This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

Resource Scheduling for eMBB and URLLC Multiplexing in NOMA-Based VANETs: A Dual Time-Scale Approach

Yujie Zhang, *Student Member*, *IEEE*, Nan Cheng, *Member*, *IEEE*, Yanpeng Dai, *Member*, *IEEE*, Zhisheng Yin, *Member*, *IEEE*, Wei Quan, *Member*, *IEEE*, Yi Zhou, *Member*, *IEEE*, Ning Zhang *Senior Member*, *IEEE*

***Abstract*—Enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) are two critical services in vehicular networks. However, the presence of both services creates a difficult resource allocation problem due to their heterogeneous requirements. To address the challenge of simultaneously providing eMBB and URLLC services in vehicular networks, we propose a resource allocation approach that maximizes eMBB rate while ensuring that both URLLC latency and reliability requirements are satisfied. Our approach utilizes non-orthogonal multiple access (NOMA) technology, where the resource for eMBB services is allocated by slot and the traffic of URLLC service is accommodated using mini-slots. To solve this dual time-scale problem, we employ a dual decomposition and sub-gradient method to solve the power allocation and resource block assignment of eMBB services, while the Vogel's approximation method (VAM) and modified distribution method (MODI) are proposed to solve the URLLC resource allocation problem. Additionally, we present two low-complexity heuristic algorithms for the URLLC sub-problem. Simulation results indicate that our proposed approach surpasses baseline methods in terms of both eMBB rate and fairness.**

***Index Terms*—Vehicular networks, network slicing, NOMA, eMBB, URLLC, resource allocation.**

I. Introduction

Vehicle-to-everything (V2X) is a crucial technology for intelligent transportation systems (ITS) that facilitates comprehensive integration of vehicles, pedestrians, basic traffic facilities, and network cloud platforms [1]. This integration enables effective and precise exchange of information, leading to improved road safety and traffic efficiency. The proliferation of V2X technology has given birth to a plethora of businesses, each with their own unique set of requirements and demands. Two typical applications are entertainment services and security services [2]. The former places a heavy emphasis on bandwidth and transmission speed, with the corresponding indicators falling squarely within the ambit of enhanced mobile broadband (eMBB) scene. In contrast, the latter demands strict adherence to the tenets of latency and reliability, and as such, is a prime example of ultra-reliable and low-latency communications (URLLC) [3]. A homogeneous network may not adequately address the varied quality of service (QoS) requirements, which can ultimately result in a reduction in user satisfaction or even service failure. Network slicing technology can create diversified network slices supporting various services, delivering tailored services for users with different needs [4]. However, high throughput, low latency, and reliable connectivity are in conflict with one another, making simultaneous resource allocation for network slicing a difficult task. Achieving an optimal allocation of resources for these three key performance indicators is a critical challenge in the design and implementation of network slicing.

The infrequent nature of URLLC slicing traffic requires prompt and efficient handling. One strategy is to preserve resources for URLLC slicing, but this could lead to underutilization of radio resources [5]. To address this issue, the 3rd generation partnership project (3GPP) recommends the use of superposition or puncturing mechanisms based on short transmission time interval (TTI) for dynamically reusing wireless resources [6]. Currently, many related studies are investigating the perforation scheme, which imposes URLLC service transmission on wireless resources that are already assigned to eMBB services [7]–[14]. With the concept of mini-slots, scheduling cycles of eMBB services are set at one time slot whereas scheduling cycles of URLLC services are set at mini-slots. However, such perforation for URLLC services comes at the cost of reduced transmission speed for eMBB services, despite having improved utilization of wireless resources. The superposition approach can potentially offer a solution to improve radio resource sharing between eMBB and URLLC slices in a more efficient manner. By leveraging the superposition technique, both slices can share the same radio resources simultaneously, while still maintaining their respective QoS demands.

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No.62071356 (Corresponding author: Nan Cheng)

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Yujie Zhang and Nan Cheng are with the State Key Laboratory of ISN and School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail:yjzhang_5@stu.xidian.edu.cn; dr.nan.cheng@ieee.org).

Yanpeng Dai is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: yanpdai@gmail.com).

Zhisheng Yin is with the State Key Laboratory of ISN and School of Cyber Engineering, Xidian University, Xi'an 710071, China (e-mail: zsyin@xidian.edu.cn).

Wei Quan is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: weiquan@bjtu.edu.cn).

Yi Zhou is with the International Joint Research Laboratory for Cooperative Vehicular Networks of Henan, Henan University, Kaifeng 475001, China (e-mail: zhouyi@henu.edu.cn).

Ning Zhang is with Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: ning.zhang@uwindsor.ca).

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

Non-orthogonal multiple access (NOMA) is a multi-user access technology that leverages the power domain to maximize frequency resource utilization by supporting service superposition in mixed eMBB and URLLC scenarios [15]. This capability is crucial in increasing the network's overall capacity and accommodating an increased number of users with varying QoS requirements. Furthermore, recent research efforts have explored the potential of NOMA as an enabler technology for URLLC, which requires support for time-critical applications [16], [17]. Jaya et al. [18] proposed a novel approach for satisfying the stringent delay limitations of applications with time-sensitive requirements in the uplink NOMA context. Their approach is based on two user clustering methods and resource slicing. Specifically, the proposed approach leverages user clustering to group users with similar delay constraints and dynamically allocate RBs based on the specific delay requirements of each cluster. Notably, this approach represents a significant advancement over previous research, as it effectively resolves the trade-off between maximizing capacity and satisfying delay constraints in URLLC scenarios. Additionally, the utilization of NOMA technology in [19] involves categorizing resources into shared and private resources, where shared resources are utilized first to satisfy the transmission and delay needs of users. In cases where shared resources are insufficient, private resources are then allocated to meet these requirements. The proposed algorithm takes into account the delay limits as well as throughput maximization, thereby rendering it suitable for URLLC services.

This paper distinguishes itself from prior research by concurrently addressing two heterogeneous services. It further advances by devising transmission strategies for both services, operating within the framework of heterogeneous NOMA signal modeling. Despite the inherent interdependencies, our design's central tenet emphasizes frequency reuse, resulting in a significant enhancement in spectral efficiency. Moreover, we introduce an innovative problem formulation that seamlessly integrates multi-dimensional resource optimization, thus ensuring optimal performance for both services. Consequently, the intricate nature of the problem necessitates a series of transformative steps, culminating in the identification of two subproblems possessing broader relevance and applicability.The core focus of this study remains firmly anchored in practical utility. Given the heterogeneous characteristics inherent in the transmission formats of the two services, the algorithms presented herein exhibit a pronounced suitability for deployment within heterogeneous systems.

In this paper, we endeavor to tackle a complex and heterogeneous downlink scenario that involves the provision of two distinct types of V2X services. Our primary objective is to optimize the utilization of system resources while ensuring that the varying QoS demands are satisfied through the strategic employment of various techniques, including bandwidth and power allocation, slicing superposition, and delay guaranteeing strategies. To this end, we adopt NOMA to support the superposition of eMBB and URLLC slicing, which is a crucial step in facilitating efficient resource utilization. The eMBB slicing traffic is meticulously scheduled at time slot boundaries to maximize throughput and take into account the rate loss incurred from sharing radio resources with URLLC slicing. In contrast, the URLLC slicing is allocated on a mini-slot basis to ensure that strict QoS demands are upheld and that latency constraints are adhered to by allocating URLLC users immediately upon arrival. Moreover, to ensure utmost reliability, sufficient power is allocated during NOMA superposition. Our contributions are summarized below.

- Our proposed method that enables the management of traffic from eMBB and URLLC services by modifying bandwidth, frequency and power allocations, ensuring that both service types meet their respective QoS requirements. This approach allows for more flexible allocation of resources, thus overcoming the challenge of resource competition and allocation that arise when these two service types coexist.
- To address the challenge of maximizing the throughput of eMBB slicing while meeting QoS requirements in URLLC slicing, we define an optimization framework for resource allocation. We introduce a two-phase framework that includes eMBB slicing resource allocation in a slot and URLLC slicing scheduling in the mini-slot, the latter of which can handle dynamic URLLC traffic.
- In a slot, the eMBB slicing resource allocation is performed using a dual decomposition and sub-gradient method. Whereas in each mini-slot of a slot, we redefine the URLLC resource allocation as a minimization problem and propose the Vogel's approximation method (VAM) and modified distribution method (MODI) to solve the URLLC slicing resource allocation subproblem.
- We also introduce two low-complexity algorithms, namely the minimal resource block reuse (MRBR) scheduling strategy and weight-based joint scheduling strategy, to tackle the URLLC resource allocation subproblem. Our simulation results demonstrate that our proposed algorithms can effectively enhance spectrum efficiency while complying with the delay and reliability requirements of URLLC services.

The organization of this paper is as follows. The literature review is presented in Section II, while the system model and resource allocation problem formulation are introduced in Section III. Our two-timescale resource allocation scheme is discussed in Section IV, and Section V demonstrates the efficacy and feasibility of our proposed algorithms through extensive simulations. Section VI concludes the paper by summarizing our findings and discussing the implications of our work.

II. Related Works

A. Coexistence of eMBB and URLLC services

In the scenario of overlapping eMBB and URLLC services, the challenge lies in achieving the sharing of spectrum resources while satisfying the QoS demands of both eMBB and URLLC users simultaneously. Researchers have utilized information-theoretic approaches to evaluate the efficacy of eMBB and URLLC traffic, as evidenced in [8]. Because

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

URLLC service data packets arrive in bursts, reserving resources for them may result in an unnecessary waste of resources. In [12], the authors introduce a punching scheduling strategy designed to handle sporadic URLLC service. This scheme obviates the need to reserve resources for URLLC services by preempting eMBB service resources when needed. Additionally, an eMBB service transmission interruption recovery mechanism and a punching scheduling strategy are proposed in [12] to counteract eMBB data rate loss. In [14], the authors formulate a joint scheduling problem to optimize the utilization of network resources for both eMBB and URLLC traffic. The primary goal is optimizing the utility of eMBB services while meeting the stochastic demand for URLLC services. The authors propose three models to evaluate the influence of superposition and puncturing on eMBB users, including linear, convex, and threshold-based schemes. In [20], the authors propose a risk-conscious strategy for distributing resources to URLLC services to reduce eMBB transmission uncertainty caused by potential interference. They propose using the conditional value at risk method, which provides an estimate of eMBB traffic uncertainty, for distributing RBs in way that takes into account the risk posed by such interference.

In coexistence service scenarios, eMBB and URLLC services often have distinct QoS requirements. Consequently, many researchers have explored various methods to fulfill these diverse service-level demands. In [21], he authors present a bimodal dynamic spectrum allocation model characterized by a dual-loop methodology. Within this model, resource allocation is dynamically achieved through a fusion of Many-to-Many matching algorithms and the implementation of Advanced K-Granularity Coloring algorithms. In [22], the authors analyze the MAC layer's models of delay and reliability for URLLC traffic. They introduce a resource allocation algorithm that takes into account URLLC service delay constraints in the punching scheduling model, with the aim of enhancing system efficiency. In [23], a two-layer optimization model is introduced to address the issue of data rate loss that eMBB service faces under the overlay scheme. The proposed model optimally allocates spectral resources and power for each eMBB and URLLC user pair, and then optimizes the user-pairing strategy. Moreover, a growing body of scholars is progressively adopting deep learning approaches to effectively tackle the intricate resource allocation dilemmas that arise from intermittent URLLC traffic patterns. The present study in [24] introduced a framework grounded in optimization-assisted deep reinforcement learning (DRL) techniques, aiming to strategically distribute URLLC traffic among eMBB users. This framework capitalizes on the DRL algorithm to mitigate the potential disruptions stemming from real-time scheduling of URLLC traffic, thereby enhancing the stability of eMBB services. In [25], the authors introduced a resource allocation mechanism rooted in event-driven DRL. This approach employed a quartet of distinct DRL techniques, effectively addressing the challenge of stochastic arrival in URLLC services. In [26], the authors presented a novel resource allocation strategy employing DRL, strategically harnessing the capabilities of network slicing techniques. This approach was designed to effectively cater to the diverse and distinct demands of various services.

B. Network Slicing

Network slicing has emerged as a crucial technology in 5G networks, enabling resource allocation through customized services and flexible scheduling. Network slicing relies on network function virtualization (NFV) and software-defined networking (SDN). NFV separates network functions from the physical network and leverages a network function virtualization orchestrator to allocate virtual resources to each slice. SDN enables the data and control planes to be separated, with network control functions deployed centrally on the SDN controller [27], [28]. The primary objective of slicing in radio access networks (RANs) is the flexible allocation of the protocol stack and wireless resources. RAN slicing allows for the integrated management of resources, such as spectrum, power, and air interface, to enable more efficient and flexible utilization of resources for delivering distinct services that meet the varied needs of multiple applications and user groups. Hence, the problem of scheduling resources for RAN slicing is receiving more attention [29]–[31].

In [32], the authors proposed a framework and study how to achieve dynamic resource sharing through network slicing, especially considering non-elastic users with minimum rate requirements. To enable 5G service provisioning, the authors propose a radio resource slicing mechanism. The author introduced a communication-theoretic model that takes into account the unique needs and characteristics of various services [33]. The model demonstrates that slicing can achieve superior performance by by harmonizing the requirements of various services. In [34], the authors investigated the energy efficiency problem in resource scheduling for network slicing, and developed a dynamic optimization model that takes into account both power consumption and service quality. A two-timescale algorithm was devised for achieving bandwidth allocation and service control. In [35], the authors presented an intent-aware reinforcement learning methodology for inter-slice resource allocation, grounded in the explicit characterization of QoS intentions pertinent to individual slices. In [36], the authors introduced a network slicing strategy characterized by a dual-level control granularity, leveraging model-free deep reinforcement learning techniques. This strategy aims to enhance long-term QoS for network services while concurrently optimizing spectrum efficiency within the context of slicing operations. In summary, employing network slicing technology is an efficient strategy to manage the limited wireless resources and the diversity of services in the vehicular communication network.

C. Non-orthogonal Multiple Access

By enabling different users to share resources, NOMA technology has the ability to substantially improve spectrum efficiency and meet the rising demand for data transmission in vehicular networks. Several studies have investigated the use of NOMA to facilitate URLLC services in the context of 5G systems. For example, an optimal resource allocation strategy for NOMA-enabled URLLC services that supports both uplink and downlink transmissions is proposed in [37]. The scheme

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

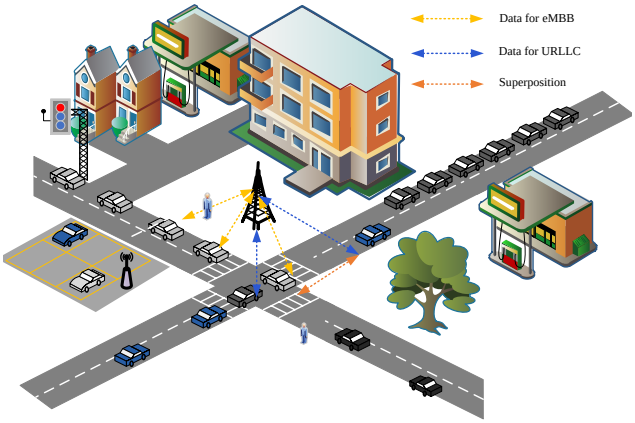


Fig. 1. Heterogeneous service scenarios in vehicular networks

involves developing a queue-based optimization model where power allocation and transmission block length are jointly optimized to minimize the total error probability and enhance transmission reliability. Moreover, the approach accounts for reliability requirements and transmission block length. In [38], the authors investigate the utilization of NOMA for URLLC traffic retransmission while satisfying its constraints. Two schemes, constant power and constant resources, are proposed to reduce the average power consumed by each packet provided URLLC constraints. These studies demonstrate the potential of NOMA in addressing the challenges of URLLC services in 5G systems.

To optimize the system's overall performance, hybrid service transmission schemes based on NOMA must consider various service characteristics and demand disparities. One crucial challenge is to satisfy the stringent demands of URLLC traffic without compromising spectral efficiency [39]. To address this, the authors proposed using unlicensed access for URLLC traffic in order to minimize transmission delay, reserving licensed access for wait-tolerant connectivity [40]. In [41], a cooperative scheduling plan for uplink URLLC and eMBB services in cellular networks was developed, leveraging the idea of communication opportunities for URLLC services. Specifically, the authors proposed a pre-optimized arrangement of distinct transmission options for URLLC users, that prevents conflicts and meets the latency and reliability requirements. To further enhance system performance, overlapping eMBB user traffic opportunities are scheduled using NOMA, and a combined power control technique is employed to reduce the transmit power of eMBB services, thereby avoiding potential impact from URLLC transmissions.

III. System Model And Problem Formulation

A. Vehicular Network Slicing Model

The system model we consider in this paper is a downlink cellular V2X communication system, as illustrated in Fig. 1. The next-generation base station (gNB) provides two distinct services to the randomly distributed vehicle users throughout the system, namely information and entertainment services that require high-rate connectivity and security services with high requirements for latency and reliability. To cater to different QoS demands of diverse services, we have implemented network slicing technology resulting in two service slices: eMBB and URLLC. The set of vehicle users in the eMBB slice can be represented as $\mathcal{K} = \{1, 2, ..., K\}$, and the set of vehicle users in the URLLC slice can be represented as $\mathcal{N} = \{1, 2, ..., N\}$. The RB set of the base station can be denoted by $\mathcal{B} = \{1, 2, ...B\}$, with each RB occupying 12 subcarriers in frequency domain. These RBs are shared by two service slices, providing services to vehicle users within each slice.

The network employs NOMA technology to serve users belonging to different slices while ensuring the orthogonality of RBs in each slice. Users belonging to URLLC slice are scheduled with short TTI, also known as mini-slots, whereas users in the eMBB slice are scheduled with a larger TTI size, such as a slot of 1 millisecond duration, as shown Fig. 2. This approach guarantees no delay in URLLC traffic while providing flexibility to optimize eMBB transmissions. Table I presents the essential notations utilized in this paper.

B. Vehicular Network Communication Model

1) eMBB Slice: The eMBB slice is designed to offer vehicle users with information and entertainment services, consisting of a set of RBs. A crucial principle of eMBB slice allocation is that each RB can only be assigned to one user at any given slot $t \in T$. This makes the RBs orthogonal to each other, and their allocation is performed at the slot boundaries. The achieved rate of vehicle user $k \in \mathcal{K}$ in eMBB slicing at RB $b \in \mathcal{B}$ is expressed by:

$$r^e_{k,b}(t) = f_b \log_2 \left(1 + \gamma_{k,b}(t)\right), \tag{1}$$

where $\gamma_{k,b}(t) = \frac{p_{k,b}(t) h_{k,b}(t)}{\sigma^2}$ represents the signal-to-noise ratio (SNR) calculated as the product of the downlink transmission power $p_{k,b}(t)$, channel gain $h_{k,b}(t)$, and the inverse of the noise power σ^2. The bandwidth of the RB b is denoted by f_b.

To satisfy QoS requirements specified in eMBB slice, gNB needs to assign multiple RBs to users to meet their minimum data rate requirements. Therefore, the achieved rate of the eMBB user $k \in \mathcal{K}$ in time slot t is expressed by:

$$r^e_k(t) = \sum_{b \in \mathcal{B}} x_{k,b}(t) r_{k,b}(t), \tag{2}$$

where $x_{k,b}(t)$ is the RB allocation indicator of eMBB slicing at slot t, which can be defined as:

$$x_{k,b}(t) = \begin{cases} 1, & \text{if the RB } b \text{ is allocated to user } k, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

2) URLLC Slice: In URLLC service slice, data packet arrival times for vehicle users are uncertain. To meet strict QoS requirements, we propose to allocate the incoming URLLC data packet immediately to the next available mini-slot. Additionally, the traditional Shannon's capacity model may not be suitable due to the typically small size of the vehicle user data packets in URLLC slices. Instead, we utilizes the theory

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

TABLE I
LIST OF ABBREVIATIONS

Symbol	Meaning
$\mathcal{K}$	The collection of eMBB slicing users
$\mathcal{N}$	The collection of URLLC slicing users
$\mathcal{B}$	The collection of RBs
f_b	The bandwidth of the RB
P_{max}	The transmission power of the gNB
$D_m(t)$	The number of URLLC slice user packets arriving in mini-slot m.
$\gamma_{kb}(t)$	SNR of eMBB user $k \in K$ in time slot t
$h_{k,b}(t)$	Channel gain of eMBB slicing user $k \in K$ at RB b
$p_{k,b}(t)$	Downlink transmission power of the gNB
σ^2	Noise power
x	Resource allocation vector for URLLC user
N_u	Block length of URLLC packets
ε_u	Error probability of URLLC signal decoding
z	Resource allocation vector for N at mini-slot m of time slot t
α	Identification of strong user among paired eMBB user k and URLLC user n in RB b
ζ_n	The required delay of URLLC
φ	Resource allocation vector for URLLC user
λ	Meaning of Possion arrival process
ϵ	URLLC Reliability probability
$R^e_{k,ach}(m,t)$	Acieved rate in the mini-slot m at eMBB slicing user
$R^u_{n,ach}(m,t)$	Acieved rate in the mini-slot m at URLLC slicing user

of finite block length channel coding to find the rate of vehicle user $n \in \mathcal{N}$ in URLLC slicing at mini-slot m of slot t under the conditions where the block length is N_u and the decoding error probability is ε_u [42]:

$$\begin{aligned} r^u_n(m,t) = &\sum_{b\in\mathcal{B}} z^{m,t}_{n,b} f_b \log_2\left(1+\gamma_{n,b}(m,t)\right) \\ &+ \sqrt{\frac{V_u}{N_u}} Q^{-1}\left(\varepsilon_u\right), \end{aligned} \tag{4}$$

where $\gamma_{n,b}(m,t)$ represents the SINR, $Q^{-1}(\cdot)$ represents the inverse of the Gaussian Q-function, and the channel dispersion is expressed as $V_u = (1-(1+\gamma_{n,b}(m,t))^{-2})$. $z^{m,t}_{n,b}$ is the RB allocation indicator of URLLC slicing at mini-slot m, which can be defined as:

$$z^{m,t}_{n,b} = \begin{cases} 1, & \text{if RB b is allocated to user } n, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

3) NOMA Superposition: In this paper, we investigate the application of the superposition technique, which leverages NOMA, for scheduling overlapping URLLC traffic and eMBB services. Specifically, we employ successive interference cancellation (SIC) to benefit users with stronger channel gain. Suppose that the eMBB user $k \in \mathcal{K}$ has greater channel gain than the URLLC user $u \in \mathcal{N}$ in a given mini-slot m of slot t. Under this circumstance, we calculate the SINRs of the two users as:

$$\gamma_{k,b}(m,t) = \frac{p_{k,b}(t)h_{k,b}(t)}{\sigma^2}, \tag{6}$$

$$\gamma_{n,b}(m,t) = \frac{p_{n,b}(m)h_{n,b}(t)}{p_{k,b}(t)h_{n,b}(t)+\sigma^2}, \tag{7}$$

To satisfy the latency and reliability demands of the URLLC slice, with consideration for NOMA superposition, the URLLC user $u \in \mathcal{N}$ may receive multiple RBs in mini-slot m. The SINR of URLLC user $n \in \mathcal{N}$ at mini-slot m of slot t is given by:

$$\gamma_n(m,t) = \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}} \frac{p_{n,b}(m)h_{n,b}(t)}{\left(1-\alpha^b_{k,n}(m)\right) z^{m,t}_{n,b} p_{k,b}(t)h_{n,b}(t)+\sigma^2}, \tag{8}$$

where $\alpha^b_{k,n}(m)$ is the identification of strong user among paired eMBB user $k \in \mathcal{K}$ and URLLC user $n \in \mathcal{N}$ in RB $b \in \mathcal{B}$. If eMBB user k is the strong user, $\alpha^b_{k,n}(m) = 0$ and if URLLC user n is the strong user, $\alpha^b_{k,n}(m) = 1$.

According to (4) and (8), the achieved rate of URLLC user n is defined as:

$$\begin{aligned} R^u_{n,ach}(m,t) = &\sum_{b\in\mathcal{B}} z^{m,t}_{n,b} f_b \log_2\left(1+\gamma_{n,b}(m,t)\right) \\ &+ \sqrt{\frac{V_u}{N_u}} Q^{-1}\left(\varepsilon_u\right). \end{aligned} \tag{9}$$

The payload $D^{m,t}_n$ of URLLC slicing user $n \in \mathcal{N}$ must be transmitted within the required delay ζ_n to ensure the latency requirement. This requirement is guaranteed by the following condition:

$$\varphi^{m,t}_n D^{m,t}_n \le \zeta_n R^u_{n,ach}(m,t), \tag{10}$$

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

where $\varphi_n^{m,t}$ is a mini-slot allocation indicator for URLLC user $n \in \mathcal{N}$, which can be defined as:

$$\varphi_n^{m,t} = \begin{cases} 1, & \text{if } n \text{ is served at mini-slot } m, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The objective of satisfying the transmission requirements of the URLLC user $n \in \mathcal{N}$ while meeting the hard latency constraint is achieved in our proposed method. Specifically, (10) guarantees that the obtained rate for the URLLC user is above its prescribed threshold within the mini-slot duration. To satisfy the transmission requirements, multiple RBs are allocated to the URLLC slicing user $n \in \mathcal{N}$ in a given mini-slot m as in (8), which ensures that the allocated RBs are sufficient to support the transmission of the user's data and meet its latency requirement.

The provision of URLLC services entails a strict requirement that all requests made during any given mini-slot m must be served, thus constituting a reliability constraint, which can be formulated as:

$$P\left(\sum_{n\in\mathcal{N}} \varphi_n^{m,t} < \Psi\right) \leq (1-\epsilon), \forall m \in M, t \in T. \tag{12}$$

Assuming a Poisson arrival process Ψ with mean arrival rate λ denote the traffic model of URLLC slicing in mini-slot m, and let ϵ represent the reliability probability of URLLC slicing. By setting $\epsilon = 0.999$, the probability of serving fewer URLLC requests than arrived ones must be less than 0.001, as guaranteed by (12).

Since the traffic of URLLC slicing on mini-slots can overlap with allocated eMBB slicing resources, eMBB users may experience a decrease of data rate. The rate achieved by eMBB user $k \in \mathcal{K}$ during mini-slot m can be represented as:

$$\begin{aligned} R_{k,ach}^{e}(m,t) = &\sum_{b\in\mathcal{B}} x_{k,b}(t) z_{n,b}^{m,t} f_b \log_2(1+\gamma_{k,b}(m,t)) + \\ &\sum_{b\in\mathcal{B}} x_{k,b}(t) \overline{z}_{n,b}^{m,t} f_b \log_2(1+\tilde{\gamma}_{k,b}(m,t)), \end{aligned} \tag{13}$$

where the SINR in the conditions that the eMBB user $k \in \mathcal{K}$ is with lower channel gain among paired users, and $\tilde{\gamma}_{k,b}(m,t)$ is given as:

$$\tilde{\gamma}_{k,b}(m,t) = \frac{p_{k,b}(t) h_{k,b}(t)}{\alpha_{k,n}^{b}(m) z_{n,b}^{m,t} p_{n,b}(m) h_{k,b}(t) + \sigma^2}, \tag{14}$$

The data rate of eMBB slicing user k in the duration of slot t is defined as:

$$R_{k,ach}^{e}(t) = \sum_{m\in\mathcal{M}} R_{k,ach}^{e}(m,t). \tag{15}$$

C. Problem Formulation

In the coexistence scenario of mixed vehicular network services, the resource allocation problem involves the allocation for both the eMBB slice and URLLC slice. To satisfy the data requirements of the eMBB slice and the latency and reliability requirements of the URLLC slice simultaneously, our proposed method seeks to optimize the available rate for users of eMBB slice while maintaining the QoS demands of URLLC slice. The optimization problem can be formulated as:

$$\max_{x,p,z} \sum_{k\in\mathcal{K}} R_{k,ach}^{e}(t) \tag{16}$$

$$\text{s.t. } \varphi_n^{m,t} D_n^{m,t} \leq \zeta_n R_{n,ach}^{u}(m,t), \forall n \in \mathcal{N}, m \in M, t \in T, \tag{16a}$$

$$P\left(\sum_{n\in\mathcal{N}} \varphi_u^{m,t} < \Psi\right) \leq (1-\epsilon), \forall m \in M, t \in T, \tag{16b}$$

$$R_{k,ach}^{e}(t) \geq R_{k,\min}^{e}, \forall k \in \mathcal{K}, \tag{16c}$$

$$\sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}} p_{k,b}(t) + \sum_{m\in M}\sum_{n\in\mathcal{N}}\sum_{b\in\mathcal{B}} p_{n,b}(m) \leq P_{\max}, \tag{16d}$$

$$\sum_{k\in\mathcal{K}} x_{k,b}(t) \leq 1, \forall b \in \mathcal{B}, \tag{16e}$$

$$\sum_{n\in\mathcal{N}} z_{n,b}^{m,t} \leq 1, \forall b \in \mathcal{B}, \tag{16f}$$

$$x_{k,b}(t) \in \{0,1\}, \quad \forall k \in \mathcal{K}, b \in \mathcal{B}, t \in T, \tag{16g}$$

$$z_{n,b}^{m,t} \in \{0,1\}, \quad \forall n \in \mathcal{N}, b \in \mathcal{B}, m \in M, t \in T, \tag{16h}$$

$$\alpha_{k,n}^{b}(m) \in \{0,1\}, \forall k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}, m \in M, \tag{16i}$$

$$\varphi_n^{m,t} \in \{0,1\}, \forall n \in \mathcal{N}, m \in M, t \in T. \tag{16j}$$

Constraints (16a) and (16b) ensure the latency and reliability requirements of URLLC slicing, respectively, while constraint (16c) guarantees the rate demands $R_{k,\min}^{e}$ of eMBB users. Constraint (16d) enforces power allocation, and constraints (16e) and (16f) ensure RB orthogonality between eMBB and URLLC users. Moreover, every element of $\mathbf{x}$, $\mathbf{z}$, α, and φ is binary, as indicated by constraints (16g)-(16j). Note that the problem (16) is a challenging combinatorial optimization problem with chance constraints that falls into the NP-hard category.

IV. Two-Timescale Resource Allocation Scheme

To address the high data consumption patterns exhibited by eMBB slicing, the gNB allocates all of its RBs to eMBB slicing users at the outset of each slot t and maintains this allocation throughout t. In any mini-slot m of t, the arrived traffic of URLLC slice must be served in the following mini-slot $m+1$ by the gNB, allowing for the intermingling of URLLC and eMBB traffic as illustrated in Fig. 2. Allocating a subset of RBs to URLLC traffic presents a challenge in identifying eMBB users whose data rates may be affected, as per optimization problem (16). To overcome this obstacle, we can employ a divide-and-conquer strategy to separate the joint resource allocation problem into two sub-problems: eMBB slicing on a slot basis and URLLC slicing on a mini-slot basis. This approach enables us to address the unique challenges of each type of traffic separately and ultimately arrive at a more effective solution.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

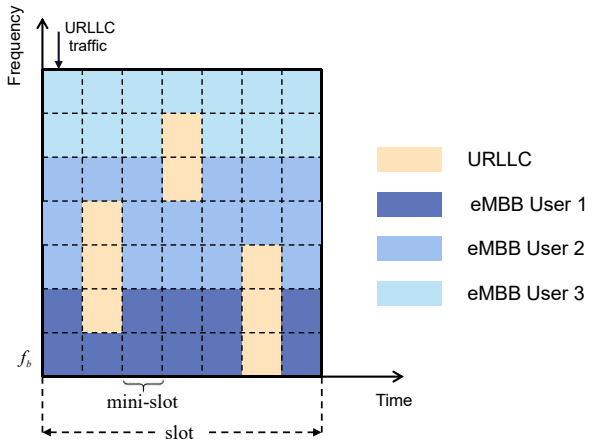


Fig. 2. The scenario where eMBB and URLLC services coexist.

A. eMBB Resource Allocation

Given a fixed feasible URLLC traffic placement z, the eMBB resource allocation sub-problem can be formulated as:

$$\max_{x,p} \sum_{k\in\mathcal{K}} R^e_{k,ach}(t) \tag{17}$$

$$\text{s.t.} \sum_{b\in\mathcal{B}} x_{k,b}(t) r_{k,b}(t) \geq R^e_{k,\min}, \tag{17a}$$

$$x_{k,b}(t) \in \{0,1\}, \forall k \in \mathcal{K}, b \in \mathcal{B}, \tag{17b}$$

$$\sum_{k\in\mathcal{K}} x_{k,b}(t) \leq 1, \forall b \in \mathcal{B}, \tag{17c}$$

$$\sum_{k\in\mathcal{K}} \sum_{b\in\mathcal{B}} x_{k,b}(t) p_{k,b}(t) \leq P_{\max}, \tag{17d}$$

$$p_{k,b}(t) \geq 0, \forall k \in \mathcal{K}, b \in \mathcal{B}. \tag{17e}$$

The Lagrangian function for the problem in (17) can be expressed as:

$$\begin{aligned} L(x,p,&\lambda_k,\mu) = \\ &\sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} R^e_{k,ach}(t) + \sum_{k\in\mathcal{K}} \lambda_k \left(\sum_{b\in\mathcal{B}} R^e_{k,ach}(t) - R^e_{k,\min} \right) \\ &+ \mu \left(P_{\max} - \sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} x_{k,b}(t) p_{k,b}(t) \right) \\ &= \sum_{b\in\mathcal{B}} \left[\sum_{k\in\mathcal{K}} (1+\lambda_k) R^e_{k,ach}(t) - \mu \sum_{k\in\mathcal{K}} x_{k,b}(t) p_{k,b}(t) \right] \\ &- \sum_{k\in\mathcal{K}} \lambda_k R^e_{k,\min} + \mu P_{\max}. \end{aligned} \tag{18}$$

The dual variables associated with the QoS constraints and power constraint are represented by λ_k and μ, respectively. The variables λ_k and μ play an important part in the optimization problem, as they help in finding the optimal solution that satisfies the constraints. Using these variables, we can write the Lagrangian dual function as:

$$G(\lambda_k,\mu) = \begin{cases} \max\limits_{x,p} & L(x,p,\lambda_k,\mu) \\ \text{s.t.} & \sum_{k\in\mathcal{K}} x_{k,b}(t) = 1, \forall b \in \mathcal{B} \\ & 0 \leq x_{k,b}(t) \leq 1, p_{k,b}(t) \geq 0 \end{cases} \tag{19}$$

Then, the dual optimization problem is given by:

$$\min_{\lambda_k,\mu\geq 0} G(\lambda_k,\mu) \tag{20}$$

Through Lagrangian relaxation, the connection between RBs may be eliminated, and (17) can be divided into B sub-problems at each RB, each of which can be solved independently. The sub-problem at each RB is given by:

$$\max_{x,p} L_b = \sum_{k\in\mathcal{K}} (1+\lambda_k) x_{kb}(t) r_{k,b}(t) - \mu \sum_{k\in\mathcal{K}} x_{k,b}(t) p_{k,b}(t) \tag{21}$$

$$\text{s.t.} \sum_{k\in\mathcal{K}} x_{k,b}(t) \leq 1, \forall b \in \mathcal{B}, \tag{21a}$$

$$x_{k,b}(t) \in [0,1], \forall k \in \mathcal{K}, b \in \mathcal{B}, \tag{21b}$$

$$p_{k,b}(t) \geq 0, \forall k \in \mathcal{K}, b \in \mathcal{B}. \tag{21c}$$

A second-level decomposition can further divide the sub-problem (21) into power and subcarrier allocation problems.

1) Power Allocation: Let RB $b \in \mathcal{B}$ be allocated to eMBB user $k \in \mathcal{K}$ and $x_{k,b}(t) = 1$. Then the optimal power allocation on this RB b that yields the best possible performance can be mathematically represented as:

$$\begin{aligned} &\max_{p} \ L_b \\ &\text{s.t. } p_{k,b}(t) \geq 0, \forall k \in \mathcal{K}, b \in \mathcal{B}. \end{aligned} \tag{22}$$

Since this is a constrained optimization problem, Karush-Kuhn-Tucker (KKT) conditions can be used to derive the optimal power allocation solution.

$$p^*_{k,b}(t) = \left[\frac{1+\lambda_k}{\mu \ln 2} - \frac{1}{\omega_{k,b}(t)} \right]^+, \tag{23}$$

where $\omega_{k,b}(t) = \frac{h_{k,b}(t)}{\sigma^2}$, $[x]^+ = \max[0,x]$.

2) RB Allocation: Once the optimal power allocation has been obtained for each RB, the dual function of equation (19) can be expressed as:

$$\begin{aligned} G(\lambda_k,\mu) = &\max_{x} \sum_{b\in\mathcal{B}} \sum_{k\in\mathcal{K}} x_{k,b}(t) H_{k,b}(\lambda_k,\mu) \\ &- \sum_{k\in\mathcal{K}} \lambda_k R^e_{k,\min} + \mu P_{\max} \\ \text{s.t.} \quad &\sum_{k\in\mathcal{K}} x_{k,b}(t) = 1, \forall b \in \mathcal{B}, \\ &x_{k,b}(t) \in [0,1], \forall k \in \mathcal{K}, b \in \mathcal{B}, \end{aligned} \tag{24}$$

where the function $H_{k,b}(\lambda_k,\mu)$ is defined as:

$$H_{k,b}(\lambda_k,\mu) = (1+\lambda_k) r_{k,b}(t) - \mu p^*_{k,b}(t). \tag{25}$$

$H_{k,b}(\lambda_k,\mu)$ expresses the rate achieved by user $k \in \mathcal{K}$ when selecting RB $b \in \mathcal{B}$ as the first term, while the second term of

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

(25) represents the power consumption cost. Thus, the profit obtained by user k by transmitting over RB b is denoted by $H_{k,b}(\lambda_k, \mu)$, and the profit vector at each RB b is represented by $\mathbf{H} = [H_{k,b}]$. To maximize the objective function (24), we need to select exactly one element of vector $\mathbf{H}$ for each RB b, such that the total profit is maximized. The optimal RB allocation can be determined by selecting the user k^* that has the highest value of $H_{k,b}(\lambda_k, \mu)$. The corresponding formula is provided as follows:

$$x_{k,b}(t) = \begin{cases} 1, k^* = \arg\max\limits_{k} H_{k,b}(\lambda_k, \mu), \\ 0, \qquad\qquad otherwise. \end{cases} \tag{26}$$

3) Variable Update: The convexity of the dual function $G(\lambda, \mu)$ implies that it can be minimized using a sub-gradient method. In this regard, the dual variables λ and μ are updated in parallel using the following procedure:

$$\lambda_k(i+1) = \left[\lambda_k(i) + \xi_1(i)(R_{k,\min} - \sum_{b\in\mathcal{B}} x_{k,b}(t) r_{k,b}(t))\right]^+, \tag{27}$$

$$\mu(i+1) = \left[\mu(i) + \xi_2(i)(\sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}} x_{k,b}(t) p_{k,b}(t) - P_{\max})\right]^+, \tag{28}$$

where $\xi_1(i)$ and $\xi_2(i)$ are diminishing step sizes and i is the iteration index. The convergence of the sub-gradient method described above to the optimal dual variables is guaranteed provided that the step sizes $\xi_1(i)$ and $\xi_2(i)$ follow a diminishing step size policy. By virtue of the mathematical formulations and derivations presented, it is possible to algorithmically compute the optimal assignment of RBs and power allocation. The complexity of of this process is expected to be $O(n^3)$. The proposed optimal scheme is formalized in pseudocode as presented in Algorithm 1.

Algorithm 1 eMBB Resource Allocation

```
Initialize: λ_k, μ
while not converged do
    for b = 1 → B do
        p*_{1:K,b} ← calculate optimal powers using (21)
        H_{1:K,b} ← calculate H_{k,b}(λ_k, μ) using (23)
    end for
    k* ← find optimal users according to (24)
    Allocate RBs to k*
    Update λ_k, μ using (25)
end while
```

B. URLLC Resource Allocation

Given the resource allocation scheme for the eMBB slice in each slot, the resource allocation subproblem for the URLLC service slice in the same slot can be expressed mathematically as:

$$\max_{z,p} \sum_{k\in\mathcal{K}} R^e_{k,ach}(t) \tag{29}$$

$$\text{s.t. } \varphi^{m,t}_n D^{m,t}_n \le \zeta_n R^u_{n,ach}(m,t), \forall n \in \mathcal{N}, m \in M, t \in T, \tag{29a}$$

$$P\left(\sum_{n\in\mathcal{N}} \varphi^{m,t}_u < \Psi\right) \le (1-\epsilon), \forall m \in M, t \in T, \tag{29b}$$

$$p'_{k,b}(m) + p_{n,b}(m) \le p_{k,b}(\mathrm{t}), \forall b \in \mathcal{B}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{29c}$$

$$\sum_{n=1}^{N} z_{n,b}(m) \le 1, \forall b \in \mathcal{B}, \tag{29d}$$

$$z_{n,b}(m) \in \{0,1\}, \forall b \in \mathcal{B}, n \in \mathcal{N}, \tag{29e}$$

$$\alpha^b_{k,n}(m) \in \{0,1\}, \forall b \in \mathcal{B}, n \in \mathcal{N}. \tag{29f}$$

The optimization problem presented in (29) is a challenging combinatorial problem that demands a careful allocation of resources among URLLC users. Specifically, it is necessary to determine the RB selection process of each URLLC user and the reallocation of power with eMBB users on the RB. The fractional power allocation (FTPA) algorithm, which takes into account the channel quality between different users and allocates higher power to users with poor channel status to meet their rate requirements by adaptively adjusting the power allocation factor according to channel gain, could be utilized to calculate the power assigned to each URLLC user, considering the allocated RBs among eMBB slice users had been determined already at the start of each time slot. Specifically, when a subset of $N^{'} \le N$ URLLC users has been served, (29b) holds for a particular value of ϵ. Assuming $\mathcal{N}^{'} = \{1, 2, ..., N^{'}\}$ and $\varphi^{m,t}_n = 1, \forall n \in \mathcal{N}^{'}$, we can construct a cost matrix $C = (c_{n,b})$, where $n \in \mathcal{N}^{'}, b \in \mathcal{B}$ to determine the required RBs for all $n \in \mathcal{N}^{'}$ as specified by the upper bound in (29a). We can also define $S = [s_1, s_2, ..., s_{N'}]$ represents the vector of the number of resource blocks required by all URLLC slice users. Accordingly, the problem (29) can be equivalently rewritten in a more explicit form as:

$$\min_{z} \sum_{n\in\mathcal{N}}\sum_{b\in\mathcal{B}} c_{n,b}(m) z_{n,b}(m) \tag{30}$$

$$\text{s.t. } \sum_{b=1}^{B} z_{n,b}(m) = s_n, \forall n \in \mathcal{N}^{'}, \tag{30a}$$

$$\sum_{n=1}^{N^{'}} z_{n,b}(m) \le 1, \forall b \in \mathcal{B}, \tag{30b}$$

$$\sum_{n=1}^{N^{'}} s_n \le B, \tag{30c}$$

$$z_{n,b}(m) \in \{0,1\}, \forall b \in \mathcal{B}, n \in \mathcal{N}. \tag{30e}$$

The objective of (30) is to minimize the rate loss of eMBB slicing by identifying a matrix $\mathcal{Z}$ that complies with the problem's constraints. This problem is a well-known unbalanced transportation model and a linear programming problem. To convert the inequality constraints in equations (30b) and (30c)

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

to equalities, we introduce slack variables $z_{N'+1,b}, \forall b \in \mathcal{B}$ and $s_{N'+1}$, resulting in the following reformulated problem.

$$\sum_{n=1}^{N'} z_{n,b}(m) + z_{N'+1,b} = 1, \forall b \in \mathcal{B}. \tag{31}$$

$$\sum_{n=1}^{N'} s_n + s_{N'+1} = B. \tag{32}$$

The problem in (30) has been transformed into a balanced transportation model (BTM). There are numerous methods to obtain an initial feasible solution for the BTM, including Northwest corner (NWC) and Vogel's Approximation Method (VAM). To reach the optimal solution, one can use the Modified Distribution (MODI) or stepping-stone methods. In the subsequent sub-section, we combine VAM and MODI to acquire the optimal solution.

1) Acquiring Initial Feasible Solution by VAM: VAM is a technique that accounts for the potential impact of using supply and sales points with high freight rates in the subsequent dispatch process after initially dispatching cells based on a certain minimum unit freight rate. This has the potential to increase the overall transportation cost. VAM works by calculating the difference between the minimum unit freight rate and the next-smallest unit freight rate in each row and column of the freight rate table. This difference between the two unit freight rates is known as the penalty amount. The transportation is then arranged according to the minimum unit freight rate for the maximum penalty amount. However, if the penalty amount is particularly high, the transportation is not planned in accordance with the minimum freight rate, resulting in significant freight loss. The VAM method consists of the following steps:

Step1: Determine the row penalty and column penalty for each row and column in C, which are known as the lowest loss and the next-lowest loss, respectively;

Step2: Select the cell with the lowest loss in the row or column where the largest of these penalties is located, and cross the row or column;

Step3: In each row or column that has not been crossed out, the above steps are repeated until the last cell is also allocated and the initial feasible solution is obtained.

2) Finding an Optimal Solution by MODI Method: An initial basic feasible solution for the transportation problem is obtained, and then the optimality of the solution is evaluated using the potential method. This method involves solving the test number of non-basic variables in the simplex table, and to do this, an additional left-hand column l_n and top row m_b with the cost matrix C are added. The values for these new entries are calculated, and then measured for all cells corresponding to the allocation in $\mathcal{Z}$, as illustrated below:

$$l_n + m_b = c_{n,b}, \forall z_{n,b} \neq 0. \tag{33}$$

In order to further progress, equation (33) is solved to acquire all l_n and m_b. If necessary, one of the unknowns is assigned a value of zero to facilitate the solution. Subsequently, the empty cells of $\mathcal{Z}$ are evaluated as follows:

$$\sigma_{n,b} = c_{n,b} - l_n - m_b, \forall z_{n,b} \neq 0. \tag{34}$$

To obtain the optimal allocation $\mathcal{Z}$, we first select the $\sigma_{n,b}$ corresponding to the most negative value and determine the closed path for the corresponding cell to obtain the reallocation amount. Next, we allocate the maximum permissible to the empty cell of $\mathcal{Z}$ based on the selected $\sigma_{n,b}$, l_n, and m_b values. Then, we compute the cost of the empty cells in $\mathcal{Z}$ using (34) and recompute the C and $\mathcal{Z}$ values using (33). This process is repeated until there are no negative $\sigma_{n,b}$ values left, resulting in the optimal allocation $\mathcal{Z}$. The MODI method consists of the following steps:

Step 1: Add the potential column and potential row to the table of feasible initial basis solution $\mathcal{Z}$ obtained from the VAM method;

Step 2: Add a row l_n and a column m_b to the value matrix C based on (33);

Step 3: Calculate $\sigma_{n,b}$ using (34);

Step 4: Determine the closed path by identifying the cell that corresponds to the minimum $\sigma_{n,b}$ found in Step 3 and allocate resource blocks accordingly;

Step 5: Reiterate Step 2 to 4 until all $\sigma_{n,b} \geq 0$.

C. Two Low-Complexity Heuristic Algorithms for URLLC Resource Allocation

Although the VAM and MODI methods are able to achieve optimal solutions for sub-problem (30), the final number of RBs allocated to URLLC users may exceed the actual requirements due to the scaling of variables in the conversion of sub-problem (29) into the transportation model. This issue becomes more pronounced with an increasing number of URLLC users, potentially leading to a significant reduction in the eMBB user rate. To mitigate this problem, we introduce two heuristic algorithms.

1) Strategy I : Minimal resource block reuse (MRBR) scheduling strategy

The aim of MRBR is to minimize the reuse of resources between eMBB and URLLC slices. For this purpose, we partition the set of available RBs based on the channel status of each users into two subsets, namely B_1 and B_2. RBs in set B_1 with the highest channel status are initially assigned to the URLLC slicing user n to meet their QoS requirements and minimize the interference due to sharing the RBs with eMBB slicing users. URLLC service slicing users prioritize using the resource pool B_1 to meet their latency and reliability requirements and preferentially choose RB capable of completing transmission tasks using a single RB.

Algorithm 2 Minimal resource block reuse (MRBR) scheduling strategy

```
Initialize: B_1, B_2, p, x
for n = 1 → N do
    Power allocation following the FTPA algorithm
    Calculate the achieved rated of eMBB user using (13)
    Calculate the achieved rate of URLLC user n using (9)
    if ζ_n D_n^{m,t} ≤ r_{n,b}^u then
        Allocate RB b* to URLLC user n
    end if
    Update B_1 and B_2
end for
```

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

TABLE II
SIMULATION DEFAULT PARAMETER VALUES

Parameter	Value	Parameter	Value
K	20	N	10
f_b	180kHz	M	7
B	50	ζ	32bytes
N_0	-114dBm	λ	1,2,3,4,5,6

2) Strategy II: Weight-based joint scheduling strategy

We employ a weight matrix $W = \{w_{n,b} = \alpha r^e_{k,b}(m) + (1-\alpha) r^u_{n,b}(m), n \in \mathcal{N}, k \in \mathcal{K}, b \in \mathcal{B}\}$, as opposed to MRBR, to explain how URLLC multiplexing affects eMBB user rate. gNB allocate resources to URLLC user $n \in N$ according to the value of weight, and calculate the total revenue that the system can obtain under this allocation scheme.

Algorithm 3 Weight-based joint scheduling strategy

1: Initialize: B_1, B_2, p, x
2: **for** $n = 1 \rightarrow N$ **do**
3: Power allocation following the FTPA algorithm
4: Calculate the weight matrix W
5: Calculate the achieved rate of eMBB slicing using (15)
6: Allocate RB by selecting the maximum element in W
7: Update the weight matrix W
8: **end for**

V. NUMERICAL RESULTS AND DISCUSSION

In this section, simulations are carried out to evaluate and analyze the performance of the proposed scheme. The SINRmax-based algorithm was adopted as the baseline algorithm, which mainly allocates RBs to the URLLC user with the highest SINR value. Consider a downlink vehicular communication system with a base station coverage range of 500 meters. Two types of service slice users are distributed randomly. The eMBB traffic adopts a full-buffer traffic model and the URLLC traffic can be modeled through a Poisson process with a mean λ. The scheduling cycle for the eMBB service slice is 1 millisecond, with each time slot having the same duration. Moreover, each time slot is subdivided into seven equally-spaced sub-slots, and the start of each sub-slot is earmarked to schedule URLLC service slice users. A summary of key simulation parameters is presented in Table II.

Fig. 3 illustrates the effect of varying the weight parameter α on the performance of the weight-based joint scheduling strategy. The simulation results indicate that the influence of α on the eMBB slice user rate is insignificant when $\lambda = 1$, whereas when $\lambda = 6$, α has a considerable impact. The weight-based joint scheduling strategy successfully integrates the QoS requirements of URLLC users and mitigates the impact of the superposition mechanism on the eMBB slice user rate. In addition, it can be observed that the average eMBB rate increases firstly and then decreases as the weight factor α from 0 to 1. Particularly, there exists a selectable weight about $\alpha = 0.4$ maximizing the eMBB rate. Thus, we choose $\alpha = 0.4$ as the weight factor for the weight-based joint scheduling strategy in the subsequent simulations.

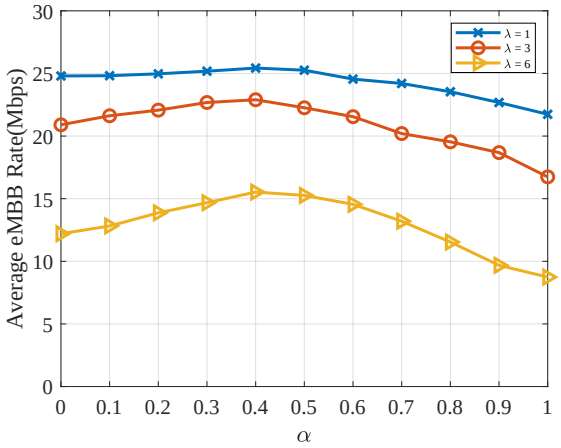


Fig. 3. Average eMBB data rate for different value of α.

Fig. 4 presents a comparison of the empirical cumulative distribution function (ECDF) values of the average rate of eMBB slice users with the increasing URLLC packet arrival rate. The results indicate that the transportation model-based URLLC scheduling scheme outperforms other schemes when $\lambda = 1$. Specifically, the probability of achieving average rate for eMBB slice users at least 10 Mbps is 0.2 with the transportation model algorithm, 0.168 with the weight-based joint scheduling strategy, and about 0.3 with the MRBR algorithm. These findings suggest that the weight-based joint scheduling strategy takes into account the impact on eMBB users when allocating resources to URLLC slice users, resulting in a higher reachable rate of eMBB users under the same conditions. Thus, the weight-based joint scheduling strategy is considered to be more effective in improving the performance of the system for eMBB slice users.

Fig. 5 depicts the mean eMBB data rate as a function of λ. As λ increases, the average eMBB data rate drops due to the increase in resources allocated to URLLC users. Upon comparing the simulation results, it is observed that the URLLC scheduling scheme solved by TM can enhance the system's performance, particularly when the URLLC packet arrival rate λ is small. However, as λ increases, the performance of TM gradually deteriorates, although it still outperforms the SINRmax-based scheduling scheme overall. The use of the minimum data rate when allocating resource blocks during transmission model establishment explains this observation. With a rise in data packets, the number of resource blocks assigned to users exceeds the actual requirement, adversely affecting the performance of eMBB slice users. Notably, the two proposed heuristic algorithms exhibit superior performance compared with the SINRmax-based scheduling scheme, and the weight-based joint scheduling strategy performs better than MRBR. This is attributed to the fact that the weight-based joint scheduling strategy considers both the URLLC's QoS constraints and the effect of superposition on eMBB user rates.

Moreover, we employ superposition mechanisms to multiplex the resources of the two sliced services. Punching is

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

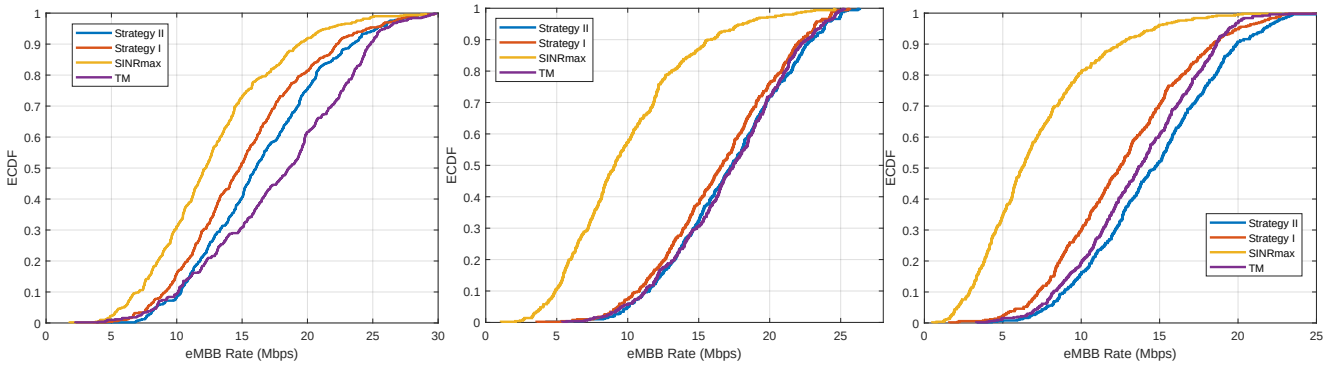


Fig. 4. Comparison of ECDF for (a) $\lambda = 1$, (b) $\lambda = 3$, and (c) $\lambda = 6$.

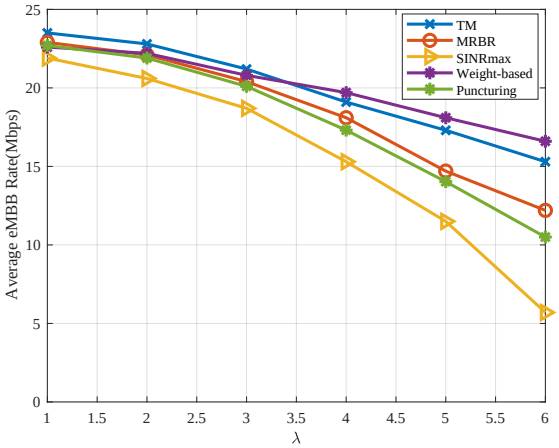


Fig. 5. Average eMBB data rate for different value of λ.

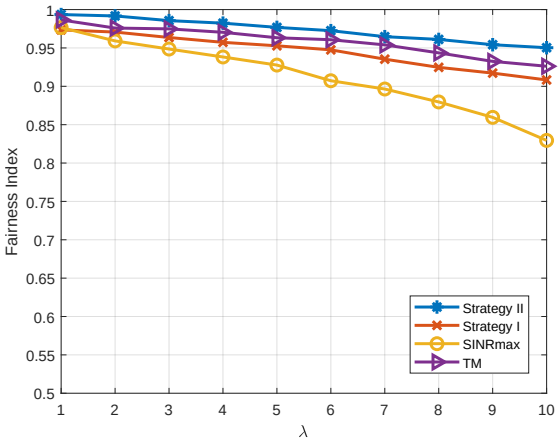


Fig. 6. Fairness index for different value of λ.

another common multiplexing approach, prioritizing URLLC packets by temporarily suspending the data transmission of eMBB users. Fig. 5 presents a comparative analysis of the impact of URLLC packet arrival rates on the average data rate of eMBB users under the two multiplexing mechanisms. The results demonstrate that NOMA significantly outperforms punching in terms of system performance due to its ability to enable URLLC users to multiplex the same RB as eMBB users simultaneously by leveraging different transmission powers to extract signals from different users. The NOMA technique thus reduces the effect of URLLC users on the eMBB data rate without completely terminating eMBB services, leading to an improved eMBB slice user rate when compared to the punching method.

The simulation results, depicting fairness scores using Jain's fairness index across a range of URLLC packet arrival rates, are visually represented in Figure 6. It is observed that the weight-based joint scheduling strategy outperforms the other schemes across the different values of λ. Furthermore, the results indicate that the average data rate experiences a decline as the value of λ increases. This reduction is attributed to the additional resource allocation required for URLLC traffic. The fairness score experiences a similar decline as the RB allocation to URLLC users becomes a priority over eMBB users, ultimately leading to a reduction in eMBB user rate.

Fig. 7 compares the impact of total transmit power of the base station on the eMBB slice data rate using different power allocation schemes. The proposed FTPA algorithm outperforms the conventional FPA algorithm which distributes the transmitted power to each user in a fixed manner. The eMBB slice data rate increases gradually as the total transmit power of the base station increases, but the growth trend slows down as the FTPA algorithm dynamically allocates transmit power based on the channel quality of user pairs on the same resource block. This allows the algorithm to allocate more power to users with poor channel conditions, ensuring their data rate and achieving better overall performance.

The impact of URLLC QoS requirements on the rate of eMBB slices is contrasted in Fig. 8. It shows that the data rate of eMBB slices increases when the delay requirement of URLLC ζ_n is increased from 1 mini-slot to 4 mini-slots while it still falls as the arrival rate of URLLC packets rises. This is a result of better spectrum resource utilization caused by scheduling burst incoming URLLC packets in more mini-slots.

VI. Conclusion

In this paper, we proposed a joint scheduling scheme on dual time-scales that utilizes network slicing and NONA technology to address the resource allocation problem in the multi-service coexistence scenario of the vehicular network. Specifically,

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

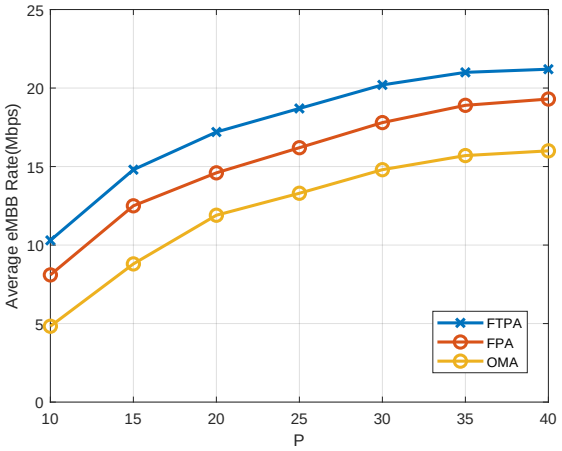


Fig. 7. Average eMBB data rate for different value of P.

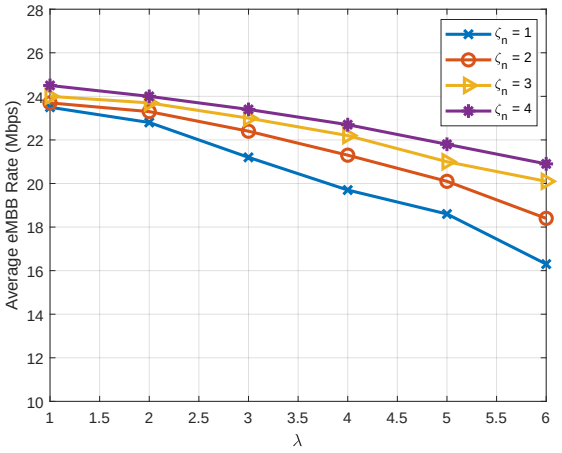


Fig. 8. Average eMBB data rate for different value of delay.

the decomposition strategy divides the problem into a slot-based eMBB slice resource allocation problem and a mini-slot-based URLLC slice resource scheduling problem to cope with dynamic URLLC traffic and channel variations. Simulation results have demonstrated that the proposed approach comprehensively considers the diverse QoS demands of different services and outperforms the baseline method with regards to throughput and latency. This study provides insightful guidance to the practical deployment of vehicular networks. In future work, we aim to investigate other aspects of resource allocation in conjunction with machine learning, including dynamic traffic pattern and energy efficiency optimization.

REFERENCES

[1] A. Nanda, D. Puthal, J. J. P. C. Rodrigues, and S. A. Kozlov, "Internet of autonomous vehicles communications security: Overview, issues, and directions," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 60–65, 2019.

[2] H. Zhou, W. Xu, J. Chen, and W. Wang, "Evolutionary V2X technologies toward the internet of vehicles: Challenges and opportunities," *Proc. IEEE*, vol. 108, no. 2, pp. 308–323, 2020.

[3] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 2021.

[4] 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, 2017.

[5] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, X. S. Shen, and W. Zhuang, "AI-native network slicing for 6G networks," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 96–103, 2022.

[6] F. Song, J. Li, C. Ma, Y. Zhang, L. Shi, and D. N. K. Jayakody, "Dynamic virtual resource allocation for 5G and beyond network slicing," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 215–226, 2020.

[7] N. Cheng, N. Lu, N. Zhang, X. Zhang, X. S. Shen, and J. W. Mark, "Opportunistic WiFi offloading in vehicular environment: A game-theory approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1944–1955, 2016.

[8] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 489–502, 2019.

[9] Y. Kai, K. John, M, N. Toshizo, Y. Zhanping, and J. Sheng, "Coexistence of enhanced mobile broadband communications and ultra-reliable low-latency communications in mobile front-haul," *Proc. SPIE*, vol. 10559, 2018.

[10] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, 2018.

[11] Z. Wu, F. Zhao, and X. Liu, "Signal space diversity aided dynamic multiplexing for eMBB and URLLC traffics," *in Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, pp. 1396–1400, 2017.

[12] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, pp. 1–6, 2017.

[13] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, "A matching based coexistence mechanism between eMBB and URLLC in 5G wireless networks," *in Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, p. 2377–2384, 2019.

[14] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *in Proc. INFOCOM*, p. 1970–1978, 2018.

[15] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[16] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[17] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, "Space/Aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, 2019.

[18] N. Imtiaz Jaya and M. F. Hossain, "RAN resource slicing and sharing with NOMA for latency reduction in uplink URLLC networks," *in Proceedings of the 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, 2020.

[19] H. Deng, F. Luo, and Q. Li, "A hybrid resource allocation method for URLLC based on NOMA," *in Proceedings of the 2021 IEEE 21st International Conference on Communication Technology (ICCT)*, pp. 902–905, 2021.

[20] 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, 2019.

[21] N. Chen, Z. Cheng, Y. Zhao, L. Huang, X. Du, and M. Guizani, "Joint Dynamic Spectrum Allocation for URLLC and eMBB in 6G Networks," *IEEE Transactions on Network Science and Engineering*, pp. 1–14, 2023.

[22] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1080–1093, 2021.

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

[24] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585–4600, 2021.

[25] Y.-H. Hsu and W. Liao, "embb and urllc service multiplexing based on deep reinforcement learning in 5g and beyond," *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1467–1472, 2022.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

[26] K. Suh, S. Kim, Y. Ahn, S. Kim, H. Ju, and B. Shim, "Deep Reinforcement Learning-Based Network Slicing for Beyond 5G," *IEEE Access*, vol. 10, pp. 7384–7395, 2022.

[27] Y. Zhou, N. Cheng, N. Lu, and X. S. Shen, "Multi-UAV-Aided networks: Aerial-Ground cooperative vehicular networking architecture," *IEEE Veh. Technol. Mag.*, vol. 10, no. 4, pp. 36–44, 2015.

[28] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, and X. Shen, "Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, 2021.

[29] K. Xiong, S. Leng, J. Hu, X. Chen, and K. Yang, "Smart network slicing for vehicular fog-RANs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3075–3085, 2019.

[30] S. Zhang, W. Quan, J. Li, W. Shi, P. Yang, and X. Shen, "Air-ground integrated vehicular network slicing with content pushing and caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2114–2127, 2018.

[31] X. Ge, "Ultra-reliable low-latency communications in autonomous vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5005–5016, 2019.

[32] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6419–6432, 2018.

[33] 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.

[34] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, "Dynamic resource allocation with RAN slicing and scheduling for URLLC and eMBB hybrid services," *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[35] C. V. Nahum, V. H. Lopes, R. M. Dreifuerst, P. Batista, I. Correa, K. V. Cardoso, A. Klautau, and R. W. Heath, "Intent-aware Radio Resource Scheduling in a RAN Slicing Scenario using Reinforcement Learning," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[36] J. Mei, X. Wang, K. Zheng, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, "Intelligent Radio Access Network Slicing for Service Provisioning in 6G: A Hierarchical Deep Reinforcement Learning Approach," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6063–6078, 2021.

[37] X. Xie, X. Ou, H. Lu, and Q. Huang, "Joint uplink and downlink resource allocation in NOMA for end-to-end URLLC services," *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3942–3946, 2021.

[38] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "How URLLC can benefit from NOMA-based retransmissions," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1684–1699, 2021.

[39] T. Ma, H. Zhou, B. Qian, N. Cheng, X. Shen, X. Chen, and B. Bai, "UAV-LEO integrated backbone: A ubiquitous data collection approach for B5G internet of remote things networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3491–3505, 2021.

[40] S. Doğan, A. Tusha, and H. Arslan, "NOMA with index modulation for uplink URLLC through grant-free access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1249–1257, 2019.

[41] A. Anand, G. de Veciana, D. Malak, A. Elezabi, and A. Venkatakrishnan, "Opportunistic overlapping: Joint scheduling of uplink URLLC/eMBB traffic in NOMA based wireless systems," *in Proc. 19th Int. Symp. Modeling Optim. Mobile, Ad hoc, Wireless Netw. (WiOpt)*, pp. 1–8, 2021.

[42] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

Yujie Zhang received the B.E. degree from the School of Computers Science and Engineering, Northeastern University, Shenyang, China, in 2020, and the M.S. degree from the School of Telecommunications Engineering, Xidian University, Xi'an, China, in 2023. Her research interests include network slicing and resource allocation in vehicular networks.

Nan Cheng received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo in 2016, and B.E. degree and the M.S. degree from the Department of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2009 and 2012, respectively. He worked as a Post-doctoral fellow with the Department of Electrical and Computer Engineering, University of Toronto, from 2017 to 2019. He is currently a professor with State Key Lab. of ISN and with School of Telecommunications Engineering, Xidian University, Shaanxi, China. He has published over 90 journal papers in IEEE Transactions and other top journals. He serves as associate editors for IEEE Transactions on Vehicular Technology, IEEE Open Journal of the Communications Society, and Peer-to-Peer Networking and Applications, and serves/served as guest editors for several journals. His current research focuses on B5G/6G, AI-driven future networks, and space-air-ground integrated network.

Yanpeng Dai (Member, IEEE) received the B.Eng. degree from Shandong Normal University, Jinan, China, in 2014, and the Ph.D. degree from Xidian University, Xi'an, China, in 2020. He is currently an Associate Professor with the Information Science and Technology College, Dalian Maritime University, Dalian, China. He was a Visiting Student at the University of Waterloo, Waterloo, ON, Canada. His research interests include resource management and interference coordination for heterogeneous wireless networks and maritime communications.

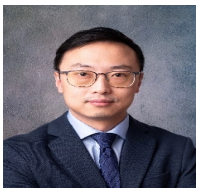
Zhisheng Yin (M'20) received his Ph.D. degree from the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2020, and the B.E. degree from the Wuhan Institute of Technology, the B.B.A. degree from the Zhongnan University of Economics and Law, Wuhan, China, in 2012, and the M.Sc. degree from the Civil Aviation University of China, Tianjin, China, in 2016. From Sept. 2018 to Sept. 2019, Dr. Yin visited in BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently an Assistant Professor with School of Cyber Engineering, Xidian University, Xi'an, China. He is also an Associate Editor of IEEE Internet of Things Journal. His research interests include space-air-ground integrated networks, wireless communications, digital twin, and physical layer security.

Wei Quan (Senior Member, IEEE) received the Ph.D. degree in communication and information system from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014. He is currently a Full Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University (BJTU), Beijing, China. He has coauthored more than 50 papers in prestigious international journals and conferences, including IEEE Journal on Selected Areas in Communications, IEEE Communications Magazine, IEEE Wireless Communications, IEEE Network, IEEE Transactions on Vehicular Technology, IEEE Communications Letters. His research interests focus on reliable transmission in mobile networks, vehicular networks and Industrial IoT. Dr. Quan is an Associate Editor for the IEEE Transactions on Vehicular Technology, Peer-to-Peer Networking and Applications, Journal of Internet Technology, and IEEE Access. He is a winner of 2022 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award and a principle investigator (PI) of National Key Research and Development Program of China.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. 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/TVT.2023.3337250

Yi Zhou (Member, IEEE) received the B.S. degree in electronic engineering from the First Aeronautic Institute of Air Force, China, in 2002, and the Ph.D. degree in control system and theory from Tongji University, China, in 2011. He is currently a full professor and deputy dean with the School of Artificial Intelligence, Henan University, China. He is also the Director of International Joint Research Laboratory for Cooperative Vehicular Networks, Henan, China. His research interests include vehicular cyber physical systems and multi-agent collaboration.



Ning Zhang is an Associate Professor and Canada Research Chair in the Department of Electrical and Computer Engineering at University of Windsor. He received the Ph.D degree in Electrical and Computer Engineering from University of Waterloo, Canada, in 2015. After that, he was a postdoc research fellow at University of Waterloo and University of Toronto, respectively. His research interests include connected vehicles, mobile edge computing, wireless networking, and security. He is a Distinguished Lecturer of IEEE ComSoc, a Highly Cited Researcher (Web of Science), as well as the Vice Chair for IEEE Technical Committee on Cognitive Networks and IEEE Technical Committee on Big Data. He serves/served as an Associate Editor of IEEE Transactions on Mobile Computing, IEEE Communications Surveys and Tutorials, IEEE Internet of Things Journal, and IEEE Transactions on Cognitive Communications and Networking. He also serves/served as a TPC/General chair for numerous conferences and workshops, such as IEEE ICC, VTC, INFOCOM Workshop, and Mobicom Workshop. He received several Best Paper Awards from conferences and journals, such as IEEE Globecom, IEEE ICC, IEEE ICCC, IEEE WCSP, and Journal of Communications and Information Networks.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on December 21,2023 at 06:12:43 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.