Synergistic Benefits in IRS- and RS-enabled C-RAN with Energy-Efficient Clustering

Kevin Weinberger*, Alaa Alameer Ahmad*, Aydin Sezgin*, Alessio Zappone†

Abstract—The potential of intelligent reflecting surfaces (IRSs) is investigated as a promising technique for enhancing the energy efficiency of wireless networks. Specifically, the IRS enables passive beamsteering by employing many low-cost individually controllable reflect elements. The resulting change of the channel state, however, not only increases the signal quality but also the interference at the users. To counteract this negative side effect, we employ rate splitting (RS), which inherently is able to mitigate the impact of interference. We facilitate practical implementation by considering a Cloud Radio Access Network (C-RAN) at the cost of finite fronthaul-link capacities, which necessitate the allocation of sensible user-centric clusters to ensure energy-efficient transmissions. Dynamic methods for RS and the user clustering are proposed to account for the interdependencies of the individual techniques. Numerical results show that the dynamic RS method establishes synergistic benefits between RS and the IRS. Additionally, the dynamic user clustering and the IRS cooperate synergistically, reflected by increased individual gains for the dynamic clustering. Interestingly, with an increasing fronthaul capacity, the gain of the dynamic user clustering decreases, while the gain of the dynamic RS method increases. Around the resulting intersection, both methods affect the system concurrently, improving the energy efficiency drastically.

Index Terms—Beyond 5G (B5G), intelligent reflecting surface (IRS), reconfigurable intelligent surface (RIS), rate splitting (RS), successive interference cancellation (SIC), dynamic clustering, cloud radio access networks (C-RAN), interference management, resource allocation, energy efficiency (EE), imperfect channel state information (CSI), successive convex approximations (SCA).

I. INTRODUCTION

With the recent introduction of solutions based on the Internet of Things (IoT) on various application areas, Beyond 5G (B5G) mobile wireless networks are expected to support an abundance of devices, while simultaneously providing a 1000 times capacity increase at a similar or lower power consumption than current cellular systems [2]. To this end, spatial reuse is utilized, effectively densifying the deployment of remote radio units (RRUs) compared to previous networks. More specifically, by decreasing the distance between user and RRU, the employment of the spatial reuse technique achieves this goal as shorter distances result in direct channels of higher quality. However, this not only introduces additional operating costs to the network but also increases the interference in the network because of the proximity of the RRRUs within the area of operation. It follows that the interference also becomes a limiting factor for achieving high efficiency in modern communication networks.

In order to further improve the communication quality under these conditions without increasing the necessary power consumption, this work focuses on enhancing the energy efficiency (EE) of the network cost-effectively with the deployment of an intelligent reflecting surface (IRS) [3–7]. An IRS is a metasurface, consisting of many low-cost reflect elements, which are able to induce real-time changes to the reflected signals. As the induced change of each reflect element to the reflected signal is individually adjustable, the IRS enables passive beamsteering, which results in an increased efficiency of the communication network [3–7]. In a multiuser network, however, each phase shift, induced at the reflect elements, affects the reflected channel path of each user. This results in suboptimal phase shifts for some users and entails additional interfering links within the network. Due to this reason, the utilization of the passive beamsteering of the IRS in a multiuser scenario inherently introduces an increase in interference at the users.

Under these circumstances, the strategy of treating interference as noise (TIN) becomes an unattractive solution as it is known to be a suboptimal strategy, especially in high interference regimes [8, 9]. Thus, considering the interference caused by spatial reuse and the IRS, we deploy rate splitting (RS) [10, 11] as an efficient interference mitigation strategy in the context of our paper. In the RS strategy, the message of each user is split into a private part that is supposed to be decoded at the intended user only, and a common part, which can be decoded by a subset of users. Leveraging on this concept of RS, receivers can adopt successive interference cancellation (SIC) [12]. This not only enables the mitigation of the interference caused by spatial reuse but also inherently mitigates the increased interference caused by the improved channel gains of the IRS. Note that the use of RS alone in a classical multiuser network can achieve gains of up to 97% in terms of the EE [13], if compared to the baseline scheme of TIN [14, 15]. As it will turn out, with the addition of the IRS, a synergistic interaction between the IRS and the RS technique arises. In fact, the concurrent use of both techniques results in an EE gain, which is beyond the sum of the EE gains, each technique is able to achieve individually.

To enable the practical implementation of the RS strategy for a 5G network, this work utilizes a Cloud Radio Ac-
cess Network (C-RAN), in which a central processor (CP) centrally splits and encodes the users’ messages and coordinates the transmissions within the network. Moreover, utilizing the C-RAN architecture enables the utilization of coordinated multi-point (CoMP) transmissions based on user-centric clustering. Thus, each user can potentially be served by multiple RRUs, enhancing the user throughput and EE [16, 17]. However, utilizing this data-sharing strategy [18] can also have a negative impact on the EE of the network as each RRU is connected to the CP via fronthaul links of finite capacity. More precisely, in order to realize the CoMP transmissions, the CP is required to send the same message to multiple RRUs. This effectively introduces redundancy to the fronthaul, which potentially decreases the total achievable rate within the network. Especially in the fronthaul limited regime, this can have a major impact on the EE. Therefore, it becomes equally important to assess the sets of RRUs, which serve each user, while jointly designating the transmission mode of each user (i.e., private, common, or both) [19, 20].

In this work, we aim to obtain the full benefit from the cooperation of the RS technique and the IRS in order to maximize the EE of a fronthaul-constrained C-RAN. We intend to achieve this by investigating the coupling of the parameters and their interplay within the context of the user-centric clustering. However, capitalizing on this mutual interaction between these techniques poses difficult challenges, due to the dependencies among the parameters of the individual techniques.

A. Related Works

The contributions of this paper relate to works on RS, user clustering, IRS as well as active and passive beamforming; topics studied in the literature of wireless communication individually and jointly.

The research on RS-enabled C-RANs has recently received considerable attention when considering dense wireless communication systems, prioritizing different goals, i.e., sum-rate maximization [11, 21–23], reduction in power consumption [24], max-min fairness [25] and EE [26]. For EE, the objective of RS in C-RAN is to cooperatively reduce the interference at the users, therefore facilitating higher rates without increasing the transmit power. However, the authors in [26] demonstrated that the achievable benefits not only depend on the rate allocation but also on the available fronthaul capacity at the cloud. Addressing this issue, [22] introduces a successive convex approximation (SCA)-based user clustering approach, taking the limited fronthaul capacity into account by proposing an $\ell_0$-norm relaxation technique. The authors show that utilizing such a dynamic clustering approach in a RS-enabled C-RAN increases the system’s performance significantly. The above works, i.e., [11, 21–26], however, do not address IRS-assisted scenarios and do not account for a dynamic selection of common message sets. Thus, this paper focuses on the study of an IRS-assisted system as the passive beamforming makes it an attractive solution from an energy consumption standpoint [4, 5, 27]. More precisely, the authors in [4, 5] show that an IRS-assisted multiuser system can outperform a relay-assisted one in terms of EE, if a sufficiently large IRS is deployed. Moreover, the authors in [27] extend the above system with RS and employ an SCA-based approach to determine optimal phase shifters, showing that RS achieves up to 12% gains in terms of EE, when compared to a SIC-enabled non-orthogonal multiple access (NOMA). All these works, however, do not address cloud-enabled scenarios as they ignore the physical-layer considerations induced by RS in C-RANs. Furthermore, [27] assumes static sets of common messages that do not facilitate a cooperation between RS and the IRS. At this point, it becomes essential to investigate how a cooperative interaction between RS and the IRS can influence the EE in a C-RAN setup. To the best of the authors’ knowledge this is the first work, which studies the application of both, RS and IRS in a C-RAN setup, while facilitating cooperation between these techniques through dynamic allocation of user clusters and common-message-decoding (CMD) sets.

B. Contributions

In order to deal with the interdependencies between the individual techniques, this paper considers the problem of maximizing the EE of the network subject to per-user quality-of-service (QoS) constraints, per-RRU fronthaul capacity constraints and per-RRU transmission power constraints. Moreover, the IRS is constrained to only induce phase shifts to the reflected signals. The goal of this optimization is to jointly determine the CMD sets of each user, the corresponding clusters of RRUs serving each user’s common and private message and the associated beamformers and rates of each user’s private and common message, as well as an efficient alignment of the phase shifts at the IRS. To deal with the dependencies among the variables, we propose solving the problem in an alternating fashion with the aim of obtaining more tractable subproblems by decoupling the problem. The main contributions can be summarized as follows:

- **Centralized Optimization:** This paper considers a centralized approach, which splits the intractable EE problem into tractable subproblems. Since the emerging subproblems are still dependent on each other, we propose an algorithm that updates specific control-parameters between the optimizations of the individual subproblems. These parameters directly influence the solutions to the subproblems, enabling the cooperation towards the same objective.

- **Dynamic Common Message Decoding Set Allocation:** We propose a dynamic procedure that determines CMD sets, which cooperate well with the IRS as they are updated dynamically throughout the alternating optimization algorithm. Because the effectiveness of the RS technique to mitigate interference is based on the CMD sets, it is vital that the sets are updated with regards to the phase shifters of the IRS.

- **Dynamic User-Centric Clustering:** The EE of the system is dependent on sensible user clusters, the choice of which highly depends on the parameters of both, RS and the IRS. Thus, we adopt a formulation of the EE problem that enables the dynamic allocation of clusters (for both,
Figure 1: IRS-assisted multiuser C-RAN system with fronthaul links. The IRS-assisted channels carry both, intended and interfering signals to each user.

The private and common transmissions), which represent the perfect tradeoff between sending redundant messages to multiple RRUs and the system’s power consumption.

- Phase Shift Optimization: In order to guarantee the convergence of the alternating optimization approach, we determine phase shifters, which exclusively improve both, the achievable private rate and common rates at each user. The proposed approach, not only ensures the convergence of the alternating optimization but also facilitates cooperation of the IRS with the other techniques.

II. SYSTEM MODEL

In this work, we consider the system model depicted in Figure 1, which consists of an IRS-assisted and RS-enabled C-RAN downlink system. More precisely, the network consists of a set of multi-antenna RRUs $N = \{1, 2, \ldots, N\}$, each of which is equipped with $L \geq 1$ antennas. A set of single-antenna users $K = \{1, 2, \ldots, K\}$ is served by the RRUs. We consider the deployment of an IRS, composed of $R$ passive real-time controllable reflect elements, in the communication environment to support the RRUs in their transmissions to the users. We assume two cases of channel state information (CSI) knowledge at the CP of the C-RAN, i.e., perfect and imperfect CSI knowledge. The CP of the C-RAN is connected to each RRU $n \in N$ via orthogonal fronthaul links of finite capacity $C_{n}^{FH}$, while the IRS is connected to the CP via a separate fronthaul link. Each user $k$ has a QoS target, which is represented by a minimum data rate $C_{k}^{min}$.

A. Perfect CSI Scenario

The direct channel link between RRU $n$ and user $k$ is denoted by $h_{n,k}^{BU} \in \mathbb{C}^{L \times 1}$. Moreover, $H_{n}^{BI} \in \mathbb{C}^{L \times R}$ denotes the channel link from IRS to the IRS, while $h_{k}^{BU} \in \mathbb{C}^{R \times 1}$ denotes the channel link from IRS to the user $k$. Similarly, we denote $h_{n,k} = [(h_{1,k}^{BU})^T, (h_{2,k}^{BU})^T, \ldots, (h_{N,k}^{BU})^T]^T \in \mathbb{C}^{NL \times 1}$ as the aggregate direct channel vector of user $k$, $H_{n}^{BI} = [(H_{1}^{BI})^T, (H_{2}^{BI})^T, \ldots, (H_{N}^{BI})^T]^T \in \mathbb{C}^{NL \times R}$ as the aggregate channel matrix from the RRUs to the IRS and $x = [x_{1}^T, x_{2}^T, \ldots, x_{N}^T]^T \in \mathbb{C}^{NL \times 1}$ as the aggregate transmit signal vector. Using this notation, the received signal at user $k$ can be written as the sum of the direct and IRS-assisted link, namely

$$y_{k} = h_{k}^{H} x + (H_{k}^{BI} \Theta h_{k}^{BU})^H x + n_{k},$$

(1)

where $n_{k} \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) and $\Theta = \text{diag}(\upsilon) \in \mathbb{C}^{R \times R}$ is a diagonal reflection coefficient matrix representing the response of the reflect elements. We define the phase shift vector as $\upsilon = [A_{1} e^{j \theta_{1}}, \ldots, A_{R} e^{j \theta_{R}}]^T$, where each phase shift $\theta_{r} \in [0, 2\pi]$ and reflect amplitude $A_{r} \in [0, 1]$ induced at the $r$-th reflect element is represented by the corresponding reflection coefficient defined as $\upsilon_{r} = A_{r} e^{j \theta_{r}}$. This notation facilitates denoting the combination of the aggregated direct and reflected channel vectors of user $k$ as an effective channel, namely

$$h_{k}^{eff}(\upsilon) = h_{k} + H_{k} \upsilon,$$

(2)

where $H_{k} = H_{k}^{BI} \text{diag}(h_{k}^{BU})$. The received signal (1) at user $k$ can thus be rewritten as

$$y_{k} = (h_{k}^{eff}(\upsilon))^H x + n_{k}.$$

(3)

B. Imperfect CSI Scenario

In this section, we introduce a channel estimation protocol to the system, which is necessary to compute the transmit and reflect beamforming vectors at the CP. In practice, channel estimation has a negative impact on the achievable rates at the users, since it not only introduces channel estimation errors but also shortens the duration of the transmission phase. In this work, we adopt the channel estimation protocol introduced in [28] in order to study the impact of channel estimation errors on the performance of the proposed algorithms. To this end, we utilize the proposed minimum mean squared error (MMSE) estimators [28, Lemma 1, Lemma 2], which, by invoking the orthogonality property of the MMSE estimates, facilitate the decomposition of the effective channel $h_{k}^{eff}(\upsilon)$ into

$$h_{k}^{eff}(\upsilon) = (\hat{h}_{k} + \Delta h_{k}) + (\hat{H}_{k} + \Delta H_{k}) \upsilon,$$

(4)

where $\hat{h}_{k}$ is the MMSE estimate of $h_{k}$ and $\Delta h_{k} = [(\Delta h_{1,k}^{BU})^T, \ldots, (\Delta h_{N,k}^{BU})^T]^T$, where $\Delta h_{n,k}^{BU} \sim \mathcal{CN}(0, \Xi_{n,k}^{BU})$ is the uncorrelated estimation error with

$$\Xi_{n,k}^{BU} = \sigma_{n,k}^{BU} 1_{L} - (\sigma_{n,k}^{BU})^2 1_{L} (\alpha_{n,k}^{BU} 1_{L} + \sigma_{R}^{2} P_{C} \tau_{c})^{-1} 1_{L},$$

(5)

where $\sigma_{n,k}^{BU}$ is the pathloss factor of the $n$-th RRU-to-user $k$ link, $1_{L}$ is the $L \times L$ identity matrix, $\sigma_{R}$ is the variance of received noise at the RRUs, $P_{C}$ is the transmit power of each user when transmitting the pilot sequences and $\tau_{c} = S \tau_{s}$ is the total channel estimation period, which consists of $S$ training sub-phases $1$, each of length $\tau_{s}$. Similarly, $\hat{H}_{k}$ denotes the MMSE estimate of $H_{k}$ and $\Delta H_{k} = [(\Delta H_{1,k})^T, \ldots, (\Delta H_{N,k})]^T$, with $\Delta h_{r,k} = [(\Delta h_{1,k})^T, \ldots, (\Delta h_{N,k})]^T$, where

1Note, that we choose $S = R + 1$ since it has been shown to be optimal when utilizing this protocol.
\[ \Delta h_{n,r,k} \sim \mathcal{CN}(0, \Xi_{n,r,k}) \] represents the uncorrelated estimation error of the cascaded channels from RRU \( n \) to user \( k \) over reflect element \( r \) with

\[
\Xi_{n,r,k} = \alpha_{r,k}^U h_{n,r,k}^B (h_{n,r,k}^B)^H - (\alpha_{r,k}^U)^2 h_{n,r,k}^B (h_{n,r,k}^B)^H
\]

(6)

where \( \alpha_{r,k}^U \) is the path loss factor from reflect element \( r \) to user \( k \) and \( h_{n,r,k}^B \) denotes the \( r \)-th column of \( H_{n,r,k}^B \) and are both assumed to be known at the CP because they need to be determined only once as the position of the IRS and RRUs is fixed.

### C. Rate Splitting

The RS scheme can be adopted with different approaches \([10, 29]\). The scheme adopted in this paper is based on \([10]\), in which the CP splits the requested message of user \( k \), denoted as \( u_k \), into two sub-messages, namely a private part \( u_k^p \) and a common part \( u_k^c \). The CP subsequently encodes the respective parts into the private and common symbols denoted as \( s_k^p \) and \( s_k^c \), respectively. It is assumed that these messages form an independent and identically distributed (i.i.d.) Gaussian codebook. Afterwards, the CP distributes the private symbols \( s_k^p \) and common symbols \( s_k^c \) with a cluster of RRUs that exclusively transmits the beamformed private or common symbols to user \( k \). We denote the subsets of users that are served by RRU \( n \) with a private or common message as \( \mathcal{K}_n^p, \mathcal{K}_n^c \subset \mathcal{K} \), respectively. With the knowledge of these subsets, the CP is able to create the beamformers \( \omega_{n,k}^p, \omega_{n,k}^c \), used by RRU \( n \) to send \( s_k^p \) and \( s_k^c \), respectively, and forwards them to RRU \( n \) through the fronthaul link \( C_{FH} \) along with the respective private symbols \( \{s_k^p \mid k \in \mathcal{K}_n^p\} \) and common symbols \( \{s_k^c \mid k \in \mathcal{K}_n^c\} \). Due to the finite fronthaul capacity limit \( C_{FH} \), the achievable transmission rate is subjected to the following fronthaul constraint:

\[
\sum_{k \in \mathcal{K}_n^p} \xi_k + \sum_{k \in \mathcal{K}_n^c} \xi_k \leq C_{FH}^n, \quad \forall n \in \mathcal{N},
\]

(7)

where \( \xi_k \) and \( \xi_k \) are the private and common rate of user \( k \), respectively, and thus \( \xi_k = \xi_k^p + \xi_k^c \), where \( \xi_k \) is the rate of user \( k \). Moreover, we ignore the overhead of sending the beamformers since they are neglectable compared to the size of the data symbols.

After receiving the symbols with the corresponding beamformers, each RRU \( n \) constructs the overall transmit signal vector \( x_n \), which is defined as

\[
x_n = \sum_{k \in \mathcal{K}_n^p} \omega_{n,k}^p s_k^p + \sum_{k \in \mathcal{K}_n^c} \omega_{n,k}^c s_k^c
\]

(8)

and subject to the following power constraint:

\[
\mathbb{E}\{x_n^H x_n\} \leq P_{n,\text{Max}}, \quad \forall n \in \mathcal{N},
\]

(9)

where \( P_{n,\text{Max}} \) represents the maximum transmit power that is available at RRU \( n \). By denoting the aggregate beamforming vectors as \( \omega_k^c = \left[ (\omega_{n,k}^p)^T, (\omega_{n,k}^c)^T, \ldots, (\omega_{n,K}^c)^T \right]^T \in \mathbb{C}^{NL \times 1} \), \( \forall o \in \{p, c\} \) associated with \( s_k^o \), the aggregate transmit signal vector can be constructed as

\[
x = \sum_{k \in \mathcal{K}_n^p} \omega_{k}^p s_k^p + \sum_{k \in \mathcal{K}_n^c} \omega_{k}^c s_k^c.
\]

(10)

Using the definition of the transmit signal vector (10), the power constraint (9) can be rewritten as

\[
\sum_{k \in \mathcal{K}_n} ||\omega_{n,k}^p||^2 + ||\omega_{n,k}^c||^2 \leq P_{n,\text{Max}}, \quad \forall n \in \mathcal{N}.
\]

(11)

Note that if RRU \( n \) does not participate in the cooperative transmission of the private or common symbol of user \( k \), the respective beamformers are set to zero, i.e., \( \omega_{n,k}^p = 0_L \) or \( \omega_{n,k}^c = 0_L \), where \( 0_L \) denotes a column vector of length \( L \) with all zero entries. This can be equivalently expressed in terms of the indicator function \([16]\), i.e.,

\[
\mathbb{I}\left\{ ||\omega_{n,k}^p||^2 \geq 0 \right\}, \quad \forall o \in \{p, c\}.
\]

(12)

Without loss of generality, this indicator function can be written as the \( \ell_0 \)-norm, i.e., \( \mathbb{I}\left\{ ||\omega_{n,k}^p||^2 \geq 0 \right\} = \mathbb{I}\left\{ ||\omega_{n,k}^p||^2 = 0 \right\} \) because the power, that RRU \( n \) transmits to user \( k \), is a positive scalar, i.e., \( ||\omega_{n,k}^p||^2 \in \mathbb{R}^+ \). Using this definition, the subset of users \( \mathcal{K}_n^p \) and \( \mathcal{K}_n^c \) can be written as

\[
\mathcal{K}_n^p = \{ k \in \mathcal{K} \mid ||\omega_{n,k}^p||^2 = 0 \},
\]

(13)

\[
\mathcal{K}_n^c = \{ k \in \mathcal{K} \mid ||\omega_{n,k}^c||^2 = 0 \}.
\]

(14)

Using the expressions (13) and (14), the fronthaul capacity constraints in (7) can be reexpressed in the following form:

\[
\sum_{k \in \mathcal{K}_n} \left( ||\omega_{n,k}^p||^2 + ||\omega_{n,k}^c||^2 \right) \leq C_{FH}, \quad \forall n \in \mathcal{N}.
\]

(15)

### D. Achievable rates

In this work, the influence of the CMD scheme, adopted by the users, is utilized for the purpose of interference mitigation, especially focusing on mitigating the additional interference caused by the IRS-enhanced channels. Each user decodes a subset of common messages with a fixed decoding order in addition to their private message. It follows that choosing the decoding order for user \( k \), decoding its common and private messages, is vital for the performance of using RS as interference mitigation technique. Hence, we adopt a successive decoding strategy, in which user \( k \) decodes the common message of the user, whose interference is the strongest, first. This enables each user to decode and efficiently cancel parts of the received interference. Thus, let \( \mathcal{M}_k \) be the set of users decoding \( s_k^c \):

\[
\mathcal{M}_k = \{ j \in \mathcal{K} \mid \text{user } j \text{ decodes } s_k^c \}.
\]

(16)

In addition, let the set of users \( \Phi_k \), whose common messages are decoded by user \( k \), and the set of users \( \Phi_k \)'s, whose common messages are not decoded by user \( k \), be defined as

\[
\Phi_k = \{ j \in \mathcal{K} \mid k \in \mathcal{M}_j \}, \quad \Phi_k = \{ j \in \mathcal{K} \mid k \notin \mathcal{M}_j \},
\]

(17)

in which \( \Phi_k \) and \( \Phi_k \) are two disjoint subsets from the set of active users \( \mathcal{K} \), while the cardinality of \( \Phi_k \) is bounded by the
number of layers $D$ in the successive decoding strategy, i.e., $|\Phi_k| \leq D$. To decode the common messages in the set $\Phi_k$, user $k$ follows a decoding order, which is defined as

$$\pi_k(j) : \Phi_k \rightarrow \{1, 2, \ldots, |\Phi_k|\}$$

(18)

and represents a bijective function of the set $\Phi_k$ with cardinality $|\Phi_k|$. Thus, $\pi_k(j)$ is the successive decoding step, in which the message $j \in \Phi_k$ is decoded at user $k$. Accordingly, the expression $\pi_k(j_1) > \pi_k(j_2)$ signifies that user $k$ prioritizes decoding the common message of user $j_1$ before decoding the common message of user $j_2$, assuming $j_1 \neq j_2$.

Using the expressions above, the received signal at user $k$ can be expressed as

$$y_k = \left( h_k^{\text{eff}}(v) \right)^H \Omega_k + \sum_{j \in \Phi_k} \left( h_k^{\text{eff}}(v) \right)^H \omega_j s_j + \sum_{m \in K \setminus \{k\}} \left( h_k^{\text{eff}}(v) \right)^H \omega_m s_m + \sum_{i \in \Phi_k} \left( h_k^{\text{eff}}(v) \right)^H \gamma_{i,k}^c = A_k^n + n_k.$$  

(19)

This expression captures the received signal with interfering noise and cross-coupling. Let $\gamma_k^P$ denote the signal-to-interference-plus-noise ratio (SINR) of user $k$ decoding its private message, and let $\gamma_i^c$ denote the SINR of user $i$ decoding the common message of user $k$. To assure that each user is served with its request, the QoS rates of each user following conditions are satisfied by the private and common messages

$$\begin{align*}
\Omega_{i,k} = \{m \in \Phi_i | \pi_i(k) > \pi_i(m)\}. \\
\text{maximize} & \quad E_{\text{Qos}}(\omega, v, S) \\
\text{subject to} & \quad (11), (15), (23), (24), (25),
\end{align*}$$

(26)

in which $B$ denotes the transmission bandwidth and the variable

$$S = \{M_k, \Phi_k, \Phi_k, \{\Omega_{i,k}\}_{i \in K}\}_{k \in K}$$

(27)

is introduced, which represents the set variables of the RS and the successive decoding strategy.

### III. Problem Formulation

In this work, we are interested in the joint optimization of the power control, private and common rate allocations, CMD sets, clusters and their associated beamformer design for each user. In addition, we aim to jointly determine the phase shift design for the IRS, which maximizes the EE of the network subject to per-user QoS constraints and shared per-RRU fronthaul and power constraints. To this end, let the total transmit power $P^T$ be defined as

$$P^T(\omega) = \sum_{k \in K} (\omega^P_n k)^2 + (\omega_n^C k)^2$$

(26)

and the total rate be defined as

$$\begin{align*}
\xi^P(\omega, v, S) = & \sum_{k \in K} (T^P_k(\omega, v, S) + T^C_k(\omega, v, S)) \\
= & \sum_{k \in K} B \left( \log_2 \left( 1 + \gamma_k^P(\omega, v) \right) + \min_{i \in M_k} \left( \log_2 \left( 1 + \gamma_i^c(\omega, v) \right) \right) \right).
\end{align*}$$

(28)

The EE of the network can then be defined as

$$E_{\text{Qos}}(\omega, v, S) = \frac{\xi^P(\omega, v, S)}{\nu_{PA} P^T(\omega) + P^CIRC(\omega, v, S)},$$

(29)

where $\nu_{PA}$ denotes the efficiency of the transmit power amplifier (PA) and $P^CIRC(\omega, v, S) = P^{PR} + P^{IRS} + P^{FH}(\omega, v, S)$ represents the required processing power of the network. More precisely, it is defined by the signal processing circuitry of the RRsUs and the processing power of the CP, denoted as $P^{PR}$, as well as the required power to perform phase shifting on the impinging signal $P^{IRS} = \rho_{\text{r}} ^{IRS}$, where $\rho_{\text{r}}$ represents the power consumption of each phase shifter [4]. Moreover, $P^{FH}(\omega, v, S)$ represents the power that is required for the fronthaul traffic and can be defined as

$$P^{FH}(\omega, v, S) = P^{MBS} \sum_{k \in K, r \in N} \left( ||\omega^P_n k||^2 + ||\omega^C_n k||^2 \right) + \left( ||\gamma_k^C||^2 \right) T^C_k(\omega, v, S),$$

(30)

where $P^{MBS}$ is an EE factor representing the power consumption per mega bit rate [30]. By denoting the stacked private and common rate of each user $k$ as $\xi_k = [\xi^P_k, \xi^C_k]^T$ and defining $\xi = [\xi^P_1, \ldots, \xi^P_K]^T \in \mathbb{R}_{+}^{2K}$ to represent all rates, the problem can be mathematically formulated as

$$\begin{align*}
\text{maximize} & \quad E_{\text{Qos}}(\omega, v, S) \\
\text{subject to} & \quad (11), (15), (23), (24), (25),
\end{align*}$$

(31)

where the unit-modulus constraints in (31) represent the phase shift constraints if optimal reflect amplitudes are assumed, i.e., $0 \leq \theta_r \leq 2\pi, A_r = 1, \forall r \in \{1, \ldots, R\}$. The problem is generally difficult to solve, due to the non-convex constraints.
in combination with the set design problem. Additionally, the optimization variables $\omega, v$ and $S$ are coupled in $E_{\text{Obj}}$, (24) and (25).

IV. ALTERNATING OPTIMIZATION

To facilitate practical implementation, the problem is solved in an alternating fashion by proposing an alternating optimization framework. This decouples the variables $v, \omega$ and $S$ and results in three subproblems originating from (P1), namely a beamforming design problem, a phase shift design problem and a set design problem, each of which is solved individually with a different framework. In order to obtain the most efficient combination of IRS phase shifts $v$ and set variables $S$, we propose to optimize the phase shifts to prioritize the sum-rate gain in the beginning of the optimization process. This results in potentially higher efficiency of the phase shifts at the cost of high interference at some users, which we will mitigate with RS by designing sensible set variables $S$. Moreover, as elaborated later, the non-smooth $\ell_0$-norm is approximated with a smooth approximation [11]. However, even with this approximation, the optimization problem is still defined by a non-convex feasible set. To address this problem, we derive surrogate upper-bound functions for all non-convex functions, thereby approximating the non-convex set with a convex one. The approximations are then iteratively improved until convergence by using the SCA approach [13].

A. Relaxation of the $\ell_0$-norm

To approximate the non-smooth, non-convex $\ell_0$ norm, we utilize a smooth and convex approximation function. To this end, we consider the function

$$f_A(x) = \frac{1}{\pi} \arctan \left( \frac{x}{A} \right), \quad x \geq 0,$$

which is frequently used to approximate the $\ell_0$-norm and where $A > 0$ is the smoothness parameter controlling the quality of the approximation [31]. Note that by only relaxing the binary constraints in (15) and (30), e.g., by using the $\ell_1$ relaxation to the $\ell_0$-norm, the problem still remains non-convex due to the non-convexity of $E_{\text{Obj}}(\omega, v, S)$ as a function of the beamforming vectors. With the help of function (32) and equations (24) and (25), we are able to derive the upper-bound functions for (15) and (30) by replacing the non-convex $\ell_0$-norm with the approximated convex function, namely

$$\sum_{k \in K} f_A(\|\omega_{n,k}\|^2 / 2) \xi_k^p + f_A(\|\omega_{n,k}\|^2) \xi_k \leq C_{\text{PH}}^n, \forall n \in N, \quad (33)$$

$$P_{\text{Amp}} = \sum_{k \in K} \sum_{n \in N} (f_A(\|\omega_{n,k}\|^2 / 2) \xi_k^p + f_A(\|\omega_{n,k}\|^2) \xi_k). \quad (34)$$

B. Beamforming Design

Due to the alternating optimization approach, $v$ and $S$ are assumed to be fixed during the beamforming design procedure. This facilitates rewriting the optimization problem (P1) as

$$\max_{\omega, t, \xi} \left( \sum_{k \in K} (\xi_k^p + \xi_k) \right)$$

subject to (11), (23), (33),

$$\sum_{k \in K} \xi_k^p - B \log_2(1 + t_k^p) \leq 0, \quad \forall k \in K, \quad (35)$$

$$\sum_{k \in K} \xi_k - B \log_2(1 + t_k^p) \leq 0, \quad \forall k \in K, \quad (36)$$

$$t_k \geq 0, \quad \forall k \in K, \quad (37)$$

$$\xi_k \geq 0, \quad \forall k \in K, \quad (38)$$

$$t_k^p \leq \gamma_k^p(\omega), \quad \forall k \in K, \quad (39)$$

$$\xi_k \leq \gamma_{ik}^p(\omega), \quad \forall k \in K, \forall i \in M_k, \quad (40)$$

where the slack variables $t_k = [t_k^p, t_k^c]^T$ are introduced to obtain a convex representation of the rate expressions and $P_{\text{Amp}}^\text{Circ}(\omega, \xi)$ denotes the utilization of (34) instead of (30). The notation $\xi_k \geq 0$ and $\xi_k \geq 0$ indicates that vector $t_k$ and matrix $\xi_k$ are greater than or equal to 0 in a component-wise manner.

Problem (P2.1) is a fractional programming problem and is conventionally solved with the Dinkelbach algorithm [32]. However, the feasible set of problem (P2.1) is non-convex as the constraints (39) and (40) still define a non-convex feasible set, which makes it computationally inefficient to apply the Dinkelbach transformation directly to solve the problem [33]. We overcome this challenge by applying the SCA approach in combination with the Dinkelbach transformation [13]. In order to obtain a convex representation of the constraints (39) and (40), they can be rewritten into the following form

$$\sum_{k \in K} \left| (h_k^e)^H \omega_m^p \right|^2 + \sum_{\ell \in \Phi_k} \left| (h_k^e)^H \omega_m^c \right|^2 + \sigma^2 - \frac{| (h_k^e)^H \omega_m^p |^2}{t_k^p} \leq 0, \forall k \in K, \quad (41)$$

$$T_i + \sum_{\ell \in \Phi_k} \left| (h_k^e)^H \omega_m^c \right|^2 + \sum_{m \in M_i,k} \left| (h_k^e)^H \omega_m^c \right|^2 - \frac{| (h_k^e)^H \omega_m^c |^2}{t_k^c} \leq 0, \forall k \in K, \forall i \in M_k, \quad (42)$$

with an abuse of notation by skipping the dependency of $h_k^e$ on $v$. Next, the constraints can be approximated by using the first-order Taylor approximation around a feasible point $(\hat{\omega}, \hat{v})$ [34] as they are represented by a difference of convex functions. By applying the first-order Taylor approximations to the fractional terms in (41) and (42), upper bounds for these constraints can be derived. Consequently, if $(\hat{\omega}, \hat{v})$ is a feasible point of problem (P2.1) then it holds [11] that

$$\frac{| (h_k^e)^H \omega_m^p |^2}{t_k^p} \geq 2 \text{Re} \left\{ (\hat{\omega}_m^p)^H (h_k^e)^H \omega_m^p \right\} - \frac{| (h_k^e)^H \omega_m^p |^2}{t_k^p}, \forall k \in K, \quad (43)$$

$$\frac{| (h_k^e)^H \omega_m^c |^2}{t_k^c} \geq 2 \text{Re} \left\{ (\hat{\omega}_m^c)^H (h_k^e)^H \omega_m^c \right\} - \frac{| (h_k^e)^H \omega_m^c |^2}{t_k^c}, \forall k \in K, \forall i \in M_k, \quad (44)$$

where $\text{Re} \{ \cdot \}$ denotes the real part of a complex-valued number. By utilizing the approximations (43) and (44), inner-
convex approximations of the constraints (41) and (42) can be established by substituting the corresponding terms with their respective upper bound, namely

\[
0 \geq \sum_{j \in \mathcal{K} \setminus \{k\}} |(h_{k_j}^e)^H \omega_j|^2 + \sum_{t \in \mathcal{P}_k} |(h_{k_t}^e)^H \omega_t|^2 + \sigma^2 + \left| (h_{k_k}^e)^H \Omega_k^e \right|^2 \left( \frac{p_k}{\bar{t}_k^e} \right)^2 - 2 \text{Re}\left\{ (\Omega_k^e)^H h_{k_k}^e (h_{k_k}^e)^H \omega_k^e \right\}, \quad \forall k \in \mathcal{K}, \quad (45)
\]

\[
0 \geq T_i + \sum_{t \in \mathcal{F}_i} |(h_{k_t}^e)^H \omega_t|^2 + \sum_{n \in \mathcal{N}_i} |(h_{k_n}^e)^H \omega_n|^2 + |(h_{k_k}^e)^H \omega_k^e|^2 \left( \frac{p_k}{\bar{t}_k^e} \right)^2 - 2 \text{Re}\left\{ (\omega_k^e)^H h_{k_k}^e (h_{k_k}^e)^H \omega_k^e \right\}, \quad \forall k \in \mathcal{K}, \quad \forall i \in \mathcal{M}_k. \quad (46)
\]

Next, we focus on finding a convex representation for constraint (33). To this end, we introduce the slack variables \( q_k = [(p_k^0, \bar{q}_k^0)^T, q = [(q_1^T, \ldots, q_{T_k}^T)^T, \quad d_k = [d_1^k, d_1^k, \ldots, d_{N,k}^T, d_{N,k}^T] \) and \( d = [(d_1^T, \ldots, d_{T}^T)^T \) in order to split the constraint into five simpler constraints [11, Proposition 1] as follows:

\[
\begin{align*}
\text{maximize} & \quad \sum_{k \in \mathcal{K}} (\xi_k^0 + \xi_k^e)
\text{subject to} & \quad \text{(11), (23), (35) – (38), (45), (46)}, \\
& \quad f_A \left( \left| (\omega_p^k) \right|^2 \right) \leq d_{n,k}^p, \quad \forall n \in \mathcal{N}, \quad \forall k \in \mathcal{K}, \\
& \quad f_A \left( \left| (\omega_{n,k}) \right|^2 \right) \leq d_{n,k}^e, \quad \forall n \in \mathcal{N}, \quad \forall k \in \mathcal{K}, \\
& \quad \log_2(1 + t_k^e) \leq q_k^0, \quad \forall k \in \mathcal{K}, \\
& \quad \log_2(1 + t_k^e) \leq d_{n,k}^e, \quad \forall n \in \mathcal{N}, \quad \forall k \in \mathcal{K}, \\
& \quad \sum_{k \in \mathcal{K}} (d_{n,k}^p q_k^0 + d_{n,k}^e q_k^0) \leq C_{\mathcal{F}}^n / B, \forall n \in \mathcal{N}. \quad (51)
\end{align*}
\]

The same procedure is applied to the constraints (49) and (50) by linearizing the concave functions \( \log_2(1 + t_k^e) \) around \( \bar{t}_k^e \), namely

\[
\log_2(1 + t_k^e) + \left( \frac{t_k^e - \bar{t}_k^e}{1 + \bar{t}_k^e} \right) \ln(2) \leq q_k^0, \forall k \in \mathcal{K}, \forall o \in \{p, e\}. \quad (55)
\]

With the approximations defined above, we are able to formulate an approximate optimization problem as:

\[
\begin{align*}
\text{maximize} & \quad \sum_{k \in \mathcal{K}} (\xi_k^0 + \xi_k^e)
\text{subject to} & \quad \text{(11), (23), (35) – (38), (45), (46), (54), (55)}, \\
& \quad \sum_{k \in \mathcal{K}} \left( g_{n,k}(d, q, \tilde{d}, \tilde{q}) \right) \leq C_{\mathcal{F}}^n / B, \forall n \in \mathcal{N}. \quad (56)
\end{align*}
\]

Problem (P2.3) is convex and can be solved with the Dinkelbach algorithm [13, 32].

Let \( A = [w^T, t^T, \xi^T, d^T, q^T]^T \) be a vector stacking the optimization variables of (P2.3), \( A_z = [\tilde{w}_z^T, \tilde{t}_z^T, \tilde{\xi}_z^T, \tilde{d}_z^T, \tilde{q}_z^T]^T \) be the variables that are the optimal solution of problem (P2.3) computed at iteration \( z \) and \( A = [\tilde{w}^T, t^T, \xi^T, d^T, q^T]^T \) be the point, around which the approximations are computed. Moreover, let \( A \) be the convex feasible set of problem (P2.3). To solve the problem in an iterative manner, the algorithm starts by initializing the vector \( A_1 \). More specifically, we first initialize the algorithm with feasible maximum ratio transmission (MRT) beamformers \( \tilde{w}_z \) for the users. Next, \( \tilde{t}_z \), \( \tilde{d}_z \) and \( \tilde{q}_z \) are initialized by replacing the inequalities with equalities in equation (20) and (47)–(50), respectively. Using this initialization, problem (P2.3) in iteration \( z \) can be solved to obtain the vector \( A_z \). If the current solution \( A_z \) is not stationary, it is used for computing \( A_{z+1} \) for the next iteration, i.e., \( A_{z+1} = A_z + \varphi \left( A_z - A_z \right), \) for some \( \varphi \in (0, 1] \), until a stationary solution is found. The detailed steps are outlined in Algorithm 1 below.

**Algorithm 1**: Procedure to determine the optimal beamforming vector \( \omega^* \) of problem (P2.1)

**Input**: \( A_0 \in A, \ z^{\text{max}} \in \mathbb{N} \)

**Initialize**: \( z \leftarrow 0, \ A_0 \leftarrow A_0 \)

**while** \( A_z \) is not a stationary solution of problem (P2.3) and \( z < z^{\text{max}} \) **do**

\( \lambda_z \leftarrow \max_{k \in \mathcal{K}} \{ \xi_k^0 + \xi_k^e \} \)

Solve problem (P2.3) approximated around \( A_{z-1} \) to obtain \( A_z \) as:

\( A_z = \arg \max_{A \in A} \left\{ \sum_{k \in \mathcal{K}} (\xi_k^0 + \xi_k^e) - \lambda_z (P_{\mathcal{F}}(\omega) + P_{\mathcal{C}}^\text{bic}(d, q, d_z, q_z)) \right\} \)

\( A_z \leftarrow A_{z-1} + \varphi \left( A_z - A_{z-1} \right), \) for some \( \varphi \in (0, 1] \)

**end while**

**Output**: \( A_z = [\tilde{w}_z^T, \tilde{t}_z^T, \tilde{\xi}_z^T, \tilde{d}_z^T, \tilde{q}_z^T]^T \)

**C. Phase Shift Design**

Given the vector \( A = [w^T, t^T, \xi^T, d^T, q^T]^T \), which is assumed to be fixed for the duration of optimizing the phase shift vector \( \nu \), Problem (P1) is suitable to be reformulated as a quadratically constrained quadratic programming (QCQP) problem [35]. To this end, problem (P1) can first be

\(2\) In case of infeasibility, we relax the minimum-rate constraints in (23) by introducing a non-negative slack variable, which is highly penalized in the objective function, to the constraints.
reexpressed as the following feasibility detection problem:

\[
\begin{align*}
\text{find} & \quad \mathbf{v} \\
\text{subject to} & \quad \gamma_k^p(v) \geq t_k^p, \quad \forall k \in K, \\
& \quad \gamma_{k,i}^o(v) \geq t_k^o, \quad \forall k \in K, \forall i \in M_k, \\
& \quad |v_p| = 1, \quad \forall r \in \{1, \ldots, R\}
\end{align*}
\]  

(P3.1)

as only the SINR constraints, (24) and (25), and the unit-modulus constraints are dependent on the phase shift vector. However, the feasible set of Problem (P3.1) is still non-convex. In order to obtain a convex representation of the phase-shift-dependent SINR constraints (57) and (58), they can be expressed as

\[
| (h_k + H_k v)^H \omega_k^o |^2 \geq t_k^o \left( \sum_{j \in K \setminus \{k\}} | (h_k + H_k v)^H \omega_j |^2 + \sum_{\ell \in \Phi_k} | (h_k + H_k v)^H \omega_\ell |^2 + \sigma^2 \right), \forall k \in K,
\]

(60)

\[
| (h_i + H_i v)^H \omega_i^c |^2 \geq t_k^c T_i + \sum_{\ell \in \Phi_i} | (h_i + H_i v)^H \omega_i |^2 + \sum_{m \in \Omega_i} | (h_i + H_i v)^H \omega_m |^2), \forall k \in K, \forall i \in M_k,
\]

(61)

with the use of the SINR expressions (20) and (21). By denoting

\[
b_{k,i}^o = h_k^H \omega_i^o, \quad M_{k,j}^o = \left[ \begin{array}{cc} a_{k,j}^o a_{k,j}^o^H \hfill & b_{k,j}^o \\
0 & 0 \end{array} \right], \quad \bar{v} = [v \ h]^T,
\]

(62)

where \( o \in \{p, c\} \) and \( s \) is an auxiliary variable, it holds that if a feasible solution \( \bar{v}^* \) is found, the solution \( v^* \) can be retrieved by \( v^* = [\bar{v}^T, \bar{v}^T, 1:(R+1), \bar{v}^T]^T \), where \( \bar{x} \in [1:R] \) denotes the first \( \ell \) elements of vector \( x \) and \( x_{R+1} \) denotes the \( r \)-th element of vector \( x \) [35]. With the above definitions, we are able to formulate a convex representation of the constraints (60) and (61) by utilizing the matrix lifting technique, namely \( \mathbf{V} = \bar{v} \bar{v}^H \) [36]. This facilitates the reformulation of problem (P3.1) into

\[
\begin{align*}
\text{maximize} & \quad \sum_{k \in K} \zeta_k^p + \zeta_k^c \\
\text{s.t.} & \quad |b_{k,k}^o|^2 + \text{Tr} \left( M_{k,k}^o \mathbf{V} \right) \geq \eta t_k^p \left( \sum_{j \in K \setminus \{k\}} |b_{j,k}^o|^2 + \text{Tr} \left( M_{j,k}^o \mathbf{V} \right) \right) \\
& \quad + \sum_{\ell \in \Phi_k} |b_{\ell,k}^o|^2 + \text{Tr} \left( M_{\ell,k}^o \mathbf{V} \right) + \sigma^2 \right) + \zeta_k^p, \quad \forall k \in K, \\
& \quad |b_{i,i}^c|^2 + \text{Tr} \left( M_{i,i}^c \mathbf{V} \right) \geq \eta t_k^c \left( \sum_{j \in K \setminus \{k\}} |b_{i,j}^c|^2 + \text{Tr} \left( M_{j,i}^c \mathbf{V} \right) \right) + \sigma^2 \\
& \quad + \sum_{\ell \in \Phi_i} |b_{\ell,i}^c|^2 + \text{Tr} \left( M_{\ell,i}^c \mathbf{V} \right) + \sum_{m \in \Omega_i} |b_{i,m}^c|^2 + \text{Tr} \left( M_{i,m}^c \mathbf{V} \right) \\
& \quad + \zeta_k^c, \quad \forall k \in K, \forall i \in M_k, \\
& \quad V_{r,r} = 1, \quad \forall r \in \{1, ..., R+1\}, \\
& \quad V \geq 0, \\
& \quad \text{rank} (\mathbf{V}) = 1.
\end{align*}
\]  

(P3.2)

where \( \zeta_k^p \) and \( \zeta_k^c \) can be understood as SINR residuals of user \( k \) in phase shift optimization [35] and \( \eta \in [0, 1] \) is a parameter, which regulates the objective of the phase shift optimization. More precisely, for \( \eta = 0 \) the resulting phase shifters maximize the sum-path gain, which ignores the users’ interference-and-noise term of the private and common SINRs and therefore comes at the cost of increased interference within the network. The rationale behind this approach is to maximize the potential improvements of the phase shifters on the channels, while subsequently utilizing the dynamic CMD set allocation to cancel the resulting increase in interference. The other extreme case, namely \( \eta = 1 \), finds phase shifters that strictly increase the private and common SINRs at the users. This approach also aids in increasing the EE of the network because increasing the rate at the users results in a higher total rate of the network \( \xi_i \). Alternatively, if the fronthaul is already at full capacity, it also enables the subsequent beamforming optimization to scale the beamformers \( \omega \) down, lowering the power consumption. Problem (P3.2) is still non-convex due to the rank-one constraint (67). However, the current formulation enables solving the problem with the semidefinite relaxation (SDR) technique, in which the rank-one constraint is dropped in order to obtain a matrix optimization problem that can be solved with existing solvers. As the obtained solution is not necessarily rank-one, it only acts as an upper-bound of the unrelaxed problem. For this reason, this work proposes the combination of two approaches to find suitable and high quality rank-one solutions: 1) The Gaussian randomization technique [37] and 2) a difference-of-convex (DC) programming approach, which exploits the fact that the nuclear norm and the spectral norm of a positive semidefinite (PSD) rank-one matrix have the same values [36]. To this end, (67) can be reformulated using the following proposition:

\[
\text{Proposition 1: For a PSD matrix } \mathbf{X} \in \mathbb{C}^{N \times N} \text{ and } \| \mathbf{X} \|_s \geq 0, \text{ it holds [36] that}
\]

\[
\text{rank} (\mathbf{X}) = 1 \iff \| \mathbf{X} \|_s - \| \mathbf{X} \|_2 = 0,
\]

(68)

where \( \| \mathbf{X} \|_s \) denotes the nuclear norm and \( \| \mathbf{X} \|_2 \) denotes the spectral norm of \( \mathbf{X} \).

To tackle the DC form of the reformulated rank-one expression, the first-order Taylor approximation of the spectral norm \( \| \mathbf{V} \|_2 \) around point \( \mathbf{V}^0 \) can be derived as

\[
\| \mathbf{V} \|_2 \geq \| \mathbf{V}^0 \|_2 + \langle \partial_{\mathbf{V}} \| \mathbf{V} \|_2, (\mathbf{V} - \mathbf{V}^0) \rangle,
\]

(69)

where \( \partial_{\mathbf{V}} \) is the subgradient of \( \| \mathbf{V} \|_2 \) with respect to \( \mathbf{V} \) at \( \mathbf{V}^0 \) and where the inner product is defined as \( \langle \mathbf{X}, \mathbf{Y} \rangle = \text{Re} \left\{ \text{Tr} \left( \mathbf{X}^H \mathbf{Y} \right) \right\} \) as stated by Wirtinger’s calculus [38] in the complex domain. Using these definitions, the spectral norm in (68) can be replaced with the approximation given in (69) to obtain a convex approximation of the rank-one constraint. The resulting expression is suitable to be added as a penalty term to problem (P3.2) and results in the following optimization problem

\[
\begin{align*}
\text{maximize} & \quad \rho \left( \sum_{k \in K} \zeta_k^p + \zeta_k^c \right) \\
& \quad - (1 - \rho) \left( \| \mathbf{V} \|_s - \| \mathbf{V}^0 \|_2 - \langle \partial_{\mathbf{V}} \| \mathbf{V} \|_2, (\mathbf{V} - \mathbf{V}^0) \rangle \right) \\
& \quad \text{subject to } \quad (63), (64), (65), (66).
\end{align*}
\]  

(P3.3)

where \( \rho \in [0, 1] \) regulates the trade-off between a high-quality and a rank-one solution. It is worth noting that the subgradient...
\[ \partial \sqrt{\| V \|_2} \] can be efficiently computed by using the following proposition [36].

**Proposition 2:** For a given PSD matrix \( X \in \mathbb{C}^{N \times N} \), the subgradient \( \partial \sqrt{\| X \|_2} \) can be computed as \( e_1 e_1^H \), where \( e_1 \in \mathbb{C}^N \) is the leading eigenvector of matrix \( X \).

Problem (P3.3) is convex and solvable with existing solvers. After obtaining the solution \( V^* \) of problem (P3.3), the Gaussian randomization technique is applied to determine a feasible solution of high quality to problem (P3.1). To this end, the singular value decomposition (SVD) of \( V^* \) is calculated as \( V^* = U \Sigma U^H \), where \( U \in \mathbb{C}(R+1) \times (R+1) \) and \( \Sigma \in \mathbb{C}(R+1) \times (R+1) \) are the unitary matrix and the diagonal matrix, respectively. A rank-one candidate solution \( \tilde{v}_g \) to problem (P3.1) can be generated with the use of the SVD components [39], namely

\[ \tilde{v}_g = U \Sigma^2 z_g \in \mathbb{C}(R+1) \times 1, \] (70)

where \( z_g \sim \mathcal{CN}(0, I_{R+1}) \) denotes a random vector drawn independently from a circularly-symmetric complex Gaussian distribution. After \( G \) randomized solutions are generated, each randomization \( \tilde{v}_g \) results in a potential candidate solution \( v_g \) for problem (P3.1), namely

\[ v_g = \exp \left( j \arg \left( \frac{\tilde{v}_g}{\tilde{v}_g^{R+1}} \right) \right). \] (71)

If a potential candidate solution \( v_g \) satisfies the constraints of problem (P3.1), it is declared to be a feasible candidate solution. The best performing solution out of all feasible candidates, i.e., the one with the highest achievable total rate of the network, is chosen as the final solution \( v^* \) to problem (P3.1). If no feasible solution can be determined with the given amount of Gaussian randomizations \( G \) and \( \eta \neq 1 \), the vector \( u \in \mathbb{C}(R+1) \times 1 \) associated with the leading singular value is adapted by applying equation (71) and chosen as the solution \( v^* \) instead. Algorithm 2 summarizes the necessary steps towards the goal of obtaining a solution to problem (P3.1).

**D. Low-complexity Phase Shift Design**

In this section, we devise a suboptimal but computationally efficient version of the phase shift optimization problem by formulating the problem in a form, which makes it applicable for the same SCA framework, on which the beamforming design is based on. To this end, we show \( \| (\tilde{h}_{k,i}^{oH}(v))^{H} \omega_\kappa^{2} \|^{2} = \| \tilde{h}_{k,i}^{oH}\|^{2} \omega_\kappa^{2} \), where \( \tilde{h}_{k,i}^{oH} = (\omega_\kappa^{oH})^{H} h_i \) and \( H_{i,k}^{oH} = (\omega_\kappa^{oH})^{H} H_i \). Hence, constraint (57) and (58) can be rewritten as

\[ \sum_{m \in \mathcal{K} \backslash \{k\}} | \tilde{h}_{k,m}^{oH} + \tilde{H}_{k,m}^{oH} v^{2} + \sum_{\ell \in \mathcal{F}_k} \tilde{h}_{k,\ell}^{oH} + \tilde{H}_{k,\ell}^{oH} v^{2} + \sigma^2 - \frac{| \tilde{h}_{k,k}^{oH} + \tilde{H}_{k,k}^{oH} v^{2} |^{2}}{\eta \tilde{v}_k^{oH}} \leq 0, \forall k \in \mathcal{K}, \] (72)

\[ \tilde{T}_i + \sum_{\ell \in \Phi_i} \| \tilde{h}_{i,\ell}^{cH} + \tilde{H}_{i,\ell}^{cH} v^{2} + \sum_{m \in \mathcal{C}_k} \| \tilde{h}_{i,m}^{cH} + \tilde{H}_{i,m}^{cH} v^{2} \|^{2} - \frac{| \tilde{h}_{i,k}^{cH} + \tilde{H}_{i,k}^{cH} v^{2} |^{2}}{\eta \tilde{v}_k^{cH}} \leq 0, \forall k \in \mathcal{K}, \forall i \in \mathcal{M}_k, \] (73)

in which \( \tilde{T}_i = \sum_{j \in \mathcal{K}} | \tilde{h}_{j,i}^{cH} + \tilde{H}_{j,i}^{cH} v^{2} + \sigma^2 \). Similarly as for problem (P1), by introducing slack variables, the optimization over \( v \) for given \( \omega \) can be expressed as

\[ \maximize \sum_{v,\{\xi_k,\xi_c\}} \left( \xi_k + \xi_c \right) - 2k \left( \sum_{r=1}^{K} s_r^{(i-1)} \left( v_r - v_r^{(i-1)} \right) \right) \] (P3.4)

subject to (23), (35) – (38), (72), (73),

where (72), (73) denote the first-order Taylor approximations of (72) and (73), respectively. Furthermore, the unit-modulus constraint (59) is added as a penalty term to the objective function [27], where \( k \) is a large positive constant and the superscript \((i-1)\) denotes the value of the variable at the \((i-1)\)-th iteration, while the objective function itself is relaxed to maximize the sum-rate. This is achieved by relaxing the impact of \( P_{\text{AIRC}}(\omega, \xi) \), which results in an approach similar to the one in problem (P3.2), where the energy efficiency is implicitly improved by maximizing the SINR residuals, thus increasing the achievable rates. In addition, the fronthaul constraint (15) is relaxed, in order to find phase shifters, which obtain the highest possible rates. The rationale behind these relaxations is that in contrast to problem (P3.4), problem (P2.3) is able to optimize the relaxed expressions jointly through \( \omega \), resulting in higher quality solutions. Problem (P3.4) is convex and can be solved with the SCA framework similarly to problem (P2.1). It is initialized by the output \( A \) of Algorithm 1.

**E. Determining the CMD sets**

In this work, the RS technique is adopted to mitigate the increased impact of the interference in the network, which is caused by the spatial reuse technique and the IRS. Determining the optimal CMD sets for the scenario at hand is vital for the performance of RS because the common rate \( \xi_c^k \) of user \( k \) is chosen such that its decodable at all users in the corresponding set \( \mathcal{M}_k \). To elaborate, suppose that there is a user \( j \) in the group \( \mathcal{M}_k \) (users decoding the common message of user \( k \)) and the user’s SINR \( \gamma_{j,k} \) of receiving the common message of user \( k \) is notably lower than for the rest of the group. It follows that the resulting transmission of the common message \( s_{k}^{c} \) becomes inefficient. This is because the common rate \( \xi_c^k \) of the group \( \mathcal{M}_k \) decreases to the rate user \( j \) is able to decode \( s_{k}^{c} \), and subsequently reduces the impact of the SIC at all users in \( \mathcal{M}_k \). Specifically for IRS-assisted networks, finding suitable CMD sets becomes a problem as current methods for determining the CMD sets, e.g., with maximum ratio combining (MRC) beamformers [11], would depend on the initial (sub-optimal) IRS phase shifters. Since the phase shifters are subject to change during the optimization process, it becomes inefficient to determine the CMD groups based on the initial phase shifters. Moreover, the optimization of the phase shifters in problem (P3.3) is also dependent on the CMD sets in constraint (64), which increases the difficulty of finding a good combination of effective CMD sets and a phase shift vector that efficiently supports the private and common transmissions of the network.

To bridge this mutual dependency of the phase shift vector optimization and the allocation of CMD sets, we propose
the following method: First, we prioritize the maximization of the achievable sum-path gain in the first few phase shift optimizations, which comes at the cost of potentially increasing the interference at some users. After each phase shift optimization, efficient CMD groups are formed, based on the received interference at the users and the potential achievable common rates within the groups. We implement this idea, by initializing the alternating optimization algorithm with a low-value for the parameter \( \eta \in [0, 1] \), which increases for each successive iteration of the phase shift optimization so that convergence can be assured.

To this end, let \( \tilde{\gamma}_{i,k}^p \) denote the SINR of user \( i \) decoding the private message of user \( k \) as:

\[
\tilde{\gamma}_{i,k}^p(\omega, v) = \sum_{m \in \mathcal{K}_i(k)} \frac{|\langle H_m^p(\omega)v \rangle H_m^p(\omega)v|^2}{\|H_m^p(\omega)v\|^2 + \|H_m^p(\omega)v\|^2 + \sigma^2},
\]

and let \( \tilde{\gamma}_{i,k}^{p+c} = \tilde{\gamma}_{i,k}^p + \tilde{\gamma}_{i,k}^c \) denote the sum-SINR of user \( i \) decoding the private and common message of user \( k \). Moreover, let \( \Gamma_k^p, \forall \ell \in \{p, c\} \) denote a vector, whose entries represent the ratio of each user decoding the private/common message of user \( k \) and user \( k \) decoding his private/common message:

\[
\Gamma_k^p = \begin{bmatrix} \tilde{\gamma}_{i,k}^p(\omega, v) & 1, \ldots, \tilde{\gamma}_{j,k}^p(\omega, v) & 1 \end{bmatrix}^T, \quad \Gamma_k^c = \begin{bmatrix} \tilde{\gamma}_{i,k}^c(\omega, v) & 1, \ldots, \tilde{\gamma}_{j,k}^c(\omega, v) & 1 \end{bmatrix}^T,
\]

and let \( \Gamma^o = [\Gamma_1^p, \ldots, \Gamma_K^p] \). Note that each positive non-zero value in \( \Gamma_k^p \) represents a user, which is able to decode the respective message of user \( k \) with a higher rate than user \( k \) itself. The algorithm starts by calculating \( \Gamma^o \) given the beamforming vectors \( \omega \) and phase shift vector \( v \). To ignore the users \( j \in \mathcal{M}_k \), which are already decoding the common message of user \( k \), the corresponding \( (j, k) \)-th entries in \( \Gamma^o \) are replaced with an arbitrarily low value, i.e., \(-\infty\). Next, the highest value \( \Gamma_{j,k}^o \) of \( \Gamma^o \) is determined. Should this \( (j, k) \)-th entry \( \Gamma_{j,k}^o \) be above \( \epsilon_{CMD} \in [-1, \infty] \) then user \( j \) is considered as an addition to the group \( \mathcal{M}_k \), if the number of decoding layers \( D \) of user \( j \) is not at full capacity, i.e., \( |\mathcal{F}_j| < D \). The effectiveness of this procedure is verified by temporarily adding the user to the respective CMD sets in \( \mathcal{S} \), resulting in \( \mathcal{S}_o \). In order to obtain high quality sets for \( \{\Omega_{j,k}\}_{j \in \mathcal{K}} \) in \( \mathcal{S}_o \), the decoding order \( \pi_j \) of user \( j \) is updated according to the descending order of the sum-SINRs of user \( j \) decoding the private and common message of all other users as follows:

\[
\gamma_{i,\pi_j(1)}^p(\omega, v) \geq \cdots \geq \gamma_{i,\pi_j(2)}^p(\omega, v) \geq \gamma_{i,\pi_j(1)}^c(\omega, v).
\]

Should the \( (j, k) \)-th entry \( \Gamma_{j,k}^o \) be part of \( \Gamma^p \), i.e., \( o \in \{p\} \), the beamforming vector \( \omega_{j,k} \), which is associated with the common message of user \( k \), is set as \( \omega_{j,k}^p = \omega_{j,k}^c + \omega_{j,k}^p \). Should the resulting value \( \gamma_{j,k}^o(\omega, v) / \gamma_{j,k}^o(\omega, v) - 1 \), which is determined with the temporary CMD sets \( \mathcal{S}_o \), still be above a certain threshold \( \epsilon_{Thr} \in [-1, \infty] \), the temporary CMD sets \( \mathcal{S}_o \) are adopted and the corresponding values in \( \mathcal{A} \) updated. Otherwise, the changes to \( \omega_{j,k}^c \) are discarded and the corresponding entry of \( \Gamma^o \) is set to \(-\infty\). This process is repeated until there are no values in \( \Gamma^o \) above \( \epsilon_{CMD} \). The detailed steps for this procedure are outlined in Algorithm 3.

The proposed alternating optimization algorithm for solving problem (P1) is outlined in Algorithm 4, where problems (P2.3) and either (P3.3) or (P3.4) are solved alternatively, while the objective of the first few outer-loop iterations is to find suitable CMD sets by prioritizing the sum-path gain in the phase shift optimization. We initialize the algorithm with random phase shifters \( \psi_0 \) and set \( \{\phi_k = \{k\}\}_{k \in \mathcal{K}} \), from which \( \mathcal{S}_0 \) can be inferred.

### F. Complexity Analysis

The proposed procedure to determine \( \tilde{A}^* \), \( v^* \) and \( S^* \) is split into three stages, each of which consists of applying Algorithm 1, 2 or 3, respectively. The complexity of the proposed Algorithms is described in the following.

At each iteration of Algorithm 1 the convex problem (P2.3) needs to be solved. Since the objective function is solved with the Dinkelbach approach, the problem can be easily cast as a second-order cone program (SOCP), see [40] and references therein. Thus, the total complexity of the SCA method for solving problem (P2.3) is \( O((NK)^{3.5}) \), when using general purpose solvers, e.g., MOSEK. Similarly, each iteration of the phase shifting optimization problem (P3.4) using the SCA framework is given as \( O((RK)^{3.5}) \) using similar arguments as above. In contrast, the computational cost for using the SDP approach for solving the phase shift optimization problem is \( O(R^6) \). Regarding Algorithm 3, the overall complexity is \( O(K^3 N LR) \), characterized by the calculation of \( \Gamma^o \) and \( \Gamma^c \). As a result, the overall complexity of Algorithm 4 is \( O(S_{Max}(Z_{Max}(NK))^3 + R^6 + K^3 N LR)) \), when utilizing the SDP approach and \( O(S_{Max}(Z_{Max}(NK)^3 + (RK)^{3.5} + K^3 N LR)) \), when utilizing the SCA approach, where \( S_{Max} \) and \( Z_{Max} \) are the worst-case fixed number of iterations needed for Algorithm 4 and the SCA problems, respectively.

### V. Numerical Simulations

For the simulation results we consider a C-RAN consisting of one CP, which is connected to 3 RRUs via finite capacity
fronthaul links. Each RRU is equipped with \( L = 2 \) antennas with \( \frac{\lambda}{2} \) spacing, where \( \lambda \) is the wavelength of the carrier frequency \( f = 2 \) GHz, and serving 6 single-antenna users in the area of operation, sized as \([-500,500] \times [-500,500] \) m\(^2\). For simplicity of the analysis, we assume that all users require a QoS \( \xi_{\text{Min}} = 3 \) Mbps, \( \forall k \in \mathcal{K} \) and set the PA efficiency to \( \eta_{\text{PA}} = 1 \).

For the allocation of the fronthaul capacities, we distinguish between two regions: 1) the partially-connected regime and 2) the fully-connected regime. In the partially-connected regime, where the available fronthaul capacity at each RRU is insufficient to serve all users from each individual RRU. For this reason, the allocation of the fronthaul capacities is determined based on the assumption that \( \xi_{\text{Min}} \) is an integer divisor of \( C_{\text{PH}} \). The rationale behind this approach is to enable the maximum number of CoMP transmissions within the network, given the available resources, because the utilization of this strategy is able to increase the EE of the network substantially in this available resources, because the utilization of this strategy is able to increase the EE of the network substantially in this regime.

To better illustrate the impact of the CoMP transmissions, we introduce the level of supportive connectivity (LoSC) as a metric, which captures the number of RRU-to-user links within the network. Note that due to the non-zero QoS constraints, each user is required to be served by at least one RRU. At the transition point (TP), \( C_{\text{TP}} = NK \xi_{\text{Min}} \) (here \( C_{\text{TP}} = 54 \) Mbps), however, the available resources becomes sufficient to allocate enough resources to each RRU such that each individual RRU is able to serve every user with \( \xi_{\text{Min}} \). It is therefore at the TP, where the transition into the fully-connected regime takes place, in which we assume a symmetric allocation of the fronthaul capacities, namely \( C_{\text{PH}} = C_{\text{PH}} / N, \forall n \in \mathcal{N} \).

To aid the RRUs in their transmissions, an IRS is assumed to be deployed in the center of the area, which consists of \( R = 15 \) reflect elements. The users and RRUs are positioned uniformly and independently within the area of operation. We deploy the 3GPP LTE model from [41], where the channel coefficients between the RRUs, users and IRS are subject to the path-loss model consisting of three components: 1) path-loss as \( PL_{x,y} = 148.1 + 37.6 \log_{10}(d_{x,y}) \) dBm, where \( d_{x,y} \) is the distance between device \( x \) and device \( y \) in km; 2) log-normal shadowing with 8 dB standard deviation and 3) Rayleigh channel fading with zero-mean and unit-variance. The channel bandwidth is set to \( B = 10 \) MHz and the noise power spectrum is set to \(-169\) dBm/Hz. Furthermore, the number of Gaussian randomizations is set to \( G = 25 \), the smoothness parameter to \( \Lambda = 10^{-6} \), the SCA step size is set to \( \epsilon_{\text{s}} = 0.9 \), \( \forall z \), the maximum amount of SCA iterations to \( Z_{\text{Max}} = 6 \) and the trade-off parameter is set to \( \rho = 0.9 \), which facilitate a high detection rate for feasible rank-one solutions when solving problem (P3.3). The maximum number of successive decoding layers at the users are set to \( D = K \), as the number of layers and the decoding order at each user is determined dynamically (we choose \( D = 2 \) for the baseline static CMD set selection). The dynamic CMD set thresholds are set to \( \epsilon_{\text{CMD}} = -0.4 \) and \( \epsilon_{\text{Br}} = -0.5 \), which means that the users, whose SINR of potentially decoding another user’s message are at least 60% of the current SINR of decoding the message at the intended user, are considered candidates to this user’s CMD set and are accepted, if their SINR is still at 50% when the SIC order at this user is considered. We assume the maximum transmit power per RRU \( P_{\text{Max}} \) to be 35 dBm, the signal processing circuitry of the network is set to \( P_{\text{circ}} = 37 \) dBm. Moreover we assume \( P_{\text{IRS}} = 10 \) dBm and \( P_{\text{RF}} = 0.3 \) W/Mbps. In addition to the dynamic clustering algorithm, we also consider a static clustering algorithm [16, Algorithm 3] as baseline clustering approach, which can also be extended accordingly to support the RS scenario [11].

### A. Impact of the dynamic clustering on the performance

First the performance of the dynamic clustering is evaluated against the static clustering in a setup with and without an IRS, respectively. Figure 2 depicts the EE of the network as...
Figure 2: The EE of all studied schemes, where the prefix “d-” represents the dynamic and the prefix “s-” the static user clustering scheme. The horizontal lines represent the equivalent broadcast channel scenarios of s-RS+IRS and d-RS+IRS, in which the fronthaul capacity is unlimited. Note, that the d-RS+IRS scheme additionally employs the dynamic CMD set allocation. The dynamic CMD set allocation obtains sets, which enable synergistic benefits between RS and the IRS.

B. Impact of the dynamic CMD set allocation on the performance

As we move towards the higher capacity regions, we note that the gains of the dynamic user clustering in Figure 2 decrease, while the gains of the RS-assisted schemes improve. Especially in the interference-limited regime, the gains of the RS technique become more pronounced. This is expected as the RS scheme is specifically designed to mitigate interference, hence the increase in its performance. Interestingly, the importance of the dynamic clustering becomes less significant at higher available fronthaul capacities. This causes the curves, which represent the RS-enabled non-IRS scenarios with the statically-clustered (s-RS) and dynamic-clustered user allocations (d-RS), to converge to the same value. In other words, the performance gain of RS over TIN at higher fronthaul capacities is solely dependent on the quality of the CMD sets. The figure shows two important facts for the IRS-assisted scenarios in that regard: 1) the curve, representing the static CMD set allocation (s-RS+IRS) does not converge to the same value as the curve, representing the dynamic CMD set allocation (d-RS+IRS), but in fact converges at a much lower value; 2) the gain of the dynamic CMD sets over the TIN case is higher in the IRS-assisted scenario than in the non-IRS assisted scenario. In essence, by using the dynamic CMD sets allocation, the algorithm is able to determine sets, which are of higher quality and more importantly, interacting synergistically with the IRS (d-RS+IRS), as opposed to the statically determined CMD sets (s-RS+IRS), which only show a marginal improvement over the respective TIN scenario (d-TIN+IRS). This synergistic behaviour is highlighted by arrows on the right-hand side of Figure 2, which illustrate the gains of the statically and dynamically determined CMD sets. We can see that the arrow representing the gain of the dynamic CMD set allocation is larger than the arrow of the statically allocated counterpart, indicating a synergistic interaction between RS with the dynamic CMD sets and the IRS. In fact, if the fronthaul capacity goes to infinity (equivalent to the broadcast channel), the gain between the d-RS+IRS scheme gets even more pronounced if compared to the s-RS+IRS scheme.
C. The Role of Rate Splitting

To illustrate the impact of RS and the selected CMD sets, we plot the achievable sum rate of the common rate $\xi_c = \sum_{k \in K} \xi_k^c$ in proportion to the total achievable rate $\xi^s$ of the network in Figure 5. Note, that the impact of the dynamic user clustering decreases in the high fronthaul capacity region, which causes the s-RS and d-RS curves to converge to the same value. However, by comparing the s-RS curve with the d-RS curve in the low fronthaul capacity region, we can see a decrease in the proportion of the common rates within the network. This implies that the dynamic user clustering algorithm mainly detects the RRU-to-user allocations of the common message transmissions as inefficient and removes the corresponding RRU-to-user links. Moreover, the figure also shows that the proportion of the common rate is decreasing, as the fronthaul capacity increases towards the TP for all schemes. This signifies, that using CoMP transmissions to reduce the required transmit power within a fronthaul-limited network is more efficient than the utilization of RS in this regime, which fits with the fact that RS performs best in the interference-limited regime. In opposition to the d-RS scheme, the d-RS+IRS scheme counteracts the removal of inefficient common-rate links by finding sensible CMD sets dynamically.

The more efficient sets result in an overall higher common rate percentage, especially in the low fronthaul regime, if compared to the s-RS+IRS case and also causes the convergence to a higher value, representing the higher quality of the determined sets. Interestingly, the curves, which represent the dynamic user allocation schemes, are characterized by a peak that begins right after the TP. This peak is a result of the interaction between the two techniques that are employed in these scenarios, namely RS and the dynamic user clustering.

In fact, as the fronthaul increases after the TP, the dynamic clustering gain decreases, while the effectiveness of the rate splitting increases. This leads to an intersection, around which both techniques are able to interact with each other. Therefore, it is also around this intersection, where the user allocations, representing the common rate links, get more efficient than the respective private allocations, while the clustering algorithm is still able to remove these inefficient private links. To illustrate this behaviour, we plot the normalized gains of the individual techniques as a function of the fronthaul capacity in Figure 3. The figure shows that the intersection, at which the gains of the individual techniques meet, overlaps with the maximum of the respective peak in Figure 5. It is also at this point, where the growth of the energy efficiency of the system (see Figure 2) for the d-RS+IRS scenario increases significantly. Interestingly, this synergistic interaction between these techniques requires the IRS to be optimized, i.e., adjusted to engage favourably with the other techniques. This can be deduced by comparing the d-RS+IRS scheme with its phase-unoptimized version in Figure 2 because it does not display the characteristic behaviour discussed above.

D. Outages and Scalability

Figure 6 compares the EE and outage performance difference of the d-RS+IRS scheme between solving the phase shifting problem with the semidefinite programming (SDP) (P3.1) or the SCA approach (P3.4). Overall, the SCA approach is able to generate results with similar performance to the SDP approach regarding the EE, while exhibiting overall lower computational complexity. The approximations made in both approaches, cause them to converge to locally optimal solutions making their performance dependent on the initialized phase shifting vector, which is chosen randomly. However, the approximations of the SCA approach exhibit a worse outage percentage compared to the SDP approach, especially in the partially-connected regime, where the quality of the acquired phase shifters has a higher impact on achieving the rate requirements because of the scarcity of other resources.

In order to study the scalability of the proposed algorithms, we utilize the low-complexity SCA approach in Figure 7, which depicts the dynamic (d-)schemes for different system configurations with $L = 8$ antennas per RRU and unlimited fronthaul capacities. The figure shows that adding users to each configuration of the network causes the EE to rise up to a peak at a specific amount of users, after which the EE decreases. Increasing the number of reflect elements deployed in the network not only increases the overall EE but also shifts these peaks to a higher quantity of users. It appears, that utilizing $L = 8$ antennas enables the algorithms to find active and passive beamformers that effectively mitigate interference between users up to the EE-peaks, around which the network’s performance becomes interference-limited and enables the RS approaches to start outperforming the respective TIN schemes.

Interestingly, the performance gain for the d-RS scheme (i.e., $R = 0$) in these configurations becomes negligible compared to the gains that RS is able to achieve in the d-RS+IRS schemes, i.e., when an IRS is present in the network, which indicates that the synergistic benefits between RS and the IRS are also present in large scale networks.

E. Impact of Higher Losses

In order to study the impact on the EE when higher losses are introduced to the system, we define a more “lossy” scenario by lowering the PA efficiency to $\eta_{PA}^{-1} = 1.2$ and introducing a 1 dB power loss due to reflection at the IRS [42], which translates into setting the reflect amplitudes to
Figure 4: The decrease in the level of supportive connectivity (LoSC) as a function of the fronthaul capacity. The algorithm is able to discard more RRU-to-user links in the IRS-assisted scenario, causing a synergistic effect between the dynamic user clustering and the IRS.

Figure 5: The proportion of the common rate $\xi_c$ in the total achievable network rate $\xi$. Both, the d-RS+IRS and the s-RS+IRS case display a peak after the fronthaul capacity overcomes the TP (at $C_FH = 54$), which is caused by the interaction between the RS and dynamic user allocation technique.

Figure 6: The EE and outage percentages of the d-RS+IRS scheme when optimizing the phase shifters $v$ with the SDP approach (P3.1) or the SCA approach (P3.4).

Figure 7: The EE for the dynamic (d-)schemes with $L = 8$ and unlimited fronthaul capacity for different system configurations using the SCA approach (P3.4).

Figure 9: Impact of the channel estimation error on the proposed algorithms for different schemes.

The synergistic benefits decrease as well because the overall interference in the network is reduced. In Figure 7, the “lossy” scenario depicts an overall reduction to the EE of the network, where a similar behaviour for both dynamic schemes can be observed when comparing the shape of the curves to the lossless scenarios.

**F. Convergence Performance and Impact of Imperfect CSI**

To investigate the impact on the convergence performance for different configurations, Figure 8 compares the convergence behaviour of Algorithm 4 for the d-RS+IRS schemes for different setups. It is shown, that the algorithm converges for both, SCA and SDP approaches after around 15 iterations for small-scaled systems. For large-scaled systems the figure shows distinct jumps at the 7-th and 13-th iteration of Algorithm 1, representing the improvements induced by the other algorithms due to $Z_{\text{Max}} = 6$.

In order to study the influence of the channel estimation errors, we replace the cascaded channels $h_{k}^{\text{eff}}(v)$ with their estimates, i.e., $\hat{h}_{k}^{\text{eff}}(v) = h_k + \hat{H}_k v$. To this end, we assume $\sigma_{\text{RRU}} = -100 \text{ dBm}$ and $P_C = 1\text{ W}$. Moreover, the channel coherence time is set to $\tau = 0.05\text{s}$, while the length of the
sub-phases is assumed to be $\tau_S = 50K \mu s$. The number of sub-phases is assumed to be $S = R + 1$, resulting in $S = 1$ if there is no IRS present in the network. To account for the rate loss due to the time needed to train the channels, we define the net achievable rate of user $k$ as

$$
\begin{align*}
\gamma^\text{net}_k \equiv \left(1 - \frac{S\tau_S}{\tau}\right) \left\{ \log_2 \left(1 + \gamma^p_k (\omega, v) \right) + B \min_{i \in M_k} \left\{ \log_2 \left(1 + \tilde{\gamma}^c_{i,k} (\omega, v) \right) \right\} \right\},
\end{align*}
$$

(77)

where $\gamma^p_k (\omega, v)$ and $\tilde{\gamma}^c_{i,k} (\omega, v)$ denote the replacement of $h^\text{eff}_k (v)$ with $h^p_k (v)$ in (20) and (21), respectively. The results are plotted in Figure 9 and depict the achievable EE with and without channel estimation errors. The figure shows that with an increasing fronthaul capacity, the EE loss also rises. This behaviour can be explained with the help of Figure 4, which shows that the LoSC in the fully-connected regime is lower, for smaller fronthaul capacities. This results in a lower amount of channels that are used, consequently inducing an overall smaller estimation error. Similarly, the additional channels in the IRS-assisted scenarios not only increase the rate loss due to a longer estimation period $\tau = S\tau_S$ but also increase the overall estimation error, causing a higher overall EE loss. However, even with the increased EE loss for the IRS-assisted scenarios, utilizing the IRS still outperforms the non-IRS scenarios, even when perfect CSI is available. Moreover, the figure illustrates that even with imperfect CSI conditions, the synergistic benefits between RS and IRS remain present.

### VI. Conclusion

Modern communication networks are forecast to require an increased amount of transmit power as more users and RRUUs connect to the network, causing an increase in interference. Motivated by these predictions, this work studies the deployment of a passive IRS in a C-RAN with the aim to increase the energy efficiency. To mitigate the additional increase in interference, caused by spatial reuse and the IRS, the impact of the interaction between the RS scheme and the IRS on the EE, under the assumption of finite fronthaul capacities, is studied. An optimization problem is formulated, which includes a dynamic user clustering and a dynamic CMD set allocation, in order to take the changes of the phase shifts throughout the optimization process into account. As the optimization variables are highly dependent on each other, the problem is split into three subproblems, which are solved in an alternating fashion. A strategy is devised, in which the outer loop is able to influence the solutions of the individual subproblems by regulating outer-loop parameters, resulting in high quality solutions due to cooperation between the techniques.

Numerical results show that the deployment of the IRS in combination with the dynamically-determined CMD sets displays a synergistic benefit with regards to the EE, if used together with the RS technique. In addition to that, the dynamic clustering is also displaying a synergistic interaction with the IRS, which becomes evident by the increased the individual gain of the dynamic clustering from 71% to 88%, resulting in a total gain of up to 125% over the baseline approach. The results, moreover, show that with an increasing available fronthaul capacity, the gain of the dynamic clustering decreases, while the gain of RS simultaneously increases. Around the resulting intersection, the growth of the energy efficiency of the network increases notably as both, the dynamic user clustering and the dynamic CMD set allocation technique are able to interact with each other. The proposed methods offer an attractive upgrade path to existing networks as these can be easily and cost efficiently extended by deploying an IRS in a RS-enabled C-RAN to realize future communication networks.

### References

[1] K. Weinberger, A. A. Ahmad, and A. Sezgin, “On synergistic benefits of rate splitting in IRS-assisted cloud radio access networks,” in *IEEE ICC*, 2021, pp. 1–6.

[2] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, “A survey of energy-efficient techniques for 5G networks and challenges ahead,” *IEEE Jrn. on Sel. A. in Com.*, vol. 34, no. 4, pp. 697–709, 2016.

[3] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, “Asymptotic max-min SINR analysis of reconﬁgurable intelligent surface assisted MISO systems,” *IEEE Tr. on Wirel. Com.*, vol. 19, no. 12, pp. 7748–7764, 2020.

[4] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable intelligent surfaces for energy efﬁciency in wireless communication,” *IEEE Tr. on Wirel. Com.*, vol. 18, no. 8, pp. 4157–4170, 2019.

[5] K. Ntontin, M. Di Renzo, and F. Lazarakis, “On the rate and energy efﬁciency comparison of reconﬁgurable intelligent surfaces with relays,” in *IEEE SPAWC*, 2020, pp. 1–5.

[6] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, “Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities,” *IEEE Tr. on Cogn. Commun. and Netw.*, vol. 6, no. 3, pp. 990–1002, 2020.

[7] M. Jung, W. Saud, M. Debbah, and C. S. Hong, “On the optimality of reconﬁgurable intelligent surfaces (RISs): Passive beamforming, modulation, and resource allocation,” *IEEE Tr. on Wirel. Com.*, vol. 20, no. 7, pp. 4347–4363, 2021.

[8] T. Han and K. Kobayashi, “A new achievable rate region for the interference channel,” *IEEE Tr. on IT*, vol. 27, no. 1, pp. 49–60, 1981.

[9] A. Chaaban and A. Sezgin, “The approximate capacity region of the symmetric k-user Gaussian interference channel with strong interference,” *IEEE Tr. on IT*, vol. 62, no. 5, pp. 2592–2621, 2016.

[10] T. Gou and S. A. Jafar, “Capacity of a class of symmetric SIMO Gaussian interference channels within O(1),” in *IEEE Intern. Symp. on Info. Theory*, 2009, pp. 1924–1928.

[11] A. A. Ahmad, H. Dahrouj, A. Chaaban, A. Sezgin, and M. Alouini, “Interference mitigation via rate-splitting and common message decoding in cloud radio access networks,” *IEEE Acc.*, pp. 80 350–80 365, 2019.

[12] E. Che, H. D. Tuan, H. H. M. Tam, and H. H. Nguyen, “Successive interference mitigation in multiuser MIMO channels,” *IEEE Tr. on Com.*, vol. 63, no. 6, pp. 2185–2199, 2015.

[13] A. A. Ahmad, B. Matthiesen, A. Sezgin, and E. Jorswieck, “Energy efﬁciency in C-RAN using rate splitting and common message decoding,” in *IEEE ICCC*, 2020, pp. 1–6.

[14] M. Charafeddine, A. Sezgin, Z. Han, and A. Paulraj, “Achievable and crystallized rate regions of the interference channel with interference as noise,” *IEEE Tr. on Wirel. Com.*, vol. 11, 11 2011.

[15] B. Bandemer, A. Sezgin, and A. Paulraj, “On the noisy interference regime of the MISO Gaussian interference channel,” in *42nd Asilomar Conf. on Sign., Sys. and Comp.*, 2008, pp. 1098–1102.

[16] B. Dai and W. Yu, “Sparse beamforming and user-centric clustering for downlink cloud radio access network,” *IEEE Acc.*, vol. 2, pp. 1326–1339, 2014.

[17] G. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, “Multi-cell MIMO cooperative networks: A new look at interference,” *IEEE Jrn. on Sel. A. in Com.*, vol. 28, no. 9, pp. 1380–1408, 2010.

[18] P. Marsch and G. Fettweis, “On downlink network MIMO under a constrained backhaul and imperfect channel knowledge,” 01 2010, pp. 1–6.

[19] Y. Shi, J. Zhang, and K. B. Letaief, “Group sparse beamforming for green Cloud-RAN,” *IEEE Tr. on Wirel. Com.*, vol. 13, no. 5, pp. 2809–2823, 2014.
[20] A. Alameer and A. Sezgin, “Joint beamforming and network topology optimization of green cloud radio access networks,” in 9th ISTC, 2016, pp. 375–379.

[21] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, “Rate splitting multiple access in C-RAN,” in IEEE PIIMC, 2020, pp. 1–6.

[22] A. A. Ahmad, J. Kakar, H. Dahrouj, A. Chaaban, K. Shen, A. Sezgin, T. Y. Al-Naffouri, and M.-S. Alouini, “Rate splitting and common message decoding for MIMO C-RAN systems,” in IEEE SPAWC, 2019, pp. 1–5.

[23] A. Alameer Ahmad, H. Dahrouj, A. Chaaban, and M. Alouini, “Interference mitigation via rate-splitting in collision cloud radio access networks,” in IEEE WCNC, 2018, pp. 1–5.

[24] A. A. Ahmad, H. Dahrouj, A. Sezgin, T. Y. Al-Naffouri, and M.-S. Alouini, “Power minimization via rate splitting in downlink cloud-radio access networks,” in IEEE ICC Workshops, 2020, pp. 1–6.

[25] D. Yu, J. Kim, and S.-H. Park, “An efficient rate-splitting multiple access scheme for the downlink of C-RAN systems,” IEEE Wirel. Com. Letters, vol. 8, no. 6, pp. 1555–1558, 2019.

[26] A. A. Ahmad, B. Matthiesen, A. Sezgin, and E. Jorswieck, “Energy efficiency in C-RAN using rate splitting and common message decoding,” in IEEE ICC Workshops, 2020, pp. 1–6.

[27] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, “Energy efficient rate splitting multiple access RSMA with reconfigurable intelligent surface,” in IEEE ICC Workshops, 2020, pp. 1–6.

[28] Q. Li, I. F. A. Nadeem in 2019, C. Zhou, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, “Intelligent reflecting surface-assisted multiuser MISO communication: Channel estimation and beamforming design,” IEEE IOCOM, vol. 1, pp. 661–680, 2020.

[29] Y. Mao, B. Clerckx, and V. Li, “Rate-splitting multiple access for downlink communication systems: Bridging, generalizing and outperforming SDMA and NOMA,” EURASIP J. on Wirel. Com. and Netw., 2018.

[30] R. Jiao, X. Wen, Z. Lu, Y. Chen, H. Shao, and W. Jing, “Dynamic user-centric clustering algorithm based on energy efficiency in Cloud-RAN,” in 24th ICT, 2017, pp. 1–7.

[31] M. Tao, E. Chen, H. Zhou, and W. Yu, “Content-centric sparse multicast beamforming for cache-enabled cloud RAN,” IEEE Tr. on Wirel. Com., vol. 15, no. 9, pp. 6118–6131, 2016.

[32] W. Dinkelbach, “On nonlinear fractional programming,” Management Science, vol. 13, no. 7, pp. 492–498, 1967.

[33] A. Zappone, E. Björnson, L. Sanguineti, and E. Jorswieck, “Globally optimal energy-efficient power control and receiver design in wireless networks,” IEEE Tr. on SP, vol. 65, no. 11, pp. 2444–2459, 2017.

[34] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, “Secure and energy-efficient beamforming for simultaneous information and energy transfer,” IEEE Tr. on Wirel. Com., vol. 16, no. 11, pp. 7523–7537, 2017.

[35] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” IEEE Tr. on Wirel. Com., vol. 18, no. 11, pp. 5394–5409, 2019.

[36] K. Yang, T. Jiang, Y. Shi, and Z. Ding, “Federated learning via over-the-air computation,” IEEE Tr. on Wirel. Com., vol. 19, no. 3, pp. 2022–2035, 2020.

[37] W. K. Ma, T. N. Davidson, K. M. Wong, Z. Luo, and P. Ching, “Quasi-maximum-likelihood multuser detection using semi-definite relaxation,” IEEE Tr. on SP, vol. 50, no. 4, pp. 912–922, 2002.

[38] J. Dong and Y. Shi, “Nonconvex demixing from bilinear measurements,” IEEE Tr. on SP, vol. 66, no. 19, pp. 5152–5166, 2018.

[39] B. Lyu, D. T. Hoang, S. Gong, and Z. Yang, “Intelligent reflecting surface assisted wireless powered communication networks,” in IEEE WCNCW, 2020, pp. 1–6.

[40] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, “Applications of second-order cone programming,” Linear Algebra and its Applications, vol. 284, no. 1, pp. 193–228, 1998.

[41] Further Advancements for E-UTRA Physical Layer Aspects (Release 9), document TS 36.814, 3GPP, Mar. 2010.

[42] D. Ma, M. Ding, and M. Hassan, “Enhancing cellular communications for UAVs via intelligent reflective surface,” in IEEE WCNC, 2020, pp. 1–6.

Kevin Weinberger received his B.Sc., in 2017, and M.Sc. degree, in 2020, in electrical engineering and information technology both from the Ruhr Universität Bochum, Germany. He is currently pursuing the Ph.D. degree with the Institute of Digital Communication Systems, Ruhr-Universität Bochum. His research interests include optimization and signal processing in wireless networks with focus on reconfigurable intelligent surfaces and rate splitting.

Alaa Alameer Ahmad received his B.Sc. degree in electrical engineering from the Higher Institute of Applied Sciences and Technology (Hiast), Damascus, Syria in 2008, his M.Sc. degree in information and communication technology from TU-Darmstadt, Darmstadt, Germany, in 2015, and his Ph.D. in wireless communication from Ruhr University Bochum. He received the AKDN scholarship for the year 2011-2012. He is currently working at Volkswagen Infotainment in Bochum, Germany. His research interests include optimization of wireless communication systems, signal processing in communication, and machine learning applications for resource management.

Aydin Sezgin (S’01-M’05–SM’13) received the Dipl.Ing. (M.S.) degree in communications engineering from Technische Fachhochschule Berlin (TFH), Berlin, in 2000, and the Dr. Ing. (Ph.D.) degree in electrical engineering from TU Berlin, in 2005. From 2001 to 2006, he was with the Heinrich-Heine-University, Dusseldorf. From 2006 to 2008, he held a postdoctoral position, and was also a lecturer with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA, USA. From 2008 to 2009, he held a postdoctoral position with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA. From 2009 to 2011, he was the Head of the Emmy-Noether- Research Group on Wireless Networks, Ulm University. In 2011, he joined TU Darmstadt, Germany, as a professor. He is currently a professor of information systems and sciences with the Department of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, Germany. He is interested in signal processing, communication, and information theory, with a focus on wireless networks. He has published several book chapters more than 60 journals and 180 conference papers in these topics. He has coauthored a book on multi-way communications. Aydin is a winner of the ITG-Sponsorship Award, in 2006. He was a first recipient of the prestigious Emmy-Noether Grant by the German Research Foundation in communication engineering, in 2009. He has coauthored papers that received the Best Paper Award at the IEEE Communication Theory Workshop, in 2011, the Best Paper Award at ICCSPA, in 2015, and the Best Paper Award at ICC, in 2019. He has served as an Associate Editor for the IEEE Transactions on Wireless Communications (2009-2014), and as an area editor for the Elsevier Journal of Electronics and Communications (2010-2011). He was also the General Co-chair of the 2018 International ITG Workshop on Smart Antennas, the program co-chair of the 2018 IEEE Signal Processing Letters Conference, and the workshop co-chair of the 2022 WCNC workshop on rate-splitting and next generation multiple access.

Alessio Zappone (SM’16) received his M.Sc. and Ph.D. both from the University of Cassino and Southern Lazio (Cassino, Italy). In 2012 he has worked with the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT) in the framework of the FP7 EU-funded project TREND. From 2012 to 2016 he has been with the Dresden University of Technology, managing the project CEMRIN on energy-efficient resource allocation in wireless networks, funded by the German research foundation (DFG). In 2017 he was the recipient of the H2020 Marie Curie IF BEMSART fellowship for experienced researchers, carried out at the LANEAS group of the CentraleSupelec (Gif-sur-Yvette, France). Since 2019, he is a tenured professor at the University of Cassino and Southern Lazio. His research interests lie in the area of communication theory and signal processing, with main focus on optimization techniques for resource allocation and energy efficiency maximization. For his research he received the Marconi Award of the IEEE Communications Society in 2021. Alessio serves as senior area editor for the IEEE SIGNAL PROCESSING LETTERS and has served as guest editor for the IEEE JOURNAL ON SELECTED AREAS ON COMMUNICATIONS (Special Issues on Energy-Efficient Techniques for 5G Wireless Communication Systems and on Wireless Networks Empowered by Reconfigurable Intelligent Surfaces).