1}^{8} e^{\tilde{z_k}+\tilde{r_l}^{(k)}} + \frac{1}{B}e^{\tilde{L}_j}} \\
\frac{\partial g}{\partial L_l} &= \frac{\frac{1}{B}e^{\tilde{L}_l}}{\sum_{k=1}^{8} e^{\tilde{z_k}+\tilde{r_l}^{(k)}} + \frac{1}{B}e^{\tilde{L}_l}}
\end{aligned}$$

The full problem can be written as:

$$\begin{aligned}
\underset{\tilde{\mathbf{z}},\tilde{\mathbf{p}},\tilde{\mathbf{L}},\tilde{\mathbf{f}},\tilde{y}}{\mathbf{min}} \quad & \tilde{y} \\
\mathbf{s.t.} \quad & \mathrm{C}_1 : \log\left(\tfrac{\tilde{N}_i}{B}\right) - g_i(\mathbf{x_{i_0}}) - \nabla g_i(\mathbf{x_{i_0}})(\mathbf{x_i} - \mathbf{x_{i_0}}) \leq 0 \\
& \mathrm{C}_2 : \log\left(\tfrac{\tilde{N}_j}{B}\right) - g_i(\mathbf{x_{j_0}}) - \nabla g_i(\mathbf{x_{j_0}})(\mathbf{x_j} - \mathbf{x_{j_0}}) \leq 0 \\
& \mathrm{C}_3 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k}+\tilde{p_i}^{(k)}} + k_i X_i e^{\tilde{L}_i + 2\tilde{f}_i}\right) - \log\left(E_i\right) \leq 0, \\
& \mathrm{C}_4 : \log\left(\textstyle\sum_{k=1}^{8} e^{\tilde{z_k}+\tilde{p_j}^{(k)}} + k_i X_i e^{\tilde{L}_j + 2\tilde{f}_j}\right) - \log\left(E_j\right) \leq 0, \\
& \mathrm{C}_5 : \log\left(\textstyle\sum_{k=1}^{8} e^{z_k}\right) - \tilde{y} \leq 0, \\
& \mathrm{C}_6 : \tilde{L}_i - \tilde{f}_i + \log X_i - \tilde{y} \leq 0, \\
& \mathrm{C}_7 : \tilde{L}_i - \tilde{f}_i + \log X_i - \tilde{y} \leq 0, \\
& \mathrm{C}_8 : \tilde{f}_i - \log\left(f_{max}\right) \leq 0, \\
& \mathrm{C}_9 : \tilde{f}_j - \log\left(f_{max}\right) \leq 0, \\
& \mathrm{C}_{(10+k)} : \log\left(\tfrac{1}{g_{ik}} e^{-\tilde{p_i}^{(k)}} + \tfrac{g_{jk}}{g_{ik}} e^{\tilde{p_j}^{(k)} - \tilde{p_i}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_i}^{(k)}}} - 1\right) \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(14+k)} : \log\left(2^{e^{\tilde{p_j}^{(k)}}} - 1\right) - \tilde{p_j}^{(k)} - \log g_{jk} \leq 0, \forall k = 1, .., 4 \\
& \mathrm{C}_{(14+k)} : \log\left(\tfrac{1}{g_{jk}} e^{-\tilde{p_j}^{(k)}} + \tfrac{g_{ik}}{g_{jk}} e^{\tilde{p_i}^{(k)} - \tilde{p_j}^{(k)}}\right) + \log\left(2^{e^{\tilde{r_j}^{(k)}}} - 1\right) \leq 0, \forall k = 5, .., 8 \\
& \mathrm{C}_{(17+k)} : \log\left(2^{e^{\tilde{p_i}^{(k)}}} - 1\right) - \tilde{p_i}^{(k)} - \log g_{ik} \leq 0, \forall k = 5, .., 8
\end{aligned} \tag{32}$$

This problem can be solved effectively ,in the MEC server, by iteratively solving a convex problem with different starting points $\tilde{x}_{i_o}, x_{j_o}$ until suitable convergence is reached.The algorithm describing this procedure is given below:

Algorithm 2 Solution of DC optimization problem (32)

Initialize: $A, e, \tilde{\mathbf{z}}, \tilde{\mathbf{r}}$
  while $A > e$ **do**
Solve: optimization problem (9) for $\mathbf{z}^*, \mathbf{r}^*, \mathbf{L}^*$
  $\tilde{\mathbf{x}} = [\tilde{\mathbf{z}}, \tilde{\mathbf{r}}, \mathbf{L}]$
  $A = \|\tilde{\mathbf{x}} - \mathbf{x}^*\|_2^2$
    $\tilde{\mathbf{z}} \leftarrow \mathbf{z}^*$
    $\tilde{\mathbf{r}} \leftarrow \mathbf{r}^*$
  end while
Output: $\tilde{\mathbf{z}^*}, \tilde{\mathbf{p}^*}, \tilde{\mathbf{L}^*}, \tilde{\mathbf{f}^*}, \tilde{\mathbf{y}^*}$

3.4 Simulation Results

For the simulations, we assume that $B = 1$ MHz, the number of bits for each computational task is $N_i = N_j = 0.5$ Mbits, and $E_i = E_j$ for each user. The parameters that vary in the simulation are the SNR(db) and the metric $K = \frac{\frac{E}{BN_o}}{p}$ where p is the maximum SNR used. This metric is measured in seconds and can be interpreted as the time the system can operate at a given maximum SNR (denoted as p), with the available energy E.

4 Benchmark Description

As a benchmark measure for the full offloading problems, a simpler solution methodology is used based on the following observation: for the first four operation scenarios(decoding orders) of the system, user i is decoded first, and for the remaining four, user j is decoded first. That is, during time T of time sharing, one of the user's messages is always decoded first, regardless of channel conditions. According to this logic, and using a parameter c_1, where $c_1 \in [0, 1]$, it can be formulated that for user i, for the first 4 scenarios where it is decoded first, the transmission power and the corresponding SNR are $c_1 P_{max}$, $c_1 p_{max}$, and for the remaining 4 scenarios where it is decoded second, using another parameter c_2, with $c_2 \in [0, 1]$, it is $c_2 P_{max}$ and $c_2 p_{max}$.

With the exactly corresponding logic, for user j, for the first 4 scenarios where it is decoded second, the transmission power and the corresponding SNR are $c_2 P_{max}$, $c_2 p_{max}$, and for the remaining 4 where it is decoded first, it is $c_1 P_{max}$ and $c_1 p_{max}$. The problem is formulated in terms of the variables z_k and the parameters c_1, c_2, and is solved by selecting the best delay values $\sum_{k=1}^{8} z_k$ for $c \in [0, 1]$ with a discretization step of 0.1.Given a fixed value for the parameters c, the problem reduces to a linear program with respect to the variables z_k:

$$
\begin{aligned}
\underset{\mathbf{z}}{\textbf{min}} \quad & \textstyle\sum_{k=1}^{8} z_k \\
\textbf{s.t.} \quad & \mathrm{C}_1 : z_1 BR_i^{(1)} + \ldots + z_4 BR_i^{(4)} + \ldots + z_5 BR_i^{(7)} + \ldots + z_8 BR_i^{(8)} \geq \tilde{N}_i, \\
& \mathrm{C}_2 : z_1 BR_j^{(1)} + \ldots + z_4 BR_j^{(4)} + \ldots + z_5 BR_j^{(5)} + \ldots + z_8 BR_j^{(8)} \geq \tilde{N}_j, \\
& \mathrm{C}_3 : z_1 c_1 p_{max} + \ldots + z_4 c_1 p_{max} + \ldots + z_5 c_2 p_{max} + \ldots + z_8 c_2 p_{max} \leq \tfrac{E_i}{BN_0}, \\
& \mathrm{C}_4 : z_1 c_2 p_{max} + \ldots + z_4 c_2 p_{max} + \ldots + z_5 c_1 p_{max} + \ldots + z_8 c_1 p_{max} \leq \tfrac{E_j}{BN_0}, \\
& \mathrm{C}_5 : z_k, c_1, c_2 \geq 0,
\end{aligned}
\tag{33}
$$

Following a similar logic, the problem of delay-constrained energy minimization, is formulated as:

$$
\begin{aligned}
\min_{\mathbf{z}} \quad & (c_1 * w_i + c_2 * w_j) * pmax * \sum_{k=1}^{4} z_k + (c_2 * w_i + c_1 * w_j) * pmax * \sum_{k=5}^{8} z_k \\
\text{s.t.} \quad & \mathrm{C}_1 : z_1 BR_i^{(1)} + \ldots + z_4 BR_i^{(4)} + \ldots + z_5 BR_i^{(7)} + \ldots + z_8 BR_i^{(8)} \geq \tilde{N}_i, \\
& \mathrm{C}_2 : z_1 BR_j^{(1)} + \ldots + z_4 BR_j^{(4)} + \ldots + z_5 BR_j^{(5)} + \ldots + z_8 BR_j^{(8)} \geq \tilde{N}_j, \\
& \mathrm{C}_3 : \sum_{k=1}^{8} z_k \leq Tmax
\end{aligned}
\tag{34}
$$

5 Simulations for delay minimization

Impact of Maximum SNR

For given values of K in the interval $[0.2, 1]$ and $SNR = 10, 20, 30$, we have:

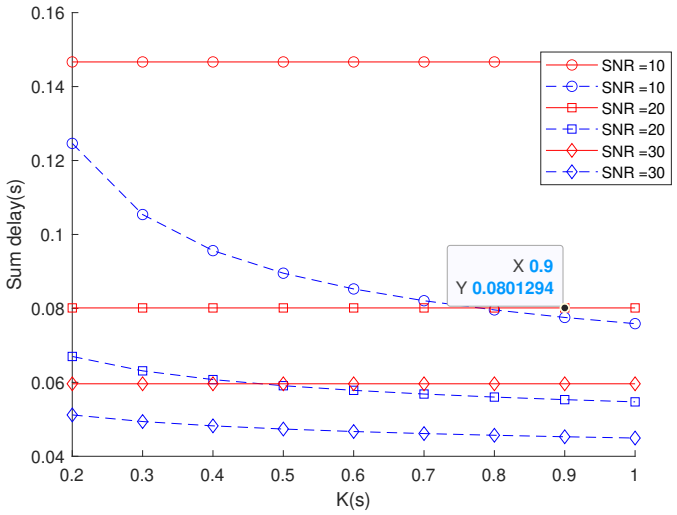

Figure 5: Variation of K

and for K in the interval $[1, 12]$

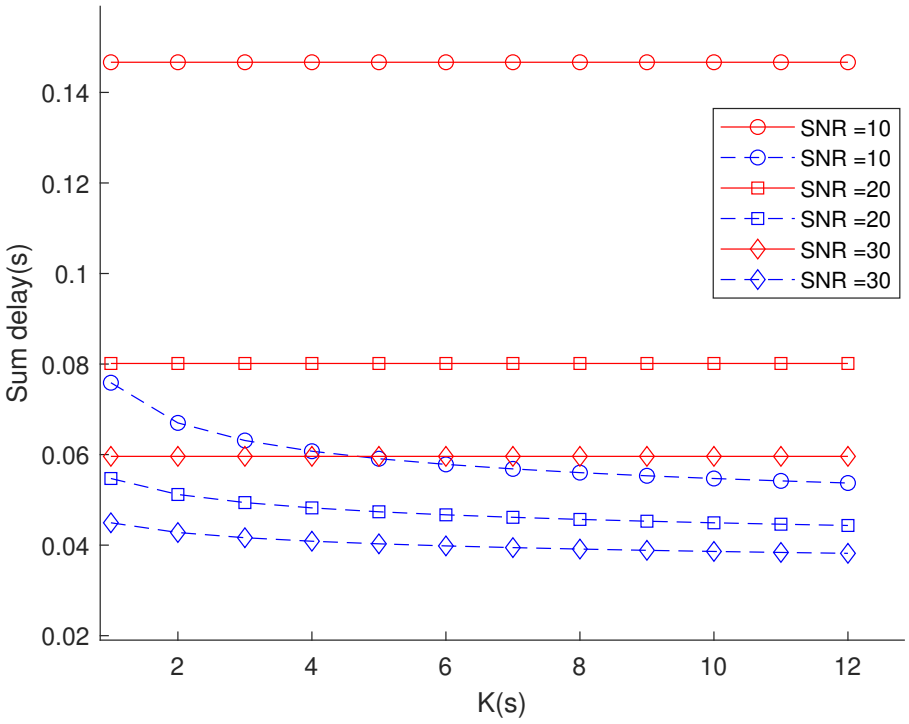

Figure 6: Variation of K

We observe that for each maximum SNR value, the geometric and DC programming methodology achieves lower values of total delay compared to linear programming methodology as the available transmission time increases, with greater gains for lower SNR values. The benchmark achieves optimal solutions that are not affected by changes in K, as the energy constraint has a linear relationship with K and a non-linear, decreasing relationship with SNR as $\frac{E}{BN_o} = K * pmax$. The proposed methodology better exploits the available energy, and we observe that for K greater than 0.6 s, for each SNR, the optimal delay converges.

Impact of SNR

For given values of $K \in 0.1, 0.2, 0.5$ and $SNR \in [10, 30]$, we have:

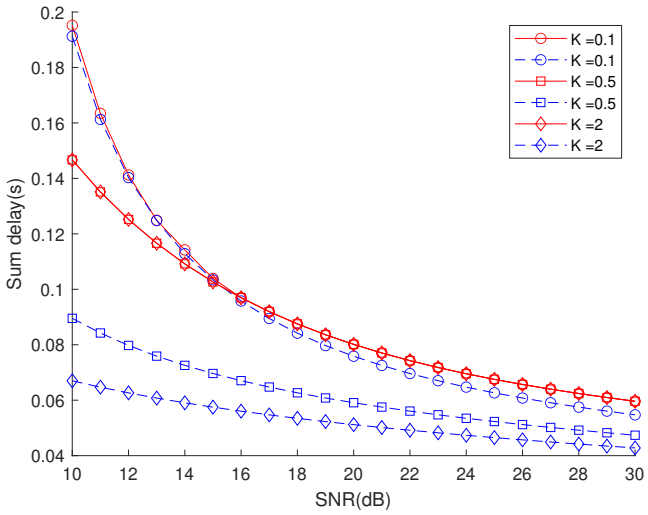

Figure 7: Variation of SNR

and for $K \in 0.5, 5, 12$ seconds,

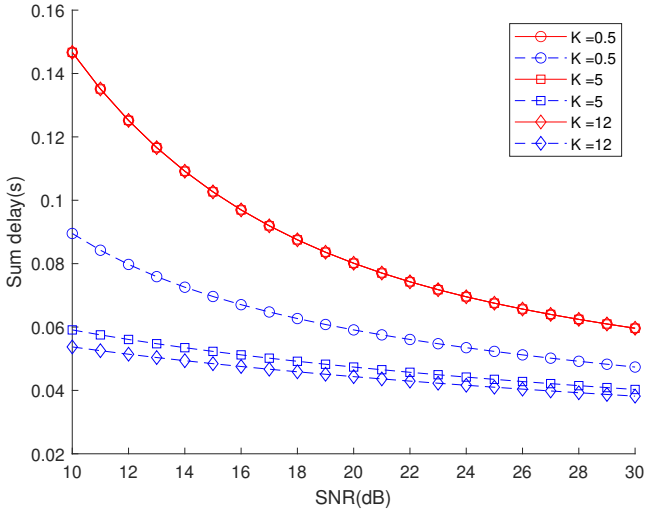

Figure 8: Variation of SNR

Similar to the simulations for fixed SNR, the choice of K does not affect the optimal solution of the linear programming methodology, and we observe that for higher K and SNR greater than 22 dB, the delay converges to a value up to 30% greater than the linear solution. For $K = 0.1$, the solutions of the two methods converge, especially for small SNR values.

Distances

The following diagrams are obtained with constant channel gains, assuming Rayleigh fading, and initially compare 3 different scenarios for user placement. The distance between user l and RRH m is denoted as d_{lm} for all $l \in i, j$ and $m \in a, b$, the path loss factor as w_{lm} for all $l \in i, j$ and $m \in a, b$, and the different scenarios are formulated as follows:

- The constant scenario in which all distances d_{lm} are equal and normalized to 1.
- The distant scenario in which the ratios of distances are $\frac{d_{ia}}{d_{ib}} = \frac{d_{jb}}{d_{ja}} = 1/5$, $d_{ia} = d_{jb}$, and the ratios of path loss factors are $\frac{w_{im}}{w_{jm}} = \left(\frac{d_{im}}{d_{jm}}\right)^3$ for all $m \in a, b$, which is equivalent to a path loss exponent of 3.
- The close scenario in which the distances are $\frac{d_{ia}}{d_{ib}} = \frac{d_{ja}}{d_{jb}} = 1/5$, $d_{ia} = d_{ja}$, with a path loss exponent of 3.

In Figures 9, 10, and 11, the effect of user distances on the overall delay can be observed by setting K and varying the maximum SNR achieved by users from 20 to 30 dB.

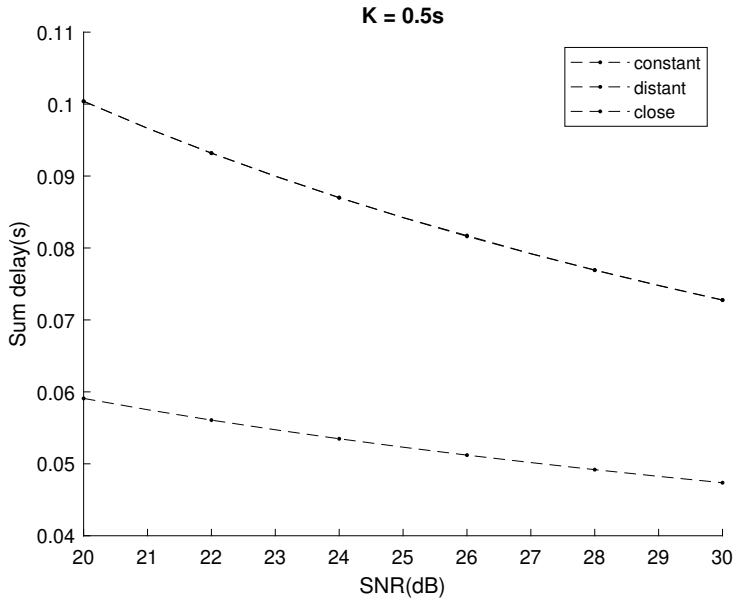

Figure 9: Distances Comparison

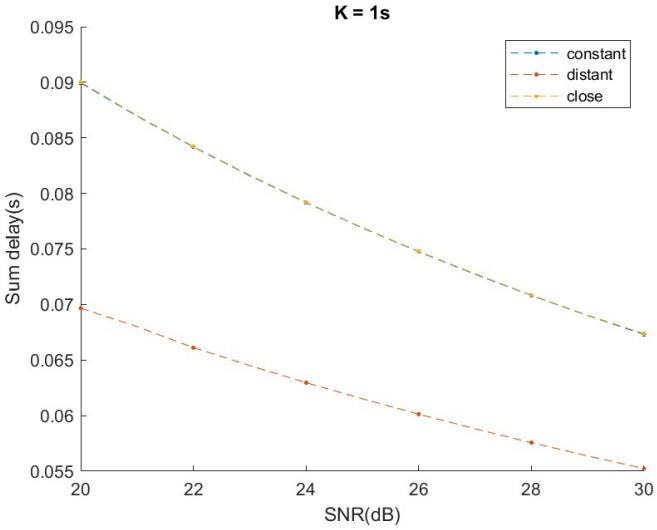

Figure 10: Distances Comparison

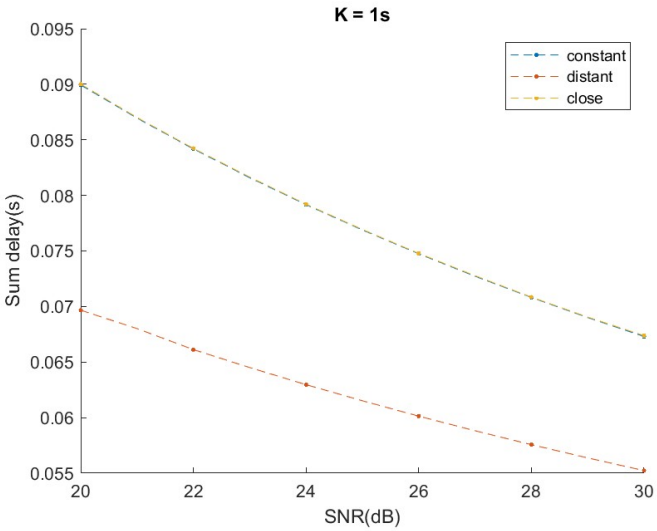

Figure 11: Distances Comparison

ugmenting the maximum normalized power per user decreases the overall delay, as expected, because in the optimal selection of decoding order during time sharing, the optimization process can more easily satisfy the rate constraints while keeping the overall energy low. Additionally, by comparing the plots, we observe that for larger values of K, smaller delays are achieved in all possible arrangements.

It is interesting to note that in the distant user arrangement, the achieved total delay is smaller compared to the constant and close arrangements, while in the cases of constant and close arrangements, the total delay is identical for all SNR values. A possible explanation for the difference in optimal solutions regarding distances could be the comparative performance advantage of NOMA systems over OMA when users have significant channel condition differences.

Similar behavior of the optimal solution is observed when we vary K while keeping SNR constant. In Figures 12, 13, and 14, the result of increasing the provided maximum transmission time at maximum power, K, from 0.1 to 2 seconds with constant SNR at 20, 25, and 30 dB, respectively, is presented.

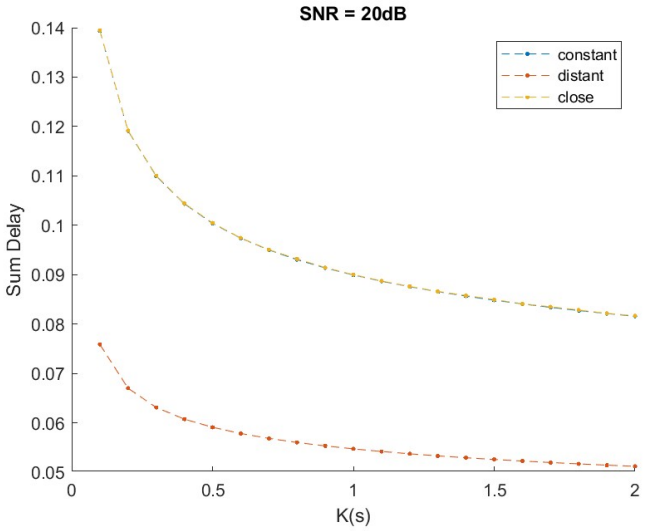

Figure 12: Distances comparison

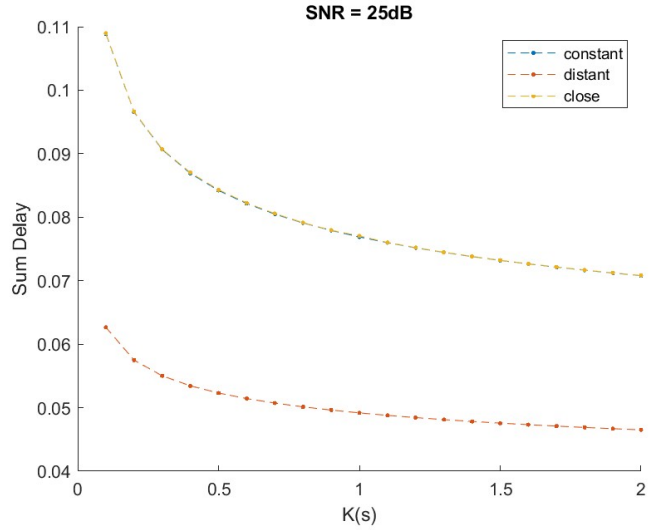

Figure 13: Distances comparison

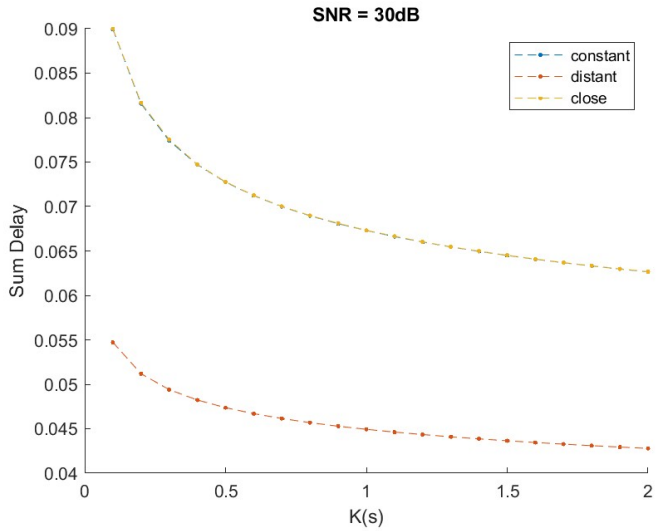

Figure 14: Distances comparison

It is evident that for small values of the parameter K, i.e., less available transmission time at maximum power, the assignment becomes more challenging and requires significantly more time, in milliseconds, than the minimum possible in these simulations. The relationship between the arrangement and the optimal solution remains the same, and we observe that for $K > 0.6$ seconds, increasing K does not significantly affect the maximum

delay. Comparing the plots, we observe that the minimum delay is achieved with increasing SNR, confirming the results of the above analysis.

Monte Carlo

General observations regarding energy-constrained delay minimization can be confirmed through a Monte Carlo simulation, where we assume that in each trial, the distances of users from the RRHs are different, resulting in different channel gains. Here, small-scale fading is taken from the random variable $h_l \sim \mathcal{CN}(0,1);\ \forall l \in \{i,j\}$.

Figure 15 presents the results of 50 trials, with a fixed $K = 0.5$, which we observed is sufficient to characterize the system behavior for most values:

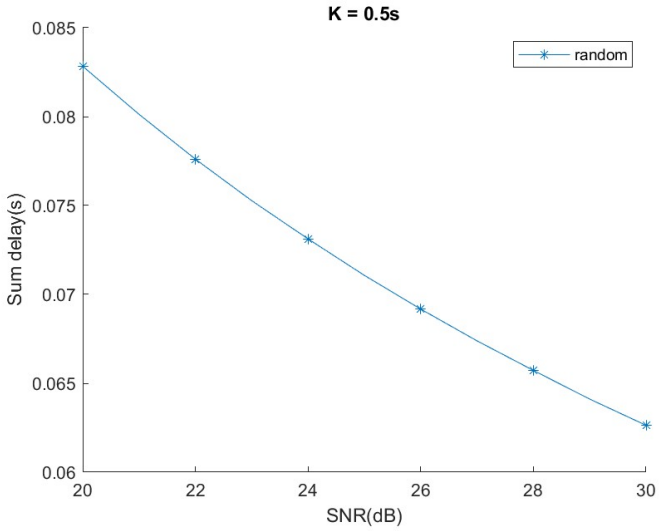

Figure 15: Monte Carlo

We notice that the "average" delay for arbitrary channel gain values is close to the delay for the "distant" user arrangement.

Accordingly, the average behavior of the system across 50 trials with $SNR = 20$ is illustrated in Figure 16:

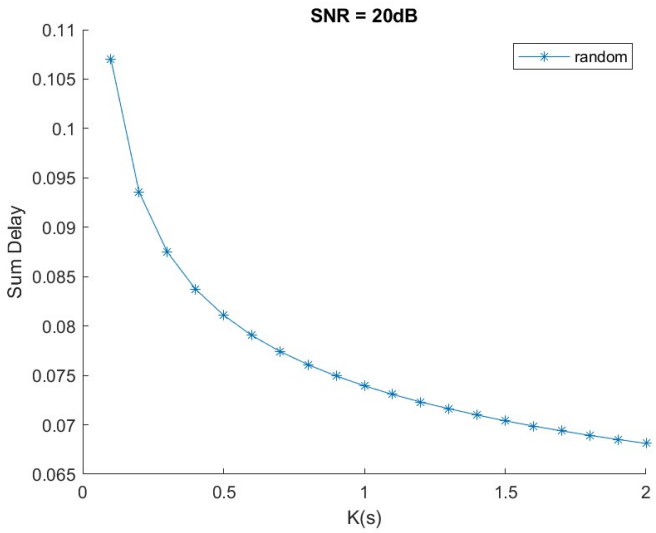

Figure 16: Monte Carlo

6 Simulations for energy minimization

In the case of energy consumption minimization, simulations are conducted with fixed values of acceptable delay $T_{\max}$, different values of weight constants w_i, w_j controlling the energy balance of users in the overall system consumption, and identical user arrangements as in the previous section. In all the following diagrams, the energy axis represents the energy normalized by the noise spectral density, $\frac{E}{B \cdot N_0}$.

Impact of Acceptable Delay *Dmax*

For SNR values of 10, 20, and 30, acceptable delay $Dmax$ in the range $[0.2, 1]$, and $w_i = w_j = 0.5$, we have:

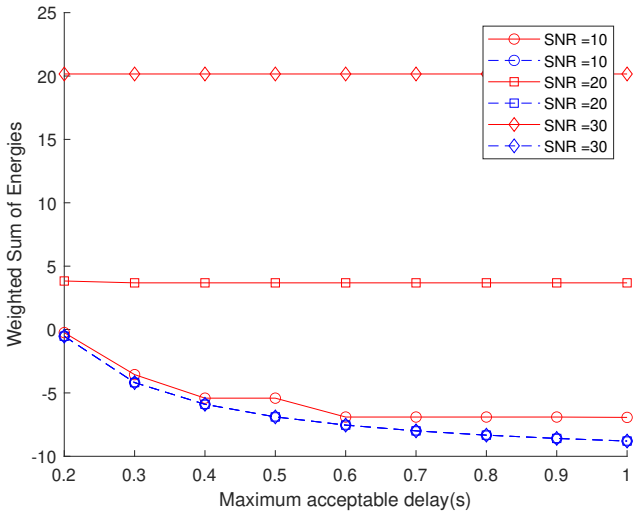

Figure 17: Variation of Dmax

We observe that for any value of the maximum SNR, the geometric-DC programming methodology achieves significantly lower overall energy consumption orders of magnitude compared to linear programming methodology. The increase in acceptable delay improves the energy consumption, as expected.

Impact of Distances

The constant and close arrangements present identical optimal values and are depicted on a common curve. Specifically, for $Tmax \in [0.2, 1]$, Figure 4.9

illustrates the total energy consumption E_{FO} for $w_i = w_j = 0.5$ and delay values $Tmax$ ranging from 0.2 to 2 s:

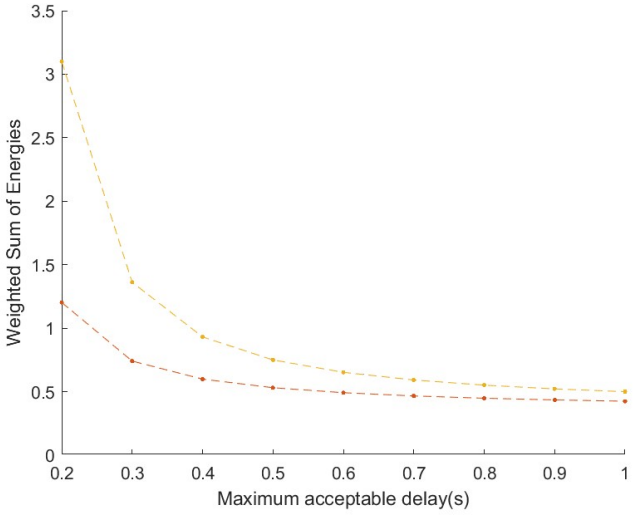

Figure 18: Distance Comparison

For $w_i = 0.1, w_j = 0.9$, prioritizing the consumption needs of only one user (in this case, j):

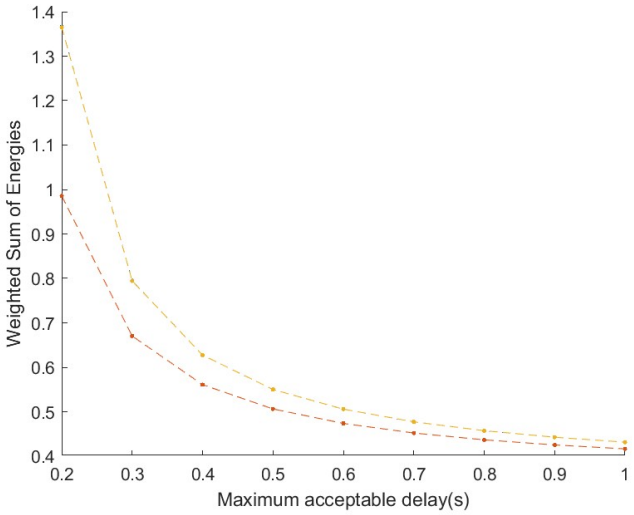

Figure 19: Distance Comparison

Finally, for $w_i = 0.4, w_j = 0.7$:

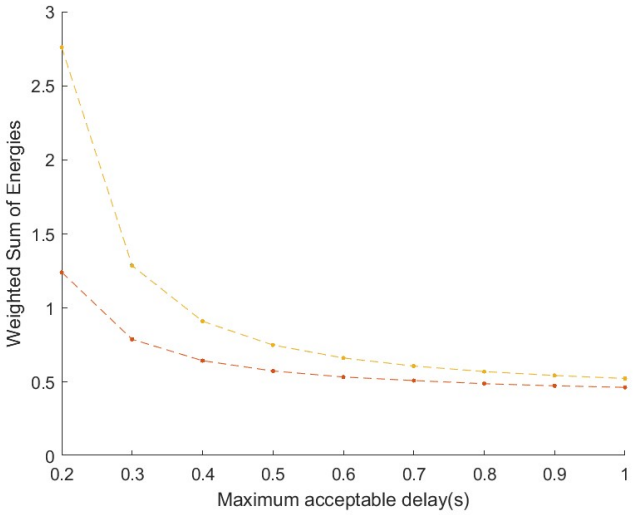

Figure 20: Distance Comparison

The red line represents the distant configuration and the yellow one the common close/constant. In the above figures, it is evident that the distant configuration benefits more from the NOMA technique in reducing the overall energy consumption of users compared to the other configurations, due to the significant difference in channel conditions to the RRHs. This is more apparent for $Tmax$ less than 0.4s.

Comparing the diagrams among them, we observe that by giving greater weight to one user ($w_i = 0.1, w_j = 0.9$, Figure 4.9), the minimum energy consumption is significantly lower compared to cases where consumption is more evenly distributed among users. Additionally, the influence of user configuration in space decreases as the acceptable delay increases (the diagrams have much smaller deviation). A possible explanation for this observation could be that the optimization process prioritizes satisfying the constraints of the dominant user ($w_j = 0.9$) and then finds satisfactory solutions for the energy consumption of the other user ($w_i = 0.1$).

Monte Carlo

General observations regarding delay-constrained energy minimization can be confirmed through a Monte Carlo simulation, where we assume that, in each trial, the distance of users from the RRHs varies, resulting in different channel gains, with small-scale fading being represented by the random variable $h_l \sim \mathcal{CN}(0, 1);\ \forall l \in i, j$. The plot is provided for 50 trials, with $w_i = w_j = 0.5$,

which, as noted, is sufficient to characterize the system's behavior for most values:

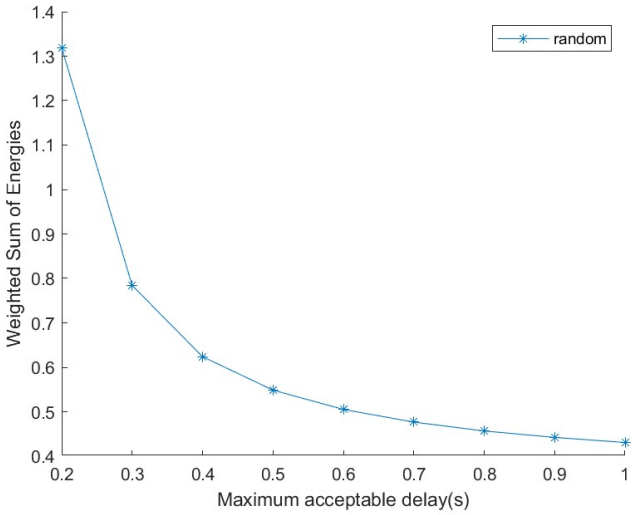

Figure 21: Monte Carlo

The average behavior regarding energy consumption is decreasing, as expected, and remains close to the delay values of the distant user arrangement.

References

[1] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5g networks for the internet of things: Communication technologies and challenges," *IEEE Access*, vol. 6, no. 10, pp. 3619–3647, 2018.

[2] N. Al-Falahy and O. Y. Alani, "Technologies for 5g networks: Challenges and opportunities," *IT Professional*, vol. 19, no. 1, pp. 12–20, Jan.-Feb. 2017.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[4] K.-C. Chen, T. Zhang, R. D. Gitlin, and G. Fettweis, "Ultra-low latency mobile networking," *IEEE Network*, vol. 33, no. 2, pp. 181–187, 2019.

[5] D. Soldani and A. Manzalini, "Horizon 2020 and beyond: On the 5g operating system for a true digital society," *IEEE Vehicular Technology Magazine*, vol. 10, no. 1, pp. 32–42, 2015.

[6] M. Alsabah, M. A. Naser, B. M. Mahmmod, S. H. Abdulhussain, M. R. Eissa, A. Al-Baidhani, N. K. Noordin, S. M. Sait, K. A. Al-Utaibi, and F. Hashim, "6g wireless communications networks: A comprehensive survey," *IEEE Access*, vol. 9, pp. 148 191–148 243, 2021.

[7] S. Zhang, C. Xiang, and S. Xu, "6g: Connecting everything by 1000 times price reduction," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 107–115, 2020.

[8] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[9] P. D. Diamantoulakis and G. K. Karagiannidis, "Performance analysis of distributed uplink noma," *IEEE Communications Letters*, vol. 25, no. 3, pp. 788–792, 2021.

[10] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5g networks:

Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[11] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in lte and 5g networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.

[12] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[13] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, 2010.

[14] "The road towards green ran," *China Mobile Res. Inst., Beiging, China, White Paper*, Oct.2011.

[15] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of noma: Current status and open research challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, 2020.

[16] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-noma for cloud radio access networks," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2274–2277, 2017.

[17] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on c-ran centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[18] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of ran architectures toward 5g mobile communication system," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019.

[19] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5g perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, 2014.

[20] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[21] H. Li, K. Ota, and M. Dong, "Learning iot in edge: Deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[22] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6g wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[23] H. Li, K. Ota, and M. Dong, "Eccn: Orchestration of edge-centric computing and content-centric networking in the 5g radio access network," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 88–93, 2018.

[24] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[25] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal *et al.*, "Mobile-edge computing introductory technical white paper," *White paper, mobile-edge computing (MEC) industry initiative*, pp. 1089–7801, 2014.

[26] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2017.

[27] K. Li, "Computation offloading strategy optimization with multiple heterogeneous servers in mobile edge computing," *IEEE Transactions on Sustainable Computing*, pp. 1–1, 2019.

[28] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[29] P. D. Diamantoulakis, P. S. Bouzinis, P. G. Sarigannidis, Z. Ding, and G. K. Karagiannidis, "Optimal design and orchestration of mobile edge computing with energy awareness," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 2, pp. 456–470, 2022.

[30] A. A. Al-Habob, O. A. Dobre, A. G. Armada, and S. Muhaidat, "Task scheduling for mobile edge computing using genetic algorithm and conflict graphs," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8805–8819, 2020.

[31] A. A. Al-Habob, A. Ibrahim, O. A. Dobre, and A. G. Armada, "Collision-free sequential task offloading for mobile edge computing," *IEEE Communications Letters*, vol. 24, no. 1, pp. 71–75, 2020.

[32] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of noma over oma in uplink communication systems," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 536–568, 2020.

[33] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 4–11, 2006.

[34] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8422–8436, 2016.

[35] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5g nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, 2016.

[36] N. Ye, H. Han, L. Zhao, and A.-h. Wang, "Uplink nonorthogonal multiple access technologies toward 5g: A survey," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[37] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 375–390, 2019.

[38] A. Kiani and N. Ansari, "Edge computing aware noma for 5g networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, 2018.

[39] J. Du, W. Liu, G. Lu, J. Jiang, D. Zhai, F. R. Yu, and Z. Ding, "When mobile-edge computing (mec) meets nonorthogonal multiple access (noma) for the internet of things (iot): System design and optimization," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7849–7862, 2021.

[40] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for noma-mec offloading," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1875–1879, 2018.

[41] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for noma–mec offloading," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6207–6211, 2019.

[42] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268–4282, 2016.

[43] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient noma-based mobile edge computing offloading," *IEEE Communications Letters*, vol. 23, no. 2, pp. 310–313, 2019.

[44] H. Li, F. Fang, and Z. Ding, "Joint resource allocation for hybrid noma-assisted mec in 6g networks," *Digital Communications and Networks*, vol. 6, no. 3, pp. 241–252, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864819304274

[45] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal task partition and power allocation for mobile edge computing with noma," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[46] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 392–407, 2019.

[47] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3435–3447, 2017.

[48] X. Diao, J. Zheng, Y. Wu, and Y. Cai, "Joint computing resource, power, and channel allocations for d2d-assisted and noma-based mobile edge computing," *IEEE Access*, vol. 7, pp. 9243–9257, 2019.

[49] F. Wang, J. Xu, and Z. Ding, "Multi-antenna noma for computation offloading in multiuser mobile edge computing systems," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2450–2463, 2019.

[50] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[51] S. Boyd, S.-J. Kim, L. Vandenberghe, and et al., "A tutorial on geometric programming," *Optimization and Engineering*, vol. 8, pp. 67–127, 2007. [Online]. Available: https://doi.org/10.1007/s11081-007-9001-7

[52] H. A. Le Thi and T. Pham Dinh, "Dc programming in communication systems: challenging problems and methods," *Vietnam Journal of Computer Science*, vol. 1, pp. 15–28, 2014. [Online]. Available: https://doi.org/10.1007/s40595-013-0010-5