Design and Evaluation of Multi-layer NOMA on NR Physical Layer for 5G and Beyond

Md Shantanu Islam, Graduate Student Member, IEEE, Raouf Abozariba, Member, IEEE, De Mi, Senior Member, IEEE, Mohammad Patwary, Senior Member, IEEE, Dazhi He, Member, IEEE, A. Taufiq Asyhari, Senior Member, IEEE,

Abstract-This paper investigates the integration of multilayer non-orthogonal multiple access (N-NOMA) into a 5G New Radio (NR) compliant transceiver model, aiming to reveal the full potential of the NOMA technology in practical scenarios. We propose an N-NOMA-aided 5G NR physical layer (PHY) design, where a simplified multi-layer NOMA multiplexer with a one-shot multiplexing technique is developed to reduce the transmitter complexity and thus potential delay for processing the additional NOMA layers. Our design offers a new perspective for the NOMA technology to address various challenging use cases, such as massive machine type communication (mMTC) and enhanced mobile broadband (eMBB) under low signal-to-noise ratio (SNR) regimes. Then, in order to provide a comprehensive error performance evaluation of the proposed N-NOMA PHY design, we take into account various system configurations, e.g., different modulation and coding schemes (MCSs) with lowdensity parity-check (LDPC) code and different multi-input multi-output (MIMO) configurations. During the evaluation of the proposed design, we uncovered key factors missing from the existing bit error rate (BER) analytical models literature, e.g., the imperfect successive interference cancellation (SIC). The derived BER expressions capture the effect of the SIC errors, which is consistent in our analytical and simulation performance comparison. Through the simulation, we also comprehensively evaluate and discuss the link-level performance of the proposed PHY design.

Index Terms—Non-orthogonal multiple access (NOMA), multiple-input multiple-out (MIMO), low-density parity-check (LDPC), New Radio (NR), bit error rate (BER), successive interference cancellation (SIC)

I. INTRODUCTION

FUTURE cellular networks aim to support unprecedented user connectivity with a certain quality of experience met evenly, to enable services such as live sports and video

M. S. Islam, R. Abozariba, and D. Mi are with the School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK (e-mails: shantanuislam@ieee.org, r.abozariba@ieee.org, de.mi@bcu.ac.uk).

M. Patwary is with the Department of Computer Science, University of Wolverhampton, Wolverhampton, UK (e-mail: patwary@wlv.ac.uk).

D. He is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hedazhi@sjtu.edu.cn).

A. T. Asyhari is with the Data Science Department, Monash University, Indonesia Campus, BSD City, Tangerang 15345, Indonesia (email: taufiq-a@ieee.org).

game broadcasting [1], or to deliver massive machine-tomachine communications (mMTC), e.g., large-scale software updates for IoT devices [2]-[5]. Supporting different service or device categories is a challenging task, due to diverse data rates, power efficiency, latency and accuracy requirements [6]. Non-orthogonal multiple access (NOMA), which allows transmission of multiple data streams over the same timefrequency resources, has shown a potential to support massive connectivity, providing greater spectrum efficiency and system capacity [7]-[9]. Advanced television systems committee (ATSC) 3.0, a major standard for television broadcasting, uses a 2-layer division multiplexing (LDM), a subset of the power domain NOMA introduced in 2012 [10]-[12]. These two layers, namely upper layer (UL) and lower layer (LL), provide services to portable or mobile receivers with low data rates, and services (e.g., UHDTV/HDTV) to fixed receivers, respectively. Then at the receivers, successive interference cancellation (SIC) technique plays a key role in eliminating the higher-layer interference from the composite signal by subtracting the received higher-layer data after reconstruction from the composite signal [13], [14]. Previous studies have shown that using NOMA in ATSC 3.0 physical layer (PHY) outperforms orthogonal multiplexing counterparts, e.g., timedivision multiplexing (TDM), in terms of throughput for mobile services [11], [12]. Also, the application of NOMA in ATSC 3.0 highlights its adaptivity to varying demands of the receiver heterogeneity [15], [16]. The integration of NOMA into the ATSC 3.0 has paved avenues for its application in 5G New Radio (NR) as shown in [17], [18].

Over time, extensive literature [19]–[24] has developed on exploring the intricate nature of the NOMA-based transmitter and the effect of variable channel conditions on system performance. The authors of [25] combined spatial modulation with multiple-input and multiple-output (MIMO) and LDM, suggesting that increasing the MIMO order could potentially enhance the delivery of broadcast/multicast services. However, this model does not incorporate key elements such as lowdensity parity-check (LDPC) forward error correction and practical MIMO channel modeling. In [26] it suggested that using low-rate LDPC codes in the UL improves the overall performance and achieves higher transmission efficiency. Motivated by these research gaps, we consider a practically novel transceiver model, incorporating a series of NOMA functionalities for heterogeneous services to a 5G NR-complaint PHY, including MIMO precoding and LDPC coding, with tapped delay line (TDL) or clustered delay line (CDL) channel

This work was supported in part by the U.K. West Midlands Innovation Accelerator WMHTIA, in part by the P.R.C. Ministry of Education under Grant HZKY20220106 No. 202202120 and the P.R.C. Shenzhen Science and Technology Innovation Commission under Grant GJHZ20220913143013024, and in part by the National Natural Science Foundation of China Program 62271316 and the 111 Plan of China under Grant BP0719010. (*Corresponding authors: Raouf Abozariba, De Mi*).

model. We also take into account critical NOMA techniques for incorporation and integration, including symbol rate synchronization by aligning signals and optimal placement of NOMA PHY processing functions in the 5G transceiver chain, to reduce latency and complexity.

While the two data layers LDM or NOMA structure remains limited capacity to support services such as mMTC and enhanced mobile broadband (eMBB) with massive concurrent users, multi-layer NOMA (N-NOMA) (> 2) can play a significant role here to enhance system capacity [26]. In recent work, Kim et al. [26] explored the potential of a time-shared 3-layer LDM system for ATSC 3.0, and tested various multi-layer systems and investigated their capacity enhancement potential. The time-shared 3-layer NOMA is a way to provide 3 services using a 2-layer LDM design where the lower layer is shared between two services due to the performance degradation and complexity associated with vertical integration. However, the integration of N-NOMA, e.g., more than 3 layers, will introduce severe challenges to the practical system, e.g., the requirement of high transmission power and the increased complexity of the transceiver design. As demonstrated in [27], [28], adding more NOMA layers, while possible, the minimum signal-to-noise ratio (SNR) requirement increases, which is challenging in practical scenarios, particularly if the microWave/mmWave spectrum is used. Compared to previous generations, 5G NR PHY performance at low SNR regimes offers a higher quality of service (OoS) owed to the inclusion of technical enablers such as LDPC and MIMO. It is of great interest to see if a combination of NOMA and NR technologies can reduce the minimum SNR requirements when adding new layers. On the other hand, the work in [29] showed that the complexity at the transmitter side arises from data processing, NOMA multiplexing, power allocation, and user grouping. The main issues at the receiver side are the multiple SIC operations (n-1) SIC operation for the *n*th data layer) and maintaining the minimum SNR levels for successful signal detection [30]. Resource allocation in N-NOMA transmission is a complex problem concerning power allocation and user pairing as data layers increase [31], [32]. Note that, unlike point-to-point unicast, point-to-multipoint (PTM) transmission does not require exclusive user pairing, lowering resource allocation complexity and thus making N-NOMA more attractive for PTM service delivery.

Our motivation is not only to explore the feasibility of N-NOMA in conjunction with 5G NR technical enablers, but also potentially address some key implementation issues of the N-NOMA-aided PHY. For example, one of the drawbacks of including more NOMA layers in a standard sequential combiner is the added complexity of the additional signal processing functions [26]. Our proposal is to design a one-shot technique that multiplexes all NOMA layers using a single function. Another example is that, during the SIC operation, an imperfect detection of higher layer data can add interference to the lower layer signal. This important factor has not been considered in the previous bit error rate (BER) analytical models, leading to an overoptimistic estimation of error. As the scale of the NOMA order increases, the estimation becomes even more deflated with the SIC operation in the system. An accurate BER analysis, essential for validating NOMA architectures in practice, should incorporate the impact of SIC into the error modeling, e.g., considering both residual and non-residual errors in case of the perfect or imperfect SIC, respectively. In this context, we define the residual error events, which occur when the symbol from the preceding layer is incorrectly decoded, leading to additional interference during SIC operation. On the other hand, non-residual error events apply to cases where the preceding layer symbol is correctly decoded, resulting in NOMA interference solely from the impact of all subsequent NOMA layers. In [33], a closed form BER was derived for NOMA with receiver diversity for several modulation orders. However, the authors did not capture the impact of SIC in their proposed solution, and they assumed a single antenna transmission which has certain limitations in 5G. The authors in [31] derived a theoretical BER expression for N-NOMA-quadrature amplitude modulation (QAM) with equal power distribution between data layers. However, they did not consider orthogonal frequencydivision multiplexing (OFDM) or MIMO processing in their expression. Furthermore, although they explored SIC operation for the BER performance, they did not quantify the additional interference caused by the imperfect SIC. With the above in perspective, we aim to derive a closed-form BER analytical expression by accounting for both residual and non-residual errors from SIC at the receiver, incorporating MIMO gain, using TDL/CDL model and thus expanding its scope. Our

approach to derive the error probability follows the Q function estimation, where we assume uncoded OFDM system and Gaussian distributed interference for residual and non-residual errors.

The contributions of this paper are summarized as follows.

- We propose a new N-NOMA-aided 5G NR-complaint transceiver design, incorporating a series of NOMA and MIMO functions to the NR PHY to enable heterogeneous transmission, combining multiple NOMA layers over a common communication channel.
- We provide an improved N-NOMA multiplexer design by developing a one-shot technique to reduce the complexity and latency, compared to the standard sequential combiners.
- We derive a closed-form BER expression, accounting for both residual and non-residual errors from SIC, and incorporating MIMO gain and practical channel model, which introduces a new tool for understanding and analyzing NOMA systems. In our expression, we quantify the maximum non-residual error as the maximum pairwise distance between square QAM (SQAM) symbols capturing its maximum possible impact on the error probability.
- We conduct a series of BER measurements for different modulation schemes at various NOMA layers and normalized SNR levels under a TDL-MIMO channel. As such, the practical applicability and effectiveness of the proposed transceiver design and analytical model are verified through rigorous simulation results.

The structure of the remainder of this paper is as follows: Section II proposes a system model incorporating a 2-layer



Fig. 1. Illustrates a 5G transceiver model integrating a 2-layer NOMA scheme. The shaded areas denote proposed modifications for NOMA integration, while the remaining components adhere to the standard 3GPP 5G transceiver model.

NOMA-aided 5G PHY and a transceiver model. This model is extended in Section III into an N-NOMA transceiver model, allowing the transmission of over 2 data layers using a single traffic channel. This model is further simplified to reduce latency and complexity. Section IV presents a closed-form BER expression to determine the maximum error for any layer of data in N-NOMA transmission. The results of our system and analytical models are presented and evaluated in Section V. Finally, Section VI summarizes our findings and discusses potential future work.

II. 2-LAYER NOMA-AIDED 5G NR-COMPLIANT PHY

We start with a 2-layer NOMA to present the proposed modular PHY design, in which we integrate NOMA-aided transceiver blocks into the standard 5G NR-compliant downlink data transmission model, as specified in 3GPP TR 38.901 [34]. As illustrated in Fig. 1, besides the standardized transceiver blocks such as LDPC-based encoding and OFDM modulation, the model also combines NOMA multiplexing and SIC, MIMO precoding and CDL/TDL channel models, among other functions.

A. Transmitter design

5G NR downlink transmitter commonly has the data input (in binary bits) in the PHY intended for single-user cases. Our design considers multi-user cases, e.g., with the data streams for two users separated into two data layers (e.g., L1 and L2) using NOMA multiplexing, which is merged with the standard data processing chain to combine and transmit data blocks for two independent users using a single traffic channel.

Since the processing of individual data blocks can be executed in parallel before multiplexing, the only added latency in the transmitter side is the time required for NOMA integration. The multiplexing can be achieved on a bit or symbol level, while for both data layers, the length of the processed bit/symbol needs to be identical to carry out this operation. As such, the challenges of our design are: 1) synchronizing the bit/symbol rate of both layers before multiplexing and 2) determining the optimum position of the NOMA multiplexer in the transmitter chain.

Recall in our system description that two input sources are defined as L1 data and L2 data, which can be expressed as:

$$\mathbf{x}_{l1} = (x_{l1,1}, \cdots, x_{l1,n_{l1}}), \tag{1}$$

$$\mathbf{x}_{l2} = (x_{l2,1}, \cdots, x_{l2,n_{l2}}), \tag{2}$$

where $x_{l1,i}, x_{l2,i} \in \{0, 1\}, \forall i$, and n_{l1}, n_{l2} represent the total number of L1 and L2 bits to be transmitted, respectively. To effectively synchronize the symbol rate for transmission, we derive n_{l1} and n_{l2} based on the OFDM transmission symbol rate n_s , which is the number of data symbols in each frame, and the respective coding rates ρ_{l1}, ρ_{l2} , and bits per modulated symbol, m_{l1}, m_{l2} , determined by the modulation order (*M*) by $m = \log_2 M$. We then calculate the bit rate of both layers using the following expressions:

$$n_{l1} = n_s \cdot \rho_{l1} \cdot m_{l1} \tag{3}$$

$$n_{l2} = n_s \cdot \rho_{l2} \cdot m_{l2} \tag{4}$$

Both layers of data are first processed through the downlink shared channel encoding (DL-SCH Encoding) block, where \mathbf{x}_{l1} and \mathbf{x}_{l2} are encoded according to the coding rates, ρ_{l1} and ρ_{l2} . The DL-SCH block uses LDPC error-correcting encoding, in line with 3GPP TR 38.901 Release 16 [34]. The output of the DL-SCH block can be expressed as

γ

$$\mathbf{x}_{dl1} = (x_{dl1,1}, \cdots, x_{dl1,n_{dl1}}), \text{ and}$$
 (5)

$$\mathbf{x}_{dl2} = (x_{dl2,1}, \cdots, x_{dl2,n_{dl2}}), \tag{6}$$

where the size $n_d = (n/\rho)$ for both UL and LL. Next, \mathbf{x}_{dl1} and \mathbf{x}_{dl2} are modulated into QPSK or QAM symbols as per packet data shared channel (PDSCH) block. The output is in the following format:

$$\mathbf{s}_{l1} = (s_{l1,1}, \cdots, s_{l1,n_s}),\tag{7}$$

$$\mathbf{s}_{l2} = (s_{l2,1}, \cdots, s_{l2,n_s}),$$
 (8)

where $s_{l1,i}, s_{l2,i} \in \{s_1, s_2, \cdots, s_M\}, \forall M_{l1} \text{ and } M_{l2}$.

Data integration into the NOMA signal could occur at multiple stages within the 5G transmitter framework, leading to varying degrees of complexity and latency. As both data layers undergo simultaneous processing until they reach the point of multiplexing, implementing a bit-level NOMA multiplexer reduces the total count of operational blocks within the transmitter, decreasing the overall computational complexity. Conversely, symbol-level NOMA multiplexing allows the system to assign different bit rates to each layer, where symbols are synchronized using varying code rates ρ and modulation orders M. Furthermore, the placement of the NOMA integration block within the transmitter also determines the location of SIC at the receiver and the number of additional blocks required for L1 reconstruction. This factor is crucial in determining the overall latency at the L2 signal detection.

In [20], a symbol level 2-Layer NOMA integration was implemented within a simplified 5G transceiver model, incorporating a hybrid automatic repeat request (HARQ). Similarly, [16] designed a low complexity LDM-enabled transceiver model for ATSC 3.0, also applying symbol level integration. In line with these studies, we positioned the NOMA integration block after the PDSCH block, representing the earliest data symbol form in the 5G PHY as per 3GPP TR 38.901 [34]. This design approach offers the benefit of individual coding rates ρ and modulation orders M to perform synchronized symbol rates, enabling transmission of different bit rates for each layer while minimizing latency. In a sense, this is the optimum position for NOMA multiplexer block.

The symbols from both layers are multiplexed together using the following relation:

$$s_{NOMA,i} = s_{l1,i} + g \cdot s_{l2,i},$$
 (9)

 $s_{l1,i}$ and $s_{l2,i}$ are *i*th L1, L2 QAM symbols, and $s_{NOMA,i}$ is the *i*th NOMA multiplexed symbol. The combined data \mathbf{x}_{NOMA} can be expressed as:

$$\mathbf{x}_{NOMA} = (s_{NOMA,1}, \cdots, s_{NOMA,n_s}). \tag{10}$$

We assume the transmitter is equipped with a uniform linear array pattern of n_t antennas each have 0 dBi gain, and its precoding vector is in the form of:

$$\mathbf{p}_w = (w_1, \cdots, w_{n_t}). \tag{11}$$

 \mathbf{x}_{NOMA} is then multiplied with \mathbf{p}_w to spread the data according to the precoding weight as:

$$\mathbf{X}_{pre} = \mathbf{x}_{NOMA}^T \times \mathbf{p}_w,\tag{12}$$

where \mathbf{x}_{NOMA}^T is the transpose of the \mathbf{x}_{NOMA} . The matrix that represents the preceded signal from (12) can be expressed as follows:

$$\mathbf{X}_{pre} = \begin{bmatrix} s_{p,11} & s_{p,12} & \dots & s_{p,1n_s} \\ s_{p,21} & s_{p,22} & \dots & s_{p,2n_s} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p,n_t1} & s_{p,n_t2} & \dots & s_{p,n_tn_s} \end{bmatrix}, \quad (13)$$

These signals are then mapped into the resource grid and converted to OFDM-modulated signals using a series of signal processing blocks (e.g., IFFT, DAC), ready for radio transmission. The OFDM symbol matrix **X** can be expressed as:

$$\mathbf{X} = \begin{bmatrix} s_{o,11} & s_{o,12} & \dots & s_{o,1n_o} \\ s_{o,21} & s_{o,22} & \dots & s_{o,2n_o} \\ \vdots & \vdots & \ddots & \vdots \\ s_{o,n_t1} & s_{o,n_t2} & \dots & s_{o,n_tn_o} \end{bmatrix},$$
(14)

where the index

$$n_o = \frac{\text{Sampling Rate}}{\text{Non-Uniform Fast Fourier Transform (NFFT) size}},$$
(15)

is the frequency resolution in Hz.

B. Channel model

The most accepted 5G channel models between a transmitter and a receiver with MIMO and mMIMO transmissions are TDL and CDL models, respectively, as set by the ETSI TR 138 900 V14.2.0. We have considered both of these channel models in our system model. 1) TDL channel model: The TDL channel models are defined for a total frequency range from 0.5 to 100 GHz with a maximum bandwidth of 2 GHz. TDL model is useful in MIMO systems because it captures the effects of multipath propagation, and these channel models are categorized into TDL-A, TDL-B, and TDL-C to represent different non-line of sight (NLOS) channel profiles and TDL-D and TDL-E for line of sight (LOS) channel profile. The channel impulse response of a TDL channel, $h(t, \tau) \in \mathbb{C}$, for N_{tap} number of taps is given by

$$h(t,\tau) = \sum_{k=1}^{N_{tap}} a_k(t) \ \delta(\tau - \tau_k).$$
(16)

where $a_k(t)$ is the amplitude with τ_k delay and $\delta(\tau - \tau_k)$ is the Dirac delta function representing all signals with τ_k delay [35].

2) CDL channel model: The CDL models are defined for the same frequency range and maximum bandwidth and are more suitable for representing MIMO transmission with beamforming. Similarly, CDL-A, CDL-B and CDL-C represent the NLOS channel models, whereas CDL-D and CDL-E are used to represent the LOS channel models. Each CDL model can be scaled in delay and angles to achieve the desired RMS delay spread and angle spreads. The channel impulse response of the CDL channel is described in [36]

$$H_{u,s}(t,\tau) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left(a_{u,s,n,m} \sqrt{\frac{P_n}{M}} \,\delta(\tau - \tau_n) \right), \quad (17)$$

where u is the antenna element index of the receiver, s is the antenna element index of the transmitter, N is the number of clusters between the base station (BS) and the user equipment (UE), and M is the number of rays within each cluster. The gain coefficient $a_{u,s,n,m}$ is calculated as:

$$a_{u,s,n,m} = \mathbf{f}_{\mathbf{UE},u,n,m} \ \mathbf{f}_{\mathbf{BS},s,n,m},\tag{18}$$

where $\mathbf{f}_{\mathbf{UE},u,n,m}$ and $\mathbf{f}_{\mathbf{BS},s,n,m}$ denote the antenna gains of the antenna elements u and s in the UE and BS, respectively.

The channel correlation matrix **H** between n_o transmission paths and n_r received antenna can be written as:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n_o} \\ h_{21} & h_{22} & \dots & h_{2n_o} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_r 1} & h_{n_r 2} & \dots & h_{n_r n_o} \end{bmatrix}.$$
 (19)

Each element of \mathbf{H} combines signals of transmit antenna element using (16), indicating how related the signals received at different antennas.

C. Receiver design

The radio signal is received by n_r antenna elements, and the linear channel model Y can be defined as:

$$\mathbf{Y} = \mathbf{H} \times \mathbf{X} + \mathbf{N},\tag{20}$$

where N is the zero-mean variance- σ^2 additive white Gaussian noise (AWGN). The observation matrix Y is expressed as

$$\mathbf{Y} = \begin{bmatrix} s_{r,11} & s_{r,12} & \dots & s_{r,1n_o} \\ s_{r,21} & s_{r,22} & \dots & s_{r,2n_o} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r,n_r1} & s_{r,n_r2} & \dots & s_{r,n_rn_0} \end{bmatrix}.$$
 (21)

The receiver first applies a path filter to the received signal **Y**, synchronizing the time and mitigating the effects of varying path delays of **H**. Following channel equalization, the signal then proceeds to OFDM demodulation, passing through a channel filter to reduce noise and path gain. To extract NOMA symbols \mathbf{y}_{NOMA} as a vector from **Y**, the received signal is processed through timing synchronization, OFDM demodulation, channel estimation and equalization blocks, sequentially. The output from the channel estimation and equalization blocks \mathbf{y}_{NOMA} , is given by:

$$\mathbf{y}_{NOMA} = (s_{r,1}, \cdots, s_{r,n_s}). \tag{22}$$

The L1 and L2 data are required to be extracted from the received NOMA symbols y_{NOMA} . Unlike the transmitter chain, the receiver can not process both data layers in parallel due to the SIC operation, which requires the L1 data to be detected first. To perform the SIC, the L1 data is detected from y_{NOMA} , with the L2 data being treated as residual interference during this detection phase. Concurrently, y_{NOMA} is stored in a local buffer for L2 layer detection, which will subsequently be utilized for SIC during NOMA subtraction.

For the detection of L1, the y_{NOMA} undergoes processing via PDSCH demodulation (L1 layer) and produces encoded L1 bits as:

$$\mathbf{y}_{l1e} = (y_{l1e,1}, \cdots, y_{l1e,n_{dl1}}),$$
 (23)

where $y_{l1e,i} \in \{0,1\} \forall i$. The signal is then passed through the L1-DL-SCH decoder $(g_{e,l1})$, to remove channel coding and the received L1 data bits y_{l1} can be expressed as:

$$\mathbf{y}_{l1} \leftarrow g_{e,l1}(\mathbf{y}_{l1e}) \\ \mathbf{y}_{l1} = (y_{l1,1}, \cdots, y_{l1,n_{l1}}).$$

$$(24)$$

The received y_{l1} will be processed using the same DL-SCH and PDSCH encoding module used for L1 encoding at the transmitter. The reconstructed L1 layer symbols y_{rl1} can be expressed as:

$$\mathbf{y}_{rl1} = (s_{rl1,1}, \cdots, s_{rl1,n_s}).$$
 (25)

At this stage, both \mathbf{y}_{NOMA} and \mathbf{y}_{rl1} are input into the NOMA subtraction, which executes SIC by eliminating \mathbf{y}_{rl1} from \mathbf{y}_{NOMA} . If the reconstructed symbol $s_{rl1,i}$ is different from the original L1 transmit symbol, then the SIC operation adds additional interference in the L2 symbols, which is defined as non-residual interference. The result of this SIC operation is subsequently multiplied by (1/g) to compensate for the L2 power reduction implemented at the transmitter and given by:

$$\mathbf{y}_{l2s} = \frac{\mathbf{y}_{NOMA} - \mathbf{y}_{rl1}}{g}.$$
 (26)

The retrieved L2 symbols can be expressed as:

$$\mathbf{y}_{l2s} = (sl2_1, \cdots, sl2_{n_s}). \tag{27}$$



Fig. 2. An extended 5G transceiver model, incorporating an N-NOMA scheme. The shaded areas denote the proposed modifications for N-NOMA integration, allowing for the multiplexing of N data layers. The remaining components adhere to the standard high-level 3GPP 5G transceiver model.

Now, \mathbf{y}_{l2s} is demodulated using L2 PDSCH demodulation to extract encoded L2 bits, as

$$\mathbf{y}_{l2e} = (y_{l2e,1}, \cdots, y_{l2e,n_{dl2}}),$$
 (28)

which will be processed by L2 DL-SCH decoder $(g_{e,l2})$, to remove channel coding bits and retrieve L2 data. The L2 received bits can be expressed as:

$$\mathbf{y}_{l2} \leftarrow g_{e,l2}(\mathbf{y}_{l2e})
 \mathbf{y}_{l2} = (y_{l2,1}, \cdots, y_{l2,n_{l2}}).$$
(29)

III. N-NOMA-AIDED 5G NR-COMPLIANT PHY DESIGN

In this section, we expand the system model shown in Fig. 1 to incorporate N-NOMA as shown in Fig. 2. In this design, the N-NOMA multiplexing uses the same sequential combiner used in [26]. The proposed system model integrates N data layers into a single channel using NOMA multiplexing. Therefore, the total input for the system can be defined as a vector of size n_i for the *u*th data layer, represented as:

$$\mathbf{x}_u = (x_{u,1}, \cdots, x_{u,n_u}). \tag{30}$$

Each data layer is then processed according to its own DL-SCH and PDSCH configuration, as explained in Section II. The processed data symbols for the *u*th data layer can be expressed as:

$$\mathbf{x}_{su} = (s_{u,1}, \cdots, s_{u,n_{su}}). \tag{31}$$

Fig. 2 shows NOMA multiplexing using a sequential combiner technique, which multiplexes 2 lowest data layers, and then the combined signal is multiplexed with a third data layer, and so on. An N-NOMA transmission must execute n - 1multiplexing blocks following this technique. If the the power



Fig. 3. Depicts the one-shot N-NOMA integration in the 5G NR PHY, demonstrating the multiplexing of all data layers in a single block



Fig. 4. Latency comparison between the sequential combiner NOMA multiplexing and one-shot NOMA multiplexing, emphasizing the efficiency of the latter approach. (Simulation device: Intel core i7-13700K, 32 GB DDR5 (5600 MHz) RAM, NVIDIA RTX 4090)

ratio between *u*th and (u-1)th data layer is g_u , then we can express the combined NOMA symbols \mathbf{x}_{NOMA} as:

$$\mathbf{x}_{NOMA} = g_1(\mathbf{x}_{s1} + \dots + g_{n-1}(\mathbf{x}_{sN-1} + g_n \mathbf{s}_N))$$
$$= \sum_{i=1}^N \left[\prod_{j=1}^i g_j\right] \mathbf{x}_{si}.$$
(32)

The NOMA signal can now be processed for transmission according to the 5G PHY as described in the previous section.

Using a one-shot multiplexing technique, data symbols from all N layers can be seamlessly multiplexed, following (32). This one-shot multiplexer for N-NOMA transmission is shown in Fig. 3. To test the efficiency of this new multiplexer, we comprehensively evaluate the latency associated with both multiplexing systems, compared in Fig. 4. These findings indicate a substantial reduction in latency as additional data layers are incorporated into the system. For a 2-Layer system, the latency remains the same, as the process of 2-layer integration is identical in both systems. However, from three to higher layer multiplexing, the latency from the one-shot multiplexer is significantly lower than the sequential combiner multiplexer, as shown in Fig. 4.

From (32), the total power of the transmitted signal is the

sum of powers from each data layer, as:

$$P_{total} = \sum_{u=1}^{N} \left[P_t \prod_{i=1}^{u} g_i \right].$$
(33)

The data is transmitted using a single traffic channel following the transmitter model in Fig. 2 with the same principle of a 2-NOMA system. The received signal is initially synchronized and then demodulated to remove the OFDM components and corrected based on channel estimation while the NOMA symbols are as in (22). The data of the first layer is detected using standard NR PD-SCH demodulation and DL-SCH decoding blocks, as shown in Fig. 2. During this signal detection, no prior knowledge of NOMA data layers is required, and the sum power of all subsequent layers is treated as noise.

Detecting any *u*th data layer where u > 1 requires SIC to remove u-1 layer symbols from the received NOMA symbols. Therefore, the L_N data can not be detected before the L_{N-1} data is detected, and so on. This operation can be explained for *u*th data layer using (26). The generalized equation for *u*th layer SIC is written as:

$$\mathbf{y}_{su} = \frac{\mathbf{y}_{s(u-1)} - \mathbf{y}_{r(u-1)}}{g_u}, \ \forall \ 2 \le u \le N.$$
(34)

IV. BIT ERROR RATE ANALYSIS FOR N-NOMA SYSTEM

This section presents a closed-form error probability expression, incorporating MIMO gain, using the TDL model and accounting for residual and non-residual errors from SIC.

A. Error probability for M-QAM transmission

We begin with a standard symbol error function for a QAM-OFDM system with AWGN channel. The universal formula for the symbol error rate (SER) of a QAM-OFDM system can be expressed as [37]:

$$P_{rs} \le 4 Q \left[\sqrt{\frac{3 \gamma_s}{M-1}} \right], \tag{35}$$

where $Q[\cdot]$ denotes the Gaussian Q-function, and γ_s represents the SNR of the transmitted symbol, s. In the context of an M-QAM OFDM system over an AWGN channel with transmit power P_t , the upper bound of error probability is expressed as:

$$\widehat{P}_{rs} = 4 \left[\sqrt{\frac{3 P_t}{N_0(M-1)}} \right], \tag{36}$$

where N_0 denotes the noise spectral density.

B. Derivation of the effective channel gain G_u for 5G PHY MIMO systems

To factor in the effects of MIMO transmission, we now define the effective channel gain, denoted as G_u , for a PHY MIMO system. MIMO transmission channel employs n_t transmit antennas and n_r receive antennas, and we utilize the TDL channel model, as defined in (16) to calculate the spatial channel gain. Initially, we formulate the channel matrix **H** for $n_t \times n_r$ transmission. Subsequently, we apply singular value decomposition (SVD) to calculate the total channel gain, G_{total} , at receiver u_i . 1) Channel matrix formation: The process begins with constructing a $n_r \times n_t$ channel matrix **H**, where the element $h_{i,j}(t)$ represents channel response at time t, expressed as:

$$h_{i,j}(t) = \sum_{k=1}^{N_{tap}} a_{i,j,k}(t) \ e^{-j2\pi f_c \tau_k},$$
(37)

where f_c is the carrier center frequency, and $a_{i,j,k}(t)$ is the complex gain of the *k*th tap between the *i*th receive antenna and the *j*th transmit antenna at time *t*. The channel matrix **H** is expressed as:

$$\mathbf{H} = \begin{bmatrix} h_{1,1}(t) & h_{1,2}(t) & \dots & h_{1,n_t}(t) \\ h_{2,1}(t) & h_{2,2}(t) & \dots & h_{2,n_t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_r,1}(t) & h_{n_r,2}(t) & \dots & h_{n_r,n_t}(t) \end{bmatrix}.$$
 (38)

2) Total MIMO channel gain G_{total} using SVD: To derive the effective channel gain from H, we use SVD to approximate the eigendecomposition as follows:

$$\mathbf{H} = \mathbf{U} \cdot \boldsymbol{\Sigma} \cdot \mathbf{V}^H, \tag{39}$$

where U is an $n_r \times n_r$ unitary matrix, Σ is an $n_r \times n_t$ matrix with non-negative real numbers on the diagonal, and \mathbf{V}^H is Conjugate transpose of $n_t \times n_t$ unitary matrix. $diag(\Sigma) = \sigma_1, \sigma_2, \cdots, \sigma_{\min(n_r, n_t)}$, are the singular values of H and represent the effective channel gains for each spatial mode in the MIMO system.

We obtain the effective channel gains for all spatial modes in the MIMO system as follows:

$$G_k = |\sigma_k|^2, \quad k = 1, 2, \cdots, \min(n_r, n_t).$$
 (40)

To calculate the total effective MIMO channel gain, we assume transmit diversity and combine the coefficient gains as follows:

$$G_{\text{total}} = \sum_{k=1}^{\min(n_r, n_t)} G_k.$$
(41)

We next incorporate G_{total} into (36) to obtain the impact of MIMO transmission with transmit diversity as follows:

$$\widehat{P}_{rs} = 4 Q \left[\sqrt{\frac{3 G_{\text{total}} P_t}{N_0 (M-1)}} \right].$$
(42)

C. Symbol error probability for N-NOMA system with SIC

In an N-NOMA system that utilizes SIC, the total error probability for a layer is a function of both non-residual and residual errors that arise from the SIC. Let E_u be the decoding error event at the *u*th layer and C_u as the successful decoding event at the *u*th layer. Using Bayes theorem, the symbol error probability for the *u*th layer is therefore defined using the weighted sum of non-residual and residual error probability, which can be expressed mathematically as

$$P_{rs;u} = \underbrace{\Pr\{E_u | C_{u-1}\}}_{\text{Non-residual error}} \Pr\{C_{u-1}\} + \underbrace{\Pr\{E_u | E_{u-1}\}}_{\text{Residual error}} \Pr\{E_{u-1}\}.$$
(43)

1) Non-residual error probability: The impact of nonresidual noise on NOMA layers can be measured as the conditional error probability that the *u*th layer symbol is incorrectly decoded, given that the previous layer, (u - 1), has been successfully decoded. We then calculate the probability of a correctly decoded symbol in (u - 1) layer as $(1 - \hat{P}_{rs;(u-1)})$, and the interference from the non-residual error as the sum power of the subsequent NOMA layer. The non-residual symbol error probability for *u*th layer is derived in (45) by considering the transmit power of the *u*th layer

$$P_u = P_t \prod_{i=1}^u g_i, \tag{44}$$

extracted from (33), and the total interference (N_0 + non-residual interference).

power,

When u = 1, the first layer detection does not have any dependency on the SIC performance, and we can consider $\left[1 - \hat{P}_{rs;(u-1)} = 1\right]$. In equation (52), interference from other layers is accounted for in the denominator of the expression inside the $Q[\cdot]$ by assuming the interference from subsequent data layers to follow a Gaussian distribution. The interference term

$$\sum_{i=u+1}^{N} P_t \prod_{j=1}^{i} g_j,$$
 (46)

defines the cumulative interference from all other NOMA layers with an index greater than u. This term is added to the noise spectral density N_0 in the denominator, which improves the quality of symbol error probability estimation.

2) Residual error probability: We define $[\Pr\{E_u|E_{u-1}\}]$ as the residual error probability for uth layer, $\forall u > 1$ when the (u-1) layer symbol is decoded incorrectly. To determine the maximum impact of the residual error, we consider the maximum pair-wise distance between the transmitted and detected (u-1) layer QAM symbol. Assuming a SQAM constellation for (u-1) layer with transmit power P_{u-1} , the average symbol power P_{avg} is expressed as:

$$P_{avg} = \frac{2 P_t \prod_{i=1}^{u-1} g_i}{3 (M_{u-1} - 1)}.$$
(47)

The minimum pairwise distance, which is the distance between two adjacent symbols along the axis, denoted as d_{min} , is calculated from P_{avg} as follows:

$$d_{min} = \frac{2\sqrt{P_{avg}}}{\sqrt{M_{u-1}} - 1}.$$
(48)

Now consider an SQAM constellation of order M. The maximum pairwise distance is defined as the Euclidean distance between diagonally opposite QAM symbols. We can determine the distance between two edge points on the axis, denoted as d_{axis} , using $d_{axis} = d_{min}(\sqrt{M} - 1)$ as there are \sqrt{M} points along each axis, and the distance between adjacent points is d_{min} . Then we can derive the maximum pairwise distance, d_{max} , as follows:

$$d_{max} = \sqrt{2} \left(\sqrt{M_{u-1}} - 1 \right) d_{min}.$$
 (49)

$$\Pr\{E_u|C_{u-1}\}\Pr\{C_{u-1}\} = 4 Q \left[\sqrt{\frac{3 G_u P_t \prod_{i=1}^u g_i}{\left(N_0 + \sum_{i=u+1}^N P_t \prod_{j=1}^i g_j\right) (M_u - 1)}} \right] \left(1 - P_{rs;(u-1)}\right).$$
(45)

Then, we simplify the right term of the equation (49) as

$$2\sqrt{2P_{avg}}.$$
 (50)

Substituting the value of P_{avg} from (47) in (50) we get the maximum pairwise distance for u - 1 layer symbols as:

$$d_{max_{u-1}} = 4\sqrt{\frac{P_t \prod_{i=1}^{u-1} g_i}{3(M_{u-1} - 1)}}$$
(51)

 $d_{max_{u-1}}$ in (51) is the interference from the residual error for the *u*th NOMA layer. Assuming a Gaussian distribution for $d_{max_{u-1}}$, this term is added to the denominator with N_0 and residual interference in (52) leading to (45).

The symbol error probability for the *u*th NOMA layer is obtained from (43), incorporating both residual and non-residual error probabilities from (52) and (45), respectively. The first NOMA layer, independent from SIC, does not exhibit residual error. As such, the generalized expression for the SER for any NOMA layer in a QAM-MIMO-OFDM transmission in an AWGN channel is derived in (53), enhancing error estimation by accounting for interference from adjacent NOMA layers.

Assuming a Grey coding, we derive the BER upper bound from (53) as follows:

$$P_{b;u} = \frac{\widehat{P}_{s;u}}{\log_2 M_u}.$$
(54)

The total probability of error from all NOMA layers is then expressed as:

$$P_{bNOMA} = \sum_{i=1}^{N} P_{b;u}.$$
 (55)

V. RESULTS AND ANALYSIS

We present the outcomes of our comprehensive analysis and link-level simulations of 2-NOMA and N-NOMA system models using MATLAB. The simulation model uses the standard 5G NR PHY technologies, such as DL-SCH, PDSCH, LDPC, and MIMO precoding, in line with 3GPP TR 38.901 [34]. We incorporate 2- and multi-layer NOMA within the standard 5G transceiver framework for end-to-end link-level simulation to obtain specific KPIs (e.g., BER and throughput). The results of these simulations correlate the KPIs with transmission configuration metrics such as NOMA power ratio, modulation schemes, MIMO order, and LDPC coding rate with various code lengths. The OFDM carrier parameters are fixed as per Table I. We analyze the performance of the NOMA-MIMO-OFDM system according to the developed BER model in (54) and compare its performance with the simulation model, cementing the hypothesis that high-performance operation may be achieved by 5G PHY under low-SNR.

TABLE I Fixed carrier parameters

Carrier information	Value
Sub-carrier spacing	15 kHz
# resource blocks	52
# symbols per slot	14
# slots per frame	10
# slots per sub-frame	1

A. Performance comparison of NOMA L1 with a standard OMA transmission

The NOMA layer 1 is equivalent to an OMA transmission, with higher NOMA layers introducing additional interference. Figure 5 highlights the degradation as a result of additional layers on L1 performance, compared to a single-user OMA transmission. At -10 dB power ratio, the performance of L1 is almost the same as that of OMA transmission for both QPSK and 16QAM transmission. However, allocating higher power for L2 (-5 dB) causes significant performance degradation in L1 16QAM transmission as it achieves the same BER at a 3 dB higher SNR channel.

Although NOMA requires a higher SNR to attain the same BER, additional data layers can be transmitted using the same traffic channel. Fig. 5 shows that a low rate L1 is more robust and provides a higher degree of freedom in NOMA power distribution without significant performance degradation. This means L1 is more suited for low data rate PTM mMTC transmission, where the devices are distributed over large geographical areas with low or poor signal quality.

B. NOMA performance analysis

This subsection analyzes the performance of NOMA layers and the services that are achievable by individual layers. We conduct simulations on 2-layer and 3-layer NOMA transmissions to investigate the correlation between NOMA layer performance and various configuration parameters, including NOMA power ratio, modulation order, MIMO, and channel coding. As we add more layers to the NOMA system, increasing its complexity, this study reveals how the power ratio, choice of modulation, and MIMO configuration impact the performance of each NOMA layer. The results presented in this section illustrate these impacts, and the analysis presented here clearly demonstrates the degree of freedom in transceiver configuration to meet the required individual QoS of each receiver in the NOMA cluster.

1) NOMA power ratio vs throughput: Fig. 6 compares the throughput of NOMA layers for different power ratios in different L1 and L2 receivers' channel conditions. The throughput here is the ratio of successfully transmitted bits and the total transmitted bits per frame, which is the relative

$$\Pr\{E_u|E_{u-1}\}\Pr\{E_{u-1}\} = 4Q\left[\sqrt{\frac{3G_uP_t\prod_{i=1}^u g_i}{\left(N_0 + 4\sqrt{\frac{P_t\prod_{i=1}^{u-1} g_i}{3(M_{u-1}-1)}} + \sum_{i=u+1}^N P_t\prod_{j=1}^i g_j\right)(M_u - 1)}}\right]P_{rs;(u-1)}, \quad (52)$$

$$\widehat{P}_{s;u} = \begin{cases}
4Q \left[\sqrt{\frac{3G_u P_t \prod_{i=1}^u g_i}{(N_0 + \sum_{i=u+1}^N P_t \prod_{j=1}^i g_j)(M_u - 1)}} \right], & \text{if } u=1 \\
4Q \left[\sqrt{\frac{3G_u P_t \prod_{i=1}^u g_i}{(N_0 + \sum_{i=u+1}^N P_t \prod_{j=1}^i g_j)(M_u - 1)}} \right] (1 - P_{s;(u-1)}) \\
+ 4Q \left[\sqrt{\frac{3G_u P_t \prod_{i=1}^u g_i}{(N_0 + 4\sqrt{\frac{P_t \prod_{i=1}^{u-1} g_i}{3(M_u - 1)}} + \sum_{i=u+1}^N P_t \prod_{j=1}^i g_j})(M_u - 1)} \right] e^{(53)} \\$$



Fig. 5. Comparison of BER between OMA and L1 NOMA transmission under QPSK and 16 QAM modulation, utilizing an $[8 \times 1]$ MIMO order and (490/1024) LDPC code rate in a TDL-D channel model.

efficiency achieved during data transmission, as expressed in [38]:

Throughput (%) =
$$\frac{N_{Rx} - N_{eRx}}{N_{Tx}} \times 100,$$
 (56)

where N_{Rx} is the number of received bits, N_{eRx} is the number of erroneously received bits and N_{Tx} is the number of transmitted bits. Fig. 6 present a surface graph highlighting the impact of different power ratio on L1 and L2 throughput within the channel SNR range.

The first NOMA pair, depicted in Fig. 6(a,b), employs QPSK for both L1 and L2 with both receivers having the same channel SNR range of -10 to 0 dB. Fig. 6(a) demonstrates that L1 achieves the same throughput at a power ratio of -3 dB, compared to -10 dB, with nearly the same channel SNR. Specifically, L1 only requires a 1 dB higher SNR at the -3 dB power ratio to achieve the same throughput. Fig. 6 (b) shows improvement in L2 throughput as the power of L2 increases,

and in this SNR range, L2 achieves 100% throughput when the power ratio is higher than -5 dB. The first two subplots, Fig. 6-(a), (b), show a minimum power limit in L2 to achieve 100% throughput on a given SNR range. This minimum power requirement for successful data detection comes from the modulation power and order, which determines the Euclidean distance between modulation symbols.

The second NOMA pair, depicted in Fig. 6(c,d), employs 16 and 256 QAM for L1 and L2 with both receivers having a different channel SNR range of -5 to 5 and 5 to 15 dB. Both L1 simulation shown in Fig. 6 (a,c) shows that L1 achieves higher throughput as we decrease the power of L2 in NOMA multiplexing. This is because reducing the power ratio allocated to L2 relative to L1 improves the performance of L1 at a given SNR value, as L1 detection is independent of L2 detection and does not involve SIC. Fig. 6 (c) shows that at a power ratio over -4 dB, the L1 fails to detect its symbol even in high channel SNR. This failure to detect L1 also results in a drop in L2 throughput in the same power ratio, as shown in Fig. 6(d). This shows the maximum power that can be allocated to L2 as L1 and L2 fail to achieve high throughput at -3 dB power ratio. This is because if the interference from L2 to L1 is more than the maximum noise tolerance of L1 modulation, then L1 symbols can not be detected correctly, irrespective of the user's channel condition. As the L2 symbol detection depends on SIC, the L2 performance decreases despite a higher power allocated to the symbols.

Fig. 6 demonstrates the impact of power ratio on each layer's throughput, emphasizing the need for local optimization that considers the specific channel conditions, modulation, and MIMO order of individual NOMA layers to maximize the overall system performance. We have also observed that at a -5 dB power ratio, both layers achieved maximum throughput for all four different data rate combinations. Building on this



Fig. 6. Impact of power ratio on the throughput of L1: $[8 \times 1]$ and L2: $[8 \times 2]$ transmission with (490/1024) LDPC code rate and TDL-D channel model for two different data rate combinations (a-b) and (c-d): (a) L1-QPSK (-10 to 0 dB) (b) L2-QPSK (-10 to 0 dB) (c) L1-16QAM (-5 to 5 dB) (d) L2-256QAM (5 to 15 dB)

result, we focused on the -5 dB power ratio in the following analysis.

2) Modulation order of NOMA layers: We analyze various data rate transmissions, including QPSK, 16, 64, and 256 QAM for L2, and QPSK and 16QAM for L1, with MIMO configurations of $[8 \times 1]$ for L1 and $[8 \times 2]$ for L2. Fig. 7 shows that the BER of L1 remains consistent for both QPSK and 16 QAM, irrespective of the L2 modulation order, as all L1 BER lines are grouped together in both subplots. Similarly, the performance of L2 remains unaffected for QPSK and 16 QAM L1 transmissions. The performance of L1 and L2 has no dependency on each other's modulation order. This allows us to leverage any layer's improved channel condition by providing a high data rate service. This characteristic makes NOMA highly resilient in fulfilling diverse user requirements of mMTC and eMBB transmissions.

Fig. 7-(b) also reveals that for 16 QAM L1 and QPSK L2 transmissions, L2 performs better under low-SNR conditions due to higher MIMO order and lower data rate. The detection of L1 data SIC at L2 requires a lower SNR than at L1. This means that even when L2 is in worse channel conditions than L1, low data rate transmission of L2 is still feasible.

3) *MIMO:* Fig. 8 demonstrates the effect of MIMO order on the BER performance of L1 and L2. Given the resource constraints of mobile receivers, we assume one receiver antenna for L1 (representing mMTC devices) and two receiver antennas for L2 (representing eMBB devices). We used 1, 2, 4 and 8 transmit antennas to establish various MIMO orders.

The findings from Fig. 7, highlight the performance enhancement of L2 due to an additional receive antenna. These results echo the same potential for varying MIMO orders. In all four transmission scenarios depicted in Fig. 8-(a,b,c,d), the BER improves as the MIMO order increases. The BER of L1 improves by 6 dB for an $[8 \times 1]$ configuration from a $[1 \times 1]$ configuration for both data rate transmissions. Meanwhile, the BER of L2 improves by 8 dB, indicating that a higher MIMO order can significantly enhance the performance of lower orders. The higher MIMO order makes NOMA PTM transmission promising in downlink scenarios.

4) LDPC coding length: LDPC codes have been proven to be asymptotically optimal for wireless channels. However, practical issues of code rates across different NOMA layers should be examined closely. To quantify the impact of LDPC channel coding on the BER performance of NOMA, we used six different coding rates as shown in Fig. 9. We display the BER of NOMA layers L1, L2, and L3, altering only the coding rate of the respective layer while maintaining the other two at (490/1024). The code length impacts the BER of all three layers, but the impact is more pronounced in lower layers with higher-order modulation.

Fig. 9-(a,b,c) illustrates that the BER line follows the same curve for all code lengths, with shorter code lengths reducing the BER at lower SNR. The findings suggest that a shorter coding rate can enhance the noise tolerance margin, which can be used to offset additional interference from NOMA layers.

From Fig. 9-(b), we note that L2 fails to detect its signal when (850/1024) coding rate is used, while both L1 and L3 successfully detect signals under all coding rates. The inability to detect a signal even at high SNR points can be attributed to the events when the interference residual noise surpasses the maximum noise tolerance. In such cases, NOMA transmission of that layer and any subsequent lower layers will fail.



Fig. 7. Comparison of BER across various data rates in a 2-NOMA layer utilizing (490/1024) LDPC code rate, examined in a TDL-D channel



Fig. 8. Impact of MIMO configuration on BER performance of L1 and L2 for different data rate transmission with (490/1024) LDPC coding and -5dB NOMA power ratio

10

10

BER 10-

10

10⁻⁴ -20

-15 -10

(a) L1

150/1024

320/1024

490/1024

680/1024 850/1024

950/1024

10

10-4

-5 0



Fig. 9. Impact of LDPC coding rate on the BER of 3 NOMA $[8 \times 2]$ transmission with -5 dB power ratio between L1-L2 and L2-l3 using TDL-D channel: (a) L1-QPSK (b) L2-16QAM (c) L3-64QAM (d) L2-16QAM, L3-64QAM

10 15 20

SNR (dB)

10

10

10 15 20 25 30 35

SNR (dB)

We validate this hypothesis by simulating a different setup in (b), where we reduce the power ratio between L2 and L3 to -10 dB. The resulting BER performance with (850/1024) is 4 dB lower than with (680/1024) at a -5 dB NOMA power ratio. Based on these findings, one can optimize the performance to offset the residual error's impact, potentially leading to a successful multi-layer NOMA configuration.

0

-10 -5 SNR (dB)

Fig. 9-(b,c) shows that L3 with (150/1024) coding rate achieves an acceptable BER at a lower SNR than L2 with (680/1024) coding rate. This enables us to examine the effect of SIC and higher layer coding rate on lower layer performance, as depicted in Fig. 9-(d), where (150/1024) is used for L3, while (490/1024) and (680/1024) are used for L2 in *config* 1 and 2 respectively. The results reveal that the L3 receiver fails to detect signals at lower SNR when combined with the L2 (680/1024) coding rate. In this setup, the L3 BER line begins to fall after the L2 BER line dips. We also note that, for each code rate, the required SNR to achieve the same BER monotonically increases as the modulation order increases.

5) Channel Models: Figure 10 demonstrates the BER performance across different layers of NOMA under various channel conditions. Simulations utilized four distinct TDL and CDL channel models, capturing transmission in both urban and rural environments. The result shows consistency among all three NOMA layers. For instance, the TDL-A model required maximum SNR, whereas CDL-C achieved the same BER at minimum channel SNR. In our system model, NOMA achieves better performance in CDL channel environments.

In a study conducted by Iradier et al. [20], a 3-layered NOMA transmission was designed for 5G networks, accomplishing a BER of 10^{-4} at SNR of 11, 15, and 17 dB for

layers L1, L2, and L3, respectively. However, their work has not incorporated all the relevant 5G PHY technologies such as LDPC channel coding, MIMO, and CDL channel model. Our system model on the other hand incorporates all the 5G PHY technologies with N-NOMA, and their combined impact achieves 10^{-4} BER at -12 to -5 db SNR for L1, 5 to 12 db SNR for L2 and 15 to 22 db SNR for L3. These new performances represent substantial improvement over previous systems in the literature.

10

10

5

10 15

SNR (dB)

C. Theoretical BER evaluation of N-NOMA

Figure 11 provides a comparative analysis between the BER of a 3-NOMA system, following equation (54), and results obtained from a complete end-to-end simulation model, which incorporates an uncoded OFDM with MIMO AWGN channels designed for a 3-layer NOMA transmission. In this model, QPSK, 16QAM, and 64QAM are employed for L1, L2, and L3 layers, with 2, 4, and 8 receive antenna elements to capture the heterogeneity of devices in a NOMA cluster. However, our analytical and simulation models can accommodate a wide range of MIMO configurations. Power ratios for L2 and L3 are set at -5 dB and -10 dB.

As demonstrated in Fig. 11, theoretically determined BER closely matches with the simulated results, especially for higher NOMA layers, solidifying the validity of the derived equation. We also applied the Jaccard Similarity Indices to quantify the similarity between the theoretical and simulated BER, and our results show L2 curves match by about 0.90 while L3 gives 0.92.

Config 2

20 25



Fig. 10. BER for 3-layer NOMA in TDL and CDL channels. Transceiver: $[8 \times 2]$ MIMO, (490/1024) LDPC. Modulations: QPSK (L1), 16QAM (L2), 64QAM (L3).



Fig. 11. BER analysis illustrating the theoretical performance of a 3-layer NOMA system as per equation (54). The comparison is obtained using an end-to-end simulation model of the same system. The transmission layers are as follows: Layer 1 (L1) comprises an $[8 \times 2]$ QPSK, Layer 2 (L2) utilizes an $[8 \times 4]$ 16QAM, and Layer 3 (L3) deploys an $[8 \times 8]$ 64QAM.

VI. CONCLUSION

In this paper, we proposed a new 5G NR-compliant NOMA design, which incorporates a set of optimally placed NOMAaided transceiver functionalities, followed by a comprehensive performance evaluation. Technically, the transceiver allows us to quantitatively assess the performance of N-NOMA transmission, considering a series of parameters, including forward error correction, MIMO precoding, modulation schemes, and carrier specification, which opens new perspectives for the non-orthogonal access technology, particularly pointing at new optimization criteria and configuration dependency to the complex MIMO-NOMA detection process. To reduce transmission latency and transceiver complexity, we have simplified the multi-layer standard sequential combiners with a one-shot N-NOMA multiplexing technique. We have also developed a new and more accurate analytical BER expression to assess the impact of residual and non-residual SIC errors across NOMA layers on the error probability performance. Simulation-based evaluations have been carried out on a 5G NR-compliant link level simulator, which validated the proposed PHY design and the developed BER analytical results and also shown that N-NOMA has the potential to accommodate more heterogeneous users.

REFERENCES

 C. C. González, S. Pizzi, M. Murroni, and G. Araniti, "Multicasting over 6G non-terrestrial networks: a softwarization-based approach," *IEEE* Vehicular Technology Magazine, 2023.

- [2] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.
- [3] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, Č. Stefanović *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE access*, vol. 6, pp. 28 969–28 992, 2018.
- [4] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE signal processing letters*, vol. 21, no. 12, pp. 1501–1505, 2014.
- [5] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "A mathematical proof of the superiority of NOMA compared to conventional OMA," *arXiv preprint arXiv:1612.01069*, 2016.
- [6] A. Banchs, D. M. Gutierrez-Estevez, M. Fuentes, M. Boldi, and S. Provvedi, "A 5G mobile network architecture to support vertical industries," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 38–44, 2019.
- [7] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Nonorthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.
- [8] J. P. Lemayian and J. M. Hamamreh, "Recurrent neural network-based channel prediction in mMIMO for enhanced performance in future wireless communication," in 2020 International Conference on UK-China Emerging Technologies (UCET). IEEE, 2020, pp. 1–4.
- [9] Y. Yuan, S. Wang, Y. Wu, H. V. Poor, Z. Ding, X. You, and L. Hanzo, "NOMA for next-generation massive IoT: Performance potential and technology directions," *IEEE Communications Magazine*, 2021.
- [10] Y. Wu, B. Rong, K. Salehian, and G. Gagnon, "Cloud transmission: A new spectrum-reuse friendly digital terrestrial broadcasting transmission system," *IEEE Transactions on Broadcasting*, vol. 58, no. 3, pp. 329– 337, 2012.
- [11] M. Earnshaw, K. Shelby, H. Lee, Y. Oh, and M. Simon, "Physical layer framing for ATSC 3.0," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 263–270, 2016.
- [12] L. Zhang, W. Li, Y. Wu, K. Salehian, S. Laflèche, Z. Hong, S.-I. Park, H. M. Kim, J.-Y. Lee, N. Hur *et al.*, "Using layered-divisionmultiplexing to deliver multi-layer mobile services in ATSC 3.0," *IEEE Transactions on Broadcasting*, vol. 65, no. 1, pp. 40–52, 2018.
- [13] Y. Liu, W. Yi, Z. Ding, X. Liu, O. A. Dobre, and N. Al-Dhahir, "Developing NOMA to next generation multiple access: Future vision and research opportunities," *IEEE Wireless Communications*, 2022.
- [14] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE, 2013, pp. 611–615.
- [15] C. Regueiro, J. Montalban, J. Barrueco, M. Velez, P. Angueira, Y. Wu, L. Zhang, S.-I. Park, J.-Y. Lee, and H. M. Kim, "LDM core services performance in ATSC 3.0," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 244–252, 2016.
- [16] S. I. Park, J.-Y. Lee, S. Myoung, L. Zhang, Y. Wu, J. Montalbán, S. Kwon, B.-M. Lim, P. Angueira, H. M. Kim *et al.*, "Low complexity layered division multiplexing for ATSC 3.0," *IEEE Transactions on broadcasting*, vol. 62, no. 1, pp. 233–243, 2016.
- [17] L. Zhang, Y. Wu, W. Li, K. Salehian, S. Lafleche, X. Wang, S. I. Park, H. M. Kim, J.-y. Lee, N. Hur *et al.*, "Layered-division multiplexing: An enabling technology for multicast/broadcast service delivery in 5G," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 82–90, 2018.
- [18] J. Zhao, O. Simeone, D. Gunduz, and D. Gómez-Barquero, "Nonorthogonal unicast and broadcast transmission via joint beamforming and LDM in cellular networks," in 2016 IEEE Global Communications Conference (GLOBECOM). IEEE, 2016, pp. 1–6.
- [19] L. Zhang, W. Li, Y. Wu, Y. Xue, E. Sousa, S.-I. Park, J.-Y. Lee, N. Hur, and H.-M. Kim, "Using non-orthogonal multiplexing in 5G-MBMS to achieve broadband-broadcast convergence with high spectral efficiency," *IEEE Transactions on Broadcasting*, 2020.
- [20] E. Iradier, J. Montalban, L. Fanari, P. Angueira, L. Zhang, Y. Wu, and W. Li, "Using NOMA for enabling broadcast/unicast convergence in 5G networks," *IEEE Transactions on Broadcasting*, 2020.
- [21] Y. Xue, A. Alsohaily, E. Sousa, W. Li, L. Zhang, and Y. Wu, "Using layered division multiplexing for mixed unicast-broadcast service delivery in 5G," in 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2019, pp. 1–6.

- [22] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, 2015.
- [23] M. K. Naeem, R. Abozariba, M. Asaduzzaman, and M. Patwary, "Towards the mobility issues of 5G-NOMA through user dissociation and re-association control," in 2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). IEEE, 2020, pp. 427–432.
- [24] V. D. Khairnar and S. Pradhan, "Mobility models for vehicular adhoc network simulation," in 2011 IEEE Symposium on Computers & Informatics. IEEE, 2011, pp. 460–465.
- [25] S. Zhang, K. Peng, and J. Song, "Performance analysis and power allocation for spatial modulation-aided MIMO-LDM with finite alphabet inputs," *IEEE Transactions on Broadcasting*, 2022.
- [26] H. Kim, J. Kim, S.-I. Park, J.-y. Lee, S. Kwon, and N. Hur, "Capacity analysis and improvement of LDM-based Multiple-PLP configurations in ATSC 3.0," *IEEE Transactions on Broadcasting*, 2021.
- [27] S.-I. Park, J.-Y. Lee, B.-M. Lim, S. Kwon, J.-H. Seo, H. M. Kim, N. Hur, and J. Kim, "Field comparison tests of LDM and TDM in ATSC 3.0," *IEEE Transactions on Broadcasting*, vol. 64, no. 3, pp. 637–647, 2018.
- [28] D. Gómez-Barquero and O. Simeone, "LDM versus FDM/TDM for unequal error protection in terrestrial broadcasting systems: An information-theoretic view," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 571–579, 2015.
- [29] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1037–1071, 2022.
- [30] A. Akbar, S. Jangsher, and F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques," *Computer Networks*, vol. 190, p. 107950, 2021.
- [31] E. C. Cejudo, H. Zhu, and J. Wang, "Resource allocation in multicarrier NOMA systems based on optimal channel gain ratios," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 635–650, 2021.
- [32] S. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 40–47, 2018.
- [33] H. Yahya, E. Alsusa, and A. Al-Dweik, "Exact BER analysis of NOMA with arbitrary number of users and modulation orders," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6330–6344, 2021.
- [34] "5G; Study on channel model for frequencies from 0.5 to 100 GHz," 2020, 3GPP TR 38.901 version 16.1.0 Release 16. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/ 16.01.00_60/tr_138901v160100p.pdf
- [35] G.-R. Barb, M. Otesteanu, G. Budura, and C. Balint, "Performance evaluation of TDL channels for downlink 5G MIMO systems," in 2019 International Symposium on Signals, Circuits and Systems (ISSCS). IEEE, 2019, pp. 1–4.
- [36] A. M. Pessoa, B. Sokal, C. F. e Silva, T. F. Maciel, A. L. de Almeida, D. A. Sousa, Y. C. Silva, and F. R. P. Cavalcanti, "CDL-based channel model for 5G MIMO systems in remote rural areas," in 2019 16th International Symposium on Wireless Communication Systems (ISWCS). IEEE, 2019, pp. 21–26.
- [37] V. Meghdadi, "BER calculation," Wireless Communications, 2008.
- [38] MATLAB, "NR PDSCH Throughput," https://uk.mathworks.com/help/ 5g/ug/nr-pdsch-throughput.html, accessed: May 25, 2023.



Md Shantanu Islam (Graduate Student Member, IEEE) completed his BSc (Hons) in Electronic and Telecommunication Engineering, followed by an MS in Telecommunication Engineering, both at East West University in Bangladesh. Shantanu is pursuing his PhD degree at Birmingham City University, UK, in the College of Computing. He contributed to several projects, including as the lead researcher in the digital pathfinder project, developing automated systems for radio mapping of 4G/5G networks and was also involved the 5GCF. His current research

interests include future wireless physical layer architecture design, with an emphasis on NOMA, resource optimization and mobility management.



Raouf Abozariba (Member, IEEE) received a PhD degree from Staffordshire University in 2017 and was a Senior Research Associate with the School of Computing and Communications at Lancaster University, UK. He joined Birmingham City University, UK, in 2019 where he is currently a Senior Lecturer with the college of computing. He is frequently invited as a reviewer for many prestigious and first-tier journals in communications. He has led large-scale national research projects, such as a £0.26m UK project (5GRIT) trialling TV White Space tech-

nology to test the potential for shared spectrum radio to deliver 5G services to rural areas. Raouf was a Co-Investigator of the £1m DCMS-funded project, 5GCF, developing and evaluating new sensor technologies over 5G networks in forest environments, leveraging uncrewed vehicles such as robots and drones. His current research focuses on network QoS evaluation, NOMA technology, dynamic resource allocation and real-time communication for machine learning tasks and applications.



Dazhi He (Member, IEEE) is currently a Professor in the Cooperative Medianet Innovation Center of Shanghai Jiao tong University and Associate Editor of IEEE Transaction on Broadcasting. He has had in-depth research experience for the technology and standardization of terrestrial digital TV in China (DTMB), direct satellite broadcasting in China (ABS-S) and the newest global digital TV standard (ATSC3.0) . He had published more than 50 papers in the journal of IEEE and applied more than 70 patents (Including 20 PCT patents). He ever

won second prize of national scientific and technological progress award in China, first prize of the technology innovation award of National Radio and Television Administration in China, and first prize of science and technology progress award in Shanghai. His currently main research interests include 5G broadcasting, media communications, and heterogeneous network.



De Mi (Senior Member, IEEE) received the B.Eng. degree from the Beijing Institute of Technology, China, the M.Sc. degree from the Imperial College London and the Ph.D. degree from the University of Surrey, U.K.. He is currently an Associate Professor in future communication systems and the Head of the Future Information Networks (FINET) Research Cluster with the School of Computing and Digital Technology at Birmingham City University. He has also been a visiting professor at various universities and institutes, e.g., Shanghai Jiao Tong University,

Shenzhen Institute of Information Technology and Zhengzhou University, China. He was the recipient of the IEEE Broadcast Technology Society 2020 Scott Helt Memorial Award (Best Paper) and IEEE UCET 2020 best paper award. His research interests and expertise span a range of areas in wireless communications and signal processing, including B5G/6G radio access techniques, hypercomplex signal processing, next-generation broadcast and multicast communications and future open networks.



Mohammad N. Patwary (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical and electronic engineering from the Chittagong University of Engineering and Technology, Bangladesh, in 1998, and the PhD degree in telecommunication engineering from The University of New South Wales at Sydney, Sydney, NSW, Australia, in 2005. He was a Research and Development Engineer with General Electric Company, Bangladesh, from 1998 to 2000, and Southern Poro Communications, Sydney, from 2001 to 2002. He was a Lecturer with The University

of New South Wales at Sydney from 2005 to 2006 and then a Senior Lecturer with Staffordshire University, UK, from 2006 to 2010. He was then a Full Professor of wireless systems and digital productivity and the Chair of the Centre of Excellence on Digital Productivity with Connected Services at Staffordshire University until 2016. He was a Full Professor of telecommunication networks and digital productivity and the Head of the Intelligent Systems and Networks (ISN) Research Group, School of Computing and Digital Technology, Birmingham City University, UK, from 2017 to 2020. He was the research lead for the world's first '5G Connected Forest project, accelerating destination branding for the visitor economy in the UK. He also served as Principal Data Architect for a multi-city 5G testbed in the UK that aimed at accelerating digital productivity & develop an urban, connected community within the UK. He is currently a full Professor of Telecommunications and Director of the Centre for Future Networks & Autonomous Systems at the Faculty of Science & Engineering at the University of Wolverhampton, UK. His current research interests include wireless communication networks & systems design and optimization, signal processing and energy-efficient systems, intelligent & autonomous systems, the future generation of cellular network architecture and business modelling for Data-economy.



Taufiq Asyhari (Senior Member, IEEE) received the B.Eng. (with Hons.) degree in Electrical and Electronic Engineering from Nanyang Technological University, Singapore, in 2007, and the Ph.D. degree in Information Engineering from the University of Cambridge, U.K., in 2012. He is currently Professor of Data Science and Information Engineering with Monash University, (Indonesia Campus). He has been an Academic Visitor with Coventry University, U.K. and Cranfield University, U.K. He previously held full-time/visiting positions with National Chiao

Tung University, Hsinchu, Taiwan, Bell Laboratories, Stuttgart, Germany, and the University of Stuttgart, Germany. His research interests include the areas of information theory, communication and coding theory, and signal processing techniques with applications to wireless and molecular communication networks, the Internet of Things, and data analytics.

Dr. Asyhari is a Fellow of the Higher Education Academy, U.K. He has been the Member of the Editorial Board and the Technical Program Committee in numerous leading international journals and conferences. He is the Academic Lead for the multi-million pound 5G Connected Forest project funded by the U.K. DCMS and Lead Principal Investigator of British Council projects on digital forestry and environmental conservation. He was the recipient of the notable Samsung Global Research Outreach Award, in 2017, Silver Medal at the International Trade Fair iENA 2017, IEEE-EURASIP Best Paper Award at ISWCS 2014, National Science Council of Taiwan Starting Grant, in 2013, and the Cambridge Trust - Yousef Jameel Scholarship.