Effective Design of Multi-User Reception and Fronthaul Rate Allocation in 5G Cloud RAN

Dora Boviz, Chung Shue Chen, Sheng Yang

Effective Design of Multi-User Reception and Fronthaul Rate Allocation in 5G Cloud RAN

Dora Boviz, *Student Member, IEEE,* Chung Shue Chen, *Senior Member, IEEE,* and Sheng Yang, *Member, IEEE*

***Abstract*—Cloud Radio Access Network (C-RAN) is becoming more than ever timely in 5G for dense network deployments, but its design has to be adapted to 5G requirements. To ensure high uplink throughput in cell-edge regions affected by inter-cell interference, C-RAN enables multi-user reception techniques among cells. In this paper, we propose an efficient end-to-end uplink transmission scheme dealing with implementation and deployment constraints on both communication interfaces in C-RAN, i.e., the wireless one between the users and the Remote Radio Heads (RRHs), and the fronthaul links between the RRHs and the central processing unit. Multi-cell non-orthogonal multiple access (NOMA) can improve spectral efficiency and cell-edge throughput, but its design needs to fit requirements regarding receiver complexity and delay. Optimizing the fronthaul rate allocation to maximize the benefit of the transmissions allows to exploit the fronthaul links effectively. Our model considers the throughput of NOMA transmissions in C-RAN and the expense related to fronthaul usage in various deployment scenarios. When the available fronthaul rate allows accurate transmission, optimizing fronthaul rate allocation results in 10% higher transmission benefit than uniform allocation. It also makes possible to involve less users in NOMA while keeping the benefit. The proposed strategy enables uplink multi-cell multi-user processing in a cost-effective manner in 5G C-RAN deployments.**

***Index Terms*—5G, Cloud RAN, Multi-User Reception, Fronthaul, Resource Allocation, Cost-aware Optimization.**

I. Introduction

As some operators have already announced the deployment of 5G mobile networks in the coming years [1], it becomes essential to propose high performance and efficient network design that supports new services. We are entering the era of ambient connectivity requiring highly increased network capacity and reliability. Enhancing network capacity on the uplink becomes essential for services including real-time personal content sharing. In this paper, we will address radio access network for 5G and describe a practical solution that allows to achieve high throughput on the uplink even in the presence of many cell-edge users, when processing capability and network infrastructure are constrained. Our solution is based on Cloud Radio Access Network (C-RAN) which aims to improve both the performance on the radio interface and the efficiency of utilizing the computational resources and infrastructures in 5G networks.

A. C-RAN Deployment Challenges

An important part of throughput improvement in 5G is planned to be realized through the deployment of more radio access points, e.g., in a dense urban deployment, the distance between two macro base stations (BS) should be 200 m [2], significantly lower then with LTE, where usually we have at least 1 km between them. Consequently, the coverage area of each neighboring radio access point can simply overlap and due to frequency reuse between cells, inter-cell interference occurs and needs to be managed. Besides, it is also practical to use light-weight access points executing only a part of radio access network (RAN) processing: C-RAN is a key architecture in 5G [3]. The distributed Remote Radio Heads (RRHs) are connected to a Central Office (CO) where signal processing for several cells takes place allowing low latency multi-cell cooperation. Furthermore, today's general purpose processors and optimized real-time software enable a virtualized implementation of C-RAN in cloud components. The flexibility and scalability introduced by this new architecture allows to shift the focus of processing chains from the conventional cell-centric one to user-centric, making multi-cell transmission and reception techniques a major element

A part of this work has been presented at IEEE Wireless Communications and Networking Conference 2017.

D. Boviz is with Software Defined Mobile Networks Department, Nokia Bell Labs, 91620 Nozay, France and the Laboratoire des Signaux et Systèmes, CentraleSupélec, 91190 Gif-sur-Yvette, France (e-mail: dora.boviz@nokia-bell-labs.com).

C. S. Chen is with the department of Mathematics of Dynamic Systems, Nokia Bell Labs, 91620 Nozay, France (e-mail: chung_shue.chen@nokia-bell-labs.com).

S. Yang is with the Laboratoire des Signaux et Systèmes, CentraleSupélec, 91190 Gif-sur-Yvette, France. (e-mail: sheng.yang@centralsupelec.fr).

The work of D. Boviz and C. S. Chen was supported in part by the ANR project IDEFIX under the grant number ANR-13-INFR-0006.

of 5G. However, from an economic perspective, cost-effective network design is needed to allow sustainable operation. Fronthaul design is crucial to C-RAN and its optimization is indispensable. Our work is dedicated to the fronthaul resource allocation optimization for the purpose of 5G multi-user reception in C-RAN. We will discuss related work in the following.

B. Related Work

Aiming to achieve higher spectral efficiency and satisfy 5G requirements, non-orthogonal multiple access (NOMA) techniques are considered both for the downlink [4] and the uplink [5]. NOMA is also helpful in a multi-cell environment to manage inter-cell interference [6]. To deal with practical constraints such as receiver complexity, it has been proposed in [7] to perform the uplink multi-user detection needed in NOMA over several smaller subsets of users, while each subset transmits on a different channel. Besides, low-complexity multi-user receivers based on the approximation of optimal Minimum Mean Square Error (MMSE) detection are studied in [8]–[10]. In [11], the characteristics of the multi-user Multiple Input Multiple Output (MIMO) channel with fewer users than receive antennas are exploited to design an efficient receiver.

In addition, to realize low-latency data exchange required for multi-cell NOMA reception on the uplink (UL), a part of the physical layer processing of the cooperating cells has to be centralized, i.e., Cloud RAN architecture is needed. C-RAN is also necessary to accommodate large bandwidth planned 5G [2] and enable to serve a high number of users per cell, however, the interface between the RRHs and the CO has to be redefined [12], [13]. Signal compression in C-RAN aiming to satisfy fronthaul constraint is studied in various configurations, e.g., in [14], the authors propose a distributed compression method to maximize the sum rate. An efficient strategy consisting in jointly compressing the uplink signals and the channel state information (CSI) is described in [15]. In a previous work, we have defined a user-centric interface enabling multi-cell processing with affordable fronthaul traffic between the RRHs and the CO [16]. For legacy C-RAN and orthogonal multiple-access on the uplink, the fronthaul allocation optimization problem is investigated in [17]. Joint power allocation and adaptive quantization can substantially improve the system throughput compared to uniform distribution of available fronthaul, indicating the interest of channel-aware fronthaul allocation.

C. Contributions

In this paper, we consider 5G C-RAN uplink and study the fronthaul optimization problem in an end-to-end view. The contributions of the paper can be summarized as follows:

- We propose an uplink multiple-access scheme for C-RAN that improves cell-edge throughput and spectral efficiency while requiring affordable receive processing. It consists in associating UL Coordinated MultiPoint (CoMP) joint reception with NOMA, i.e., scheduling on the same frequency resource users located at the edge of neighboring cells. This allows cell-edge users benefit from multi-cell diversity and improved spectral efficiency. To ensure low receiver complexity, we distribute the users into several smaller groups and orthogonal resources are attributed to each group.
- We study the partitioning of users in groups which are required for the low-complexity NOMA scheme. We design an iterative algorithm for creating user groups based only on information that can be acquired without degrading the overall performance (e.g., the channel statistics), and evaluate its result compared to other user grouping strategies.
- We consider various deployment scenarios and the fronthaul allocation strategy that maximizes the net gain of operators from uplink transmissions in each scenario. We point out the significant improvement of the benefit of uplink transmissions in C-RAN thanks to optimal fronthaul allocation. Numerical evaluations illustrate that combining the proposed NOMA scheme with fronthaul optimization allows to deploy a practically implementable receiver satisfying 5G requirements without sacrificing the operators' benefits.

The remainder of the paper is organized as follows. In Section II, we describe the system model. In Section III, we propose an uplink NOMA scheme taking into account practical constraints regarding receiver complexity. In Section IV, we formulate the optimization problem allowing to adapt fronthaul rate in order to maximize the benefit that we can obtain from uplink transmissions using the NOMA strategy. In Section V, we give numerical results to illustrate the gain of fronthaul optimization. Finally, Section VI contains some concluding remarks.

II. Uplink MIMO C-RAN System Model

The system model that we use for uplink multi-user MIMO transmissions in C-RAN architecture is depicted in Figure 1. We consider M RRHs located at the cell sites with a antennas each, thus the total

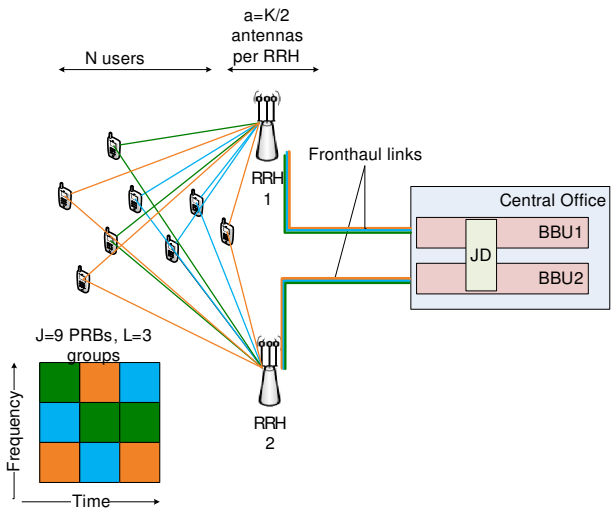

Fig. 1: System model with several user groups (example M = 2 RRHs), where JD stands for joint detection and BBU stands for base-band unit.

number of antennas is $K = a \cdot M$. All of the RRHs are connected via digital communication links to the same CO which can perform signal processing. We have N single-antenna users and consider that each user is allowed to communicate with all the M RRHs. The users are already present in the system and remain present, arrival or departure of users is not in the scope of the present work. We assume that the system architecture allows for all the N users to centralize the user-specific physical layer functions, i.e., the ones after demapping user signals from the attributed physical resource blocks (PRBs). This enables to perform multi-cell multi-user joint detection (JD) on the uplink [16] when NOMA is applied. We will describe in Section III the details of the NOMA receiver required before the user-PHY processing.

The RRHs, after receiving the uplink signals sent by the users, quantize them and forward through the fronthaul links to the CO. To send the signal received by each antenna via the fronthaul link, a given quantization rate is used when frequency modulated radio signals are compressed to digital base-band symbols. The transmission rate over the fronthaul link depends on the quantization rate, which is included in our optimization problem.

In the system, we consider that there is a total of J orthogonal physical resource blocks (PRBs) available commonly for all the RRHs at each channel use. Note that if the same data is sent by a user on several PRBs, the signal received by a given antenna will be combined after PRB demapping and then forwarded to the CO. Block fading channel model is used throughout the paper, and the coherence time is assumed to be long enough to consider adapting the fronthaul rate following uplink channel realizations. In terms of user mobility, e.g., for pedestrian users moving at 5 km/h, the channel changes approximately every 720 ms, in this time window we can update network parameters such as scheduling or quantization rate. Optimization is performed in the CO and results are fed back to the RRHs that can apply them on the mobile radio interface and the fronthaul interface respectively.

We use the following notational conventions: for random variables, uppercase letters with non-italic fonts, e.g., S, for scalars, bold and non-italic fonts, e.g., $\mathbf{V}$, for vectors, and bold and sans serif fonts, e.g., $\mathbf{M}$, for matrices. Deterministic variables are denoted with italic letters, e.g., a scalar x or N (for quantities), lowercase bold for a vector $\boldsymbol{v}$, and uppercase bold letters for a matrix $\boldsymbol{M}$. Logarithms are in base 2 and superscript $(.)^H$ denotes the conjugate transpose of a vector or a matrix.

The N users are uniformly distributed in the region covered by every RRH. The channel of each user n in the set $\mathcal{S}_N = \{1, ..., N\}$ towards all the antennas among the RRHs is denoted by $\boldsymbol{h}_n$, which is a K dimensional vector following the Gaussian distribution $\mathcal{N}(0, \boldsymbol{R}_n)$ with $\boldsymbol{R}_n \in \mathbb{C}^{K \times K}$. In the numerical evaluations, the channel covariances are computed following the one-ring scatterer model [18], thus they reflect the position of each User Equipment (UE)[1] with respect to the RRHs. Note that one can consider that the N users are not uniformly distributed in a region; this can be a subject of future work.

The UEs transmitting on the same PRB are said to form a group that has channel matrix given by the juxtaposition of the users' channel vectors. The number of user groups is denoted by L (with $L \leq J$). There are s_l users in the group denoted by Π_l, with $l \in \{1, ..., L\}$. The multi-user channel of the group Π_l, comprising the users with indexes $\pi_i^l \in \mathcal{S}_N$ with $i = 1, ..., s_l$, towards the K antennas is the $K \times s_l$ dimensional matrix $\boldsymbol{H}_l = \left[\boldsymbol{h}_{\pi_1^l} ... \boldsymbol{h}_{\pi_{s_l}^l}\right]$. The complete set of the groups $\{\Pi_1, ..., \Pi_L\}$ is a partition of the set of the users $\mathcal{S}_N$. We assume that the channel is perfectly known at the receiver. Though, channel estimators that are generally used result in an estimation error, its impact on our model is negligible.

The Gaussian channel noise vector is denoted by $\boldsymbol{n}_l \sim \mathcal{N}(0, \sigma_z^2 \boldsymbol{I}_K)$. Power of the input signal is normalized, so

[1]In 3GPP terminology, a mobile user is also called User Equipment (UE). In the paper, we will use the term user and UE interchangeably.

that noise covariance is $\sigma_z^2 = \frac{1}{\mathsf{SNR}}$. The signal received by the whole set of antennas for group Π_l is the K-dimensional vector $\boldsymbol{y}_l$ and it is given by the superposition of the signals sent by all of the s_l users in the group such that

$$\boldsymbol{y}_l = \sum_{i=1}^{s_l} \boldsymbol{h}_{\pi_i^l} x_{\pi_i^l} + \boldsymbol{n}_l. \quad (1)$$

The UL signal from group Π_l received at the CO is denoted by $\hat{\boldsymbol{y}}_l = (\hat{y}_{l1}, ..., \hat{y}_{lK})^T$, it is the quantized form of $\boldsymbol{y}_l$. The fronthaul rate used for transmitting $\hat{\boldsymbol{y}}_l$ is $\boldsymbol{c}^{(l)} = (c_{l1}, ..., c_{lK})^T$, where c_{lk} with $k \in \{1, ..., K\}$ denotes the fronthaul rate attributed to the forwarding of the component of $\boldsymbol{y}_l$ received by the antenna k, i.e., y_{lk}. The total fronthaul rate of group Π_l over the link between the RRH m and the CO is denoted by $c_m^{(l)}$ with $l \in \{1, ..., L\}$ and $m \in \{1, ..., M\}$. We can write $c_m^{(l)}$ using fronthaul rates $\{c_{lk}\}$ as follows:

$$c_m^{(l)} = \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk}. \quad (2)$$

If the capacity of the fronthaul links is limited, we denote the maximum capacity available for the whole set of user groups on the link m by $\bar{c}_m$.

III. Partial Multi-Cell NOMA

In this section, we describe the multi-cell NOMA scheme that is adapted to implementation requirements. This transmission strategy is applied in C-RAN architecture for the study of fronthaul rate allocation in Section IV.

A. Uplink NOMA

While with orthogonal multiple-access techniques we attribute to each UE a different time-frequency resource, in NOMA, users are intentionally scheduled in a way that they use the same PRBs. NOMA allows to increase the overall spectral efficiency if multi-user reception is applied for uplink, but requires this additional signal processing on the receiver side. Fortunately, in C-RAN, with multiple antennas, we can apply MIMO techniques, such as linear MMSE, for multi-user detection.

To increase the overall throughput while serving the whole set of users, in order to ensure (some) fairness, the best scheduling strategy is allowing to as many users as possible to transmit on all of the available PRBs. We will call this strategy full NOMA in the coming discussions. Indeed, the sum rate of s_l users transmitting together over J PRBs is higher than the sum of their rates while each of them uses alone J/s_l PRBs. Furthermore, frequency diversity is improved by scheduling all of the UEs on all of the subcarriers. However, regarding the implementation and the execution of receiver processing, it is not necessarily the best choice to apply full NOMA for many users.

B. Practical limitations: receiver complexity and data exchange

We aim to provide an end-to-end system design integrating uplink NOMA and ensure both high throughput and efficient implementation. For example, even if linear MMSE receiver has significantly lower complexity compared to maximum likelihood detection, it still requires a matrix inversion that has a complexity of $O(N^3)$ with N co-channel users. Several approximations of the MMSE detector have been proposed (e.g., [10], [11]). However, their complexity is still around $O(N^2)$. Furthermore, since the design of the low-complexity algorithms exploits the properties of tall MIMO channel matrices, with many users, the performance of the receiver decreases and would result in relatively high block error rates.

The physical layer processing of MIMO receivers has usually a parallel structure which bears separate streams for the processing of the data received at each antenna, and combines them before the decoding. Thus, to include NOMA in such an implementation, data sharing between these streams is needed, both in single-cell and multi-cell configurations. The fastest way of communicating between these simultaneously executed physical layer signal processing functions is to store the data that needs to be used by several streams (i.e., threads or processes) in a memory segment accessible to all of them. In order to keep the parallel structure of processing units and reduce the delay due to synchronization between the data streams that provide the inputs for computing MMSE matrices, it is more suitable and convenient to perform multi-user detection over several smaller subsets of users, with each subset scheduled on different PRBs.

C. Partial NOMA scheme

An ideal tradeoff between the throughput and the complexity of receive processing is to schedule users by groups and attribute different PRBs to each group in order to ensure orthogonality between them [7]. We will call this strategy partial NOMA, which aims to achieve significantly higher overall throughput with respect to single user transmission. Its advantage compared to full NOMA is to require multi-user receive processing over a smaller number of users for each group, thus, receiver complexity remains low and the signals of different groups can be processed simultaneously.

We define as a user group the UEs that are granted to transmit over the same set of PRBs. In the following, our goal is to illustrate the interest of partial NOMA on a simple model. Therefore, we assume that each of the N users in our system is included in one group exactly (see conditions (3b)-(3c) below) and for the sake of fairness, each group contains the same number of users when possible or at most one user more than another group (see condition (3d)). Also, we attribute the same number of PRBs to each group.

The following function $g(.)$ is used to denote the partitioning of the set of users $\mathcal{S}_N$ into L groups under the conditions (3b)-(3d):

$$g : \mathcal{S}_N \mapsto \{\Pi_1, ..., \Pi_L\}, \tag{3a}$$

$$\bigcup_l \Pi_l = \mathcal{S}_N, \tag{3b}$$

$$\Pi_i \cap \Pi_j = \emptyset \ \forall i, j \mid i \neq j, \tag{3c}$$

$$\mid s_i - s_j \mid \leq 1 \ \forall i, j \mid i \neq j. \tag{3d}$$

The proposed partial NOMA scheme allows to maintain the advantages of full NOMA, i.e., frequency and antenna diversity, at a reduced computational cost on the receiver side. In principle, the throughput of partial NOMA increases linearly with the number of users in each group, while the complexity of the MMSE receiver used for the detection in each group will increase proportionally to the square of the number of users. This would make MMSE receive processing inefficient and add significant computational delay when handling many users. Since in 5G systems, low transmission round trip time is targeted [19] and retransmission protocols (such as Hybrid Automatic Repeat reQuest) require low receive processing latency, full NOMA receiver is not very practical for 5G, for example to support enhanced Mobile BroadBand (eMBB) service. With partial NOMA, we can achieve high throughput compared to single-user scheduling and keep computational complexity and delay low.

D. Multi-cell partial NOMA and user grouping

For cells with the same CO, we apply the partial NOMA to multiple cells and turn inter-cell interference to useful signal by exchanging received data and CSI inside the CO. Similarly to uplink CoMP reception, the signals sent by cell-edge users can be received by several cells. If scheduler coordination is not used between neighboring cells involved in joint reception, we can benefit only from additional antenna diversity, but the overall spectral efficiency does not increase linearly. Furthermore, in some cases, cell-edge users can find themselves scheduled on the same PRB so that they are affected by the interference created by each other. To perform uplink NOMA, the scheduling needs to be coordinated between cells, and multi-user JR has to be realized accordingly. Low-latency data exchange enabled by C-RAN allows on one hand to apply multi-user reception techniques, and on the other hand to support it by scheduling coordination in order to assign the same PRB to the cooperating users, thus improve the throughput and spectral efficiency.

The C-RAN configuration is also more favorable for NOMA reception, thanks to better channel diversity created by the placement of the receive antennas at several spatially distributed RRHs. Users can have channel gain mainly contained in different matrix subspaces, i.e., their channels are not much correlated. Scheduling such users in the same group results in higher sum rate. However, in a practical system, it would introduce very high signaling overhead and require high-complexity processing to find the partitioning which gives the maximal throughput. Indeed, one needs the CSI between each user and each antenna, on all of the frequency resources. In addition, the problem of joint user grouping and scheduling optimization of N users into L groups scheduled on J PRBs would have an extremely high complexity, especially if we may deal with a significant number of users. As such an optimization must be realized in a fraction of the channel coherence time, high processing time prevents its usage in real systems.

To take into account constraints of real-time operation and avoid the overhead introduced by the pilot sequences for channel estimation, we can use the second order channel statistics, i.e., the channel covariance, that also does not change with the carrier frequency. Assuming that previous channel estimates are available for each user and these were computed for different subcarrier frequencies, the covariance can be easily computed and characterizes the user's channel independently from the scheduling decision. However, statistics describe less precisely the channel compared to real-time CSI. Using statistical CSI, we can then approximate the group sum rates and find a grouping scheme that is likely to achieve higher throughput, even without the knowledge of concrete channel realizations. This method also reduces the computational complexity, since the grouping can be realized separately from the scheduling.

Since we perform the user grouping step regardless to the limited fronthaul, we can assume unconstrained fronthaul in this step. In fact, the optimization of the fronthaul allocation (see Section IV) is performed once the user groups are determined and scheduled. We also assume that each group transmits over a resource orthog-

onal to the resources used by the other groups, thus there is no interference between them. The achievable sum rate of the user group Π_l with uniform power allocation is given by

$$\mathbb{E}[\log\ \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \mathbf{H}_l \mathbf{H}_l^H)]. \tag{4}$$

Note that for equation (4) we assume that the number of PRBs is equal to the number of user groups, i.e., $J = L$. If more PRBs are available then the number of user groups resulting in a group size convenient for receive processing, we can allocate in average J/L PRBs to each group and scale the sum rate consequently. To provide a simple comparison between the overall performance with different group sizes, we can write the following relation for $L_1 < L_2$ and $s_{l1} = \frac{N}{L_1} > s_{l2} = \frac{N}{L_2}$,

$$\begin{aligned} &\frac{J}{L_1}\mathbb{E}[\log\ \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \mathbf{H}_{l1} \mathbf{H}_{l1}^H)] > \\ &\quad \frac{J}{L_2}\mathbb{E}[\log\ \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \mathbf{H}_{l2} \mathbf{H}_{l2}^H)]. \end{aligned} \tag{5}$$

We can see that for larger user groups the sum rate for a single PRB is higher, since more users transmit, thus the total power is higher, in addition, they can use more PRBs, consequently, the total sum rate increases with the group size. This confirms that in theory full NOMA gives better performance, but for the reasons detailed in Section III-B it is not convenient for being used in practical systems.

Using Jensen's inequality, we get the following function that is an upper bound of the sum rate depending only on the channel covariance $\boldsymbol{R}_l = \mathbb{E}[\mathbf{H}_l \mathbf{H}_l^H]$:

$$\hat{f}(\Pi_l) = \log\ \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \boldsymbol{R}_l). \tag{6}$$

We can compute the value $\hat{f}(.)$ based on the CSI available in a practical system, although it only approximates the sum rate. Optimizing $\hat{f}(.)$ cannot guarantee that the actual sum rate is maximal, but it is likely to increase it.

A user pairing method maximizing the total sum rate was proposed in [20], it optimizes the user association by solving the assignment problem for a bipartite graph where users are denoted as the nodes and the value associated to the edge between two nodes is the sum rate that they can achieve together. Since the problem of finding the best user grouping with more than two users in each group is very complex to evaluate by exhaustive search, we can use an iterative algorithm built on the maximum sum assignment in bipartite graphs [21]. In our proposed scheme, we will solve the assignment problem N/L times to find a user grouping scheme which is expected to achieve high sum rate. The details of the proposed user grouping method are described in Algorithm 1, in the following. Obviously, the proposed user grouping method is not optimal, since in addition to the approximation of the sum rate, we assign the users to the groups in several iterations and keep the decisions of the previous iterations unchanged.

Input: Sets of channel vectors $\{\boldsymbol{h}_1, ... \boldsymbol{h}_N\}$ enabling to compute $\boldsymbol{R}_l$
Result: User grouping solution $g(\mathcal{S}_N)$
Initialize $\Pi_l \leftarrow \emptyset\ \forall l \in \{1, ..., L\}$;
for $i = 1 : \lceil \frac{N}{L} \rceil$ **do**
  Set $W = \{w_{k,m}\}_{\substack{1 \le k \le L \\ 1 \le m \le N-(i-1)L}} = \{0\}_{L \times (N-(i-1)L)}$;
  for $l = 1 : L$ **do**
    for each user n not included in any group **do**
      $\mathbf{H}_l^{(n)} \leftarrow [\mathbf{H}_l, \boldsymbol{h}_n]$;
      $\boldsymbol{R}_l^{(n)} \leftarrow \mathbb{E}[\mathbf{H}_l^{(n)} \mathbf{H}_l^{(n)^H}]$;
      $w_{ln} \leftarrow \hat{f}(\Pi_l \cup \{n\})$;
    end
  end
  Find $\{n_1^*, ..., n_L^*\} \leftarrow \underset{1 \le n_l \le s_l}{\operatorname{argmax}} \sum_{l=1}^{L} w_{ln_l}$ using the Hungarian algorithm [21];
  for $l = 1 : L$ **do**
    $\Pi_l \leftarrow \Pi_l \cup \{n_l^*\}$;
    $\mathbf{H}_l \leftarrow [\mathbf{H}_l, \boldsymbol{h}_{n_l^*}]$;
  end
end

Algorithm 1: User grouping algorithm based on channel statistics

We have evaluated the sum rate of the grouping scheme resulted by Algorithm 1 over a set of $N = 12$ users forming $L = 4$ groups and transmitting towards $M = 2$ RRHs with $a = 4$ antennas at each of them. We have compared the performance obtained by Algorithm 1 to exhaustive search, as well as to the average sum rate and the worst case grouping that we get by creating groups randomly. Table I shows the comparison of sum rates and also the complexity of each method.

We can see from Table I that the sum rate of the grouping scheme obtained by our proposed iterative algorithm that has the advantage of exploiting input data available in real radio access network deployments, is able to ensure a throughput only 1.2% lower than the globally best solution. Although the throughput achieved in average with random grouping is only slightly lower than the one with our solution, by the proposed deterministic grouping we can avoid a drop of the throughput by 5% in some cases. In use cases where throughput requirements are tight, it can be interesting to use the proposed algorithm although it requires more computation than random grouping. We can remark that the result of any practically implementable method exploiting realistic

Grouping strategy	Single-user	Random grouping (worst case)	Random grouping (average)	Iterative algorithm (proposed)	Exhaustive search
Average throughput (bits per channel use)	157.27	441.38	455.66	458.88	464.45
Processing complexity	1	1	1	$O(\frac{3N^4}{4L}) \sim 4000$	$O(\frac{L^N}{L!}) \sim 700000$

TABLE I: Comparison of different user grouping methods in terms of throughput and computational complexity. Evaluation is done for $N = 12$ users with random location and $J = 12$ PRBs available. The partial NOMA scheme is realized with $L = 4$ groups of 3 users, each of them transmitting over 3 PRBs towards $K = 8$ receive antennas equally distributed between the $M = 2$ RRHs.

input data similarly to the one described above can serve to fill the relatively small gap between random grouping and the optimal scheme.

IV. Cost-Effective Fronthaul Allocation

To propose an efficient end-to-end design of multi-cell NOMA in C-RAN, besides the implementation constraints in the CO described in the previous section, we have to improve the data transfer over the fronthaul links between the RRHs and the CO. We adopt a practical optimization scheme of fronthaul rates on the uplink which aims to maximize the net benefit that operators get from an uplink transmission. Since the user grouping decision is taken before scheduling the users, while the fronthaul quantization is set after the uplink partial NOMA transmission, for fronthaul optimization we are not restricted to use only statistical CSI. For given scheduling decision, real-time channel realizations (assumed perfectly known) within a channel coherence period are available for being used in the optimization of fronthaul rate. The technical details are given below.

A. Net benefit of the uplink multi-user transmission

In the C-RAN architecture, the cost of the data transfer between the RRHs and the CO is added to the usual exploitation costs of mobile networks. The gain of uplink transmissions increases with the rate, since high throughput allows mobile users to realize more data traffic that generates (direct or indirect) incomes for the mobile network operator. At the same time, the transmission requires some operational costs, among which we will focus on the one related to fronthaul usage. In fact, each uplink data transmission from a UE to the network results in a benefit that is assumed proportional to the instantaneous transmission rate, and each of them requires also to use fronthaul links for a given cost depending on the actual data rate. Obviously, the end-to-end rate of a user depends on the applied fronthaul rate, as it is determined by the level of quantization which impacts the quality of the reception at the CO. Finding an optimal tradeoff between the overall throughput and the fronthaul usage allows to maximize the final benefit that the operator gets. In the following, we define the net benefit of an uplink multi-user transmission as the instantaneous sum rate minus the fronthaul cost.

Proposition 1. *The partial NOMA scheme in C-RAN architecture with fronthaul quantization the achievable sum rate of a group Π_l with a known channel realization $\boldsymbol{H}_l$ is given by*

$$\sum_{i=1}^{s_l} r_i \geq log\ det\left(I_{s_l} + \boldsymbol{H}_l^H \boldsymbol{V}_{s_l}^{-1} \boldsymbol{H}_l\right) \tag{7}$$

with $\boldsymbol{V}_{s_l}$ the equivalent noise covariance:

$$\boldsymbol{V}_{s_l} = \sigma_z^2 \boldsymbol{I}_K + \underset{k=\{1,\ldots,K\}}{diag}\left(\frac{\sigma^2_{y_{lk}|\boldsymbol{H}_l} 2^{-c_{lk}}}{1 - 2^{-c_{lk}}}\right) \tag{8}$$

where $\sigma^2_{y_{lk}|\boldsymbol{H}_l}$ is the covariance of the signal received by the RRH at antenna k and c_{lk} is the number of bits that we use over the fronthaul link to forward this signal, i.e., the number of quantization bits.[2]

The proof of Proposition 1 is provided in Appendix A; it shows the impact of fronthaul quantization on the sum rate.

We use the achievable sum rate (7) to formulate the objective function allowing to maximize the end-to-end benefit of the user group Π_l with s_l users towards the M RRHs with a antennas each. The parameters of this function are the following:

[2]Note that as both the channel noise and the equivalent quantization noise are assumed independent (between the antennas) and Gaussian, equation (7) gives the minimum sum rate that we can achieve [22] in the defined system model.

- The Gaussian channel noise variance σ_z^2.
- The average received signal power from group l at antenna k given the channel estimate of the group: $\sigma^2_{y_{lk}|\mathbf{H}_l}$.
- The fronthaul capacity c_{lk} used for forwarding to the CO the symbols y_{lk}.

The following function characterizes the net benefit of the transmission of group Π_l towards the whole set of receive antennas when the fronthaul rate allocated to the group is $\boldsymbol{c}^{(l)}$:

Given the parameters $\sigma_z^2, \sigma^2_{y_{lk}|\mathbf{H}_l}, \forall k \in \{1, ..., K\}$,

$$f(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) = \text{logdet}\left(I_{s_l} + \boldsymbol{H}_l^H \boldsymbol{V}_{s_l}^{-1} \boldsymbol{H}_l\right) - q(\boldsymbol{c}^{(l)}) \quad (9)$$

where $\boldsymbol{V}_{s_l}$ is defined in (8). The first term of the function $f(.)$ in (9) gives the instantaneous sum rate during the coherence time block where the channel matrix $\boldsymbol{H}_l$ holds and the second term is the total cost of the fronthaul transmission for group Π_l over the whole set of fronthaul links connecting the RRHs to the CO, denoted by the cost function $q(.)$ defined in equation (13), see later.

B. Fronthaul deployments and cost models

To enable data processing in the CO, it has to be connected to the RRHs through high capacity and reliable communication links. Several technologies (e.g., microwave) are considered depending on the cell size and the deployment area, however, the mainstream technology for fronthaul connection of macrocells in 5G remains the fiber-based Carrier Ethernet. It is also the most suitable for a generic C-RAN architecture, where the distance between the RRHs and the CO can reach a few kilometers. Network operators have several options to connect their cell-sites to the CO [23]:

a) Scenario 1 (Fronthaul leasing): A given capacity of fiber Ethernet can be leased from its owner who provides it either through a point-to-point link in some cases or through a switched network infrastructure. This scenario is modeled by limited fronthaul link capacity and a constant cost-per-bit $\lambda_k^{(1)}$. Assuming that the leased network capacity is accurately dimensioned, the cost-per-bit of the fronthaul transmissions in this model reflects the part of the total leasing cost for a unit rate. This linear model allows to dispatch the overall investment between all the transmissions over the leased link proportionally to the fronthaul rate that they use. As in our system model the antennas are located at several RRHs, due to differences between the fronthaul connections at each RRH, the cost can vary in function of the location of the antennas (we attribute a cost to an antenna k instead of an RRH in order to keep the framework independent of the number of antennas).[3] Then the total cost of the fronthaul transmission between the K antennas and the CO can be written as

$$q_1(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} \lambda_k^{(1)} c_{lk}. \quad (10)$$

b) Scenario 2 (Owned point-to-point links): The network operator can install its own point-to-point fiber link fully dedicated to the communication between a given RRH and the CO, thus the transmission cost is the consequence of the investment realized for the deployment and the operational costs such as energy consumption. In this case, the fronthaul capacity is limited and the cost-per-bit can be modeled as in (11). In this formulation, the first term decreases with the rate used, its role is to represent a portion of deployment costs. The more fronthaul is used, the lower is the cost-per-bit, since the constant deployment cost is distributed over a higher total rate. The second term $\lambda_k^{(2)}$ accounts for constant operational costs such as energy consumption.

$$q_2(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} \left(\frac{\mu_k^{(2)}}{c_{lk}} + \lambda_k^{(2)} c_{lk} \right) \quad (11)$$

with $\mu_k^{(2)}$ a real-valued constant. We use $\frac{\mu_k^{(2)}}{c_{lk}}$ as a decreasing function in our model, since the more straightforward is to distribute investment costs over the occurring transmissions proportionally to the rate used. However, note that any other positive decreasing function can be applied to represent a different way of distributing link deployment costs.

c) Scenario 3 (Owned converged infrastructure): Some operators own large fiber network infrastructures that are shared by various services and multiple sites. We expect that the cost model suitable to this scenario is when per-link fronthaul capacity is unlimited and the cost-per-bit of the fronthaul usage includes a penalty for preventing other services to use the given amount of rate. The first term of the cost then increases with the allocated rate and the second one represents the constant price. In this model, the more fronthaul rate is used, the higher is the price factor, since the allocated rate becomes unavailable for other services that may generate additional revenue. Here, we model the penalty pricing linearly with respect to the fronthaul rate (resulting in a quadratic term in the total cost), but other positive increasing functions can be also suitable. Note that the

[3]It is also possible to attribute different cost coefficients $\lambda_k^{(1)}$ to each group, for example if there is a priority ordering between the user groups. For simplicity, we do not consider this case, but our method and results remain valid as long as all the cost coefficients are positive.

second term captures operational costs in the constant $\lambda_k^{(3)}$. The following equation describes this model where $\mu_k^{(3)}$ is a real-valued constant

$$q_3(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} (\mu_k^{(3)} c_{lk} + \lambda_k^{(3)}) \cdot c_{lk}. \tag{12}$$

The cost functions (10)-(12) model various considerations of the fronthaul capacity, pricing, investment cost, and fronthaul infrastructure sharing. They can be generalized as follows:

$$q(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} (\mu_k c_{lk}^n + \lambda_k c_{lk}) \tag{13}$$

where n is a real-valued exponent, μ_k and λ_k are non-negative coefficients.[4]

When $n = 1$, $q(\boldsymbol{c}^{(l)}) = q_1(\boldsymbol{c}^{(l)})$ with $\lambda_k^{(1)} = \lambda_k + \mu_k$; $n = -1$ gives $q(\boldsymbol{c}^{(l)}) = q_2(\boldsymbol{c}^{(l)})$ with $\mu_k = \mu_k^{(2)}$ and $\lambda_k = \lambda_k^{(2)}$; while $n = 2$ covers Scenario 3 with $\mu_k = \mu_k^{(3)}$ and $\lambda_k = \lambda_k^{(3)}$.

C. Optimal fronthaul allocation

With the above definitions of the net benefit of uplink partial NOMA transmission and the fronthaul cost, we can determine the fronthaul allocation that results in a maximal benefit for the network operator, by the following optimization problems.

1) With per-link fronthaul constraint: The fronthaul allocation scheme that maximizes the net benefit for the whole set of L users groups is the one that gives the highest sum of the metric (9) over all groups. In Scenario 1 and Scenario 2, where the available fronthaul rate is limited, we have to solve the following constrained optimization problem:

$$\begin{aligned} &\text{Find } \{\boldsymbol{c}^{(1)*}, ..., \boldsymbol{c}^{(L)*}\} = \underset{\{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)}\}}{\operatorname{argmax}} \sum_{l=1}^{L} f(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) \\ &\text{subject to } \sum_{l=1}^{L} c_m^{(l)} \leq \overline{c}_m, \ \forall m \in \{1, ..., M\}. \end{aligned} \tag{14}$$

Let us recall that $c_m^{(l)} = \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk}$, so the above constraint can also be written as:

$$\sum_{l=1}^{L} \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk} \leq \overline{c}_m, \ \forall m \in \{1, ..., M\}. \tag{15}$$

[4]We can remark that although (12) is a general polynomial function, the optimization problem that includes such a cost function can be solved similarly with any non-negative convex cost function.

Proposition 2. *The problem (14) is concave, thus admits a unique solution that gives the optimal capacity allocation scheme.*

The proof of Proposition 2 is provided in Appendix B.

2) Without fronthaul constraint: In Scenario 3, we assume that by the usage of a converged network infrastructure serving for the fronthaul, the available fronthaul capacity can be considered unlimited, and the cost function includes a term that stands for the cost related to network sharing. In this case, the following unconstrained optimization problem, which is also to maximize the the sum of (9) over all groups, needs to be solved for finding the best fronthaul allocation scheme:

$$\begin{aligned} &\text{Find } \{\boldsymbol{c}^{(1)*}, ..., \boldsymbol{c}^{(L)*}\} = \underset{\{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)}\}}{\operatorname{argmax}} \sum_{l=1}^{L} f(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) \\ &\text{with } 0 < c_{lk}^*, \ \forall k \in \{1, ..., K\}, \ \forall l \in \{1, ..., L\}. \end{aligned} \tag{16}$$

Note that the possible fronthaul rate values have to be positive, however, this does constrain the optimization in practice. We have the following statement extended from Proposition 2. The proof is very similar and thus omitted.

Proposition 3. *The unconstrained optimization problem (16) is concave and admits a unique solution, as it has the same objective function as problem (14).*

V. Performance Evaluation

In this section, we aim to highlight by numerical results the benefit of cost-aware fronthaul allocation applied to the proposed partial multi-cell NOMA scheme. We have evaluated the results of the fronthaul allocation optimization in the 3 deployment scenarios described in Subsection IV-B with $N = 40$ users transmitting towards $M = 2$ RRHs located at 500 meters from each other. We have $K = 8$ antennas equally distributed between the RRHs. Note that the optimization problem can be solved efficiently using standard convex programming [24]. Channel gain is modeled using independent one-ring scatterer model for each user [18].

To provide an accurate comparison of the scenarios that we have defined in Subsection IV-B, we assume the following relation among the price factors $\lambda_k^{(i)}$ in order to illustrate the characteristics of the fronthaul architectures and commercial models respectively

$$\lambda_k^{(2)} \leq \lambda_k^{(3)} \leq \lambda_k^{(1)}. \tag{17}$$

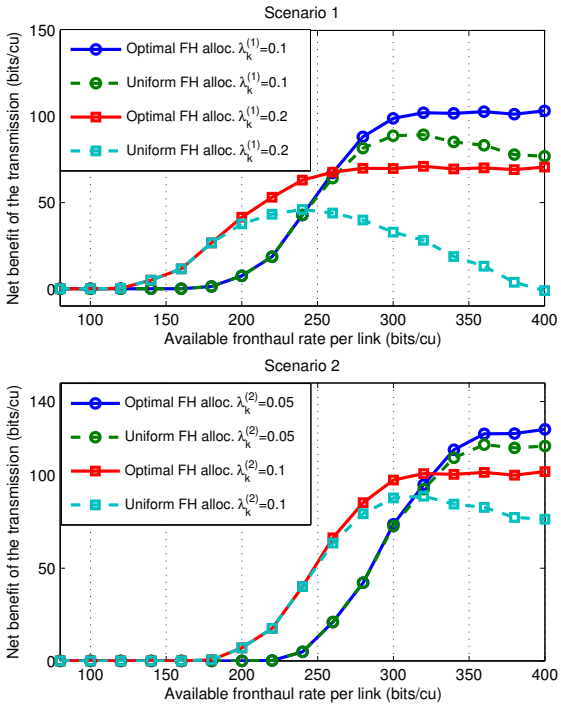

Fig. 2: Net benefit of uplink transmission with constrained fronthaul

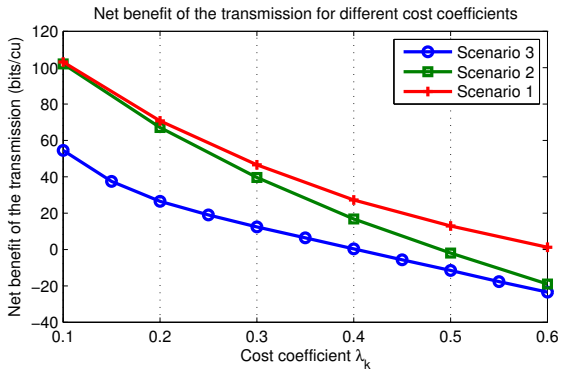

Fig. 3: Net benefit of uplink transmission for different cost coefficient values

A. Limited fronthaul rate available

In Scenarios 1 and 2, the fronthaul rate that we can allocate per link is constrained. The gain provided by optimal fronthaul allocation strategy depends on the amount of available fronthaul rate and the fronthaul cost.

We show in Figure 2 the benefit realized by uplink multi-user partial NOMA transmissions with $L = 10$ user groups following the metric defined in (9). For Scenario 2, we used $\mu_k = \frac{\lambda_k^{(2)}}{2}$ for this evaluation in order to model the fact that constant operational costs are higher than the cost related to the investment which is shared among all the transmissions occurring. We compare the optimal fronthaul allocation scheme to uniform fronthaul allocation for different amounts of available fronthaul capacity. Note that in uniform fronthaul allocation, available fronthaul capacity is equally distributed to all groups and all antennas.

When the available fronthaul rate is low, both uniform and optimized allocation result in similar efficiency, since the constraint does not allow to achieve higher sum rate. With sufficient fronthaul rate, optimized fronthaul allocation allows to achieve higher transmission benefit, since the sum rate of each group can be improved by allocating more fronthaul to the received signals with higher powers. In other words, fronthaul allocation is adapted to the variations of channel gains for different users and antennas. In Scenario 2, when a point-to-point fronthaul link is owned by the network operator, we can see that the benefit of the transmission is higher when more fronthaul rate is available, since the investment cost term is reduced thanks to higher sum rate, therefore higher gain.

B. Fronthaul allocation with different cost values

We have evaluated the maximal net benefit that we get by optimizing the fronthaul rate allocation for different cost coefficients. For Scenario 2, we have used $\mu_k^{(2)} = \lambda_k^{(2)}/2$ and for Scenario 3 $\mu_k^{(3)} = \lambda_k^{(3)}/4$.

We can observe in Figure 3 that the benefit in Scenario 2 is close to the one of Scenario 1 when exploitation costs are low, and for high cost it approaches the (lower) benefit obtained in Scenario 3. Also, the benefit of the transmission decreases quickly when the cost increases. The benefit of the transmission can even happen to be negative despite optimization, whereas the cost of fronthaul usage can be higher than the total sum rate. Obviously, in this case it is better not to transmit or change the system parameters, e.g., the size of the NOMA groups.

C. Optimization for various group sizes

As we have detailed in Section III, for practical considerations, the best choice is not necessarily to schedule as many users as possible on the same PRBs. However, regarding the fronthaul, one can expect that by reducing the number of user groups, we use less fronthaul and get higher benefit from the transmission. To quantify this benefit, we have evaluated the result of optimal fronthaul allocation for various group sizes. The number of PRBs

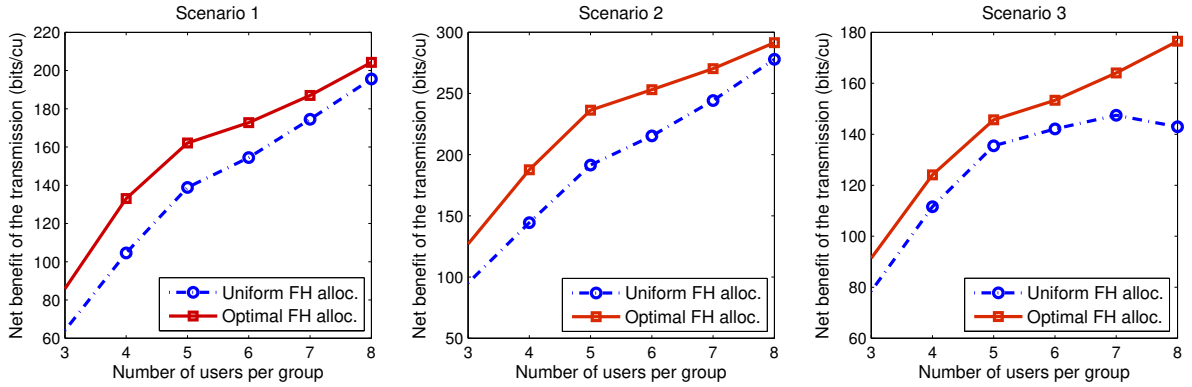

Fig. 4: Net benefit of uplink partial NOMA transmission for different group sizes. For Scenario 1 $\lambda_k^{(1)} = 0.2,\ \forall k \in \{1, ..., K\}$ and the available fronthaul rate is 400 bits/channel use/link. For Scenario 2, $\lambda_k^{(2)} = 0.1$ and $\mu_k^{(2)} = 0.05,\ \forall k \in \{1, ..., K\}$ and the available fronthaul rate is also 400 bits/channel use/link. For Scenario 3, $\lambda_k^{(3)} = 0.12$ and $\mu_k^{(2)} = 0.03,\ \forall k \in \{1, ..., K\}$. These values are set to follow the differences between deployment and operational cost as described in (17).

is fixed to $J = 20$ and the $N = 40$ users are partitioned in groups varying from $L = 20$ to $L = 5$.

We can see in Figure 4 that thanks to fronthaul optimization, with groups of 4 users, we already get around 70% of the gain that we can get with 8 users per group (note that the latter requires much higher computational cost). In Scenarios 1 and 2, we have a gap of about 10% of net benefit between optimized and uniform fronthaul allocation. In these scenarios, since the dominant cost term is the one with $\lambda_k^{(i)}$ which models exploitation costs, fronthaul allocation improves more the transmission gain for medium group size than for large group size. We can achieve a given value of net benefit for a partial NOMA transmission with less users when fronthaul allocation is optimized.

In Scenario 3, when more fronthaul is allocated per group, the cost term with $\mu_k^{(3)}$ that aims equity between the various services sharing the same fronthaul infrastructure, becomes dominant for large user group. Consequently, optimizing the fronthaul allocation gives more improvement compared to smaller group size. However, with optimal fronthaul allocation and a group size of 5 users, we can achieve the maximal net benefit possible with uniform fronthaul allocation for any group size.

We compare in Figure 5 the efficiency of fronthaul usage, i.e., the ratio of the net benefit of the transmission and the total fronthaul rate used, in the different deployment scenarios. We can see again that optimizing fronthaul allocation improves the performance of transmissions for any group size in all of the 3 scenarios. With the different cost models used, the fronthaul is exploited

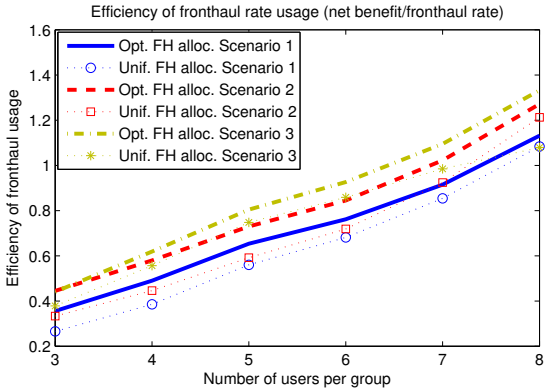

Fig. 5: The efficiency of fronthaul usage by uplink partial NOMA transmission for different group sizes in the 3 deployment scenarios defined in Subsection IV-B

with the highest efficiency in Scenario 3. In comparison, Scenario 2 can achieve higher efficiency than Scenario 1, in both the uniform and optimized fronthaul allocation cases. We can observe in Figure 5 that by optimizing the fronthaul allocation in Scenarios 1 and 2, the efficiency of the fronthaul usage becomes close to the one that we get from uniform allocation under Scenarios 2 and 3, respectively.

VI. Conclusion and Perspectives

In the first part of the paper, we have described practical limitations of uplink multi-user detection for NOMA and identified a solution that allows to improve the spectral efficiency with respect to single-user trans-

missions, despite the tradeoff between high throughput and implementation complexity. We have proposed to apply the partial NOMA scheme defined as a mixture between orthogonal and non-orthogonal scheduling to Cloud RAN architecture where both control and traffic data can be shared between several cells. By partitioning the users into groups that use different PRBs, we can significantly reduce computational complexity and processing latency, although we would get slightly lower data rate. We have also studied the possibility to create the user groups in a deterministic manner based on channel statistics accessible in real RAN deployments. We could observe that with a practical grouping method only a small improvement of the sum rate can be achieved with respect to random grouping, thus it is useful to apply it only in services for which even a small improvement of the throughput can be important.

In the second part of the paper, we have focused on optimizing the fronthaul rate used for forwarding the signals received by the RRHs to the CO in the case of partial multi-cell NOMA where these signals are the superposition of UEs that are in the same user group. Our aim is to adapt the fronthaul rate of each group following its instantaneous channel conditions, in order to maximize the net benefit of the transmission (i.e., the gain that we get from achievable sum rate minus the cost of using a given amount of fronthaul rate). We have described various fronthaul cost models, each of them corresponding to a different deployment and exploitation scenario. The optimization problems that we can define allow to find the optimal allocation scheme that gives the highest net benefit of the uplink transmission.

Finally we have shown the improvement of the net benefit of uplink partial NOMA transmissions by the proposed optimal fronthaul allocation optimization. We have found that the more fronthaul rate is available on the link between the RRHs and the CO, the more the optimal allocation improves the benefit compared to uniform allocation. The comparison of deployment scenarios for different exploitation cost coefficients has shown that leasing fronthaul infrastructures can be the most beneficial except when the cost of deploying new links is negligible (e.g., very long term investments). We can also confirm that independently of the cost model, fronthaul allocation is useful for any group size and can compensate the loss of benefit due to partial NOMA instead of using full NOMA. By optimizing fronthaul rates, we can achieve the efficiency of exploiting fronthaul links, for example optimal allocation in simpler models (Scenarios 1 and 2) can be as efficient as the more evolved one (Scenario 3) but using uniform allocation. These show that by combining partial multi-cell NOMA on the C-RAN radio interface with cost-aware fronthaul allocation on the RRH-CO interconnection, we are able to ensure high spectral efficiency and throughput, by using affordable and practically implementable multi-user receiver and maximizing the benefit that operators get despite the fronthaul usage cost that is considered as the main limitation of Cloud RAN.

An interesting work to study the fronthaul allocation optimization is to use channel measurements coming from real network deployments in order to evaluate its effect with real-world channel realizations instead of modeling the channel. A future work is also to consider other channel models including microwave and millimeter wave. Applying similar optimization strategies in heterogeneous networks can be also an interesting future work.

APPENDIX A
SUM RATE OF A MULTI-USER CHANNEL IN C-RAN

Proof:

Quantization noise: To model the limited available fronthaul rate, we use the notion of equivalent quantization noise which is defined for a transmission by s_l users towards K receive antennas as follows.

We define the distortion d_{lk} between the signal y_{lk} received by the RRH and the compressed signal $\hat{y}_{lk}$ received by the CO as the squared-error distortion between y_{lk} and $\hat{y}_{lk}$.

$$d_{lk} = D(y_{lk}, \hat{y}_j) \coloneqq \mathbb{E}[\mid y_{lk} - \hat{y}_{lk} \mid^2 \mid \mathbf{H}] \tag{18}$$

The minimum achievable rate of a signal quantized with distortion D is given by the mutual information between the initial (received) signal and the quantized (forwarded) one [25, Theorem 10.2.1]. If this rate is lower than the capacity of the fronthaul link, the used quantization allows an accurate transmission where the distortion does not exceed a given variance $\sigma^2_{d_{lk}}$. We can write for a point-to-point link lk:

$$\begin{aligned} r_{lk} &\leq c_{lk} \\ r_{lk} &\geq \min_{p_{\hat{y}_{lk}|y_{lk}}: D \leq \sigma^2_{d_{lk}}} I(Y_{lk}; \hat{Y}_{lk} \mid \mathbf{H}_l). \end{aligned} \tag{19}$$

The following relation between the received signal power denoted by $\sigma^2_{y_{lk}|\mathbf{H}_l}$ and $\sigma^2_{d_{lk}}$, the maximum variance of distortion noise that we assume Gaussian[5] can be derived according to [25, Theorem 10.3.2].

$$\sigma^2_{d_{lk}} \leq \sigma^2_{y_{lk}|\mathbf{H}_l} 2^{-c_{lk}} \tag{20}$$

[5]Other distributions of the distortion noise would result in a higher achievable rate.

where

$$\sigma^2_{y_{lk}|\mathbf{H}_l} = \sum_{i=1}^{s_l} (\mid h_{k\pi_i^l} \mid^2) + \sigma_z^2. \quad (21)$$

We use a scaling factor α_{lk} in order to adapt the power of forwarded signal to the fronthaul capacity used, i.e.,

$$\alpha_{lk} = \frac{\sigma^2_{y_{lk}|\mathbf{H}} - \sigma^2_{d_{lk}}}{\sigma^2_{y_{lk}|\mathbf{H}_l}}, \ \forall k \in \{1, ..., K\}. \quad (22)$$

Scaling factors for each antenna form the matrix $\boldsymbol{A}_l = \underset{k=1,...,K}{\text{diag}} (\alpha_{lk})$. The distortion has then the following upper bound which, when the equality is satisfied, gives the optimal point-to-point quantization scheme:

$$\frac{\sigma^2_{d_{lk}}}{\alpha_{lk}} = \frac{\sigma^2_{d_{lk}} \sigma^2_{y_{lk}|\mathbf{H}_l}}{\sigma^2_{y_{lk}|\mathbf{H}_l} - \sigma^2_{d_{lk}}} \leq \frac{\sigma^2_{y_{lk}|\mathbf{H}_l} 2^{-c_{lk}}}{1 - 2^{-c_{lk}}}. \quad (23)$$

Achievable sum rate: We compute the achievable sum rate from the received signal affected by the quantization noise for a given user group Π_l using the mutual information between the signal sent by all users in the group and the one received in the CO,

$$\sum_{i=1}^{s_l} r_i \geq I(\mathbf{X}_l; \hat{\mathbf{Y}}_l \mid \mathbf{H}_l) = h(\mathbf{X}_l) - h(\mathbf{X}_l \mid \hat{\mathbf{Y}}_l, \mathbf{H}_l). \quad (24)$$

We compute both entropy terms in order to find the lower bound of the sum rate. The first term describes the quantity of information sent by the users, and thus depends on the transmission power (defined as unit in our system model):

$$h(\mathbf{X}_l) = \log(\det(2\pi e \mathbb{E}[\mathbf{X}_l \mathbf{X}_l^H])) = \log((2\pi e)). \quad (25)$$

The second term represents the loss of information between the UEs and the CO. We can compute its upper bound using linear MMSE covariance $\mathcal{C}_e$ that corresponds to the case where $\hat{\mathbf{Y}}_l$ would be Gaussian.

$$h(\mathbf{X}_l \mid \hat{\mathbf{Y}}_l, \mathbf{H}_l) \leq \log(\det(2\pi e \mathcal{C}_e)). \quad (26)$$

For this, we use the definition of received signal by the CO:

$$\hat{\boldsymbol{y}}_l = \boldsymbol{A}_l \left(\sum_{i=1}^{s_l} \boldsymbol{h}_{\pi_i^l} \boldsymbol{x}_i + \boldsymbol{n}_l \right) \quad (27)$$

where $\boldsymbol{n} = \boldsymbol{z} + \boldsymbol{d}$ is the equivalent noise containing Gaussian channel noise and the quantization noise. The covariance matrix of this equivalent noise is $\mathcal{C}_N = \sigma_z^2 I_K + \underset{k=1,...,K}{\text{diag}} \left(\frac{\sigma^2_{d_{lk}}}{\alpha_{lk}} \right)$.

We can compute the linear MMSE covariance based for a given channel realization

$$\mathcal{C}_e = \boldsymbol{I}_{s_l} - \boldsymbol{H}_l^H (\boldsymbol{H}_l \boldsymbol{H}_l^H + \mathcal{C}_N)^{-1} \boldsymbol{H}_l. \quad (28)$$

Then we apply the inversion lemma on the lower bound of the mutual information, finally, substituting the quantization noise by its upper bound following (23) completes the proof of Proposition 1. ■

APPENDIX B
PROOF OF PROPOSITION 2

Proof: The constraint in (14) is linear, and subtracted cost function $q(.)$ is assumed to be convex, thus the concavity of the first term of $f(.)$ is sufficient to show that the problem is concave. The function $\log\det(\boldsymbol{A})$ is concave if and only if the matrix $\boldsymbol{A}$ is non-negative definite. The sum of two non-negative definite matrices is also non-negative definite. Since the identity matrix satisfies this condition, we only need to show that the second term of the argument of the $\log\det(.)$ in (9) is non-negative definite. The equivalent SNR matrix $\boldsymbol{V}_{s_l}$ is diagonal with positive elements which are its eigenvalues, thus it is positive definite. This property stands also for its inverse.

If a positive definite matrix $\boldsymbol{M}$ is multiplied by another matrix and its hermitian as $\boldsymbol{B}^H \boldsymbol{M} \boldsymbol{B}$, the result is also positive definite if $\boldsymbol{B}$ is full rank. This is true for $\boldsymbol{H}_l^H \boldsymbol{V}_{s_l}^{-1} \boldsymbol{H}_l$ since the columns of $\boldsymbol{H}$ are independent, thus $\text{rank}(\boldsymbol{H}) = s_l$. Consequently, the matrix being the argument of $\log\det(.)$ is positive definite and also non-negative definite, thus the first term of (9) which implies with the above reasons that (14) is concave. ■

ACKNOWLEDGEMENT

The authors would like to thank Laurent Roullet, Jakob Hoydis and Imran Latif of Nokia Bell Labs for their valuable comments and discussions in this work.

REFERENCES

[1] InterDigital, Inc., "How will the olympics shape 5G?" 2016.

[2] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies ," 3rd Generation Partnership Project (3GPP), TSG-RAN TR.38.913, 2017.

[3] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, March 2015.

[4] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE 77th Vehicular Technology Conference (VTC-Spring)*, June 2013, pp. 1–5.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, September 2015.

[6] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *arXiv preprint arXiv:1611.01607*, 2016.

[7] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 781–785.

[8] P. Torres and A. Gusmao, "Detection issues with many BS antennas available for bandwidth-efficient uplink transmission in a MU-MIMO system," in *IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[9] B. Yin, M. Wu, C. Studer, J. R. Cavallaro, and C. Dick, "Implementation trade-offs for linear detection in large-scale MIMO systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 2679–2683.

[10] S. Wang, Y. Li, and J. Wang, "Low-complexity multiuser detection for uplink large-scale MIMO," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2014, pp. 224–229.

[11] L. Fang, L. Xu, and D. D. Huang, "Low complexity iterative MMSE-PIC detection for medium-size massive MIMO," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 108–111, Feb 2016.

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

[13] China Mobile Research Institute, Alcatel-Lucent, Nokia Networks, ZTE Corporation, Broadcom Corporation, Intel China Research Center, "White paper of NGFI (Next Generation Fronthaul Interface)," october 2015.

[14] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 692–703, Feb 2013.

[15] J.-K. Kang, O. Simeone, J. Kang, and S. Shamai, "Joint signal and channel state information compression for uplink network MIMO systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2013, pp. 875–878.

[16] D. Boviz and Y. El Mghazli, "Fronthaul for 5G: low bit-rate design enabling joint transmission and reception," in *IEEE Global Telecommunications Conference (Globecom), 5G RAN Design Workshop*, December 2016.

[17] L. Liu, S. Bi, and R. Zhang, "Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4097–4110, Nov 2015.

[18] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, 2000.

[19] Nokia, "5G use cases and requirements," 2015.

[20] E. Viterbo and A. Hottinen, "Optimal user pairing for multiuser MIMO," in *IEEE 10th International Symposium on Spread Spectrum Techniques and Applications*, 2008, pp. 242–246.

[21] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[22] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, April 2003.

[23] E. Wireless, "Economics of Backhaul," http://www.exaltcom.com/Economics-of-Backhaul.aspx, 2016, [Online].

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.