Quasi-Isolated Network Slicing for Multi-Access Edge Computing

Vasiliki I. Koutsioumpa, Nikos A. Mitsiou, *Graduate Student Member, IEEE*, Sotiris A. Tegos, *Senior Member, IEEE*, Panagiotis D. Diamantoulakis, *Senior Member, IEEE*, Panagiotis G. Sarigiannidis, *Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—Network slicing via next-generation multiple access techniques and multi-access edge computing (MEC) are considered key enablers for meeting the heterogeneous quality of service requirements of the sixth-generation (6G) networks. Thus, in this work, we investigate the coexistence of further enhanced mobile broadband (feMBB) and ultra-massive machine-type communications (umMTC) devices in a quasi-isolated (QI) heterogeneous uplink MEC network, where users of both services share the same resources, interfering with each other. The feMBB users can partially offload their data to the MEC server utilizing the rate-splitting multiple access (RSMA) protocol, while the umMTC users perform only full offloading. We formulate and optimally solve the problem of maximizing the number of umMTC devices subject to data processing time and data rate constraints by adjusting both the decoding order of the users and the power splitting factor of the RSMA, while a closed-form expression for the optimal partial offloading factor of the feMBBs devices is derived. Simulation results verify that utilizing the RSMA protocol, the QI MEC network has the potential to support more umMTC devices compared to the isolated one.**

***Index Terms*—QI network, MEC, RSMA, feMBB, umMTC**

I. Introduction

Although the global deployment of fifth-generation (5G) networks is still in its infancy, the development of sixth-generation (6G) networks has already intrigued the scientific community. It is expected that 6G will build on the capabilities of 5G and provide further enhancements [1]. Existing services will be expanded to include further enhanced mobile broadband (feMBB), ultra-massive machine-type communications (umMTC), and extremely reliable low-latency communications (ERLLC). In more detail, feMBB focuses on extremely high data rates, enhanced connectivity, and higher user mobility with high reliability. The focus of umMTC is to provide connectivity for a large number of occasionally active Internet of Things (IoT) devices that transmit small data payloads and are constrained by cost and power consumption, while the goal of ERLLC is to enable fast and reliable transmission of small data packets with extremely low latency.

V. I. Koutsioumpa, N. A. Mitsiou, S. A. Tegos, and P. D. Diamantoulakis are with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mails: vasioakou@ece.auth.gr, nmitsiou@ece.auth.gr, tegosoti@auth.gr, padiaman@auth.gr).

P. G. Sarigiannidis is with the Department of Electrical and Computer Engineering, University of Western Macedonia, 50100, Kozani, Greece (e-mail: psarigiannidis@uowm.gr).

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece and also with the Artificial Intelligence & Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (e-mail: geokarag@auth.gr).

This work has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) European Union's Horizon Europe research and innovation programme under Grant Agreement No 101096456 (NANCY).

To support customized, on-demand network services, the concept of network slicing has been proposed [2], which enables the coexistence of different use cases with heterogeneous requirements on the same network infrastructure. Furthermore, to bring intelligence closer to the end user, there has been a significant shift to multi-access edge computing (MEC) instead of centralized mobile cloud computing, providing computing capabilities at the edge of the network [3]. This allows compute-intensive and latency-sensitive applications to run on mobile devices with limited resources, improving end-to-end network latency and user quality of service (QoS).

Multiple access is a key enabler to efficiently manage and allocate resources to different users and services sharing the same network infrastructure, while non-orthogonal protocols have been proposed to complement the orthogonal ones due to their increased connectivity. In particular, non-orthogonal multiple access (NOMA) in a heterogeneous uplink network with one eMBB and multiple URLLC or mMTC users was studied in [4], [5] and it was shown that NOMA-aided slicing outperforms orthogonal multiple access (OMA)-aided slicing. In [6], network slicing was proposed in a NOMA-enabled MEC network that provides improvements in latency, energy, and spectral efficiency over its OMA counterpart. Rate-splitting multiple access (RSMA) is also a promising technique that can handle heterogeneity in wireless networks [7]–[12] and can achieve the entire capacity region [13]. RSMA can provide significant gains in terms of massive connectivity [7], ergodic rate [8], user fairness and outage probability [9]. A cognitive radio (CR)-inspired RSMA-MEC scheme was designed in [10], which outperformed its NOMA counterpart in terms of successful offload probability. Furthermore, in [11], [12], an RSMA-based slicing was studied for one eMBB and multiple mMTC users, and for one eMBB and multiple URLLC users, respectively. Therefore, although the use of RSMA for heterogeneous network slicing [11], [12] and NOMA for heterogeneous MEC systems [6] have been studied, the use of RSMA for heterogeneous MEC with network slicing has yet to be investigated.

To this end, we investigate the RSMA scheme for an uplink quasi-isolated (QI) MEC network with two active feMBB users capable of partial offloading and multiple umMTC users performing full offloading. Due to the fact that users of both services coexist and interfere with each other, we introduce the term QI to describe such heterogeneous networks. Our primary interest is to maximize the number of active umMTC devices under given QoS and data processing time requirements, introduced by MEC, of all users. For the RSMA scheme, the optimal decoding order and the optimal power sharing factor of the feMBB devices are investigated, while also deriving a closed-form expression for the offloading factor of the feMBB users. Simulation results show that by utilizing RSMA, the QI edge network has the potential to support more umMTC devices compared to the isolated network, while the feMBB devices can offload a larger part of their computation tasks.

II. System model

We assume a QI heterogeneous uplink network consisting of two permanently active feMBB users, multiple umMTC devices, and a base station (BS) co-located with a MEC server. The number of active umMTC devices is a random variable denoted by n_{M} and follows the Poisson distribution, hence $n_{\mathrm{M}} \sim \text{Poisson}(\lambda_{\mathrm{M}})$, where λ_{M} is the mean and is referred to as the arrival rate of umMTC devices. Both feMBB and umMTC users are assigned to a single frequency channel with bandwidth B. This frequency channel is assumed to be within the time and frequency coherence interval of the channel, thus the channel coefficients are assumed to be constant. The channel coefficients of feMBB user i and umMTC user m are given by $h_{\mathrm{B}_i} \in \mathbb{C}$, where $i \in \{1,2\}$, and $h_{\mathrm{M}_m} \in \mathbb{C}$, where $m \in \{1,\dots,n_M\}$, respectively. Rayleigh fading is assumed, i.e., $h_{\mathrm{B}_i} \sim \mathcal{CN}(0,\Gamma_{\mathrm{B}})$ and $h_{\mathrm{M}_m} \sim \mathcal{CN}(0,\Gamma_{\mathrm{M}})$, and the channel gains, denoted by $G_{\mathrm{B}_i} = |h_{\mathrm{B}_i}|^2$ and $G_{\mathrm{M}_m} = |h_{\mathrm{M}_m}|^2$, follow the exponential distribution. The constants Γ_{B} and Γ_{M} are the average channel gains, which include the path loss.

The feMBB devices can also offload part of their computing tasks to the MEC server. Each task is represented by a tuple $\mathcal{T}_i = \{L_i,\ X_i,\ d_{\max_i}\}$, $i \in \{1.2\}$, where L_i is the task input data size in bits, X_i is the computational workload/intensity in CPU cycles per bit, and $d_{\max_i}$ is the required completion time in seconds, which is also considered as a QoS requirement and is associated with a rate threshold $r_{\mathrm{B}}^{\mathrm{th}}$. The umMTC devices are assumed to perform only full offloading, and their QoS requirement is defined by a given rate threshold $r_{\mathrm{M}}^{\mathrm{th}}$.

A. Achievable rates of feMBB and umMTC users

The umMTC and feMBB devices are sorted by their channel gains in a descending order, i.e., $G_{\mathrm{M}_1} \geq G_{\mathrm{M}_2} \geq \dots \geq G_{\mathrm{M}_{n_{\mathrm{M}}}}$ and $G_{\mathrm{B}_1} \geq G_{\mathrm{B}_2}$, respectively, and their messages are encoded into streams s_{M_m}, $m \in \{1,\dots,n_{\mathrm{M}}\}$, and s_{B_i}, $i \in \{1,2\}$, and transmitted to the MEC server. The umMTC users' messages are decoded with the above order. For the feMBB devices, utilizing the RSMA protocol, only one user splits the message in an isolated network [13]. However, in this QI network we assume that both users can split their messages into two sub-streams, $s_{\mathrm{B}_{i,1}}$, $s_{\mathrm{B}_{i,2}}$, $i \in \{1,2\}$, and we assume that $s_{\mathrm{B}_{i,1}}$ is always decoded first. The total rate of feMBB users is also split into two sub-rates, $r_{\mathrm{B}_{i,1}}$ and $r_{\mathrm{B}_{i,2}}$, respectively. For feMBB users' messages, there are six possible decoding orders, given in (1). The SINR expressions are derived using the first decoding order and can be adapted to any other order.

Let $\alpha \in [0,1]$ be the power factor for $s_{\mathrm{B}_{1,1}}$ and $\beta \in [0,1]$ for $s_{\mathrm{B}_{2,1}}$, so that $s_{\mathrm{B}_{1,2}}$ and $s_{\mathrm{B}_{2,2}}$ are transmitted with $1-\alpha$ and $1-\beta$ of the total power, respectively. The received signal at the base station is

$$\begin{aligned} y = &\sqrt{\alpha}\, h_{\mathrm{B}_1} s_{\mathrm{B}_{1,1}} + \sqrt{1-\alpha}\, h_{\mathrm{B}_1} s_{\mathrm{B}_{1,2}} + \sqrt{\beta}\, h_{\mathrm{B}_2} s_{\mathrm{B}_{2,1}} \\ &+ \sqrt{1-\beta}\, h_{\mathrm{B}_2} s_{\mathrm{B}_{2,2}} + \sum_{m=1}^{n_{\mathrm{M}}} h_{\mathrm{M}_m} s_{\mathrm{M}_m} + n, \end{aligned} \tag{2}$$

where n is the additive white Gaussian noise (AWGN). We assume that the noise power is normalized.

The SINR expressions of all feMBB users' streams can be written as

$$\gamma_{\mathrm{B}_{1,1}} = \frac{\alpha G_{\mathrm{B}_1}}{1 + G_{\mathrm{B}_2} + (1-\alpha) G_{\mathrm{B}_1} + \sum_{m=m_1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{3a}$$

$$\gamma_{\mathrm{B}_{2,1}} = \frac{\beta G_{\mathrm{B}_2}}{1 + (1-\beta) G_{\mathrm{B}_2} + (1-\alpha) G_{\mathrm{B}_1} + \sum_{m=m_2}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{3b}$$

$$\gamma_{\mathrm{B}_{2,2}} = \frac{(1-\beta) G_{\mathrm{B}_2}}{1 + (1-\alpha) G_{\mathrm{B}_1} + \sum_{m=m_3}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{3c}$$

$$\gamma_{\mathrm{B}_{1,2}} = \frac{(1-\alpha) G_{\mathrm{B}_1}}{1 + \sum_{m=m_4}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{3d}$$

where i) umMTC users indexed by 1 to $m_1 - 1$ are decoded before all feMBB users, ii) users m_1 to $m_2 - 1$ are decoded between $s_{\mathrm{B}_{1,1}}$ and $s_{\mathrm{B}_{2,1}}$, iii) users m_2 to $m_3 - 1$ are decoded between $s_{\mathrm{B}_{2,1}}$ and $s_{\mathrm{B}_{2,2}}$, iv) users m_3 to $m_4 - 1$ are decoded between $s_{\mathrm{B}_{2,2}}$ and $s_{\mathrm{B}_{1,2}}$, and v) users m_4 to n_{M} are decoded after all feMBB users. Using these expressions, the corresponding rates are given by

$$r_{\mathrm{B}_n} = \log_2(1+\gamma_{\mathrm{B}_n})\ (\text{bps/Hz}),\ n \in \{1{,}1,\ 2{,}1,\ 2{,}2,\ 1{,}2\}. \tag{4}$$

The BS uses successive interference cancellation (SIC) to decode the feMBB and umMTC users' messages. First, the umMTC devices' messages are decoded. If a message cannot be decoded, the BS decodes $s_{\mathrm{B}_{1,1}}$ and then retries to decode the umMTC devices' messages. If the umMTC message cannot be decoded again, the BS decodes $s_{\mathrm{B}_{2,1}}$ and then continues with the umMTC devices. If the message of any umMTC device cannot be decoded, $s_{\mathrm{B}_{2,2}}$ is tried to be decoded. If the message of feMBB user 2 is successfully decoded, the BS continues decoding the remaining umMTC devices. If a message cannot be decoded again, the BS tries to decode $s_{\mathrm{B}_{1,2}}$. If the message of feMBB user 1 is successfully decoded, the decoding of the remaining umMTC users follows. The decoding procedure terminates, if the message of any feMBB or umMTC device cannot be decoded or if the messages of all devices have been decoded. The feMBB users are decoded, when $r_{\mathrm{B}_i} = r_{\mathrm{B}_{i,1}} + r_{\mathrm{B}_{i,2}} \geq r_{\mathrm{B}}^{\mathrm{th}}$, $i \in \{1,2\}$, while a umMTC device m_0 is decoded, if $r_{\mathrm{M}_{m_0}} \geq r_{\mathrm{M}}^{\mathrm{th}}$ holds. For convenience, the decoding procedure is summarized in Algorithm 1.

As mentioned above, there are five possible decoding cases for umMTC users. Thus, the SINR of umMTC user m_0 in each possible decoding case is given by

$$\gamma_{\mathrm{M}_{m_0}}^{\mathrm{i}} = \frac{G_{\mathrm{M}_{m_0}}}{1 + G_{\mathrm{B}_1} + G_{\mathrm{B}_2} + \sum_{m=m_0+1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{5a}$$

$$\begin{aligned} &\mathrm{r}_1\!: s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{2,2}} \to s_{\mathrm{B}_{1,2}}, \quad \mathrm{r}_2\!: s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{1,2}} \to s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{2,2}}, \quad \mathrm{r}_3\!: s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{1,2}} \to s_{\mathrm{B}_{2,2}}, \\ &\mathrm{r}_4\!: s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{1,2}} \to s_{\mathrm{B}_{2,2}}, \quad \mathrm{r}_5\!: s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{2,2}} \to s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{1,2}}, \quad \mathrm{r}_6\!: s_{\mathrm{B}_{2,1}} \to s_{\mathrm{B}_{1,1}} \to s_{\mathrm{B}_{2,2}} \to s_{\mathrm{B}_{1,2}}. \end{aligned} \tag{1}$$

This article has been accepted for publication in IEEE Communications Letters. 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/LCOMM.2024.3374816

Algorithm 1 Decoding procedure at the BS

Attempt to decode the messages of umMTC devices first
if $r_{\mathrm{M}_{m_0}} < r_{\mathrm{M}}^{\mathrm{th}}$ **then**
  Decode $s_{\mathrm{B}_{1,1}}$ at a rate $r_{\mathrm{B}_{1,1}}$ and continue with the messages of the remaining umMTC devices
  if $r_{\mathrm{M}_{m_0}} < r_{\mathrm{M}}^{\mathrm{th}}$ **then**
    Attempt to decode $s_{\mathrm{B}_{2,1}}$ at a rate $r_{\mathrm{B}_{2,1}}$ and continue with the remaining umMTC devices
    if $r_{\mathrm{M}_{m_0}} < r_{\mathrm{M}}^{\mathrm{th}}$ **then**
      Attempt to decode $s_{\mathrm{B}_{2,2}}$ at a rate $r_{\mathrm{B}_{2,2}}$
      if $r_{\mathrm{B}_{2,1}} + r_{\mathrm{B}_{2,2}} < r_{\mathrm{B}}^{\mathrm{th}}$ **then**
        The decoding procedure terminates
      else
        Continue with the messages of the remaining umMTC devices
        if $r_{\mathrm{M}_{m_0}} < r_{\mathrm{M}}^{\mathrm{th}}$ **then**
          Calculate $r_{\mathrm{B}_{1,2}}$
          if $r_{\mathrm{B}_{1,1}} + r_{\mathrm{B}_{1,2}} < r_{\mathrm{B}}^{\mathrm{th}}$ **then**
            The decoding procedure terminates
          else
            Continue with the messages of the remaining umMTC devices
            if Another umMTC device cannot be decoded or all devices have been decoded **then**
              The decoding procedure terminates

$$\gamma_{\mathrm{M}_{m_0}}^{\mathrm{ii}} = \frac{G_{\mathrm{M}_{m_0}}}{1+(1-\alpha)G_{\mathrm{B}_1}+G_{\mathrm{B}_2}+\sum_{m=m_0+1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{5b}$$

$$\gamma_{\mathrm{M}_{m_0}}^{\mathrm{iii}} = \frac{G_{\mathrm{M}_{m_0}}}{1+(1-\alpha)G_{\mathrm{B}_1}+(1-\beta)G_{\mathrm{B}_2}+\sum_{m=m_0+1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{5c}$$

$$\gamma_{\mathrm{M}_{m_0}}^{\mathrm{iv}} = \frac{G_{\mathrm{M}_{m_0}}}{1+(1-\alpha)G_{\mathrm{B}_1}+\sum_{m=m_0+1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{5d}$$

$$\gamma_{\mathrm{M}_{m_0}}^{\mathrm{v}} = \frac{G_{\mathrm{M}_{m_0}}}{1+\sum_{m=m_0+1}^{n_{\mathrm{M}}} G_{\mathrm{M}_m}}, \tag{5e}$$

and its rate is given by

$$r_{\mathrm{M}_{m_0}} = \log_2(1+\gamma_{\mathrm{M}_{m_0}}^{c}) \ \ (\mathrm{bps/Hz}), \ \ c \in \{\mathrm{i, \ ii, \ iii, \ iv, \ v}\}. \tag{6}$$

Let $D_{\mathrm{M}} \in \{0, ..., n_{\mathrm{M}}\}$ be the random variable denoting the number of successfully decoded umMTC users. We define the average connectivity ratio of umMTC users as the ratio of the expectation of successfully decoded users to the arrival rate of all active users and is given by

$$\mathrm{ACR} = \frac{\mathbb{E}[D_{\mathrm{M}}]}{\lambda_{\mathrm{M}}}, \tag{7}$$

with $\mathbb{E}[\cdot]$ denoting expectation.

B. Partial offloading of feMBB users

Let $\theta_i \in [0,1]$, $i \in \{1,2\}$, be the offloading factor of an feMBB device and $1-\theta_i$ be the amount of the locally processed task. The data transmission time of the i-th feMBB user is given by

$$d_{\mathrm{t}_i} = \frac{\theta_i L_i}{r_{\mathrm{B}_i} B}, \tag{8}$$

while the local computation time can be written as

$$d_{\mathrm{c}_i} = \frac{(1-\theta_i) L_i X_i}{f_i}, \tag{9}$$

where f_i is the frequency of the CPU cycles. Since the data transmission between an feMBB user and the MEC server can occur in parallel with local execution, the total time to complete the processing of the task is then calculated as

$$d_i = \max\{d_{\mathrm{t}_i}, d_{\mathrm{c}_i}\}, \tag{10}$$

In addition, MEC systems require that the offloading and the local computations occur within a certain time period, hence

$$d_i \leq d_{\max_i} \implies d_{\mathrm{t}_i} \leq d_{\max_i} \ \text{ and } \ d_{\mathrm{c}_i} \leq d_{\max_i}. \tag{11}$$

The per-device probability that data processing time does not exceed the required completion time can be written as

$$\Pr(E_{\mathrm{B}_i}) = \Pr(d_i \leq d_{\max_i}), \tag{12}$$

where E_{B_i}, $i \in \{1,2\}$ is the event that the processing time of the i-th feMBB user's data is below a threshold.

III. Number of umMTC devices maximization

One of the key characteristics of network slicing is the isolation among different services, according to which scaling of resources for a certain slice will not come at the minimum impact on the services of another slice. A straightforward way to fully isolate the different slices is through the isolation of orthogonal resources, e.g., time or bandwidth. However, to improve the overall QoS while retaining a sufficient level of isolation between the different slices, the use of the QI network slicing is proposed, which allows different slices to interfere with each other, as long as the QoS constraints are met. Thus, we aim to maximize the total number of umMTC devices that the QI edge network can support considering the QoS requirements of both services as well as the data processing time requirements of the feMBB users. The connectivity requirement of umMTC users and the reliability level of feMBB users are ϵ_{M} and ϵ_{B}, respectively. The optimization problem is then formulated as

$$\begin{aligned} \max_{\lambda_{\mathrm{M}}, \alpha, \beta, \theta} \quad & \lambda_{\mathrm{M}} \geq 0 \\ \text{s.t.} \quad & C_1: \ \mathrm{ACR} \geq \epsilon_{\mathrm{M}} \\ & C_2: \ \Pr(E_{\mathrm{B}_i}) \geq \epsilon_{\mathrm{B}}, \quad i \in \{1,2\} \\ & C_3: \ d_i \leq d_{\max_i}, \qquad i \in \{1,2\} \\ & C_4: \ \alpha, \beta, \theta \in [0,1]. \end{aligned} \tag{13}$$

Replacing (8) and (9) in constraint C_3, we can obtain an upper and a lower bound of the offloading factor θ_i as

$$1 - \frac{d_{\max_i} \ f_i}{L_i \ X_i} \leq \theta_i \leq \frac{r_{\mathrm{B}_i} \ B \ d_{\max_i}}{L_i}. \tag{14}$$

The transmission rate of a user can be increased by increasing the SINR. At fixed transmit power, an increase in the SINR of an feMBB user can be achieved by reducing the interference it experiences, which implies a reduction in the number of active umMTC users. However, our goal is to maximize the number of umMTC devices. It should be noted that the more umMTC devices that can be supported in the system, the more interference they cause, and the rate of feMBB devices is equal to the threshold. Thus, for a fixed data processing time equal to the maximum, the optimal offloading factor must take the minimum value, which is given by (14) as

$$\theta_i^* \ = \ 1 - \frac{d_{\max_i} \ f_i}{L_i \ X_i}. \tag{15}$$

Thus, from (14), the rate of feMBB users can be bounded by

$$r_{\mathrm{B}_i} \geq \frac{1}{B}\left(\frac{L_i}{d_{\max_i}} - \frac{f_i}{X_i}\right), \tag{16}$$

where equality holds for the rate threshold $r_{\mathrm{B}_i}^{\mathrm{th}}$. It can be observed that the rate of feMBB users depends on the maximum task completion time and vice versa. Therefore, the probability in (12) can be equivalently expressed as the probability that the feMBB users' rate exceeds a threshold and is given by

$$\Pr(E_{\mathrm{B}_i}) = \Pr\left(r_{\mathrm{B}_i} \geq \frac{1}{B}\left(\frac{L_i}{d_{\max_i}} - \frac{f_i}{X_i}\right)\right). \tag{17}$$

To obtain the optimal λ_{M}^*, the formulated optimization problem can be written as

$$\begin{aligned} \max_{\lambda_{\mathrm{M}},\alpha,\beta} \quad & \lambda_{\mathrm{M}} \geq 0 \\ \text{s.t.} \quad & C_1: \frac{\mathbb{E}[D_{\mathrm{M}}]}{\lambda_{\mathrm{M}}} \geq \epsilon_{\mathrm{M}} \\ & C_2: \Pr\left(r_{\mathrm{B}_i} \geq \frac{1}{B}\left(\frac{L_i}{d_{\max_i}} - \frac{f_i}{X_i}\right)\right) \geq \epsilon_{\mathrm{B}}, \ \ i \in \{1,2\} \\ & C_3: \alpha, \beta \in [0,1]. \end{aligned} \tag{18}$$

We note that problem (18) is not tractable by any known optimization method or machine learning tool, since the decoding order for the feMBB and umMTC users is unknown a priori and can change between two different Monte Carlo runs. The decoding order affects the data rate of each user, thus leading to a different optimization problem each time. Nevertheless, to solve this challenging optimization problem, an efficient algorithm is provided and illustrated in Algorithm 2. The optimal solution λ_{M}^* can be obtained by a Monte Carlo approach combined with a line search over sets $\mathcal{M}$ and $\mathcal{S}$, which contain all possible values of the number of umMTC users and the splitting factors of feMBB users, respectively.

Algorithm 2 Solution of problem (18)

```
1:  Initialize simulation parameters, M, S r_B^th, d_max
2:  for n ∈ M do
3:      for α ∈ S do
4:          for β ∈ S do
5:              for each Monte Carlo iteration do
6:                  Follow Algorithm 1 and calculate D_M
7:              Compute E[D_M] and Pr(E_Bi), i ∈ {1,2}
8:              if constraints C1 and C2 are satisfied then
9:                  λ ← n
10:             else
11:                 λ ← 0
12:     λ_M ← max{λ}
13: λ_M^* ← max{λ_M}
```

The computational complexity of Algorithm 1 is $\mathcal{O}(n)$ and of Algorithm 2 is $\mathcal{O}(|\mathcal{S}|^2 \mathcal{I} n^2)$, where $|\cdot|$ denotes the cardinality of set $\mathcal{S}$ and $\mathcal{I}$ is the number of Monte Carlo iterations. However, these two constants can be omitted, hence the complexity can be written as $\mathcal{O}(n^2)$. We note that the optimal α and β depend on the distribution of channel coefficients and not on the instantaneous channel gains, thus for given system parameters and topology, the algorithm can be utilized only once, which reduces the system's complexity.

IV. Numerical results and discussion

In this section, we evaluate the performance of the proposed RSMA scheme in the considered QI heterogeneous MEC system and compare it with the isolated counterpart, which is considered as benchmark, and where feMBB users are allocated to the same resources for a time fraction $\tau \in [0,1]$ and can utilize the RSMA and OMA protocols. In both systems, the NOMA protocol appears as a special case of the RSMA, when $\alpha, \beta \in \{0,1\}$ concurrently, and therefore it is not considered in comparison. An efficient benchmark from literature does not also exist. The parameters used for the simulations are mostly based in [11] and given in Table I. We assume that the tasks of both feMBB users are represented

TABLE I: Simulation parameters.

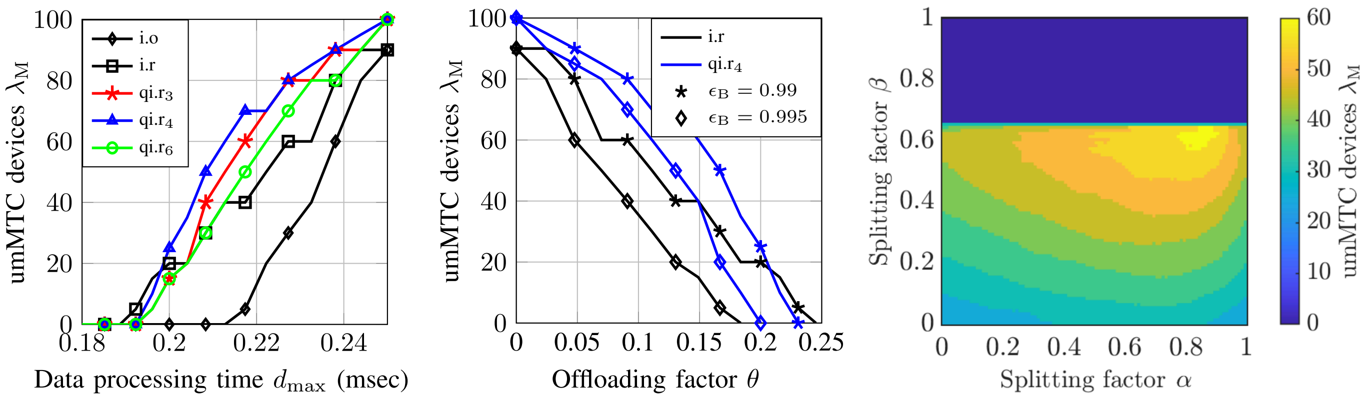
Parameter	Value	Parameter	Value
Γ_{B}	25 dB	L	1 kbit
Γ_{M}	10 dB	X	250 CPU cycles/bit
ϵ_{B}	0.99	f	1 GHz
ϵ_{M}	0.9	$\mathcal{I}$	10^4
$r_{\mathrm{M}}^{\mathrm{th}}$	0.04 bps/Hz	$\mathcal{M}$	$\{1,2,5,10,...,100\}$
B	1 MHz	$\mathcal{S}$	$\{0, 0.01, ..., 1\}$

by the same tuple $\mathcal{T}$, i.e., they have the same amount of data, computational workload, and maximum latency, as well as the same CPU cycle frequency and target rate. In the following figures, "i" and "qi" refer to the isolated and the QI MEC network, respectively, while "o" and "r_k", $k \in \{1,\ldots,6\}$, denote that the feMBB users utilize the OMA and the RSMA protocol with the decoding order k as declared in (1), respectively.

In Fig. 1a, the number λ_{M} of umMTC devices that can be supported utilizing RSMA with the best three decoding orders in this QI heterogeneous system is plotted for different values of the maximum task completion time $d_{\max}$ and compared to the isolated system utilizing OMA and RSMA. The data processing time and the rate are inversely proportional. Note that in the isolated network the RSMA scheme outperforms the OMA counterpart, because the resources can be utilized more efficiently compared to OMA. Moreover, it is observed that in the QI system the optimal decoding order in the RSMA scheme is r_4, while r_6 is equal to the optimal according to literature, where $\beta = 0$ and only one user splits its message. Both services share the same resources, thus the differences between the two services' channel gains and QoS requirements can be exploited at the BS and more umMTC devices can be supported. Furthermore, for $0.185 < d_{\max} < 0.2$, the RSMA scheme in the isolated network is more efficient than in the QI, since the rate threshold of the feMBB devices is relatively high. In the QI network, the feMBB devices cannot tolerate much interference from umMTC devices, thus to meet their QoS requirement, the number of active umMTC devices is reduced, while, in the isolated counterpart, there is no umMTC interference to the feMBB devices.

In Fig. 1b, the relationship between the task offloading factor of feMBB users and the arrival rate of umMTC users is examined for the RSMA protocol in both the QI and the isolated system for two different values of the feMBB users' reliability requirement ϵ_{B}. In general, a larger amount of data has to be offloaded in the QI MEC network compared to the

This article has been accepted for publication in IEEE Communications Letters. 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/LCOMM.2024.3374816



(a) Number of active umMTC devices vs maximum task completion time.

(b) Number of active umMTC users vs feMBBs users' offloading factor.

(c) Number of umMTC devices in the QI-MEC network vs the power splitting factors.

Fig. 1: Numerical evaluation of the proposed QI system.

isolated counterpart, for a given number of λ_{M}. It should be noted that an increase in the offloading factor implies a decrease in the number of active umMTC devices. This is in agreement with (15), since with a decrease in the data processing time the offloading factor increases, and with the insights of Fig. 1a, where it appears that the arrival rate of umMTCs decreases with a decrease in the data processing time. Moreover, it is illustrated that for a given value of θ, increasing the value of ϵ_{B}, the arrival rate of the umMTC users decreases. In this case, it is also observed that the coexistence of both services in the QI network is always more efficient compared to the isolated counterpart, as regards the number of active umMTC devices.

Fig. 1c illustrates the number of active umMTC devices versus both splitting factors of the feMBB devices utilizing the RSMA scheme with the optimal decoding order in the QI MEC network. We assume $r_{\mathrm{B}}^{\mathrm{th}} = 0.7$ bps/Hz and $d_{\max} = 0.2128$ msec. This figure consists of level curves with different values of λ_{M}. It is noted that for $\beta \geq 0.66$ very few or no umMTC devices can be supported. This is mainly because for large values of β, s_{B_2} degenerates into $s_{\mathrm{B}_{2,1}}$ which suffers from high interference. Thus, feMBB user 2 cannot satisfy its QoS requirements to decode its message leading to the absence of umMTC devices in the QI network. Additionally, $r_{\mathrm{B}}^{\mathrm{th}} = 0.7$ bps/Hz is relatively high and some feMBB or umMTC devices do not have enough SINR to be decoded. By adjusting the power splitting factor of feMBB devices, the SINR of both feMBB and umMTC users is also adjusted, allowing more umMTC devices to be supported in this system.

V. Conclusion

In this work, the coexistence of two feMBB and multiple active umMTC users in a QI uplink MEC network with RSMA has been investigated. By splitting messages, adjusting the splitting factor, and selecting the optimal decoding order and the proper system parameters, part of the interference caused by the feMBB users can be canceled, so that a larger number of umMTC devices can be supported in this network compared to the isolated counterpart. A closed-form expression for the offloading factor of the feMBB users was also derived, as well as a trade-off between the amount of their offloaded data and the maximization of the number of umMTC devices.

References

[1] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[2] S. Zhang, "An Overview of Network Slicing for 5G," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, Jun. 2019.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[4] 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, Sep. 2018.

[5] A. A. Tegos, S. A. Tegos, D. Tyrovolas, P. D. Diamantoulakis, P. Sarigiannidis, and G. K. Karagiannidis, "Breaking Orthogonality in Uplink With Randomly Deployed Sources," *IEEE Open J. Commun. Soc.*, 2023.

[6] M. A. Hossain and N. Ansari, "Network Slicing for NOMA-Enabled Edge Computing," *IEEE Trans. Cloud Comput*, vol. 11, no. 1, pp. 811–821, 2023.

[7] S. A. Tegos, P. D. Diamantoulakis, and G. K. Karagiannidis, "On the Performance of Uplink Rate-Splitting Multiple Access," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 523–527, Mar. 2022.

[8] Y. Xiao, S. A. Tegos, P. D. Diamantoulakis, Z. Ma, and G. K. Karagiannidis, "On the Ergodic Rate of Cognitive Radio Inspired Uplink Multiple Access," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 95–99, Jan. 2023.

[9] H. Liu, T. Tsiftsis, K. J. Kim, K. Kwak, and H. V. Poor, "Rate Splitting for Uplink NOMA With Enhanced Fairness and Outage Performance," *IEEE Trans. Wireless Commun.*, vol. 26, Mar. 2020.

[10] P. Chen, H. Liu, Y. Ye, L. Yang, K. J. Kim, and T. A. Tsiftsis, "Rate-Splitting Multiple Access Aided Mobile Edge Computing With Randomly Deployed Users," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1549–1565, May 2023.

[11] Y. Liu, B. Clerckx, and P. Popovski, "Network Slicing for eMBB, URLLC, and mMTC: An Uplink Rate-Splitting Multiple Access Approach," *IEEE Trans. Wireless Commun.*, pp. 1–1, Jul. 2023.

[12] E. J. D. 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, Dec. 2021.

[13] 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.