Network Slicing for eMBB, URLLC, and mMTC: An Uplink Rate-Splitting Multiple Access Approach

Yuanwen Liu®, Bruno Clerckx®, Fellow, IEEE, and Petar Popovski®, Fellow, IEEE

Abstract—There are three generic services in 5G: enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC). To guarantee the performance of heterogeneous services, network slicing is proposed to allocate resources to different services. Network slicing is typically done in an orthogonal multiple access (OMA) fashion, which means different services are allocated non-interfering resources. However, as the number of users grows, OMA-based slicing is not always optimal, and a non-orthogonal scheme may achieve better performance. This work aims to analyse the performances of different slicing schemes in uplink, and a promising solution based on rate-splitting multiple access (RSMA) is studied. RSMA can provide a more flexible decoding order and theoretically has the largest achievable rate region than OMA and non-orthogonal multiple access (NOMA) without time-sharing. Hence, RSMA has the potential to increase the rate of users requiring different services. In addition, it is not necessary to decode the two split streams of one user successively, so RSMA lets suitable users split messages and designs an appropriate decoding order depending on the service requirements. This work shows that for network slicing RSMA can outperform NOMA counterpart, and obtain significant gains over OMA in some regions.

Index Terms—eMBB, mMTC, network slicing, rate-splitting multiple access, URLLC.

I. INTRODUCTION

5G features three generic services: enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC). These three services have various applications and requirements [1], [2]. eMBB aims to provide extremely high data rates with high reliability, for applications such as augmented reality and remote presence. mMTC aims to provide connectivity for numerous devices which are cost and energy-constrained. The objective of URLLC is to provide a service featuring ultra-reliability and low latency while the rate is relatively low compared to mMTC. Some typical use cases are shown in Table I [3], [4], [5], [6].

Network slicing is a promising technique to satisfy different service requirements [7], [8]. This technique allocates communication resources to the users requiring different services to guarantee their varied requirements. Usually, network slicing is done by OMA-based slicing. However, as the number of users grows, a more flexible and efficient scheme may be explored, and these three services have different requirements, so network slicing may leverage these differences to get better performance. We note that this work is limited to the wireless part of the network slicing, and sometimes it is referred to as spectrum slicing [9].

NOMA-based network slicing scheme has the potential to improve the performance in the uplink, and it was studied in [10], [11], and [12]. In [10], a NOMA-based slicing scheme was explored, and this work showed that NOMA-based slicing is more suitable than OMA-based slicing in some scenarios. In [11], network slicing for eMBB and URLLC coexistence was analysed. There was one eMBB user and multiple URLLC users in this scenario. When URLLC users had better channel conditions than the eMBB user, NOMA-based slicing always outperformed OMA-slicing. While when eMBB users had better channel conditions, NOMA-based slicing only had better performance when the eMBB rate was relatively high. In [12], the eMBB and mMTC coexistence scenario was analysed. There was one eMBB user and multiple mMTC users, and the arrival process of mMTC traffic followed Poisson distribution. This paper analysed the achievable pairs of eMBB rate and mMTC arrival rate. According to the results, NOMA-based slicing can achieve the pairs that OMA-based slicing cannot, and vice versa. As the number of receive antennas at the

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
base station (BS) increased, NOMA-based slicing benefited more from this increase than OMA-based slicing. In [13], the performance trade-offs between the eMBB rate and the URLLC latency were investigated in NOMA-based slicing with both puncturing and successive interference cancellation (SIC) methods.

Applying NOMA-based slicing in downlink has also been studied and showed lower power consumption and more effective utilization. In [14], [15], [16], [17], [18], [19], and [20], eMBB and URLLC coexistence was studied. In [14], NOMA-based slicing was employed to minimize the transmission power and it had lower power consumption than OMA-based slicing. References [15], [16], [17], and [18] showed that NOMA-based slicing can leverage resources more effectively, and it has promising performance when resources allocation and scheduling are designed properly, and [19] proposed a learning-based approach for resource allocation. In [20], a scheduling algorithm for eMBB and URLLC coexistence in multiple-input multiple-output (MIMO) NOMA systems was studied, and it brought a higher spectrum utilization than the OMA counterpart.

RSMA has drawn a lot of attention recently for both downlink and uplink communications [21], [22]. In the downlink, the message of one or multiple user(s) is split into a common part and a private part, and all the common parts are encoded jointly into one common stream while the private parts are encoded independently into private streams. Each user can then reconstruct the original message from the decoded common stream and decoded private stream. By splitting messages at the transmitter, RSMA has been shown to unify and outperform NOMA and space division multiple access (SDMA) [23], [24]. RSMA was shown to be not only more spectrally efficient but also more energy efficient [25], [26] showed that RSMA is a powerful technique in multi-user MIMO scenarios. In [27], RSMA was shown to obtain the same performance as SDMA and NOMA with shorter blocklength, therefore opening the door to low latency applications. Reference [28] showed that RSMA is more robust than SDMA to the degrading effects of mobility and latency in CSI acquisition. With these advantages, RSMA is a competitive candidate, and [29] and [30] have shown that RSMA is a very promising technique for future networks.

RSMA has promising performance not only in the downlink but also in the uplink. For uplink RSMA, a user splits its message into two streams and transmits a superposed message to the BS [21], [31]. How the power is split between the two streams is a critical aspect because it dictates the performance. We use the splitting power fraction to refer to the power fraction allocated to the first stream without loss of generality. Since the number of streams increases, RSMA can provide a more flexible decoding order at the BS. By wisely deciding the decoding order and adjusting the splitting power fraction, RSMA can achieve all boundary points of the capacity region with SIC without the need for time-sharing [31], [32], improve the user fairness and outage performance [33], [34], and have the potential to simplify the implementation for uplink by avoiding the need for user pairing [35], [36] demonstrated that RSMA could increase connectivity and reliability in a semi-grant-free transmission scenario, which are significant expectations for future networks. A RSMA-based slicing scheme for eMBB and URLLC coexistence was proposed in [37], and this work showed that RSMA could achieve a larger rate region than NOMA-based and OMA-based slicing when the power splitting factor is properly configured.

In this paper, we explore the performance of RSMA-based slicing1 in eMBB and URLLC coexistence and eMBB and mMTC coexistence. For eMBB and URLLC coexistence, the contributions of this paper are summarised as follows:

- This paper investigates eMBB and URLLC coexistence with RSMA and analyses the performances in different channel conditions. The achievable rate regions and the relation between the splitting power fraction and the sum-rate of URLLC users are shown. Although [37] analysed the eMBB and URLLC coexistence when channel conditions of URLLC users were better than eMBB, and showed that RSMA always outperformed OMA-based slicing2 and NOMA-based slicing3 when users of the same services shared the resource in a non-orthogonal fashion, it did not analyse the opposite scenario when an eMBB user experiences a better channel than URLLC users. In the previous scenario, it would be easier for RSMA to exploit flexible decoding order; while in the later one, sometimes RSMA may boil down to NOMA. Interestingly, the later scenario shows when RSMA can outperform OMA and reveals a suitable scenario for RSMA. Our results show that RSMA can always achieve a larger rate region compared to NOMA, while it outperforms OMA in some scenarios. Adjusting splitting power fraction appropriately can improve the performance, and the lower the interference from the eMBB user, the higher the improvements RSMA can make.

- This paper gives a detailed analysis of the suitable scenario for RSMA. The main superiority of RSMA in uplink comes from the flexible decoding order since the split streams do not have to be decoded successively. Hence, for the user with the lower rate, one of its split streams can be decoded after the streams of other users, so it has lower interference and achieves a higher rate; for the user with a higher rate, its stream can be decoded early to deal with more interference. Thus, all the users can achieve a higher rate. In this way, a higher decoding order flexibility leads to higher achievable rate pairs. This is the reason why RSMA has better performance when eMBB interference is not that high. The achievable rate regions of NOMA and RSMA in eMBB and URLLC coexistence are also simulated to demonstrate the suitable scenario of RSMA.

For eMBB and mMTC coexistence, the contributions are summarised below:

- This paper proposes a RSMA scheme for eMBB and mMTC coexistence, and to the best of our knowledge,
RSMA has not been applied to this scenario before. In this scheme, the message of the eMBB user is split into two streams. By adjusting the splitting power fraction of eMBB and scheduling the decoding order, the signal-to-interference-plus-noise ratio (SINR) of mMTC devices is also adjusted, since the first stream of eMBB message can be cancelled before decoding some mMTC devices. Consequently, more mMTC devices can have sufficient SINR to satisfy the target rate compared to NOMA, which means a higher arrival rate of mMTC can be supported using RSMA for a given reliability requirement.

- This paper analyses the performance of eMBB and mMTC coexistence with RSMA. The achievable pairs of the eMBB rate and the arrival rate of mMTC and the relation between the splitting power fraction of eMBB and the arrival rate of mMTC are shown. According to the simulation results, RSMA can achieve rate pairs non-achievable by NOMA and OMA, and RSMA is shown to leverage the difference between reliability requirements of different services better compared to NOMA.

The rest of this paper is organised as follows. Section II gives the introduction to the system model. Section III analyses the network slicing performances of eMBB and URLLC coexistence. Section IV introduces eMBB and mMTC coexistence. Section V presents the numerical results, and Section VI is the conclusions.

Notations: $\mathbb{C}$ represents the set of complex numbers. $\mathbb{E}[\cdot]$ refers to the statistical expectation. $\mathcal{CN}(\delta, \sigma^2)$ represents a complex Gaussian distribution with mean $\delta$ and variance $\sigma^2$. $\text{Poisson}(\lambda)$ represents a Poisson distribution with mean $\lambda$. $\Gamma(s, x)$ represents the upper incomplete gamma function which is defined as $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t} \, dt$. $Q(x)$ represents Gaussian $Q$ function which is $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{t^2}{2} \right) \, dt$.

II. SYSTEM MODEL

This section introduces the model of 5G services individually, namely eMBB, URLLC and mMTC, and a general signal model of RSMA. Different services can share the same radio resources when they communicate to a common BS as shown in Fig. 1. We consider that there are $F$ frequency channels indexed by $f \in \{1, \ldots, F\}$, and $S$ time slots indexed by $s \in \{1, \ldots, S\}$, as shown in Fig. 2. Fig. 2a is the illustration of OMA. Each service occupies individual resources without overlap, but users of the same service can share the resources, i.e., one URLLC user can be allocated an individual resource denoted by light blue, and multiple URLLC users can also be allocated to the same resources represented by dark blue. Fig. 2b shows the general non-orthogonal slicing scenario, which includes NOMA and RSMA. Different services can share the same resource, and the sharing part is denoted by the overlap.

We assume that eMBB users are allocated to single or multiple radio resources at one or several given frequencies, while URLLC users can occupy several radio frequencies and a single mini-slot because of the stringent latency requirement.\textsuperscript{4} and mMTC users are allocated a specified frequency resource. The numbers of URLLC users and active mMTC devices are $n_U$ and $n_M$, respectively. The arrival process of mMTC follows a Poisson distribution, $n_M \sim \text{Poisson}(\lambda_M)$, where $\lambda_M$ is the mean value and referred to as the arrival rate of mMTC users.

As in [10], we assume that the radio access and competition among eMBB devices have been solved prior to the considered time slot, and there is only one eMBB device in each time slot. Each frequency channel $f$ is assumed to be within the time-and-frequency-coherence interval of the wireless channel, so the wireless channel coefficients are seen as constant. The channel coefficients of eMBB user, URLLC user $u$ and mMTC user $m$ at frequency $f$ are denoted by $h_{B,f} \in \mathbb{C}$, $h_{U,u,f} \in \mathbb{C}$ and $h_{M,m,f} \in \mathbb{C}$, respectively, where $u = 1, \ldots, n_U$ and $m = 1, 2, \ldots, n_M$, and they fade independently. The channels are considered as Rayleigh fading channels, i.e. $h_{B,f} \sim$

\textsuperscript{4} Besides latency requirement, high reliability is also an important requirement of URLLC and diversity is critical to achieve it [38]. Therefore, transmitting with multiple frequencies would be more reasonable to increase diversity. In this way, increasing diversity compensates for fades in wireless channels [39], [40] and guarantees reliability.
The gains of the users of three services in radio frequency feMBB, URLLC and mMTC users, respectively. The channel where power equals SNR. The received signal at BS at time slot s of the noise at BS is normalized to one. Thus, the received CSI, while for URLLC and mMTC devices the transmission CSI, and BS also has perfect CSI as in [10]. Thus, the eMBB information (CSI), while URLLC and mMTC do not have where

\[ \Gamma \]

and frequency \( \Gamma \). Without loss of generality, the power of the noise at BS is normalized to one. Thus, the received CSI, while for URLLC and mMTC devices the transmission CSI, and BS also has perfect CSI as in [10]. Thus, the eMBB information (CSI), while URLLC and mMTC do not have

models for different coexistence scenarios will be introduced, and the detailed uplink RSMA will be briefly introduced, and the detailed uplink RSMA models for different coexistence scenarios will be introduced in Section III and Section IV.

A. eMBB

Since the eMBB device and BS are assumed to have perfect CSI, the eMBB device can adjust the transmission power \( P_B(G_{B,f}) \) according to the instantaneous channel gains to achieve a predefined SNR value. Thus, the objective is that the eMBB device transmits at the largest rate \( r_{B,f} \) at frequency \( f \) and the outage probability is lower than reliability requirements \( \epsilon_B \) under a long-term average power constraint, so the problem can be formulated as

\[
\max r_{B,f} \quad \text{s.t. } \begin{align*}
\Pr[\log_2 (1 + G_{B,f}P_B(G_{B,f})) < r_{B,f}] & \leq \epsilon_B, \\
E[P_B(G_{B,f})] & = 1.
\end{align*}
\] (2)

The optimal solution to this problem is given by truncated power inversion [41], which means that the eMBB device chooses a transmission power that is inversely proportional to the channel gain \( G_{B,f} \) if it is above a given threshold \( G^\text{min}_{B,f} \). The probability that the eMBB transmits is

\[
a_B = \Pr[G_{B,f} \geq G^\text{min}_{B,f}] = e^{-G^\text{min}_{B,f}/\Gamma_B}.
\] (3)

Without interference from other services, the only outage source is that the eMBB user does not transmit due to the insufficient SNR level, so the event that the eMBB user is in outage is actually the event that the eMBB user is inactive, i.e. not transmitting. Thus, the error probability is

\[
\Pr(E_B) = 1 - a_B.
\] (4)

Impose the reliability requirement \( \epsilon_B \)

\[
\Pr(E_B) = \epsilon_B,
\] (5)

so we can obtain the value of threshold SNR

\[
G^\text{min}_{B,f} = \Gamma_B \ln \left( \frac{1}{1 - \epsilon_B} \right).
\] (6)

With the power-inversion scheme, the transmission power is chosen as a function of the instantaneous channel gain \( G_{B,f} \) as

\[
P_B(G_{B,f}) = \begin{cases} 
G^\text{tar}_{B,f} & \text{if } G_{B,f} \geq G^\text{min}_{B,f} \\
0 & \text{if } G_{B,f} < G^\text{min}_{B,f}
\end{cases},
\] (7)

where \( G^\text{tar}_{B,f} \) is the target SNR. Imposing the average power constraint to be normalized one, we have

\[
1 = \mathbb{E}[P_B(G_{B,f})] = \int_0^{G^\text{min}_{B,f}} \Gamma_B e^{-x/\Gamma_B} P_B(x) dx,
\] (8)

and the target SNR can be obtained as

\[
G^\text{tar}_{B,f} = \frac{\Gamma_B}{\Gamma(0, G^\text{min}_{B,f}/\Gamma_B)}.
\] (9)

Thus, the largest transmission rate of eMBB can be obtained as

\[
r_{B,f} = \log_2 (1 + G^\text{tar}_{B,f}) \text{ (bits/s/Hz)}.
\] (10)

From this result, we know that \( r_{B,f} \) only depends on \( \Gamma_B \) and \( \epsilon_B \), so for each frequency channel the largest transmission rate is the same, and the index \( f \) can be omitted. This largest transmission \( r_B \) is referred to as \( r^\text{orth}_B \),

\[
r^\text{orth}_B = \log_2 (1 + G^\text{tar}_{B,f}) \text{ (bits/s/Hz)}.
\] (11)

B. URLLC

URLLC devices transmit messages across \( F_U \) frequency resources, where \( F_U \leq F \). We assume that URLLC devices do not have CSI, and the transmission power is fixed to 1. BS has the CSI of URLLC users and uses SIC to decode the messages. In the SIC procedure, BS receives the superposition of URLLC signals, successively detects the received data symbols, and then subtracts them from the superposition of signals. URLLC users are intermittent and each user is active with the probability \( a_U \), so the number of active URLLC users is random in a given spanning time resource with multiple frequency resources, which we refer to as a frequency strip. This is a generalization of URLLC setup in [10], which assumes a single URLLC user. Here we consider the worst case that there are always \( n_U \) URLLC users transmitting simultaneously in the given frequency strip, so the outage...
probability in practice should be lower than the estimation which will be given then. Let $G_{U,u,f}$ denote the channel gain of user $u$ in frequency channel $f$, and the user decoded after user $u$ is denoted by $j > u$, so the SINR of user $u$ in frequency $f$

$$\sigma_{U,u,f} = \frac{G_{U,u,f}}{1 + \sum_{j > u} G_{U,j,f}}.$$  

(12)

Thus the outage probability is

$$\Pr(E_U) = \frac{1}{n_U} \sum_{u=1}^{n_U} \Pr \left( \frac{1}{F_U} \sum_{f=1}^{F_U} \log_2 (1 + \sigma_{U,u,f}) < r_U \right),$$

(13)

where $r_U$ is the target rate of URLLC.

The decoding order is decided by the sum of the mutual information across $F_U$ frequency resources [11]. The users are successively decoded, and BS picks the device with the largest value of the sum of mutual information $I_{sum}$ to decode at each step, i.e., for user $u$ it is

$$I_{sum} = \frac{1}{F_U} \sum_{f=1}^{F_U} \log_2 \left(1 + \frac{G_{U,u,f}}{1 + \sum_{j > u} G_{U,j,f}}\right).$$

(14)

Thus, we can impose the reliability requirement $\Pr(E_U) = \epsilon_U$ and obtain the maximum $r_U$, and the sum-rate of URLLC users is

$$r_U^{sum} = n_U r_U \text{ (bits/s/Hz).}$$

(15)

Increasing $F_U$ can enhance the frequency diversity, so it is possible to obtain a higher $r_U$ while satisfying the reliability requirement $\epsilon_U$ with larger $F_U$.

C. mMTC

The arrival process of mMTC devices is random, so we assume that it follows a Poisson distribution with the arrival rate $\lambda_M$, and the rate of mMTC devices is fixed at $r_M$. CSI is unknown for mMTC, and the transmission power is fixed to 1. Since mMTC devices are allocated a specified frequency resource, the index $f$ is omitted.

BS uses SIC to decode mMTC devices. The decoding order is decided by the channel gains of users. Let $m$ denote the mMTC device with the $m$-th largest channel gain and the channel gains are sorted in descending manner $G_1 \geq G_2 \geq \ldots G_{n_M}$, so the decoding order is $1, 2 \ldots n_M$. Since there is no interference from other services, for a random user $m_0$, the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + \sum_{m=m_0+1}^{n_M} G_m}.$$  

(16)

If $\log_2(1 + \sigma_{m_0}) \geq r_M$ holds, this mMTC device can be decoded successfully and then cancelled. Otherwise, it cannot be decoded and the procedure terminates. Let $D_M$ denote the number of successfully decoded devices, and the error probability measures the ratio of successfully decoded devices $D_M$ to all active devices $n_M$. Thus the error probability is

$$\Pr(E_M) = 1 - \frac{\mathbb{E}[D_M]}{\lambda_M},$$

(17)

where $\mathbb{E}[D_M]$ is the expectation of $D_M$, and $\lambda_M$ can represent the mean value of $n_M$. Thus, the problem can be formulated as

$$\lambda_M^{\text{opt}}(r_M) = \max \{\lambda_M : \Pr(E_M) \leq \epsilon_M\},$$

(18)

and the value of $\lambda_M^{\text{opt}}(r_M)$ can be computed by Monte Carlo method.

D. Signal Model of Rate-Splitting in the Uplink

Rate-splitting method for the multiple-access channel was first introduced in [31]. The main idea of rate-splitting is to divide a message of one user into two parts, which is equivalent to adding a virtual user, and in this way provide a more flexible decoding order. For a general RSMA network slicing model in uplink, we assume that there is one eMBB user, two URLLC users and $n_{mMTC}$ mMTC users. This setup can effectively capture the essence of the problem, because in a given cell the number of connected mMTC devices can be very large compared to the number of connected eMBB and URLLC users in each time slot. Since the arrival process of mMTC is random and mMTC devices are energy and cost-constrained, we assume that mMTC does not split the messages and the transmission power of the signal is fixed to 1. Hence, eMBB and URLLC users can have their message split, in contrast to mMTC users. Thus, the received signal of BS at a given time slot $t$ and frequency $f$ is ($s$ and $f$ are omitted for simplicity)

$$y = h_B \sum_{i=1}^{2} \sqrt{F_{B,i} x_{B,i}} + \sum_{u=1}^{2} \sum_{i=1}^{2} h_{U,u,i} \sqrt{F_{U,u,i} x_{U,u,i}}$$

$$+ \sum_{m=1}^{n_M} h_{M,m} x_{M,m} + z,$$

(19)

where $x_{B,i}$ and $x_{U,u,i}$ are split messages of eMBB user and URLLC user $u$, respectively, and $F_{B,i}, F_{U,u,i}$ are the power allocated for $x_{B,i}$ and $x_{U,u,i}$, respectively. For eMBB and URLLC, the power constraints are $\mathbb{E} \left[ \sum_{i=1}^{2} \sqrt{F_{B,i}} \right] = 1$ and $\sum_{i=1}^{2} \sum_{u=1}^{2} \sqrt{F_{U,u,i}} = 1$, respectively.

We note that RSMA can unify NOMA and OMA in a single multiple access framework. When the transmission power is all allocated to one of the split streams and the other stream is not allocated power, RSMA boils down to NOMA; while when all the transmission power is turned off except one service, RSMA becomes OMA. Thus, OMA and NOMA can be contained in RSMA if full power control is applicable for all the users. However, since URLLC and mMTC do not have the CSI and the transmission power is fixed, they do not have the full power control, so for network slicing, OMA is not a subset of RSMA.

Although RSMA increases the number of messages, the decoding order is more flexible and it can achieve any boundary points of the rate region as shown in [31]. Other methods that achieve the same region are NOMA with time-sharing and joint encoding/decoding [21]. However, implementing time-sharing requires multiple time slots [32] and induces communication overhead. A joint encoding/decoding approach is not practical to implement due to the high decoding complexity of random codes [21], [31]. Since URLLC has
stringent latency requirements and mMTC is aimed at cost and power-constrained devices, NOMA with time-sharing and joint encoding/decoding can be problematic in network slicing, so they will not be considered here.

III. SLICING FOR eMBB AND URLLC

In this section, the coexistence of eMBB and URLLC will be discussed, and mMTC traffic will not be considered here. Since the radio access and competition among eMBB devices are assumed to be solved, there is only one eMBB device and two URLLC devices in each time slot.

A. OMA for eMBB and URLLC

In the OMA scenario, we assume that the eMBB device and URLLC devices are allocated individual frequency resources. Let \( F \) denote the total number of the frequency resources, and URLLC devices are allocated \( F_U \) frequency resources.

The performance of this system is specified in terms of \((r^\text{sum}_B, r^\text{sum}_U)\) of eMBB sum-rate \( r^\text{sum}_B \) and sum-rate of URLLC \( r^\text{sum}_U \) while reliability requirements \((\epsilon_B, \epsilon_U)\) are satisfied. The sum-rate of eMBB is obtained as

\[
r^\text{sum}_B = (F - F_U)r^\text{th}_B \text{ (bits/s/Hz)}, \tag{20}
\]

where \( r^\text{th}_B \) is obtained in (11). The rate of URLLC devices can be computed from (13) while satisfying \( \Pr(E_U) \leq \epsilon_U \), and the sum-rate can be computed from (15).

B. NOMA for eMBB and URLLC

In the NOMA scenario, eMBB and URLLC devices share all \( F \) frequency resources. Here BS uses SIC to decode eMBB and URLLC users, and it is shown in [10] that the SIC method outperforms the puncturing method. Like the scenario mentioned in Section II-B, the worst case that there is always URLLC traffic when eMBB is transmitting is considered. Since URLLC devices are sensitive to latency, URLLC devices should be decoded prior to the eMBB device. The reliability requirement of URLLC is much more stringent than eMBB. If decoding eMBB first and this eMBB device fails, decoding URLLC users will fail. Thus, decoding eMBB first will require eMBB to satisfy the same reliability requirement as URLLC, and it is actually not possible for eMBB to satisfy it, so decoding eMBB first can deteriorate the reliability of URLLC. Hence, the decoding priority should be given to URLLC because it can leverage the reliability requirements difference. In addition, eMBB aims to provide the service with a high data rate, so decoding eMBB after URLLC will let eMBB users experience less interference and achieve a higher rate. Thus, the procedure is that URLLC devices are always decoded first. If it is successful, then an eMBB device will be decoded. Otherwise, the procedure terminates, because the interference from URLLC devices cannot be cancelled.

Since there is always interference from the eMBB device to URLLC devices, it would be better to choose a smaller eMBB target SNR to get better overall performance, i.e.

\[
G^\text{tar}_{B,f} \leq \frac{\Gamma_B}{\Gamma(0, \frac{G^\text{sum}_B}{R_B})}, \tag{21}
\]

so the achievable rate for eMBB is

\[
r_{B,f} = \log_2 \left(1 + G^\text{tar}_{B,f} \right) \text{ (bits/s/Hz)}, \tag{22}
\]

and the sum-rate is

\[
r^\text{sum}_B = \sum_{f=1}^{F} r_{B,f}. \tag{23}
\]

Given the rate of eMBB \( r_{B,f} \), the objective is to find the highest achievable rate \( r_U \) while reliability requirements are satisfied. The error probabilities of eMBB and URLLC are calculated individually. For eMBB, the error probability can be computed by the Monte Carlo method, which is

\[
\Pr(E_B) = 1 - E[D_B], \tag{24}
\]

where \( D_B \) is the number of successfully decoded eMBB devices.

Then the error probability of URLLC is formulated. The decoding order of URLLC can be decided by the same method as in Section II-B, and assuming it is 1, 2, . . . , \( n_U \), so the SINR of the URLLC device is

\[
\sigma_{U,u,f} = \frac{G_{U,u,f}}{1 + \sum_{j>u} G_{U,J,f} + \delta_f G_{B,f}^\text{tar}}, \tag{25}
\]

where \( \delta_f \) is an independent Bernoulli random variable with parameter \( a_B \) which is the active probability of the eMBB device. Since the reliability requirement of the eMBB user is also stringent, i.e. \( 10^{-3} \), it means that the probability that the eMBB device is inactive is extremely low. Thus, we consider the most stringent scenario that eMBB is always active, which means \( \delta_f = 1 \) and there is always interference from the eMBB device. Thus, the error probability of URLLC can be presented by (13), where \( \sigma_{U,u,f} \) is obtained from (25) and \( F_U = F \). Then, impose \( \epsilon_U, \Pr(E_U) \leq \epsilon_U \), and \( r_U \) can be computed by the Monte Carlo method, and the sum-rate is

\[
r^\text{sum}_U = n_U r_U \text{ (bits/s/Hz)}. \tag{26}
\]

C. RSMA for eMBB and URLLC

Since URLLC devices are very sensitive to latency and have high reliability requirements, i.e. decoding of URLLC should not depend on decoding of other services, they should be decoded first, so rate-splitting is only used for URLLC devices. Here we consider the setup in Section II-D with one eMBB user and two URLLC users and eMBB is always decoded after URLLC. If any URLLC message cannot be decoded, the procedure terminates and the eMBB message is lost.

The eMBB part is the same as in Section III-B. \( G^\text{tar}_{B} \) is chosen by the same method as in (21), and the rate can be obtained from (22) and (23). The objective is to find the highest achievable rate pairs \( (r^\text{sum}_B, r^\text{sum}_U) \) while satisfying the reliability requirements of eMBB and URLLC. The error probability of eMBB can be obtained from (24). The decoding

5 Actually, from an information theory perspective, all users but one (e.g. eMBB user and one of the two URLLC users, or the two URLLC users) split their messages to achieve the capacity region. However, in network slicing, latency and reliability should also be considered. Hence, due to similar reasons as in Section III-B, URLLC will always be decoded before eMBB and RSMA is only applied to URLLC users.
order of URLLC is the same as the one in Section II-B. Let URLLC user 1 denote the device decoded first, and URLLC user 2 denote the other. The message from URLLC user 1 is divided into two parts, and the power factor for the first part is $\beta, \beta \in [0, 1]$. The received signal at BS at frequency $f$ is

$$y = h_{B,f} x_{B,f} + \sqrt{\beta} h_{U,1,f} x_{U,1,1,f} + \sqrt{1 - \beta} h_{U,1,f} x_{U,1,2,f} + h_{U,2,f} x_{U,2,f} + z,$$

(26)

where $x_{U,1,1,f}$ and $x_{U,1,2,f}$ are the signals from the first and second part of the message of URLLC user 1, respectively. Without loss of generality, we assume that $x_{U,1,1,f}$ is decoded first. The decoding order is set to $x_{U,1,1,f}, x_{U,2,f}$ and $x_{U,1,2,f}$, since this order can obtain all the boundary points of rate region by adjusting $\beta$. Thus, the SINR of $x_{U,1,1,f}$ is

$$\sigma_{U,1,1,f} = \frac{\beta G_{U,1,f}}{1 + (1 - \beta) G_{U,1,f} + G_{U,2,f} + G_{B}^{tar}},$$

(27)

the SINR of $x_{U,2,f}$ is

$$\sigma_{U,2,f} = \frac{G_{U,2,f}}{1 + (1 - \beta) G_{U,1,f} + G_{B}^{tar}},$$

(28)

and the SINR of $x_{U,1,2}$ is

$$\sigma_{U,1,2,f} = \frac{(1 - \beta) G_{U,1,f}}{1 + G_{B}^{tar}}.$$  

(29)

Then we get the corresponding rate

$$r_{U,1,1,f} = \log_2(1 + \sigma_{U,1,1,f}) \text{ (bits/s/Hz)},$$

(30)

$$r_{U,1,2,f} = \log_2(1 + \sigma_{U,1,2,f}) \text{ (bits/s/Hz)},$$

(31)

$$r_{U,2,f} = \log_2(1 + \sigma_{U,2,f}) \text{ (bits/s/Hz)},$$

(32)

and the sum-rate of URLLC user 1 is

$$r_U = \frac{1}{F} \sum_{f=1}^{F} (r_{U,1,1,f} + r_{U,1,2,f}) \text{ (bits/s/Hz)}.$$  

(33)

Similarly, the rate of URLLC user 2 is

$$r_{U,2} = \frac{1}{F} \sum_{f=1}^{F} r_{U,2,f} \text{ (bits/s/Hz)}.$$  

(34)

The error probability is formulated as

$$Pr(E_U) = \frac{1}{2} \sum_{u=1}^{2} Pr(r_{U,u} < r_U),$$

(35)

and impose $\epsilon_U, Pr(E_U) \leq \epsilon_U$, and $r_U$ can be computed by Monte Carlo method, and $r_{U,ortho}$ can be obtained from (15).

Remark 1: The scenario with two URLLC users can also give insight into a more general scenario with $n_U$ URLLC users. First, we consider a generic two-URLLC-user case, and we do not consider NOMA with time-sharing for the reasons mentioned in II-D. We assume that user 1 splits the message into two streams, and user 2 does not split. If BS decodes the split streams of user 1 not successively, this decoding order can achieve any point on the diagonal line of the rate region. Fig. 3 is an illustration of rate regions of time-division multiple access (TDMA), frequency-division multiple access (FDMA), NOMA and RSMA for a generic two-user case, and the average channel gain of both users is 20 dB. With the nonsuccessive decoding order, RSMA can achieve the points on the yellow diagonal line. Similarly, in URLLC and eMBB coexistence, RSMA could achieve all the boundary points of rate region by adjusting $\beta$. Thus, intuitively, when the number of users is $n_U$, these users can be ordered by mutual information in descending manner, and the order is $1, 2, \ldots, n_U$. Let the first $n_U - 1$ users split messages and decode these split streams not successively, this decoding order should also bring freedom to obtain any boundary points of the rate region, i.e. the decoding order could be $x_{U,1,1}, x_{U,2,1}, \ldots, x_{U,n_U-1,1}, x_{U,n_U,1}, x_{U,n_U-1,2}, \ldots, x_{U,2,2}, x_{U,1,2}$.

IV. SLICING FOR eMBB and mMTC

In this section, the coexistence of eMBB and mMTC will be discussed. The arrival process of mMTC follows Poisson distribution as in Section II-C, and URLLC traffic will not be considered. Similarly, there is one eMBB device and multiple mMTC devices, because we assume that the radio access and competition among eMBB devices are solved. Since only a single frequency channel is considered, the frequency indices $f$ are omitted in this section.

A. OMA for eMBB and mMTC

In the OMA scenario, eMBB and mMTC users share the resources in a time-sharing manner. Let $\alpha, \alpha \in [0, 1]$, denote the time faction allocated to eMBB users, and the time allocated to mMTC users is $1 - \alpha$. Thus, the achievable rate of eMBB is

$$r_B = \alpha r_B^{orth} \text{ (bits/s/Hz)},$$

(36)
where $r_{eMBB}^{opt}$ is obtained in (11). The arrival rate $\lambda_M$ can be computed by a method similar to that in Section II-C. It is equivalent to the scenario in Section II-C with mMTC rate at $r_M^{\frac{\alpha}{1-\alpha}}$, so

$$\lambda_M = \lambda_M^{opt} \left( \frac{r_M}{1-\alpha} \right).$$

(37)

Then, $\lambda_M$ can be obtained from (18).

B. NOMA for eMBB and mMTC

In the NOMA scenario, eMBB and mMTC devices share the same frequency resources at the same time. The decoding procedure is the same as in [10]. Trying to decode the eMBB message before mMTC messages may be an option since the high rate requirement of eMBB can cause high interference when decoding mMTC messages. However, this is not the optimal option, because the error probability of mMTC is actually measuring the ratio of successfully decoded mMTC devices to all active mMTC devices instead of the decoding probability for each device. There may be some mMTC devices with very high channel gains, so it would be better to decode them before decoding the eMBB device.

Based on the analysis above, [10] gave the exact procedure. The mMTC devices are ordered in a descending manner, and BS tries to decode mMTC devices one by one. If an mMTC device cannot be decoded, BS tries to decode the eMBB device. If this eMBB device is successfully decoded, then BS continues decoding mMTC devices. Otherwise, the decoding procedure terminates, because eMBB interference cannot be cancelled.

For the eMBB device, similar to the eMBB and URLLC coexistence scenario, it would be better to choose a lower target SNR for eMBB $G_{eMBB}^{tar}$ which satisfies (21), so the rate of eMBB device is (22). Because the error of decoding the eMBB device is caused by the interference from mMTC devices and inactive state due to the insufficient SNR level, we must set a higher active probability for the eMBB device to satisfy the reliability requirement $\epsilon_B$. For example, if there is not any other interference, like the scenario in Section II-A, $a_B = 1 - \epsilon_B$, i.e., $\epsilon_B = 10^{-9}$, $a_B = 0.999$, while for eMBB and mMTC coexistence scenario, the active probability $a_B$ should satisfy $a_B > 0.999$. Here we assume that $a_B = 1$, which means there is always eMBB interference for mMTC devices, and the error of eMBB is only caused by decoding error. Thus, the SINR of an mMTC device decoded before the eMBB device is

$$\sigma_{m0} = \frac{G_{m0}}{1 + \sum_{m=m0+1}^{nM} G_{m} + G_{eMBB}^{tar}}.$$  

(38)

If $\log_2(1 + \sigma_{m0}) \geq r_M$, this mMTC device can be decoded. Otherwise, it failed and BS will try to decode the eMBB device. The SINR of the eMBB device is

$$\sigma_B = \frac{G_{eMBB}^{tar}}{1 + \sum_{m=m0+1}^{nM} G_{m}}.$$  

(39)

If $\log_2(1 + \sigma_B) \geq r_B$, the eMBB device can be decoded, and then BS will try to decode other mMTC devices as in Section II-C. Otherwise, it fails and the procedure terminates.

Let $D_M$ and $D_B$ denote the number of decoded mMTC devices and eMBB device, respectively. The error probability of mMTC is the ratio of expectation of decoded devices to all active devices, $1 - \mathbb{E}[D_M]/\lambda_M$, and the error probability of eMBB device is $1 - \mathbb{E}[D_B]$. In order to find the achievable pairs $(r_B, \lambda_M)$, the problem can be formulated as

$$\lambda_M^{NOMA}(r_B) = \max\{\lambda_M \geq 0 : \exists G_{eMBB}^{tar} \}
\text{s.t. } \mathbb{E}[D_M]/\lambda_M \geq 1 - \epsilon_M
\text{ and } \mathbb{E}[D_B] \geq 1 - \epsilon_B\}.$$  

(40)

For a given $r_B$, $\lambda_M^{NOMA}$ can be computed by Monte Carlo method.

C. RSMA for eMBB and mMTC

According to the discussion in NOMA for eMBB and mMTC, the decoding procedure can terminate when neither mMTC device nor eMBB device can be decoded. This implies that mMTC device and eMBB do not have sufficient SINR to achieve the target rate, which is caused by the relatively high interference from each other. Then the decoding procedure terminates because it is ‘lodged’.

To mitigate this effect, we apply RSMA in the eMBB and mMTC coexistence scenario. The eMBB message is split into two streams, so BS can decode one stream at a relatively low rate, and then continue decoding mMTC devices. Since part of the interference from eMBB can be cancelled, the decoding procedure avoids getting “lodged” to some extent. The whole decoding procedure is described below. The decoding order of the mMTC devices is defined according to the descending order of their channel gains. Firstly, BS tries to decode mMTC devices. If a mMTC device cannot be decoded, BS tries to decode the first stream of eMBB at a relatively low rate $r_{B_1}$, and then continues decoding mMTC devices. Similarly, if another mMTC device cannot be decoded, BS tries to decode the second stream of eMBB at rate $r_{B_2}$. If $r_{B_1} + r_{B_2} < r_B$, the eMBB device failed, and the decoding procedure terminates. Otherwise, BS continues decoding mMTC devices as in Section II-C. Note that $r_{B_1}$ is a non-negative value, so it is always possible to decode the first stream with a rate not higher than $r_{B_1}$, so we assume that BS could always decode the first stream regardless of whether the second stream can be decoded successfully. Similar to Section IV-B, the active probability of eMBB is $a_B = 1$, and the power factor for the first part of the eMBB message is $\beta$, $\beta \in [0, 1]$. The received signal is

$$y = \sqrt{\beta} h_B x_{B_1} + \sqrt{1 - \beta} h_B x_{B_2} + \sum_{m=1}^{nM} h_m x_m + z,$$  

(41)

where $x_{B_1}$ and $x_{B_2}$ are the signals from the first and second streams of the eMBB message, respectively, and $h_m$ is the channel coefficient of mMTC user $m$. The target SNR of eMBB is still $G_{eMBB}^{tar}$, and it satisfies (21). We assume that $x_{B_1}$ is always decoded first, and its SINR at BS is

$$\sigma_{B_1} = \frac{\beta G_{eMBB}^{tar}}{1 + (1 - \beta) G_{eMBB}^{tar} + \sum_{m=m_1}^{nM} G_m},$$

(42)
where $m_1$ represents the index of the mMTC device which will be decoded after $x_{B_1}$, and the rate is

$$r_{B_1} = \log_2 (1 + \sigma_{B_1}) \text{ (bits/s/Hz)}.$$  

(43)

The SINR of $x_{B_2}$ is

$$\sigma_{B_2} = \frac{(1 - \beta)G_{tar}^B}{1 + \sum_{m=m_2}^{m-M} G_m},$$

where $m_2$ represents the number of the mMTC device which will be decoded after $x_{B_2}$, and the rate is

$$r_{B_2} = \log_2 (1 + \sigma_{B_2}) \text{ (bits/s/Hz)}.$$  

(45)

For mMTC devices decoded before all eMBB streams, the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + (1 - \beta)G_{tar}^B + \sum_{m=m_0+1}^{m-M} G_m}.$$  

(46)

For mMTC devices decoded between $x_{B_1}$ and $x_{B_2}$, the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + (1 - \beta)G_{tar}^B + \sum_{m=m_0+1}^{m-M} G_m}.$$  

(47)

For mMTC devices decoded after all eMBB streams, the SINR is as same as (16).

Let $D_M$ and $D_B$ denote the number of decoded mMTC devices and eMBB device, respectively. In this scenario, $D_B = 1$ holds only when $r_{B_1} + r_{B_2} \geq r_B$ holds. Otherwise, $D_B = 0$. For mMTC devices, they can be decoded when $\log_2 (1 + \sigma_{m_0}) \geq r_M$ holds. Similarly, in order to find the achievable pairs $(r_B, \lambda_M)$, the problem can be formulated as

$$\lambda_M^{RMSA} (r_B) = \max \{ \lambda_M \geq 0: \exists G_{tar}^B \text{ and } \beta$$

s.t. $\frac{D_M}{\lambda_M} \geq 1 - \epsilon_M$ and $\frac{D_B}{\lambda_M} \geq 1 - \epsilon_B \}.$$  

(48)

$\lambda_M^{RMSA}$ can be computed by the Monte Carlo method.

V. NUMERICAL RESULTS

This section presents the simulation results of eMBB and URLLC coexistence and eMBB and mMTC coexistence. For eMBB and URLLC coexistence, the achievable rate pairs of eMBB and URLLC and the relation between splitting power fraction and sum-rate of URLLC in different channel conditions are shown and analysed. The rate region of two URLLC users in eMBB and URLLC coexistence is simulated to give a clearer analysis of what is a suitable scenario for RSMA. For eMBB and mMTC coexistence, the achievable pairs of eMBB rate and mMTC arrival rate and the relation between splitting power fraction and mMTC arrival rate are presented and analysed. The suitable situations for RSMA are also discussed in this section.

A. eMBB and URLLC Coexistence

The simulation results when $n_U = 2$ are shown. Rate regions $(r_{U}^{sum}, r_{B}^{sum})$ are presented in Fig. 4 and Fig. 5. There is one eMBB user and two URLLC users, and the total number of frequency resources is $F = 10$. Note that to make the figures more intuitive these rate regions are presented as enclosed, and each achievable rate pair is denoted by a marker.

In Fig. 4, $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$. This result shows that NOMA always outperforms OMA, and RSMA outperforms these two schemes. In this scenario, $\Gamma_U > \Gamma_B$, non-orthogonal slicing schemes can leverage the difference between channel gains of URLLC and eMBB to perform SIC, so NOMA and RSMA outperform OMA, especially at higher rate of eMBB, and RSMA can always achieve a larger rate region than NOMA. The gain of RSMA comes from the decoding order flexibility. By allocating power to split streams appropriately, both URLLC users can achieve a higher $r_U^B$ and then a higher sum-rate $r_U^{sum}$ can be obtained.

We can imagine a simple situation to get the intuition. Assuming that the channel gains of URLLC user 1 and URLLC user 2 are 21 dB and 19 dB, respectively. The target SNR of eMBB is assumed to be 10 dB. If NOMA is applied and the decoding order is URLLC user 1, URLLC user 2 and eMBB, we can easily calculate the rate for URLLC user 1 and URLLC user 2 are 1.26 (bits/s/Hz) and 3.04 (bits/s/Hz). Obviously, $r_U^B$ should not be higher than 1.26 (bits/s/Hz) if we want both the two users to be decoded. If RSMA is applied and URLLC user 1 splits its message into two streams and the splitting power fraction is 0.8, we can use the decoding order discussed in Section III-C and then calculate the rate for URLLC user 1 and URLLC user 2, and the results are 2.62 (bits/s/Hz) and 1.68 (bits/s/Hz), respectively. Now $r_U^B$ can be 1.68 (bits/s/Hz), so $r_U^{sum}$ also increases.

In Fig. 5, $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$. This simulation result shows that OMA outperforms NOMA except when $r_B^{sum}$ is relatively high. RSMA can outperform NOMA for the same reason in Fig. 4, and outperform OMA when $r_B^{sum}$ is in the range from $0 - 5$ (bits/s/Hz) and in the range from $25 - 40$ (bits/s/Hz). When $\Gamma_B > \Gamma_U$, it is more difficult to satisfy the reliability requirements of URLLC devices when interference from the eMBB device exists, so OMA is preferable in this scenario. However, if the aim is to achieve a relatively high $r_B^{sum}$, NOMA can perform better, because eMBB device can take advantage of...
frequency diversity. Similarly, RSMA can outperform OMA, and the improvement made by RSMA is more obvious when interference from eMBB device is relatively low, and this is also shown in Fig. 4.

Fig. 6 and Fig. 7 present the relation between splitting power fraction $\beta$ and the sum-rate of URLLC $r_{U}^{\text{sum}}$. URLLC users occupy $F_U = 5$ frequency resources in the OMA case, and the eMBB user occupies another $F - F_U = 5$ frequency resources. In both NOMA and RSMA cases, the eMBB rate is fixed at the rate which is the same as the one in OMA case. Since there is no power splitting in OMA and NOMA, the lines corresponding to OMA and NOMA are horizontal lines. In Fig. 6, $\Gamma_B = 10$ dB, $\Gamma_U = 20$ db, and $r_{U}^{\text{sum}}$ achieves the maximal point at $\beta = 0.95$. In this scenario, the URLLC channel condition is better than the eMBB one, so RSMA has more space to adjust the splitting power fraction. However, in Fig. 7, $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $r_{U}^{\text{sum}}$ increases as $\beta$ increases, and it achieves the maximum point at $\beta = 1$, which means RSMA boils down to NOMA. The reason is that due to the relatively high eMBB interference, it is difficult for URLLC user 2 to achieve $r_{U}$ while having the interference from a split stream of URLLC user 1, so splitting power does not make that much difference in this situation. This explains why RSMA and NOMA have the same performance when the rate of eMBB is relatively high in Fig. 5.

The reason why RSMA outperforms NOMA is that it can provide a more flexible SIC decoding order to achieve all the boundary points of the rate region theoretically. Thus, a higher $r_{U}$ and $r_{U}^{\text{sum}}$ could be found in RSMA. In eMBB and URLLC coexistence, the rate regions of two URLLC users in RSMA and NOMA are shown in Fig. 8, and the eMBB rate is the same as the rate in Fig. 6 and Fig. 7. In Fig. 8, the boundary points are the average of achievable rate pairs which satisfy the reliability requirements. NOMA without time-sharing and joint encoding/decoding can only achieve the points in the region bounded by the red line, while RSMA can achieve the points in the region bounded by the blue line, which includes the region of NOMA. The essence is that SINR of both URLLC users is adjustable by changing the splitting power fraction, so that RSMA can achieve all the points on the diagonal line instead of just two corner points in the rate region, and this can be seen as the user with higher rate compensates the user with lower rate to achieve the desired rate. When $\Gamma_B = 10$ dB and $\Gamma_U = 20$ dB, the SIR of URLLC users is high, the two corner points of NOMA are distant and the difference of regions of NOMA and RSMA is big, which is shown in Fig. 8a. In this situation, RSMA has more room to adjust the splitting power fraction to obtain the points on the diagonal line, and a higher $r_{U}$ can be found. However, when $\Gamma_B = 20$ dB and $\Gamma_U = 10$ dB and the SIR of URLLC users is low, the two corner points of NOMA can be very close due to the high interference from eMBB, so the difference between NOMA and RSMA regions becomes smaller, which is shown in Fig. 8b. In this scenario, RSMA only has a small room to play its role, and the user with a higher rate also struggles to satisfy the rate requirement, so it is more difficult to compensate the user with a lower rate. Thus, RSMA obtains better performance than both OMA and NOMA when $r_{U}^{\text{sum}}$ is relatively low, while obtains the same performance as NOMA when $r_{U}^{\text{sum}}$ is high.
B. eMBB and mMTC Coexistence

The simulation results of achievable pairs \((r_B, \lambda_M)\) are presented. In Fig. 9, \(\Gamma_B = 20\) dB, \(\Gamma_M = 5\) dB, \(r_M = 0.04\) (bits/s/Hz), \(\epsilon_B = 10^{-3}\), \(\epsilon_M = 10^{-1}\). This result shows that when \(r_B > 1.8\) (bits/s/Hz), OMA can achieve higher \(\lambda_M\) than NOMA; while when \(0.5 < r_B < 1.8\) (bits/s/Hz), NOMA achieves higher \(\lambda_M\). RSMA always outperforms NOMA and outperforms OMA when \(0.5 < r_B < 2.5\) (bits/s/Hz), but it still cannot achieve the region that OMA achieves when \(r_B > 2.5\) (bit/s/Hz). The curve of OMA is approximately linear, while for NOMA and RSMA the curves drop dramatically at \(r_B = 1.5\) (bit/s/Hz) and \(r_B = 2\) (bit/s/Hz), respectively.

When \(r_B\) is relatively low, i.e. \(0 - 0.5\) (bits/s/Hz), the performances of these three schemes are very similar, because for NOMA and RSMA the interference from eMBB device is relatively low, so it almost does not affect mMTC traffic, and for OMA, a large time fraction is allocated to mMTC traffic and eMBB only occupies a small one. When \(r_B\) is medium, i.e. \(0.5 - 2\) (bits/s/Hz), RSMA and NOMA outperform OMA.

At this eMBB rate, BS can leverage the differences between channel gains and reliability requirements of eMBB and mMTC devices when NOMA or RSMA is applied, but OMA cannot leverage these differences. When eMBB rate is relatively high, i.e. \(2 - 4\) (bits/s/Hz), OMA outperforms NOMA, and it also outperforms RSMA when \(r_B > 2.5\) (bits/s/Hz). eMBB device will cause high interference to mMTC devices due to its relatively high rate, and almost no mMTC devices can be decoded before the eMBB device. Because of the power constraint of the eMBB device, the eMBB device cannot tolerate much interference from mMTC devices, so \(\lambda_M\) drops to satisfy \(\epsilon_M\). While for OMA, eMBB and mMTC traffic are isolated, so mMTC will not be affected by eMBB so much compared to NOMA and RSMA.

RSMA mainly outperforms NOMA at a medium eMBB rate, i.e. \(1.5 - 3\) (bits/s/Hz). At this rate region, some mMTC devices must be decoded after eMBB due to the insufficient SINR, so if eMBB cannot be decoded, the decoding procedure is ‘lodged’ and then terminates. As analysed in the previous section, using RSMA can let part of the eMBB interference be cancelled, and in this way, the decoding procedure avoids ‘lodged’ to some extent, so RSMA has gain in this situation. For low eMBB rate and high eMBB rate, RSMA does not
make much difference, because at a low rate almost all mMTC devices can be decoded before the eMBB device, and at a high rate mMTC devices must be decoded after eMBB device, and RSMA boils down to NOMA.

Fig. 10 shows the relation between mMTC arrival rate \( \lambda_M \) and power factor \( \beta \) when eMBB rate is fixed at 2 (bits/s/Hz). Since OMA and NOMA do not split power, the corresponding curves are horizontal. For RSMA, the arrival rate achieves the maximum point at \( \beta = 0.45 \), and is the same as NOMA when \( \beta = 0 \) and \( \beta = 1 \). As shown in Fig. 9, \( r_B = 2 \) (bits/s/Hz) is a relatively high rate, and either mMTC devices or eMBB device may not have enough SINR to be decoded, which causes decoding procedure to terminate. By adjusting the splitting power fraction of the eMBB device, the SINR of mMTC devices is also adjusted, so that more mMTC devices can be decoded. Thus, \( \lambda_M \) increases as \( \beta \) increases till \( \beta = 0.45 \), and then decreases to the point which is as same as NOMA at \( \beta = 1 \).

VI. CONCLUSION AND FUTURE WORK

In this paper, we investigate eMBB and URLLC coexistence and eMBB and mMTC coexistence with OMA, NOMA and RSMA. For RSMA scheme designs, it is important to decide properly which user splits the message and design the power allocations. Given the different requirements of eMBB, URLLC and mMTC, different splitting strategies are used. For eMBB and URLLC coexistence, RSMA is applied between URLLC users to fulfill the requirements of eMBB and URLLC. The performances in different channel conditions are simulated and analysed, and then we discuss the suitable scenario for RSMA. For eMBB and mMTC coexistence, a novel RSMA scheme is proposed. RSMA is applied between eMBB and mMTC because of service requirements and the nature of mMTC arrival process, and the performance is simulated and analysed.

According to the simulation results, RSMA always outperforms NOMA in terms of achievable pairs for both coexistence scenarios. RSMA and NOMA can achieve the achievable pairs that OMA cannot; while OMA can also achieve the pairs that RSMA or NOMA cannot. The advantages of RSMA are that it can leverage the differences between channel gains and reliability requirements, and it achieves a larger rate region. The splitting power fraction of RSMA is a very significant factor to achieve the optimal sum-rate, and this paper gives insights on how the splitting power fraction can impact the performances. For eMBB and URLLC coexistence, the lower the interference from eMBB, the higher the improvement RSMA can make, because RSMA has more room to adjust the power allocation and obtain the optimal rate pairs while guaranteeing the service requirements. For eMBB and mMTC coexistence, due to splitting the messages, it is possible for RSMA to let part of the eMBB interference be cancelled, so that it can support a higher arrival rate of mMTC for a given reliability requirement when the splitting power fraction is set properly.

This paper presents the promising results of RSMA with network slicing. These results will motivate us to explore more general setups. For eMBB and URLLC coexistence scenario, we do not consider finite blocklength, because in this work we consider that URLLC does not have instantaneous CSI, such that the factor which causes packet error is the randomness of the channel coefficient (SNR) instead of the noise instance at a given SNR [42]. Therefore, we are working with the asymptotic rate. The scenario related to finite blocklength could be explored in future work. A more general \( M_M \)-URLLC-user scenario would also be an interesting direction and the extension of Remark 1 about decoding order could be studied.

REFERENCES

[1] H. Tullberg et al., “The METIS 5G system concept: Meeting the 5G requirements,” IEEE Commun. Mag., vol. 54, no. 12, pp. 152–139, Dec. 2016.
[2] S. Zhang, “An overview of network slicing for 5G,” IEEE Wireless Commun., vol. 26, no. 3, pp. 111–117, Jun. 2019.
[3] HUAWEI. (Jan. 2021). 5G ToB service Experience Standard White Paper. [Online]. Available: https://carrier.huawei.com/~media/cnbvlg2/download/products/servies/5g-b2b-service-experience-standard-white-paper-en1.pdf
[4] 5G Americas. (Nov. 2018). New Services & Applications With 5G Ultra-Reliable Low Latency Communications. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLCC_White_Paper_Final_updatedV2.pdf
[5] J. Lorca et al., “Scenarios, KPIs, use cases and baseline system evaluation,” E2E-aware Optimizations and Advancements for Network Edge of 5G New Radio (ONE5G), 5G Infrastructure Public Private Partnership (5G PPP), Tech. Rep. ONE5G D1.2, Nov. 2017.
[6] 5G Americas. (Dec. 2022). Becoming 5G-Advanced: The 3GPP 2025 Roadmap. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2022/12/Becoming-5G-Advanced-the-3GPP-2025-Roadmap-InDesign.pdf
[7] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. M. Leung, “Network slicing based on heterogeneous mobile networks: Mobility, resource management, and challenges,” IEEE Commun. Mag., vol. 55, no. 8, pp. 138–145, Aug. 2017.
[8] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, “Network slicing in 5G: Survey and challenges,” IEEE Commun. Mag., vol. 55, no. 5, pp. 94–100, May 2017.
[9] F. Chiarotti et al., “Spectrum slicing for multiple access channels with heterogeneous services,” Entropy, vol. 23, no. 6, p. 686, 2021. [Online]. Available: https://www.mdpi.com/1099-4302/23/6/686
[10] 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. 55765–55779, 2018.
[11] E. N. Tominaga, H. Alves, R. D. Souza, J. Luiz Rebelatto, and M. Latva-aho, “Non-orthogonal multiple access and network slicing: Scalable coexistence of eMBB and URLLC,” in Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring), Apr. 2021, pp. 1–6.
[12] E. N. Tominaga, H. Alves, O. L. A. López, R. D. Souza, J. L. Rebelatto, and M. Latva-aho, “Network slicing for eMBB and mMTC with NOMA and space diversity reception,” in Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring), Apr. 2021, pp. 1–6.
[13] 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), Dec. 2018, pp. 1–6.
[14] F. Saggese, M. Moretti, and P. Popovski, “NOMA power minimization of downlink spectrum slicing for eMBB and URLLC users,” 2021.
[15] A. K. Bairagi et al., “Coexistence mechanism between eMBB and uURLC in 5G wireless networks,” IEEE Trans. Commun., vol. 69, no. 3, pp. 1736–1749, Mar. 2021.
[16] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, “eMBB-URLLC resource slicing: A risk-sensitive approach,” IEEE Commun. Lett., vol. 23, no. 4, pp. 740–743, Apr. 2019.
[17] A. Anand, G. de Veciana, and S. Shakkottai, “Joint scheduling of eMBB and URLLC in 5G wireless networks,” IEEE/ACM Trans. Netw., vol. 26, no. 12, pp. 5776–5789, Dec. 2018.
[18] R. Kassab, O. Simeone, and P. Popovski, “Coexistence of URLLC and eMBB services in 5G wireless networks,” IEEE Trans. Commun., vol. 69, no. 3, pp. 1736–1749, Mar. 2021.
[19] A. Niyato, D. I. Kim, and Z. Han, “Network slicing for eMBB and mMTC in 5G wireless networks,” IEEE Commun. Lett., vol. 24, no. 5, pp. 658–661, May 2020.
Yuanwen Liu received the B.Eng. degree in telecommunications engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2019, and the M.S. degree in electrical engineering from Northwestern University, Evanston, USA, in 2021. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, Imperial College London, U.K. Her research interests include rate splitting multiple access.

Bruno Clerckx (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from Universit Catholique de Louvain, Belgium, and the Doctor of Science (D.Sc.) degree from Imperial College London, U.K. He is currently a (Full) Professor, the Head of the Wireless Communications and Signal Processing Laboratory, and the Deputy Head of the Communications and Signal Processing Group, Electrical and Electronic Engineering Department, Imperial College London. He is also the Chief Technology Officer (CTO) of Silicon Austria Laboratories (SAL), where he is responsible for all research areas of Austria’s top research center for electronic-based systems. Prior to joining Imperial College in 2011, he was with Samsung Electronics, Suwon, South Korea, where he actively contributed to 4G (3GPP LTE/LTE-A and IEEE 802.16m). He has authored two books on MIMO wireless communications and MIMO wireless networks, 250 peer-reviewed international research papers, and 150 standards contributions. He is an 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 interests include 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 the Royal Academy of Belgium, multiple awards from Samsung, IEEE Best Student Paper Award, and the European Association for Signal Processing (EURASIP) Best Paper Award 2022. He is a fellow of the IET and an IEEE Communications Society Distinguished Lecturer.

Petar Popovski (Fellow, IEEE) received the Dipl.-Ing. and M.Sc. degrees in communication engineering from S. Cyril and Methodius University in Skopje and the Ph.D. degree from Aalborg University in 2005. He is currently a Professor with Aalborg University, where he heads the Section on Connectivity and a Visiting Excellence Chair of the University of Bremen. He authored the book "Wireless Connectivity: An Intuitive and Fundamental Guide." His research interests include wireless communication and communication theory. He received an ERC Consolidator Grant in 2015, the Danish Elite Researcher Award in 2016, the IEEE Fred W. Ellersick Prize in 2016, the IEEE Stephen O. Rice Prize in 2018, the Technical Achievement Award from the IEEE Technical Committee on Smart Grid Communications in 2019, the Danish Telecommunication Prize in 2020, and Villum Investigator Grant in 2021. He was a Member at Large at the Board of Governors in IEEE Communication Society (2019–2021). He is also an Editor-in-Chief of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and the Chair of the IEEE Communication Theory Technical Committee. He was the General Chair of IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019.