NOMA-Aided Multi-Way Massive MIMO Relay Networks

Shashindra Silva, Gayan Amarasuriya*, Masoud Ardakani and Chintha Tellambura
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6G 2V4
Email: {jayamuni, ardakani, chintha}@ece.ualberta.ca
*Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, USA 62901
Email:gayan.baduge@siu.edu

***Abstract*—We propose a novel transmission protocol for multi-way relay networks (MWRNs) in which the number of time-slots required for full mutual multi-way data exchange among K user nodes can be reduced to just two from $\lceil (K-1)/2 \rceil + 1$ in the current state-of-the-art. The proposed MWRN adopts superposition-coded transmission, successive interference cancellation (SIC) reception, power-domain non-orthogonal multiple-access (NOMA) and linear detection/precoding facilitated by massive multiple-input multiple-output (MIMO). First, the user nodes transmit their signals to a massive MIMO-enabled relay, where a linear detector based on maximal ratio combining criterion is used for signal reception. Next, the relay composes a superposition-coded signal for each user node and transmits towards the user nodes by using a linear precoder based on maximal ratio transmission criterion. User nodes perform SIC-based decoding for retrieving symbols sent by the remaining K user nodes. Thus, our proposed MWRN protocol completes the full mutual multi-way data exchange among all users within two time-slots. We derive the achievable sum rate of it via the so-called worst-case Gaussian approximation and show that a significant spectral efficiency gain can be achieved over the existing MWRN counterparts.**

I. Introduction

Non-orthogonal multiple access (NOMA) [1] has been identified as one of the main enabling components of future 5G wireless systems. The main idea of NOMA is to use the same frequency-time resources to accommodate multiple access. NOMA can be categorized into power-domain NOMA and code-domain NOMA [2]–[4]. In power-domain NOMA, different power levels are used to differentiate the signals transmitted by different users in the same time-frequency-space resource. The transmitter superimposes the signal and the receiver uses successive interference cancellations to decode the data. NOMA offers improved spectral efficiencies, reduced latency, and massive connectivity [5].

Sub-6 GHz massive multiple-input multiple-output (MIMO) has also identified as a key enabling technology for next generation wireless standards [6]. In massive MIMO, an unprecedented amount of spatial degrees-of-freedom provided by massive antenna arrays are exploited to serve many user nodes simultaneously in the same time-frequency resource by virtue of spatial multiplexing [7]. The higher number of base-station antennas (between $128 - 256$) provides higher spectral efficiencies and significantly improved energy efficiencies. Multi-way relay networks (MWRNs) are used in scenarios, where a set of users need to mutually exchange their own data [8]. In this case, a relay node will act as the intermediate node to accommodate the data transfer. With the emergence of Internet-of-Things (IoT) for 5G systems, the practical applications for MWRN systems will increase. MWRNs can be identified as a natural extension of two-way relay networks [9].

Prior related research: The fundamental performance limits of MIMO-enabled MWRNs have been established in [**?**]. This analysis reveals that the K user nodes can mutually exchange their symbols within K orthogonal time-slots via an intermediate MIMO-enabled relay. The coexistence of massive MIMO and MWRNs has already been explored in [10]–[12]. Reference [10] adopts the MWRN transmission protocol proposed in [8] and computes the achievable rates in the presence of imperfectly estimated channels at the relay in a multi-cell set-up. In [11], an efficient transmit power allocation scheme has been proposed for the same massive MIMO system model proposed in [10]. Although the MWRN transmission protocol studied in [10], [11] accomplishes the multi-way mutual symbol exchange with low probability of bit errors, it requires K time-slots when the system is comprised of K user nodes. Thus, the overall spectral efficiency is low due to the fractional pre-log factor, which is introduced by K time-slots, in achievable rate expressions. As a remedy, [12] proposes a new transmission protocol for massive MIMO MWRNs in which the required number of time-slots for full mutual data exchange is reduced from K to $\lceil (K-1)/2 \rceil + 1$. The MWRN transmission protocol in [12] reduces the number of time-slots by approximately a half by adopting linear processing, self-interference cancellation, and successive cancellation decoding. However, the achievable sum rate of this MWRN protocol still suffers by a low/fractional pre-log factor.

Our contribution: Having been motivated by the aforementioned studies on massive MIMO MWRNs, in this paper, we propose a novel MWRN protocol, which can reduce the number of time-slots from $\lceil (K-1)/2 \rceil + 1$ (or approximately $K/2$) in [12] to just 2 by adopting the concept of power-domain NOMA into MWRNs. Thus, our proposed MWRN protocol can provide a significant gain in spectral efficiency over the previously proposed massive MIMO MWRN protocols in [10]–[12]. Specifically, we achieve a time-slot reduction of $(1 - 4/K) \times 100\%$ over the current state-of-the-art MWRN protocol [12]. This reduction of time-slots directly translates into a significant spectral efficiency gain over all existing MWRN counterparts [10]–[12]. The proposed MWRN protocol

leverages the benefits of power-domain NOMA and massive MIMO by adopting superposition coding, successive interference cancellation (SIC) and linear detection/precoding.

We analyze a system with multiple users that exchange their data among themselves. Massive MIMO enabled relay facilitates this data transfer during two time slots. In the first time-slot, all user nodes transmit simultaneously to the massive MIMO relay, which in turn applies a linear detector designed based on the maximal ratio combining (MRC) for reception of a superimposed signal. Then, the relay generates a superposition-coded signal, applies an amplification factor, and transmits back to the user nodes by using a linear precoder designed based on maximal ratio transmission (MRT) criterion. Then each user node performs SIC to decode the symbols belonging to remaining $K-1$ users messages. Thus, this proposed MWRN transmission protocol accomplishes full mutual multi-way data exchange in just two time-slots. As well, we investigate the performance of this NOMA-aided massive MIMO MWRN protocol by deriving the achievable sum rate in closed-form. Thereby, we show that it achieves a sum rate gain of $K/2$ (approximately) over the current state-of-the-art counterpart in [12]. Interestingly, this spectral-efficiency gain scales-up with the number of user nodes (K). Moreover, in order to enable an efficient SIC operation at each user node, a simple, suboptimal transmit power allocation scheme is proposed for generating the superposition-coded signal at the massive MIMO relay. A set of rigorous numerical results are presented to compare the achievable sum rate performance of the proposed MWRN protocol with competing counterparts. Thereby, we show that the proposed NOMA-aided massive MIMO MWRN achieves a spectral efficiency gain, which scales-up with the number of user nodes. On the contrary, the spectrum efficiency of all related prior research [10]–[12] exhibits a demising return when the number of user nodes increases. Thus, our proposed system model can be used to support massive connectivity with growing spectral efficiency gains.

Notation: $\mathbf{Z}^H$, $\mathbf{Z}^T$, $[\mathbf{Z}]_k$, and $[\mathbf{Z}]_{k,k}$ denote the Hermitian-transpose, transpose, the kth row, and the kth diagonal element of the matrix, $\mathbf{Z}$, respectively. A complex Gaussian random variable X with mean μ and standard deviation σ is denoted as $X \sim \mathcal{CN}\left(\mu, \sigma^2\right)$.

II. System, channel, and signal model

A. System and channel model

Our system model consists of K users which are represented as S_k where $k \in \{1, \cdots, K\}$. All the users are single antenna terminals. The user node S_k needs to transmit its data to the remaining $K-1$ users and needs to receive data from all the $K-1$ users. The relay is represented by R consists of M antennas and is massive MIMO enabled (i.e. $M \gg K$). The direct channels between the users are not available due to heavy fading [9], [13] or not utilized in-order to control the interferences among the users. Thus, R will accommodate the data transfer between the users.

The $M \times 1$ channel between S_k and R is represented as $\mathbf{h}_k = \sqrt{\beta_k}\tilde{\mathbf{h}}_k$ where β_k is the path-loss component and the small scale fading $\tilde{\mathbf{h}}_k$ is given as

$$\tilde{\mathbf{h}}_k \sim \mathcal{CN}\left(0, \mathbf{I}_M\right). \tag{1}$$

The $M \times K$ matrix $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \cdots, \tilde{\mathbf{h}}_K]$ represent combined small scale fading channel matrix from all the users towards the relay. Thus, the channel from the users to the relay can be written as

$$\mathbf{H} = \tilde{\mathbf{H}}\mathbf{D}^{1/2}, \tag{2}$$

where $\mathbf{D} = diag(\beta_1, \ldots, \beta_K)$.

B. Signal model

The data transmission between the users is done using two time-slots. In the first time-slot, all the users transmit to the relay and R receives the data by using receive beamforming. User S_k transmit

$$x_{t,k} = \sqrt{\alpha_k P}x_k, \tag{3}$$

where x_k is the data that it needs to transmit, P is the available maximum power (assumed to be equal for all the users) and $0 < \alpha_k \leq 1$ is the power scaling factor. The received signal at the relay is given as

$$\mathbf{y}_r = \sqrt{P}\mathbf{H}\alpha^{1/2}\mathbf{x} + \mathbf{n}_R, \tag{4}$$

where $\mathbf{x} = [x_1, \cdots, x_K]^T$, $\alpha = diag(\alpha_1, \ldots, \alpha_K)$, and $\mathbf{n}_r$ is $M \times 1$ additive white Gaussian noise (AWGN) vector at the relay satisfying $\mathcal{E}\left[\mathbf{n}_R^H\mathbf{n}_R\right] = \mathbf{I}_M\sigma_R^2$. The received signal after the combining at the relay can be given as

$$\begin{aligned}\mathbf{y}_p &= \mathbf{W}_R\mathbf{y}_r \\ &= \mathbf{W}_R\left(\sqrt{P}\mathbf{H}\alpha^{1/2}\mathbf{x} + \mathbf{n}_R\right),\end{aligned} \tag{5}$$

where $\mathbf{W}_R$ is the beamforming matrix at the relay. Then, the relay will make a superimposed signal to be transmitted to the users as follows:

$$\mathbf{y}_t = \beta\mathbf{W}_T\mathbf{\Lambda}\mathbf{y}_p = \beta\mathbf{W}_T\begin{bmatrix}\lambda_{1,1} & \lambda_{1,2} & \cdots \\ \lambda_{1,2} & \lambda_{2,2} & \cdots \\ \vdots & \ddots & \\ \lambda_{K,1} & \cdots & \lambda_{K,K}\end{bmatrix}\mathbf{W}_R\mathbf{y}_r. \tag{6}$$

where $\mathbf{W}_T$ is the precoding matrix at the relay, $\mathbf{\Lambda}$ is the $K \times K$ power allocation matrix at the relay, and β is the power control factor. The total power constraint at the relay is given as

$$P_R = \mathrm{Tr}\left(\mathbf{y}_t\mathbf{y}_t^H\right) = \beta^2 P\mathrm{Tr}\left(\mathbf{V}\mathbf{H}\alpha\mathbf{H}^H\mathbf{V}^H\right) + \beta^2\sigma_R^2\mathrm{Tr}\left(\mathbf{V}\mathbf{V}^H\right), \tag{7}$$

where we define $\mathbf{V} = \mathbf{W}_T\mathbf{\Lambda}\mathbf{W}_R$ as the total precoding/beamforming vector at the relay, and the relay gain β is computed to constrain the average transmit power as

$$\beta = \sqrt{\frac{P_R}{P\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}\mathbf{H}\alpha\mathbf{H}^H\mathbf{V}^H\right)\right] + \sigma_R^2\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{V}\mathbf{V}^H\right)\right]}}. \tag{8}$$

$$\mathcal{R}_{f_k(n),k}=\frac{1}{2}\log\left(1+\frac{\beta^2 P\alpha_{f_k(n)}|\mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right]|^2}{\beta^2 P\alpha_{f_k(n)}\mathbb{V}ar\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right]+\beta^2 P\sum\limits_{m=n+1}^{K-1}\alpha_{f_k(m)}\mathbb{E}\left[|\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(m)}|^2\right]+\beta^2\sigma_R^2\mathbb{E}\left[\|\mathbf{h}_k^T\mathbf{V}\|^2\right]+\sigma_k^2}\right) \tag{13}$$

$$\beta=\sqrt{\frac{P_R}{P\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{H}\alpha\mathbf{H}^H\mathbf{H}\mathbf{\Lambda}^H\mathbf{H}^T\right]\right)+\sigma_R^2\mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{H}\mathbf{\Lambda}^H\mathbf{H}^T\right]\right)}}=\sqrt{\frac{P_R}{PL_1+\sigma_R^2L_2}}. \tag{18}$$

Based on (6), the intended transmitted signal for S_k is given as $[\mathbf{y}_t]_k$, which is the kth row of $\mathbf{y}_t$. The received signal at S_k is given as

$$\begin{aligned}y_k&=\mathbf{h}_k^T\mathbf{y}_t+n_k\\&=\beta\mathbf{h}_k^T\mathbf{V}\sum_{m=1}^{K}\sqrt{P\alpha_m}\mathbf{h}_m x_m+\beta\mathbf{h}_k^T\mathbf{V}\mathbf{n}_R+n_k,\end{aligned} \tag{9}$$

where n_k is the AWGN at S_k with power σ_k^2.

III. Achievable sum rate analysis for MRC/MRT

We next obtain the achievable sum rate between each user pairs by using the worst-case Gaussian approximation [14] for a system with MRC/MRT beamforming. The received signal at S_k before after decoding $n-1$ users is written as

$$\begin{aligned}y_{k,n}&=\beta\sqrt{P\alpha_{f_k(n)}}\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\\&+\beta\mathbf{h}_k^T\mathbf{V}\sum_{m=n+1}^{K-1}\sqrt{P\alpha_{f_k(m)}}\mathbf{h}_{f_k(m)}x_{f_k(m)}+\beta\mathbf{h}_k^T\mathbf{V}\mathbf{n}_R+n_k\\&=\beta\sqrt{P\alpha_{f_k(n)}}\mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right]+\hat{\mathbf{n}}_{\mathbf{k},\mathbf{n}},\end{aligned} \tag{10}$$

where $\hat{\mathbf{n}}_{\mathbf{k},\mathbf{n}}$ is the effective noise and the first term is the desired signal. The function $f_k(n)$ contains the index of the user that will be decoded in the nth order. For MRC/MRT beamforming, $\mathbf{V}$ is given as

$$\mathbf{V}=\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H. \tag{11}$$

The noise term in (10) can be expressed as

$$\begin{aligned}\hat{\mathbf{n}}_{\mathbf{k},\mathbf{n}}=&\underbrace{\beta\sqrt{P\alpha_{f_k(n)}}\left(\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}-\mathbb{E}\left[\mathbf{h}_k^T\mathbf{V}\mathbf{h}_{f_k(n)}x_{f_k(n)}\right]\right)}_{\text{detection uncertainty}}\\&+\underbrace{\beta\mathbf{h}_k^T\mathbf{V}\sum_{m=n+1}^{K-1}\sqrt{P\alpha_{f_k(m)}}\mathbf{h}_{f_k(m)}x_{f_k(m)}}_{\text{interference from other users}}\\&+\underbrace{\beta\mathbf{h}_k^T\mathbf{V}\mathbf{n}_R}_{\text{amplified noise}}+\underbrace{n_k}_{\text{AWGN noise at the receiver}}.\end{aligned} \tag{12}$$

Based on (10) and (12), and by assuming additive noise as independently distributed Gaussian noise having the same variance [14], a tight approximation for the achievable sum rate can be given as (13) at the top of this page. By evaluating the variance and expectation terms in (13), the achievable rate can be computed as (see Appendix A for derivation)

$$\mathcal{R}_{f_k(n),k}=\frac{1}{2}\log\left(\frac{P\alpha_{f_k(n)}\beta^2M_{k,m}^2}{P\alpha_{f_k(n)}\beta^2N_{k,m}+\beta^2\sum\limits_{m=n+1}^{K-1}P_{k,m}+\beta^2Q_k+\sigma_k^2}\right). \tag{14}$$

where $M_{k,m}$, $N_{k,m}$, $P_{k,m}$, and Q_k are given as

$$M_{k,m}=\mathbb{E}\left[\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}\right], \tag{15a}$$

$$N_{k,m}=\mathbb{V}ar\left[\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}\right], \tag{15b}$$

$$P_{k,m}=P\alpha_{f_k(m)}\mathbb{E}\left[|\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}|^2\right], \tag{15c}$$

$$Q_k=\sigma_R^2\mathbb{E}\left[\|\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\|^2\right]. \tag{15d}$$

The closed-form evaluations of (15) are provided in Appendix A. Based on (14), the total achievable sum rate at S_k is given as

$$\mathcal{R}_k=\sum_{n=1}^{K-1}\mathcal{R}_{f_k(n),k}. \tag{16}$$

The total achievable sum rate of the system is obtained as

$$\mathcal{R}=\sum_{k=1}^{K}\sum_{n=1}^{K-1}\mathcal{R}_{f_k(n),k}. \tag{17}$$

A. Calculation of β

By using (8), the value for β, when perfect CSI is available, can be rewritten as (18) at the top of this page and values of L_1 and L_2 is given in Appendix B.

IV. Comparison with MWRN operations in [10], [12]

We next analyze a normal MWRN for comparison purposes. We consider a system model which requires K time slots to transmit the data of all the users to all other users. The use of time-slots can be listed as follows [10].

1) **Time-slot 1** All the users transmit to the relay. This step is similar to the first step in NOMA protocol.
2) **Time-slot 2 to K** In this time slot, relay transmits to all the users using digital beamforming. However, instead of sending a superimposed signal of all the other received signals, data of a single user is transmitted to each user. Thus, $K-1$ time slots are required to send the data of all the users to all other users.

The beamforming matrix at the relay for the jth time slot ($2\leq j\leq K$) is given as

$$\mathbf{V}_j=\mathbf{H}^*\mathbf{\Lambda}_j\mathbf{H}^H. \tag{19}$$

here in (19), Λ_j is a permutation matrix in which each row consists of a single one and all zeros. The location of the number one, decides the transmitted signal of the initial set of users. We can obtain the approximation for the end-to-end data rate between each pair of users by using the same steps as the previous case. However, when calculating the achievable data rate, we have to use the pre-log factor $\frac{1}{K}$, as K total time slots are required for the data transmission. Moreover, we compare the performance of the proposed MWRN with that of [12], which utilizes $\lceil (K-1)/2 \rceil + 1$ time-slots. These two references [10], [12] are used as a benchmark to test the performance of the proposed NOMA-aided massive MIMO MWRN.

V. Sum Rate Maximization

We next focus on designing $\mathbf{\Lambda}$ and $f_k(n)$ to maximize the achievable sum rate while satisfying the relay power constraints. The optimization problem can be formulated as

$$\underset{\mathbf{\Lambda}, f_k(n), \alpha}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{n=1}^{K-1} \mathcal{R}_{f_k(n),k} \tag{20}$$

Basically, (20) requires SIC order function $f_k(n)$ and the power allocation coefficient matrix at the relay. Due to space limitations, we only consider a sub-optimal power allocation matrix and a decoding order. Without the loss generality, we assume that users are ordered according to the descending order of path-loss between them and the relay (i.e. $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_K$). Thus, first we order the users according to the average path-loss (i.e. β_k) and use this order to decide the function $f_k(x)$. The decoding order is decided as decoding the users according to their natural order. This can be mathematically defined as

$$f_k(n) = \begin{cases} n & n < k \\ n+1 & n \geq k \end{cases} \tag{21}$$

Regarding power allocation matrix, we observe the following.

- Each row in $\mathbf{\Lambda}$ corresponds to the power allocation factors to each user. Thus, to compensate for the downlink path-losses, we allocate more power to the users which have the highest path-loss by multiplying each row of $\mathbf{\Lambda}$ by the inverse of the path-loss component of each user.
- Diagonal entries of $\mathbf{\Lambda}$ should be zeros to avoid the self interference.
- The ratios between the non-zero coefficients in a single row determines the data rate between the users. In this paper, we designed the ratios in the form of $\sqrt{\frac{1}{K}}, \sqrt{\frac{2}{K}}, \cdots, \sqrt{\frac{K-1}{K}}$.

Based on these points, a suboptimal power allocation matrix is designed and used as

$$\mathbf{\Lambda} = \mathbf{D}^{-1} \begin{bmatrix} 0 & \sqrt{\frac{1}{K}} & \cdots & \sqrt{\frac{K-1}{K}} \\ \sqrt{\frac{1}{K}} & 0 & \cdots & \sqrt{\frac{K-1}{K}} \\ \vdots & \ddots & & \vdots \\ \sqrt{\frac{1}{K}} & \cdots & \sqrt{\frac{K-1}{K}} & 0 \end{bmatrix}. \tag{22}$$

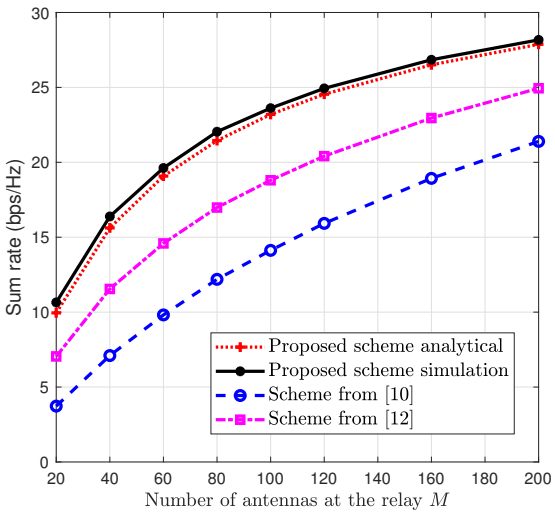

Fig. 1: Total sum rate for $K = 8$ against the number of relay antennas.

Remark 1: Our simulations show that the design of $\mathbf{\Lambda}$, $f_k(n)$, and α affects the achievable sum rate between each user pair. By manipulating these three degrees of freedoms, various data rate combinations can be achieved between the users. Further analysis on the design of these parameters will be forthcoming in an extended version of this paper.

VI. Numerical Results

Here, we present simulation results to demonstrate the performance gains of proposed NOMA scheme. We use the power allocation matrix defined in the previous section with $K = 8$, $\mathbf{D} = \text{diag}(1, 0.875, 0.75, 0.625, 0.5, 0.375, 0.25, 0.125)$, and $\alpha_k = 1$ for $1 \leq k \leq K$.

In Fig. 1, we plot the achievable sum rate of the proposed system and schemes in [10], [12] against the number of relay antennas. As evident from the figure, the proposed scheme provides a higher achievable sum rate compared to other two schemes. As an example with 100 relay antennas, the proposed scheme provides a sum rate of 24 bps/Hz while the sum rates of [10] and [12] are only 14 bps/Hz and 19 bps/Hz. This is an increase of 71% and 26% respectively. Furthermore, it can be seen that increasing the number of relay antennas results in increased sum rates of the system. In Fig. 2 we plot the received sum rates for S_2 against the number of relay antennas. This shows the same performance improvements in sum rate for a single user and also justifies our closed form solutions.

In Fig. 3, we plot the achievable sum rate against the number of users (K). As K increases, the proposed NOMA MWRN provides higher data rate compared to other two schemes. The achievable sum rate gap between our method and those two increases as the number of users are increased. As an example with 6 users, the gap is 4 bps/Hz which is a 21% increase while the gap at 20 users is 12 bps/Hz which corresponds to a 31% increase compared to [12]. This justifies the use of proposed NOMA MWRN with large numbers of users.

In Fig. 4, we plot achievable sum rate per user ($\mathcal{R}/K$) for the proposed system and two other MWRN schemes. While the

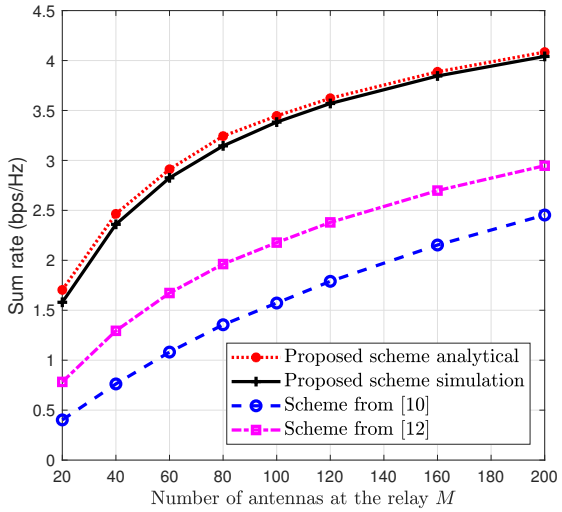

Fig. 2: Sum rate of S_2 for $K = 8$ against the number of relay antennas.

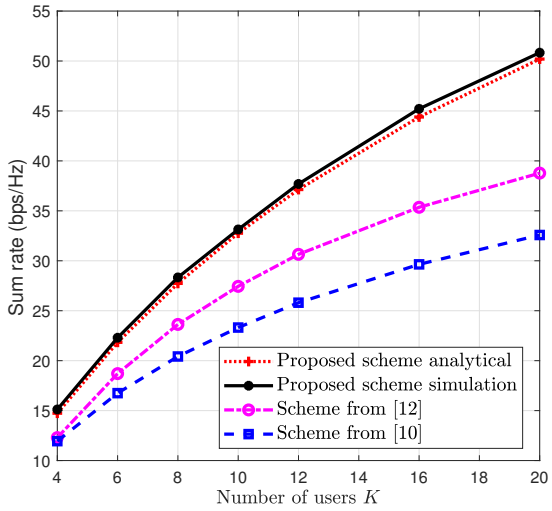

Fig. 3: Total sum rate versus K for $M = 128$.

achievable sum rate per user decreases, our method outperforms [10] and [12]. With 10 users, it provides 3.4 bps/Hz achievable sum rate per user, while the [10] and [12] can only provide sum rates of 2.1 bps/Hz and 2.5 bps/Hz per user. Of course, the sum rate per user decreases with K, but due to different reasons. In our method, the decrease of the achievable sum rate per user is due to the increased interferences among the users as K increases, while in other two schemes the decrease is due to the increased number of time slots used in the data transfer process.

VII. Conclusion

We proposed a NOMA-aided massive MIMO MWRN, which enables full mutual multi-way data exchange between K users within only two time slots. This is a drastic reduction of $\lceil (K-1)/2 \rceil - 1$ time-slots with respect to the current state-of-the-art solution for MWRN. This in turn translates directly

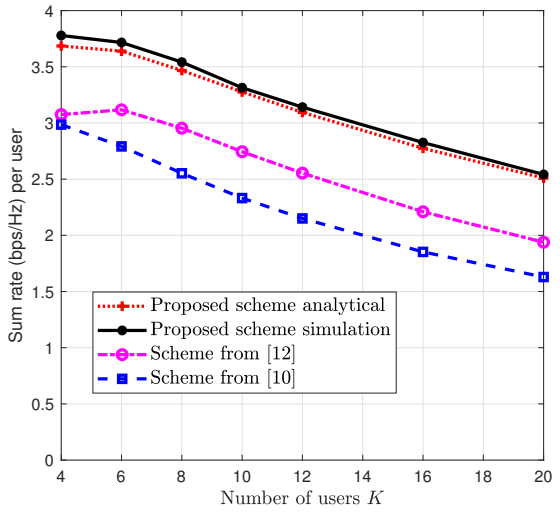

Fig. 4: Average sum rate per user versus K for $M = 128$.

into a significant spectral efficiency gain. In the first time-slot, all user nodes transmit simultaneously to the relay, which in turn applies a linear MRC detector. In the second time-slot, relay constructs a superposition-coded signal consisting of data symbols belonging to all other users to be transmitted by using linear MRT precoding to each user. Upon receiving this superposition-coded signal, users adopt SIC to decode the data from each user. We derived the the achievable sum rate in closed-form and compared the underlying performance gains with competing MWRN literature. Numerical results show that our proposed MWRN provides a significantly higher achievable sum rate (approximately a gain of $(1 - 4/K) \times 100\%$) compared to a typical MWRN. This performance increase is more significant when the number of users (K) is increased. This is the key benefit of our proposed MWRN because all the related prior results achieve diminishing spectral efficiency gains with the increasing number of users. Thus, our proposed scheme may enable high-speed massive connectivity in next-generation wireless applications.

Appendix A
Proof of Limits for SINR

A. Derivation of $M_{k,m}$

To calculate $M_{k,m}$, we first rewrite the term inside the expected value by using basic matrix multiplication steps as

$$\mathbf{h}_k^T \mathbf{H}^* \mathbf{\Lambda} \mathbf{H}^H \mathbf{h}_{f_k(m)} = \sum_{i=1}^{K} \sum_{j=1}^{K} \lambda_{i,j} \mathbf{h}_k^T \mathbf{h}_i^* \mathbf{h}_j^H \mathbf{h}_{f_k(m)}. \quad (23)$$

Next, by considering different i and j combinations (i.e. $i = k$, $j = f_k(m)$ and $j = k$, $i = f_k(m)$ and by using the fact that $k \neq f_k(m)$) in the double summation in (23), the expected value can be written as

$$\mathbb{E}\left[\sum_{i=1}^{K} \sum_{j=1}^{K} \lambda_{i,j} \mathbf{h}_k^T \mathbf{h}_i^* \mathbf{h}_j^H \mathbf{h}_{f_k(m)} \right] = \lambda_{k,f_k(m)} \|\mathbf{h}_k\|^2 \|\mathbf{h}_{f_k(m)}\|^2 + \lambda_{f_k(m),k} |\mathbf{h}_k \mathbf{h}_{f_k(m)}|^2. \quad (24)$$

$$\Phi_i = \beta_i \left[\lambda_{i,i}^2\beta_i + 2\lambda_{k,i}\lambda_{i,k}\beta_k + 2\lambda_{m,i}\lambda_{i,m}\beta_m + M\lambda_{k,i}^2\beta_k + \lambda_{m,i}^2\beta_m + \lambda_{i,k}^2\beta_k + (M+2)\lambda_{i,m}^2\beta_m\right]. \quad (27a)$$

$$\Theta_{i,j} = \lambda_{i,j}^2\beta_i\beta_j + (M+1)\left((M+2)\left(\lambda_{k,k}^2\beta_k^2 + \lambda_{m,m}^2\beta_m^2\right) + \lambda_{k,m}\lambda_{m,k}(M+1)\beta_m\beta_k + \lambda_{k,m}^2 M(M+1)\beta_m\beta_k + 2\lambda_{m,k}^2\beta_m\beta_k\right) \quad (27b)$$

$$Q_k = M\beta_k\left((M+1)\sum_{i=1,i\neq k,m}^{K}\lambda_{i,i}^2\beta_i^2 + (M+1)(M+2)\lambda_{k,k}^2\beta_k^2 + M\sum_{i=1,i\neq k}^{K}\sum_{j=1,j\neq k,i}^{K}\left(M\lambda_{i,j}^2 + \lambda_{i,j}\lambda_{j,i}\right)\beta_i\beta_j + (M+1)\beta_k\sum_{i=1,i\neq k}^{K}\left(2\lambda_{k,i}\lambda_{i,k} + M\lambda_{k,i}^2 + \lambda_{i,k}^2\right)\beta_i\right). \quad (29)$$

$$L_1 = M\left((M+1)\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\left(\alpha_j\lambda_{i,i}^2\beta_i + \alpha_i\lambda_{i,j}^2\beta_i + \alpha_i\lambda_{i,j}\lambda_{j,i}\beta_i + \alpha_j\lambda_{i,j}\lambda_{j,i}\beta_j + M\alpha_j\lambda_{i,j}^2\beta_j\right)\beta_j\beta_i + (M+1)(M+2)\sum_{i=1}^{K}\alpha_i\lambda_{i,i}^2\beta_i^3 + \sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\sum_{m=1,m\neq i,j}^{K}\alpha_m\lambda_{i,j}^2 M\beta_i\beta_j\beta_m\right). \quad (30)$$

Finally, by using the expected value results M_1 is given as

$$M_{k,m} = \left(\lambda_{k,f_k(m)}M + \lambda_{f_k(m),k}\right)M\beta_k\beta_{f_k(m)}. \quad (25)$$

B. Derivation of $N_{k,m}$

In order to calculate $N_{k,m}$, we first look at $\bar{N}_{k,m} = \mathbb{E}\left[|\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}|^2\right]$, where the argument of the expectation operator can be expanded as

$$\left|\mathbf{h}_k^T\mathbf{H}^*\mathbf{\Lambda}\mathbf{H}^H\mathbf{h}_{f_k(m)}\right|^2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{l=1}^{K}\sum_{n=1}^{K}\lambda_{i,j}\lambda_{n,l}\mathbf{h}_k^T\mathbf{h}_i^*\mathbf{h}_j^H\mathbf{h}_{f_k(m)}\mathbf{h}_{f_k(m)}^H\mathbf{h}_l\mathbf{h}_n^T\mathbf{h}_k^*. \quad (26)$$

The expected value of (26) or $\bar{N}_{k,m}$ can be calculated by considering different parameter settings for i, j, l, and n and is given i as

$$\bar{N}_{k,m} = M\beta_k\beta_m\left((M+1)\sum_{\substack{i=1\\ i\neq k,m}}^{K}\Phi_i + M\sum_{\substack{i=1\\ i\neq k,m}}^{K}\sum_{\substack{j=1\\ j\neq k,m,i}}^{K}\Theta_{ij}\right), \quad (27)$$

where Φ_i and Θ_{ij} in (27) are given by (27a) and (27b), respectively at the top of this page. Due to space limitations we have omitted the intermediate steps of this calculation. The value of $N_{k,m}$ is obtained as

$$N_{k,m} = \bar{N}_{k,m} - M_{k,m}^2. \quad (28)$$

C. Derivation of $P_{k,m}$ and Q_k

$P_{k,m}$ can be calculated by using the results obtained in (27) and Q_k is calculated by using similar steps as in $N_{k,m}$. Value for Q_k is written as (29) at the top of this page.

APPENDIX B
CALCULATION OF β

By using similar steps as in Appendix A, L_1 is obtained in (30) at the top of this page and L_2 is given as

$$L_2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\left(\lambda_{i,j}M + \lambda_{j,i}\right)M\lambda_{i,j}\beta_i\beta_j. \quad (31)$$

REFERENCES

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

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

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

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

[5] H. V. Cheng, E. Bjrnson, and E. G. Larsson, "Performance analysis of NOMA in training-based multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 372–385, Jan. 2018.

[6] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[7] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[8] G. Amarasuriya, C. Tellambura, and M. Ardakani, "Multi-way MIMO amplify-and-forward relay networks with zero-forcing transmission," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4847–4863, Dec. 2013.

[9] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 379–389, Feb. 2007.

[10] D. P. Kudathanthirige and G. A. A. Baduge, "Multicell multiway massive MIMO relay networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6831–6848, Aug. 2017.

[11] C. D. Ho, H. Q. Ngo, M. Matthaiou, and T. Q. Duong, "On the performance of zero-forcing processing in multi-way massive MIMO relay networks," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 849–852, Apr. 2017.

[12] C. D. Ho, H. Q. Ngo, M. Matthaiou, and L. D. Nguyen, "Power allocation for multi-way massive MIMO relaying," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4457–4472, Oct. 2018.

[13] G. Amarasuriya, C. Tellambura, and M. Ardakani, "Performance analysis of zero-forcing for two-way MIMO AF relay networks," *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, pp. 53–56, Apr. 2012.

[14] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.