Enhanced User Fairness and Performance for eMBB-URLLC Uplink Traffic with Rate-Splitting based Super-positioning

Mayur Katwe , *Member, IEEE,* Keshav Singh, *Member, IEEE*, Chih-Peng Li , *Fellow, IEEE*, Shankar Prakriya, *Member, IEEE*, Bruno Clerckx, *Fellow, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—This paper investigates an unconventional superposition scheme, i.e., rate-splitting multiple access (RSMA) to maximize the overall user fairness and high system performance gain for ultra-reliable low-latency communication (URLLC), enhanced mobile-broadband (eMBB) traffic coexistence in uplink scenarios. In particular, we focus on maximizing the worst-case performance of uplink eMBB and URLLC users when multiplexed in a given resource block using an effective rate-splitting approach among multiple sub-messages. Subsequently, a multi-objective optimization problem (MOOP) is formulated to jointly maximize the worst-case rate and minimize the worst-case packet-error probability (PEP) for eMBB and URLLC users, respectively, using effective power splitting and successive interference cancellation (SIC) decoding of the sub-messages. To solve the non-convexity of the formulated MOOP, we adopt a priori articulation scheme combined with the weighted product approach to transforming the MOOP into a single objective optimization problem (SOOP) and later, solve it using a low complex differential evolution (DE)-based meta-heuristic algorithm. We derive an optimal decoding strategy for sub-messages to ensure better user fairness among eMBB-URLLC traffic. Numerical simulations demonstrate the superiority of the considered RSMA-based superposition for hybrid eMBB-URLLC traffic over conventional slicing and superposition techniques. Moreover, the adopted weighted product method-based DE algorithm outperforms the state-of-art solutions.**

***Index Terms*—Rate-splitting (RS), uplink (UL) communication, enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) traffic multiplexing, worst-case performance maximization.**

I. Introduction

FUTURE wireless networks are expected to bring a new paradigm of services beyond the capabilities of existing cellular architecture such as ultra-reliable low-latency communication (URLLC), enhanced mobile-broadband (eMBB) and massive-machine type communication (mMTC) [1], [2]. In general, the URLLC services render short packet for low data rate and emphasize low latency (e.g., below 1msec) communication with low packet error probability (PEP) (e.g., below 10^{-6}) constraints. On the contrary, eMBB services aim to capture high data rates up to 10 Gbps while relaxing latency and reliability constraints. The hybrid framework architectures involving the coexistence of these services, generally termed as multiplexing/ super-positioning, have gained significant research interest in the past few years as they cater to the quality of services (QoS) of heterogeneous traffic in terms of rate, latency, and reliability [2]–[5]. Nonetheless, the proliferation of heterogeneous traffic and the inevitable requirements of ultra-high spectral efficiency and reliability along with optimal user fairness brings formidable challenges for integrated catering of eMBB and URLLC traffic in next-generation wireless networks.

The most straightforward multiplexing approach involves orthogonal resource slicing which allocates distinct radio resources to each eMBB and URLLC user [2]. However, the inherited sporadic transmission characteristics of URLLC users may lead to wasteful spectral utilization in orthogonal slicing for a long duration and thus leads to poor spectral efficiency. As per 3GPP specifications, the fifth-generation new radio (5G-NR) accommodates pre-emptive or puncturing multiplexing which halts the current eMBB operation for short time duration and schedules the URLLC services on top of it [4], [5]. Despite its simplicity, the puncturing techniques encounter high rate-throughput loss and severe re-transmission overhead for eMBB users. Another multiplexing technique referred to as super-positioning, endeavors to serve eMBB and URLLC users in the same resource block (time-frequency) [6]–[9] through power-domain non-orthogonal multiple access (NOMA) scheme and successive interference cancellation(SIC) at the receiver. Typically, the super-positioning schemes significantly rely on effective inter-user interference management and sometimes may lead to sub-optimal performance and user fairness due to heavy traffic conditions. Notably, these schemes do not guarantee to impart a high degree of freedom (DoF) in performance gain due to their sub-optimal performance, and thus, they may not be

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC 112-2218-E-110-004 and NSTC 112-2221-E-110-029-MY3 and in part by the Sixth Generation Communication and Sensing Research Center funded by the Higher Education SPROUT Project, the Ministry of Education of Taiwan. (*Corresponding author: Keshav Singh*)

M. Katwe, K. Singh and C.-P. Li are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (E-mail: mayurkatwe@gmail.com, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw).

S. Prakriya is with the Department of Electrical Engineering, Indian Institute of Technology Delhi, India (E-mail: shankar@ee.iitd.ac.in).

B. Clerckx is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom (E-mail: b.clerckx@imperial.ac.uk).

G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece and also with Artificial Intelligence Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (E-mail: geokarag@auth.gr).

considered competent solutions in future wireless networks. Indeed, it is crucial to depart from the conventional approaches and pursue a spectral and power-efficient scheme for improved system performance in hybrid eMBB-URLLC traffic which constitutes the prime motivation of this work.

As a possible candidate, rate-splitting multiple access (RSMA) has recently been recognized as a promising solution due to its efficient spectrum utilization and a larger degree of freedom (DoF) in uplink (UL) as well as in downlink (DL) scenarios when compared to NOMA and orthogonal multiple access (OMA) schemes [10]–[12]. Particularly, the multi-layer UL RSMA flexibly splits the individual message of each user into multiple sub-messages and then transmits them using power-domain super-positioning [13], [14]. This unconventional approach of a configurable split in UL RSMA exploits the available resources and can ensure optimal DoF for the rate and PEP performance for eMBB and URLLC traffic without employing time scheduling among users, as opposed to conventional multiplexing [15]. Owing to its advantages of spectral efficiency, high user fairness, and low latent characteristics [10], [16], RSMA can be contemplated as a potential technique to ameliorate performance gain in hybrid eMBB-URLLC superposition algorithms.

A. Previous Works

The anticipated heterogeneous QoS in 5G-NR and beyond wireless systems has attracted tremendous research attention on eMBB-URLLC traffic multiplexing both in academia and industry [2]–[9], [17]–[24]. These works studied various multiplexing approaches that focused on the resource allocation design to improve key performance indicators (KPI)- rate for eMBB users PEP/latency or availability metrics for URLLC users. Precisely, they formulate an optimization problem seeking to maximize the performance of one service while fulfilling the provisions of the other. Many of these works such as [3]–[5], [22], [23] concluded that the performance amelioration for both the services is contradictory and the optimal DoF captivating a better trade-off between services becomes an challenging aspect regardless of any multiplexing scheme.

Besides, most recent works [2], [6], [7], [21] focused on spectrum slicing for hybrid OMA-NOMA architectures to achieve considerable diversity in the rate-throughput and PEP for eMBB and URLLC traffic, respectively. In general, the performance of NOMA becomes inferior to that of the puncturing method under high traffic and low channel-gain disparity conditions i.e., when the channel gains of users are similar [4]. Moreover, the fixed decoding order adopted in NOMA may not facilitate better system performance in asymmetric channel conditions [25]. An unconventional NOMA with time-sharing scheme was introduced in recent works [25]–[28] which resolves the issue of optimal DoF for NOMA especially in asymmetric channel conditions using time-sharing or time-scheduling approaches. However, it has issues of high implementation complexity and time-synchronization due to multiple time-slot transmission. For more competent hybrid traffic multiplexing, a few researchers have proposed to amalgamate puncturing and super-positioning techniques to harness their individual advantages [22]–[24]. Recent works showcase the merits of RSMA implementations with regard to high spectral efficiency, complexity, latency, robustness, reliability and other QoS metrics as compared to conventional multiple access techniques [11], [12], [16], [29]–[33]. There exist pioneer research works [14], [34]–[39] which address the captious problem of user fairness in UL scenario using RSMA. Interestingly, the diversified power-splitting in sub-messages and their optimal SIC decoding at the BS yield peculiar high-performance characteristics for both multi-layer and single-layer UL RSMA systems [14], [34]–[37]. For instance, the authors in [36], [37] showed that the optimal DoF w.r.t. rate can be attained in one-layer RSMA, and also derived the closed-form expressions for the outage probabilities for a set of near-far users. The work in [30] analyzed the superiority of single-layer RSMA over NOMA w.r.t rate and PEP subject to given QoS. However, their contribution was limited to only rate analysis of a two-user uplink URLLC scenario. Motivated by the benefits of RSMA, the authors in [39] explored the problem of spectral-efficiency maximization of URLLC users in presence of eMBB traffic using SIC decoding and resource slicing, the diverse DoF characteristics for eMBB and URLLC users w.r.t. rate and PEP were not highlighted. Further, the authors in [40] studied the different uplink network slicing techniques based on RSMA, NOMA, and OMA schemes and showed that one-layer RSMA renders better multiplexing for all core services due to its adjustable splitting power fraction. Nevertheless, the authors in [40] carried out a simplistic two URLLC and one eMBB user rate analysis. Moreover, the worst-case performance analysis of the rate-PEP region and effective power allocation scheme for eMBB and URLLC for better DoF is quite challenging and remains unaddressed throughout the literature.

B. Motivations

There exist several challenges associated with multiplexing eMBB and URLLC traffic such as

1) Latency Considerations: URLLC applications often have stringent latency requirements, demanding ultra-low latency communication. Multiplexing with eMBB services can introduce additional latency due to scheduling and complexities.
2) Interference management: The eMBB and URLLC multiplexing in the uplink can lead to increased interference between the two services. URLLC typically requires stringent latency and reliability, which can be adversely affected by interference from eMBB traffic.
3) Traffic prioritization: Prioritizing URLLC traffic over eMBB traffic is necessary to meet the strict latency and reliability requirements of URLLC traffic. However, this needs to be done without significantly impacting the throughput and quality of service (QoS) of eMBB traffic.
4) Multi-user access: The simultaneous transmission of eMBB and URLLC traffic by multiple users requires effective multiple access schemes that can ensure reliable and low-latency communication for both types of traffic.
5) Improved latency-reliability and rate trade-off: This is because achieving high reliability and low latency often

requires sacrificing some throughput, while achieving high throughput may require sacrificing some reliability or increasing latency.

6) Radio Resource Management: Efficiently managing radio resources for eMBB and URLLC in the multiplexed uplink is essential. This includes power control, and scheduling strategies that optimize the utilization of available resources while meeting the diverse requirements of both services.

Notably, orthogonal and puncturing techniques become more challenging in uplink scenarios due to the low latency and high spectral performance requirements of heterogeneous devices. Moreover, the existing super-positioning techniques such as NOMA rely on sophisticated user-pairing schemes and may not provide high user fairness. Interestingly, RSMA can potentially address these technical challenges associated with multiplexing eMBB and URLLC traffic. RSMA can effectively manage interference by splitting its original message into multiple sub-messages. Overall, RSMA-based superpositioning can optimize the trade-off between latency, reliability, and throughput by using adaptive rate splitting without traffic prioritization. This allows for dynamic adjustment of the transmission rate between the various sub-messages of the transmission based on the QoS requirements of the traffic type, ensuring that the best possible trade-off is achieved between latency, reliability, and throughput. Despite this, one of the main challenges in considering an RSMA-enabled eMBB-URLLC multiplexing is achieving efficient resource allocation, power control, and optimal decoding order strategy under specific QoS requirements which serves as a prime motivation of this work.

C. Contributions

On that account, we consider a two-layer UL RSMA system for hybrid eMBB-URLLC traffic multiplexing in this paper and study its effectiveness over conventional multiple access schemes in terms of worst-case achievable rate and PEP via effective power allocation and decoding order schemes. In contrast to previous research contributions, the prime objective of the considered hybrid eMBB-URLLC model is to simultaneously maximize the worst-case rate and minimize the worst-case PEP for eMBB and URLLC users respectively and thus ensure optimal fairness among them. The main challenge is to find an optimal point such that it provides the best trade-off among different performances requirement for eMBB and URLLC users. Specifically, enhancing the performance for eMBB users may negatively impact the performance of URLLC users and vice versa. Consequently, it becomes crucial to find an optimal solution that strikes the best trade-off between the different performance requirements of eMBB and URLLC users. This unique consideration of jointly optimizing the worst-case rate and worst-case PEP while accounting for the conflicting performance improvements due to inter user-interference in our work apart from previous studies in the field. By finding an optimal point that offers the best trade-off among the diverse performance requirements of eMBB and URLLC users, this study introduces a novel approach to tackle the challenges associated with the simultaneous optimization of these two critical parameters in a hybrid eMBB-URLLC scenario under given latency requirement.

To the best of our knowledge, this work represents the first attempt to address this specific eMBB-URLLC multiplexing problem. Existing works for eMBB-URLLC multiplexing have primarily focused on specific aspects such as user scheduling [5], spectrum allocation, or individual optimization of either eMBB user rate or URLLC performance [17], [23], [41] and some studies have explored the performance trade-off between rate and reliability in the context of hybrid eMBB-URLLC systems [2], [21]. However, our research distinguishes itself from these existing works by addressing a broader scope. We go beyond the traditional approaches and consider the simultaneous optimization of both the worst-case rate for eMBB users and the worst-case PEP for URLLC users. While RSMA offers advantages in terms of spectrum utilization and degrees of freedom, it also introduces challenges in resource allocation, i.e., power allocation, sub-message decoding, and computational complexity. This necessitates the development of advanced algorithmic designs, optimization techniques, and careful trade-off considerations. Notably, the major novelty of the proposed work lies not in designing a novel mathematical optimization framework, but rather in addressing the simultaneous maximization of the worst-case rate for eMBB users and the minimization of the worst-case PEP for URLLC users. The main contributions of this work are as follows:

1) We investigate a two-layer UL RSMA system for co-existing eMBB-URLLC traffic, analyzing the achievable rate-PEP regions under different power allocations and decoding orders. Our focus is on optimizing power allocation and SIC decoding order to enhance rate-throughput and reliability fairness while maximizing heterogeneous QoS for each user via effective management of inter-user interference.
2) We formulate a MOOP to maximize rate and minimize PEP for eMBB and URLLC users, respectively, with optimal power allocation and decoding order. To tackle the non-convex nature of the MOOP, we relax it into a single-objective problem (SOOP) using weighted product method and solve it with a low-complexity differential evolution-based algorithm.
3) Extensive simulations compare our proposed RSMA system with OMA and NOMA schemes, considering various parameters. Results demonstrate the effectiveness of our proposed solution over conventional schemes and the significance of decoding order selection in UL RSMA for achieving high-reliability, low-latency, and enhanced rate-throughput characteristics. Furthermore, the simulation results also show that the adopted weighted product to solve MOOP method significantly outperforms the other state-of-art method such as weighted Chebyshev method, weighted sum method, convex approximation and others.

II. Uplink RSMA for eMBB and URLLC Traffic

A. Assumptions in the Model and Analysis:

The considered system and the proposed analysis considered following assumptions:

1) Our analysis is carried out under the strict assumption of wideband transmission which is associated with the mmWave spectrum to counterbalance the path loss effects. Achieving a completely frequency-flat channel can be complex, and it is a pertinent factor that may affect the system's performance. However, advanced signal processing techniques, adaptive modulation and coding, and sophisticated antenna technologies may be employed to mitigate the effects of frequency-selective fading; however, these aspects are not the prime focus of this work.
2) The channel state information (CSI) of all channels involved is perfectly known at the BS [3]–[6], [40] and we consider a block fading scenario such that the channel condition is invariant over one complete subframe transmission time interval. The results in this paper serve as theoretical performance upper bounds for the considered system which can provide a benchmark for the system design under imperfect CSI.
3) For the sake of simplicity, we adopt matched-filter based beamforming at the BS for uplink transmission owing to its low-computational complexity. Although this approach may not be optimal in certain scenarios, such as high SINR, imperfect CSI estimation, or multi-user deployment, our main focus is on studying the spectral performance of the considered system in terms of rate and URLLC parameters.
4) Further, it is assumed that the BS can decode all the messages with perfect SIC such that all the sub-messages are decoded as per the defined decoding order. To validate the impact of SIC, the results are shown in the simulation section.
5) We focused on superposition multiplexing for the entire resource block with bandwidth B and mini time slot duration T. Alternatively, the URLLC users are assumed to be superimposed with eMBB users operating only for a resource block for duration T and with a set of sub-carriers (or carriers). Without loss of generality, it is assumed that the single transmission [1] for URLLC multiplexing is fixed using any contention scheme [44], [45].
6) In our analysis, we did not specify a specific modulation scheme due to the flexibility and applicability of the proposed system across various modulation schemes compatible with the 6G framework. Instead, our focus was on evaluating the performance of the system architecture, particularly in terms of resource allocation and scheduling within the context of a resource block and mini-slot framework, independent of modulation scheme intricacies.

[1]While it is acknowledged that retransmission mechanisms, such as automatic repeat request (ARQ) or others, can potentially enhance PEP [42], we deliberately exclude them in our study. Our rationale is rooted in the desire to provide a nuanced exploration of different superposition strategies. The fixed multiplexing duration allows us to focus on worst-case achievable PEP under different superpositioning scenarios, contributing valuable insights into the trade-offs associated with each strategy [43].

B. System Model

Let us consider a multi-user UL scenario where a set of K users communicate with a single-antenna BS. We consider two-layer[2] multi-user UL RSMA where each user splits its own message into two parts (sub-messages) and transmits them to the BS simultaneously i.e., at the same time and in the same frequency slot [15]. Then, the BS uses SIC to decode the sub-messages from all the users using a predefined decoding order.

The transmitted signal for any k^{th} user can be given for two-layer UL RSMA as [34], $x_k = \sqrt{p_{k1}}s_{k1} + \sqrt{p_{k2}}s_{k2}$, where $\{s_{k1}, s_{k2}\}$ is a set of sub-messages of the k^{th} user such that $\mathbb{E}\left[|s_{k1}|^2\right] = \mathbb{E}\left[|s_{k2}|^2\right] = 1$ and $\{p_{k1}, p_{k2}\}$ are their corresponding power allocation values. Note that the total transmitted power of each user is limited to $p_k^{\max}$, i.e., $p_{k1} + p_{k2} \leq p_k^{\max}$. So, the total received signal at the BS is given as

$$y = \sum_{k=1}^{K} h_k x_k + n = \sum_{k=1}^{K} h_k \left(\sum_{j=1}^{2} \sqrt{p_{kj}} s_{kj}\right) + n, \quad (1)$$

where h_k are the channel gain between BS and k^{th} user and $n \sim \mathcal{N}\left(0, \sigma^2\right)$ is a zero-mean complex Gaussian noise received at the BS with noise power σ^2.

Assuming that the SIC decoding order of sub-messages at the BS is denoted as the set $\boldsymbol{\pi} = \{\pi_{kj} : k \in \mathcal{K} \triangleq \{1, \ldots, K\}, j \in \mathcal{J} \triangleq \{1, 2\}\}$ in which the first element is decoded first, the second element is decoded second, and so on. $\pi_{kj} \in \mathcal{M} \triangleq \{1, \ldots, 2K\}$ denotes the decoding order of the sub-message s_{kj}. The permutation $\boldsymbol{\pi}$ belongs to the set $\boldsymbol{\Pi}$ which is the set of all the possible decoding orders of sub-messages. Particularly, for the sub-message s_{kj}, the BS successfully decodes and eliminates all the sub-messages that have a lower decoding order than s_{kj} and treats the remaining sub-messages as interference (other than s_{kj}). So, the signal to interference-noise-ratio (SINR) for the sub-message s_{kj} can be given as

$$\gamma_{kj}(\boldsymbol{p}, \boldsymbol{\pi}) = |h_k|^2 p_{kj} \Big/ \left(\sum_{(u,v)\in\mathcal{Q}_{kj}} |h_u|^2 p_{uv} + \sigma^2\right), \quad (2)$$

where $k \in \mathcal{K}, j \in \mathcal{J}$ and $\boldsymbol{p} = \{p_{kj}\}$ and $\mathcal{Q}_{kj}$ is a set of all the sub-messages which have greater decoding order than s_{kj} i.e., $\mathcal{Q}_{kj} = \left\{(u, v) : \pi_{uv} > \pi_{kj}\right\}$. In other words, the sub-messages from the set $\mathcal{Q}_{kj}$ becomes IUI for sub-message s_{kj}.

The maximum achievable rate for finite block-length coding can be approximately given as [46]

$$r_k = \frac{L_k}{T_k B} = \sum_{j=1}^{2} \left\{\log_2\left(1+\gamma_{kj}\right) - \sqrt{\frac{V(\gamma_{kj})}{T_k B}} \frac{Q^{-1}\left(\epsilon_{kj}\right)}{\log_e 2}\right\}, \quad (3)$$

where $k \in \mathcal{K}$, $V(\gamma_{kj}) \triangleq 1 - (1 + \gamma_{kj})^{-2}$ is the dispersion parameter, L_k indicates the number of intended transmitted bits for short-packet communication, T_k is the packet-transmission time of the k^{th} user, B is the transmission bandwidth and ϵ_{kj} is PEP for the k^{th} UL user. In practice, the block-length size and minimum required PEP vary for eMBB and URLLC devices.

[2]For multi-layer RSMA, when the transmitted signal is divided into two sub-messages, then the adopted system is called as two-layer RSMA.

C. eMBB Traffic

Let us consider that all users in the given network are eMBB devices. The maximum achievable rate expression for k^{th} eMBB user can be given as[3]

$$r_k^e \approx \sum\nolimits_{j=1}^{2} \log_2\left(1+\gamma_{kj}\right), \forall k \in \mathcal{K}, \tag{4}$$

The following lemma illustrates the criterion to achieve maximum achievable rate in the considered multiple access channel.

Lemma 1: Each eMBB user should operate at maximum transmit power budget to achieve maximum rate capacity (i.e., maximum sum-rate) in two-layer UL RSMA irrespective of any decoding order scheme.

Proof: Refer to Appendix A ∎

RSMA and NOMA both achieve high sum-rate [34], however, RSMA captivates the benefit of higher user fairness better as compared to NOMA. In order to validate this, we consider a two-user UL scenario and determine the rate-tuple for all possible power allocation and decoding orders of the sub-messages under given channel conditions as shown in Fig. 1a. The region OAPRQBO corresponds to the rate region of RSMA[4]. The yellow region represent the rate tuples with $p_{11}+p_{12} < p_1^{\max}$ and $p_{21}+p_{22} < p_2^{\max}$, while the, blue lines represent $p_{11}+p_{12} = p_1^{\max}$ and $p_{21}+p_{22} = p_2^{\max}$. Besides, the regions OAPP'O, OQ'QBO, and OABO are the rate region for NOMA with user-2 decoded first, NOMA with user-1 decoded first, and OMA, respectively. Clearly, the NOMA and OMA are subsets of RSMA and it also validates Lemma 1 as stated previously.

Interestingly, the variation of decoding order and power allocation to sub-messages allows RSMA to select the rate tuple on line segment PQ. On the other hand, NOMA can achieve only P and Q when user-2 is decoded first and second, respectively. The following theorem illustrates the optimal decoding order scheme for the two-layer RSMA scheme for enhanced user fairness and sum rate.

Theorem 1: The maximum worst-case rate between any pair[5] of eMBB users k and k' in multi-user scenario when decoded successively can be achieved in two-layer UL RSMA using the following decoding order scheme

$$\bar{\boldsymbol{\pi}} = \left\{(x_{1,1} \rightarrow x_{2,1} \rightarrow x_{1,2} \rightarrow x_{2,2}) : |h_1| \geq |h_2|\right\}, \tag{5}$$

subject to the maximum transmit power operation of each user. Defining $\underline{\boldsymbol{\pi}} = \{(x_{k'1} \rightarrow x_{k'2} \rightarrow x_{k1} \rightarrow x_{k2}), k \neq k'\}$ as the subset of decoding orders which does not belong to $\overline{\boldsymbol{\pi}}$, i.e., $\underline{\boldsymbol{\pi}} = \boldsymbol{\Pi}/\overline{\boldsymbol{\pi}}$.

[3]Without loss of generality, it is assumed that eMBB users implement long packet communication (with transmission time $T_k \geq 10$ msec) and with low reliability (with PEP $\geq 10^{-2}$). Hence, we neglect infinite block-length transmission scheme by neglecting the reliability term i.e., second term from achievable rate expression (3).

[4]The complete shaded regions (yellow +blue) indicate the performance region of RSMA

[5] Based on Theorem 1 and its proof in Appendix B, the decoding order for a multi-user scenario can be given as

$$\bar{\boldsymbol{\pi}} = \big\{(x_{1,1} \rightarrow x_{2,1} \rightarrow \ldots \rightarrow x_{K,1} \rightarrow x_{1,2} \rightarrow x_{2,2}) \ldots \rightarrow x_{K,2} : |h_1| \geq |h_2| \geq \cdots \geq |h_K|\big\}.$$

Proof: Refer to Appendix B ∎

In Fig. 1a, R corresponds to optimal point for given channel condition and decoding order $x_{11} \rightarrow x_{21} \rightarrow x_{12} \rightarrow x_{22}$ with power allocation $\boldsymbol{p}_1^*$ and $\boldsymbol{p}_2^*$ and this validates Theorem 1.

D. URLLC Traffic

Next, we consider a URLLC traffic scenario where all the devices implement short-packet transmission and analyze the PEP region for the two-layer RSMA system. In particular, the sub-message of each user is divided into short-packets with fixed coding rate and then the BS decodes all the messages using perfect SIC. We fix the achievable coding rate for each user by fixing the URLLC packet parameters, i.e., T_k and L_k, and then determine the minimum achievable PEP for each user under all possible decoding order and power allocation [47]. So, the minimum achievable PEP for k^{th} user under fixed coding rate can be formulated by rearranging (3) as [48]

$$\epsilon_{kj}^u = Q\left(\frac{\sqrt{T_k B}\log_e 2\left(\sum_{j=1}^{2}\log_2\left(1+\gamma_{kj}\right) - \frac{L_k}{T_k B}\right)}{\sqrt{V(\gamma_{kj})}}\right) \tag{6}$$

Fig. 1b illustrates the achievable PEP region for two-layer UL RSMA in UL scenario which is determined using all possible power allocation and decoding orders of the sub-messages under given channel conditions. The region OAPRQBO in Fig. 1b corresponds to the PEP region of RSMA in which the yellow shaded portion represents $p_{11}+p_{12} < p_1^{\max}$ and $p_{21}+p_{22} < p_2^{\max}$, while the blue shaded portion represents $p_{11}+p_{12} = p_1^{\max}$ and $p_{21}+p_{22} = p_2^{\max}$. The regions OAPP'O, OQ'QBO and OABO are the PEP region for NOMA with user-2 decoded first, NOMA user-1 decoded first and OMA, respectively. Thus, NOMA and OMA are subsets of RSMA and RSMA therefore outperforms NOMA and OMA in terms of minimum achievable PEP.

Also, the variation in selection of decoding order and power allocation of sub-messages allows RSMA to select the PEP on the curve PQ, while, the NOMA can achieve only P and Q. We provide the following theorem to illustrate the optimal decoding order scheme for two-layer RSMA scheme to achieve minimum worst-case PEP in URLLC traffic.

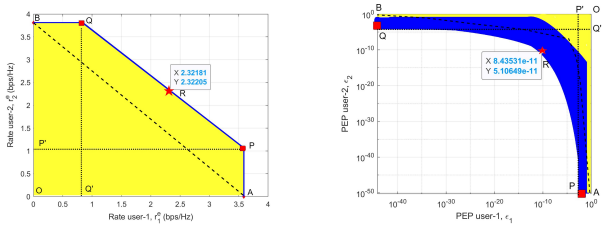
Theorem 2: The minimized worst-case PEP among the pair of URLLC users k and k' can be achieved w.r.t. PEP using $\overline{\boldsymbol{\pi}}$ as defined in (5) when compared to $\underline{\boldsymbol{\pi}}$ decoding order scheme subject to maximum transmit power operation of each user.

Proof: Refer Appendix C ∎

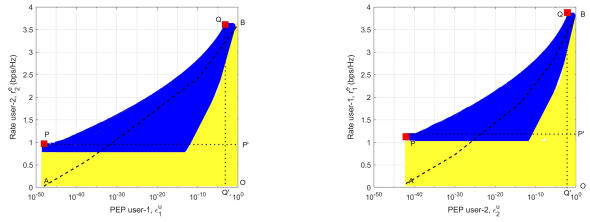
In Fig. 1b, R corresponds to the optimal point which attains the minimum worst-case PEP for given channel conditions and decoding order $x_{11} \rightarrow x_{21} \rightarrow x_{12} \rightarrow x_{22}$ with power allocation $\boldsymbol{p}_1^*$ and $\boldsymbol{p}_2^*$ (this validates Theorem 2).

III. System Model: Uplink RSMA for Hybrid eMBB-URLLC Traffic

Here, we extend our analysis for the more practical scenario of hybrid eMBB-URLLC traffic scenario where there exists K_e eMBB and K_u URLLC single-antenna users which simultaneously communicate with a multi-antenna BS using two-layer UL RSMA. Let us denote $\mathcal{K}_e = \{1, 2, \ldots, K_e\}$ and



(a) Rate Region (b) PEP Region



(c) u-1 = URLLC, u-2 = eMBB (d) u-1 = eMBB, u-2 = URLLC

Fig. 1: Achievable rate/PEP region for two-layer UL when $|h_1|^2/\sigma^2 = 11.14$ dB, $|h_2|^2/\sigma^2 = 10.79$ dB, $p_1^{\max} = p_2^{\max} = 1$ W.

$\mathcal{K}_u = \{K_e + 1, K_e + 2, \ldots, K_e + K_u\}$, as the set of eMBB and URLLC users, respectively, such that $K \triangleq K_e + K_u$.

Remark 1: In practice, acquiring accurate and timely CSI for URLLC communication poses challenges due to factors such as sporadic traffic, fast-changing channel conditions, and the stringent latency requirements of URLLC applications. For CSI estimation for URLLC users, one practical scenario involves the use of periodic channel measurements during idle or low-traffic periods with pilot signal transmission and sounding reference signals. However, this periodic estimation may not capture instantaneous variations in channel conditions during high-traffic or sporadic URLLC communication. Furthermore, the concept of "semi-persistent" scheduling can be considered, where certain resources are allocated persistently for URLLC communication. This ensures that even during periods of intermittent traffic, resources are available for sporadic URLLC transmissions. It is crucial to acknowledge that our assumption of perfect CSI is an idealization for analytical tractability, and we recognize that real-world implementations would need to incorporate sophisticated CSI acquisition strategies. We intend to include a discussion on the imperfect CSI and highlight its impact on the eMBB and URLLC user fairness in the simulation results.

The SINR at the k^{th} user in hybrid eMBB-URLLC traffic can be re-expressedas

$$\gamma_{kj}(\boldsymbol{p}, \boldsymbol{\pi}) = \frac{||h_k||^2 \, p_{kj}}{\underbrace{\sum_{(a,b)\in Q^e_{kj}} ||h_a||^2 p_{ab}}_{\text{IUI from } \mathcal{K}_e} + \underbrace{\sum_{(c,d)\in Q^u_{kj}} ||h_c||^2 p_{cd}}_{\text{IUI from } \mathcal{K}_u} + \sigma^2}, \tag{7}$$

where $k \in \mathcal{K}_e \cup \mathcal{K}_u$, $j \in \mathcal{J}$ and Q^e_{kj} and Q^u_{kj} are the set of all the sub-messages from the users belonging to $\mathcal{K}_e$ and $\mathcal{K}_u$, respectively, which have greater decoding order than s_{kj}. The rate and PEP expressions for eMBB and URLLC users are obtained in (3) and (6), respectively.

We now analyze the rate-PEP region of two-layer UL RSMA for hybrid eMBB-URLLC traffic with one eMBB and one URLLC user. Fig.1c illustrates the rate-PEP region when user-1 is URLLC and user-2 is eMBB, while, Fig. 1d illustrate the vice-versa scenario. For both scenarios, the blue shaded-region which corresponds to full power allocation i.e., $p_{k1} + p_{k2} \leq p_k^{\max}$ possesses better rate-PEP characterstics for eMBB-URLLC traffic than the yellow shaded region (i.e., $p_{k1} + p_{k2} < p_k^{\max}$). Hence, the user in hybrid eMBB-URLLC should operate with maximum power-budget to maintain rate and PEP fariness among eMBB and URLLC traffics, respectively. OAPP'O rate-PEP region is achieved with NOMA when the message of user 2 is decoded first. Similarly, OQ'QBO and OABO are the rate-PEP regions with NOMA user-1 decoded first and OMA, respectively. Specifically, NOMA and OMA are subsets of RSMA in terms of rate-PEP region and RSMA outperforms NOMA and OMA in terms of achievable rate-PEP region. Moreover, the variation in selection of decoding order and power allocation of sub-messages allows RSMA to select the rate-PEP on the curve PQ, while NOMA can achieve only P and Q. Conclusively, this induces the following theorem for decoding order to maximize worst-case performance for hybrid eMBB-URLLC traffic using the two-layer RSMA scheme.

Theorem 3: The optimal user fairness point between a pair of eMBB and URLLC users which corresponds to a better trade-off point for maximized worst-case rate and minimized worst-case PEP among eMBB and URLLC, respectively, in two-layer RSMA can be achieved using $\bar{\boldsymbol{\pi}}$ decoding order as defined in (5) subject to the maximum transmit power operation of each user.

Proof: Refer Appendix D ■

IV. Problem Formulation: Fairness Maximization

Generally, the URLLC system either focuses on latency minimization under fixed packet-error probability (PEP) or PEP minimization under latency minimization. Nevertheless, few works also focussed on the joint optimization of latency and PEP [49]. Primarily, minimization of both latency and PEP simultaneously for given rate requirements is quite contradictory as minimizing latency will increase PEP for given rate QoS requirements and vice versa. So, their joint minimization is impractical and there will always exist trade-offs in their joint minimization which may not be beneficial in many scenarios. For instance, let us consider a smart factory scenario, where a large number of sensors and machines are deployed to monitor and control the production process. These devices generate a massive amount of data, which needs to be transmitted to the central control unit with high reliability at a given specific interval.

The prime objective of the considered hybrid eMBB-URLLC model is to simultaneously maximize the worst-case rate and minimize the worst-case PEP for eMBB and URLLC users respectively and thus ensure optimal fairness among them. Since all the users adopt two-layer RSMA, the performance amelioration for both sets of eMBB and URLLC users conflict with each other due to IUI. To this end, a

optimization problem of worst-case performance maximization of eMBB-URLLC can be formulated as the following MOOP:

$$\begin{aligned}
&(\text{O1}): \max_{\boldsymbol{p},\boldsymbol{\pi}} \quad \min_{k\in\mathcal{K}_e} \{r_k^e(\boldsymbol{p},\boldsymbol{\pi})\} \\
&(\text{O2}): \min_{\boldsymbol{p},\boldsymbol{\pi}} \quad \max_{k\in\mathcal{K}_u} \{\epsilon_{kj}^u(\boldsymbol{p},\boldsymbol{\pi})\} \\
&\text{s.t.} \quad (\text{C1}): r_k^e \geq r_k^{\min}, \forall k \in \mathcal{K}_e, \\
&\qquad (\text{C2}): \epsilon_{kj}^u \leq \epsilon_k^{\max}, \forall k \in \mathcal{K}_u, \\
&\qquad (\text{C3}): 0 \leq p_{k1} + p_{k2} \leq p_k^{\max}, \forall k \in \mathcal{K}, \\
&\qquad (\text{C4}): \boldsymbol{\pi} \in \boldsymbol{\Pi}, \qquad (8)
\end{aligned}$$

where $\boldsymbol{p} = \{p_{k1}, p_{k2}\}$ is a set of power allocation values for all the sub-messages, r_k^e is the rate for the k^{th} eMBB user and ϵ_{kj}^u is the PEP for the k^{th} eMBB user which are expressed as (3) and (6), respectively. The objective functions (O1) corresponds to worst-case rate maximization for eMBB users, while (O2) corresponds to worst-case PEP minimization for URLLC users and the constraints (C1) and (C2) impose minimum QoS constraint for each eMBB and URLLC users with minimum rate threshold $r_k^{\min}$ and maximum PEP threshold $\epsilon_k^{\max}$, respectively. Besides, the constraint (C3) is the power allocation constraint for the sub-messages of any k^{th} user and the constraints (C4) restricts the decoding order of sub-messages into the set of optimal decoding order $\bar{\boldsymbol{\pi}}$.

Remark 2: We utilize max-min (or min-max) i.e., worst-case performance maximization as the definition of user fairness[6] among eMBB and URLLC fairness. While not a traditional fairness index, worst-case performance maximization can be applied to ensure that different features (or users) are on a similar scale, preventing one from dominating the others. Most commonly adopted proportion fairness maximization, which is generally considered for eMBB rate fairness [3], [52], may not be suitable when there are inherent trade-offs between the performance metrics of eMBB and URLLC users. On the other hand, worst-case rate maximization allows for a more flexible approach that considers the specific needs of each user type, addressing the heterogeneity in performance requirement which has been highlighted in the simulation section.

Remark 3: Remarkably, our analytical approach extends to scenarios focused on worst-case latency minimization and worst-case rate maximization for URLLC and eMBB users, respectively [55]. In such instances, the analysis aligns with grant-free URLLC and eMBB multiplexing, emphasizing the minimization of time-slot duration to facilitate uplink eMBB-URLLC transmission with minimal latency. However, it's important to note that the exploration of this intriguing problem is deferred to future research endeavors.

A. Feasibility Condition of the problem (8)

The feasibility condition for the optimization problem in (8) is discussed as follows.

[6]Different fairness indices, such as Jain's fairness index [50], entropy-based index (Theil index, and the Atkinson index) [51], max-min fairness [14], [48], [52], proportional fairness [3], [53], [54], Gini coefficient, etc, provide quantitative measures to assess the fairness of resource allocation among users for different scenarios. Overall, user fairness is a broad concept, and the choice of a specific fairness index depends on the context of the problem, the characteristics of the users, and the goals of the system.

Lemma 2: The problem in (8) is feasible only if the maximum transmit power budget for any k^{th} user (eMBB or URLLC) in hybrid eMBB-URLLC traffic satisfies

$$p_k^{\max} \geq \frac{\left(2^{q_k^{\min}} - 1\right) 2^{\left(\sum_{k'=k+1}^{K} q_{k'}^{\min}\right)} \sigma^2}{||h_k||^2}, \qquad (9)$$

$$q_k^{\min} = \begin{cases} r_k^{\min}, k \in \mathcal{K}_e, \\ \widetilde{\epsilon}_{kj}^{\min} \triangleq \frac{L_k}{TB} + \frac{(2\log_2 e)\, Q\left(\epsilon_k^{max}\right)}{\sqrt{TB}}, k \in \mathcal{K}_u, \end{cases} \qquad (10)$$

Proof: See Appendix E ∎

However, the MOOP in (8) is non-convex in nature as the decoding order constraint and strong coupling of power allocation in SINR expressions makes the problem of mixed-integer non-linear programming (MINLP). Consequently, the formulated MOOP in (8) is hard to solve and moreover, there exist no standard methods to solve it.

V. Proposed Solution

To tackle the multi-objective problem in (8), we first transform it into an equivalent single objective optimization problem (SOOP) using the weighted-product method [56] and later, the SOOP is solved using the differential evolution (DE) algorithm with given decoding order scheme.

A. Problem Transformation

First, we transform the minimization of PEP into an equivalent maximization problem. Since the Q-function is monotonic decreasing function, the worst-case minimization of PEP for any k^{th} user corresponds to the worst-case maximization of the q_k such that

$$\begin{aligned}
q_k^u(\boldsymbol{p},\boldsymbol{\pi}) &= Q^{-1}(\epsilon_{kj}^u) \\
&= \frac{\sqrt{T_k B}(\log_e 2)\left(\sum_{j=1}^{2}\log_2\left(1+\gamma_{kj}\right) - \frac{L_k}{T_k B}\right)}{\sqrt{V(\gamma_{kj})}}
\end{aligned}$$

While considering the maximum power allocation constraint to satisfy the feasibility conditions in (9) and (10), we reformulate the MOOP in (8) as follows

$$\begin{aligned}
&(\text{O1}): \max_{\boldsymbol{p},\boldsymbol{\pi}} \quad \min_{k\in\mathcal{K}_e} \{r_k^e(\boldsymbol{p},\boldsymbol{\pi})\} \\
&(\text{O2}): \max_{\boldsymbol{p},\boldsymbol{\pi}} \quad \min_{k\in\mathcal{K}_u} \{q_k^u(\boldsymbol{p},\boldsymbol{\pi})\} \\
&\text{s.t.} \quad (\text{C3}), (\text{C4}). \qquad (11)
\end{aligned}$$

such that the $p_k^{\max}$ is lower bounded as (9) and (10) which satisfy (C1) and (C2).

Now, the MOOP in (11) can be approximated as SOOP using weighted-product method [56] as

$$\begin{aligned}
&(\widetilde{\text{O}}): \max_{\boldsymbol{p},\boldsymbol{\pi}} \quad U_\omega \triangleq R^\omega Q^{(1-\omega)} \\
&\text{s.t.} \quad (\text{C3}), (\text{C4}), \qquad (12)
\end{aligned}$$

where $R = \min_{k\in\mathcal{K}_e}\{r_k\}, Q = \min_{k\in\mathcal{K}_u}\{q_k\}$ ω is a positive weighing coefficient that takes values between 0 and 1 ($0 \leq \omega \leq 1$). A large value of ω (i.e., $\omega > 0.5$) will prioritize the maximization of worst-case rate for eMBB users first,

while $\omega \leq 0.5$ will prioritize worst-case PEP minimization for URLLC users. Here, we fix ω and then determine the optimal power allocation and decoding order.

Remark 4: A most common method in prior preferences approach is weighted sum method in which transferring MOOP into a SOOP is accomplished by pre-multiplying each objective owing to its simplicity. A generic variant of the weighted sum method is the weighted product method. The weighted product method is often used for min-max optimization problems, where in the minimization (maximization) strategy objectives to be maximized (minimized) as a composite function [56]. Due to the distinct characteristics of objectives in terms of rate and reliability in the problem (11), both, weighted Chebyshev method and weighted sum method may not be the right choice for MOOP transformation. In particular, both weighted Chebyshev method and weighted sum method may not be suitable when the objective functions are inherently distinct and do not have a meaningful common scale or unit. In such cases, normalizing the objectives and comparing them using weights can lead to misleading results or incorrect trade-off assessments.

Following Theorem 3, we set the power allocation of each user to its maximum transmit power as $p_{k1}+p_{k2} = p_k^{\max}, \forall k \in \mathcal{K}$ and select a decoding order $\boldsymbol{\pi} \in \bar{\boldsymbol{\pi}}$ which guarantees the maximization of the worst-case rate-PEP performance. So, the SOOP in (12) can be further approximated as

$$\begin{aligned} (\widetilde{\mathrm{O}}): \max_{\boldsymbol{\alpha},\boldsymbol{\pi}} \quad & U_\omega \triangleq R^\omega Q^{(1-\omega)} \\ \text{s.t.} \quad & (\widetilde{\mathrm{C3}}): \alpha_k \leq p_k^{\max}, \forall k \in \mathcal{K}, \\ & (\widetilde{\mathrm{C4}}): \boldsymbol{\pi} \in \bar{\boldsymbol{\pi}}, \end{aligned} \tag{13}$$

where α_k is the power allocation coefficient for the first sub-message of k^{th} user and it signifies that $p_{k1} = \alpha_k, p_{k2} = p_k^{\max} - \alpha_k$.

The definition of Pareto optimality for the considered resource allocation for the problem (8) can be given as follows

Definition 1: An action profile , $\mathcal{A}^* \triangleq \{\boldsymbol{p}^*, \boldsymbol{\pi}^*\}$ which corresponds to set of distinct solution of decoding orders and power allocation, is set to be Pareto optimal if and only if there exist no other action profile, say $\widetilde{\boldsymbol{A}} \triangleq \{\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{\pi}}\}$, which gives better rate and reliability fairness, i.e., $\boldsymbol{U}_\omega(\boldsymbol{p}^*, \boldsymbol{\pi}^*) \geq \boldsymbol{U}_\omega(\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{\pi}})$.

In words, an action profile (or resource allocation) is Pareto optimal if there exists no other action profile that makes the performance of eMBB and URLLC users better without making the other users' performance worse. Basically, the Pareto front describes the set of efficient potential operating points, while the network designer is responsible for selecting the point that seems more appropriate for fulfilling the network requirements.

Lemma 3: Following from Theorem 3, the decoding order $\bar{\boldsymbol{\pi}}$ attains a set of Pareto-optimal front for the MOOP in (8) subject to the maximum transmit power operation of each user.

The bi-level optimization problem in (13) is still non-convex due to decoding order constraint. Primarily, any decoding order scheme among $\overline{\boldsymbol{\pi}}$ can achieve the optimal solution for the problem in (13). Nevertheless, the power allocation for the sub-messages will be different for different decoding orders. Hence, we solve only the power allocation problem under any given decoding order among $\overline{\boldsymbol{\pi}}$. However, the optimal power allocation policy is imperative to attain better performance gain for both the sets of eMBB and URLLC users.

B. Differential Evolution (DE) based Power Allocation

To solve the problem in (13), we propose a low-complexity DE-based meta-heuristic search algorithm[7] to attain optimal power allocation for all users under given decoding order from the decoding order set $\overline{\boldsymbol{\pi}}$, as defined in (5). In particular, the DE algorithm is an effective population-based meta-heuristic search algorithm that globally optimizes a problem over continuous spaces [58], [59]. It provides robust evolutionary process which aims to improve a candidate solution (population) in iterative manner under given quality of measure.

Let $\boldsymbol{P}^{(i)} \triangleq \left\{\boldsymbol{\alpha}^{(i,t)}, t = \{1, \ldots, N_P\}\right\}$ indicate a population, i.e., (a set of N_P individuals) at the i^{th} iteration (generation) where $\boldsymbol{\alpha}^{(i,t)}$ is the index of the t^{th} individual. Initially, the population with N_P number of power allocation coefficient for each user is randomly selected under the condition $(\widetilde{\mathrm{C3}})$ and later each individual is evolved using differential mutation. The complete procedure is described as follows:

Firstly, we select any two random individual $\boldsymbol{\alpha}^{(i,t_1)}$ and $\boldsymbol{\alpha}^{(i,t_2)}$ from the given population at the i^{th} iteration and generate the difference vector $\boldsymbol{\alpha}^{(i,t_1)} - \boldsymbol{\alpha}^{(i,t_2)}$.Unlike the classic DE algorithm, we avoid degenerate combinations by ensuring a distinct selection of individuals $t \neq t_1 \neq t_2$. Further, a mutant solution $\boldsymbol{\alpha}^{(i,t)}$ is produced by applying the scaled difference vector to the base vector. Instead of randomly choosing the base vector, we generate the base vector anywhere on the line segment between $\boldsymbol{\alpha}^{(i,t)}$ and $\boldsymbol{\alpha}^{(i,t^*)}$. where $\boldsymbol{\alpha}^{(i,t^*)}$ indicates the best point in the given population at the i^{th} iteration. Overall, the mutant solution can be expressed as

$$\widetilde{\boldsymbol{\alpha}}^{(i,t)} = \boldsymbol{\alpha}^{(i,t)} + \widetilde{\boldsymbol{s}}, \tag{14}$$

$$\widetilde{\boldsymbol{s}} = \zeta\left(\boldsymbol{\alpha}^{(i,t^*)} - \boldsymbol{\alpha}^{(i,t)}\right) + \left(\boldsymbol{s} \odot \left(\boldsymbol{\alpha}^{(i,t_1)} - \boldsymbol{\alpha}^{(i,t_2)}\right)\right), \tag{15}$$

$\zeta \in [0,1]$ and $\boldsymbol{s} \sim \mathcal{N}\left(s_o, \sigma_s^2 \boldsymbol{I}_K\right)$, and $\boldsymbol{s} \in \mathbb{R}^K$ is the scaling parameter modeled as multi-variable Gaussian random variable with mean $\boldsymbol{s}$ and covariance $\sigma_s^2 \boldsymbol{I}_K$. Note that the scaling parameter is dynamically selected[8] such that it performs a search operation with diverse set of possible mutant solution, and thus reduces the possibility of local convergence of the algorithm.

Now, the offspring population results are obtained using discrete combination operation between the current population

[7]The evolutionary algorithms are primarily utilized in situations when other usual methods fail to find the optimized state. For instance the commonly used gradient methods in convex approximation usually do not provide the global minimum when the cost function has a lot of local minima. In these cases the gradient methods are prone to converge to a local minimum and their result strongly depends on the choice of the starting point [57]. Moreover, the evolutionary algorithms excel at exploring a wide range of possible solutions.

[8]The spatial distribution of population varies with utility function which also vary the orientation and sizes of difference vector. The large scaling function may degrade the convergence and optimality [60]. Hence, for the scaling variable $\boldsymbol{s}$, the mean value $\boldsymbol{s}_o$ must be small to restrict large changes (alignment) in the angle of difference vector.

$\boldsymbol{P}^t$ and mutant population $\widetilde{\boldsymbol{P}}^t \triangleq \left\{\widetilde{\boldsymbol{\alpha}}^{(i,t)}, t = \{1, \ldots, N_P\}\right\}$ such that the

$$\bar{\alpha}_k^{(i,t)} = \begin{cases} \widetilde{\alpha}_k^{(i,t)}, k = k_t \text{ or } \mathcal{U}(0,1) \leq C \\ \alpha_k^{(i,t)}, \text{otherwise} \end{cases} \tag{16}$$

where $k_t \in \mathcal{K}$ is the random index which ensures that at least one of the decision variable inherits from the mutants and $C \in [0,1]$ is controlling parameter which controls the fraction of decision variables to be updated using mutant variables. Alternatively, in the first iterations of the algorithm, the controlling parameter $C = 1$, which enables a higher degree of exploration of the solution space by allowing the mutant solutions to prevail in the test population. As the algorithm progresses C is reduced, at a rate ρ, to enhance exploitation, favoring a more local search.

Finally, the utility function, U_ω is evaluated for each variable under given weight ω and the condition that i.e., the problem in (8) is feasible and the constraints (C1) and (C2) are satisfied. And, the individual for the next generation are determined as

$$\boldsymbol{\alpha}^{(i+1,t)} = \begin{cases} \bar{\boldsymbol{\alpha}}^{(i,t)}, U_\omega\left(\bar{\boldsymbol{\alpha}}^{(i,t)}\right) \geq U_\omega\left(\boldsymbol{\alpha}^{(i,t)}\right) \\ \boldsymbol{\alpha}^{(i,t)}, \text{ otherwise} \end{cases} \tag{17}$$

For each iteration, the best value among all the population, defined as $\boldsymbol{\alpha}^{i,t*}$, is selected such that it provides maximum utility function. This process continues for maximum N_I iteration or until is convergence is achieved (when there is no improvement in the evaluation function within certain tolerance[9] ξ or there is no improvement in the utility function over N_J generations). Algorithm 1 summarizes the proposed DE-based power allocation scheme.

Algorithm 1 DE algorithm for Power Allocation

Input: $N_I, N_J, \xi, C, s_o, N_p, \rho, \sigma_s^2,$ and **Initialize:** $i = 0, j = 0$
while $\left(\left|\max\left(\boldsymbol{\alpha}^{(i,t^*)} - \boldsymbol{\alpha}^{(i,(t-1)^*)}\right)\right| \geq \xi\right)$ && $(i < N_I)$ && $(j < N_J)$ **do**
  Determine the best value among the current population, $\boldsymbol{\alpha}^{(i,t^*)}$
  for $t = 1$ to N_P **do**
    Choose scaling parameter $\boldsymbol{s} \sim \mathcal{N}\left(s_o, \sigma_s^2 \boldsymbol{I}_K\right)$
    Calculate mutants $\widetilde{\boldsymbol{\alpha}}^{(i,t)}$ using (14) and (15)
    while $\widetilde{\boldsymbol{\alpha}}^{(i,t)}$ does not satisfy constraint (C3) **do**
      $\widetilde{\boldsymbol{s}} = 0.1\widetilde{\boldsymbol{s}}$,
      Calculate mutants $\widetilde{\boldsymbol{\alpha}}^{(i,t)}$ using (14) and (15)
    end while
    Choose any random index $k_t \in \mathcal{K}$
    Determine the offspring $\bar{\boldsymbol{\alpha}}^{(i,t)}$ using (16) an calculate the next generation $\boldsymbol{\alpha}^{(i+1,t)}$ using (17)
  end for
  $i = i + 1$
  if $((i \,\%100) == 0)$ **then** $C = \rho C$
  end if
  if $\left(j \geq 0.5 N_j\right)$ **then** $C = 1.1C$
    if $(C > 1)$ **then** $C = C$
    end if
  end if
  if $\left(\left|\max\left(\boldsymbol{\alpha}^{(i,t^*)} - \boldsymbol{\alpha}^{(i,(t-1)^*)}\right)\right| < \xi\right)$ **then** $j = j + 1$
  end if
end while
Output: $\boldsymbol{\alpha}^{(i,t^*)}$

C. Computational Complexity and Global Convergence Analysis of Algorithm 1

To assess the computational complexity, we can consider the number of arithmetic operations executed by the DE algorithm. In our case, the DE algorithm performs a total of $4K^2N_p + 12KN_p + 3$ arithmetic operations for N_I iterations, assuming that the algorithm converges within a maximum of N_I iterations. Therefore, the worst-case computational complexity of the DE algorithm can be expressed as

$$O\left(N_I\left(4K^2N_p + 12KN_p + 3\right)\right), \tag{18}$$

where N_p denotes the population size.

Notably, guaranteeing the optimality of the DE algorithm is challenging because it is a stochastic, population-based meta-heuristic that aims to find good solutions but does not provide guarantees of finding the globally optimal solution. The major reasons of sub-optimality of DE algorithms are local search, exploration-exploitation trade-off, parameter selection, and problem dependency [61]. However, the proposed DE algorithm involves significant modifications of classic DE algorithm w.r.t. stopping criteria, scaling/tuning paremeter and mutant generations which improve the overall likelihood of achieving near global optimal point for DE algorithm. Primarily, the convergence analysis of the DE algorithm can be conducted using theoretical analysis, empirical evaluation, or a combination of both approaches. In the case of the proposed DE algorithm, its convergence can be established by drawing parallels with the proof presented in [57, see Theoreem 6.1]. This provides a theoretical foundation for understanding the convergence behavior of the DE algorithm and its ability to reach a satisfactory solution. The empirical evaluation, on the other hand, entails running the algorithm on various problem instances and analyzing its convergence behavior based on predefined convergence criteria. To assess the optimality gap of the DE algorithm, the most prominent way is to compare the obtained solution's objective function value to known or benchmark solutions such as brute-force search (BFS) algorithm as explained shortly.

[9]Based on the numerical simulations, a tolerance value of $\xi = 10^{-5}$ was identified as providing better resolution for fine-tuning the convergence criterion. Note that the appropriate value of ξ may vary depending on the problem characteristics and the desired level of convergence accuracy.

VI. Numerical Simulation and Discussion

This section examines the performance behaviour of the proposed two-layer UL RSMA system over conventional schemes through extensive computer simulations. The simulation results are averaged over 2000 Monte-Carlo simulations.

A. Parameter Settings

For simulation scenario, we uniformly deploy $K_e = 3$ eMBB and $K_u = 3$ URLLC users in 100×100 m^2 region and fix the position of BS at the origin, i.e., $[0,0]m$. We adopted UMi-Street Canyon path loss model such that $PL(dB) = 32.4 + 20\log_{10} f_c + 31.9\log_{10} d$, where $PL(dB)$ represents the path loss in decibels, f_c is the carrier frequency, and d is the

distance between the communicating nodes. We consider the carrier frequency to be transmitted at the 28 GHz mmWave band with bandwidth of $B = 1$Ghz. To account for small-scale fading, we consider all channels as Rician distributed due to the presence of both LOS and NLOS components. Thus, the channels are expressed as $G = \sqrt{\frac{\kappa_f}{\kappa_f+1}}\boldsymbol{G}_L + \sqrt{\frac{1}{\kappa_f+1}}\boldsymbol{G}_N$, where $\kappa_f = 10$ is the Rician factor, $\boldsymbol{G}_L = 1$ is the LOS component and $\boldsymbol{G}_N$ is the NLOS component that follows Rayleigh distribution with parameter $\ell = 1$. The elements of G are multiplied by the square root of distance-dependent path-loss model. The noise power σ^2 is set to -100 dBm. The maximum transmit UL power[10] for all eMBB and URLLC users is set as $p_1^{\max} = p_2^{\max} = \cdots = p_K^{\max} = p^{\max} = 1$ W. The decoding order is set as $\boldsymbol{\pi} \in \bar{\boldsymbol{\pi}}$. The URLLC parameters, i.e., transmission time and rate of each URLLC user are set as $T_k = T_K = T = 5$ μsec and $\overline{L}_k = L = 40$ bits, where $k \in \mathcal{K}_u$. For the DE algorithm, we set number of populations $N_p = 20$, maximum controlling parameter $C = 1$, controlling parameter decrease rate $\rho = 0.9$, arithmetic recombination factor $\zeta = 0.5$, nominal value of scaling $s_o = 0.5$, maximum number of DE generations $N_I = 10000$, maximum number of generations without improvement $N_J = 1000$, tolerance $\xi = 10^{-5}$ and variance for scaling $\sigma_s^2 = 0.5$.

B. Baseline Schemes

We have considered following baseline schemes as the performance benchmarks for the comparative analysis with our weighted product method (WPM) based proposed DE algorithm for RSMA aided eMBB-URLLC system:

1) Proposed Solution (WSM): Here, we implement a weighted sum-method based MOOP transformation and adopt our proposed DE algorithm for power allocation in considered eMBB-URLLC UL RSMA system:
2) Proposed Solution (WCM): Here, we implement a weighted Chebyshev-method (WCM) based MOOP transformation and adopt our proposed DE algorithm for power allocation in considered system. The WCM is often utilize to solve the pareto-optimal problems [63].
3) Convex Approximation (WCM): Here, we implement a weighted Chebyshev-method based MOOP transformation and adopt equivalent convex relaxation approach for power allocation in considered system. The detail derivation of this schemes is provided in Appendix F. The methods to solve MOOP other than except WCM does not make the problem to be solve the MOOP [63]. Hence, to solve the MOOP using conventional method, we use WCM
4) NOMA: This scheme compares a counterpart NOMA based superposition for eMBB-URLLC system and weighted product method based proposed DE algorithm based power allocation. The RSMA with decoding order $\boldsymbol{\pi} \in \underline{\boldsymbol{\pi}}$ attains performance equivalent to NOMA as discussed earlier.
5) OMA: This scheme compares a counterpart OMA based slicing for eMBB-URLLC system and weighted product method based proposed DE algorithm-based power allocation. Here, we consider that each user is operating under fixed and uniform bandwidth (B/K).
6) Proposed Solution (PF+WSM): In particular, we compare the widely adopted proportion fairness [53] maximization with the considered worst-case maximization with under weighted sum approach.

We perform a thorough evaluation and provide detailed comparisons with these baselines to enhance the robustness and credibility of the considered RSMA-based hybrid eMBB-URLLC system and the proposed solution.

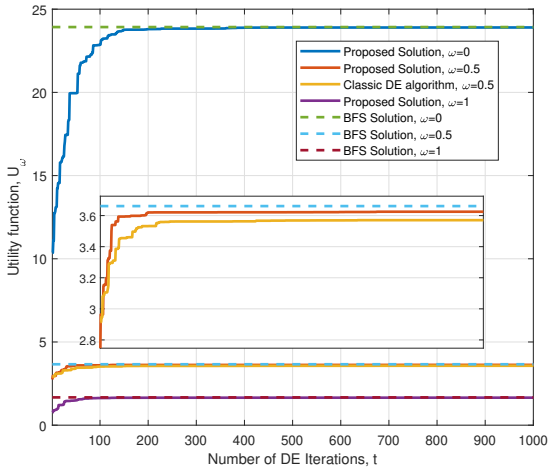
C. Convergence and Optimality Analysis

Firstly, we examine the convergence behaviour and optimality of the proposed DE algorithm for power allocation as shown in Fig. 2a. We illustrate the convergence behaviour of the weighted product method based DE algorithm for $\omega = 0, \omega = 1$ and $\omega = 0.5$ which corresponds to the scenarios of performance maximization for eMBB traffic only, URLLC traffic only and hybrid eMBB-URLLC traffic, respectively. We analyze the utility function w.r.t. varying number of generations (iterations) for DE algorithm. Fig. 2a validates the fast convergence of the proposed DE algorithm to a global stationary point which corresponds to the solution achieved by brute-force search (BFS) scheme. BFS executes $K^{S^{\max}}$ search operations and it complexity cost, $O\left(K^{S^{\max}}\right)$ rises exponentially with increase in users where $S^{\max}$ is the sample size for search operation. Although, the worst-case computational complexity for the proposed algorithm depends upon the popularization size, it computational cost is significantly lower than the BFS algorithm. Fig. 2a also confirms that proposed DE algorithm converges to the global point with substantially higher probability than classic DE algorithm. Overall the results validate that the adopted modifications in proposed DE algorithm closely attains the performance equivalent to BFS algorithm at the cost of reduced complexity.

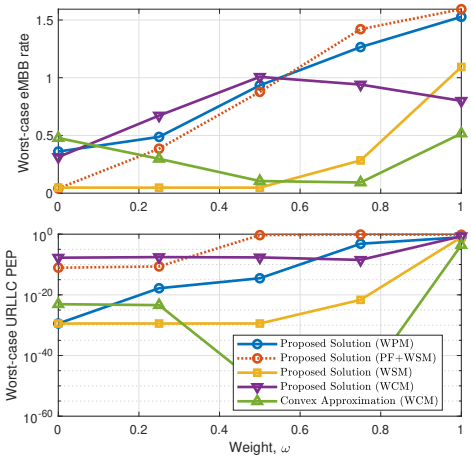
Next, we discuss the impact of the varying weight ω on the performance behaviour of proposed solution for the MOOP for system performance in hybrid eMBB-URLLC traffic. Fig. 2b shows that a low value of weight parameter ω in (12) improves the PEP for URLLC users at the expense of reduced eMBB rate. On the other hand, the high value of ω improves the performance for eMBB traffic as the at the expense of compromised URLLC performance. Consequently, the appropriate setting of ω leads to efficient trade-off between the performance of eMBB and URLLC traffic. Clearly, the appropriate selection of ω relies on the application requirement. The proposed solution with $\omega = 0.5$ can achieve the worst-case rate and PEP up to 1 Gbps and 10^{-10} for eMBB and URLLC users. In our proposed approach, we aim to enhance fairness for both eMBB and URLLC users. The term "both" underscores the inclusivity of our approach, emphasizing that it is designed to benefit both user categories. This is a critical

[10] The elevated transmission power in mmWave communication is often a result of the need to overcome higher propagation losses, leverage directional antennas, and ensure reliable links in challenging propagation environments. The specifications for mmWave transmission power are subject to ongoing research, standardization efforts, and regulatory considerations in the evolving landscape of wireless communication technologies [62].

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2024.3392929

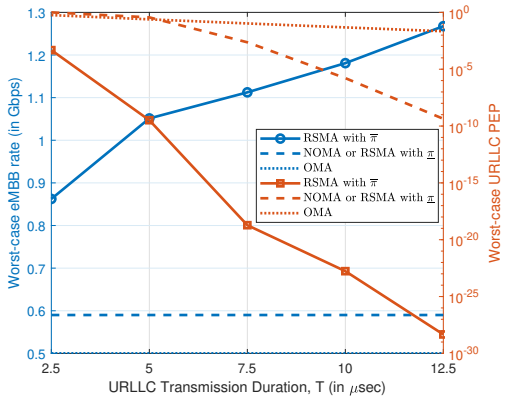


(a)

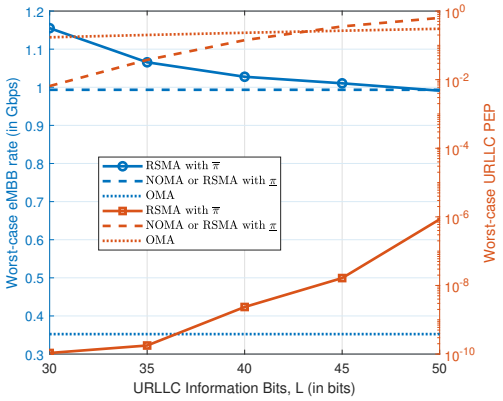


(b)

Fig. 2: a) Convergence analysis of the DE algorithm and b) Impact of weight on the system performance when $K_e = K_u = 3, T = 5$ μsec, $L = 40$ bits, $B = 1$ Ghz, $S^{\max} = 20$ and $p^{\max} = 1$W



(a) $L = 40$ bits



(b) $T = 5\mu$sec

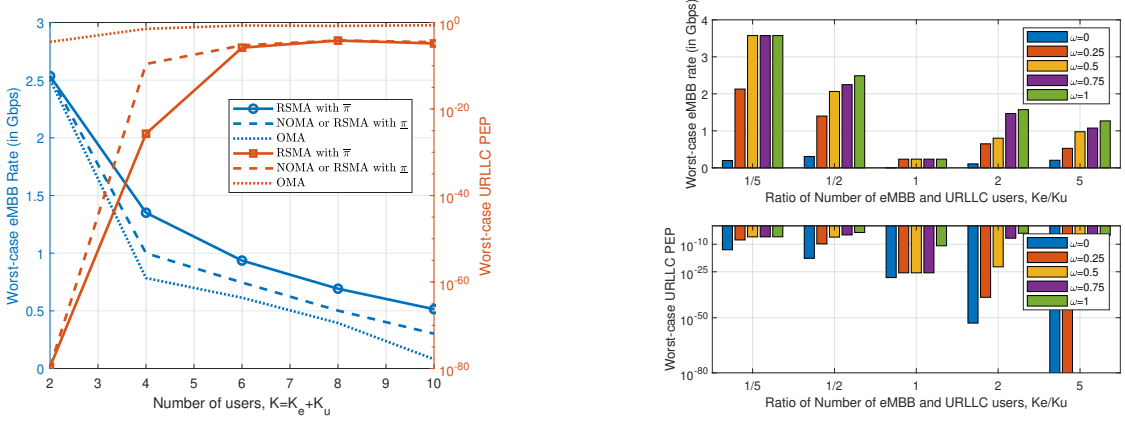
Fig. 3: Impact of URLLC packet a) transmission time and b) information bits on the system performance when $\omega = 0.5, K_e = K_u = 3$, $B = 1$ Ghz and $p^{\max} = 1$W

aspect of our methodology as it addresses the diverse and distinct requirements of eMBB and URLLC applications. The fairness enhancement is achieved by optimizing the allocation of resources, considering the specific needs and characteristics of both eMBB and URLLC users [17].

The results in Fig. 2b show the superiority of the proposed weighted product method over the weighted sum method and weighted Chebyshev method [56] in terms of user fairness for eMBB and URLLC users. The weighted product method explores a wider solution space, assigning higher weights to critical objectives using exponential weights. This allows for a more comprehensive search across the objective space, leading to diverse trade-off solutions. The proposed differential evolution-based search algorithm outperforms conventional convex relaxation solutions, which can degrade system fairness performance. Also, we examine the performance comparison of the proposed scheme (based on worst-case) and the counterpart proportional fairness maximization scheme in Fig. 2b. Overall, the rate performance for eMBB users and PEP performance for URLLC users proposed is better for our considered objective function when compared to the proportional fairness maximization. Clearly, the worst-case rate performance for both the schemes is quite close; however, the worst-case PEP performance gap corresponding to URLLC users between the two schemes is high. In fact, when a higher priority (higher weight) to rate performance is preferred, proportional fairness maximization is slightly better than worst-case rate maximization as it ensures good resource utilization among users of the same kind.

D. Impact of URLLC parameters

Fig. 3a and Fig. 3b examines the impact of varying URLLC parameters i.e., packet transmission time and intended data in URLLC packets on the system peformance in the hybrid eMBB-URLLC traffic, respectively. Intuitively, the increase in transmission time will improve the reliability performance of URLLC users for all schemes. However, the increase



(a) $\omega = 0.5, K_e = K_u = 3$ (b) $K_e + K_u = 6$

Fig. 4: Impact of a) uniform user density and b) selection of ω under uniform density on the system performance when μsec, $L = 40$ bits, $B = 1$ Ghz and $p^{\max} = 1$W.

in PEP improves the achievable rate-PEP region for hybrid eMBB-URLLC traffic in RSMA. In other words, the proposed algorithm selects the optimal power allocation strategy such that the performance gain for URLLC traffic is slightly relaxed to captivate better rate performance for eMBB traffic. Hence, the system performance point (rate-PEP) for hybrid eMBB-URLLC traffic is improved with the increase in URLLC packet transmission time as shown in Fig. 3a. Besides, the increase in packet-size (transmission bits) leads to higher error probability for all the schemes and hence the PEP[11] increases as shown in Fig 3b. In order to balance this PEP deterioration, the RSMA minimizes the worst-case PEP for the URLLC traffic at the cost of reduction in eMBB traffic. For both the scenarios, the RSMA with decoding order $\overline{\pi}$ provide significant performance amelioration for both eMBB and URLLC traffic in terms of rate and PEP, respectively when compared to NOMA and OMA. In RSMA, the users have the capability to transmit multiple sub-messages, and the system employs a power-splitting mechanism to effectively manage interference. This enables adaptive power utilization among sub-messages, ensuring optimal SIC and decoding order. The flexibility provided by RSMA in handling interference contributes to maintaining high QoS for each user, resulting in superior user fairness. Conversely, NOMA allocates the transmit power to a single message for each user, aiming to minimize interference in the desired UL signal communication. While NOMA is a valuable scheme, it can be considered a specific case of RSMA when the number of sub-messages is set to 1. The limitation of NOMA arises from its fixed power allocation strategy, which might not adapt as effectively to varying network conditions.

[11]Noteworthy, the obtained PEP values may not be feasible to realize in practice as they correspond to trivial/zero packet errors. Indeed, this may not be realistic, since there may exist some packet loss due to many unfavorable conditions such as weak signal strength, interference from other devices, multipath fading, channel conditions, noise, congestion, protocol design, wireless medium characteristics, environmental factors, and device mobility. The results in this paper serve as theoretical performance upper bounds for the considered system which can provide a benchmark for the system design while considering all the physical and real parameters for the packet loss.

E. Impact of user density

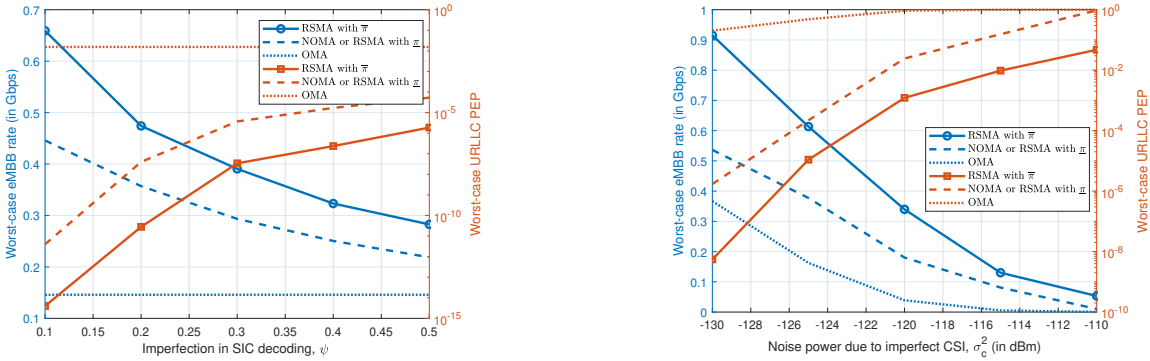
Further, we validate the performance of the behaviour of the proposed solution with varying user density of eMBB and URLLC traffic. Fig. 4a illustrates the impact of increase in users for both eMBB and URLLC users. As the number of users competing for the fixed resources increases, the IUI for each user increases which deteriorates the SINR for each transmitting user and hence the achievable rate-PEP region for hybrid eMBB-URLLC traffic is reduced. Fig. 4b clearly indicates the performance of the proposed solution for non-uniform traffic-density of eMBB and URLLC traffic. The lower weight value in the proposed solution is desirable when eMBB users are lower in number than URLLC users, while, a higher value of ω is preferable when the number eMBB users is higher than the number of URLLC users. It is due to the fact that the low and high value of ω in the proposed solution prioritizes the performance amelioration for eMBB and URLLC traffic, respectively, as validated in Fig. 4b. Hence, the appropriate (dynamic) selection of weight ω in the proposed solution ensures optimal system performance for dissimilar user number in the considered hybrid eMBB-URLLC traffic.

F. Impact of imperfect SIC and imperfect CSI

Importantly, performance of the proposed RSMA system relies on the perfect SIC decoding at the receiver. Here, we discuss the impact of imperfect SIC decoding operation on the performance of eMBB and URLLC users in hybrid multiplexing as shown in Fig. 5a. Under imperfect SIC decoding at each user (as in [64]), the SINR for the k^{th} user can be expressed for hybrid eMBB-URLLC traffic multiplexing as (19), shown on top of next page where $\eta_{k,a} = \psi, k > a, \eta_{k,a} = 1, k < a$ and $\eta_{k,c} = \psi, k > c, \eta_{k,c} = 1, k < c$ such that $\psi = 0$ indicates perfect SIC, while $\psi = 1$ refers to no SIC, and any value between 0 and 1 represents imperfect SIC.

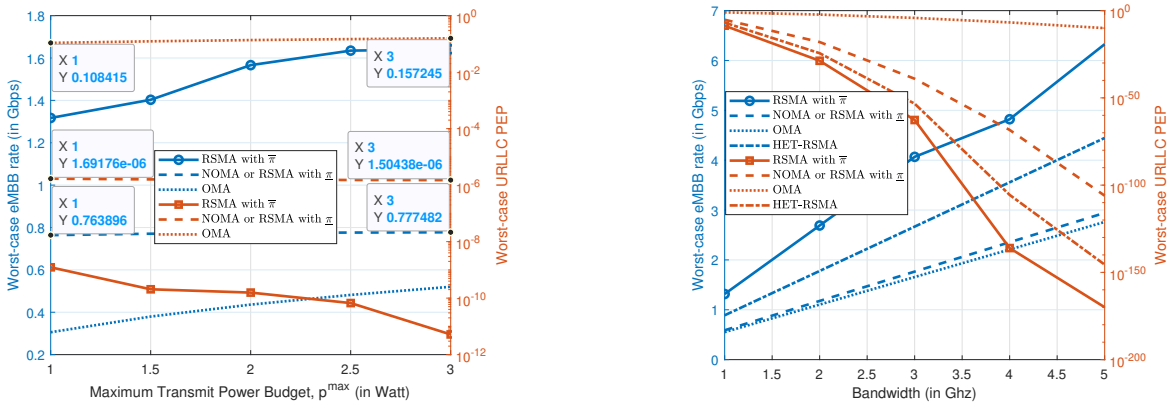
Based on (19), we now discuss the impact of imperfect SIC decoding operation on the performance of eMBB and URLLC users in hybrid multiplexing as shown in Fig. 5a. Intuitively,

$$\gamma_{kj}(\boldsymbol{p},\boldsymbol{\pi}) = \frac{||h_k||^2 p_{kj}}{\sum_{(a,b)\in\mathcal{Q}^e_{kj}} \eta_{k,a} ||h_a||^2 p_{ab} + \sum_{(c,d)\in\mathcal{Q}^u_{kj}} \eta_{k,c} ||h_c||^2 p_{cd} + \sigma^2}, \tag{19}$$



(a) (b)

Fig. 5: Impact of a) imperfect SIC decoding and b) imperfect CSI when $\omega = 0.5, K_e = K_u = 3, T = 5$ μsec, $L = 40$ bits, $B = 1$ Ghz and $p^{\max} = 1$W



(a) $B = 1$ Ghz (b) $p^{\max} = 1$W

Fig. 6: Impact of a) transmission power and b) transmission bandwidth on the system performance when when $\omega = 0.5, K_e = K_u = 3, T = 5$ μsec, $L = 40$ bits

the performance of RSMA and NOMA degrades with an increase in SIC imperfection ψ; however, the performance of OMA remains unaltered as it does not involve any SIC decoding. Notably, the performance of RSMA degrades much faster than NOMA with an increase in SIC imperfection. It is because the BS inherently performs $2K$ SIC decoding under the RSMA scheme; while the BS performs only K SIC decoding under the NOMA scheme. At lower SIC imperfection, the RSMA significantly outperforms NOMA in terms of rate and reliability. As the imperfection increases, the performance of RSMA decreases rapidly, and at higher imperfection, the performance of RSMA is closer to NOMA. Next, Fig. 5b illustrates the impact of imperfect CSI on all the considered schemes. For sake of simplicity, we model the imperfection in CSI as a zero-mean Gaussian random variable with the variance of σ_c^2. The results in Fig. 5b follow a similar trend as Fig. 5a. Conclusively, in a high noise environment, the performance of RSMA approaches the NOMA's performance as the channel noise power dominates the performance. Generally, a high imperfection in the SIC and CSI leads to higher performance degradation for superposition multiplexing approaches [4]. Nevertheless, the performance of RSMA is superior to OMA and NOMA throughout.

G. Impact of available resources

Fig. 6a and Fig. 6b depict the performance behaviour of the proposed solution with varying maximum transmit power and transmission bandwidth, respectively. It is obvious that increase in transmit power and transmission bandwidth for each user will increase the system performance for hybrid

eMBB-URLLC traffic in terms of both rate and PEP [12]. The increase in resource capability increases the SINR which in turn improves the rate and PEP for eMBB and URLLC users. Importantly, performance amelioration for RSMA is significantly better than for NOMA and OMA scheme. The performance improvement for NOMA and OMA with increase in transmit power is trivial as validated in Fig. 6a. Overall, the RSMA (with optimal decoding order) when compared to other conventional schemes achieves high spectral efficiency for eMBB uses and high reliability for URLLC users in accordance with efficient resource utilization. Consequently, the RSMA system with optimal decoding order brings promising potential for the core services in future wireless networks in terms of URLLC constraints (i.e., low latency and high reliability) as well as high data-rate.

VII. Conclusion

We investigated a two-layer UL RSMA system for hybrid eMBB-URLLC traffic to captivate high quality heterogeneous for each eMBB and URLLC users in terms of rate and reliability, respectively. In particular, we formulated a MOOP for coherent maximization and minimization of worst-case rate and worst-case PEP for eMBB and URLLC users, respectively. The formulated non-convex and (NP)-hard MOOP was firstly relaxed into a weighted product approach to relax the MOOP as SOOP and then solved it using a low complex differential evolution algorithm under a given decoding order. Simulation results validated fast convergence of the proposed DE algorithm. It was shown that the optimal resource allocation design in UL RSMA plays an exceptional role for rendering high-reliability, low-latency and enhanced rate-throughput characteristics. For instance, the worst-case rate and PEP were achieved up to 1Gbps and 10^{-20} for eMBB and URLLC users, respectively, in hybrid multiplexing using two-layer uplink RSMA with transmission bandwidth of 1GHz. Overall, the achieved results in this work demonstrated the effectiveness of considered RSMA system for hybrid eMBB-URLLC traffic over conventional orthogonal slicing and superposition multiplexing schemes.

Appendix A: Proof of Lemma 1

The sum-rate for UL RSMA system can be formulated as

$$r_t = \sum_{k=1}^{K} r_k^e = \sum_{k=1}^{K} \sum_{j=1}^{2} \log_2\left(1+\gamma_{kj}\right) \tag{A.1}$$

However, the sum-rate expression in (A.1) can be simplified as

$$r_t = \sum_{m=1}^{M} \log_2\left(1+\frac{|h_m|^2 p_m}{\sum_{m'=m+1}^{M}|h_{m'}|^2 p_{m'}+\sigma^2}\right). \tag{A.2}$$

where $M = 2K$, h_m and p_m corresponds to the channel gain and the power allocation of m^{th} sub-message in the decoding order where $m \in \{1,\ldots,M\}$.

[12]Indeed, the PEP values as low as 10^{-60} or 10^{-200} may seem unrealistic, and we want to clarify that these extreme values are illustrative of an ideal scenario where no packet error loss occurs, particularly in situations with higher bandwidths exceeding 1 GHz. These results are presented to provide insights into the performance implications of different bandwidths on rate reliability.

Using telescoping product, the sum-rate expression in (A.2) can be re-expressed as

$$\begin{aligned} r_t &= \log_2\left(1+\frac{\sum_{m=1}^{M}|h_m|^2 p_m}{\sigma^2}\right) \\ &= \log_2\left(1+\frac{\sum_{k=1}^{K}\sum_{j=1}^{2}|h_k|^2 p_{kj}}{\sigma^2}\right). \end{aligned} \tag{A.3}$$

It can be observed that the sum-rate expression in (A.3) does not depend upon decoding order. Moreover, the sum-rate expression in (A.3) is monotonically increasing function w.r.t. p_{k1} and p_{k2}. Hence, each user should operate with maximum transmit power budget to ensure maximum achievable sum-rate throughput irrespective of any decoding order scheme.

Appendix B: Proof of Theorem 1

Defining $\underline{\boldsymbol{\pi}} = \{(x_{k'j} \to x_{k'j'} \to x_{kj} \to x_{kj'}), k \neq k', j \neq j'\}$ as the subset of decoding orders which does not belong to $\overline{\boldsymbol{\pi}}$, i.e., $\underline{\boldsymbol{\pi}} = \boldsymbol{\Pi}/\overline{\boldsymbol{\pi}}$. Now, the achievable rate-throughput for any pair of users (k and k') when successively decoded under decoding scheme $\underline{\boldsymbol{\pi}}$ can be given as $r_k^{(\underline{\pi})} \in \{\underline{r}_k^{(\underline{\pi})}, \overline{r}_k^{(\underline{\pi})}\}$ and $r_{k'}^{(\underline{\pi})} \in \{\underline{r}_{k'}^{(\underline{\pi})}, \overline{r}_{k'}^{(\underline{\pi})}\}$, respectively, such that

$$\underline{r}_i^{(\underline{\pi})} = \log_2\left(\frac{|h_i|^2(p_{i1}+p_{i2})+|h_{i'}|^2(p_{i'1}+p_{i'2})+N_{ii'}+\sigma^2}{|h_{i'}|^2(p_{i'1}+p_{i'2})+N_{ii'}+\sigma^2}\right), \tag{B.1}$$

$$\overline{r}_i^{(\underline{\pi})} = \log_2\left(\frac{|h_i|^2(p_{i1}+p_{i2})+N_{ii'}+\sigma^2}{N_{ii'}+\sigma^2}\right), \tag{B.2}$$

where $i, i' \in k, k', i \neq i'$ and $N_{ii'}$ is the IUI from the users which are decoded after i and i' users. Since $p_{k1}+p_{k2}$ is the total transmit power, the rate-throughput expressions in (B.1) and (B.2) also correspond to the rate-throughput for NOMA scheme when k^{th} and $(k')^{th}$ users are decoded in ascending/descending order.

Now, let us consider decoding order belonging to $\overline{\boldsymbol{\pi}}$ such that sub-messages of k^{th} user are decoded first and then the sub-messages of $(k')^{th}$ user are decoded i.e., $(x_{kj} \to x_{kj'} \to x_{k'j} \to x_{k'j'})$. So, the rate expression for k^{th} and $(k')^{th}$ users can be given as

$$\begin{aligned} r_k^{\overline{\pi}} &= \sum_{j=1}^{2}\log_2(1+\gamma_{kj}) = \log_2\left(\frac{|h_k|^2 p_{k2}+|h_{k'}|^2 p_{k'2}+\sigma_{kk'}^2}{|h_{k'}|^2 p_{k'2}+\sigma_{kk'}^2}\right) \\ &\quad + \log_2\left(\frac{|h_k|^2(p_{k1}+p_{k2})+|h_{k'}|^2(p_{k'1}+p_{k'2})+\sigma_{kk'}^2}{|h_k|^2 p_{k2}+|h_{k'}|^2(p_{k'1}+p_{k'2})+\sigma_{kk'}^2}\right) \\ r_{k'}^{\overline{\pi}} &= \sum_{j=1}^{2}\log_2(1+\gamma_{k'j}) = \log_2\left(\frac{|h_{k'}|^2 p_{k'2}+\sigma_{kk'}^2}{\sigma_{kk'}^2}\right) \\ &\quad + \log_2\left(\frac{|h_k|^2 p_{k2}+|h_{k'}|^2(p_{k'1}+p_{k'2})+\sigma_{kk'}^2}{|h_{k'}|^2(p_{k'1}+p_{k'2})+\sigma_{kk'}^2}\right) \end{aligned}$$

where $\sigma_{kk'}^2 = N_{kk'}^2 + \sigma^2$.

Following Lemma (1), it can be obtained that when all users operate at maximum transmit power, i.e., $\sum_j p_{kj} = p_k^{\max}, \forall k \in \mathcal{K}$

$$r_k^{\overline{\pi}} + r_{k'}^{\overline{\pi}} = \underline{r}_k^{(\underline{\pi})} + \overline{r}_{k'}^{(\underline{\pi})} = \underline{r}_{k'}^{(\underline{\pi})} + \overline{r}_k^{(\underline{\pi})} = c = r_t - \widetilde{r}_{kk'}. \tag{B.3}$$

where c is the constant[13] value, r_t is the sum-rate and $\widetilde{r}_{kk'}$ is the rate of all i^{th} users such that $i \in \mathcal{K}/\{k,k'\}$.

Importantly, the expression (B.3) corresponds to a locus of an equation such that $y + x = c$. So, the amount of decrease in rate for k^{th} user from its maximum rate is equivalent to an increase in the rate for $(k')^{th}$ with same amount and vice-versa for decoding order $\overline{\boldsymbol{\pi}}$. On the contrary, the decoding order $\underline{\boldsymbol{\pi}}$ takes only fixed values of rate-PEP. So, a point can be achieved where $r_k^{\overline{\pi}} = r_{k'}^{\overline{\pi}} = 0.5c$ which maintain optimal fairness for both the users. Apparently, while performing SIC for multiple sub-messages at the receiver, the interference starts decreasing i.e., the interference for the first decoded sub-message of user k will be always higher than the second decoded message of k'. So, in order to ensure enhanced user fairness i.e., reduced rate-difference among any two users, the user $|h_k| \geq \left|h_k'\right|$. Overall, the decoding order $\overline{\boldsymbol{\pi}}$ can achieve maximum user fairness as compared to $\underline{\boldsymbol{\pi}}$.

Appendix C: Proof of Theorem 2

Following (B.1) and (B.2), it can be concluded that ϵ_{kj}^u can take only two possible values using (6), i.e., $\underline{\epsilon}_{kj}^{(\boldsymbol{\pi})}$ and $\overline{\epsilon}_{kj}^{(\boldsymbol{\pi})}$ for decoding order $\underline{\boldsymbol{\pi}}$ which is equivalent to NOMA scheme. However, ϵ_{kj}^u can take many possible values for decoding order $\overline{\boldsymbol{\pi}}$ when users are operated at maximum transmit power. Mainly $T_k = T_{k'}$ and $r_k^{\overline{\pi}} = r_{k'}^{\overline{\pi}}$ can achieve same value for ϵ_{kj}^u and $\epsilon_{k'}$ with the decoding order $\overline{\boldsymbol{\pi}}$. Hence, the decoding order $\overline{\boldsymbol{\pi}}$ under maximum transmission power constraint for each user power can achieve maximum fairness with better PEP.

Appendix D: Proof of Theorem 3

Following the proof of Theorem (1) and Theorem (2), it can be concluded that the rate of k^{th} eMBB user, r_k^e, and the PEP of $(k')^{th}$ URLLC user, $\epsilon_{k'}$, can take only two values when k^{th} eMBB and $(k')^{th}$ URLLC are decoded successively using $\underline{\boldsymbol{\pi}}$ under maximum transmit power, which is equivalent to the NOMA scheme. However, the rate of k^{th} user, r_k^e and PEP for $(k')^{th}$ user, $\epsilon_{k'}^u$ can take many possible values for decoding order $\overline{\boldsymbol{\pi}}$. From (B.3), an optimal fairness point among eMBB and URLLC can be achieved with $\overline{\boldsymbol{\pi}}$ when compared to $\underline{\boldsymbol{\pi}}$.

Appendix E: Proof of Lemma 2

Let us consider the worst-case scenario where the sub-messages are decoded with decoding order $\underline{\boldsymbol{\pi}}$. Using (6) and the constraint (C2) in (8), we can obtain

$$\frac{L_k}{TB} + \frac{\left(\sum_j \sqrt{V(\gamma_{kj})}\log_2 e\right) Q\left(\epsilon_k^{max}\right)}{\sqrt{TB}} \leq \log_2\left(1 + \frac{|h_k|^2 p_k^{\max}}{\sum_{k'\in\mathcal{K}/k}|h_{k'}|^2 p_{k'}^{\max} + \sigma^2}\right), k \in \mathcal{K}_u, \quad \text{(E.1)}$$

Under the assumption of high SINR, $\sum_j \sqrt{V(\gamma_{kj})}$ can be approximated to unity. This simplifies (E.1) as

$$\widetilde{\epsilon}_k^{\min} \triangleq \frac{L_k}{TB} + \left\{\frac{\left(\log_2 e\right) Q\left(\epsilon_k^{max}\right)}{\sqrt{TB}}\right.$$

$$\left.\log_2\left(1 + \frac{|h_k|^2 p_k^{\max}}{\sum_{k'\in\mathcal{K}/k}|h_{k'}|^2 p_{k'}^{\max} + \sigma^2}\right), k \in \mathcal{K}_u, \right\} \quad \text{(E.2)}$$

$$q_k^{\min} \leq \log_2\left(1 + \frac{|h_k|^2 p_k^{\max}}{\sum_{k'\in\mathcal{K}/k}|h_{k'}|^2 p_{k'}^{\max} + \sigma^2}\right), k \in \mathcal{K}, \quad \text{(E.3)}$$

where $q_k^{\min} = r_k^{\min}, \in \mathcal{K}_e, q_k^{\min} = \widetilde{\epsilon}_k^{\min}, \in \mathcal{K}_u$.

The last decoded pair sub-messages (either URLLC or eMBB) satisfy that

$$\sum_{j=1}^{2}\log_2\left(1+\gamma_{Kj}\right) = \log_2\left(1 + \frac{|h_K|^2 p_K^{\max}}{\sigma^2}\right) \geq q_K^{\min}, \quad \text{(E.4)}$$

which implies that

$$p_K^{\max} \geq \left(2^{q_K^{\min}} - 1\right)\sigma^2 \Big/ |h_K|^2. \quad \text{(E.5)}$$

Utilizing the lower threshold on the maximum transmit power for the last decoded K^{th} user derived in (E.5), we derive the lower threshold on the maximum transmit power constraint for the pen-ultimate i.e., $(K-1)^{th}$ user. To minimize the interference from the last decoded sub-messages and increase the overall signal strength, the minimum transmit power should be utilized by the last decoded sub-message. Therefore, the second last decoded sub-messages of user satisfy

$$\sum\nolimits_{j=1}^{2}\log_2\left(1+\gamma_{(K-1)j}\right) = \frac{\left|h_{(K-1)}\right|^2 p_{(K-1)}}{|h_K|^2 p_K + \sigma^2} \geq \gamma^{\min}, \quad \text{(E.6)}$$

$$\implies p_{(K-1)}^{\max} \geq \left(2^{q_{(K-1)}^{\min}} - 1\right)\sigma^2 2^{q_K^{\min}} \Big/ \left(\left|h_{(K-1)}\right|^2\right) \quad \text{(E.7)}$$

Similarly, the sub-message s_{kj} with $\pi_{kj} = M-2$ decoding order follows

$$p_{(K-2)}^{\max} \geq \left(2^{q_{(K-2)}^{\min}} - 1\right)\sigma^2 2^{q_{(K-1)}^{\min}} 2^{q_K^{\min}} \Big/ \left(\left|h_{(K-2)}\right|^2\right) \quad \text{(E.8)}$$

Thus, the lower threshold on the maximum transmit power for the sub-messages of any k^{th} user can be obtained using the induction method as

$$p_k^{\max} \geq \left(2^{q_k^{\min}} - 1\right) 2^{\left(\sum_{k'=k+1}^{K} q_{k'}^{\min}\right)}\sigma^2 \Big/ |h_k|^2, \quad \text{(E.9)}$$

which proves (9).

Appendix F: Weighted Chebyshev method and Convex optimization Framework

Under fixed decoding order (as per Theorem 3), the MOOP in (11) can be transformed into SOOP using weighted Chebyshev method [56] as

$$\begin{aligned}(\widetilde{\mathrm{O}}_c): \max_{\boldsymbol{p},t} \quad & t \\ \text{s.t.} \quad & \text{(C3)}, \\ & \text{(C5)}: \omega\left(R - R^*\right) \geq t, \\ & \text{(C6)}: (1-\omega)\left(Q - Q^*\right) \geq t, \end{aligned} \quad \text{(F.1)}$$

$R\left(\boldsymbol{p}, \boldsymbol{\pi} = \overline{\boldsymbol{\pi}}\right) = \min_{k\in\mathcal{K}_e}\{r_k\}, Q\left(\boldsymbol{p}, \boldsymbol{\pi} = \overline{\boldsymbol{\pi}}\right) = \min_{k\in\mathcal{K}_u}\{q_k\}$ and ω is a positive weighing coefficient that takes values

[13]Under given fixed decoding of all $i \in \mathcal{K}/\{k,k'\}$, it is straightforward to consider $r_k^{\overline{\pi}} + r_{k'}^{\overline{\pi}}$ as constant.

between 0 and 1 ($0 \leq \omega \leq 1$) and R^* and Q^* are the optimal objective which are assumed to be known. Note the optimization problem $(\widetilde{\mathrm{O}}_c)$ in (F.1) is non-convex owing to its constraints. To make it tractable, we approximate the dispersion function $V(\gamma_{kj})$ by 1, i.e., $V(\gamma_{kj}) \approx 1, \forall k$ and $\forall j$. Owing to the consideration of high SINR regime, the minimum achievable PEP can be approximated as

$$\epsilon_k^u \approx \widetilde{\epsilon}_k^u = Q\left(\sqrt{T_k B}\log_e 2\left(\sum\nolimits_{j=1}^{2}\log_2\left(1+\gamma_{kj}\right)-\frac{L_k}{T_k B}\right)\right) \quad \text{(F.2)}$$

Ultimately, the constraint (C6) can be rewritten as $(1-\omega)\left(\widetilde{Q}-Q^*\right)$, where

$$\widetilde{Q} = \min_k\left\{\sqrt{T_k B}\log_e 2\left(\sum\nolimits_{j=1}^{2}\log_2\left(1+\gamma_{kj}\right)-\frac{L_k}{T_k B}\right)\right\} \quad \text{(F.3)}$$

Using (F.3) and auxiliary variables ρ and κ, the problem in (F.1) can be simplifed as

$$\begin{aligned}
&\max_{\boldsymbol{p},t,\rho,\kappa} \quad t\\
&\text{s.t.} \quad \text{(C3)},\\
&\text{(C5)}: \omega\left(\rho - R^*\right) \geq t,\\
&\text{(C6)}: (1-\omega)\left(\kappa - Q^*\right) \geq t,\\
&\text{(C7)}: T_k B\sum\nolimits_{j=1}^{2}\log_2(1+\gamma_{kj}) \geq \rho, \forall k \in \mathcal{K},\\
&\text{(C8)}: \sqrt{T_k B}\log_e 2\left(\sum_{j=1}^{2}\log_2\left(1+\gamma_{kj}\right)-\frac{L_k}{T_k B}\right) \geq \kappa, \forall k. \quad \text{(F.4)}
\end{aligned}$$

In order to solve (F.4), we approximate the γ_{kj} by ν_{kj} using successive convex approximation as

$$\begin{aligned}
\nu_{kj} \leq\ & \frac{2\left|\left|h_k\right|\right|^2 p_{kj}}{\sum_{(a,b)\in\mathcal{Q}_{kj}}\left|\left|h_u\right|\right|^2 {p_{uv}}^{(a)}+\sigma^2}\\
&-\frac{\left|\left|h_k\right|\right|^2 p_{kj}^{(a)}\left(\sum_{(u,v)\in\mathcal{Q}_{kj}}\left|\left|h_u\right|\right|^2 p_{uv}+\sigma^2\right)}{\left(\sum_{(a,b)\in\mathcal{Q}_{kj}}\left|\left|h_u\right|\right|^2 {p_{uv}}^{(a)}+\sigma^2\right)^2}, \qquad \text{(F.5)}
\end{aligned}$$

where a denotes the iteration. Overall, the problem in (F.4) can be transformed into its equivalent convex form as

$$\begin{aligned}
&\max_{\boldsymbol{p},t,\rho,\kappa,\boldsymbol{\nu}} \quad t\\
&\text{s.t.} \qquad \text{(C3), (C5), (C6), (F.5), (C7), (C8)}, \qquad \text{(F.6)}
\end{aligned}$$

where $\boldsymbol{\nu} = \{\nu_{kj}\}$. The problem is now convex in nature and can be solved iteratively untill convergence.

References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, 2019.

[2] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[3] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, 2021.

[4] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, 2020.

[5] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, 2020.

[6] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, and H. Alves, "Network slicing for URLLC and eMBB with max-matching diversity channel allocation," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 658–661, 2019.

[7] F. Saggese, M. Moretti, and P. Popovski, "Power minimization of downlink spectrum slicing for eMBB and URLLC users," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 11 051–11 065, 2022.

[8] E. N. Tominaga, H. Alves, R. D. Souza, J. L. Rebelatto, and M. Latva-Aho, "Non-orthogonal multiple access and network slicing: Scalable coexistence of eMBB and URLLC," in *Proc. IEEE Veh. Technol. Conf. (VTC2021-Spring)*. IEEE, 2021, pp. 1–6.

[9] Y. Ruan, G. Nie, W. Ni, H. Tian, and J. Ren, "Efficient Traffic Scheduling for Coexistence of eMBB and uRLLC in Industrial IoT Networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*. IEEE, 2022, pp. 1431–1436.

[10] M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Rate splitting multiple access for sum-rate maximization in IRS aided uplink communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2246–2261, Apr. 2023.

[11] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, and G. Caire, "Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access techniques," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1310–1343, 2021.

[12] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4816–4828, Jul. 2022.

[13] B. Clerckx, Y. Mao, E. A. Jorswieck, J. Yuan, D. J. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1265–1308, 2023.

[14] J. Zeng, T. Lv, W. Ni, R. P. Liu, N. C. Beaulieu, and Y. J. Guo, "Ensuring max–min fairness of UL SIMO-NOMA: A rate splitting approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11 080–11 093, 2019.

[15] Z. Yang, M. Chen, W. Saad, W. Xu, and M. Shikh-Bahaei, "Sum-rate maximization of uplink rate splitting multiple access (RSMA) communication," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2596–2609, July 2022.

[16] C. Xu, B. Clerckx, S. Chen, Y. Mao, and J. Zhang, "Rate-splitting multiple access for multi-antenna joint radar and communications," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1332–1347, 2021.

[17] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1080–1093, 2020.

[18] M. Almekhlafi, M. Chraiti, M. A. Arfaoui, C. Assi, A. Ghrayeb, and A. Alloum, "A downlink puncturing scheme for simultaneous transmission of URLLC and eMBB traffic by exploiting data similarity," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13 087–13 100, 2021.

[19] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, and Y. Zhang, "Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks," *IEEE Syst. J.*, vol. 15, no. 1, pp. 1333–1343, 2020.

[20] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, 2022.

[21] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[22] M. Darabi, V. Jamali, L. Lampe, and R. Schober, "Hybrid puncturing and superposition scheme for joint scheduling of URLLC and eMBB traffic," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1081–1085, 2022.

[23] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2021, pp. 1–6.

[24] A. Manzoor, S. A. Kazmi, S. R. Pandey, and C. S. Hong, "Contract-based scheduling of URLLC packets in incumbent EMBB traffic," *IEEE Access*, vol. 8, pp. 167 516–167 526, 2020.

[25] Z. Wei, J. Guo, D. W. K. Ng, and J. Yuan, "Fairness comparison of uplink NOMA and OMA," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Nov. 2017, pp. 1–6.

[26] P. D. Diamantoulakis and G. K. Karagiannidis, "Maximizing proportional fairness in wireless powered communications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 202–205, 2017.

[27] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, 2019.

[28] M. Vaezi and H. Vincent Poor, "NOMA: An information-theoretic perspective," *Multiple access techniques for 5G wireless Netw. and beyond*, pp. 167–193, 2019.

[29] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, 2016.

[30] J. Xu, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for short-packet uplink communications: A finite blocklength analysis," *IEEE Commun. Lett.*, 2022.

[31] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *arXiv preprint arXiv:2201.03192*, 2022.

[32] O. Dizdar, Y. Mao, Y. Xu, P. Zhu, and B. Clerckx, "Rate-splitting multiple access for enhanced urllc and embb in 6g," in *17th Int. Sym. Wireless Commun. Syst. (ISWCS)*. IEEE, 2021, pp. 1–6.

[33] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming sdma and noma," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, pp. 1–54, 2018.

[34] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[35] Y. Zhu, X. Wang, Z. Zhang, X. Chen, and Y. Chen, "A rate-splitting non-orthogonal multiple access scheme for uplink transmission," in *Proc. Int. Conf. Wireless Commun. and Signal Process. (WCSP)*, 2017, pp. 1–6.

[36] H. Liu and K. S. Kwak, "Adaptive rate splitting for uplink non-orthogonal multiple access systems," in *Proc. Int. Conf. on Ubiquitous and Future Netw. (ICUFN)*, 2019, pp. 158–163.

[37] H. Liu, T. A. Tsiftsis, K. J. Kim, K. S. Kwak, and H. V. Poor, "Rate splitting for uplink NOMA with enhanced fairness and outage performance," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4657–4670, 2020.

[38] Z. Yang, M. Chen, W. Saad, W. Xu, and M. Shikh-Bahaei, "Sum-rate maximization of uplink rate splitting multiple access (RSMA) communication," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2596–2609, 2020.

[39] E. J. Dos Santos, R. D. Souza, and J. L. Rebelatto, "Rate-Splitting Multiple Access for URLLC Uplink in Physical Layer Network Slicing With eMBB," *IEEE Access*, vol. 9, pp. 163 178–163 187, 2021.

[40] Y. Liu, B. Clerckx, and P. Popovski, "Network slicing for eMBB, URLLC, and mMTC: An uplink rate-splitting multiple access approach," *arXiv preprint arXiv:2208.10841*, 2022.

[41] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A ran resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45 674–45 688, 2020.

[42] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Nonorthogonal HARQ for URLLC: Design and analysis," *IEEE Internet of Things J.*, vol. 8, no. 24, pp. 17 596–17 610, 2021.

[43] M. Katwe, K. Singh, C.-P. Li, and Z. Ding, "Ultra-high rate-reliability fairness in grant-free massive URLLC NOMA system: Joint power and channel allocation using meta-heuristic search," *IEEE Trans. Veh. Technol.*, vol. Early Access Article, pp. 1–17, 2023.

[44] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, 2017.

[45] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for urllc over 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 896–904, 2019.

[46] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Info. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[47] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for URLLC in a factory automation scenario," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1786–1801, 2019.

[48] A. A. Nasir, "Min-max decoding-error probability-based resource allocation for a URLLC system," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2864–2867, 2020.

[49] S. Pala, K. Singh, M. Katwe, and C.-P. Li, "Joint optimization of URLLC parameters and beamforming design for multi-RIS-aided MU-MISO URLLC system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 148–152, 2023.

[50] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and Jain's fairness index in resource allocation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3496–3509, 2013.

[51] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

[52] R. Jiao and L. Dai, "On the max-min fairness of beamspace mimo-NOMA," *IEEE Trans. Signal Proces.*, vol. 68, pp. 4919–4932, 2020.

[53] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless lans," in *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, 2008, pp. 1004–1012.

[54] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, 2016.

[55] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli, and P. Mogensen, "System level analysis of eMBB and grant-free URLLC multiplexing in uplink," in *IEEE 89th Veh. Technol. Conf. (VTC2019-Spring)*. IEEE, 2019, pp. 1–5.

[56] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct. and Multidisciplinary Optim.*, vol. 26, no. 6, pp. 369–395, 2004.

[57] R. Knobloch, J. Mlỳnek, and R. Srb, "The classic differential evolution algorithm and its convergence properties," *Applications of Mathematics*, vol. 62, pp. 197–208, 2017.

[58] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evolutionary Computation*, vol. 15, no. 1, pp. 4–31, 2010.

[59] R. Storn and K. Price, "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[60] K. V. Price, "Differential evolution," in *Handbook of optimization*. Springer, 2013, pp. 187–214.

[61] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE Trans. Evolutionary Computation*, vol. 13, no. 2, pp. 398–417, 2008.

[62] A. Esmaeily, H. K. Mendis, T. Mahmoodi, and K. Kralevska, "Beyond 5G resource slicing with mixed-numerologies for mission critical URLLC and eMBB coexistence," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 727–747, 2023.

[63] M. Luque, F. Ruiz, and K. Miettinen, "Global formulation for interactive multiobjective optimization," *Or Spectrum*, vol. 33, pp. 27–48, 2011.

[64] M. Katwe, K. Singh, C.-P. Li, and Z. Ding, "Spectral-efficient downlink systems under imperfect SIC and CSI: MC-NOMA or Partial NOMA?" *IEEE Wireless Commun. Lett.*, vol. 13, no. 1, pp. 133–137, 2024.

Mayur Katwe (Member, IEEE) received the B.E. degree in electronics and telecommunication from the SGBAU University, Amravati, India, in 2013, the M.Tech degree in Digital System from the Govt. College of Engineering, Pune, India, in 2016, and the Ph.D. degree in Electronics and Communication Engineering from Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 2021. Since February 2024, he has been working as an Assistant Professor at the National Institute of Technology (NIT), Raipur, India. He held the position of Research Scientist at Temasek Laboratories, Nanyang Technological University (NTU), Singapore from April 2023 to January 2024. Also, he held the position of a Postdoctoral Researcher with the Institute of Communications Engineering, National Sun Yat-sen University (NSYSU), Taiwan from 2021 to 2023. His current research interests include radio localization, full duplex radios, non-orthogonal multiple access, rate-splitting multiple access, eMBB-URLLC traffic multiplexing, re-configurable intelligent surfaces (RIS), integrated sensing and communication, simultaneous transmission and reflecting RIS (STAR-RIS) and unmanned aerial vehicles assisted communication.

Keshav Singh (Member, IEEE) received the M.Sc. degree in Information and Telecommunications Technologies from Athens Information Technology, Greece, in 2009, and the Ph.D. degree in Communication Engineering from National Central University, Taiwan, in 2015. He currently works at the Institute of Communications Engineering, National Sun Yat-sen University (NSYSU), Taiwan as an Assistant Professor. Prior to this, he held the position of Research Associate from 2016 to 2019 at the Institute of Digital Communications, University of Edinburgh, U.K. From 2019 to 2020, he was associated with the University College Dublin, Ireland as a Research Fellow. He had chaired workshops on conferences like IEEE GLOBECOM 2023 and IEEE WCNC, 2024. He also serves as leading guest editor of IEEE Transactions on Green Communications and Networking Special Issue on Design of Green Near-Field Wireless Communication Networks. He leads research in the areas of green communications, resource allocation, transceiver design for full-duplex radio, ultra-reliable low-latency communication, non-orthogonal multiple access, machine learning for wireless communications, integrated sensing and communications, non-terrestrial networks, and large intelligent surface assisted communications.

Chih-Peng Li (Fellow, IEEE) received the B.S. degree in Physics from National Tsing Hua University, Hsin Chu, Taiwan, and the Ph.D. degree in Electrical Engineering from Cornell University, NY, USA. Dr. Li was a Member of Technical Staff with Lucent Technologies. Since 2002, he has been with National Sun Yat-sen University (NSYSU), Kaohsiung, Taiwan, where he is currently a Distinguished Professor. Dr. Li has served various positions with NSYSU, including the Chairman of Electrical Engineering Department, the VP of General Affairs, the Dean of Engineering College, and the VP of Academic Affairs. His research interests include wireless communications, baseband signal processing, and data networks. He is now the Director General with the Engineering and Technologies Department, National Science and Technology Council, Taiwan.

Dr. Li is currently the Chapter Chair of IEEE Broadcasting Technology Society Tainan Section. Dr. Li has also served as the Chapter Chair of IEEE Communication Society Tainan Section, the President of Taiwan Institute of Electrical and Electronics Engineering, the Editor of IEEE Transactions on Wireless Communications, the Associate Editor of IEEE Transactions on Broadcasting, and the Member of Board of Governors with IEEE Tainan Section. Dr. Li has received various awards, including the Outstanding Research Award of Ministry of Science and Technology. Dr. Li is a Fellow of the IEEE.

Shankar Prakriya (SM'02) received the B.E. degree (Hons) in electronics and communication engineering from the Regional Engineering College, Bharathidasan University, Tiruchirappalli, in 1987, and the M.A.Sc. (Engg.) and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada, in 1993 and 1997, respectively. He worked for the Indian Space Research Organization for three years. He joined IIT Delhi in 1997, where he is currently a professor with the Department of Electrical Engineering. He was the Jai Gupta research chair professor for five years until September 2017. He holds three US and some Indian patents. His research interests include wireless energy harvesting, NOMA, cognitive radio, and full-duplex communication. He is a senior member of the IEEE. He has served in the technical program committees of prominent IEEE international conferences. He is currently serving as an editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Bruno Clerckx (Fellow, IEEE) is a (Full) Professor, the Head of the Communications and Signal Processing Group, and the Head of the Wireless Communications and Signal Processing Lab, within the Electrical and Electronic Engineering Department, Imperial College London, London, U.K. He received the MSc and Ph.D. degrees in Electrical Engineering from Université Catholique de Louvain, Belgium, and the Doctor of Science (DSc) degree from Imperial College London, U.K. He spent many years in industry with Silicon Austria Labs (SAL), Austria, where he was the Chief Technology Officer (CTO) responsible for all research areas of Austria's top research center for electronic based systems and with Samsung Electronics, South Korea, where he actively contributed to 4G (3GPP LTE/LTE-A and IEEE 802.16m) and. He has authored two books on "MIMO Wireless Communications" and "MIMO Wireless Networks", 300 peer-reviewed international research papers, and 150 standards contributions, and is the inventor of 80 issued or pending patents among which several have been adopted in the specifications of 4G standards and are used by billions of devices worldwide. His research spans the general area of wireless communications and signal processing for wireless networks. He received the prestigious Blondel Medal 2021 from France for exceptional work contributing to the progress of Science and Electrical and Electronic Industries, the 2021 Adolphe Wetrems Prize in mathematical and physical sciences from Royal Academy of Belgium, multiple awards from Samsung, IEEE best student paper award, and the EURASIP (European Association for Signal Processing) best paper award 2022. He is a Fellow of the IEEE and the IET.

George K. Karagiannidis ((M'96-SM'03-F'14) is currently Professor in the Electrical & Computer Engineering Dept. of Aristotle University of Thessaloniki, Greece and Head of Wireless Communications & Information Processing (WCIP) Group. He is also Faculty Fellow in the Cyber Security Systems and Applied AI Research Center, Lebanese American University. His research interests are in the areas of Wireless Communications Systems and Networks, Signal processing, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications & Signal Processing for Biomedical Engineering.

Dr. Karagiannidis was in the past Editor in several IEEE journals and from 2012 to 2015 he was the Editor-in Chief of IEEE Communications Letters. From January 2024 he is the Editor-in-Chief of IEEE Transactions on Communications.

He has received three prestigious awards: The 2021 IEEE ComSoc RCC Technical Recognition Award, the 2018 IEEE ComSoc SPCE Technical Recognition Award and the 2023 Humboldt Semior Research Award from Alexander von Humboldt Foundation.

Dr. Karagiannidis is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Highly-Cited Researcher in the nine consecutive years 2015-2023.