IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 21, 2021, accepted August 25, 2021, date of publication September 16, 2021, date of current version September 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113363

Nonlinear Forwarding Strategy for Firefly Ultra Dense Networks With mmWave Fronthaul Links

KATHARINA ACKERMANN[1], (Member, IEEE), VAHID JAMALI[1], (Member, IEEE), WOLFGANG GERSTACKER[1], (Senior Member, IEEE), JOCELYN AULIN[2], (Member, IEEE), AND ROBERT SCHOBER[1], (Fellow, IEEE)

[1]Friedrich-Alexander University (FAU) Erlangen-Nürnberg, 91054 Erlangen, Germany
[2]Huawei Technologies Sweden AB, 164 40 Kista, Sweden

Corresponding author: Katharina Ackermann (katharina.ackermann@fau.de)

This work was supported by Huawei Technologies Sweden AB.

ABSTRACT We consider the uplink of firefly ultra dense networks which combine the promising features of ultra dense deployment and centralized processing. In these networks, a large number of remote radio units which we denote as firefly nodes (FNs) are spatially distributed over an area. The mobile devices (MDs) in the coverage area are simultaneously connected via sub-6 GHz radio frequency links to all FNs. Unlike the cloud radio access network (C-RAN) architecture, in firefly ultra dense networks, the FNs forward the MDs' data through multi-hop millimeter-wave (mmWave) links to one or multiple root nodes since the coverage radius of each mmWave link is limited. These root nodes then forward the data via optical fiber links further to a central unit (CU), where the MDs' signals are decoded. The amount of data that is received at each FN is potentially huge, and hence, efficient signal processing is needed at each FN before the received signals can be forwarded to other FNs. Therefore, we propose a nonlinear processing strategy, which quantizes the received signals at each FN. In particular, we formulate an optimization problem for a local design strategy for the nonlinear forwarding at the FNs, and present an optimal solution by exploiting strong duality and using the Lagrangian method to convert the optimization problem into an unconstrained problem via its dual formulation. A closed-form solution for the primal variables and a bisection algorithm for finding the optimal dual variables are presented. Moreover, based on the cut-set bound, we develop an upper bound on the achievable sum rate of the considered firefly network. The proposed nonlinear forwarding strategy is shown to outperform a benchmark linear forwarding strategy and to approach the performance upper bound in relevant transmit power regimes at the expense of a higher computational complexity. Our results reveal that having more root nodes in the topology improves the performance of linear and nonlinear forwarding but requires additional optical fiber links to the CU.

INDEX TERMS Cloud radio access network (C-RAN), millimeter wave (mmWave), multi-hop system, nonlinear forwarding, ultra dense network (UDN).

I. INTRODUCTION

It is expected that the number of devices that will use the fifth generation (5G) of wireless communication technology will reach tens or even hundreds of billions world wide, and the total data volume demand is predicted to increase to over 175 Zettabytes until 2025 [1]. To support this tremendous demand for wireless data, there are two promising key strategies, namely ultra dense deployment and centralized processing [2]–[6]. Here, ultra dense deployment refers to the use of low power, dense small cell networks, where the distance between the base stations (BSs) and the mobile devices (MDs) is reduced significantly compared to the current networks which leads to higher achievable data rates. However, as the network becomes denser, the multiple access interference caused by the MDs in the uplink increases which results in more complex interference scenarios [2], [3]. Here, we are considering the uplink and to overcome the problem of multiple access interference, centralized processing can be

The associate editor coordinating the review of this manuscript and approving it for publication was Haipeng Yao.

The downlink of similar networks is discussed in [7]–[9], where the main design goals are to ensure fairness among the MDs, to increase the system throughput, to improve the network coverage probability, to cope with the huge multi-hop backhaul traffic, and to optimize the downlink resource allocation.

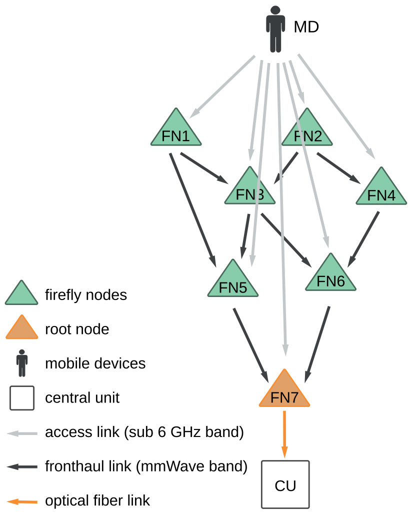

FIGURE 1. **Illustration of the considered firefly network architecture, where FN 7 is the root node. Only a single MD is shown for clarity of illustration.**

used, where all received signals in the network are processed at a central unit (CU), and thus, interference can be efficiently mitigated. Cloud radio access networks (C-RANs) are combining these two features, where several remote radio units are typically directly connected to the CU [4]–[6]. However, to cover a certain area and to reach high data rates, C-RAN architectures mostly rely on expensive optical fiber links to connect each remote radio unit to the CU [6], [10]. To overcome this cost issue, we consider a network of spatially distributed remote radio units, which we refer to as firefly nodes (FNs). The MDs in the coverage area are simultaneously connected to several FNs via sub-6 GHz links and the FNs are in turn connected via multi-hop millimeter-wave (mmWave) links (fronthaul links). A subset of the FNs serve as root nodes and are connected to other FNs via optical fiber links to the CU where the joint decoding of the MDs' messages is performed. We refer to such a network as firefly ultra dense network. The two different frequency bands ensure that the MDs' access channels do not interfere with the multi-hop channels between the FNs. However, unlike in C-RAN architectures, in the considered firefly ultra dense network, the FNs have to forward the received signals to the root nodes via other intermediate FNs over capacity-constrained mmWave links. Thereby, each FN may be connected to several other FNs, see Figure 1. The intermediate connections between the FNs are necessary since the coverage range of mmWave links is typically limited [11]–[14]. Since the positions of the FNs are fixed, beamforming can be used to achieve high-rate inter-FN links and to avoid inter-link interference [11]–[14]. Thus, this multi-hop mmWave topology allows the coverage of a certain area and can support high MD rates. Expensive optical fiber links are employed only to connect the FNs serving as root nodes to the CU, which reduces the overall cost. Having multiple root nodes reduces the required number of mmWave link hops, and hence, depending on the topology, may lead to a lower end-to-end delay for the MDs.

One of the main challenges of firefly networks is that the amount of data that is received at each FN is potentially huge, and hence, efficient signal processing is needed at each FN before the signals can be forwarded over the mmWave links to the next FNs such that the transmit data rate does not exceed the mmWave link capacity. In [15], we investigated linear processing techniques for FNs employing linear filter matrices. We proposed several locally-designed linear forwarding schemes and evaluated their performance. In this paper, we will investigate nonlinear forwarding schemes since we expect a better performance at the expense of an increased computational complexity. In [16], a similar network architecture was studied and a data compression strategy for the intermediate nodes was designed to reduce the amount of data to be processed in the network. More precisely, the authors of [16] investigated a decompress-process-and-recompress (DPR) scheme that applies linear processing of the decompressed signals before further nonlinear forwarding. In this paper, we investigate a compress-and-forward (CF) approach, whereby compared to DPR, the decompressed signals are multiplexed, instead of being linearly processed, before recompression and forwarding, and hence, the computational complexity is reduced. In addition, the authors of [16] assumed a fixed link capacity between the intermediate nodes of the network, whereas the mmWave links in the considered firefly ultra dense network are susceptible to fading.

In this paper, we investigate nonlinear processing techniques based on vector quantization at the FNs, where each FN exploits the correlation between the MD signals received at different antennas and jointly quantizes the corresponding inphase and quadrature (IQ) data streams [17], [18]. The main contributions of this paper can be summarized as follows:

- We derive an upper bound, based on the cut-set bound [19], on the achievable sum rate of the considered firefly network. This upper bound is valid for any linear and nonlinear strategy and provides significant insights regarding the performance bottlenecks of the considered communication system.
- We consider central and local design strategies for the nonlinear forwarding at the FNs and formulate corresponding optimization problems. Due to the high complexity of the optimization problem for the central design strategy, we focus on solving the optimization problem for the local design strategy. In particular, exploiting strong duality, we present an optimal solution using the Lagrangian method to convert the optimization problem to an unconstrained problem via its dual formulation. More precisely, a closed-form solution for the primal variables and a bisection algorithm for finding the optimal dual variables are presented [20].
- We show via simulations that the performance of the resulting locally-designed nonlinear forwarding scheme

outperforms the linear forwarding schemes in [15] and approaches the derived upper bound for the achievable rate for relevant transmit power regimes.
- We compare the overall delay of different firefly network topologies and investigate the impact of the number of expensive root nodes on the delay. Furthermore, we compare the complexity of the linear forwarding schemes reported in [15] with that of the proposed nonlinear forwarding schemes. This allows us to evaluate the trade-off that the different forwarding strategies and firefly network topologies offer in terms of achievable rate, delay, and cost.

The remainder of this paper is organized as follows. In Section II, we describe the system model and the channel models for the radio access and fronthaul links. In Section III, a performance upper bound for the achievable sum rate of the firefly ultra dense network is derived. In Section IV, we introduce the general problem formulation for the design of the proposed nonlinear forwarding schemes based on vector quantization. Then, the optimization problems for centrally-designed and locally-designed nonlinear forwarding are formulated and an efficient solution for the latter problem is presented. Furthermore, in Section V, the performance of the proposed nonlinear forwarding scheme is evaluated for several firefly network topologies. Finally, the main results are summarized in Section VI.

Notations: We use the following notations throughout this paper: Bold upper-case and bold lower-case letters are used for matrices and vectors, respectively. $\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{T}}$ are the Hermitian transpose and transpose of matrix $\mathbf{A}$, respectively, whereas a^* denotes the complex conjugate of a. Furthermore, $\mathbf{I}_n$ represents an $n \times n$ identity matrix, and $\mathbf{0}_{n\times m}$ is an $n \times m$ all-zero matrix. $\mathbb{E}\{\cdot\}$ stands for the expectation operator, $|\cdot|$ represents the determinant of a matrix or the cardinality of a set, and $\mathbb{C}$ is the set of complex numbers. Moreover, $\phi = \{\}$ denotes the empty set. Additionally, let $\mathcal{I}$ be an arbitrary set of integers, then Blockdiag$(\mathbf{A}_n, n \in \mathcal{I})$ is a block-diagonal matrix with $\mathbf{A}_n$, $\forall n \in \mathcal{I}$, on the main diagonal, diag$(a_n, n \in \mathcal{I})$ is a diagonal matrix with a_n, $\forall n \in \mathcal{I}$, on the main diagonal, and $\Sigma_x := \mathbb{E}\left\{\mathbf{x}\mathbf{x}^{\mathrm{H}}\right\}$ denotes the covariance matrix of zero-mean random vector $\mathbf{x}$. Furthermore, $\mathcal{CN}(\mathbf{a}, \Phi)$ represents a complex Gaussian distributed random variable (RV) with mean vector $\mathbf{a} \in \mathbb{C}^{m\times 1}$ and covariance matrix $\Phi \in \mathbb{C}^{m\times m}$. In addition, $I(X; Y)$ represents the mutual information between RVs X and Y, $f(x)$ is the probability density function of RV X, and $f(x, y)$ denotes the joint probability density function of RVs X and Y. Vercut$(\mathbf{A}_n, n \in \mathcal{I}) := [\mathbf{A}_{n_1}^{\mathrm{T}}, \ldots, \mathbf{A}_{n_{|\mathcal{I}|}}^{\mathrm{T}}]^{\mathrm{T}}$ with a set $\mathcal{I} = \left\{n_1, \ldots, n_{|\mathcal{I}|}\right\}$ and matrices $\mathbf{A}_n, n \in \mathcal{I}$. Moreover, $\lfloor\cdot\rfloor$ denotes the floor function.

II. SYSTEM AND CHANNEL MODELS

We consider an uplink communication system where K single-antenna MDs send their data to a CU via M FNs, see Figure 1. We assume that no direct connection between the MDs and the CU is available. The MDs are connected to the FNs via sub-6 GHz radio frequency (RF) links and the FNs communicate with each other through mmWave links. Furthermore, we assume that a subset of the FNs, referred to as root nodes, is connected to the CU via optical fiber links. Each FN is equipped with N RF receive antennas, N_t mmWave transmit antennas, and N_r mmWave receive antennas. $\mathcal{M} = \{1, 2, \ldots, M\}$ denotes the index set of all FNs in the considered communication system. Moreover, $\mathcal{E} = \left\{(n, m)|a_{n,m} = 1\right\}$ is the set of all available edges, where $a_{n,m} = 1$ specifies that the link from FN n to FN m is available, whereas $a_{n,m} = 0$ means that FN n and FN m are not connected. Note that the notation (n, m) describes that FN n transmit its signals to FN m via mmWave link. Furthermore, $\mathcal{N} \subset \mathcal{M}$ is the index set of root nodes. The indices of root nodes are denoted as $\nu_1, \ldots, \nu_T$, i.e., $\mathcal{N} = \{\nu_1, \ldots, \nu_T\}$ with $|\mathcal{N}| = T$.

We consider a wideband multicarrier communication system with $N_f = \frac{W^{\mathrm{RF}}}{W^{\mathrm{RF}}_{\mathrm{sub}}}$ orthogonal subcarriers for the RF link and $N_\rho = \frac{W^{\mathrm{mmW}}}{W^{\mathrm{mmW}}_{\mathrm{sub}}}$ orthogonal subcarriers for the mmWave link, respectively, i.e., orthogonal frequency-division multiplexing (OFDM) is applied in the sub-6 GHz band as well as in the mmWave band. Here, W^{RF} is the total available RF bandwidth and W^{mmW} is the total available mmWave bandwidth. Moreover, $W^{\mathrm{RF}}_{\mathrm{sub}}$ and $W^{\mathrm{mmW}}_{\mathrm{sub}}$ are the bandwidths of each RF and mmWave subcarrier, respectively. In addition, due to the N RF receive antennas, N RF signals have to be processed at each FN. For the mmWave links between the FNs, we assume single-input single-output (SISO) communication, where a single stream is transmitted and the multiple antennas at the mmWave transceivers are used for beamforming and enhancing the overall link budget (see the detailed description in the following). Since generally, in ultra dense networks, we have $N_\rho > N_f$ to forward the RF symbols over a mmWave link from one FN to another FN, the RF symbols can be multiplexed across the frequency and time domain of the mmWave channel. In particular, there are $N_L = \left\lfloor \frac{W^{\mathrm{mmW}}}{W^{\mathrm{RF}}} \right\rfloor = \left\lfloor \frac{N_\rho}{N_f} \cdot \frac{T^{\mathrm{RF}}_{\mathrm{sub}}}{T^{\mathrm{mmW}}_{\mathrm{sub}}} \right\rfloor$ mmWave symbols per N-dimensional RF vector symbol available for frequency and time multiplexing. Here, $T^{\mathrm{RF}}_{\mathrm{sub}} = \frac{1}{W^{\mathrm{RF}}_{\mathrm{sub}}}$ and $T^{\mathrm{mmW}}_{\mathrm{sub}} = \frac{1}{W^{\mathrm{mmW}}_{\mathrm{sub}}}$ are the durations of an RF OFDM symbol and a mmWave OFDM symbol, respectively. Furthermore, we enumerate the RF subcarriers as $\mathcal{N}_f = \left\{1, \ldots, N_f\right\}$. For the subsequent investigations, we assume that each MD k is only active in a certain subset of RF subcarriers, $\mathcal{F}_k \subseteq \mathcal{N}_f$. Hence, MD k is active on $F_k = |\mathcal{F}_k|$ subcarriers. For subcarrier allocation, $c_{k,s}$ indicates whether MD k is active on subcarrier s or not, i.e., $c_{k,s} = 1$ means that MD k is active on subcarrier s and $c_{k,s} = 0$ implies that MD k is not active on subcarrier s. The set of MDs which are active on subcarrier s is given by $\mathcal{K}_s = \left\{k \in \mathcal{K}|c_{k,s} = 1\right\}$, where $\mathcal{K}$ is the set of all MDs in the communication system. The number of MDs active on subcarrier s is given by $K_s = |\mathcal{K}_s|$. Note that the layer structure of our network will be discussed in detail in Section V-A1.

In the following, we introduce the channel models for the radio access and mmWave links, respectively, where we assume flat fading for each subcarrier. For simplicity of presentation, the subsequent equations do not include the subcarrier index. Moreover, we do not consider the problem of subcarrier allocation. Furthermore, we assume uniform power allocation across different mmWave subcarriers and RF subcarriers, respectively.

1) RADIO ACCESS LINKS

The received RF symbol vector on subcarrier s at FN m is denoted by $\mathbf{y}_{\mathrm{FN}m}^{\mathrm{RF}} \in \mathbb{C}^{N\times 1}$, and can be modeled as

$$\mathbf{y}_{\mathrm{FN}m}^{\mathrm{RF}} = \mathbf{H}_{\mathrm{FN}m}\mathbf{x} + \mathbf{z}_{\mathrm{FN}m}, \tag{1}$$

where vector $\mathbf{x} \in \mathbb{C}^{K_s\times 1}$ contains the transmit symbols x_k of all MDs which are active on subcarrier s. The transmit symbols of the k-th MD, $x_k \in \mathbb{C}$, are independent and identically distributed (i.i.d.), zero-mean, complex Gaussian RVs. The average power of transmit symbol x_k is constrained by the maximum RF transmit power of the k-th MD, P_k^{RF}, divided by the number of used RF subcarriers. In addition, $\mathbf{z}_{\mathrm{FN}m} \in \mathbb{C}^{N\times 1}$ is the additive white Gaussian noise (AWGN) vector at FN m, which is assumed to be independent from the transmitted symbol vector $\mathbf{x}$, i.e., $\mathbf{z}_{\mathrm{FN}m} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\mathrm{RF}}^2\mathbf{I}_N\right)$, where σ_{RF}^2 is the variance of each entry of $\mathbf{z}_{\mathrm{FN}m}$. $\mathbf{H}_{\mathrm{FN}m} \in \mathbb{C}^{N\times K_s}$ is the channel coefficient matrix corresponding to the RF links between all MDs active on subcarrier s and FN m. Channel matrix $\mathbf{H}_{\mathrm{FN}m}$ can be written as $\mathbf{H}_{\mathrm{FN}m} = [\mathbf{h}_{i_1,m}, \ldots, \mathbf{h}_{i_{K_s},m}]$, where $i_1, \ldots, i_{K_s}$ are the indices of the MDs which transmit on subcarrier s. Here, $\mathbf{h}_{k,m} \in \mathbb{C}^{N\times 1}$ is a vector containing the coefficients of the channel from the k-th MD to FN m. For the average power gain of $\mathbf{h}_{k,m}$, we employ the formula in [21, Equation (2)]. Due to the short distances between the MDs and FNs, the probability of a line-of-sight (LOS) link between the MDs and the FNs, denoted as P_{LOS}, is high. Hence, for the RF links, we adopt a probabilistic model such that 80% of the channels are Rician fading and 20% of the channels are Rayleigh fading modeling LOS and non-LOS scenarios, respectively [21]–[23]. Furthermore, we assume a Rician K-factor of $K_{\mathrm{RF}} = 6$ dB [21] for the RF LOS links.

2) mmWave LINKS

Measurement results in [11] have shown that mmWave LOS channels are very directional and include only few relevant components due to reflections. The number and strength of such components decreases with decreasing distance between transmitter and receiver. Thus, we assume the presence of only one direct path from the transmitter to the receiver. In addition, we assume ideal beamforming, such that the mmWave transmit antenna array with N_t antennas and the mmWave receive antenna array with N_r antennas can be equivalently modeled as single antennas with correspondingly large antenna gains, respectively. Consequently, we model the mmWave links between the FNs as SISO links. Each FN is equipped with several shielded transmit and receive antenna arrays to connect to other FNs. Note that the number of shielded transmit and receive antenna arrays at an FN is depending on its number of incoming and outcoming mmWave links, defined by the topology. Hence, the mmWave links are assumed to be independent and to not interfere with each other. Moreover, the FNs employ full duplex transmission, where self-interference is assumed to be negligible because of the shielding and high directionality of transmission. Hence, the mmWave signal received at FN m from FN n, $\mathbf{y}_{(n,m)}^{\mathrm{mmW}}$, $\forall(n, m) \in \mathcal{E}$, is an N_L-dimensional vector containing N_L frequency and time multiplexed signals and is given by

$$\mathbf{y}_{(n,m)}^{\mathrm{mmW}} = \mathbf{G}_{(n,m)}\mathbf{x}_{(n,m)}^{\mathrm{mmW}} + \mathbf{z}_{(n,m)} \in \mathbb{C}^{N_L\times 1}, \quad \forall(n, m) \in \mathcal{E}. \tag{2}$$

Here, $\mathbf{x}_{(n,m)}^{\mathrm{mmW}} \in \mathbb{C}^{N_L\times 1}$ is Gaussian distributed and denotes the mmWave transmit vector with N_L elements which are transmitted over N_L different mmWave time-frequency resource elements from the n-th FN to the m-th FN. Furthermore, $\mathbf{G}_{(n,m)} = \mathrm{diag}(g_{(n,m)}^{(1)}, g_{(n,m)}^{(2)}, \ldots, g_{(n,m)}^{(N_L)}) \in \mathbb{C}^{N_L\times N_L}$, where $g_{(n,m)}^{(i)} \in \mathbb{C}$ is the mmWave channel gain between FN n and FN m for mmWave time-frequency resource element i. Moreover, $\mathbf{z}_{(n,m)} \in \mathbb{C}^{N_L\times 1}$ denotes the zero-mean complex AWGN vector at FN m for the mmWave link from FN n to FN m, i.e., $\mathbf{z}_{(n,m)} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\mathrm{mmW}}^2\mathbf{I}_{N_L}\right)$, where σ_{mmW}^2 denotes the variance of the mmWave AWGN. We assume that $\mathbf{z}_{(n,m)}$ is independent of the transmit signal vector $\mathbf{x}_{(n,m)}^{\mathrm{mmW}}$. Furthermore, due to the high probability of having an LOS, we assume Rician fading for the mmWave links [12] with a Rician K-factor of $K_{\mathrm{mmW}} = 10$ dB [24]. We assume uniform power allocation across different mmWave subcarriers. Hence, the average power of the mmWave transmit signal from FN n to FN m, $\mathbf{x}_{(n,m)}^{\mathrm{mmW}}$ is constrained as

$$\mathbb{E}{-1mm}\left\{(\mathbf{x}_{(n,m)}^{\mathrm{mmW}})^{\mathrm{H}}\mathbf{x}_{(n,m)}^{\mathrm{mmW}}\right\} \leq \frac{N_L P_n^{\mathrm{mmW}}}{N_\rho}, \quad \forall(n, m) \in \mathcal{E}, \tag{3}$$

where P_n^{mmW} is the maximum mmWave transmit power of the n-th FN per link.

III. PERFORMANCE UPPER BOUND

In this section, we derive an upper bound on the achievable sum rate of the firefly network. This upper bound will allow us to investigate which links of the firefly network are the performance bottlenecks. In the following, we first introduce some variables which we require to formally present the proposed performance upper bound for the firefly network. Let $\mathcal{S}$ be a subset of $\mathcal{M}$ and $\mathcal{D}$ contain all the remaining elements of $\mathcal{M}$ which are not in $\mathcal{S}$. In other words, $\mathcal{S} \cup \mathcal{D} = \mathcal{M}$ and $\mathcal{S} \cap \mathcal{D} = \phi$ hold. These two sets describes a cut, whereby the FNs are separated namely into the sets of the transmitting FNs $\mathcal{S}$ and the receiving FNs $\mathcal{D}$. As an example, Fig. 2 illustrates one cut in a firefly network with eight FNs. Furthermore, $\mathbf{h}_{k,\mathcal{S}}^{s}$ denotes a vector of length $|\mathcal{D}| \cdot N$ containing all the RF channel coefficients from MD k to the FNs in set $\mathcal{D}$ at

The design of power and subcarrier allocation schemes for firefly ultra dense networks is beyond the scope of this paper but constitutes an interesting research problem for future work.

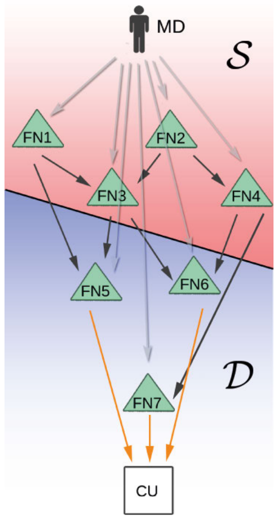

FIGURE 2. **Possible cut of the firefly network model specified by $\mathcal{S} = \{1, 2, 3, 4\}$ and $\mathcal{D} = \{5, 6, 7\}$.**

subcarrier s, whereas $g^{s}_{(m,n),\mathcal{S}}$ is the mmWave channel gain from FN m in $\mathcal{S}$ to FN n in $\mathcal{D}$ on subcarrier s. Furthermore, let $\mathbf{H}$ and $\mathbf{G}$ contain all RF and mmWave channel coefficients, respectively. Finally, let $\bar{R}_k$ denote the achievable ergodic rate of MD k to the CU. Exploiting these definitions, the proposed upper bound based on the cut-set principle [19] is provided in the following theorem.

Theorem 1: Assuming uniform power allocation across all RF and mmWave subcarriers, respectively, the achievable ergodic sum rate of the MDs to the CU for the firefly network specified by graph $(\mathcal{M}, \mathcal{E}, \mathcal{N})$ is upper bounded by

$$\sum_{k=1}^{K} \bar{R}_k \leq C^{\mathrm{MFMC}} \tag{4}$$

where

$$C^{\mathrm{MFMC}} = \mathbb{E}_{\mathbf{H},\mathbf{G}}\left\{ \min_{\forall \mathcal{S}\subset\mathcal{M}\setminus\mathcal{N}} \left(C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H}) + C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G}) \right) \right\}, \tag{5}$$

and

$$\begin{aligned} C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H}) &= \sum_{s=1}^{N_f} \frac{W^{\mathrm{RF}}}{N_f} \\ &\quad \cdot \log_2 \left| \mathbf{I}_{|\mathcal{D}|\cdot N} + \sum_{k=1}^{K} c_{k,s} \frac{P^{\mathrm{RF}}_{k}}{F_k \sigma^2_{\mathrm{RF}}} \mathbf{h}^{s}_{k,\mathcal{S}} \left(\mathbf{h}^{s}_{k,\mathcal{S}} \right)^{\mathrm{H}} \right|, \end{aligned} \tag{6}$$

$$\begin{aligned} C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G}) &= \sum_{s=1}^{N_\rho} \frac{W^{\mathrm{mmW}}}{N_\rho} \sum_{m\in\mathcal{S}} \sum_{n\in\mathcal{D}} a_{m,n} \\ &\quad \cdot \log_2 \left(1 + \frac{P^{\mathrm{mmW}}_{m}}{N_\rho \sigma^2_{\mathrm{mmW}}} |g^{s}_{(m,n),\mathcal{S}}|^2 \right), \end{aligned} \tag{7}$$

respectively.

Proof: The proof is provided in Appendix A.

Corollary 1: Under the same assumptions as in Theorem 1, C^{MFMC} is further upper bounded by

$$C^{\mathrm{MFMC}} \leq C^{\mathrm{UPP}}, \tag{8}$$

where

$$C^{\mathrm{UPP}} = \min_{\forall \mathcal{S}\subset\mathcal{M}\setminus\mathcal{N}} \left(\mathbb{E}_{\mathbf{H}}\left\{ C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H}) \right\} + \mathbb{E}_{\mathbf{G}}\left\{ C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G}) \right\} \right), \tag{9}$$

and $C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H})$ and $C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G})$ are defined as in (6) and (7), respectively.

Proof: The proof is provided in Appendix B.

Remark 1: Since both the RF and the mmWave channel coefficients are random variables, $C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H})$ and $C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G})$ vary and their minimum with respect to $\mathcal{S}$ may change from one channel realization to the next. In this paper, we assume that each codeword spans one channel realization and thus, Theorem 1 provides a realistic estimate of the data flow through the network, i.e., C^{MFMC} is the relevant upper bound. Note that by using Theorem 1, we only obtain a probabilistic measure of which links constitute the bottleneck cut. In comparison, Corollary 1 computes the expectation of the capacity of each cut, and thus, provides the performance bottleneck of the firefly network over a long time period. However, Corollary 1 ignores the instantaneous behavior of the firefly network at a given time step and thus, it is only applicable for a system design, where the instantaneous behavior of the network can be compensated, e.g., by using codewords that span many channel realizations. In this case, C^{UPP} becomes the upper bound for the achievable ergodic sum rate. To facilitate the discussion of our simulation results in Section V-D, we define further $C^{\mathrm{MAC}} = \mathbb{E}_{\mathbf{H}}\left\{ C^{\mathrm{MAC}}_{\phi}(\mathbf{H}) \right\}$ and refer to it as the *virtual MAC capacity*. Note that if $\mathcal{S} = \phi$, the set of transmitting FNs is empty and all FNs of the topology are in set $\mathcal{D}$, the set of receiving FNs. Thus, the corresponding cut describes the RF access from all MDs to all FNs, and hence, the MAC channel. In addition, we define $C^{\mathrm{FN}} = \min_{\forall \mathcal{S}\subseteq\{\mathcal{M}\setminus\nu_1,\ldots,\nu_T\}\wedge\mathcal{S}\neq\phi} \left(\mathbb{E}_{\mathbf{H}}\left\{ C^{\mathrm{MAC}}_{\mathcal{S}}(\mathbf{H}) \right\} + \mathbb{E}_{\mathbf{G}}\left\{ C^{\mathrm{mmW}}_{\mathcal{S}}(\mathbf{G}) \right\} \right)$ and refer to it as the *firefly network capacity*. Note that by excluding the cut, where $\mathcal{S} = \phi$, C^{FN} describes the capacity of the mmWave fronthaul part of the network. Thereby, the upper bound in (8) can be written as $C^{\mathrm{UPP}} = \min\{C^{\mathrm{MAC}}, C^{\mathrm{FN}}\}$ [25], [26]. This implies that when $C^{\mathrm{UPP}} = C^{\mathrm{MAC}}$ holds, the RF access part of the firefly network is the performance bottleneck. In contrast, when $C^{\mathrm{UPP}} = C^{\mathrm{FN}}$ holds, the mmWave fronthaul part of the network is the bottleneck.

IV. NONLINEAR FORWARDING VIA QUANTIZATION

In this section, we propose a nonlinear processing scheme for the FNs based on quantization of the received signals to reduce the amount of data that has to be forwarded to other FNs. The received MD symbols on different RF subcarriers are processed at each FN independently. Each FN receives N versions of the same MD symbol on a given RF subcarrier via the N RF antennas. In addition, depending on the network topology, it may also receive several already quantized versions of this MD symbol from other FNs it is connected to. On the other hand, each FN has a limited mmWave link capacity available for forwarding its received symbols to other FNs. As a result, the main challenge of nonlinear processing via quantization is how to compress the received data streams such that the resulting data rate does not exceed the fronthaul capacity of the outgoing mmWave link. Therefore, the covariance matrix of the quantization noise, the so-called distortion matrix of each FN, has to be optimized for each outgoing mmWave link.

A. NONLINEAR FORWARDING AT FN m

Each FN employs a sampling rate of $f_q \geq W_{\text{sub}}^{\text{RF}}$. However, considering a given FN m, the information content of $\mathbf{y}_{\text{FN}m}$, which is to be forwarded over the mmWave link (m, n), is in general larger than the limited fronthaul capacity, denoted as $C_{m,n}^{\text{mmW}}$, i.e.,

$$f_q \cdot H\left(\mathbf{y}_{\text{FN}m}\right) > C_{m,n}^{\text{mmW}}, \tag{10}$$

where $H\left(\mathbf{y}_{\text{FN}m}\right)$ is the entropy of $\mathbf{y}_{\text{FN}m}$. This is because $\mathbf{y}_{\text{FN}m}$ is a continuous random variable and hence, its entropy is infinitely large, see (1) and (2). However, the information content of $\mathbf{y}_{\text{FN}m}$, which originates from the MDs' transmit symbols, is finite, of course. Thus, to fully exploit the fronthaul capacity, all received data streams are jointly compressed. In particular, the overall received signal composed of the received signals from the MDs and the decompressed quantized signals received from neighboring FNs is compressed before it is forwarded to other FNs [16]. Thereby, each FN has to be informed about the adopted quantization codebook. We employ vector quantization, which is a lossy compression technique. In fact, the higher the capacity of the fronthaul links, the higher the resolution of the compressed signal can be. Thus, a key problem in vector quantization is the design of a good codebook of representative vectors which are typical for the data to be sent, and in this way, reduce the distortion caused by quantization [17]. If there are more than one outgoing mmWave link at an FN, the overall received signal at this FN is compressed for each available outgoing mmWave link separately via independent codebooks [16]. Hence, the overall received signal at FN m is compressed separately for each outgoing mmWave link as the distortion matrix $\mathbf{Q}_{(m,n)}$ caused by compressing $\mathbf{y}_{\text{FN}_m}$ for forwarding to FN n depends on the mmWave link capacity $C_{m,n}^{\text{mmW}}$, cf. Section IV-B. We assume that this results in independent quantization noises at FNs which have a mmWave link connection to the same FN [16]. Moreover, we define $\mathcal{E}_m$ as the set of all available mmWave links which influence the signal received at FN m, due to the multi-hop structure of the given topology. Furthermore, for simplicity, we assume that the considered FN topologies also exhibit at most two incoming mmWave links per FN. Thus, considering one RF time-frequency resource, cf. Subsection II-1, the overall received signal at FN m, $\mathbf{y}_{\text{FN}m}$, is given by

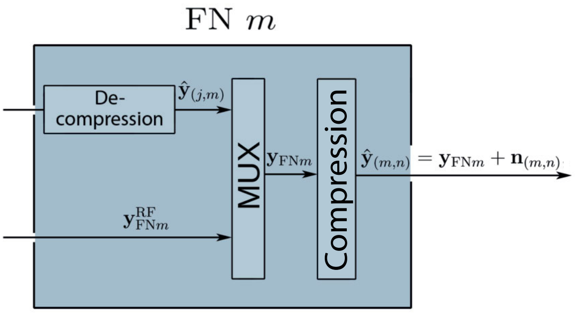
FN m
De-compression
$\hat{\mathbf{y}}_{(j,m)}$
MUX
$\mathbf{y}_{\text{FN}m}$
Compression
$\hat{\mathbf{y}}_{(m,n)} = \mathbf{y}_{\text{FN}m} + \mathbf{n}_{(m,n)}$
$\mathbf{y}_{\text{FN}m}^{\text{RF}}$

FIGURE 3. Illustration of the nonlinear processing at the m-th FN. $\hat{\mathbf{y}}_{(m,n)}$ is transmitted to FN n.

$$\mathbf{y}_{\text{FN}m} = \left[\left(\hat{\mathbf{y}}_{\text{FN}m}\right)^{\text{T}}, \left(\mathbf{y}_{\text{FN}m}^{\text{RF}}\right)^{\text{T}} \right]^{\text{T}}, \quad \forall m \in \{1, \ldots, M\}, \tag{11}$$

where the quantized signal received from other FNs at FN m, $\hat{\mathbf{y}}_{\text{FN}m}$ is defined as

$$\hat{\mathbf{y}}_{\text{FN}m} = \text{Vercut}\left(\hat{\mathbf{y}}_{(j,m)}, j \in \mathcal{M} : (j, m) \in \mathcal{E}\right). \tag{12}$$

Here, $\mathbf{y}_{\text{FN}m}^{\text{RF}} \in \mathbb{C}^{N \times 1}$ is the received RF signal at FN m defined as $\mathbf{y}_{\text{FN}m}^{\text{RF}} = \left[y_{\text{FN}m}^{\text{RF}(1)}, \ldots, y_{\text{FN}m}^{\text{RF}(N)} \right]^{\text{T}}$, where $y_{\text{FN}m}^{\text{RF}(n)} \in \mathbb{C}$, $n \in \{1, \ldots, N\}$, is the received signal at RF receive antenna n of FN m. Moreover, $\hat{\mathbf{y}}_{(j,m)}$ for $(j, m) \in \mathcal{E}$ is the quantized version of $\mathbf{y}_{\text{FN}j}$ to be forwarded over mmWave link (j, m) to FN m. Since the fronthaul capacity constraint is met, cf. Section IV-B, the decompressor at the receiving FN can identify the transmitted quantized signal in its quantization codebook. This means that at the receiving FN, the perfectly decompressed mmWave signal from a transmitting FN is identical to the quantized transmit signal of this transmitting FN. Figure 3 illustrates the forwarding scheme at FN m, where we assume that FN m has an incoming mmWave link from FN j and $\hat{\mathbf{y}}_{(m,n)}$ is nonlinearly processed for transmission to FN n. In general, $\mathbf{y}_{\text{FN}m}$ is quantized at FN m before it is forwarded over the mmWave link to FN n and based on rate-distortion theory, the quantized signal, $\hat{\mathbf{y}}_{(m,n)}$, is modeled as

$$\hat{\mathbf{y}}_{(m,n)} = \mathbf{y}_{\text{FN}m} + \mathbf{n}_{(m,n)}, \tag{13}$$

where $\mathbf{n}_{(m,n)} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times 1}$ denotes the Gaussian quantization noise of the compressed signal transmitted from FN m to FN n. $\mathbf{n}_{(m,n)}$ is independent of $\mathbf{y}_{\text{FN}m}$ and follows a Gaussian distribution where $\mathbf{n}_{(m,n)} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{(m,n)})$ with $\mathbf{Q}_{(m,n)} = \mathbb{E}\left\{\mathbf{n}_{(m,n)}\mathbf{n}_{(m,n)}^{\text{H}}\right\}$ being the distortion matrix of the compressed transmit signal of FN m for FN n [26]. Thus, the quantization noise statistics are fully characterized by $\mathbf{Q}_{(m,n)}$. Note that since we assume that all elements of $\mathbf{y}_{\text{FN}m}$ are jointly quantized, $\mathbf{Q}_{(m,n)}$ is in general non-diagonal [18].

In general, the overall received signal at FN m, $\mathbf{y}_{\text{FN}m}$, $m \in \{1, \ldots, M\}$, is an $(|\mathcal{E}_m| + 1)N$-dimensional vector, which can be modeled as

$$\mathbf{y}_{\text{FN}m} = \bar{\mathbf{H}}_{\text{FN}m}\mathbf{x} + \bar{\mathbf{z}}_{\text{FN}m} + \bar{\mathbf{n}}_{\text{FN}m}, \tag{14}$$

where $\bar{\mathbf{H}}_{\mathrm{FN}m} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times K_s}$, $\bar{\mathbf{z}}_{\mathrm{FN}m} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times 1}$, and $\bar{\mathbf{n}}_{\mathrm{FN}m} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times 1}$, $\bar{\mathbf{n}}_{\mathrm{FN}m} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{Q}}_{\mathrm{FN}_m})$, are the stacked RF channels, the stacked RF noise vectors, and the stacked quantization noise vectors of the mmWave links which have an influence on the signal received at FN m, respectively. Here, $\bar{\mathbf{Q}}_{\mathrm{FN}_m} = \mathbb{E}\left\{\bar{\mathbf{n}}_{\mathrm{FN}m}\bar{\mathbf{n}}_{\mathrm{FN}m}^{\mathrm{H}}\right\}$ is the distortion matrix of the quantization noise received at FN m. Due to the received N-dimensional RF signal, which is stacked as the "last" signal in $\mathbf{y}_{\mathrm{FN}m}$, see (11), $\bar{\mathbf{Q}}_{\mathrm{FN}m}$ is a block diagonal matrix where the "last" $N \times N$-dimensional diagonal block is equal to $\mathbf{0}_{N \times N}$.

Assuming that FN m has a mmWave link to FN n, $\hat{\mathbf{y}}_{(m,n)} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times 1}$ in (13) can be modeled as

$$\hat{\mathbf{y}}_{(m,n)} = \bar{\mathbf{H}}_{\mathrm{FN}m}\mathbf{x} + \bar{\mathbf{z}}_{\mathrm{FN}m} + \bar{\mathbf{n}}_{\mathrm{FN}m} + \mathbf{n}_{(m,n)}. \tag{15}$$

The overall received signal at the CU, $\mathbf{y}_{\mathrm{CU}}$, comprises the received signals at all available root nodes $\nu_1, \ldots, \nu_T$. At the root nodes, compression is not necessary due to the infinite capacity of the optical fiber link which connects the root nodes to the CU. Hence, $\mathbf{y}_{\mathrm{CU}}$ is defined as

$$\mathbf{y}_{\mathrm{CU}} = \left[\mathbf{y}_{\mathrm{FN}_{\nu_1}}^{\mathrm{T}}, \ldots, \mathbf{y}_{\mathrm{FN}_{\nu_T}}^{\mathrm{T}}\right]^{\mathrm{T}} = \bar{\mathbf{H}}\mathbf{x} + \bar{\mathbf{z}} + \bar{\mathbf{n}}, \tag{16}$$

where $\bar{\mathbf{H}} \in \mathbb{C}^{((|\mathcal{E}_{\nu_1}|+1)N+\cdots+(|\mathcal{E}_{\nu_T}|+1)N) \times K_s}$, $\bar{\mathbf{z}} \in \mathbb{C}^{((|\mathcal{E}_{\nu_1}|+1)N+\cdots+(|\mathcal{E}_{\nu_T}|+1)N) \times 1}$, and $\bar{\mathbf{n}} \in \mathbb{C}^{((|\mathcal{E}_{\nu_1}|+1)N+\cdots+(|\mathcal{E}_{\nu_T}|+1)N) \times 1}$, $\bar{\mathbf{n}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{Q}})$, are the stacked RF channel matrices, the stacked RF noise vectors, and the stacked quantization noise vectors of all RF channels received at the CU, respectively. Moreover, the total distortion matrix $\bar{\mathbf{Q}} = \mathbb{E}\left\{\bar{\mathbf{n}}\bar{\mathbf{n}}^{\mathrm{H}}\right\}$ is a block diagonal matrix, where the blocks corresponding to the received N-dimensional RF signal are $N \times N$ all-zero matrices, $\mathbf{0}_{N \times N}$.

B. FRONTHAUL LINK CAPACITY CONSTRAINT

As illustrated in Figure 3, at FN m, the signal received from the access network is multiplexed and then compressed jointly with the already compressed signal received from neighboring FN j which is connected to FN m. This multiplexing increases the bandwidth requirement from one FN to the next. Hence, the more the fronthaul link capacities $C_{m,n}^{\mathrm{mmW}}$, $\forall (m,n) \in \mathcal{E}$ are limited, the higher may be the needed compression rates leading to significant distortions and potentially low sum rates [29]. Therefore, the compression across the ultra dense network should be carefully optimized. This can be achieved by optimizing the distortion matrices such that the received signals are efficiently compressed before forwarding them to the next FNs in the multi-hop architecture.

Considering vector quantization, the mutual information between $\mathbf{y}_{\mathrm{FN}m}$ and $\hat{\mathbf{y}}_{(m,n)}$ is obtained as [29]

$$\begin{aligned} &I(\mathbf{y}_{\mathrm{FN}m}; \hat{\mathbf{y}}_{(m,n)}) \\ &\quad = \log_2 \frac{|\bar{\mathbf{H}}_{\mathrm{FN}m}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}_{\mathrm{FN}m}^{\mathrm{H}} + \Sigma_{\bar{\mathbf{z}}_{\mathrm{FN}m}} + \bar{\mathbf{Q}}_{\mathrm{FN}m} + \mathbf{Q}_{(m,n)}|}{|\mathbf{Q}_{(m,n)}|}, \end{aligned} \tag{17}$$

where $\Sigma_{\mathbf{x}}$ and $\Sigma_{\bar{\mathbf{z}}_{\mathrm{FN}m}}$ are the covariance matrices of the MDs' signals, $\mathbf{x}$, and the stacked RF noise vectors received at FN m, $\bar{\mathbf{z}}_{\mathrm{FN}m}$, respectively. Based on the source coding theorem [19] and the channel coding theorem [19], in order to be able to perfectly decompress $\hat{\mathbf{y}}_{(m,n)}$ at FN n, i.e., for FN n to be able to identify the quantized signal in its codebook [18], the following constraint has to be fulfilled:

$$f_q I\left(\mathbf{y}_{\mathrm{FN}m}; \hat{\mathbf{y}}_{(m,n)}\right) \leq C_{m,n}^{\mathrm{mmW}}, \quad \forall m \in \mathcal{M}\backslash\mathcal{N}, (m,n) \in \mathcal{E}, \tag{18}$$

where the fronthaul link capacity of mmWave link (m, n) is given by

$$C_{m,n}^{\mathrm{mmW}} = \frac{W^{\mathrm{mmW}}}{N_f}\log_2\left(1 + \frac{P_m^{\mathrm{mmW}}}{N_f\sigma_{\mathrm{mmW}}^2}|g_{(m,n)}|^2\right). \tag{19}$$

Note that we allocate the same mmWave bandwidth, $\frac{W^{\mathrm{mmW}}}{N_f}$, to each RF subcarrier.

In the following subsection, we will formulate optimization problems for the design of the distortion matrices $\mathbf{Q}_{(m,n)}$, $\forall (m,n) \in \mathcal{E}$.

C. PROBLEM FORMULATION AND SOLUTION

In this subsection, we propose an optimization problem for a central design strategy, where the total distortion matrix $\bar{\mathbf{Q}}$ is to be centrally designed at the CU and thus, is expected to yield the best performance results. However, the central design strategy would require global CSI at the CU. Moreover, this optimization problem is difficult to tackle and requires high complexity. Therefore, we also introduce an alternative optimization problem for a local design strategy, where the distortion matrices for each FN, $\bar{\mathbf{Q}}_{\mathrm{FN}m}$, $\forall\, m \in \mathcal{M}\backslash\mathcal{N}$, are designed locally at the corresponding FN. The latter approach entails a lower complexity. Moreover, while the central design strategy requires global CSI at the CU, for the local design strategy, a given FN m requires only full receiver CSI (CSIR) and knowledge of all $C_{m,n}^{\mathrm{mmW}}$, $\forall (m,n) \in \mathcal{E}$. However, it is expected that the local design strategy yields a lower achievable sum rate than central design strategy, of course.

1) OPTIMIZATION PROBLEM

We first consider the so-called *central design strategy*, where we assume that all distortion matrices are computed centrally at the CU and full CSI is available at the CU.

a: CENTRAL DESIGN STRATEGY

We employ the sum rate as the performance metric for optimization of $\bar{\mathbf{Q}}$. Hence, the sum rate of all MDs is maximized with respect to the distortion matrices $\mathbf{Q}_{(m,n)}$, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, $(m,n) \in \mathcal{E}$, to find the optimal compression strategies to be used at the FNs. Moreover, we fix $f_q = W_{\mathrm{sub}}^{\mathrm{RF}}$ and note that the CU performs joint decoding of the messages of all MD signals based on its overall received signal $\mathbf{y}_{\mathrm{CU}}$.

Finding quantization schemes which correspond to the determined distortion matrices, i.e., designing practical quantization code books is beyond the scope of this paper but constitutes an interesting research problem for the future.

The optimal distortion matrices, $\mathbf{Q}^{\text{opt}}_{(m,n)}$, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, $(m,n) \in \mathcal{E}$, or equivalently, the optimal total distortion matrix $\bar{\mathbf{Q}}$ maximizing the sum rate, are obtained from the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{Q}_{(m,n)},\forall m\in\mathcal{M}\backslash\mathcal{N},(m,n)\in\mathcal{E}} \quad R(\bar{\mathbf{Q}}) \\ &\quad\text{subject to}\, f_q I(\mathbf{y}_{\text{FN}m}; \hat{\mathbf{y}}_{(m,n)}) \leq C^{\text{mmW}}_{m,n}, \\ &\qquad\qquad \forall m \in \mathcal{M}\backslash\mathcal{N}, \quad \forall (m,n) \in \mathcal{E} \end{aligned} \quad (20)$$

where the rate of the proposed nonlinear forwarding scheme as a function of the total distortion matrix, denoted as $R(\bar{\mathbf{Q}})$, is calculated as

$$R(\bar{\mathbf{Q}}) = W^{\text{RF}}_{\text{sub}} \log_2 |\mathbf{I}_{(|\mathcal{E}|+1)N} + \bar{\mathbf{H}}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\text{H}}(\Sigma_{\bar{\mathbf{z}}} + \bar{\mathbf{Q}})^{-1}|. \quad (21)$$

Based on the Shannon-Hartley theorem [19], equation (21) constitutes as the formula of the channel capacity, where $\bar{\mathbf{H}}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\text{H}}(\Sigma_{\bar{\mathbf{z}}}+\bar{\mathbf{Q}})^{-1}$ describes the signal-to-noise-ratio (SNR) of the overall firefly ultra dense network. The problem in (20) is difficult to solve since jointly optimizing all distortion matrices, $\mathbf{Q}_{(m,n)}$, $\forall (m,n) \in \mathcal{E}$, entails a high complexity due to the high dimensionality of the optimization problem. Furthermore, note that each distortion matrix, $\mathbf{Q}_{(m,n)}$, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, $(m,n) \in \mathcal{E}$, depends also on all previous distortion matrices in the forwarding topology network which have an impact on the received signal at FN m. This dependency of a distortion matrix on other distortion matrices further contributes to the high complexity of the problem in (20). In addition, the problem in (20) is non-convex and would require global CSI at the CU. Hence, for practical reasons, we focus on solving the following optimization problem for the so-called *local design strategy*.

b: LOCAL DESIGN STRATEGY

In comparison to problem (20), for the local design strategy, we design the distortion matrices at each FN for each outgoing mmWave link for an efficient compression of the received signals. Therefore, an optimality criterion is needed to quantify the performance. In (20), the optimality criterion was the achievable sum rate at the CU. For the local design strategy, since the sum rate depends on the compression strategy used by the FNs via the distortion matrices $\mathbf{Q}_{(m,n)}$, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, $\forall (m,n) \in \mathcal{E}$, ideally, at FN m we would locally design the distortion matrix for each outgoing mmWave link (m,n) such that the achievable sum rate at the CU is maximized. However, using the sum rate at the CU as the objective function leads to an untractable optimization problem in terms of $\mathbf{Q}_{(m,n)}$ at FN m, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, $\forall (m,n) \in \mathcal{E}$. Thus, we adopt the maximization of the mutual information, $I(\mathbf{x}; \hat{\mathbf{y}}_{(m,n)})$, between the MDs' signals and the quantized version of the overall received signal at FN m for forwarding to FN n as optimality criterion instead. In other words, we maximize the information content of the quantized signal while meeting the fronthaul link capacity constraint. Note that f_q is again fixed to $W^{\text{RF}}_{\text{sub}}$, i.e., $f_q = W^{\text{RF}}_{\text{sub}}$.

For FN m, $\forall m \in \mathcal{M}\backslash\mathcal{N}$, the optimal distortion matrices, $\mathbf{Q}^{\text{opt}}_{(m,n)}$, $\forall\ (m,n) \in \mathcal{E}$, maximizing the mutual information between the MDs' signals and the signal quantized at FN m for forwarding to FN n, $I(\mathbf{x}; \hat{\mathbf{y}}_{(m,n)})$, are obtained from the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{Q}_{(m,n)}} \quad I(\mathbf{x}; \hat{\mathbf{y}}_{(m,n)}) \\ &\text{subject to} \quad f_q I(\mathbf{y}_{\text{FN}m}; \hat{\mathbf{y}}_{(m,n)}) \leq C^{\text{mmW}}_{m,n}, \end{aligned} \quad (22)$$

where the objective function $I(\mathbf{x}; \hat{\mathbf{y}}_{(m,n)})$ is given by

$$\begin{aligned} &I(\mathbf{x}; \hat{\mathbf{y}}_{(m,n)}) \\ &\quad = \log_2 \frac{|\bar{\mathbf{H}}_{\text{FN}m}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\text{H}}_{\text{FN}m} + \Sigma_{\bar{\mathbf{z}}_{\text{FN}m}} + \bar{\mathbf{Q}}_{\text{FN}m} + \mathbf{Q}_{(m,n)}|}{|\Sigma_{\bar{\mathbf{z}}_{\text{FN}m}} + \bar{\mathbf{Q}}_{\text{FN}m} + \mathbf{Q}_{(m,n)}|}. \end{aligned} \quad (23)$$

We assume that all $\mathbf{Q}_{(m,n)}$, $\forall (m,n) \in \mathcal{E}_m$ are optimized according to (22). Furthermore, note that (22) is formulated for each outgoing mmWave link of FN m separately. Thus, the optimization problem in (22) is easier to solve than the optimization problem in (20) since the distortion matrix of a given mmWave link is designed which requires only the CSIR at the transmitting FN and knowledge of the fronthaul link capacity of the considered mmWave link, i.e., less CSI is needed compared to (20). Moreover, the dimensionality of the optimization problem in (22) is reduced by a factor of $|\mathcal{E}|$ compared to that of (20).

2) SOLUTION VIA DUAL PROBLEM

Problem (22) is a non-convex optimization problem in $\mathbf{Q}_{(m,n)}$, as the objective function of the maximization problem is convex in $\mathbf{Q}_{(m,n)}$ instead of concave. In the following, we present a method that can handle the non-convexity of problem (22). Here, the constrained optimization problem in (22) is converted to an unconstrained problem using the dual of the problem. Then, the primal variables can be computed in closed form. We show that the dual problem is monotonic in the dual variable. Hence, we propose a bisection algorithm to find the optimal value. Moreover, the dual problem results in an optimal solution since strong duality is valid. In this subsection, we first formulate the Lagrangian, then we propose an iterative optimization algorithm for maximizing the Lagrangian.

The Lagrangian of optimization problem (22) is given by

$$\mathcal{L}(\mathbf{Q}_{(m,n)}, \mu) = F_o\left(\mathbf{Q}_{(m,n)}\right) - \mu\left(F_c\left(\mathbf{Q}_{(m,n)}\right) - C^{\text{mmW}}_{m,n}/f_q\right), \quad (24)$$

where μ is the Lagrange multiplier, the objective function is defined as

$$\begin{aligned} F_o\left(\mathbf{Q}_{(m,n)}\right) &= \log_2|\bar{\mathbf{H}}_{\text{FN}m}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\text{H}}_{\text{FN}m} + \Sigma_{\bar{\mathbf{z}}_{\text{FN}m}} + \bar{\mathbf{Q}}_{\text{FN}m} + \mathbf{Q}_{(m,n)}| \\ &\qquad \log_2|\Sigma_{\bar{\mathbf{z}}_{\text{FN}m}} + \bar{\mathbf{Q}}_{\text{FN}m} + \mathbf{Q}_{(m,n)}|, \end{aligned} \quad (25)$$

and the constraint function is given by

$$\begin{aligned} F_c\left(\mathbf{Q}_{(m,n)}\right) &= \log_2|\bar{\mathbf{H}}_{\text{FN}m}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\text{H}}_{\text{FN}m} + \Sigma_{\bar{\mathbf{z}}_{\text{FN}m}} + \bar{\mathbf{Q}}_{\text{FN}m} + \mathbf{Q}_{(m,n)}| \\ &\qquad - \log_2|\mathbf{Q}_{(m,n)}|. \end{aligned} \quad (26)$$

The dual function of (22) is stated as

$$\mathcal{D}(\mu) = \max_{\mathbf{Q}_{(m,n)} \succeq 0} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu), \quad (27)$$

and the corresponding dual problem is given by

$$\min_{\mu\geq 0}\mathcal{D}(\mu). \tag{28}$$

The overall approach for solving (22) is to find the dual function (27) and then to search in an outer loop for the optimal μ denoted as μ^* which results in

$$F_c\left(\mathbf{Q}^*_{(m,n)}\right) = C^{\mathrm{mmW}}_{m,n}/f_q. \tag{29}$$

Note that for a fixed Lagrange multiplier μ, a stationary point of the Lagrangian is found by (27) since the original problem is non-convex. μ^* must be in $[0, 1)$ since $\mu \geq 0$ holds due to the dual feasibility condition [30]. Moreover, in case of $\mu \geq 1$, $\mathbf{Q}^*_{(m,n)}$ would be infinite and $F_c\left(\mathbf{Q}^*_{(m,n)}\right) = 0$, see (48) in Appendix C. The outer loop for searching for μ^* corresponds to a one-dimensional root finding problem that can be solved by using a standard bisection method since the dual function is a convex function in μ.

Now, we provide a closed-form solution for the $\mathbf{Q}_{(m,n)}$ that maximizes the Lagrangian in (24) for a fixed $\mu \in [0, 1)$.

The Lagrangian in (24) can be rewritten as

$$\begin{aligned}\mathcal{L}(\mathbf{Q}_{(m,n)},\mu) = -\mu)\log_2|\mathbf{A}+\mathbf{Q}_{(m,n)}| + \mu\log_2|\mathbf{Q}_{(m,n)}|\\ -\log_2|\mathbf{B}+\mathbf{Q}_{(m,n)}| + \mu C^{\mathrm{mmW}}_{m,n}/f_q,\end{aligned} \tag{30}$$

where $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\mathbf{A} = \bar{\mathbf{H}}_{\mathrm{FN}m}\Sigma_{\mathbf{x}}\bar{\mathbf{H}}^{\mathrm{H}}_{\mathrm{FN}m} + \Sigma_{\bar{\mathbf{z}}_{\mathrm{FN}m}} + \bar{\mathbf{Q}}_{\mathrm{FN}m}, \tag{31}$$

$$\mathbf{B} = \Sigma_{\bar{\mathbf{z}}_{\mathrm{FN}m}} + \bar{\mathbf{Q}}_{\mathrm{FN}m}. \tag{32}$$

The key for maximizing (30) is the following simultaneous diagonalization of $\mathbf{A}$ and $\mathbf{B}$ based on [31, Corollary 7.6.5]:

Lemma 1 (Generalized Eigen-Decomposition): For Hermitian positive definite matrices $\mathbf{A} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N\times(|\mathcal{E}_m|+1)N}$ and $\mathbf{B} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N\times(|\mathcal{E}_m|+1)N}$, there exists a non-singular matrix $\mathbf{C} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N\times(|\mathcal{E}_m|+1)N}$ such that $\mathbf{C}^{\mathrm{H}}\mathbf{A}\mathbf{C} = \Lambda$ and $\mathbf{C}^{\mathrm{H}}\mathbf{B}\mathbf{C} = \mathbf{I}_{(|\mathcal{E}_m|+1)N}$, where Λ is a diagonal matrix. The diagonal elements λ_i of Λ are called the generalized eigenvalues, see the following proof for their definition. Moreover, $\lambda_i \geq 1$ for $i = 1, \ldots, (|\mathcal{E}_m| + 1)N$.

Proof: $\mathbf{A}$ and $\mathbf{B}$ are both Hermitian positive definite matrices. Thus, let $\mathbf{B}^{-1} = \mathbf{R}^{\mathrm{H}}\mathbf{R}$ be a unique Cholesky-decomposition of $\mathbf{B}^{-1}$, where $\mathbf{R}$ is an upper triangular matrix. Now, consider the eigen-decomposition $\mathbf{R}\mathbf{A}\mathbf{R}^{\mathrm{H}} = \mathbf{V}\Lambda\mathbf{V}^{\mathrm{H}}$. Note that $\mathbf{R}\mathbf{A}\mathbf{R}^{\mathrm{H}}$ is a complex normal matrix (a complex matrix $\mathbf{A}$ is normal if it satisfies $\mathbf{A}^{\mathrm{H}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{H}}$). Then, $\mathbf{C} = \mathbf{R}^{\mathrm{H}}\mathbf{V}$ satisfies both $\mathbf{C}^{\mathrm{H}}\mathbf{A}\mathbf{C} = \Lambda$ and $\mathbf{C}^{\mathrm{H}}\mathbf{B}\mathbf{C} = \mathbf{I}_{(|\mathcal{E}_m|+1)N}$. Moreover, since $\mathbf{C}$ is non-singular, $\mathbf{A} \succeq \mathbf{B}$ implies $\Lambda \succeq \mathbf{I}_{(|\mathcal{E}_m|+1)N}$. □

Algorithm 1 Algorithm of Dual Problem

```
initialize $\mu \in [0, 1)$ and error tolerance $0 \leq \varepsilon_T \ll 1$.
Compute $\mathbf{A}$ and $\mathbf{B}$ from (31) and (32).
Given $\mathbf{A}$ and $\mathbf{B}$, compute $\mathbf{C}$ and $\Lambda$ from Lemma 1.
repeat
Given $\mu$ and $\lambda_i$, compute optimal $\Sigma^{(i,i),*}_{\hat{\mathbf{Q}}}$ from (48).
Given $\Sigma^{(i,i),*}_{\hat{\mathbf{Q}}}$ and $\mathbf{C}$, compute optimal $\mathbf{Q}^*_{(m,n)}$ from (49).
Update $\mu$ using bisection.
until $F_c\left(\mathbf{Q}^*_{(m,n)}\right) - C^{\mathrm{mmW}}_{m,n}/f_q \leq \varepsilon_T$.
```

Now, we apply the approach in [20], [32], [33] to reduce the matrix optimization problem to a scalar problem and to solve the resulting scalar optimization problem. For $\mu \in (0, 1]$, Lagrangian (30) can be written as (33),

$$\begin{aligned}\mathcal{L} &= (1-\mu)\log_2\frac{|\mathbf{A}+\mathbf{Q}_{(m,n)}|}{|\mathbf{B}+\mathbf{Q}_{(m,n)}|} + \mu\log_2\frac{|\mathbf{Q}_{(m,n)}|}{|\mathbf{B}+\mathbf{Q}_{(m,n)}|}\\ &\quad+\mu C^{\mathrm{mmW}}_{m,n}/f_q\\ &= (1-\mu)\log_2\frac{|\mathbf{C}^{\mathrm{H}}(\mathbf{A}+\mathbf{Q}_{(m,n)})\mathbf{C}|}{|\mathbf{C}^{\mathrm{H}}(\mathbf{B}+\mathbf{Q}_{(m,n)})\mathbf{C}|}\\ &\quad+\mu\log_2\frac{|\mathbf{C}^{\mathrm{H}}\mathbf{Q}_{(m,n)}\mathbf{C}|}{|\mathbf{C}^{\mathrm{H}}(\mathbf{B}+\mathbf{Q}_{(m,n)})\mathbf{C}|} + \mu C^{\mathrm{mmW}}_{m,n}/f_q\\ &\overset{(a)}{=} (1-\mu)\log_2\frac{|\Lambda+\hat{\mathbf{Q}}_{(m,n)}|}{|\mathbf{I}_{(|\mathcal{E}_m|+1)N}+\hat{\mathbf{Q}}_{(m,n)}|}\\ &\quad+\mu\log_2\frac{|\hat{\mathbf{Q}}_{(m,n)}|}{|\mathbf{I}_{(|\mathcal{E}_m|+1)N}+\hat{\mathbf{Q}}_{(m,n)}|} + \mu C^{\mathrm{mmW}}_{m,n}/f_q\\ &= (1-\mu)\log_2|\Lambda\hat{\mathbf{Q}}^{-1}_{(m,n)}+\mathbf{I}_{(|\mathcal{E}_m|+1)N}|\\ &\quad-\log_2|\hat{\mathbf{Q}}^{-1}_{(m,n)}+\mathbf{I}_{(|\mathcal{E}_m|+1)N}| + \mu C^{\mathrm{mmW}}_{m,n}/f_q\\ &\overset{(b)}{\leq} (1-\mu)\log_2|\Lambda\Sigma^{-1}_{\hat{\mathbf{Q}}}+\mathbf{I}_{(|\mathcal{E}_m|+1)N}|\\ &\quad-\log_2|\Sigma^{-1}_{\hat{\mathbf{Q}}}+\mathbf{I}_{(|\mathcal{E}_m|+1)N}| + \mu C^{\mathrm{mmW}}_{m,n}/f_q,\end{aligned}$$

where (a) follows from Lemma 1 and by defining $\hat{\mathbf{Q}}_{(m,n)} = \mathbf{C}^{\mathrm{H}}\mathbf{Q}_{(m,n)}\mathbf{C}$, with $\mathbf{C}$ as in Lemma 1, and inequality (b) follows from [33, Lemma 5], where $\Sigma_{\hat{\mathbf{Q}}}$ is due to the eigen-decomposition $\hat{\mathbf{Q}}_{(m,n)} = \mathbf{U}\Sigma_{\hat{\mathbf{Q}}}\mathbf{U}^{\mathrm{H}}$. Note that in (b) equality holds for $\mathbf{U} = \mathbf{I}_{(|\mathcal{E}_m|+1)N}$. In particular, for any nondiagonal $\hat{\mathbf{Q}}_{(m,n)}$, there exists a diagonal matrix $\Sigma_{\hat{\mathbf{Q}}}$ that achieves a higher $\mathcal{L}$. Thus, without loss of optimality, $\hat{\mathbf{Q}}_{(m,n)}$ is restricted to be diagonal.

Proposition 1: For a non-singular matrix $\mathbf{C} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N\times(|\mathcal{E}_m|+1)N}$ and optimal $\Sigma_{\hat{\mathbf{Q}}}$ denoted as $\Sigma^*_{\hat{\mathbf{Q}}}$, the optimal distortion matrix $\mathbf{Q}_{(m,n)}$ is given by

$$\mathbf{Q}^*_{(m,n)} = \left(\mathbf{C}^{-1}\right)^{\mathrm{H}}\Sigma^*_{\hat{\mathbf{Q}}}\mathbf{C}^{-1}. \tag{33}$$

Proof: The proof is provided in Appendix C.

The overall iterative approach is summarized in Algorithm 1, where we initialize first μ with a suitable value, compute the matrices $\mathbf{A}$ and $\mathbf{B}$, and apply Lemma 1 to obtain $\mathbf{C}$ and Λ. Then, we iteratively compute the optimal distortion matrix $\mathbf{Q}_{(m,n)}$ and update μ by using bisection until the constraint in (22) is fulfilled within the desired error tolerance ε_T. Moreover, the following theorem holds.

Theorem 2: Algorithm 1 is guaranteed to converge within the tolerance ε_T to the optimal solution of the original problem (22) which fulfills equation (29), i.e., $F_c\left(\mathbf{Q}^*_{(m,n)}\right) = C^{\mathrm{mmW}}_{m,n}/f_q$. In addition, strong duality is valid

TABLE 1. **Computational complexity comparison, where $W := (|\mathcal{E}_m| + 1)N$, $D := M \cdot N$, $F := |\mathcal{E}|N_L$, $R := N + N_L n_m^{\text{link}}$, and $J := Nn_l^{\text{FN}} + N_L n_l^{\text{in}}$ [15].**

Forwarding Scheme	Computational Complexity
LOCAL MMSE R [15]	$\mathcal{O}(K_sR^2 + RK_s^2 + RD^2 + DJR + JR^2 + FJR + RF^2 + R^3 + K_sN_L^2 + N_LK_s^2 + N_LR^2 + K_sN_LR)$
LOCAL PCA [15]	$\mathcal{O}(K_sR^2 + RK_s^2 + RD^2 + DJR + JR^2 + FJR + RF^2 + R^3 + K_s^3 + RK_s^2 + K_sR^2 + N_LK_s^2 + N_LR^2 + K_sN_LR)$
NONLINEAR	$\mathcal{O}(K_s^2W + K_sW^2 + W^3) + \mathcal{O}(T_iW^3)$

between the primal problem in (22) and the dual problem in (28).

Proof: The proof is provided in Appendix D.

D. COMPLEXITY COMPARISON OF LOCALLY-DESIGNED NONLINEAR AND LINEAR FORWARDING

In this section, we compare the computational complexity of the proposed nonlinear forwarding scheme based on the local design strategy in Section IV-C2 with the two locally designed linear forwarding strategies proposed in [15]. For simplicity, we denote the proposed local design strategy as ‘‘NONLINEAR’’. The linear forwarding strategy, where the mean squared error (MSE) is minimized at each FN before dimension reduction via principal component analysis (PCA), is referred to as ‘‘LOCAL PCA’’ [15]. Moreover, ‘‘LOCAL MMSE R‘‘ denotes the linear forwarding scheme, where a dimension reduction is accomplished via a pre-defined combining matrix, and subsequently the MSE is minimized [15]. Table 1 summarizes the computational complexities (number of multiplications) of the three considered forwarding strategies at an FN. In Table 1, $\mathcal{O}(\cdot)$ is the big-O notation and T_i stands for the number of iterations required for Algorithm 1 to converge. In addition, n_m^{link} denotes the number of incoming mmWave links at FN m, n_l^{FN} is the number of FNs in layer l, and n_l^{in} denotes the number of incoming mmWave and delay links of layer l [15]. LOCAL MMSE requires a matrix inversion of dimension $R \times R$. Furthermore, LOCAL PCA requires a singular value decomposition (SVD) of dimension $K_s \times K_s$. NONLINEAR requires two matrix inversions, a Cholesky-decomposition, and an Eigen-decomposition, each of dimension $W \times W$. The computational complexities of matrix inversion, SVD, Cholesky-decomposition, and Eigen-decomposition are provided in [34]. Together with the computational complexities of the required matrix multiplications, we obtain the computational complexity results of the considered forwarding schemes given in Table 1. Table 1 shows that LOCAL PCA entails a slightly higher computational complexity than LOCAL MMSE R. However, the simulation results in [15] reveal that LOCAL PCA also achieves a higher performance than LOCAL MMSE R. The comparison between the complexities of the linear and nonlinear forwarding scheme is governed by the number of N_L, which does not play a role for the complexity of the nonlinear forwarding scheme. In this case, however, for NONLINEAR the number of iterations steps T_i required for Algorithm 1, according to Table 1, is not negligible, such that a higher complexity already arises for more than 160 iterations. On the other hand, our simulation results in Section V show that NONLINEAR also achieves a significantly higher performance than LOCAL PCA. Thus, the schemes LOCAL MMSE R, LOCAL PCA, and NONLINEAR offer a trade-off between performance and complexity. Moreover, the simulation results in Section V show that for power constellations, where the RF transmit power is high in respect to the mmWave transmit power, the performance gain of NONLINEAR over LOCAL PCA is significant, which justifies the high complexity of NONLINEAR.

V. NUMERICAL RESULTS

In this section, we first introduce several firefly network topologies with one and multiple root nodes, respectively. For these proposed topologies, we investigate the performance of the proposed nonlinear forwarding strategy NONLINEAR. In addition, we also investigate the performance of LOCAL PCA which, for simplicity, we refer now to as ‘‘LINEAR‘‘ in the following [15]. Moreover, we compare the performance of NONLINEAR and LINEAR to the upper bound derived in Section III. We do not consider LOCAL MMSE R, since LOCAL PCA outperforms LOCAL MMSE R as was shown in [15].

A. TOPOLOGIES FOR THE FRONTHAUL LINK

In the following, different firefly network topologies are presented for the fronthaul links. In particular, we consider different variations of a Street Canyon Scenario as well as a Small Scale Street Canyon Scenario, cf. Figure 4. However, first, we introduce a layer structure which is useful to analyze the delays in the firefly network topologies.

1) LAYER STRUCTURE

The routing layer structure of the network provides insights regarding the overall communication delay introduced by the multi-hop transmission, and hence, it is useful for comparing different topologies. We first provide the formal definition of a layer. In particular, layer γ in the network is specified by a set $\mathcal{V}_\gamma$ which contains the indices of the FNs which belong to the layer.

Definition 1 (See [16]): An FN belongs to layer 1 if it does not receive any data from other FNs. An FN belongs to layer γ if it receives data from at least one FN in layer $\gamma - 1$.

Note that an FN may belong to more than one layer. In such a case, this FN introduces an additional delay since it can forward the signals to the FNs in the next layer only after it has received the signals from all FNs in the previous layers.

Layer γ in the network is specified by a set $\mathcal{V}_\gamma$ which contains the indices of the FNs which belong to the layer, see Section V-A1 and Definition 1.

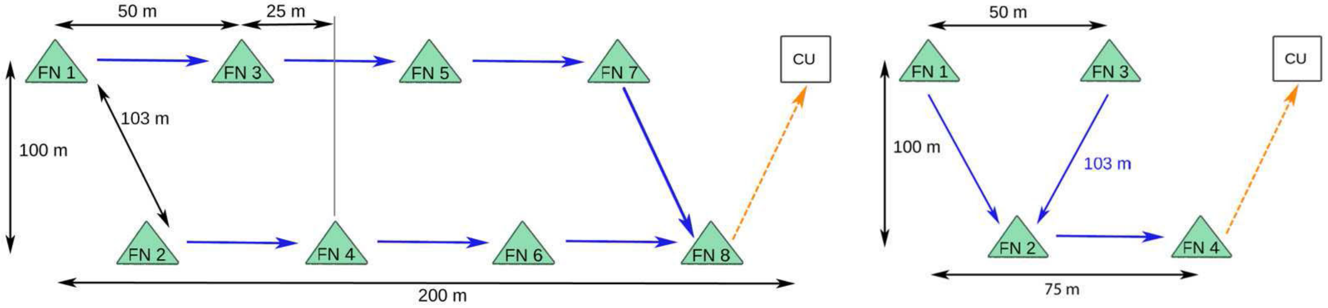

FIGURE 4. Street Canyon Scenario Variant 1 (left) and Small Scale Street Canyon Scenario (right).

TABLE 2. Data flow of MDs' messages received at FN 1 through the multi-hop structure of the Small Scale Street Canyon Scenario.

Link	Packets				
	1	2	3	4	...
MDs → FN1	W_1	W_2	W_3	W_4	...
FN1 → FN2, FN3 → FN2	-	W_1	W_2	W_3	...
FN2 → FN4	-	-	W_1	W_2	...
FN4 → CU	-	-	-	W_1	...

Moreover, we denote the layer which contains all available root nodes of a topology as layer r, i.e., $\mathcal{V}_r = \{v_1, \ldots, v_T\}$. In Table 2, the data flow of the MDs' messages received at FN1 through the multi-hop structure of the Small Scale Street Canyon Scenario in Figure 4 is illustrated. Table 2 reveals how many packet transmissions are needed for the MDs' messages transmitted in the i-th time slot, denoted as $W_i, \forall i \in \{1, 2, \ldots\}$, to arrive at the CU. For the Small Scale Street Canyon Scenario, there is a communication delay of 3 time steps, which explains also the 3 layers of this topology, given by $\mathcal{V}_1 = \{1, 3\}$, $\mathcal{V}_2 = \{2\}$, $\mathcal{V}_r = \mathcal{V}_3 = \{4\}$, see right hand side of Figure 5. Note that, in general, message W_i is also received at all other FNs in the considered topology. Hence, W_i is held at a given FN until the FN has received W_i from all the other FNs from which it receives data via an incoming link.

2) VARIATIONS OF STREET CANYON SCENARIO

The Street Canyon Scenario comprises 8 FNs, where 4 FNs are deployed on each side of the street, see Figure 4. In addition, for all variants of the Street Canyon Scenario, we assume a coverage area of 200 m × 100 m and the distance between two neighboring FNs on the same street side is equal to 50 m. As a consequence of the FN deployment, the link distance between two FNs on opposite sides of the street is 103 m following the Pythagorean theorem. Moreover, the positions of the 8 FNs are identical for all considered variants of the Street Canyon Scenario. The left hand side of Figure 4 shows the topology of Street Canyon Scenario Variant 1, where FN 8, which serves as a root node, is connected via an optical fiber link to the CU. Due to the large number of mmWave hops from the FNs to the single root node at the end of the communication chain, this topology has 5 layers, see left-hand side of Figure 5. Hence, this topology entails a large inherent delay. In Tables 3 and 4, the topology and the layer structure of the 5 considered Street Canyon Scenario Variants are given, respectively. Compared to Street Canyon Scenario Variant 1, Street Canyon Scenario Variant 2 uses FN 1 as a second root node in addition to FN 8. Therefore, FNs 2, 3, and 4 forward now their received signals to FN 1 instead of FN 8. Moreover, the links between FN 3 and FN 5 and between FN 4 and FN 5 have been removed to reduce the maximum number of hops from any FN to a root node to two. We note that for this variant, the root nodes are located at the edges of the area. One benefit of this topology compared to Street Canyon Scenario Variant 1 is that the corresponding layer structure contains only 3 layers, and thus, the inherent delay is reduced.

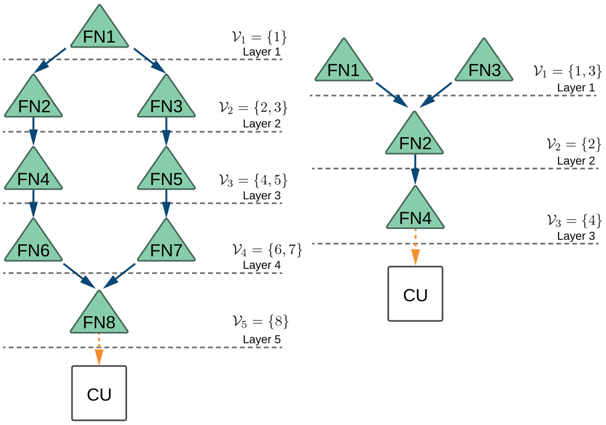

FIGURE 5. Layer structure of Street Canyon Scenario Variant 1 (left) and of Small Scale Street Canyon Scenario (right).

Street Canyon Scenario Variant 3 has also two root nodes. However, in order to be able to investigate the impact of the locations of the root nodes on the sum rate, for Street Canyon Scenario Variant 3, we choose the two FNs in the center of the considered area, i.e., FN 4 and FN 5, as root nodes. The resulting layer structure of Street Canyon Scenario Variant 3 has 3 layers, similar to Street Canyon Scenario Variant 2. For Street Canyon Scenario Variants 4 and 5, we increase

TABLE 3. Topology of Street Canyon Scenario Variants 1,2,3,4, and 5.

	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
FNs	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8	3,4,5,6
mmWave links	(1,2), (1,3), (2,4), (3,5), (4,6), (5,7), (6,8), (7,8)	(4,2), (2,1), (4,3), (3,1), (5,6), (6,8), (5,7), (7,8)	(1,2), (2,4), (1,3), (3,4), (8,6), (6,5), (8,7), (7,5)	(1,3), (2,4), (7,5), (8,6)	-
root nodes	8	1,8	4,5	3,4,5,6	3,4,5,6

TABLE 4. Layers of Street Canyon Scenario Variants 1,2,3,4, and 5.

	FNs in Layer 1	FNs in Layer 2	FNs in Layer 3	FNs in Layer 4	FNs in Layer 5
Variant 1	$\mathcal{V}_1 = \{1\}$	$\mathcal{V}_2 = \{2,3\}$	$\mathcal{V}_3 = \{4,5\}$	$\mathcal{V}_4 = \{6,7\}$	$\mathcal{V}_5 = \{8\}$
Variant 2	$\mathcal{V}_1 = \{4,5\}$	$\mathcal{V}_2 = \{2,3,6,7\}$	$\mathcal{V}_3 = \{1,8\}$	-	-
Variant 3	$\mathcal{V}_1 = \{1,8\}$	$\mathcal{V}_2 = \{2,3,6,7\}$	$\mathcal{V}_3 = \{4,5\}$	-	-
Variant 4	$\mathcal{V}_1 = \{1,2,7,8\}$	$\mathcal{V}_2 = \{3,4,5,6\}$	-	-	-
Variant 5	$\mathcal{V}_1 = \{3,4,5,6\}$	-	-	-	-

the number of root nodes to four. In Street Canyon Scenario Variant 4, the root nodes are the four FNs in the center of the area, i.e., FN 3, FN 4, FN 5, and FN 6. The other FNs forward their received signals to the nearest root node. The corresponding layer structure has only two layers. Street Canyon Scenario Variant 5 constitutes a centralized radio access network (C-RAN) configuration. Here, FNs 1, 2, 7, and 8 have been removed and only the four root nodes of Street Canyon Scenario Variant 4 remain. Note that this topology has no mmWave links at all, and hence, the transmission delay is further reduced compared to the topologies with mmWave links. This can be observed in the layer structure in Table 4 which has only one layer containing the four root nodes.

Overall, the number of distortion matrices or linear filter matrices (when applying linear forwarding as proposed in [15]) in the network is reduced when more FNs serve as root nodes. This is due to the fact that, at the root nodes, nonlinear or linear processing is not needed as the received signal can be forwarded directly to the CU over an optical fiber link with (virtually) unlimited capacity. This reduces the overall complexity of the system. The impact of the number of root nodes on the performance will be investigated in Section V-D.

3) SMALL SCALE STREET CANYON SCENARIO

Furthermore, to provide additional insights into performance, we also consider a reduced version of the Street Canyon Scenario with an area of size 75 m × 100 m, which we refer to as "Small Scale Street Canyon Scenario". The Small Scale Street Canyon Scenario and the corresponding layer structure are shown in the right hand sides of Figure 4 and Figure 5, respectively.

B. SIMULATION SETUP

For simplicity of presentation, we consider subcarrier clustering for the simulation results. Therefore, we split the RF subcarriers into N_{sub} subcarrier clusters. Each subcarrier cluster contains $N_{f,\text{sub}} = \frac{N_f}{N_{\text{sub}}}$ subcarriers. We assume that each MD is allocated to one subcarrier cluster and is only active on subcarriers which are included in this cluster. Thus, in each subcarrier cluster, $K_{\text{sub}} = \frac{K}{N_{\text{sub}}}$ MDs are active. Additionally, we assume that the MDs are uniformly distributed over the entire considered area. For the numerical results, we use the values of the parameters in Table 5 of the mmWave links and the RF links, respectively, unless stated otherwise. For the parameter values in Table 5, we obtain $N_L = 43$. Remember that N_L is the number of mmWave symbols per N-dimensional RF vector available for frequency and time multiplexing, cf. Section II. Furthermore, the Rician fading model used for the RF and mmWave link is applied as in [22] and for the average power gain based on the path-loss model, we employ the formula in [21, Equation (2)]. Note that, for the mmWave links, the average power gain contains a "natural influence" parameter β defined by $[\beta]_{\text{dB}} = -(\beta' + \beta'')d_{\text{FN}_{n,m}}$, as a multiplicative factor in linear scale [13], which models influences such as rain and atmospheric absorption and can be assumed to be constant over a long period of time. Here, β' and β'' are the rain and atmospheric absorptions in dB/m, respectively, and $d_{\text{FN}_{n,m}}$ denotes the distance between FN n and FN m. Moreover, for the RF links, $\sigma^2_{\text{RF}} = \frac{W^{\text{RF}}}{N_f} N_0^{\text{RF}} \cdot N_{\text{F}}^{\text{RF}}$, where N_0^{RF} denotes the RF noise power spectral density of the receiver, and N_{F}^{RF} is the RF receiver noise figure [21]. For the mmWave links, $\sigma^2_{\text{mmW}} = \frac{W^{\text{mmW}}}{N_\rho} N_0^{\text{mmW}} \cdot N_{\text{F}}^{\text{mmW}}$ [21]. Here, N_0^{mmW} is the mmWave noise power spectral density at the receiver, and $N_{\text{F}}^{\text{mmW}}$ is the mmWave receiver noise figure.

C. PERFORMANCE METRIC

For the locally-designed linear forwarding schemes in [15], the instantaneous achievable rate at the CU, denoted as R, is obtained as [15]

$$R = \sum_{s=1}^{N_f} W_{\text{sub}}^{\text{RF}} \log_2 |\mathbf{I}_{(N_L n_r^{\text{in}} + N n_r^{\text{FN}})} + \mathbf{P}_r^{x,s} \Sigma_{\mathbf{x}}^s \mathbf{P}_r^{x,s^{\text{H}}} \cdot (\mathbf{P}_r^{\text{RF},s} \Sigma_{\mathbf{z}^{\text{RF}}}^s \mathbf{P}_r^{\text{RF},s^{\text{H}}} + \mathbf{P}_r^{\text{mmW},s} \Sigma_{\mathbf{z}^{\text{mmW}}}^s \mathbf{P}_r^{\text{mmW},s^{\text{H}}})^{-1}|. \quad (34)$$

TABLE 5. Values of the parameters of the mmWave and RF links [13], [21].

Symbol	Definition	Value
	Fronthaul Link	
μ	Path-loss exponent	2
$d_{\text{ref}}^{\text{mmW}}$	Reference distance of the mmWave links	1 m
$G_{\text{FN}_{\text{mmW}}}^{\text{T}}$	MmWave transmit antenna gain at the FNs	23 dB
$G_{\text{FN}_{\text{mmW}}}^{\text{R}}$	MmWave receive antenna gain at the FNs	23 dB
β'	Oxygen absorption at 60 GHz band	0.020 dB/m
β''	Absorption for rain density 25 mm/hr at 60 GHz band	0.010 dB/m
N_0^{mmW}	MmWave noise power spectral density at the FNs	-174 dBm/Hz
$N_{\text{F}}^{\text{mmW}}$	MmWave receiver noise figure of FNs	10 dB
P_m^{mmW}	Transmit power of mth FN per link	21 dBm
f^{mmW}	Carrier frequency	60 GHz
N_ρ	Number of orthogonal subcarriers in mmWave link	710
$W_{\text{sub}}^{\text{mmW}}$	Bandwidth per subcarrier	6 MHz
W^{mmW}	Total bandwidth (2 channels of the 60 GHz band)	4.32 GHz
	RF Link	
κ	Path-loss exponent	3.5
$d_{\text{ref}}^{\text{RF}}$	Reference distance of the RF links	1 m
G_{FN}^{T}	RF transmit antenna gain at the MDs	0 dBi
G_{FN}^{R}	RF receive antenna gain at the FNs	8 dBi
N_0^{RF}	RF noise power spectral density at the FNs	-174 dBm/Hz
N_{F}^{RF}	RF receiver noise figure of FNs	10 dB
P_k^{RF}	Transmit power of kth MD	20 dBm
f^{RF}	Carrier frequency	3.5 GHz
K	Number of active single-antenna MDs	600
K_{sub}	Number of active single-antenna MDs in each subcarrier cluster	50
N	Number of RF antennas at the mth FN	12
N_f	Number of orthogonal subcarriers in RF links	6666
N_{sub}	Number of RF subcarrier clusters	12
$N_{f,\text{sub}}$	Number of orthogonal RF subcarriers in each subcarrier cluster	556
$W_{\text{sub}}^{\text{RF}}$	Bandwidth per subcarrier	15 kHz
W^{RF}	Total bandwidth	100 MHz
P_{LOS}	Probability of LOS RF links	80%

Here, $\mathbf{P}_r^{x,s}$, $\mathbf{P}_r^{\text{RF},s}$, and $\mathbf{P}_r^{\text{mmW},s}$ are the multi-hop channel matrices of subcarrier s [15] which affect the MD signal, the RF noise, and the mmWave noise in layer r, respectively. Furthermore, $\Sigma_{\mathbf{x}}^s$, $\Sigma_{\mathbf{z}^{\text{RF}}}^s$, and $\Sigma_{\mathbf{z}^{\text{mmW}}}^s$ are the covariance matrices of $\mathbf{x}$, $\mathbf{z}^{\text{RF}}$, and $\mathbf{z}^{\text{mmW}}$ on subcarrier s, respectively. For the proposed nonlinear forwarding scheme, R is calculated as

$$R = \sum_{s=1}^{N_f} W_{\text{sub}}^{\text{RF}} \log_2 |\mathbf{I}_{(|\varepsilon|+T)N} + \bar{\mathbf{H}}^s \Sigma_{\mathbf{x}}^s \bar{\mathbf{H}}^{s^{\text{H}}} (\Sigma_{\bar{\mathbf{z}}}^s + \Sigma_{\bar{\mathbf{n}}}^s)^{-1}|, \quad (35)$$

where $\bar{\mathbf{H}}^s$ contains the stacked RF channel matrices of subcarrier s received at the CU. Moreover, $\Sigma_{\mathbf{x}}^s$, $\Sigma_{\bar{\mathbf{z}}}^s$, and $\Sigma_{\bar{\mathbf{n}}}^s$ are the covariance matrices of $\mathbf{x}$ and the processes $\bar{\mathbf{z}}$ and $\bar{\mathbf{n}}$ received at the CU, cf. (16), on subcarrier s, respectively. Furthermore, remember that T is the number of root nodes in the considered topology. For our simulation results, we averaged the sum rates in (34) and (35) over 400 channel realizations.

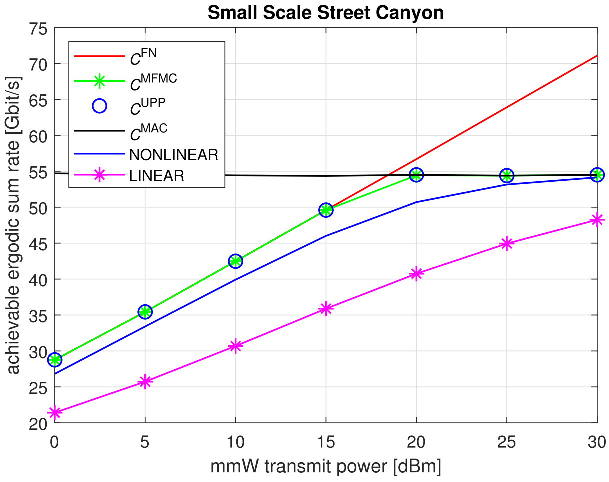

FIGURE 6. Ergodic sum rate vs. mmWave transmit power for $P_k^{\text{RF}} = 20$ dBm.

D. PERFORMANCE EVALUATION

In Figures 6 and 7, we show the achievable ergodic sum rates of the locally-designed nonlinear forwarding strategy NONLINEAR, the locally-designed linear forwarding strategy LINEAR, and the upper bounds from Section III for the Small Scale Street Canyon Scenario as functions of the mmWave transmit power and as functions of the RF transmit power, respectively.

Both figures show that NONLINEAR outperforms LINEAR in the considered power range. We observe that, for the adopted system parameters, there is a considerable gap

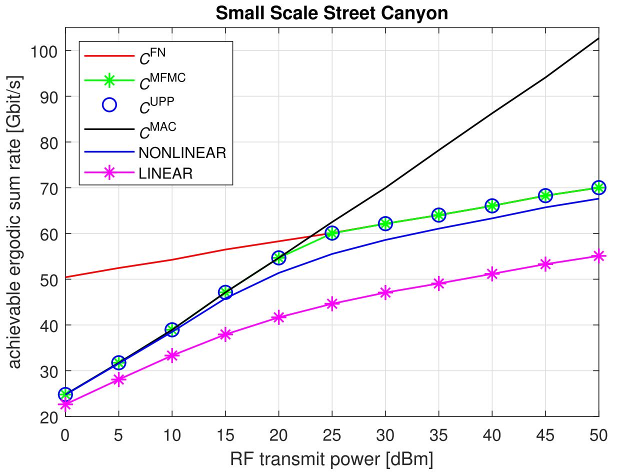

FIGURE 7. Ergodic sum rate vs. RF transmit power for $P_m^{\mathrm{mmW}} = 21$ dBm.

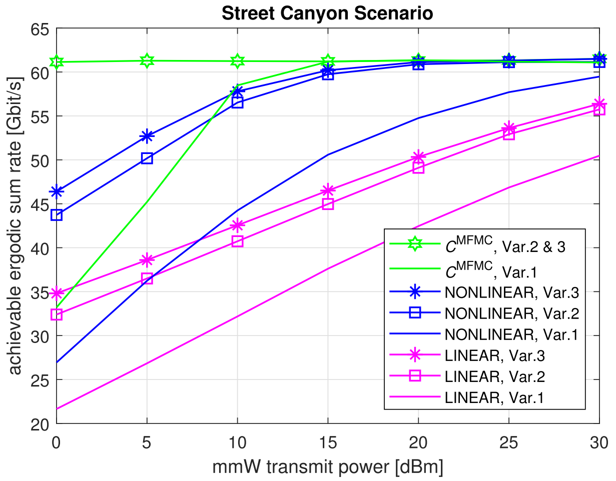

FIGURE 8. Ergodic sum rate vs. mmWave transmit power for $P_k^{\mathrm{RF}} = 20$ dBm.

between the sum rates achieved with LINEAR and the upper bounds. For example, in Figure 6, for a mmWave transmit power of 20 dBm, only around 75% of the upper bound can be achieved with LINEAR. In Figure 6, the sum rates of LINEAR and NONLINEAR increase with increasing mmWave transmit power, whereby for NONLINEAR and a mmWave transmit power exceeding 20 dBm, the sum rate approaches C^{MAC}. However, NONLINEAR achieves around 92% of the upper bound for a mmWave transmit power of 21 dBm.

In Figure 7, it can be observed that the sum rates of the locally-designed linear and nonlinear forwarding techniques increase with increasing RF transmit power. For RF transmit powers exceeding 25 dBm, the multi-hop FN network, particularly the mmWave link from FN 2 to FN 4 (due to the topology), becomes the performance bottleneck of the communication system. This can be inferred from the upper bounds since in this power region, $C^{\mathrm{UPP}} = C^{\mathrm{FN}}$. It can be observed that the slopes of the curves of the achievable rates of both the linear and the nonlinear forwarding schemes approach the slope of C^{FN} for RF transmit powers exceeding 25 dBm. In the entire considered RF transmit power range, NONLINEAR outperforms LINEAR and achieves around 93% of the upper bound for an RF power of 20 dBm. In fact, for RF transmit powers in the interval [0 dBm, 10 dBm], the performance of NONLINEAR is very close to the upper bound.

Figure 8 shows the achievable ergodic sum rate versus the mmWave transmit power for Street Canyon Scenario Variants 1, 2, and 3 for an RF transmit power of 20 dBm. In particular, in Figure 8, C^{MFMC} and the achievable sum rates for LINEAR and NONLINEAR are shown. Regarding the upper bound, the virtual MAC capacity, C^{MAC}, which is independent of the mmWave transmit power and identical to C^{MFMC} for high mmWave transmit powers, is not affected by the number of root nodes. Recall that, for C^{MAC}, all FNs and the CU are in the set of receivers $\mathcal{D}$. Hence, increasing the number of root nodes to two, while keeping the number of the FNs and their position the same, has no influence on the virtual MAC capacity. However, the firefly network capacity, C^{FN}, which is identical to C^{MFMC} for low mmWave transmit powers, increases if the number of root nodes is increased from one to two. C^{FN} is the capacity of the bottleneck cut, which is equal to the sum of the achievable rates of all RF access links from the MDs to the FNs in the set of receivers $\mathcal{D}$ plus the achievable rates of all mmWave links crossing the bottleneck cut. Moreover, in general, the bottleneck cut includes the mmWave links closest to the root nodes, i.e., the cut where all root nodes and the CU are in $\mathcal{D}$ and all remaining FNs as well as the MDs are in the set of transmitters $\mathcal{S}$. For Street Canyon Scenario Variants 2 and 3, we have two FNs in $\mathcal{D}$, i.e., the two root nodes, whereas for Variant 1 with one root node, there is only one FN in $\mathcal{D}$. This results in a higher C^{FN}, or more precisely, in a higher achievable rate for all RF access links from the MDs to the FNs in $\mathcal{D}$ for Variants 2 and 3, respectively, compared to Variant 1. In addition, in Variants 2 and 3, we have four mmWave links to the two available root nodes, whereas in Variant 1, we have only two mmWave links to the single root node.

Overall, there is a significant performance gain for Variants 2 and 3 with two root nodes compared to Variant 1 with one root node, respectively, since with the additional root node, there is one FN less, for which linear or nonlinear processing of the received data and forwarding over a mmWave link is needed. Hence, there is one FN more (the second root node), for which the received data can be simply collected and directly forwarded without any loss to the CU over the optical fiber link. Furthermore, due to the fewer mmWave multi-hop links, there is less accumulated quantization error which results in higher sum rates. In addition, for one root node in Variant 1, the performance of NONLINEAR is up to 25% worse than for two root nodes in Variants 2 and 3. However, NONLINEAR for Variant 1 still outperforms LINEAR for Variant 3 for mmWave transmit powers exceeding 8 dBm.

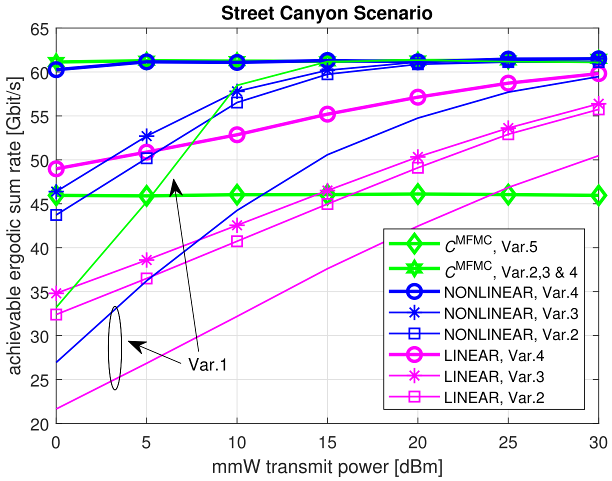

FIGURE 9. **Ergodic sum rate vs. mmWave transmit power for $P_k^{\mathrm{RF}} = 20$ dBm.**

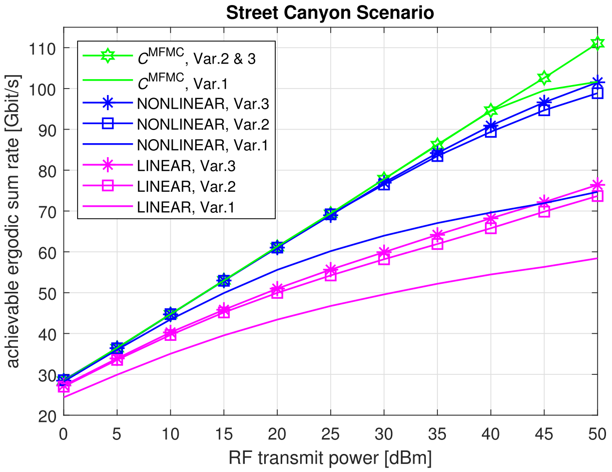

FIGURE 10. **Ergodic sum rate vs. RF transmit power for $P_m^{\mathrm{mmW}} = 21$ dBm.**

Moreover, we can observe that unlike for Variant 1, for Variants 2 and 3 the performance of NONLINEAR approaches the maximal ergodic sum rate according to C^{MFMC} for mmWave transmit powers exceeding 23 dBm. Thus, we can conclude that, in general, NONLINEAR is preferable since it outperforms LINEAR. However, for an RF transmit power lower than 8 dBm, Variant 3 with LINEAR is preferable over Variant 1 with NONLINEAR since it achieves higher rates, involves fewer hops, and employs a less complex forwarding strategy.

Comparing the results for Variants 2 and 3, we observe that Variant 3 outperforms Variant 2. This is due to the uniform distribution of the MDs over the considered area. On average the distance from the MDs to the FNs in the center of the area is smaller than that between the MDs and the FNs located at the edges. Thus, for Variant 3 with the root nodes in the center, the RF signals which can be directly forwarded to the CU by the root nodes have on average a better quality (i.e., a higher SNR due to reduced quantization errors) compared to Variant 2.

Figure 9 contains the same curves as Figure 8, but additionally contains the sum rate results for LINEAR, NONLINEAR, and C^{MFMC} for Street Canyon Scenario Variants 4 and 5. Here, a significant performance gain of Variant 4 over Variants 2 and 3, and a very large performance gain over Variant 1 can be observed. In addition, Variant 4 with NONLINEAR is very close to the maximal ergodic sum rate according to C^{MFMC} for mmWave transmit powers exceeding 5 dBm. Moreover, note that Variant 4 has only 2 layers, i.e., one packet delay is caused by this topology, see Table 4. For Variant 5 which represents a C-RAN system, only C^{MFMC} is shown, which is identical to the corresponding C^{MAC} since in Variant 5, mmWave links and forwarding are not needed. Note that, in Variant 5, all available FNs serve as root nodes, i.e., Variant 5 has only one layer and thus, causes zero delay, see Table 4. From this figure, it can be observed that Variant 4 outperforms Variant 5 since there are only four FNs in Variant 5. Recall that Variants 1, 2, and 3 employ all eight FNs, but Variant 1 has one root node, while Variants 2 and 3 both have two root nodes. Having more root nodes reduces the overall delay of the topology (Variant 1 has 5 Layers, Variants 2 and 3 have each 3 Layers), see Table 4. Nevertheless, applying LINEAR, Variants 1, 2, and 3 outperform Variant 5 only in the high mmWave transmit power regime. In particular, for $P_k^{\mathrm{RF}} = 20$ dBm and $P_m^{\mathrm{mmW}} = 21$ dBm, Variant 5 outperforms Variant 1 with LINEAR. For the same transmit powers, Variants 2 and 3 with LINEAR outperform Variant 5. Applying NONLINEAR, Variant 1 outperforms Variant 5 for mmWave transmit powers exceeding 12 dBm, Variant 2 shows the same behavior for mmWave transmit powers exceeding 2 dBm, and for Variant 3, NONLINEAR outperforms Variant 5 in the entire considered mmWave transmit power range.

In Figure 10, C^{MFMC} and the achievable ergodic sum rates for LINEAR and NONLINEAR are shown versus the RF transmit power for Street Canyon Scenario Variants 1, 2, and 3. From this figure, similar conclusions regarding the topologies with one and two root nodes can be drawn as from Figures 8 and 9. Again, the performance gain of Variants 2 and 3 over Variant 1 is large. Also, NONLINEAR yields a large performance gain compared to LINEAR. Especially for Variants 2 and 3 and RF transmit powers lower than 25 dBm, the performance of NONLINEAR is very close to C^{MFMC}. For RF transmit powers in the range [0 dBm, 45 dBm], Variant 1 with NONLINEAR outperforms LINEAR for all three Street Canyon Scenario Variants. For RF transmit powers exceeding 45 dBm, only LINEAR for Variant 3 outperforms NONLINEAR for Variant 1. In this figure, the performance upper bounds for the considered Street Canyon Scenario Variants almost coincide. This is because for the simulation parameters in Table 5, over almost the entire considered RF transmit power range, C^{FN} is larger than C^{MAC}. Hence, $C^{\mathrm{MFMC}} = C^{\mathrm{MAC}}$ for RF transmit powers

The delay of one layer and thus, only root nodes which forward their signals to the CU via fiber link is negligible.

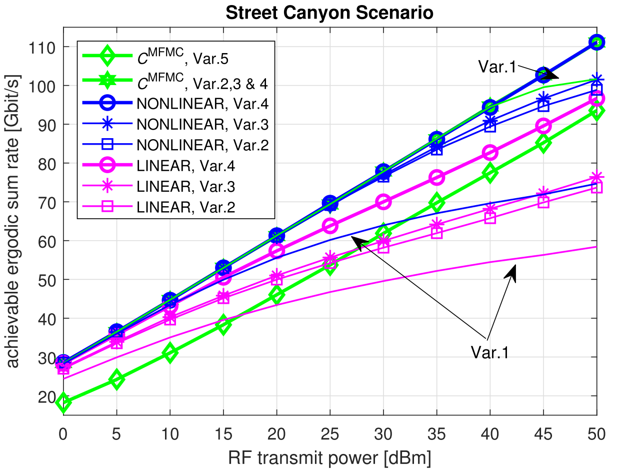

FIGURE 11. Ergodic sum rate vs. RF transmit power for $P_m^{\text{mmW}} = 21$ dBm.

in the range [0 dBm, 40 dBm]. Only for an RF transmit power exceeding 40 dBm, C^{MFMC} is smaller for Variant 1, which achieves a lower performance compared to Variants 2 and 3 since, in this transmit power regime, for Variant 1, $C^{\text{MFMC}} = C^{\text{FN}}$.

Figure 11 contains the same curves as Figure 10, but we have added the sum rates for LINEAR, NONLINEAR, and C^{MFMC} for Street Canyon Scenario Variants 4 and 5. Here, a significant performance gain of Variant 4 over Variants 2 and 3, and a very large performance gain over Variant 1 can be observed. In particular, Variant 4 with LINEAR outperforms Variant 1 with NONLINEAR for an RF transmit power larger than 15 dBm. Note that Variant 4 has 4 root nodes, whereas Variants 2 and 3 have two root nodes and Variant 1 has only one root node. Moreover, the performance of Variant 4 with NONLINEAR coincides with C^{MFMC} in the entire considered RF transmit power range.

As can be observed, over the entire considered RF transmit power range, Variant 4 outperforms Variant 5. This emphasizes that Variant 5 suffers from having only half the number of FNs compared to Variant 4, even if all FNs in Variant 5 are root nodes, i.e., Variant 5 is a C-RAN system. Moreover, in the entire considered RF transmit power range, Variants 2 and 3 applying NONLINEAR outperform Variant 5. In addition, the increased costs of having more root nodes incurred by the expensive optical fiber links from the root nodes to the CU makes C-RAN systems (as in Variant 5) less attractive than the proposed multi-hop structure.

Figures 8 and 10 reveal that increasing the number of root nodes increases the performance upper bound and the achievable sum rates of both forwarding strategies and reduces the performance gap between the upper bound and the linear/nonlinear forwarding strategies.

VI. CONCLUSION

We have studied a firefly ultra dense network where multiple MDs transmit their data over RF links to FNs and these FNs forward the received information to root nodes through multi-hop mmWave links. The root nodes forward the data further to a CU via optical fiber links. First, we derived an upper bound for the achievable sum rate of the considered network. Then, we investigated nonlinear processing strategies for forwarding the signals between the FNs. We adopted vector quantization at the FNs to efficiently exploit the correlation between signals received at different antennas for compression. Two different optimization problems for nonlinear processing were formulated, namely a central design and a local design strategy. Both proposed optimization problems are non-convex. Due to the high complexity of the central design strategy, we focused on the local design strategy. An efficient method for solving the optimization problem for the local design strategy was presented which can handle the non-convexity of the problem and resulted in an optimal solution, due to strong duality. Finally, we presented simulation results to quantify the performance of the proposed nonlinear forwarding strategy and a benchmark linear forwarding strategy for several topologies with different numbers of root nodes. Our layer structure analysis of the proposed topologies revealed that the communication delay can be reduced by increasing the number of root nodes. Our simulation results revealed that increasing the number of root nodes also improves the performance of both the linear and nonlinear forwarding strategies, whereby the performance of the nonlinear strategy approaches the proposed upper bound. However, there is a trade-off between performance and cost as root nodes have to be connected to the CU via expensive optical fiber links. This trade-off motivates the use of the proposed nonlinear forwarding strategy which unlike the linear forwarding strategy, is able to approach the performance upper bound even for a small number of root nodes. An interesting topic for future work is the design of practical quantization code books for the FNs. Moreover, the design of power and subcarrier allocation algorithms for firefly ultra dense networks requires further research.

APPENDIX A
PROOF OF THEOREM 1

The upper bound given in Theorem 1 is based on the cut-set bound [19]. A cut is defined for given sets $\mathcal{S}$ and $\mathcal{D}$ where the FNs in $\mathcal{S}$ and the MDs form the set of transmitting nodes and the FNs in $\mathcal{D}$ and the CU form the set of receiving nodes. For simplicity of presentation, we drop the subcarrier index s for the subsequent derivations. Let $\mathbf{x}_{\text{MD}}$, $\mathbf{x}_{\mathcal{A}}^{\text{mmW}}$, and $\mathbf{x}_{\mathcal{N}}^{\text{Fib}}$ denote the vectors of symbols transmitted by the MDs, the FNs in a given set $\mathcal{A}$, and the root nodes in set $\mathcal{N}$, respectively. Similarly, let $\mathbf{y}_{\text{CU}}^{\text{Fib}}$, $\mathbf{y}_{\mathcal{A}}^{\text{RF}}$, and $\mathbf{y}_{\mathcal{A}}^{\text{mmW}}$ denote the received symbols at the CU, the RF antennas of the FNs in set $\mathcal{A}$, and the mmWave receivers of the FNs in set $\mathcal{A}$, respectively. Let $I(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D})$ be the conditional mutual information between the transmitted symbols of $\mathcal{S}$ and the received symbols in $\mathcal{D}$ under the condition that the transmit symbols of the FNs in $\mathcal{D}$ are known. Thereby, for given $\mathbf{H}$ and $\mathbf{G}$, the capacity of the cut specified by $\mathcal{S}$ and $\mathcal{D}$ is given by (36) subject to feasible probability distributions satisfying the power constraints. Here, $f(\mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}})$ and $f(\mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}, \mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}})$ are the joint probability density functions of $\mathbf{x}_{\mathrm{MD}}$, $\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}$ and $\mathbf{x}_{\mathrm{MD}}$, $\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}$, $\mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}}$, respectively. Thereby, based on the max-flow min-cut theorem [26], [27], the achievable sum rate of the MDs to the CU is upper bounded by

$$\sum_{k=1}^{K} R_k(\mathbf{H}, \mathbf{G}) \leq \min_{\forall \mathcal{S} \subset \mathcal{M}} I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right), \tag{37}$$

where $R_k(\mathbf{H}, \mathbf{G})$ is the achievable rate from the k-th MD to the CU.

In the following, we simplify $I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right)$ for the firefly network introduced in Section II. First, we assume the capacity of the optical fiber links between FNs $\nu_1, \ldots, \nu_T$ and the CU to be infinite. Hence, for all cuts where $\mathcal{N} \not\subseteq \mathcal{D}$, the cut capacity $I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right)$ is infinite. Therefore, these cuts cannot be the bottleneck cut, i.e., the cut with the minimum capacity $I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right)$, and hence can be removed. Thus, in (36), as shown at the bottom of the page, we have to consider only the cuts $\mathcal{N} \subseteq \mathcal{D}$. For all $\mathcal{N} \subseteq \mathcal{D}$, we can simplify and upper bound $I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right)$ as follows

$$\begin{aligned} I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right) &\overset{(a)}{=} \max_{f(\mathbf{x}_{\mathrm{MD}})} I(\mathbf{x}_{\mathrm{MD}}; \mathbf{y}_{\mathcal{D}}^{\mathrm{RF}}) \\ &\quad + \max_{f(\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}})} I(\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}; \mathbf{y}_{\mathcal{D}}^{\mathrm{mmW}} | \mathbf{x}_{\mathcal{D}}^{\mathrm{mmW}}) \\ &\overset{(b)}{\leq} C_{\mathcal{S}}^{\mathrm{MAC}}\left(\mathbf{H}\right) + C_{\mathcal{S}}^{\mathrm{mmW}}\left(\mathbf{G}\right), \end{aligned} \tag{38}$$

subject to feasible probability distributions satisfying the power constraints, where $f(\mathbf{x}_{\mathrm{MD}})$ and $f(\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}})$ are the probability density functions of $\mathbf{x}_{\mathrm{MD}}$ and of $\mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}$, respectively. Here, equality (a) follows from the fact that the RF and mmWave links are independent parallel channels and inequality (b) follows from the assumption that the FNs in $\mathcal{D}$ can jointly decode their received signal where $C_{\mathcal{S}}^{\mathrm{MAC}}\left(\mathbf{H}\right)$ is given in (6). In particular, $C_{\mathcal{S}}^{\mathrm{MAC}}\left(\mathbf{H}\right)$ is the capacity of the MIMO MAC channel with channel matrices $[\mathbf{h}_{1,\mathcal{S}}, \mathbf{h}_{2,\mathcal{S}}, \ldots, \mathbf{h}_{K_s,\mathcal{S}}]$, cf. Section III, [28]. Note that MDs $1, \ldots, K_s$ are active on subcarrier s. Due to the directional mmWave transmission between the FNs, the mmWave links between the FNs are independent and hence, $C_{\mathcal{S}}^{\mathrm{mmW}}\left(\mathbf{G}\right)$ given in (7), is the summation of all SISO capacities of all available mmWave links from the FNs in $\mathcal{S}$ to the FNs in $\mathcal{D}$.

Inequality (37) together with inequality (38) results in the inequality

$$\sum_{k=1}^{K} R_k(\mathbf{H}, \mathbf{G}) \leq \min_{\forall \mathcal{S} \subset \mathcal{M} \setminus \mathcal{N}} \left(C_{\mathcal{S}}^{\mathrm{MAC}}\left(\mathbf{H}\right) + C_{\mathcal{S}}^{\mathrm{mmW}}\left(\mathbf{G}\right)\right), \tag{39}$$

for any channel coefficient matrix $\mathbf{H}$ and $\mathbf{G}$. A corresponding inequality holds also for the achievable ergodic sum rate:

$$\begin{aligned} \sum_{k=1}^{K} \bar{R}_k &= \mathbb{E}_{\mathbf{H},\mathbf{G}} \left\{ \sum_{k=1}^{K} R_k(\mathbf{H}, \mathbf{G}) \right\} \\ &\leq \mathbb{E}_{\mathbf{H},\mathbf{G}} \left\{ \min_{\forall \mathcal{S} \subset \mathcal{M} \setminus \mathcal{N}} \left(C_{\mathcal{S}}^{\mathrm{MAC}}(\mathbf{H}) + C_{\mathcal{S}}^{\mathrm{mmW}}(\mathbf{G}) \right) \right\}, \end{aligned} \tag{40}$$

and thus, Theorem 1 is proved.

APPENDIX B
PROOF OF COROLLARY 1

Let $\{X_i\}_{i \in \mathcal{I}}$, $\mathcal{I} \subseteq \mathbb{N}$, be a set of random variables. Then, the following well known Jensen's inequality holds [35]

$$\mathbb{E}\left\{ \min_i X_i \right\} \leq \min_i \mathbb{E}\left\{X_i\right\}, \tag{41}$$

where we exploited the concavity of the minimum function, and thus,

$$\begin{aligned} &\mathbb{E}_{\mathbf{H},\mathbf{G}} \left\{ \min_{\forall \mathcal{S} \subset \mathcal{M} \setminus \mathcal{N}} C_{\mathcal{S}}^{\mathrm{MAC}}(\mathbf{H}) + C_{\mathcal{S}}^{\mathrm{mmW}}(\mathbf{G}) \right\} \\ &\leq \min_{\forall \mathcal{S} \subset \mathcal{M} \setminus \mathcal{N}} \left(\mathbb{E}_{\mathbf{H}} \left\{ C_{\mathcal{S}}^{\mathrm{MAC}}(\mathbf{H}) \right\} + \mathbb{E}_{\mathbf{G}} \left\{ C_{\mathcal{S}}^{\mathrm{mmW}}(\mathbf{G}) \right\} \right) \end{aligned} \tag{42}$$

Hence, Corollary 1 is simply a consequence of Theorem 1 and the Jensen's inequality.

APPENDIX C
PROOF OF PROPOSITION 1

We employ the following change of variables [32]:

$$c_i = \log_2 \left(1 + \frac{\lambda_i}{\Sigma_{\hat{\mathbf{Q}}}^{(i,i)}} \right), \; i = 1, \ldots, (|\mathcal{E}_m| + 1)N, \tag{43}$$

where $\Sigma_{\hat{\mathbf{Q}}}^{(i,i)}$ is the i-th diagonal element of $\Sigma_{\hat{\mathbf{Q}}}$. Note that $\Sigma_{\hat{\mathbf{Q}}} \geq 0$ and $c_i \geq 0$. Equation (43) can be reformulated as

$$\frac{1}{\Sigma_{\hat{\mathbf{Q}}}^{(i,i)}} = \frac{2^{c_i}}{\lambda_i} - \frac{1}{\lambda_i}. \tag{44}$$

Now, using (43) and (44), we obtain from (33)

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{(|\mathcal{E}_m|+1)N} \left((1-\mu)c_i - \log_2 \left(\frac{2^{c_i}}{\lambda_i} - \frac{1}{\lambda_i} + 1 \right) \right) \\ &\quad + \mu C_{m,n}^{\mathrm{mmW}} / f_q \\ &= \sum_{i=1}^{(|\mathcal{E}_m|+1)N} \left((1-\mu)c_i - \log_2(2^{c_i} + \lambda_i - 1) \right) \end{aligned}$$

$$I\left(\mathbf{H}, \mathbf{G}, \mathcal{S}|\mathcal{D}\right) = \begin{cases} \max\limits_{f(\mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}})} I(\mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}}, \mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}; \mathbf{y}_{\mathrm{CU}}^{\mathrm{Fib}}, \mathbf{y}_{\mathcal{D}}^{\mathrm{RF}}, \mathbf{y}_{\mathcal{D}}^{\mathrm{mmW}} | \mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{mmW}}), & \mathcal{N} \subseteq \mathcal{D} \\ \max\limits_{f(\mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}, \mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}})} I(\mathbf{x}_{\mathcal{N}}^{\mathrm{Fib}}, \mathbf{x}_{\mathrm{MD}}, \mathbf{x}_{\mathcal{S}}^{\mathrm{mmW}}; \mathbf{y}_{\mathrm{CU}}^{\mathrm{Fib}}, \mathbf{y}_{\mathcal{D}}^{\mathrm{RF}}, \mathbf{y}_{\mathcal{D}}^{\mathrm{mmW}} | \mathbf{x}_{\mathcal{D}}^{\mathrm{mmW}}), & \mathcal{N} \not\subseteq \mathcal{D}, \end{cases} \tag{36}$$

$$-\log_2\left(\frac{1}{\lambda_i}\right) + \mu C^{\mathrm{mmW}}_{m,n}/f_q. \tag{45}$$

Setting the derivative of (45) with respect to c_i to zero yields

$$1 - \mu - \frac{1}{1 - \frac{1}{2^{c_i}} + \frac{\lambda_i}{2^{c_i}}} = 0 \tag{46}$$

Hence, by solving (46) with respect to c_i, the optimal c_i is given by

$$c_i^* = \left[\log_2\left(\frac{(1-\mu)(\lambda_i - 1)}{\mu}\right)\right]^+. \tag{47}$$

Consequently, the optimal $\Sigma_{\hat{\mathbf{Q}}}^{(i,i)}$ is given by

$$\Sigma_{\hat{\mathbf{Q}}}^{(i,i),*} = \begin{cases} \dfrac{\mu}{1 - \frac{1}{\lambda_i} - \mu}, & \mu < 1 - \dfrac{1}{\lambda_i} \\ +\infty, & \mu \geq 1 - \dfrac{1}{\lambda_i} \end{cases} \tag{48}$$

and the optimal $\mathbf{Q}_{(m,n)}$ is given by

$$\mathbf{Q}^*_{(m,n)} = (\mathbf{C}^{-1})^{\mathrm{H}} \Sigma^*_{\hat{\mathbf{Q}}} \mathbf{C}^{-1}. \tag{49}$$

Equation (48) implies that $\Sigma_{\hat{\mathbf{Q}}}^{(i,i)} \geq 0$. Moreover, $\mathbf{C} = \mathbf{R}^{\mathrm{H}}\mathbf{V}$, see Lemma 1, and thus, $\mathbf{Q}_{(m,n)} = (\mathbf{C}^{-1})^{\mathrm{H}}\Sigma_{\hat{\mathbf{Q}}}\mathbf{C}^{-1}$ is positive semidefinite. Hence, $\mathbf{Q}_{(m,n)} \succeq 0$ is already implicit with (48) and Lemma 1. Furthermore, from (48) and (49), we observe that with decreasing μ, where $\mu < 1 - \frac{1}{\lambda_i}$, also $\Sigma_{\hat{\mathbf{Q}}}^{(i,i),*}$ decreases and hence, $|\Sigma^*_{\hat{\mathbf{Q}}}|$ and finally $|\mathbf{Q}^*_{(m,n)}|$ decrease. Thus, for decreasing μ, $F_c\left(\mathbf{Q}^*_{(m,n)}\right)$ in (26) is a monotonically increasing function. Moreover, since the dual function is a convex function in μ, we can apply bisection to find the optimal μ such that (29) is fulfilled. Note that to avoid the second case of (48), the bisection approach, applied in line 7 of Algorithm 1, for finding the optimal μ^* can be restricted to $\mu \in \left(0, \min\limits_i(1 - \frac{1}{\lambda_i})\right]$.

APPENDIX D
PROOF OF THEOREM 2

To prove Theorem 2, we first present the solution, which is given in Algorithm 1. Subsequently, we show that this solution is both primal and dual optimal. Thus, the duality gap is zero.

A. PROPOSED SOLUTION

Let $\mu^* \in \left(0, \min\limits_i(1 - \frac{1}{\lambda_i})\right]$, $\forall i = 1, \ldots, (|\mathcal{E}_m| + 1)N$, for which the optimal $\mathbf{Q}^*_{(m,n)} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times (|\mathcal{E}_m|+1)N}$ of the optimization problem

$$\max_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu^*) \tag{50}$$

satisfies the following equality

$$F_c\left(\mathbf{Q}_{(m,n)}\right) - C^{\mathrm{mmW}}_{m,n}/f_q = 0. \tag{51}$$

Thus, $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ is the optimal solution of Algorithm 1. Note that $F_c\left(\mathbf{Q}^*_{(m,n)}(\mu)\right)$ in (26) is continuous monotonically increasing for decreasing μ, where $\left(\mathbf{Q}^*_{(m,n)}(\mu)\right)$ is the optimal solution of $\max\limits_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu)$ for a given μ (see Appendix C). Moreover, note that in general $F_c\left(\mathbf{Q}_{(m,n)}(0)\right) = \infty$ and $F_c\left(\mathbf{Q}_{(m,n)}(1)\right) = 0$. Therefore, $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ exists.

B. PRIMAL OPTIMAL SOLUTION

To show that the solution $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ is the optimal solution of the primal problem in (22), we need to show that there exists no other feasible $\tilde{\mathbf{Q}}_{(m,n)}$ for which $F_o\left(\tilde{\mathbf{Q}}_{(m,n)}\right) > F_o\left(\mathbf{Q}^*_{(m,n)}\right)$. We show this via the method of contradiction. Let us assume there is a $\tilde{\mathbf{Q}}_{(m,n)} \in \mathbb{C}^{(|\mathcal{E}_m|+1)N \times (|\mathcal{E}_m|+1)N}$ with $F_c\left(\tilde{\mathbf{Q}}_{(m,n)}\right) \leq C^{\mathrm{mmW}}_{m,n}/f_q$ which maximizes the objective function of the original problem such that

$$F_o\left(\tilde{\mathbf{Q}}_{(m,n)}\right) > F_o\left(\mathbf{Q}^*_{(m,n)}\right) = \mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*). \tag{52}$$

In this case, the following relation is valid

$$\begin{aligned} F_o\left(\tilde{\mathbf{Q}}_{(m,n)}\right) &\leq F_o\left(\tilde{\mathbf{Q}}_{(m,n)}\right) - \mu^*\left(F_c\left(\tilde{\mathbf{Q}}_{(m,n)}\right) - C^{\mathrm{mmW}}_{m,n}/f_q\right) \\ &= \mathcal{L}(\tilde{\mathbf{Q}}_{(m,n)}, \mu^*) \\ &\overset{(a)}{\leq} \mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*) \\ &= F_o\left(\mathbf{Q}^*_{(m,n)}\right), \end{aligned} \tag{53}$$

where (a) is due to $\mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*) = \max\limits_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu^*)$. Equation (53) contradicts (52), and thus, $\mathbf{Q}^*_{(m,n)}$, which fulfills $F_c\left(\mathbf{Q}_{(m,n)}\right) - C^{\mathrm{mmW}}_{m,n}/f_q = 0$, is the optimal solution of the primal problem in (22).

C. DUAL OPTIMAL SOLUTION

Now, to show that the solution $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ is dual optimal, we need to show that

$$\min_{\mu} \max_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu) = \mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*). \tag{54}$$

Since $\max\limits_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \mu)$ appears on both sides of (54), this equation does not hold only if there exists a $\hat{\mu}$ such that $\max\limits_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \hat{\mu}) < \mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*)$. Again, we show this cannot hold via contradiction. Let us assume there is a better solution than $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ for the dual problem (28), i.e., $\exists \hat{\mu} \in \left[0, \min\limits_i(1 - \frac{1}{\lambda_i})\right)$, $\forall i = 1, \ldots, (|\mathcal{E}_m| + 1)N$, $\hat{\mu} \neq \mu^*$, such that

$$\max_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \hat{\mu}) < \mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*). \tag{55}$$

Let $F_o\left(\mathbf{Q}^*_{(m,n)}\right)$ be the optimal solution of the primal problem, i.e., $\mathcal{L}(\mathbf{Q}^*_{(m,n)}, \mu^*) = F_o\left(\mathbf{Q}^*_{(m,n)}\right)$. Then, with (55) the following inequality is valid

$$\max_{\mathbf{Q}_{(m,n)}} \mathcal{L}(\mathbf{Q}_{(m,n)}, \hat{\mu}) < F_o\left(\mathbf{Q}^*_{(m,n)}\right). \tag{56}$$

Relation (56) contradicts the weak duality

$$\min_{\mu \geq 0} \mathcal{D}(\mu) \geq F_o\left(\mathbf{Q}^*_{(m,n)}\right). \tag{57}$$

Thus, $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ is the optimal solution of the dual problem. Hence, we can conclude that equality holds for the solutions of the primal problem in (22) and the dual problem in (28), and thus, strong duality holds, which proves Theorem 2.

At this point, it is worth mentioning that the statement of the proof meets the geometric interpretation of strong duality as given by Boyd and Vandenberghe (see [30], Chapter 5.3). The existence of $(\mathbf{Q}^*_{(m,n)}, \mu^*)$ with (51) is equivalent to the existence of a supporting hyperplane at $\mathcal{G}$ through the point $\left(0, F_o\left(\mathbf{Q}^*_{(m,n)}\right)\right)$, where $\mathcal{G}$ is the set of all values taken on by the constraint and the objective function [36]. Moreover, $F_c\left(\mathbf{Q}^*_{(m,n)}(0)\right) = \infty$ and $F_c\left(\mathbf{Q}^*_{(m,n)}(1)\right) = 0$ ensure that the hyperplane is non-vertical. According to [30], the existence of this hyperplane ensures the strong duality.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The digitization of the world from edge to core," Seagate, Cupertino, CA, USA, IDC White Paper US44413318, Nov. 2018.

[2] C. Galiotto, N. K. Pratas, N. Marchetti, and L. Doyle, "Effect of LOS/NLOS propagation on ultra-dense networks," in *Proc. IEEE Globecom*, Dec. 2014, pp. 3471–3476.

[3] S. Song, H. Li, Y. Fan, W. Kong, and W. Zhang, "Downlink interference rejection in ultra dense network," in *Proc. 10th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jul. 2018, pp. 361–364.

[4] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[5] M. A. Hasabelnaby, H. A. I. Selmy, and M. I. Dessouky, "C-RAN availability improvement using parallel hybrid FSO/mmW 5G fronthaul network," in *Proc. Int. Jpn.-Afr. Conf. Electron., Commun. Comput. (JAC-ECC)*, Dec. 2018, pp. 130–133.

[6] F. Tonini, C. Raffaelli, L. Wosinska, and P. Monti, "Cost-optimal deployment of a C-RAN with hybrid fiber/FSO fronthaul," *J. Opt. Commun. Netw.*, vol. 11, no. 7, pp. 397–408, Jul. 2019.

[7] P. Liu, H. Gao, W. Luo, and W. Jiang, "On secrecy performance of relay assisted millimeter wave C-RAN," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[8] Y. Yao, H. Tian, G. Nie, H. Wu, and J. Jin, "Multi-path routing based QoS-aware fairness backhaul-access scheduling in mmWave UDN," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.

[9] W. Pu, X. Li, J. Yuan, and X. Yang, "Traffic-oriented resource allocation for mmWave multi-hop backhaul networks," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2330–2333, Nov. 2018.

[10] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[11] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[12] S. K. Yong, P. Xia, and A. Valdes-Garcia, *60 GHz Technology for Gbps WLAN and WPAN: From Theory to Practice*. Hoboken, NJ, USA: Wiley, 2010.

[13] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[14] M. H. Tariq, I. Chondroulis, P. Skartsilas, N. Babu, and C. B. Papadias, "mmWave massive MIMO channel measurements for fixed wireless and smart city applications," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–6.

[15] K. Ackermann, V. Jamali, W. Gerstacker, F. Wartenberg, J. Aulin, R. Krishnan, and R. Schober, "Firefly ultra dense networks with mmWave fronthaul links," in *Proc. IEEE Globecom*, Dec. 2018, pp. 1–7.

[16] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Multihop backhaul compression for the uplink of cloud radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 2704–2708.

[17] P. K. Shah, R. P. Pandey, and R. Kumar, "Vector quantization with codebook and index compression," in *Proc. Int. Conf. Syst. Modeling Adv. Res. Trends (SMART)*, Nov. 2016, pp. 49–52.

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

[19] J. A. Thomas and T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[20] S. A. Ayoughi and W. Yu, "Optimized MIMO transmission and compression for interference mitigation with cooperative relay," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 4321–4326.

[21] V. Jamali, D. S. Michalopoulos, M. Uysal, and R. Schober, "Link allocation for multiuser systems with hybrid RF/FSO backhaul: Delay-limited and delay-tolerant designs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3281–3295, May 2016.

[22] S. Jin, W. Tan, M. Matthaiou, J. Wang, and K.-K. Wong, "Statistical eigenmode transmission for the MU-MIMO downlink in Rician fading," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6650–6663, Dec. 2015.

[23] X. Li, S. Jin, H. A. Suraweera, J. Hou, and X. Gao, "Statistical 3-D beamforming for large-scale MIMO downlink systems over Rician fading channels," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1529–1543, Apr. 2016.

[24] M. K. Samimi, G. R. MacCartney, S. Sun, and T. S. Rappaport, "28 GHz millimeter-wave ultrawideband small-scale fading models in wireless channels," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–6.

[25] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. S. Shitz, "Cooperative wireless cellular systems: An information-theoretic view," *Found. Trends Commun. Inf. Theory*, vol. 8, nos. 1–2, pp. 1–177, 2011.

[26] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[27] R. W. Yeung, *Information Theory and Network Coding*. Springer, Jan. 2008.

[28] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[29] M. Najafi, V. Jamali, D. W. K. Ng, and R. Schober, "C-RAN with hybrid RF/FSO fronthaul links: Joint optimization of fronthaul compression and RF time allocation," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8678–8695, Dec. 2019.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[31] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[32] S. Simoens, O. Munoz-Medina, J. Vidal, and A. D. Coso, "Compress-and-forward cooperative MIMO relaying with full channel state information," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 781–791, Feb. 2010.

[33] A. D. Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4698–4709, Sep. 2009.

[34] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[35] S. Loyka and A. Kouki, "On the use of Jensen's inequality for MIMO channel capacity estimation," in *Proc. Can. Conf. Electr. Comput. Eng.*, vol. 1, May 2001, pp. 475–480.

KATHARINA ACKERMANN (Member, IEEE) received the B.Sc. degree in mathematics and the M.Sc. degree in communications and multimedia engineering from Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany, in 2012 and 2015, respectively. Since 2015, she has been a Research Assistant with the Institute for Digital Communications, FAU. Her current research interests include wireless communications, ultra dense networks, mmWave communication, C-RAN, convex optimization, and MIMO systems.

VAHID JAMALI (Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in electrical engineering from the K. N. Toosi University of Technology, Tehran, Iran, in 2010 and 2012, respectively, and the Ph.D. degree (Hons.) from Friedrich-Alexander-University (FAU) of Erlangen-Nürnberg, Erlangen, Germany, in 2019. He was a Visiting Research Scholar with Stanford University, CA, USA, in 2017. He is currently a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, Princeton University. His research interests include wireless and molecular communications, Bayesian inference and learning, and multiuser information theory. He has served as a member of the Technical Program Committee for several IEEE conferences. He received several awards for his publications and research work, including the Best Paper Awards from the IEEE International Conference on Communications in 2016, the ACM International Conference on Nanoscale Computing and Communication in 2019, the Asilomar Conference on Signals, Systems, and Computers in 2020, and the IEEE Wireless Communications and Networking Conference in 2021; the Doctoral Research Grant from the German Academic Exchange Service (DAAD) in 2017; the Goldener Igel Publication Award from the Telecommunications Laboratory (LNT), FAU, in 2018; the Best Ph.D. Thesis Presentation Award from the IEEE Wireless Communications and Networking Conference in 2018; the Best Journal Paper Award (Literaturpreis) from the German Information Technology Society (ITG) in 2020; and the Postdoctoral Research Fellowship by the German Research Foundation (DFG) in 2020. He is also an Associate Editor of the IEEE Communications Letters, IEEE Open Journal of the Communications Society, and *Physical Communication* journal (Elsevier).

WOLFGANG GERSTACKER (Senior Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and the Dr.-Ing. and Habilitation degrees from Friedrich-Alexander University (FAU) Erlangen-Nürnberg, Erlangen, Germany, in 1991, 1998, and 2004, respectively. Since 2002, he has been with the Chair of Mobile Communications and Institute for Digital Communications, FAU Erlangen-Nürnberg, where he is currently a Professor. He has conducted various projects with partners from industry. His research interests include the broad areas of digital communications and statistical signal processing, THz communications, 5G and beyond, and wireless sensor networks. He is a member of the Executive Editorial Committee of IEEE Transactions on Wireless Communications. He was a recipient of several awards, including the Research Award of the German Society for Information Technology (ITG) in 2001, the EEEfCOM Innovation Award in 2003, the Vodafone Innovation Award in 2004, the Best Paper Award of EURASIP Signal Processing in 2006, and the ''Mobile Satellite & Positioning'' TrackPaper Award of VTC2011-Spring. He has been the Technical Program Co-Chair or General Co-Chair of several conferences, including BlackSeaCom 2014, VTC2013-Fall, ACMNanoCom 2016, and BalkanCom 2019. He is an Editor of *Computer Networks* (Elsevier). He has served as a member of the Editorial Board for IEEE Transactions on Wireless Communications, *Physical Communication* (Elsevier), and *EURASIP Journal on Wireless Communications and Networking*, and as a guest editor for several journals and magazines.

JOCELYN AULIN (Member, IEEE) received the Ph.D. degree in electrical engineering from Queen's University, Kingston, ON, Canada, in 2001. From 2001 to 2008, she was an Assistant Professor with the Department of Information and Computer Science, Telecommunication Theory Group, Chalmers University of Technology, Gothenburg, Sweden. From 2008 to 2013, she was with Ericsson AB, Gothenburg, working on 3GPP RAN4 standardization and base station algorithms with the Baseband Performance Group. She joined Sweden Research Center of Huawei Technologies Sweden AB, Gothenburg, in May 2013. Her research interests include security and privacy, WWW, NDSS, RAID, ESORICS, and DSN. More details can be found at: http://www.ccs.neu.edu/home/ek.

ROBERT SCHOBER (Fellow, IEEE) received the Diploma (Univ.) and Ph.D. degrees in electrical engineering from Friedrich-Alexander University of Erlangen-Nürnberg (FAU), Germany, in 1997 and 2000, respectively. From 2002 to 2011, he was a Professor and Canada Research Chair at The University of British Columbia (UBC), Vancouver, BC, Canada. Since January 2012, he has been an Alexander von Humboldt Professor and the Chair for Digital Communication at FAU. His research interests include communication theory, wireless communications, and statistical signal processing. He is a fellow of the Canadian Academy of Engineering, a fellow of the Engineering Institute of Canada, and a member of the German National Academy of Science and Engineering. He received several awards for his work, including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E.W.R. Stacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. Since 2017, he has been listed as a Highly Cited Researcher by the Web of Science. From 2012 to 2015, he served as the Editor-in-Chief of the IEEE Transactions on Communications. He also works as a member of the Editorial Board of the Proceedings of the IEEE and as a VP Publications for the IEEE Communication Society (ComSoc).