Virtualization, User Association, and Rate Maximization for Massive MIMO SD-RAN with Limited Fronthaul Resources

Homa Eghbali

***Abstract*—To facilitate the dynamic management of massive multiple-input multiple-output (MIMO) networks, centralized control of network resources via a software-defined network (SDN)-enabled control plane is required. The central controller is connected to the massive MIMO base stations (BSs) via fronthaul interfaces. In software-defined radio access networks (SD-RANs), the capacity of the fronthaul interfaces can be considerably less than the data rate required by the user equipments (UEs). This makes the fronthaul transmission the performance bottleneck of the SD-RAN. Motivated by this practical concern, we investigate the performance of fronthaul compression for massive MIMO SD-RAN, considering both perfect training and pilot contamination. We formulate a novel virtualization, user association, and rate maximization (VARM) problem based on a hybrid virtualization and compression (HVC) scheme that maximizes the aggregate data rate of the UEs while prioritizing services for the virtualized UEs. We use a Stackelberg game to characterize the resource consumption and BS association strategies of virtualized and non-virtualized UEs as well as the optimal compression noise levels. Numerical results show that compared to the conventional maximum peak rate association strategy, our proposed scheme provides higher aggregate user rate while guaranteeing the UEs their preferred quality of service (QoS).**

Index Terms: SDN, massive MIMO, limited fronthaul resources, Stackelberg game.

I. Introduction

To accommodate the surging demand for wireless data and multimedia applications, mobile operators must find solutions to provide customers with seamless and better quality of experience (QoE). Massive multiple-input multiple-output (MIMO) communication has the potential to improve the capacity of the future fifth generation (5G) mobile networks [1]. In massive MIMO, a large number of antennas simultaneously serve a number of users in the same time-frequency slot. The transmitted signals are combined in the air such that the random effects of small scale fading can be averaged out [2]. Considering a network architecture with massive MIMO base stations (BSs) that are overlaying with many small cells, one of the key system optimization problems is the optimal association of user equipments (UEs) to BSs such that the throughput of the radio access network (RAN) is maximized. Gotsis *et al.* in [3] investigated the problem of user association in a massive MIMO dense network. They formulated a problem that maximizes the worst rate across all UEs. Xu *et al.* in [4] developed centralized and distributed association algorithms for cell association in a massive MIMO enabled heterogeneous network.

Further performance enhancement of 5G mobile networks requires joint centralized control of resources through software-defined networking (SDN). The SDN paradigm has been incorporated in different wireless networking projects as a key enabler to simplify the provisioning, management, and reconfiguration of mobile networks. As an example, the MobileFlow project [5] proposed to incorporate SDN principles into the 3rd Generation Partnership Project (3GPP) evolved packet core mobile carrier networks. We envision software-defined (SD)-RAN as a crucial element of future 5G standards that facilitates dynamic implementation in software of functions such as user association and virtualization of network resources in software through advanced signal processing and fronthaul[1] traffic engineering techniques.

Despite its advantages, SD-RAN also comes with its own challenges. An important prerequisite for the effective centralized processing in the SD-RAN are high bandwidth and low latency fronthaul interfaces connecting the BSs to the central controller. Unfortunately, practical fronthaul implementation are capacity and time-delay constrained. Due to limited fiber resources, deploying a large number of fiber interfaces that directly connect the SD-RAN controller to the BSs is difficult to achieve for most operators. As a result, fronthaul transmission can be the performance bottleneck of centralized implementations such as SD-RAN [6].

To overcome the practical concerns regarding fiber fronthaul transmission of SD-RAN, various solutions have been proposed. Compression and large-scale pre-coding/de-coding can reduce consumption of fiber resources. Samardzija *et al.* in [7] employed redundancy removal in the spectral domain, block scaling, and non-uniform quantization to effectively reduce the amount of data transferred between the BSs and RAN controller. Nanba *et al.* in [8] proposed a compression scheme for baseband signals in centralized RAN to reduce

H. Eghbali is with Wireless System Engineering Technology in School of Information Communication and Engineering Technologies at Northern Alberta Institute of Technology (NAIT), Edmonton, AB, Canada (e-mail: homae@nait.ca).

I would like to thank Dr. Vincent Wong, Dr. Robert Schober, and Dr. Naofal Al-dhahir for reviewing this manuscript.

[1]It is the common nomenclature to refer to the interfaces with more stringent requirements on synchronization and latency as fronthaul interfaces. We use the term fronthaul to refer to the interfaces connecting the BSs to the SD-RAN controller. We use the term backhaul to refer to the links connecting the SD-RAN controller to the aggregation network and the mobile core network. The aggregation network provides connectivity to the adjacent RANs and the mobile core network. The mobile core network provides routing services to the geographically separated RANs in far locations and the Internet.

fiber consumption. They reported little loss of information under 50% compression ratio with an error vector magnitude (EVM) close to zero. The compression algorithm proposed in [9] can reduce the Long Term Evolution (LTE) traffic carried over the common public radio interface (CPRI) from 18 Gbps to 8 Gbps, and achieve a 44% compression ratio.

To further enhance SD-RAN deployments, real-time network virtualization solutions can be employed where network resources are managed as logical services, rather than physical resources. BS virtualization schemes are proposed to dynamically allocate the resources of BSs to meet the real-time demands of UEs [10]. By implementing virtualization for SD-RANs, real-time radio and fronthaul resources can be allocated on-demand and dynamically to the UEs. In this context, *virtualized* services are the radio and fronthaul resources that are reserved by the SD-RAN controller for the UEs to meet their minimum rate requirements. *Non-virtualized* services are the radio and fronthaul resources that are allocated to the UEs to provide them with best effort services. Furthermore, virtualized and non-virtualized UEs are the UEs which receive virtualized and non-virtualized services from the SD-RAN controller, respectively.

To the best of our knowledge, the problem of network utility maximization, virtualization, and user association for massive MIMO SD-RANs with the consideration of limited radio and fronthaul resources has not been discussed in the literature yet. To study utility maximization for massive MIMO SD-RANs, an expression for the achievable ergodic rate of the virtualized and non-virtualized UEs must be computed. To provide virtualization for massive MIMO SD-RANs and reserve the limited radio and fronthaul resources for the virtualized UEs, virtualized UEs must be prioritized over the non-virtualized UEs. To support user association for massive MIMO SD-RANs, we need to determine the fraction of transmission resources (time-frequency slots) over which the virtualized and non-virtualized UEs can be served by the BSs.

In this paper, we formulate a novel virtualization, user association, and rate maximization (VARM) problem for the downlink transmission of massive MIMO SD-RANs with limited radio and fronthaul resources. Since the existing works on user association for massive MIMO [3], [4], [11] assume that the network channel state information (CSI) is perfectly known, we derive asymptotic bounds on the achievable ergodic rate of virtualized and non-virtualized UEs for both perfect channel training and pilot contamination. These closed-form bounds do not depend only on the small-scale channel fading or the identity of the UEs. Instead, they depend on the large-scale channel fading and the number of UEs that are associated with the serving BSs. To further provide the virtualized UEs with higher quality of service (QoS) levels, we propose a hybrid virtualization and compression (HVC) method in which the SD-RAN controller transmits direct (non-compressed) messages to the virtualized UEs. The remaining radio and fronthaul resources are used to carry compressed data messages for the non-virtualized UEs. Furthermore, to allocate network resources to the virtualized UEs first, we decouple the VARM problem into two subproblems, one for virtualized UEs and one for non-virtualized UEs. These two subproblems are inter-related. In particular, the amount of interference experienced by the non-virtualized UEs due to the transmissions of the SD-RAN BSs to the virtualized UEs (and vice-versa) is the coupling parameter between the two sub-problems. This motivates the formulation of a two-stage Stackelberg game model to capture the interaction between the virtualized and non-virtualized UEs. In the first stage of the game, the virtualized UEs determine their association strategy as well as the amount of virtualized resources they want to use. In the second stage, the non-virtualized UEs determine their association strategy, the amount of non-virtualized resources they want to use, as well as the BS fronthaul noise compression levels. Thereby, the virtualized UEs are the leaders and the non-virtualized UEs are the followers of the game. The followers make their decisions according to the association/allocation strategy of the virtualized UEs. The main contributions of this work are as follows:

- We investigate the application of the compression-after-precoding technique for massive MIMO SD-RANs and derive the closed-form expressions as well as asymptotic bounds for the achievable ergodic rates of the UEs, considering both perfect training and pilot contamination.
- We formulate the VARM problem for SD-RAN while taking into account the limited resources of radio links and fronthaul interfaces.
- We propose the HVC technique which provides non-compressed services to the virtualized UEs as well as compressed and best effort services to the non-virtualized UEs.
- We decouple VARM into two subproblems, one for virtualized UEs and one for non-virtualized UEs. Since the virtualized UEs must be prioritized over the non-virtualized UEs, we use a Stackelberg game model to characterize the interplay between the optimal strategies of the virtualized UEs and the non-virtualized UEs.
- The VARM problem for virtualized and non-virtualized services is non-linear and computationally intractable. We use an exact reformulation technique to transform each subproblem into a linear form and obtain the global optimal solution with reasonable complexity.
- We provide numerical results to corroborate the effectiveness of the proposed HVC technique for massive MIMO SD-RAN. Numerical results show that our scheme outperforms the maximum peak rate association scheme [11], in which UEs are associated to the BS that provides them with the maximum rate.
- Furthermore, our results reveal that the aggregate achievable rate of the UEs follows a diminishing return pattern with respect to the maximum number of served UEs per BS.

The rest of this paper is organized as follows: We present the network model in Section II. VARM formulation for virtualized and non-virtualized services and its solution via the two-stage Stackelberg game model are presented in Section III. Numerical results are presented in Section IV. Conclusions are drawn in Section V. Throughout this paper, we use the following notations: Boldface upper case letters denote matri-

ces. Boldface lower case letters denote column vectors. Italics denote scalars. (a, b, c) denotes a column vector with three elements. $[a\ b\ c]$ denotes a row vector with three elements. $\mathbf{a}[i]$ denotes the i^{th} element of vector $\mathbf{a}$. $\mathbf{A}^*$, $\mathbf{A}^T$, and $\mathbf{A}^H$ represent the conjugate, transpose, and Hermitian transpose of matrix $\mathbf{A}$, respectively. $\mathbf{I}_M$ and $\mathbf{0}_{M,N}$ denote the M by M identity matrix and the M by N zero matrix, respectively. $\mathcal{C}^{N\times M}$ denotes the set of all $N \times M$ matrices with complex entries. $|\cdot|$ and $\|\cdot\|_p$ denote the absolute value of a complex scalar and the l_p-norm of a vector, respectively. $tr(.)$ denotes the trace operator. $\overset{a.s.}{\rightarrow}$ denotes the almost sure convergence . χ^2_S denotes a Chi-square random variable with S degrees of freedom. $\mathcal{CN}\left(\mu, \sigma^2\right)$ denotes a circularly symmetric complex Gaussian random variable with mean μ and variance σ^2.

II. Network Model

We consider an SD-RAN providing cellular services to an LTE macrocell comprising a set of small cells $\mathcal{L} = \{1, \ldots, L\}$. In the following, we simply refer to these small cells as cells. This network architecture is motivated by cell split, an advanced technique for capacity improvement, where a macrocell is split into smaller cells, as is currently practiced in the context of the 3GPP Release 12 small-cell densification enhancement [12]. Cell $l \in \mathcal{L}$ comprises M_l BSs and N_l single antenna UEs. Fig. 1 shows an example of an SD-RAN. The fronthaul part links the BSs to the SD-RAN controller. The backhaul connects the SD-RAN controller with the mobile core network. We assume that all BSs belong to the same service provider and all UEs are the subscribers of this service provider. We use the notations $m_l \in \mathcal{M}_l = \{1, \ldots, M_l\}$ and $n_l \in \mathcal{N}_l = \{1, \ldots, N_l\}$ to index the BSs and UEs in cell $l \in \mathcal{L}$, respectively.

Each BS schedules transmission over contiguous time-frequency slots called resource blocks (RBs). Each RB comprises a block of orthogonal frequency division multiplexing subcarriers and symbols. Let A_{m_l} denote the number of antennas at BS m_l. In the massive MIMO regime, independent data streams are simultaneously transmitted to multiple UEs on the same RB. Let B_{m_l} denote the number of downlink data streams that BS m_l can transmit on a given RB. The wireless channel is modeled as block-fading including both large-scale and small-scale fading effects. Let $\mathbf{h}_{m_l,n_l} \in \mathcal{C}^{A_{m_l}\times 1}$ denote the uplink channel between BS m_l and UE n_l located in cell $l \in \mathcal{L}$. The vector $\mathbf{h}_{m_l,n_l} = \sqrt{\beta_{m_l,n_l}}\tilde{\mathbf{h}}_{m_l,n_l}$ comprises the large-scale fading coefficient β_{m_l,n_l} and the small-scale fading vector $\tilde{\mathbf{h}}_{m_l,n_l}$. The large-scale fading β_{m_l,n_l} depends on the distance between the BS m_l and UE n_l and includes the effects of pathloss and shadowing. β_{m_l,n_l} is assumed to remain constant for a large number of RBs. The small-scale fading vector $\tilde{\mathbf{h}}_{m_l,n_l}[i]$, $i \in \{1, \ldots, A_{m_l}\}$, is modeled as Rayleigh fading and is assumed to remain constant within a RB but change from one RB to the next. We refer to the massive MIMO regime as the case where A_{m_l} is at least an order of magnitude greater than the number of UEs served by BS m_l, i.e., $1 \ll B_{m_l} \ll A_{m_l}$. In the following, we study uplink training and channel estimation for SD-RAN. The channel statistics provided in Section II-A will be used in Section II-B2 to characterize the asymptotic performance of the SD-RAN with imperfect channel estimates caused by pilot contamination.

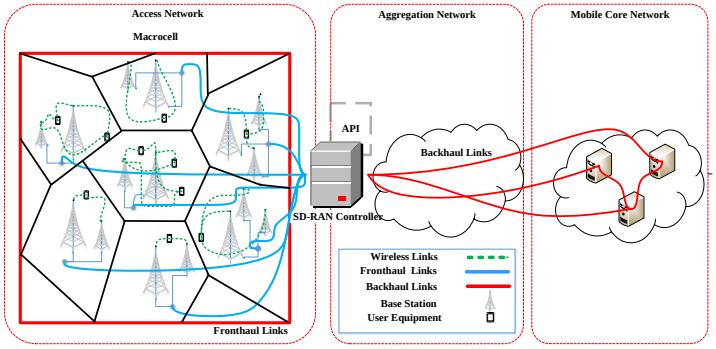

Fig. 1. System model of massive MIMO SD-RAN scenario.

A. Uplink Training and Channel Estimation

The large-scale amplitude coefficients β_{m_l,n_l}, $m_l \in \mathcal{M}_l, n_l \in \mathcal{N}_l, l \in \mathcal{L}$ are assumed to be known to the local BSs. However, the small-scale coefficient vectors $\tilde{\mathbf{h}}_{m_l,n_l}$ are not known to the local BSs and must be estimated by local BS m_l for downlink communications to UE n_l. We adopt a block-ergodic channel model, in which the small-scale fading channel coefficients are constant within the coherence period of the channel but vary in an ergodic fashion across a large number of coherence periods. We further assume uniform power allocation across the downlink data streams [11]. BS m_l employs reverse training and channel reciprocity to estimate the downlink CSI to UE n_l. We denote the channel estimate from BS m_l to UE n_l by $\hat{\mathbf{h}}_{m_l,n_l} \in \mathcal{C}^{A_{m_l}\times 1}$. In order to analyze the achievable rate for downlink transmission, we consider two scenarios: (a) perfect training where all UEs in the system transmit orthogonal pilot sequences and large pilot powers in the training phase such that $\hat{\mathbf{h}}_{m_l,n_l} = \tilde{\mathbf{h}}_{m_l,n_l}$, and (b) imperfect training where the pilot signals used in a cell are orthogonal, but all cells reuse the same set of orthogonal pilot sequences which causes pilot contamination [13].

Let $\mathbf{g}_{n_l} \in \mathcal{C}^{1\times\alpha}$ denote the pilot sequence of length α. $\sqrt{\alpha}\mathbf{g}_{n_l}$ is the pilot signal transmitted by UE n_l located in cell l. We have $\mathbf{g}_{n_l}\mathbf{g}_{n_l}^H = 1$, and $\mathbf{g}_{n_l'}\mathbf{g}_{n_l}^H = 0$, for $n_l, n_l' \in \mathcal{N}_l$, $n_l \neq n_l'$. Let p_α denote the power of the pilot signal of length α. We consider the channel training phase where the channel estimates in each cell are corrupted by pilot contamination from adjacent cells. We can statistically characterize the minimum mean-squared error (MMSE) estimate $\hat{\mathbf{h}}_{m_l,n_l}$ and the estimation error $\hat{\mathbf{e}}_{m_l,n_l}$ as $\hat{\mathbf{h}}_{m_l,n_l} \in \mathcal{CN}\left(\mathbf{0}_{A_{m_l}}, \frac{\alpha p_\alpha \tilde{\beta}_{m_l,n_l}(\mathbf{g}_{n_l})}{1+\alpha p_\alpha \sum_{j=1}^{L}\tilde{\beta}_{m_l,n_j}(\mathbf{g}_{n_l})}\mathbf{I}_{A_{m_l}}\right)$ and $\hat{\mathbf{e}}_{m_l,n_l} \in \mathcal{CN}\left(\mathbf{0}_{A_{m_l}}, \left(1-\frac{\alpha p_\alpha \tilde{\beta}_{m_l,n_l}(\mathbf{g}_{n_l})}{1+\alpha p_\alpha \sum_{j=1}^{L}\tilde{\beta}_{m_l,n_j}(\mathbf{g}_{n_l})}\right)\mathbf{I}_{A_{m_l}}\right)$, respectively. Here, $\tilde{\beta}_{m_l,n_j}(\mathbf{g}_{n_l})$ denotes the large-scale

channel coefficient between BS m_l and UE n_j from cell j that uses the pilot signal $\mathbf{g}_{n_l}$, i.e., the same pilot signal as UE n_l in cell l. Note that in the context of massive MIMO, to account for the effect of pilot contamination, we enforce the cellular concept by specifying the geographical locations of different cells. In the following, we characterize the performance behavior of SD-RAN in downlink transmission with perfect training and pilot contamination, respectively.

B. Performance of SD-RAN in Downlink Data Transmission

We initially assume that all BSs are transmitting to every UE in the macrocell. The resulting signal-to-interference and noise ratio (SINR) and rate expressions provide an insight regarding the large-scale behavior of the system and help to deduce the achievable rate of the UEs when each BS only transmits to the group of UEs that it is associated to. For notational convenience, we define $A_{BS_l} = \sum_{m_l=1}^{M_l} A_{m_l}$ as the total number of transmitting BS antennas in cell l, and $A_{BS} = \sum_{l=1}^{L} A_{BS_l}$ as the total number of transmitting BS antennas in the system. Moreover, we denote the total number of UEs in the system by $N = \sum_{l=1}^{L} N_l$ and the total number of BSs in the system by $M = \sum_{l=1}^{L} M_l$.

Let vector $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{C}^{N\times 1}$ contain the data symbols intended for all UEs in the system. Assuming simple matched filter (MF) precoding, let $\mathbf{d}_{n_{l'},m_l} = \left(\frac{\tilde{\mathbf{h}}_{m_l,n_{l'}}[1]}{\left\|\tilde{\mathbf{h}}_{m_l,n_{l'}}\right\|}, \ldots, \frac{\tilde{\mathbf{h}}_{m_l,n_{l'}}[A_{m_l}]}{\left\|\tilde{\mathbf{h}}_{m_l,n_{l'}}\right\|}\right) \in \mathcal{C}^{A_{m_l}\times 1}$ denote the beamforming vector at BS m_l for the data symbols intended for UE $n_{l'}$. We choose MF precoding since the matrix inversions required for zero-forcing (ZF) and MMSE precoding schemes are computationally expensive for the large number of users and antennas of massive MIMO systems. Let $\mathbf{d}_{n_{l'}} = \left(\mathbf{d}_{n_{l'},1}, \ldots, \mathbf{d}_{n_{l'},M}\right) \in \mathcal{C}^{A_{BS}\times 1}$ denote the precoding vector from the antennas of all BSs for the data symbols intended for UE $n_{l'}$. Let $\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_N] \in \mathcal{C}^{A_{BS}\times N}$ denote the matrix containing the precoding coefficients for the antennas of all BSs for the data symbols intended for all UEs in the system. The precoded vector transmitted from all BSs in the SD-RAN can be expressed as

$$\mathbf{s} = \mathbf{D}\mathbf{x}, \tag{1}$$

where $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_M) \in \mathcal{C}^{A_{BS}\times 1}$, and $\mathbf{s}_{m_l} \in \mathcal{C}^{A_{m_l}\times 1}$ denotes the precoded vector for BS m_l. $\mathbf{s}_{m_l}$ can be obtained as

$$\mathbf{s}_{m_l} = \mathbf{E}_{m_l}\mathbf{D}\mathbf{x}, \tag{2}$$

where $\mathbf{E}_{m_l} \in \mathcal{C}^{A_{m_l}\times A_{BS}} = \left[\mathbf{0}_{A_{m_l}\times \sum_{k=1}^{m_l-1} A_k}, \mathbf{I}_{A_{m_l}\times A_{m_l}}, \mathbf{0}_{A_{m_l}\times\left(A_{BS}-\sum_{k=1}^{m_l} A_k\right)}\right]$.

We assume that each BS m_l and the SD-RAN controller are connected by a pair of fronthaul interfaces with a fixed capacity denoted by C_{m_l} bps/Hz. One interface is for the downlink transmission between the SD-RAN controller and BS m_l. The other interface is for uplink transmission. For downlink transmission, the SD-RAN controller generates the baseband uncompressed in-phase and quadrature (IQ) precoded data samples which are then compressed by a compression module. The compressed IQ samples are transferred through the fronthaul interfaces to the BSs. In the BSs, the compressed IQ samples are decompressed through the decompression module. After decompression, the receiving BSs up-convert the decompressed baseband signals and transmit them to the UEs. The compression and decompression processing at the SD-RAN and the BSs effectively reduces the amount of data transferred through the bandwidth limited fronthaul interfaces. The proposed compression and decompression in SD-RAN is illustrated in Fig. 2. The compression module implemented in the SD-RAN controller consists of three steps: (a) removal of redundancies in the spectral domain; (b) block scaling; and (c) quantization. In the quantization step, the IQ samples are quantized using a quantizer with resolution less than the original bandwidth of the IQ data samples. This procedure introduces quantization noises. In order to model the effect of compression on the fronthaul interfaces, using standard rate distortion considerations, we adopt a Gaussian test channel and express the quantized signal received by BS m_l as

$$\hat{\mathbf{s}}_{m_l} = \mathbf{s}_{m_l} + \tilde{\mathbf{e}}_{m_l}, \tag{3}$$

where $\hat{\mathbf{s}}_{m_l}$ is the output signal after the compression and decompression processing performed at the SD-RAN controller and BS m_l, respectively, and $\tilde{\mathbf{e}}_{m_l} \in \mathcal{C}^{A_{m_l}\times 1}$ denotes the compression noise which can be modeled as a complex Gaussian vector distributed as $\mathcal{CN}(0, \mathbf{\Omega}_{m_l})$, where $\mathbf{\Omega}_{m_l} = \mathbb{E}\left[\left(\hat{\mathbf{s}}_{m_l} - \mathbf{s}_{m_l}\right)\left(\hat{\mathbf{s}}_{m_l} - \mathbf{s}_{m_l}\right)^H\right]$ constitutes the average squared error distortion between $\hat{\mathbf{s}}_{m_l}$ and $\mathbf{s}_{m_l}$. The compression error vector $\tilde{\mathbf{e}}_{m_l}$ is independent of $\mathbf{s}_{m_l}$ [14]. We assume independent quantization at each BS. This can be realized by using separate quantizers for the signals of different BSs [15]. We note that the possibility to leverage quantization noise correlation across different BSs using joint quantization techniques is explored in [6], [14]. In this work, however, we relax this point of complication as the implementation of joint quantization methods in large-scale massive MIMO SD-RANs requires the processing of very large channel matrices, leading to high computational complexity and channel estimation overhead [16]. We further assume $\mathbf{\Omega}_{m_l} = \Omega_{m_l}\mathbf{I}_{A_{m_l}}$ [17], where Ω_{m_l} denotes the quantization noise level per massive MIMO BS m_l. This is a reasonable assumption since the antennas of the massive MIMO BSs are closely positioned and have comparable system level characteristics. Considering (3), we realize that the design of the fronthaul compression for SD-RAN is equivalent to the optimization of the quantization noise variances Ω_{m_l} per BS. Let p denote the amount of power allocated to each UE from each BS. In here, we assume uniform power allocation across users. Moreover, we assume that the power of the transmitted beamforming vectors of BS m_l cannot exceed P_{m_l}. That is,

$$tr\left(p\mathbf{E}_{m_l}\mathbf{D}\mathbf{D}^H\mathbf{E}_{m_l}^T + p\Omega_{m_l}\mathbf{I}\right) \leq P_{m_l}, \tag{4}$$

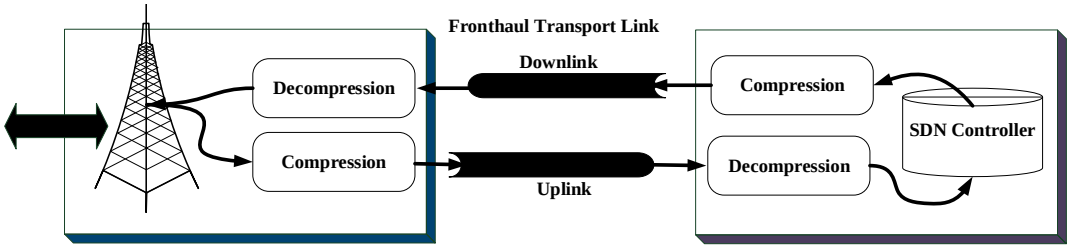

Fig. 2. SD-RAN system with fronthaul compression.

where $\mathbf{E}_{m_l}\mathbf{D} = \left[\begin{array}{ccc}\mathbf{d}_{1,m_l} & \dots & \mathbf{d}_{N,m_l}\end{array}\right] \in \mathcal{C}^{A_{m_l}\times N}$. As can be observed from (4), the quantization noise variances directly affect the transmit power of the BSs. Moreover, the rate allocated to the fronthaul interface of BS m_l is equal to $I\left(\hat{\mathbf{s}}_{m_l};\mathbf{s}_{m_l}\right)$, i.e., the mutual information between the precoded vector $\mathbf{s}_{m_l}$ and the output vector $\hat{\mathbf{s}}_{m_l}$. $I\left(\hat{\mathbf{s}}_{m_l};\mathbf{s}_{m_l}\right)$ characterizes the relationship between the quantization noise level Ω_{m_l} and the fronthaul capacity C_{m_l} as follows [18, Chapter 3]:

$$\log\det\left(\mathbf{E}_{m_l}\mathbf{D}\mathbf{D}^H\mathbf{E}_{m_l}^T + \Omega_{m_l}\mathbf{I}\right) - A_{m_l}\log\left(\Omega_{m_l}\right) \leq C_{m_l}. \tag{5}$$

In (42) in Appendix A, we have provided an asymptotic approximation for (5) when $N, A_{m_l} \to \infty$ with $\dfrac{N}{A_{m_l}} \to \nu_{m_l}$. The inequalities (4) and (42) will be used as two optimization constraints when formulating the VARM problem for massive MIMO SD-RAN.

BS m_j forwards the decompressed signal $\hat{\mathbf{s}}_{m_j}$ to the UEs. The received signal at UE n_l from cell l can be expressed as

$$\begin{aligned} y_{n_l} &= \sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sqrt{p}\mathbf{h}_{m_j,n_l}^H\left(\mathbf{s}_{m_j}+\tilde{\mathbf{e}}_{m_j}\right)+\tilde{n}_{n_l} \\ &= \sum_{k=1}^{L}\sqrt{p}\mathbf{H}_{k,n_l}^H\mathbf{s}_k + \sum_{k=1}^{L}\sqrt{p}\mathbf{H}_{k,n_l}^H\mathbf{e}_k+\tilde{n}_{n_l} \\ &= \sqrt{p}\mathbf{H}_{n_l}^H\mathbf{s}+\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{e}+\tilde{n}_{n_l} \\ &= \mathbf{H}_{n_l}^H\left(\sqrt{p}\mathbf{d}_{n_l}x_{n_l}+\sqrt{p}\sum_{k=1,k\neq n_l}^{N}\mathbf{d}_k x_k\right) \\ &\quad +\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{e}+\tilde{n}_{n_l}, \end{aligned} \tag{6}$$

where $\tilde{n}_{n_l} \in \mathcal{CN}\left(0,1\right)$ denotes the Gaussian noise at UE n_l with zero mean and unit variance, $\mathbf{H}_{k,n_l}^H = \left[\mathbf{h}_{1,n_l}^H\dots\mathbf{h}_{M_k,n_l}^H\right] \in \mathcal{C}^{1\times A_{BS_k}}$, $\mathbf{H}_{n_l}^H = \left[\mathbf{H}_{1,n_l}^H\dots\mathbf{H}_{L,n_l}^H\right] \in \mathcal{C}^{1\times A_{BS}}$, $\mathbf{s}_k = \left(\mathbf{s}_1,\dots,\mathbf{s}_{M_k}\right) \in \mathcal{C}^{A_{BS_k}\times 1}$, $\mathbf{e}_k = \left(\tilde{\mathbf{e}}_1,\dots,\tilde{\mathbf{e}}_{M_k}\right) \in \mathcal{C}^{A_{BS_k}\times 1}$, and $\mathbf{e} = \left(\mathbf{e}_1,\dots,\mathbf{e}_L\right) \in \mathcal{C}^{A_{BS}\times 1}$. We can rewrite the received signal in (6) as

$$y_{n_l} = \mathbb{E}\left[\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right]x_{n_l}+\widehat{n}_{n_l}, \tag{7}$$

where $\mathbb{E}\left[\cdot\right]$ represents the expectation operator with respect to the channel vectors and $\widehat{n}_{n_l}$ represents the effective noise. $\widehat{n}_{n_l}$ can be expressed as

$$\begin{aligned} \widehat{n}_{n_l} &= \left(\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l} - \mathbb{E}\left[\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right]\right)x_{n_l} \\ &\quad + \sum_{k=1,k\neq n_l}^{N}\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_k x_k + \sqrt{p}\mathbf{H}_{n_l}^H\mathbf{e}+\tilde{n}_{n_l}. \end{aligned} \tag{8}$$

Following (7) and (8), the SINR ψ_{n_l} can be expressed as in (9) where

$$\mathrm{var}\left(\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right) = \mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right|^2\right] - \left(\mathbb{E}\left[\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right]\right)^2. \tag{10}$$

1) Rate Analysis with Perfect Training: We proceed by statistically characterizing the terms in (9). We compute the first and second order moments of the effective channel gain and the inter-cell and intra-cell interference to obtain a simple expression for the achievable rate that solely depends on the large-scale parameters of the SD-RAN. Note that $\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_{n_l}\right|^2$ is a sum of scaled chi-square random variables and is statistically equivalent to $u_{n_l}^2$ given below

$$\begin{aligned} u_{n_l}^2 &= \sum_{i=1}^{L}\sum_{m_i=1}^{M_i} p\beta_{m_i,n_l}\left(\tilde{\mathbf{h}}_{m_i,n_l}^H\frac{\tilde{\mathbf{h}}_{m_i,n_l}}{\left\|\tilde{\mathbf{h}}_{m_i,n_l}\right\|}\frac{\tilde{\mathbf{h}}_{m_i,n_l}^H}{\left\|\tilde{\mathbf{h}}_{m_i,n_l}\right\|}\tilde{\mathbf{h}}_{m_i,n_l}\right) \\ &= \sum_{i=1}^{L}\sum_{m_i=1}^{M_i} p\beta_{m_i,n_l}\sum_{k=1}^{A_{m_i}}\left|\tilde{\mathbf{h}}_{m_i,n_l}\left[k\right]\right|^2 \\ &= \sum_{i=1}^{L}\sum_{m_i=1}^{M_i} p\beta_{m_i,n_l}x_{m_i}^2, \end{aligned} \tag{11}$$

where $x_{m_i}^2 = \sum_{k=1}^{A_{m_i}}\hat{u}_k^2 \sim \chi_{2A_{mi}}^2$ and $\hat{u}_k$ are independent identically distributed (i.i.d.) $\mathcal{CN}\left(0,1\right)$. Moreover, the interference term $\mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_k\right|^2\right]$ at the denominator of (9) can be bounded as (equation (47) in Appendix C):

$$\mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{d}_k\right|^2\right] \leq \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} pA_{m_j}\beta_{m_j,n_l}\left|\tilde{h}_{m_j,n_l}^{\max}\right|^2, \tag{12}$$

where $\left|\tilde{h}_{m_j,n_k}^{\max}\right| = \max_{1\leq i\leq A_{m_j}}\left(\left|\tilde{\mathbf{h}}_{m_j,n_k}\left[i\right]\right|\right)$. The last interference term $\mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{e}\right|^2\right]$ in (9) can be expressed as:

$$\mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H\mathbf{e}\right|^2\right] = \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} pA_{m_j}\beta_{m_j,n_l}\Omega_{m_j}. \tag{13}$$

Using (11) and (13), the SINR ψ_{n_l} in (9) can be expressed as in (14). Note that for large A_{m_j}, we have $\lim_{A_{m_j}\to\infty}\dfrac{\left(\mathbb{E}\left[x_{m_j}\right]\right)^2}{A_{m_j}} = 1$, $\lim_{A_{m_j}\to\infty}\dfrac{\mathbb{E}\left[x_{m_j}^2\right]}{A_{m_j}} = 1$, and $\lim_{A_{m_j}\to\infty}\mathrm{var}\left(x_{m_j}\right) = 0$ [13]. We assume that $\dfrac{A_{m_j}}{A_{BS}} \to \eta_{m_j}$, when $A_{m_j}\to\infty$, $m_j \in \mathcal{M}_j, j\in\mathcal{L}$ and $A_{BS}\to\infty$. Dividing the numerator and the denominator of (14) by A_{BS}, (14) can be asymptotically approximated as

$$\psi_{n_l} \overset{a.s.}{\to} \frac{\sum_{j=1}^{L}\sum_{m_j=1}^{M_j} p\eta_{m_j}\beta_{m_j,n_l}}{\sum_{j=1}^{L}\sum_{m_j=1}^{M_j} p\eta_{m_j}\beta_{m_j,n_l}\left(\sum_{k=1,k\neq n_l}^{N}\left|\tilde{h}_{m_j,k}^{\max}\right|^2+\Omega_{m_j}\right)+1}. \tag{15}$$

$$\psi_{n_l} = \frac{\left(\mathbb{E}\left[\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{d}_{n_l}\right]\right)^2}{\text{var}\left(\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{d}_{n_l}\right) + \sum\limits_{k=1,k\neq n_l}^{N} \mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{d}_k\right|^2\right] + \mathbb{E}\left[\left|\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{e}\right|^2\right] + 1}, \tag{9}$$

$$\psi_{n_l} \overset{a.s.}{\to} \frac{\sum\limits_{j=1}^{L}\sum\limits_{m_j=1}^{M_j} p\beta_{m_j,n_l}\left(\mathbb{E}\left[x_{m_j}\right]\right)^2}{\sum\limits_{j=1}^{L}\sum\limits_{m_j=1}^{M_j} p\beta_{m_j,n_l}\text{var}\left(x_{m_j}\right) + \sum\limits_{j=1}^{L}\sum\limits_{m_j=1}^{M_j} pA_{m_j}\beta_{m_j,n_l}\left(\sum\limits_{k=1,k\neq n_l}^{N}\left|\tilde{h}_{m_j,k}^{\max}\right|^2 + \Omega_{m_j}\right) + 1}. \tag{14}$$

Note that $p\eta_{m_j}\beta_{m_j,n_l}\left(\sum\limits_{k=1,k\neq n_l}^{N}\left|\tilde{h}_{m_j,k}^{\max}\right|^2 + \Omega_{m_j}\right)$ is the amount of interference that each BS contributes to the aggregate intra/inter cell interference caused to UE n_l, assuming that the BS transmits to all UEs. Therefore, if BS m_j serves a group of B_{m_j} UEs, the aggregate interference caused by this BS to the UE n_l adds up to $p\eta_{m_j}\beta_{m_j,n_l}\left(\sum\limits_{k=1,k\neq n_l}^{B_{m_j}}\left|\tilde{h}_{m_j,k}^{\max}\right|^2 + \Omega_{m_j}\right)$. Thus, assuming that each BS m_j that is transmitting to n_l, transmits to a total of B_{m_j} UEs, (15) can be expressed as follows:

$$\psi_{n_l} \overset{a.s.}{\to} \frac{\sum\limits_{j=1}^{L}\sum\limits_{m_j=1}^{M_j} p\eta_{m_j}\beta_{m_j,n_l}}{\sum\limits_{j=1}^{L}\sum\limits_{m_j=1}^{M_j} p\eta_{m_j}\beta_{m_j,n_l}\left(B_{m_j} + \Omega_{m_j} - 1\right) + 1}, \tag{16}$$

where, without loss of generality, it is assumed that $\max\limits_{k=1,k\neq n_l}^{B_{m_j}}\left|\tilde{h}_{m_j,k}^{\max}\right| = 1$ which gives the lower bound for ψ_{n_l}. Note that without compression and considering single user single antenna communications, (16) reduces to the standard well-known SINR formula $\frac{p\beta_{m_j,n_l}}{\sum\limits_{m_k\neq m_j} p\beta_{m_k,n_l} + 1}$, which accounts for large-scale channel effects only.

2) Rate Analysis with Pilot Contamination: Considering the small-scale channel vectors $\tilde{\mathbf{h}}_{m_l,n_l} = \hat{\mathbf{h}}_{m_l,n_l} + \hat{\mathbf{e}}_{m_l,n_l}$, in order to simplify (9), we note that with pilot contamination,

$$\begin{aligned}\mathbb{E}\left[\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{d}_{n_l}\right] &= \mathbb{E}\left[\left|\sqrt{p}\hat{\mathbf{H}}_{n_l}^H\right| + \sqrt{p}\hat{\mathbf{e}}_{n_l}^H \frac{\hat{\mathbf{H}}_{n_l}}{\left|\hat{\mathbf{H}}_{n_l}\right|}\right] \\ &= \sum_{i=1}^{L}\sum_{m_i=1}^{M_i}\sqrt{\frac{p\alpha p_\alpha \tilde{\beta}_{m_i,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{j=1}^{L}\tilde{\beta}_{m_i,n_j}\left(\mathbf{g}_{n_l}\right)}}\mathbb{E}\left[x_{m_i}\right],\end{aligned} \tag{17}$$

where $\hat{\mathbf{H}}_{n_l}^H = \left[\hat{\mathbf{H}}_{1,n_l}^H \ldots \hat{\mathbf{H}}_{L,n_l}^H\right] \in \mathcal{C}^{1\times A_{BS}}$, $\hat{\mathbf{H}}_{k,n_l}^H = \left[\hat{\mathbf{h}}_{1,n_l}^H \ldots \hat{\mathbf{h}}_{M_k,n_l}^H\right] \in \mathcal{C}^{1\times A_{BS_k}}$, and $\hat{\mathbf{e}}_{n_l} = \left(\hat{\mathbf{e}}_{n_l,1},\ldots,\mathbf{e}_{n_l,L}\right) \in \mathcal{C}^{A_{BS}\times 1}$, and $\hat{\mathbf{e}}_{n_l,k} = \left(\hat{\mathbf{e}}_{n_l,m_1},\ldots,\mathbf{e}_{n_l,M_k}\right) \in \mathcal{C}^{A_{BS_k}\times 1}$. Also note that

$$\begin{aligned}&\text{var}\left(\sqrt{p}\mathbf{H}_{n_l}^H \mathbf{d}_{n_l}\right) = \mathbb{E}\left[\left|\sqrt{p}\hat{\mathbf{H}}_{n_l}^H\right|^2\right] \\ &+\mathbb{E}\left[p\frac{\hat{\mathbf{H}}_{n_l}^H}{\left|\hat{\mathbf{H}}_{n_l}\right|}\hat{\mathbf{e}}_{n_l}^H\hat{\mathbf{e}}_{n_l}\frac{\hat{\mathbf{H}}_{n_l}}{\left|\hat{\mathbf{H}}_{n_l}\right|}\right] - \left(\mathbb{E}\left[\sqrt{p}\hat{\mathbf{H}}_{n_l}^H \mathbf{d}_{n_l}\right]\right)^2 \\ &= \sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{p\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)}\text{var}\left(x_{m_j}\right) \\ &+ \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} p\left(1 - \frac{\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)}\right).\end{aligned} \tag{18}$$

Therefore, the lower bound on the achievable ergodic rate can be expressed as follows:

$$\begin{aligned}\psi_{n_l} = &\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{p\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)}\left(\mathbb{E}\left[x_{m_j}\right]\right)^2 \Bigg/ \\ &\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\left(\frac{p\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)}\text{var}\left(x_{m_j}\right) + \right. \\ &+ \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} pA_{m_j}\tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)\left(\sum_{k=1,k\neq n_l}^{N}\left|\tilde{h}_{m_j,k}^{\max}\right|^2 + \Omega_{m_j}\right) \\ &\left. p\left(1 - \frac{\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)}\right) + 1\right).\end{aligned} \tag{19}$$

Note that when $A_{m_j} \to \infty$, following similar approach used to derive (15), (19) can be asymptotically expressed as in (20), where $u_{m_j} = \frac{p\alpha p_\alpha \tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)}{\left(1 + \alpha p_\alpha \sum\limits_{i=1}^{L}\tilde{\beta}_{m_j,n_i}\left(\mathbf{g}_{n_l}\right)\right)}$. Following similar analogy used to derive (16), assuming that each BS m_j that is transmitting to n_l, transmits to total B_{m_j} UEs, ψ_{n_l} can be expressed as in (21). Following (16) and (21), we arrive at the following result:

Note 1: Focusing on the massive MIMO regime, assuming that UE n_l is served by the BS m_j, the aggregate interference imposed on n_l from m_j, knowing that m_j is serving $B_{m_j} - 1$

$$\psi_{n_l} \overset{a.s.}{\to} \frac{\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\eta_{m_j}u_{m_j}}{\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\left(p-\eta_{m_j}u_{m_j}+p\eta_{m_j}\tilde{\beta}_{m_j,n_l}\left(\mathbf{g}_{n_l}\right)\left(\sum_{k=1,k\neq n_l}^{N}\left|\tilde{h}^{\max}_{m_j,n_l}\right|^2+\Omega_{m_j}\right)\right)}, \tag{20}$$

$$\psi_{n_l} \overset{a.s.}{\to} \frac{\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\eta_{m_j}u_{m_j}}{\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}\left(p-\eta_{m_j}u_{m_j}+p\eta_{m_j}\tilde{\beta}_{m_k,n_l}\left(\mathbf{g}_{n_l}\right)\left(B_{m_k}-1\right)+\Omega_{m_k}\right)+1}. \tag{21}$$

other UEs, effectively, only depends on the value of B_{m_j} and not on the identity of those UEs served by m_j.

III. Virtualization, User Association, and Rate Maximization

In this section, we propose VARM which maximizes the aggregate utility of the massive MIMO SD-RAN. Thereby, VARM associates virtualized and non-virtualized UEs to the massive MIMO BSs, and also allocates the radio and fronthaul resources of the BSs to the UEs that they are associated to. Considering ψ_{n_l} in (16) and (21), we express the achievable SINR at UE n_l as follows:

$$\psi_{n_l} = \sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\psi_{n_{l,m_j}}, \tag{22}$$

where $\psi_{n_{l,m_j}}$ denotes the achievable SINR at UE n_l which receives data from BS m_j. We use the index n_{l,m_j} to indicate that UE n_l is associated to BS m_j. If UE n_l is not associated to the BS m_j, then $\psi_{n_{l,m_j}} = 0$. Note that with perfect CSI, $\psi_{n_{l,m_j}}$ can be asymptotically expressed as follows:

$$\psi_{n_{l,m_j}} \overset{a.s.}{\to} \frac{p\eta_{m_j}\beta_{m_j,n_l}}{\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}p\eta_{m_k}\beta_{m_k,n_l}\left(B_{m_k}+\Omega_{m_k}-1\right)+1}. \tag{23}$$

With pilot contamination, $\psi_{n_{l,m_j}}$ can be expressed as in (24). In order to determine the optimized UE-BS association as well as the radio and fronthaul resource allocation decision, we propose a novel HVC technique in which virtualization, user association, rate maximization problem for massive MIMO SD-RAN is decomposed into two separate sub-problems for virtualized and non-virtualized services, respectively. UEs demanding virtualized services are prioritized over the UEs demanding non-virtualized services. This is justified by noting that virtualized users pay higher service fees to the service provider. The service provider prioritizes these UEs to meet the service level agreements (SLAs). The UE demanding virtualized services has its non-compressed message sent directly to its serving BSs. Messages intended for UEs demanding non-virtualized services with lower QoS requirements, however, are compressed and suffer from the compression error.

The amount of interference caused to the virtualized UEs by the non-virtualized UEs and vice-versa, is dependent on the UE to BS association and the resource allocation decisions made in each subproblem. Since the virtualized UEs are prioritized over the non-virtualized UEs, we formulate a two-stage Stackelberg game model to capture the interaction between the virtualized and non-virtualized UEs. In the first stage of the game, the virtualized UEs determine their BS association strategy, as well as the amount of virtualized resources they want to use. In the second stage, the association and resource consumption strategies of the non-virtualized UEs as well as the compression noise levels are determined. We assume that the virtualized UEs are the first movers and the non-virtualized UEs are the followers that make their decisions according to the association and resource consumption strategies of the virtualized UEs.

Step II: The Best Response of UEs with Non-virtualized Service Demands

In order to analyze the Stackelberg game, we first consider the second stage of the game which aims to maximize the utility of the non-virtualized UEs, given the association and resource allocation strategies of the virtualized UEs. VARM for non-virtualized UEs can be formulated as follows:

$$\underset{\Omega_{m_j},f_{n^{nv}_{l,m_j}},R_{n^{nv}_l}}{\text{maximize}} \sum_{l=1}^{L}\sum_{n^{nv}_l=1}^{N^{nv}_l}R_{n^{nv}_l} \tag{25a}$$

$$\text{subject to} \quad R_{n^{nv}_l} \leq \sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\bar{\psi}_{n^{nv}_{l,m_j}},$$

$$\forall n^{nv}_l \in \mathcal{N}^{nv}_l, l \in \mathcal{L}, \tag{25b}$$

$$\sum_{l=1}^{L}\sum_{n^{nv}_l=1}^{N^{nv}_l}\log_2\left(1+\bar{\psi}_{n^{nv}_{l,m_j}}\right) \leq \frac{\tilde{C}_{m_j}}{\varpi},$$

$$\forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{25c}$$

$$\sum_{l=1}^{L}\sum_{n^{nv}_l=1}^{N^{nv}_l}f_{n^{nv}_{l,m_j}} \leq B_{m_j} - \sum_{l=1}^{L}\sum_{n^{v}_l=1}^{N^{v}_l}f_{n^{v}_{l,m_j}},$$

$$\forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{25d}$$

$$\psi_{n_{l,m_j}} \overset{a.s.}{\rightarrow} \frac{\eta_{m_j} u_{m_j}}{\sum\limits_{k=1}^{L}\sum\limits_{m_k=1}^{M_k}\left(p-\eta_{m_k}u_{m_j}+p\eta_{m_k}\tilde{\beta}_{m_k,n_l}\left(\mathbf{g}_{n_l}\right)\left(B_{m_k}-1\right)+\Omega_{m_k}\right)+1}. \tag{24}$$

$$\sum_{j=1}^{L}\sum_{m_j=1}^{M_j} f_{n_{l,m_j}^{nv}} \leq 1, \qquad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, \tag{25e}$$

$$f_{n_{l,m_j}^{nv}} \geq 0, \quad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{25f}$$

$$\Omega_{m_j} \geq \frac{\log e}{\sqrt{\left(\frac{\tilde{C}_{m_j}}{A_{m_j}}\right)^2 + \frac{2\log e\tilde{C}_{m_j}}{A_{m_j}}} - \frac{\tilde{C}_{m_j}}{A_{m_j}}}, \qquad \forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{25g}$$

$$\sum_{l=1}^{L}\sum_{n_l^{nv}=1}^{N_l^{nv}} f_{n_{l,m_j}^{nv}} p\left(1+\eta_{m_j}\Omega_{m_j}\right) \leq P_{m_j} - \sum_{l=1}^{L}\sum_{n_l^{v}=1}^{N_l^{v}} f_{n_{l,m_j}^{v}} p, \qquad \forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{25h}$$

$$R_{n_l^{nv}} \geq 0, \qquad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, \tag{25i}$$

where n_l^{nv} is used to index the non-virtualized UEs located in cell l, n_{l,m_j}^{nv} is used to associate the non-virtualized UE n_l^{nv} to BS m_j, $f_{n_{l,m_j}^{nv}}$ denotes the scheduling activity fraction of the non-virtualized UE n_{l,m_j}^{nv} at BS m_j, i.e. the fraction of RBs over which n_{l,m_j}^{nv} is in the scheduled active subset of BS m_j, $\mathcal{N}_l^{nv}$ denotes the set of non-virtualized UEs located in cell l where $|\mathcal{N}_l^{nv}| = N_l^{nv}$, $R_{n_l^{nv}}$ denotes the achievable rate by the non-virtualized UE n_l^{nv},

$$\bar{\psi}_{n_{l,m_j}^{nv}} \overset{a.s.}{\rightarrow} \frac{p\eta_{m_j} f_{n_{l,m_j}^{nv}} \beta_{m_j,n_{l,m_j}^{nv}}}{\sum\limits_{k=1}^{L}\sum\limits_{m_k=1}^{M_k}\left(p\eta_{m_k}\beta_{m_k,n_{l,m_k}^{nv}}\left(\alpha_{m_k}^{nv}+\alpha_{m_k}^{v}+\Omega_{m_k}\right)\right)+1}, \tag{26}$$

$$\alpha_{m_k}^{nv} = \sum_{n_{z,m_k}^{nv} \neq n_{l,m_j}^{nv}} f_{n_{z,m_k}^{nv}}, \; \alpha_{m_k}^{v} = \sum_{n_{z,m_k}^{v}} f_{n_{z,m_k}^{v}},$$

$$\tilde{C}_{m_j} = C_{m_j} - \sum_{l=1}^{L}\sum_{n_l^{v}=1}^{N_l^{v}} f_{n_{l,m_j}^{v}} \gamma_{n_{l,m_j}^{v}}^{v}, \tag{27}$$

ϖ denotes the compression ratio, n_l^v is used to index the virtualized UEs located in cell l, n_{l,m_j}^v is used to associate the virtualized UE n_l^v to the BS m_j, f_{n_{l,m_j}^v} denotes the scheduling activity fraction of the virtualized UE n_{l,m_j}^v at BS m_j, i.e. the fraction of RBs over which n_{l,m_j}^v is in the scheduled active subset of BS m_j, and $\gamma_{n_l^v}^v$ denotes the minimum rate requirement of UE n_l^v. Note that the objective function (25a) follows from the definition of the aggregate throughput of user n_l over all its serving BSs. In constraint (25b), to make the problem computationally tractable, we have changed the maximum sum-rate problem to the problem of maximizing the achievable SINR at non-virtualized UEs. Constraint (25c) ensures that the allocated wireless resources through the fronthaul network does not exceed the amount of fronthaul resources available to BS m_j multiplied by inverse of the compression ratio ϖ. Constraint (25d) reflects the fact that the sum of activity fractions of non-virtualized UEs served by any BS cannot exceed the number of simultaneous downstream data streams associated with that BS subtracting the resources allocated to the virtualized UEs. Constraint (25e) ensures that the total scheduling activity fraction of any user cannot exceed 1. Constraint (25g) ensures that messages can be reliably transferred to the BSs through the fronthaul network. Constraint (25h) enforces the maximum transmit power per BS. Note that in (25d) and (27), we have limited the available SD-RAN resources per BS for non-virtualized services to the total available radio and fronthaul resources per BS subtracting the amount of aggregate resources already allocated to the virtualized UEs. Note that the problem in (25) is a mixed integer non-linear program (MINLP), which even for small dimensions, is difficult to solve while providing global optimality guarantee. To linearize (25h), we use a linearization technique which deals with products of a binary and a continuous variable [19].

Linearization Technique I: A product of a binary variable x and a continuous positive variable y can be replaced by an auxiliary continuous variable $z = xy$, along with a set of linear constraint expressions: $y - z \leq M_y(1-x)$, $z \leq y$, $z \leq M_y x$, and $z \geq 0$, where M_y is a large number guaranteed to be greater than the maximum value that y can take.

Therefore, assuming that Q is a large positive number which is at least equal to the maximum expected value for Ω_{m_j}, we introduce the auxiliary variable $u_{n_{l,m_j}^{nv}}$ to replace $f_{n_{l,m_j}^{nv}}\Omega_{m_j}$ along with the necessary linear constraint sets expressed below:

$$\Omega_{m_j} - u_{n_{l,m_j}^{nv}} \leq Q\left(1 - f_{n_{l,m_j}^{nv}}\right), \qquad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{28a}$$

$$u_{n_{l,m_j}^{nv}} \leq \Omega_{m_j}, \quad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{28b}$$

$$u_{n_{l,m_j}^{nv}} \leq Q f_{n_{l,m_j}^{nv}}, \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in L, m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{28c}$$

$$u_{n_{l,m_j}^{nv}} \geq 0, \qquad \forall n_l^{nv} \in \mathcal{N}_l^{nv}, l \in \mathcal{L}, m_j \in \mathcal{M}_j, j \in \mathcal{L}. \tag{28d}$$

Note that the reformulation technique used to eliminate the nonlinearity in constraint (25h) provides an exact transformation of the original problem variables, and no approximation or relaxation penalty is induced. The price we have to pay is the increase in the number of variables and constraint expressions caused by the introduction of

auxiliary variables. Moreover, applying a simple arrangement of terms, we obtain (29) where $R_{n_l^{nv}} \leq \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} R_{n_{l,m_j}^{nv}}$. For the term $R_{n_{l,m_j}^{nv}} f_{n_{z,m_k}^{nv}}$, we use the linearization technique I introduced previously. For the product term $R_{n_{l,m_j}^{nv}}\Omega_{m_k}$, we use a linearization technique which deals with products of two continuous variables.

Linearization Technique II: A product of a continuous positive variable x and a continuous positive variable y can be replaced by a new continuous auxiliary variable $z = xy$, along with a linear constraint expression $l_1 y \leq z \leq u_1 y$, knowing that $l_1 \leq x \leq u_1$.

Following the linearization technique II, we introduce the auxiliary variable $\tilde{u}_{n_{l,z,m_j,m_k}^{nv}}$ to replace $R_{n_{l,m_j}^{nv}}\Omega_{m_k}$ along with the necessary linear constraint expressed below:

$$\frac{\log e R_{n_{l,m_j}^{nv}}}{\sqrt{\left(\frac{\tilde{C}_{m_k}}{A_{m_k}}\right)^2 + \frac{2\log e\tilde{C}_{m_k}}{A_{m_k}}} - \frac{\tilde{C}_{m_k}}{A_{m_k}}} \leq \tilde{u}_{n_{l,m_j}^{nv}} \leq Q R_{n_{l,m_j}^{nv}}. \tag{30}$$

In order to linearize (25c), we first use the Trapezoidal Eulers Log rule [20] $\ln(1+x) \approx x\left(\frac{1+0.5x}{1+x}\right)$. Assuming that the minimum required $\bar{\psi}_{n_{l,m_j}^{v}}$ is 0 dB, for computing $\ln(2)$, the Trapezoidal Eulers Log approximation results in 5.069×10^{-2} error compared to the exact value. Other approximation methods such as the infinite series technique using the first ten terms or the $\frac{1}{3}-$SELOG technique [20] result in -5.248774×10^{-2} and -1.297264×10^{-3} error compared to the exact value, respectively. As can be seen, the Trapezoidal Eulers Log rule results in a sufficiently accurate approximation while minimizing the incurred complexity. We further apply the piecewise linear approximation of $\frac{1}{2}x^2$ which can be written as follows:

$$\begin{aligned} &\frac{\lambda_2}{2} + 2\lambda_3 + 8\lambda_4 = \frac{1}{2}x^2, \\ &\lambda_2 + 2\lambda_3 + 4\lambda_4 = x, \\ &\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1. \end{aligned} \tag{31}$$

Following (31) and (26), constraint (25c) can be linearly approximated using the following set of constraints:

$$\begin{aligned} &\left(\lambda_{2,n_{l,m_j}^{nv}} + 2\lambda_{3,n_{l,m_j}^{nv}} + 4\lambda_{4,n_{l,m_j}^{nv}}\right)\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}\left(p\eta_{m_k}\beta_{m_k,n_{l,m_k}^{nv}}\right. \\ &\left.\left(\sum_{n_{z,m_k}^{nv}\neq n_{l,m_j}^{nv}} f_{n_{z,m_k}^{nv}} + \alpha_{m_k}^{v} + \Omega_{m_k}\right)\right) + 1 \\ &= p\eta_{m_j} f_{n_{l,m_j}^{nv}}\beta_{m_j,n_{l,m_j}^{nv}}, \\ &\lambda_{1,n_{l,m_j}^{nv}} + \lambda_{2,n_{l,m_j}^{nv}} + \lambda_{3,n_{l,m_j}^{nv}} + \lambda_{4,n_{l,m_j}^{nv}} = 1, \\ &\left(\lambda_{2,n_{l,m_j}^{nv}} + 2\lambda_{3,n_{l,m_j}^{nv}} + 4\lambda_{4,n_{l,m_j}^{nv}}\right) \\ &-\left(\frac{\lambda_{2,n_{l,m_j}^{nv}}}{2} + 2\lambda_{3,n_{l,m_j}^{nv}} + 8\lambda_{4,n_{l,m_j}^{nv}}\right) \\ &\leq \frac{1+\lambda_{2,n_{l,m_j}^{nv}} + 2\lambda_{3,n_{l,m_j}^{nv}} + 4\lambda_{4,n_{l,m_j}^{nv}}}{\varpi}\tilde{C}_{n_{l,m_j}^{nv}}, \end{aligned} \tag{32}$$

where $\sum_{l=1}^{L}\sum_{n_l^{nv}=1}^{N_l^{nv}} \tilde{C}_{n_{l,m_j}^{nv}} = \tilde{C}_{m_j}$.

Step I: VARM for UEs with Virtualized Service Demands The rates requested by the virtualized UEs must be greater than a pre-specified minimum threshold. VARM for virtualized services can be formulated as follows:

$$\underset{f_{n_{l,m_j}^{v}}, R_{n_l^v}}{\text{maximize}} \quad \sum_{l=1}^{L}\sum_{n_l^v=1}^{N_l^v} R_{n_l^v} \tag{33a}$$

$$\text{subject to} \quad R_{n_l^v} \leq \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} \bar{\psi}_{n_{l,m_j}^{v}}, \forall n_l^v \in \mathcal{N}_l^v, l \in \mathcal{L}, \tag{33b}$$

$$R_{n_l^v} \geq \gamma_{n_l^v}^{v}, \qquad \forall n_l^v \in \mathcal{N}_l^v, l \in \mathcal{L}, \tag{33c}$$

$$\sum_{l=1}^{L}\sum_{n_l^v=1}^{N_l^v} \log_2\left(1+\bar{\psi}_{n_{l,m_j}^{v}}\right) \leq C_{m_j}, \forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{33d}$$

$$\sum_{l=1}^{L}\sum_{n_l^v=1}^{N_l^v} f_{n_{l,m_j}^{v}} p \leq P_{m_j}, \qquad \forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{33e}$$

$$\sum_{l=1}^{L}\sum_{n_l^v=1}^{N_l^v} f_{n_{l,m_j}^{v}} \leq B_{m_j}, \qquad \forall m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{33f}$$

$$\sum_{j=1}^{L}\sum_{m_j=1}^{M_j} f_{n_{l,m_j}^{v}} \leq G, \qquad \forall n_l^v \in \mathcal{N}_l^v, l \in \mathcal{L}, \tag{33g}$$

$$f_{n_{l,m_j}^{v}} \geq 0, \qquad \forall n_l^v \in \mathcal{N}_l^v, l \in \mathcal{L}, m_j \in \mathcal{M}_j, j \in \mathcal{L}, \tag{33h}$$

where $\mathcal{N}_l^v$ denotes the set of virtualized UEs located in cell l where $|\mathcal{N}_l^v| = N_l^v$, $\mathcal{N}_l = \mathcal{N}_l^v \cup \mathcal{N}_l^{nv}$, $\mathcal{N}_l^v \cap \mathcal{N}_l^{nv} = \emptyset$, $R_{n_l^v}$ denotes the achievable rate by the virtualized UE n_l^v,

$$\bar{\psi}_{n_{l,m_j}^{v}} \overset{a.s.}{\to} \frac{p\eta_{m_j} f_{n_{l,m_j}^{v}}\beta_{m_j,n_{l,m_j}^{v}}}{\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}\left(p\eta_{m_k}\beta_{m_k,n_{l,m_k}^{v}}\left(\tilde{\alpha}_{m_k}^{v} + \tilde{\alpha}_{m_k}^{nv}\right)\right) + 1}, \tag{34}$$

$\tilde{\alpha}_{m_k}^{v} = \sum_{n_{z,m_k}^{v}\neq n_{l,m_k}^{v}} f_{n_{z,m_k}^{v}}$, $\tilde{\alpha}_{m_k}^{nv} = \sum_{n_{z,m_k}^{nv}} f_{n_{z,m_k}^{nv}}$, and G denotes the maximum number of BSs a virtualized UE can be associated to. In (33b), to make the problem computationally tractable, we have changed the maximum sum-rate problem to the problem of maximizing the achievable SINR at virtualized UEs. Moreover, applying a simple arrangement of terms, we obtain (35) where $R_{n_l^v} \leq \sum_{j=1}^{L}\sum_{m_j=1}^{M_j} R_{n_{l,m_j}^{v}}$. To linearize the term $R_{n_{l,m_j}^{v}} f_{n_{z,m_k}^{v}}$, we use the linearization technique I introduced previously. Finally, to linearize (33d), we use the exact same procedure used to derive (32).

$$\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}\left(p\eta_{m_k}\beta_{m_k,n_{l,m_k}^{nv}}\left(\sum_{n_{z,m_k}^{nv}\neq n_{l,m_j}^{nv}} R_{n_{l,m_j}^{nv}} f_{n_{z,m_k}^{nv}} + R_{n_{l,m_j}^{nv}}\alpha_{m_k}^{v} + R_{n_{l,m_j}^{nv}}\Omega_{m_k}\right)\right) + 1 \leq p\eta_{m_j} f_{n_{l,m_j}^{nv}}\beta_{m_j,n_{l,m_j}^{nv}}, \tag{29}$$

$$\sum_{k=1}^{L}\sum_{m_k=1}^{M_k}\left(p\eta_{m_k}\beta_{m_k,n_{l,m_k}^{v}}\left(\sum_{n_{z,m_k}^{v}\neq n_{l,m_j}^{v}} R_{n_{l,m_j}^{v}} f_{n_{z,m_k}^{v}} + R_{n_{l,m_j}^{v}}\tilde{\alpha}_{m_k}^{nv}\right)\right) + 1 \leq p\eta_{m_j} f_{n_{l,m_j}^{v}}\beta_{m_j,n_{l,m_j}^{v}}, \tag{35}$$

IV. Performance Evaluation

In this section, we evaluate the performance of VARM using comprehensive simulation results. We consider an SD-RAN with $L = 4$. We consider a network topology formed by a $3000 \text{ m} \times 2000 \text{ m}$ region with BSs whose locations are fixed throughout the simulations. We assume one BS is located in the center of each cell and 20 other BSs are distributed uniformly in the region. We further assume that the number of antennas at BSs is randomly selected such that $100 \leq A_{m_j} \leq 150, m_j \in \mathcal{M}_j, j \in \mathcal{L}$. We assume that BSs can serve user sets of size 20 and the pathloss from a BS to a UE is given by $\beta_{m_l,n_{l'}} = \frac{1}{1+\left(\frac{d_{m_l,n_{l'}}}{40}\right)^4}$ [11], with $d_{m_l,n_{l'}}$ representing the distance between BS m_l and UE $n_{l'}$. We generate the location of the UEs according to a non-homogeneous Poisson point process with lower density in the central region of cells. In Fig. 3, we compare the cumulative distribution function (CDF) of the sum-rate by virtualized and non-virtualized services. We assume $N = 43$, $N_l^v = 3$, and $N_l^{nv} = 40$, $p = 50$ dBm, $B_{m_j} = 20$, $C_{m_j} = 0.5$ Mbit/s/Hz, $P_{m_j} = 30$ dB, $m_j \in \mathcal{M}_j, j \in \mathcal{L}$, and $\gamma_{n_l^v}^v \in \{0.1, 0.2, 0.3\}$ Mbit/s/Hz. For virtualized UEs, we assume that $G = 10$. Non-virtualized UEs, however, can only be associated with one BS at a time. In order to investigate the performance of the system with pilot contamination, we assume that the length of the pilot signal is $\alpha = 10$ and $p_\alpha = \frac{10 \text{ dB}}{N}$. In Fig. 3, we have used "V" to refer to virtualized services, "NV" to refer to non-virtualized services, "PC" to refer to pilot contamination, and "PCSI" to refer to the transmission scenario with perfect CSI. It can be seen that with perfect CSI, virtualized UEs receive higher QoS levels, as per the SLAs. As an example, when $\gamma_{n_l^v}^v = 0.1$ Mbit/s/Hz, non-virtualized UEs can receive service rates as high as 0.3 Mbit/s/Hz, whereas virtualized UEs are promised minimum rates of 0.446 Mbit/s/Hz. With pilot contamination, the overall performance of the SD-RAN is degraded. While the difference between the achievable aggregate rate by virtualized and non-virtualized UEs is reduced, the virtualized UEs still are guaranteed their minimum rate requirements.

In Fig. 4, we study the performance of the SD-RAN system when using the maximum peak rate association strategy [11] where the UE n_l, $l \in \mathcal{L}$ associates with BS m_j when

$$m_j = \arg\max_{m_k} R_{n_{l,m_k}}, m_k \in \mathcal{M}_k, k \in \mathcal{L}. \tag{36}$$

In the massive MIMO regime, the achievable peak rates of the UEs converge to the deterministic limits that depend only on the SINR of the UEs. Therefore, in the massive MIMO framework, it is easy to implement the peak rate association scheme. We assume $\gamma_{n_l^v}^v = 0.3$ Mbit/s/Hz. In Fig. 4, we have used "MPR" to refer to maximum peak rate association strategy. As can be seen from Fig. 4, with both perfect CSI and pilot contamination, VARM can achieve significantly higher sum-rate than the maximum peak rate strategy. As an example, with pilot contamination, virtualized UEs and non-virtualized UEs can achieve minimum rates of 0.48 Mbit/s/Hz and 0.45 Mbit/s/Hz, respectively. However, UEs can only achieve up to 0.23 Mbit/s/Hz with maximum peak rate strategy. With perfect CSI, virtualized UEs and non-virtualized UEs can achieve minimum rates of 0.59 Mbit/s/Hz and 0.45 Mbit/s/Hz, respectively. However, the SD-RAN can only provide UEs upto 0.26 Mbit/s/Hz with maximum peak rate strategy. These results corroborate the outperformance of the proposed VARM technique over the conventional maximum peak rate strategy.

Fig. 5 illustrates the sum-rate of the SD-RAN system using VARM versus maximum number of UEs that BSs can serve for $C_{m_j} = 15$ Mbit/s/Hz. We assume $\gamma_{n_l^v}^v = 0.1$ Mbit/s/Hz for virtualized UEs. We compare the performance of VARM versus the maximum peak rate (MPR) strategy with perfect CSI and pilot contamination. As can be seen from Fig. 5, compared to the maximum peak rate strategy, VARM with both perfect CSI and pilot contamination, provides the SD-RAN with significantly higher sum-rate values. Moreover, we can see a diminishing return pattern in the performance of VARM with respect to the number of UEs, the SD-RAN BSs can support. The knees of the utility curves signify the optimum number of UEs that the SD-RAN BSs should support for the considered N values. We denote this point by $B_{m_j}^*$. When $B_{m_j} > B_{m_j}^*$, further increase in B_{m_j} does not lead to any increase in the sum-rate of the system.

V. Conclusion

In this paper, we have proposed VARM which optimally allocates the limited radio and fronthaul resources of the SD-RAN to the virtualized and non-virtualized UEs. Using the proposed HVC technique, we have formulated a two-stage optimization problem for virtualized and non-virtualized UEs. In order to determine the optimum allocation/association strategies of the UEs as well as the optimum compression noise levels, we have used a Stackelberg game model. The provided solution guarantees the virtualized UEs with their preferred minimum rate requirements and provides the non-virtualized UEs with best effort services. Numerical results show that the

proposed VARM considerably outperforms the maximum peak rate association strategy for SD-RAN. Moreover, the sum-rate of the proposed VARM follows a diminishing return pattern with respect to the maximum number of UEs each BS is associated with. For future work, we will investigate pricing for the massive MIMO SD-RAN, assuming that BSs are owned by different service providers which offer their services at different usage-based prices.

APPENDIX

A. An Asymptotic Approximation for (5)

Since we have assumed MF precoding, the left hand side of (5) reduces to the following:

$$\begin{aligned} & N\log\left(1+\frac{1}{\Omega_{m_j}}-\frac{1}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)\right) \\ & +A_{m_j}\log\left(1+\frac{1}{\Omega_{m_j}}\frac{N}{A_{m_j}}-\frac{1}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)\right) \\ & -A_{m_j}\Omega_{m_j}\frac{\log e}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right), \end{aligned} \tag{37}$$

where $\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)$ is provided in (38), all logarithms are for base 10, and (37) and (38) follow from the Shannon transform of the empirical distribution of the eigenvalues of $\mathbf{E}_{m_j}\mathbf{D}\mathbf{D}^H\mathbf{E}_{m_j}^T$ using the Marcenko-Pastur law [21], assuming that $N, A_{m_j}\to\infty$ with $\frac{N}{A_{m_j}}\to\nu_{m_j}$.

In the context of massive MIMO, when $A_{m_j}\to\infty$, after some calculations ((43) in Appendix B), we obtain (39) and (5) can be approximated as follows:

$$\begin{aligned} \frac{C_{m_j}}{A_{m_j}} \geq\ & \nu_{m_j}\log\left(1+\frac{1}{\Omega_{m_j}}-\frac{1}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\nu_{m_j}\right)\right) \\ & +\log\left(1+\frac{\nu_{m_j}}{\Omega_{m_j}}-\frac{1}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\nu_{m_j}\right)\right) \\ & -\frac{\Omega_{m_j}\log e}{4}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\nu_{m_j}\right). \end{aligned} \tag{40}$$

Considering the massive MIMO regime where $N\ll A_{m_l}$, (40) can be further approximated as follows:

$$\frac{C_{m_j}}{A_{m_j}}\geq\log\left(1-\frac{1}{4}\left(1+\frac{1}{\Omega_{m_j}}\right)\right)-\frac{\Omega_{m_j}\log e}{4}\left(\frac{1}{\Omega_{m_j}}+1\right), \tag{41}$$

where we have assumed $\nu_{m_j}\approx 0$ and $\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\nu_{m_j}\right)\approx\frac{1}{\Omega_{m_j}}+1$. Using the trapezoidal rule [20] $\ln(1+x)\approx x\left(\frac{1+0.5x}{1+x}\right)$, we can solve (41) and obtain

$$\Omega_{m_j}\geq\frac{\log e}{-\frac{C_{m_j}}{A_{m_j}}+\sqrt{\left(\frac{C_{m_j}}{A_{m_j}}\right)^2+\frac{2\log eC_{m_j}}{A_{m_j}}}}. \tag{42}$$

B. Proof of Equation (39)

We have (43). Therefore, when $A_{m_j}\to\infty$, dividing (43) by A_{m_j} results in the asymptotic form provided in (44).

C. Proof of (12)

$\mathbb{E}\left[\left|\mathbf{H}_{n_l}^H\mathbf{d}_k\right|^2\right]$ can be expressed as follows:

$$\begin{aligned} & \mathbb{E}\left[\left|\mathbf{H}_{n_l}^H\mathbf{d}_k\right|^2\right] \\ & =\mathbb{E}\left[\left[\mathbf{h}_{1,n_l}^H\dots\mathbf{h}_{M_L,n_l}^H\right]\left(\frac{\mathbf{h}_{1,n_k}}{\left\|\mathbf{h}_{1,n_k}\right\|},\dots,\frac{\mathbf{h}_{M_L,n_k}}{\left\|\mathbf{h}_{M_L,n_k}\right\|}\right)\times\right. \\ & \left.\left[\frac{\mathbf{h}_{1,n_k}^H}{\left\|\mathbf{h}_{1,n_k}\right\|}\dots\frac{\mathbf{h}_{M_L,n_k}^H}{\left\|\mathbf{h}_{M_L,n_k}\right\|}\right]\left(\mathbf{h}_{1,n_l},\dots,\mathbf{h}_{M_L,n_l}\right)\right] \\ & =\mathbb{E}\left[\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\mathbf{h}_{m_j,n_l}^H\frac{\mathbf{h}_{m_j,n_k}\mathbf{h}_{m_j,n_k}^H}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\mathbf{h}_{m_j,n_l}\right]+ \\ & \underbrace{\mathbb{E}\left[\sum_{i=1}^{L}\sum_{m_i=1}^{M_i}\sum_{j=1}^{L}\sum_{m_j=1,m_j\neq m_i}^{M_j}\mathbf{h}_{m_i,n_l}^H\frac{\mathbf{h}_{m_i,n_k}\mathbf{h}_{m_j,n_k}^H}{\left\|\mathbf{h}_{m_i,n_k}\right\|^2}\mathbf{h}_{m_j,n_l}\right]}_{0}= \\ & \mathbb{E}\left[\left[\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sum_{i=1}^{A_{m_j}}\mathbf{h}_{m_j,n_l}^*[i]\frac{\mathbf{h}_{m_j,n_k}[i]\,\mathbf{h}_{m_j,n_k}^*[1]}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\mathbf{h}_{m_j,n_l}\quad\dots\right.\right. \\ & \left.\left.\dots\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sum_{i=1}^{A_{m_j}}\mathbf{h}_{m_j,n_l}^*[i]\frac{\mathbf{h}_{m_j,n_k}[i]\,\mathbf{h}_{m_j,n_k}^*\left[A_{m_j}\right]}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\mathbf{h}_{m_j,n_l}\right]\right] \\ & =\mathbb{E}\left[\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sum_{i=1}^{A_{m_j}}\mathbf{h}_{m_j,n_l}^*[i]\frac{\mathbf{h}_{m_j,n_k}[i]\,\mathbf{h}_{m_j,n_k}^*[i]}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\mathbf{h}_{m_j,n_l}[i]\right] \\ & =\mathbb{E}\left[\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sum_{i=1}^{A_{m_j}}\frac{\left|\mathbf{h}_{m_j,n_l}[i]\right|^2\left|\mathbf{h}_{m_j,n_k}[i]\right|^2}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\right]. \end{aligned} \tag{45}$$

In order to characterize the upper bound for (45), note that for positive numbers $x_1,\dots,x_N$, we have

$$\frac{\left(x_1+\dots+x_N\right)^2}{N}\leq x_1^2+\dots+x_N^2\leq N\max\left(x_1^2,\dots,x_N^2\right), \tag{46}$$

which compares the root mean square value, the arithmetic mean, and the maximum value of these positive numbers. We apply (46) to (45), and obtain

$$\begin{aligned} & \sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{\left(\sum_{i=1}^{A_{m_j}}\left|\mathbf{h}_{m_j,n_l}[i]\right|\left|\mathbf{h}_{m_j,n_k}[i]\right|\right)^2}{A_{m_j}\left\|\mathbf{h}_{m_j,n_k}\right\|^2} \\ & \leq\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\sum_{i=1}^{A_{m_j}}\frac{\left|\mathbf{h}_{m_j,n_l}[i]\right|^2\left|\mathbf{h}_{m_j,n_k}[i]\right|^2}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2} \\ & \leq\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{A_{m_j}\max\limits_{1\leq i\leq A_{m_j}}\left(\left|\mathbf{h}_{m_j,n_l}[i]\right|^2\left|\mathbf{h}_{m_j,n_k}[i]\right|^2\right)}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2} \end{aligned}$$

$$\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)= \left(\sqrt{\frac{1}{\Omega_{m_j}}\left(1+\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}-\sqrt{\frac{1}{\Omega_{m_j}}\left(1-\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}\right)^2, \tag{38}$$

$$\mathcal{F}\left(\tfrac{1}{\Omega_{m_j}},\tfrac{N}{A_{m_j}}\right)\approx \frac{4\big(\big(\Omega_{m_j}+1\big)+\nu_{m_j}\big)^2}{\Omega_{m_j}\Big(\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}+2\sqrt{\nu_{m_j}}}+\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}-2\sqrt{\nu_{m_j}}}\Big)^2}, \tag{39}$$

$$\begin{aligned} &A_{m_j}\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)= \\ &A_{m_j}\left(\sqrt{\frac{1}{\Omega_{m_j}}\left(1+\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}-\sqrt{\frac{1}{\Omega_{m_j}}\left(1-\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}\right)^2 \\ &=\frac{A_{m_j}}{\left(\sqrt{\frac{1}{\Omega_{m_j}}\left(1+\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}-\sqrt{\frac{1}{\Omega_{m_j}}\left(1-\sqrt{\frac{N}{A_{m_j}}}\right)^2+1}\right)^2} \\ &=\frac{4\big(A_{m_j}\left(\Omega_{m_j}+1\right)+N\big)^2}{\Omega_{m_j}\Big(\sqrt{\left(\Omega_{m_j}+1\right)A_{m_j}+N+2\sqrt{A_{m_j}N}}+\sqrt{\left(\Omega_{m_j}+1\right)A_{m_j}+N-2\sqrt{A_{m_j}N}}\Big)^2} \\ &=\frac{4A_{m_j}\big(\big(\Omega_{m_j}+1\big)+\nu_{m_j}\big)^2}{\Omega_{m_j}\Big(\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}+2\sqrt{\nu_{m_j}}}+\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}-2\sqrt{\nu_{m_j}}}\Big)^2}. \end{aligned} \tag{43}$$

$$\mathcal{F}\left(\frac{1}{\Omega_{m_j}},\frac{N}{A_{m_j}}\right)\overset{a.s.}{\to}\frac{4\big(\big(\Omega_{m_j}+1\big)+\nu_{m_j}\big)^2}{\Omega_{m_j}\Big(\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}+2\sqrt{\nu_{m_j}}}+\sqrt{\big(\Omega_{m_j}+1\big)+\nu_{m_j}-2\sqrt{\nu_{m_j}}}\Big)^2}. \tag{44}$$

$$\leq\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{A_{m_j}\left|h^{\max}_{m_j,n_l}\right|^2\left|h^{\max}_{m_j,n_k}\right|^2}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2},$$

where $\left|h^{\max}_{m_j,n_k}\right|=\max\limits_{1\leq i\leq A_{m_j}}\left(\left|\mathbf{h}_{m_j,n_k}[i]\right|\right)$. Considering the upper limit in (C), we have

$$\begin{aligned} &\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}\frac{pA_{m_j}\beta_{m_j,n_l}\beta_{m_j,n_k}\left|\tilde{h}^{\max}_{m_j,n_l}\right|^2\left|\tilde{h}^{\max}_{m_j,n_k}\right|^2}{\left\|\mathbf{h}_{m_j,n_k}\right\|^2}\leq \\ &\sum_{j=1}^{L}\sum_{m_j=1}^{M_j}pA_{m_j}\beta_{m_j,n_l}\beta_{m_j,n_k}\left|\tilde{h}^{\max}_{m_j,n_l}\right|^2, \end{aligned} \tag{47}$$

where $\left|\tilde{h}^{\max}_{m_j,n_k}\right|=\max\limits_{1\leq i\leq A_{m_j}}\left(\left|\tilde{\mathbf{h}}_{m_j,n_k}[i]\right|\right)$and we have used $\left|h^{\max}_{m_j,n_k}\right|^2\leq\left\|\mathbf{h}_{m_j,n_k}\right\|^2$.

REFERENCES

[1] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, R. Apelfrjd, M. Sternad, and T. Svensson, "The role of small cells, coordinated multi-point and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 45–51, May 2014.

[2] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multi-user MIMO systems," *IEEE Trans. on Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[3] A. G. Gotsis, S. Stefanatos, and A. Alexiou, "Optimal user association for massive MIMO empowered ultra-dense wireless networks," in *Proc. of IEEE International Conference on Communications (ICC)*, London, UK, June 2015.

[4] Y. Xu and S. Mao, "User association in massive MIMO HetNets," *IEEE Systems Journal*, To Appear, 2015.

[5] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 44–53, Jul. 2013.

[6] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[7] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3216–3225, Sep. 2012.

[8] S. Nanba and A. Agata, "A new IQ data compression scheme for fronthaul link in centralized RAN," in *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) Workshop on Cooperative and Heterogeneous Cellular Networks*, London, UK, Sep. 2013.

[9] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger,

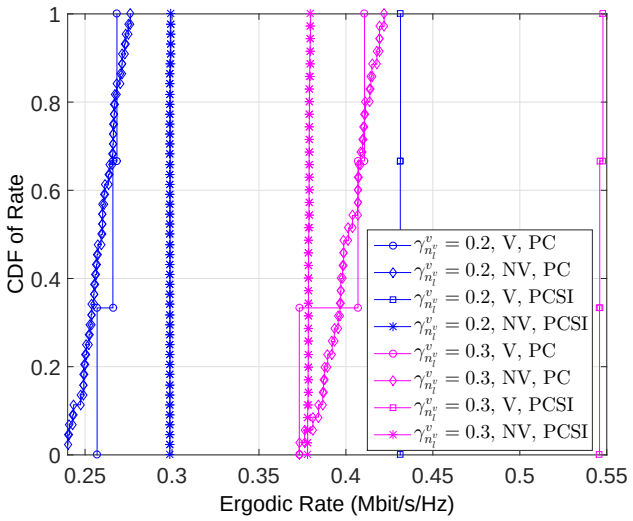

Fig. 3. CDF of aggregate ergodic rate for virtualized and non-virtualized services with perfect CSI and pilot contamination.

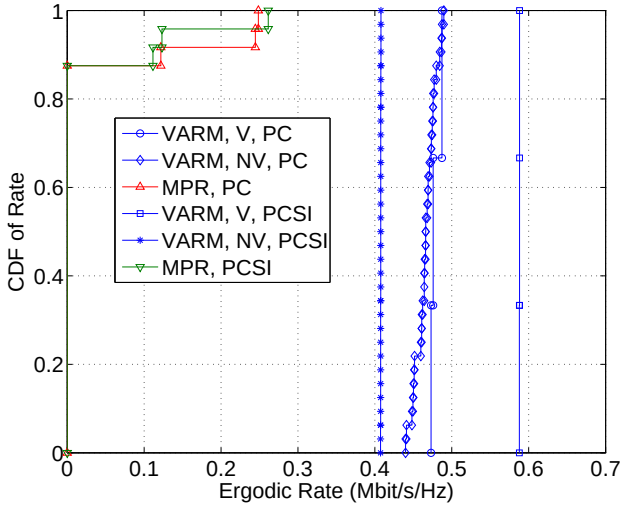

Fig. 4. CDF of ergodic rate for VARM and MPR strategy [11] with perfect CSI and pilot contamination.

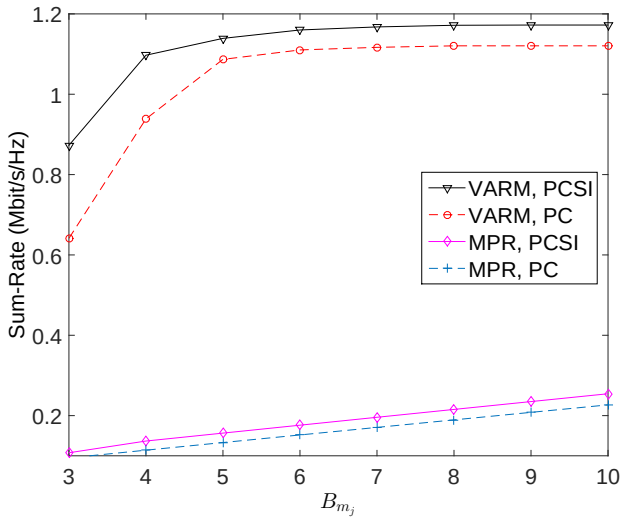

Fig. 5. System sum-rate versus B_{m_j} for VARM and MPR strategy [11].

and L. Dittmann, "Cloud RAN for mobile networks - A technology overview," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, First Quarter, 2015.

[10] K. Chen and R. Duan, "C-RAN: The road towards green RAN," China Mobile Research Institute. White Paper, Tech. Rep., Oct. 2011.

[11] D. Bethanabhotla, O. Bursalioglu, H. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *arXiv:1407.6731*, 2014.

[12] 3GPP TR 36.872 V12.1.0, "Small cell enhancements for E-UTRA and E-UTRAN Physical layer aspects (Release 12)," Tech. Rep., Dec. 2013.

[13] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[14] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[15] ——, "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," in *Proc. of Inf. Sci. Syst. Conf.*, Princeton, NJ, Mar. 2014.

[16] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Communications*, vol. 22, no. 2, pp. 152–160, April 2015.

[17] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channel," *IEEE Tran. on Vehicular Technology*, To Appear, 2015.

[18] A. El. Gamal and Y. H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[19] D. S. Chen, R. Batson, and Y. Dang, *Applied Integer Programming, Modeling and Solution*. John Wiley & Sons, 2010.

[20] S. K. Khattri, "New close form approximations of ln(1 + x)," *The Teaching of Mathematics*, vol. XII, no. 1, pp. 7–14, 2009.

[21] A. M. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communications*. Now Publishers, 2004.