Model-ML Integrated Intelligence in URLLC Towards End-to-End Delay Fulfillment Over Vehicular Networks

Yuquan Xiao, Qinghe Du, Wenchi Cheng, George K. Karagiannidis, and Zixiao Zhao

Abstract

Ultra-Reliable and Low-Latency Communication (URLLC) had been initially proposed as one of the three main application scenarios in the fifth generation of mobile telecommunications systems (5G). While URLLC is expected to support vehicular networking with low latency, current 5G infrastructures still cannot well assure about one-millisecond-level delay for various time-sensitive applications in vehicular networks. To better serve vehicular networks as well as other vertical application scenarios with stringent latency requirement, URLLC remains the hotspot for beyond-5G and 6G and is expected to dig deeper in optimization paradigm together with random-access control technologies. Model-based design principles integrating machine learning (ML) intelligence have been recognized as a competitive candidate to empower URLLC quickly into reality. In this article, we first anatomize the constitution of the end-to-end delay for URLLC towards vehicular networks and concentrate on the ways of how to apply model-ML integrated intelligence to reduce major delay components, including access delay, queuing delay, and transmission delay. Facing the challenging task, we derive an intelligent multi-tier-driven computing framework for access-delay reduction. We then introduce an efficient resource allocation approach driven by multi-deep-reinforcement-learning networks to jointly lower the queuing delay and transmission delay. Finally, we share the discussions about the open issues on latency control for future URLLC.

Introduction

Ultra-low information-exchange latency assurance is of vital importance for vehicular communication networks. The current latency-sensitive information transmissions are mainly supported by the fifth generation of wireless communications (5G) systems. However, even though 5G have been deployed for several years, the delay performances offered by the current 5G networks still cannot well meet the ever-increasing requirements of mission-critical vehicular applications. Therefore, ultra-reliable and low-latency communication (URLLC) in the vehicular networks remains to be one of the major concentrations in the study for beyond-5G and the sixth generation of mobile telecommunications systems (6G) [1–4].

In current wireless transmission standard bodies, most techniques to empower low-latency communications are derived via model-based analyses, which largely benefit from the fundamental communication theories, networking theories, and mathematical optimization methods. Model-based techniques can effectively reduce the delay control of small-scale communication scenarios. But with the rapidly increasing demands of mission-critical services, the delay quality-of-service (QoS) assurance over URLLC networks will become more challenging and even intractable. To ease delay analyses and control in such scenarios, some inevitably oversimplified assumptions are made over the model-based approach for current research, which certainly weakens the performance in practical networks [5]. In addition, some optimization problems with essential nonconvex nature and discrete variables are difficult to solve by using traditional model-based approaches. Even though some margin can be gained via sophisticated conventional optimization tools, the extra delay introduced by heavy computing load is generally large, which makes delay-QoS assurance infeasible in URLLC. Thus, model-based techniques themselves might be hardly harmonized with the future URLLC networks.

To facilitate the efficient control for URLLC, machine learning (ML) has shown the great potential as a competent candidate to complement model-based approaches [6–8]. Among diverse ML techniques, supervised deep learning (SDL) is capable of dealing with the large-scale complicated problems, which yet requires massive data to train the neutral networks. Although there have been a wealth of data sets produced by many terminal devices and base stations (BS), the generalization of the trained model fitting various scenarios is still extremely challenging. Federated learning, as a distributed learning algorithm of ML, can be used for decomposing a centralized problems into multiple distributed sub-problems, and solving them in a parallel way, meanwhile the users privacy can be better protected in light of information isolation across users. Typical delay optimization problems in URLLC are nonconvex, discrete, and built on dynamic Markov decision process (MDP). Deep reinforcement learning (DRL), one of the unsupervised learning algorithms and specializing in MDP-based problems, can be considered to tackle these highly complicated problems. There are also many non-stationary nature related issues in URLLC. Deep transfer learning, which updates part of neutral networks in real-time by using the data obtained from non-stationary environments, can be a promising processing strategy to ease the delay control. Inheriting mathematic rigorousness and power of artificial intelligence in dealing with more complex problems in communications systems and networks, the research community has recognized the importance and tendency of integrating model-based solutions with ML-based processing technologies. Particularly, mathematic modeling technologies are responsible for describing the problems characterizing essential cores, and rigorously simplifying formulation in communications networks. On the other hand, ML-based tools are used to efficiently derive

Yuquan Xiao, Qinghe Du (corresponding author), and Zixiao Zhao are with Xi'an Jiaotong University, China.

Wenchi Cheng is with Xidian University, China.

George K. Karagiannidis is with Aristotle University of Thessaloniki, Greece and also with Lebanese American University (LAU), Lebanon.

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62071373 and in part by the Innovation Capability Support Program of Shaanxi under the Grant No. 2021TD-08.

Digital Object Identifier: 10.1109/IOTM.001.2300049

2576-3180/23/$25.00 © 2023 IEEE

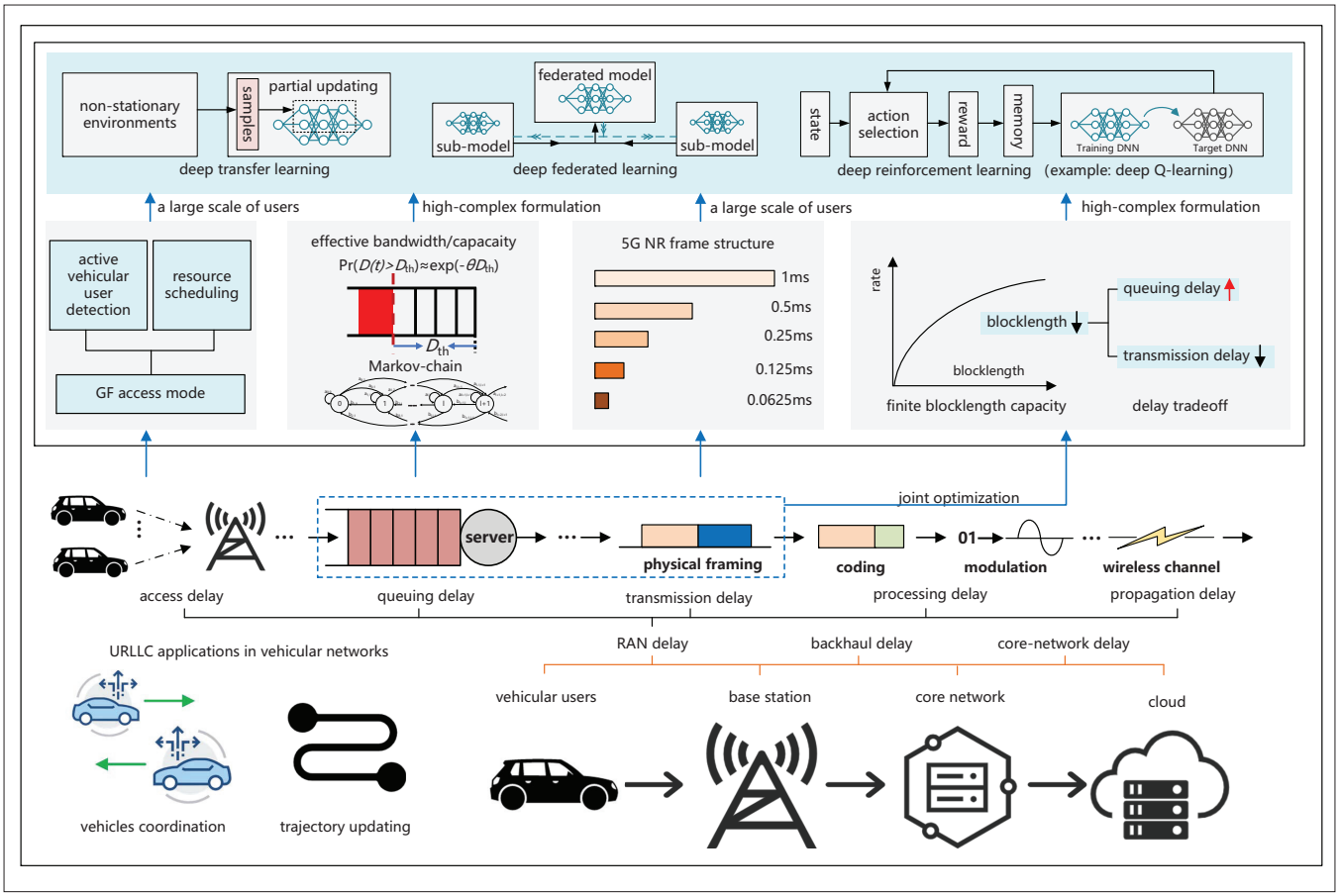

FIGURE 1. Model-ML integrated intelligence empowering URLLC in vehicular networks from the delay perspective.

the solver and identify the solution effectively. Following this mutual understanding, the model-ML integrated intelligence will become the key to unlock future URLLC.

Motivated by the above-mentioned paradigm, in this article we discuss how model-ML integrated intelligence empowers URLLC in vehicular networks from delay perspective. Specifically, we first analyze the constitution of end-to-end (E2E) delay for URLLC and then introduce model-ML integrated intelligence to help reducing major delay components, including access delay, queuing delay, and transmission delay. Next, we present an intelligent multi-tier-driven computing framework for access delay reduction, where model-based and ML-based designs are elaborated on, respectively. Furthermore, we derive the queuing delay and transmission delay tradeoff and show how to use an efficient resource allocation approach driven by multi-DRL networks to achieve the optimal balance between queuing delay and transmission delay. Finally, we share our opinions on open issues towards future URLLC and give the concluding remarks.

Model-ML Integrated Intelligence Achieving Lower Delay

There are many time-varying variables in wireless environments. Traditional model-based techniques can adapt to them to support URLLC in some small-scale vehicular networks and experimental environments, but it remains ultra challenging on fighting against or even making use of time-varying conditions to lower the E2E delay and suppress the fluctuations in large-scale vehicular networks and various environments. In this section, we firstly anatomize the constitution of E2E delay, among which radio access network (RAN) delay is the major contribution due to random access procedures as well as highly-varying channel qualities over the limited yet precious wireless resources. Then, we investigate the causes of RAN delay. Finally, we elaborate on model-ML integrated intelligence for reducing major components of RAN delay in vehicular networks.

Delay Components in URLLC

As shown in Fig. 1, E2E delay in URLLC mainly consists of RAN delay, backhaul delay, and core-network delay. Because of stochastic natures of the arrival traffics' load and highly dynamic characteristics of wireless channels, the fluctuation of RAN delay impacted by them causes a severe bottleneck for delay control in URLLC. RAN delay is composed of access delay, queuing delay, transmission delay, processing delay, and propagation delay. Specifically, access delay is the overall duration from a vehicular user equipment's (VUE) connection attempt to its success, which has to follow the competition-based random-access procedures required by the technique specifications. Queuing delay is the waiting time of each data packet in the buffer until it has been served. Processing delay is the time consumed in encoding, decoding, modulation, and demodulation. Transmission delay is defined as the amount of time used for carrying a data packet, which is inversely proportional to transmission rate. Propagation delay is defined as the duration of sending the signal from a VUE's antennas to BS's antennas over the air, which can be evaluated by dividing the distance between VUE and BS by velocity of electromagnetic waves. Since the velocity of electromagnetic waves is very high, the magnitude of propagation delay is typically far less than sub-millisecond level. In addition, the processing delay highly depends on the capability of hardware staying at almost a constant level. Processing delay and propagation delay are not the main causes of severe random delay over the RAN. As a consequence, how to control the access delay, queuing delay, and transmission delay is critical for URLLC to suppress the fluctuation as well as lower the RAN delay. To this end, we

will study model-ML integrated intelligence to reduce the access delay, queuing delay, and transmission delay, respectively.

Model-ML Based Techniques to Lower Access Delay

In current vehicular networks, which deploys the physical random access channel (PRACH) resources and enables the random access mode, the access delay caused by the standard handshake procedures will inevitably exceed 1 ms. To guarantee millisecond-level latency, URLLC employs grant-free (GF) access mode. Different from the traditional random access channel mode, the handshake phrase is skipped in GF access mode, in which the users are allowed to straight begin the transmission, and thus the access delay is reduced if no collision happens.

Active vehicular user detection is an important issue for GF access. When the number of URLLC users is small, the traditional model-based detection schemes can work well. When the user group is getting relative large, these schemes will not work effectively. In this situation, ML-based technique is a competitive candidate to improve the detection efficiency. There has been some ML-driven active user detection schemes for GF access, while most of them are designed for massive machine type communications (mMTC), where users choose to backoff once the collision happens. Clearly, this paradigm is not friendly to delay-sensitive transmissions. On one hand, URLLC shall avoid backoff due to the stringent delay requirement. On the other hand, practical URLLC scheme might often enable redundancy in occupying random access resources by users to improve the successful access probability. The representative schemes to shorten the delay of URLLC are *K*-repetition random access scheme and the variants [9, 10], where each packet is transmitted *K* times over different resource blocks and this operation can significantly reduce the access failure probability. In light of its simplicities, the *K*-repetition based schemes are a promising candidate for URLLC.

However, it costs extra resource usage. Efficient resource scheduling is another important problem to solve in GF access. Meeting the diverse QoS requirements meanwhile minimizing the overall resource consumption in URLLC is generally intractable especially when the size of URLLC user group is large. DRL can be employed to handle such complex situations lacking essential and clear mathematically-descriptive structure, which takes a try-error way to accumulate experience and finally find out the optimal scheduling policy. The fundamental challenge is how to integrate the basic elements of DRL, namely, *state*, *action*, and *reward*, with corresponding parameters in resource scheduling problems.

Towards the aforementioned topics, we present an intelligent multi-tier-driven computing framework to enhance access efficiency, where model-based and ML-based implementation will be given, respectively, with more details later.

Model-ML Based Techniques to Lower Queuing Delay

Queuing model has been extensively studied for many years. In the area of vehicular communications, Markov-chain model and large-deviation-principle (LDP)-based effective bandwidth/capacity model are the sharp weapons to analyze and control queuing delay. Specifically, Markov-chain model is often used to analyze the average queuing delay, and the effective bandwidth/capacity is powerful to bound queuing delay. With support by these tools, many efficient resource allocation solutions for URLLC are proposed, which are mostly derived by using convex optimization. Whereas, most of these schemes are base on ideal communications systems or with some necessary assumptions. For practical yet sophisticated URLLC systems, using the traditional optimization methods is hard to find the optimal solution. Although iterative algorithms are often able to track the solution closely, the duration of each iterative period is unaffordably long and thus not suitable for URLLC.

Although grant-free access technique has been a promising candidate for reducing access delay, existing works are mostly based on the fixed resource allocation strategy.

Most queuing-delay-related problems are proposed with communications domain knowledge and resolved by using machine learning (ML) techniques. Generally, the aim of tackling corresponding problems is to derive the efficient scheduling policy. Since the present and future states of queuing systems are highly affected by the scheduling policy, the formulated problems can be regarded as a Markov dynamic process (MDP). Following this characteristic, DRL can be a candidate for these intractable problems to find the optimal scheduling policy. Also, for effective bandwidth/capacity based problems, because the formulation usually leads complicated mathematics and thus the closed-form solution is not readily obtained, using deep learning networks to approach the near optimal solution can be an effective alternative, which is promising to avoid long processing delay compared with model-based iterative algorithms.

Model-ML Based Techniques to Lower Transmission Delay

Due to the fixed frame numerology in long-term evaluation (LTE) assisted vehicular networks, the transmission delay inevitably exceeds 1 ms. In order to meet strict end-to-end (E2E) delay bound outlined by the 3rd generation partnership project (3GPP) for time-sensitive services, 5G new radio (NR) is proposed and standardized, which has the potential to offer very small transmission delay. Specifically, the transmission time interval (TTI) is set to 0.5, 0.25, 0.125, and 0.0625 ms, respectively and the corresponding subcarrier spacing is 30, 60, 120, and 240 kHz. Different TTIs have its own unique advantages. Long TTI, corresponding to short subcarrier spacing, has a long cycle prefix, which is beneficial to covering large-scale cell and accommodating more users. Short TTI leads to small transmission delay and also corresponds to small blocklength. The achievable channel coding rate with respect to small blocklength is less than that of long blocklength [11]. Thus, the spectrum efficiency of short TTI is lower than that of long TTI, which results in a tradeoff between queuing delay and transmission delay. More information about this tradeoff will be shared later. Further, how to flexibly schedule these TTIs to well serve diverse URLLC traffics becomes a critical issue, which can be regarded as a classification problem. Alternatively, we can firstly generate the dataset through system-level simulation, in which the support to the diverse URLLC services can be tested. Then, the obtained dataset is used to train deep neural networks (DNN) via supervised deep learning (SDL). The well-trained DNN can then be deployed everywhere to implement the efficient and flexible TTIs schedule.

Intelligent Multi-Tier-Driven Computing Framework for High Access Efficiency

Although grant-free access technique has been a promising candidate for reducing access delay, existing works are mostly based on the fixed resource allocation strategy. However, since network load in URLLC is typically time-varying, the static allocation often leads to low resource utilization efficiency under light network load or severe collisions under heavy network load. To well match the resource to network load, we depict a multi-tier-driven computing framework, and then introduce the associated algorithms at each tier powered by model-based technique and ML technique, respectively. This multi-tier-driven computing framework can significantly improve access efficiency and is an effective solution for the issues presented later.

Multi-Tier-Driven Computing Framework

As shown in Fig. 2, the multi-tier-driven computing framework consists of three tiers, namely, network-load evaluation, network-load prediction, and adaptive resource allocation [9, 12]. Specifically, in

GF access mode, the wireless resources, i.e., resource blocks, will be randomly occupied by URLLC users who attempt to transmit data. Thus, the states of resource blocks, including success, collision, and idle, inevitably carry the information of network load in an implicit and hidden manner. Inspired by this thought, we design the network-load evaluation tier to extract the network-load information from the current states of resource blocks. The evaluated network-load information will be stored into history data pool. Then, the network-load prediction tier is to evaluate the network load for the next cycle by learning the history network-load information. Finally, the adaptive resource allocation tier yields the amount of resources desired to accommodate the coming network load, along with the QoS requirement of each one.

Model-/SDL-Based Technique Implementing Evaluation-Prediction Tier

Model-Based: Adjacent-occupation *K*-repetition GF access is considered as it is simple to implement without extra handshake. We can offer two model-based candidates for network-load evaluation. The first one is single-slot maximum likelihood with least squares (LS) estimation, derived by identifying the optimal N maximizing $P(R_{\text{suc}}, R_{\text{col}}, R_{\text{idl}} \mid N)$, where N represents the number of active vehicular users and $P(\cdot)$ represents probability observing R_{suc}, R_{col}, and R_{idl}, i.e., the numbers of resource blocks in success, collision, and idle states, respectively. Then, multiple separate slots can be fusioned to a more precise result by the LS algorithm. The computing complexity of this algorithm is comparatively low but the performance can be weakened because it does not sufficiently use the state information of other slots. The secondary one is multi-slot maximum likelihood indirect estimation, in which $\max_N P(\mathbf{R}_{\text{suc}}, \mathbf{R}_{\text{col}}, \mathbf{R}_{\text{idl}} \mid N)$ should be tackled. $\mathbf{R}_{\text{suc}}$ is the vector in the form of $(\mathbf{R}^1_{\text{suc}}, \mathbf{R}^2_{\text{suc}}, \ldots)$, which contains all information in successive states within the current *K*-repetition access cycle. Similarly, $\mathbf{R}_{\text{col}}$ and $\mathbf{R}_{\text{idl}}$ corresponds to that of collision and idle, respectively. The value evaluated using this algorithm is more accurate but bringing in much more complexity than that of the first one.

There are many model-based algorithms to predict network load, including simple equalling, moving average, exponential smoothing, and so on. For simple equalling algorithm, the next expected value is equal to the current observed value, which always falls behind the real change. The exponential smoothing algorithm is in the absence of long-term prediction. Alternatively, the moving average algorithms, especially the auto-regressive integrated moving average, which uses statistical tools to analyze the historical data and then predict future trends for stationary and part of non-stationary time series, are very suitable for sensing and predicting the burst real-time URLLC network-loads.

SDL-Based: The major challenge of using SDL to obtain the network-load function with respect to the state of resource block is how to fetch a wealth of learning samples. Generally, we can design an intermediator which takes initiative to harvest network-load information from users and state of resource from BS. Then, the state of resource in current cycle and the common network-load information corresponding the next cycle are grouped as a learning sample. Also, we can use the former model-based algorithms to generate samples then using these samples to train neutral networks in SDL. The SDL-based algorithm takes short processing delay compared with model-based algorithm when URLLC user group is large.

Model-/DRL-Based Techniques Implementing Scheduling Tier

Model-Based: For increasingly diverse services in URLLC, an effective scheduling scheme not only fits for the time-varying network load also satisfies the different QoS requirements. The URLLC traffic can be categorized into two folds:

1. Bursty traffic
2. Uniform traffic

Most of bursty traffic are critical control information, which are prior to uniform traffic, like state information. For model-based resource allocation scheme, the resource is scheduled in a priority way.

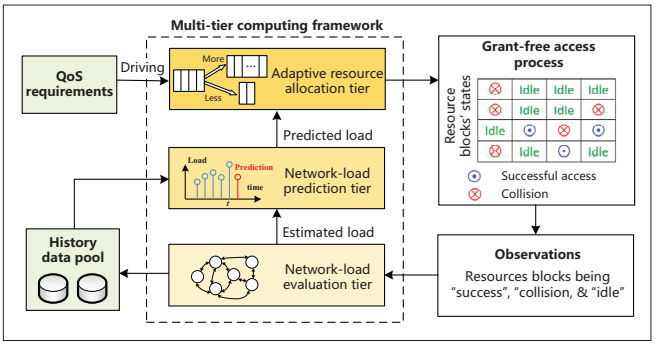

FIGURE 2. The multi-tier-driven computing framework to enhance access efficiency for URLLC.

When the resource is superfluous or the network load is slight, we set the QoS requirement of uniform traffic in hard mode. We aim to minimize the resource consumption meanwhile satisfying the QoS requirement of both bursty traffic and uniform traffic. When the resource is inadequate or the network load is heavy, the QoS requirement of uniform traffic is set in soft mode. Resource block is prior scheduled for bursty traffic and then for uniform traffic. Only when the lower-bound of QoS requirement for uniform traffic cannot be satisfied, the traffic is regarded as failed to support.

DRL-Based: We consider designing two DRL models. The first model aims to minimize overall allocated resource on the premise of assuring the QoS requirements and the second model is to not only minimize the overall allocated resource but also try the best to guarantee the QoS requirements. Specifically, for the first DRL model, each resource allocation action has been assured to satisfy the QoS requirements of bursty and uniform traffic. In contrast, the second model incorporates QoS requirements into the reward function, because the QoS requirements cannot be fully fulfilled. When traffic is low, the first DRL model is active while the second one remains dormant. When traffic is high, the second model becomes active while the first model is disabled. By alternating between the two models, our design allows us to access slight or heavy traffic loads with minimal resources, accommodating different traffic loads.

Model-DRL Integrated Intelligence for Joint Queuing Delay and Transmission Delay Minimization

In this section, we first discuss the tradeoff between queuing delay and transmission delay for URLLC, and clarify that there exists the optimal resource allocation scheme which can achieve the optimal balance between them. Then, we introduce how DRL can be used to seek the optimal resource allocation scheme.

Finite Blocklength Causing Queuing Delay and Transmission Delay Tradeoff

To lower transmission delay, short TTIs are designed in 5G. Short TTIs correspond to the finite blocklength. In the finite blocklength regime, the authors of [11] derived the more accurate limits for the channel achievable rate than Shannon capacity. Specifically, the achievable channel coding rate in finite blocklength, denoted by R, is given as follows:

$$R = \log_2(1+\gamma) - \sqrt{\frac{V}{n}} f_Q^{-1}(\varepsilon)\log_2 e + O\left(\frac{\log n}{n}\right), \qquad (1)$$

where γ is signal-to-noise (SNR) ratio, V is channel dispersion, n is blocklength, ε is block error rate,

$$O\left(\frac{\log n}{n}\right)$$

is the remainder term of

$$\frac{\log n}{n},$$

and $f_Q^{-1}(\cdot)$ is the inverse of Gaussian Q-function.

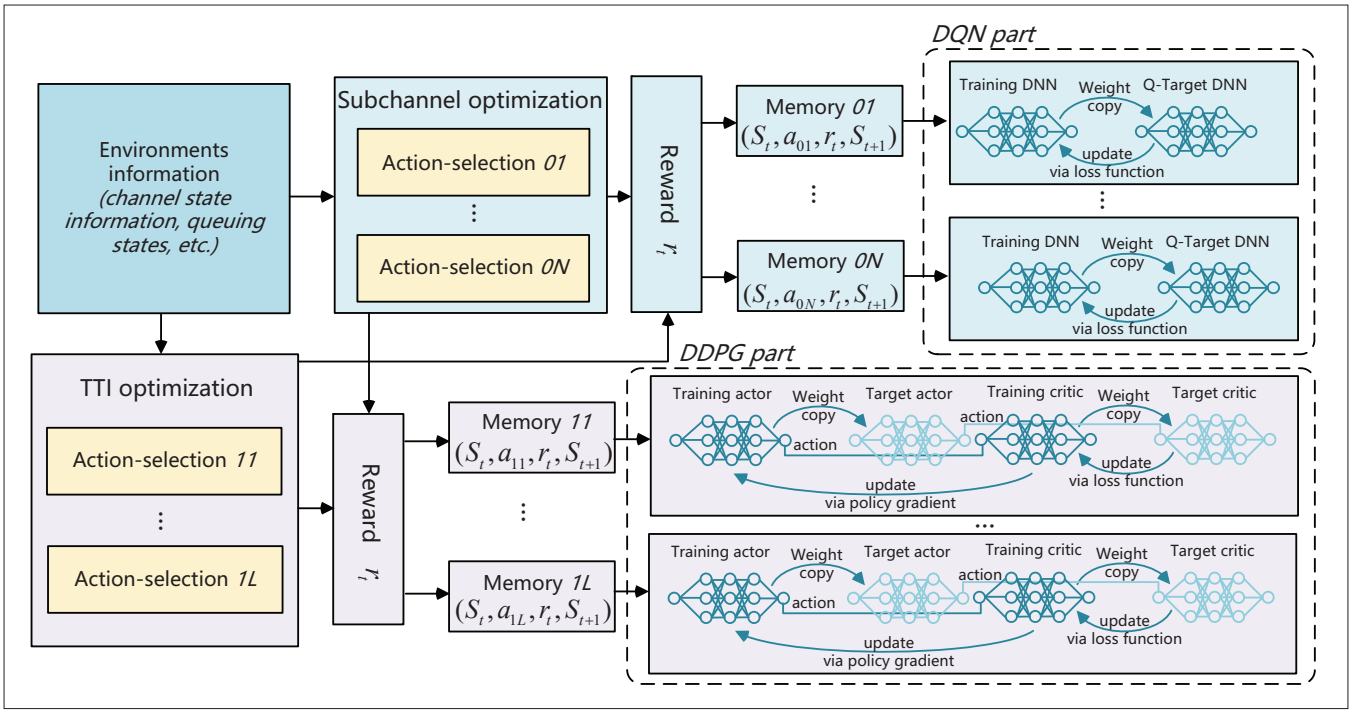

FIGURE 3. The optimal resource allocation approach driven by multi-DRL networks.

Equation 1 implies that the channel coding rate increases as blocklength increases. On one hand, as the channel coding rate increases, the service rate increases and the queuing delay accordingly decreases. On the other hand, as blocklength increases, the transmission delay increases. Short blocklength leads to small transmission delay but large queuing delay, and large blocklength results in large transmission delay but small queuing delay [15]. Therefore, there exists a tradeoff between transmission delay and queuing delay, and fine-tuning the blocklength is necessary to achieve the optimal balance, i.e., to minimize the sum of transmission and queuing delay.

It is worth mentioning that the optimal blocklength is highly affected by data traffic load. When the traffic load is slight, queuing delay often can be ignored, and thus we should shorten blocklength to reduce transmission delay. As traffic load becomes heavy, we should increase blocklength so as to obtain a large service rate helping reducing queuing delay. In practical vehicular networks, data traffic load is always time-varying. Thus, the adaptive blocklength, i.e., resource allocation, schemes should be studied to cater for the time-varying traffic load.

Optimal Resource Allocation Approach Driven by Multi-DRL Networks

Most of optimization problems related to the finite blocklength design are generally nonconvex and thus difficult to solve by using traditional model-based optimization methods. Facing this challenge, we attempt to employ DRL to resolve such problems. Specifically, deep Q-learning and deep deterministic policy gradient (DDPG) are the prevailing strategies in DRL to find the optimal solution for the formulated problems. Deep Q-learning is suitable for the discrete action space while DDPG is for continuous action space. Moreover, when the dimension of action space is large, multiple parallel DRL networks can be a competitive candidate to accelerate the convergence of learning and training process.

To achieve the optimal balance between queuing delay and transmission delay, we formulate the joint queuing delay and transmission delay minimization problem for multi-VUE downlink communications. We use the multiple DRL networks, shown in Fig. 3, to solve this problem. Since the subchannel is discrete and TTI is continuous, N deep Q-learning networks (DQN) are deployed for subchannel allocation and L DDPG networks are used for TTI optimization, respectively. Specifically, the environment variables, including queue states, i.e., the number of packets in the buffers, and CSI of wireless channel, are packaged and then informed to all action-selection modules. Then, the action-selection modules select the specified actions based on the environment variables, where the probabilistic greedy policy and the random noise policy are used for DQN and DDPG, respectively, to avoid falling into sub-optimal solutions. Next, all of selected actions are aggregated to reward modules, which take charge of calculating reward value. After that, *the environment variables, action, and reward* are grouped as a piece of experience data to store in memory. Finally, a fraction of experience data is chosen to periodically train and update DNN in DQN and the critic network in DDPG.

Figure 4 shows the performance of multi-DRL networks in solving joint queuing delay and transmission delay minimization problems. Specifically, Fig. 4a depicts the loss with respect to training iterations using different network parameters. For comparison, we also simulate the performance of using single-DQN to solve this problem. To simplify the writing, we use DDPG-DQN referring to the above multi-DRL networks. It can be seen from Fig. 4a that setting a small learning rate and a rational drop probability can make DNN fit well, where the random dropout is adopted to avoid DNN overfitting. Moreover, Fig. 4b shows the end-to-end (E2E) delay performance of DDPG-DQN, single-DQN, and the baseline schemes in 5G NR, where the drop probability is set to 0.5, the learning rate is set to 0.001, the SNR is set to 10 dB, the bandwidth is set to 10 MHz, and block error rate is set to 10^{-5}. For the different lengths of TTIs in 5G NR, shorter TTI can achieve a lower E2E delay when the average number of packets in the buffer is small. As the average number of buffering packets increases, the E2E delay corresponding to short TTI generally increases and then exceeds that corresponding to long TTI. Since the adaptive TTI is designed, DDPG-DQN can always achieve lowest E2E delay among these schemes. It can be also observed that the delay performance of single-DQN is poor, since the large step size associated with the large action space results in the performance degraded.

Open Issues Towards Future URLLC

We anatomize the constitution of the E2E delay for URLLC and elaborate on two special cases above to lower E2E delay by using model-ML integrated intelligence. There remain typically critical issues and challenges expected to be studied, which are presented as follows:

Challenge 1: Massive raw data cleaning as well as ML modeling for URLLC. Developing ML models requires a high-quality dataset. However, although there is a wealth of raw data available in vehicular networks, there are few valuable datasets regarding URLLC. Therefore, sifting through the massive raw data in vehicular networks is a critical issue. In addition, optimizing resource allocation for URLLC poses some complicated problems that can potentially be resolved by using heuristic ML algorithms. However, constructing robust neural network structures suitable for these problems while ensuring fast convergence rates to fulfill URLLC requirements in vehicular networks has not been fully understood.

Challenge 2: Metrics integrating delay and AoI. Delay performance has been extensively evaluated and enhanced in existing literatures. Recently, the concept of age of information (AoI) has been flourishing in the area of real-time status update services, which are one of typical services in URLLC. AoI focuses on the information freshness, which is impacted by not only delay but also source sampling rate. It has been demonstrated that even if the delay is small, AoI might remain large [14]. The small AoI can be achieved by fine tuning the sampling rate. However, the mathematical expression for doing so in the context of sophisticated vehicular networks is much more complex. Exploring how to facilitate URLLC with both delay and AoI metrics via ML is an avenue for future research.

Challenge 3: Security assurance for URLLC. Most existing secure transmission solutions need to introduce the data overhead as well as time budget, which are unfriendly for low-latency transmissions. How to achieve the low-latency transmission meanwhile assuring data's security in vehicular networks is an interesting issue. There is a novel technique exploiting the out-of-date characteristics of data to achieve secure transmissions [15]. Yet not all data in URLLC has a fast out-of-date rate. Ensuring secure transmissions for these data services in URLLC by fully exploiting the services' features in vehicular networks can be further researched.

Conclusions

In this article, we introduced how model-ML integrated intelligence empowers URLLC from the delay perspective. We began with anatomizing the constitution of the end-to-end delay for URLLC, and then presented an intelligent multi-tier-driven computing framework for reducing vehicular users' access delay. Furthermore, we introduced the tradeoff between queuing delay and transmission delay and proposed an efficient resource allocation approach driven by multi-DRL networks, which can achieve the optimal balance between queuing delay and transmission delay and significantly reduce E2E delay for URLLC over vehicular networks. Finally, we shared the open issues towards future URLLC, such as massive raw data cleaning as well as ML modeling for URLLC, metrics integrating delay and AoI, and security assurance for URLLC.

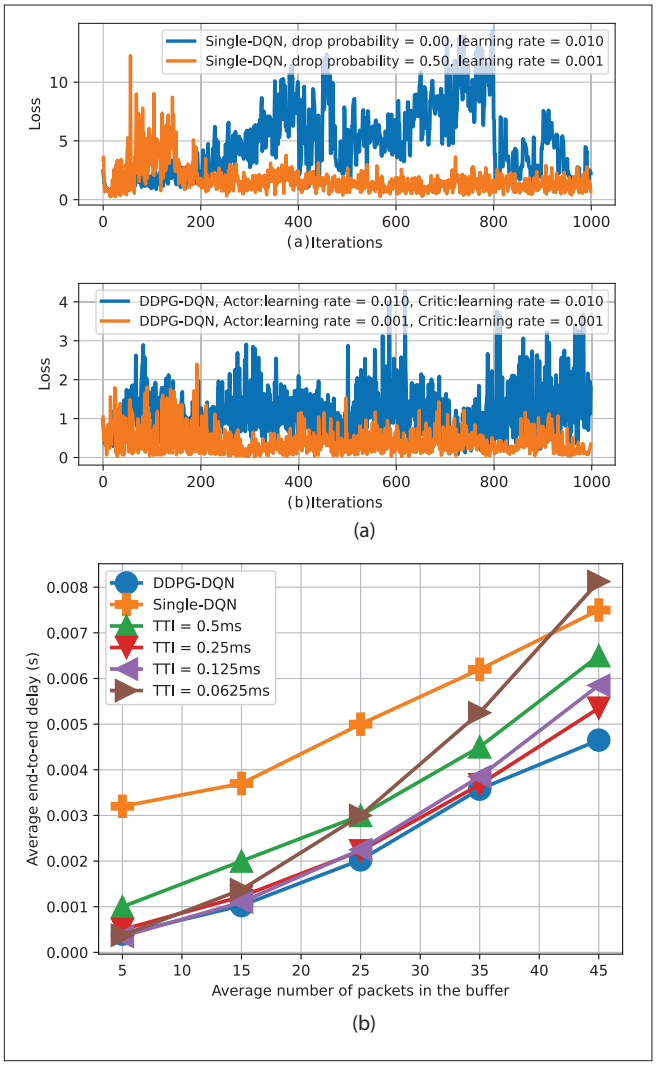

FIGURE 4. The performance of multi-DRL networks in solving joint queuing delay and transmission delay minimization problems: a) the loss with respect to training iterations using different network parameters, including drop probability and learning rate; b) the E2E delay performance of multi-DRL networks based scheme, single-DQN based scheme, and the baseline schemes in LTE and 5G NR.

References

[1] G. Ding *et al.*, "Two-Timescale Resource Management for Ultrareliable and Low-Latency Vehicular Communications," *IEEE Trans. Commun.*, vol. 70, no. 5, May 2022, pp. 3282–94.

[2] M. Giordani *et al.*, "Toward 6G Networks: Use Cases and Technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, Mar. 2020, pp. 55–61.

[3] S. Chen *et al.*, "A Vision of C-V2X: Technologies, Field Testing, and Challenges With Chinese Development," *IEEE Internet of Things J.*, vol. 7, no. 5, Feb. 2020, pp. 3872–81.

[4] D. Feng *et al.*, "Toward Ultrareliable Low-Latency Communications: Typical Scenarios, Possible Solutions, and Open Issues," *IEEE Vehic. Tech. Mag.*, vol. 14, no. 2, June 2019, pp. 94–102.

[5] C. She *et al.*, "A Tutorial on Ultrareliable and Low-Latency Communications in 6G: Integrating Domain Knowledge Into Deep Learning," *Proc. IEEE*, vol. 109, no. 3, Mar. 2021, pp. 204–46.

[6] Y. Yang *et al.*, "6G Network AI Architecture for Everyone-Centric Customized Services," *IEEE Network*, Early Access, 2022, pp. 1–10.

[7] J. Hoydis *et al.*, "Toward a 6G AI-Native Air Interface," *IEEE Commun. Mag.*, vol. 59, no. 5, Aug. 2021, pp. 76–81.

[8] D. C. Nguyen *et al.*, "Enabling AI in Future Wireless Networks: A Data Life Cycle Perspective," *IEEE Commun. Surveys & Tutorials*, vol. 23, no. 1, 1st Quarter 2021, pp. 553–5.

[9] Z. Zhao, Q. Du, and G. K. Karagiannidis, "Improved Grant-Free Access for URLLC via Multi-Tier-Driven Computing: Network-Load Learning, Prediction, and Resource Allocation," *IEEE JSAC*, vol. 41, no. 3, Mar. 2023, pp. 607–22.

[10] Y. Liu *et al.*, "Analyzing Grant-Free Access for URLLC Service," *IEEE JSAC*, vol. 39, no. 3, Aug. 2021, pp. 741–55.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Information Theory*, vol. 56, no. 5, May 2010, pp. 2307–59.

[12] Y. Yang, "Multi-Tier Computing Networks for Intelligent IoT," *Nature Electronics*, vol. 2, no. 1, pp. 4–5, 2019.

[13] W. Cheng *et al.*, "Adaptive Finite Blocklength for Ultra-Low Latency inWireless Communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, June 2022, pp. 4450–63.

[14] R. D. Yates *et al.*, "Age of Information: An Introduction and Survey," *IEEE JSAC*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[15] Q. Du *et al.*, "Statistical Security Model and Power Adaptation overWireless Fading Channels," *2015 Int'l. Conf. Wireless Communications & Signal Processing (WCSP)*, Oct. 2015, pp. 1–6.

Biographies

Yuquan Xiao (yqxiao@stu.xjtu.edu.cn) received the B.S. and M.S. degrees in information and communications engineering from Xidian University, Xi'an, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in information and communications engineering with Xi'an Jiaotong University, Xi'an. His research interests include real-time communications, physical-layer security techniques, and AI-empowered wireless communications.

Qinghe Du (duqinghe@mail.xjtu.edu.cn) received the B.S. degree in information engineering and the M.S. Degree in information and communications engineering from Xi'an Jiaotong University, China, in 2001 and 2004, respectively, and received the Ph.D. degree in computer engineering from Texas A&M University, College Station, USA, in 2010. He is currently a Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University, China. His research interests include mobile wireless communications and networking with emphasis on security assurance in wireless transmissions, AI-empowered networking technologies, 5G/B5G/6G networks and its evolution, cognitive radio networks, IoT, etc. He has published over 100 technical papers in international conferences and journals. He received the Best Paper Awards of the IEEE GLOBECOM 2007, IEEE COMCOMAP 2019, IEEE/CIC ICCC 2021, respectively, and received the Best Paper Awards of China Communications in 2017 and 2020, respectively. He has served as an Associate Editor of the *IEEE Communications Letters*, a guest Editor of *IEEE Network*, and serves as an Area Editor of KSII Transactions on Internet and Information Systems.

Wenchi Cheng (wccheng@xidian.edu.cn) received the B.S. and Ph.D. degrees in telecommunication engineering from Xidian University, Xian, China, in 2008 and 2013, respectively. He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, from 2010 to 2011. He is currently a Full Professor with Xidian University. His current research interests include B5G/6G wireless networks, emergency wireless communications, and orbital-angular-momentum-based wireless communications. He has published more than 100 international journal and conference papers in *IEEE Journal on Selected Areas in Communications*, IEEE magazines, IEEE Transactions, IEEE INFOCOM, GLOBECOM, and ICC. He received the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2021, the URSI Young Scientist Award in 2019, the Young Elite Scientist Award of CAST, and four IEEE journal/conference best papers. He has served or serving as the Wireless Communications Symposium Co-Chair for IEEE ICC 2022 and IEEE GLOBECOM 2020, the Publicity Chair for IEEE ICC 2019, the Next Generation Networks Symposium Chair for IEEE ICCC 2019, and the Workshop Chair for IEEE ICC 2019/IEEE GLOBECOM 2019/INFOCOM 2020 Workshop on Intelligent Wireless Emergency Communications Networks. He has served or serving as an Associate Editor for *IEEE Systems Journal*, *IEEE Communications Letters*, and *IEEE Wireless Communications Letters*.

George K. Karagiannidis (geokarag@auth.gr) is currently Professor in the Electrical and Computer Engineering Dept. of Aristotle University of Thessaloniki, Greece and Head of Wireless Communications and Information Processing (WCIP) Group. He is also Faculty Fellow in the Cyber Security Systems and Applied AI Research Center, Lebanese American University. His research interests are in the areas of Wireless Communications Systems and Networks, Signal processing, Optical Wireless Communications, Wireless Power Transfer and Applications and Communications and Signal Processing for Biomedical Engineering. He was in the past Editor in several IEEE journals and from 2012 to 2015 he was the Editor-in-Chief of *IEEE Communications Letters*. From September 2018 to June 2022 he served as Associate Editor-in-Chief of *IEEE Open Journal of Communications Society*. Currently, he is in the Steering Committee of IEEE Transactions on Cognitive Communications and Networks. Recently, he received three prestigious awards: The 2021 IEEE ComSoc RCC Technical Recognition Award, the 2018 IEEE ComSoc SPCE Technical Recognition Award and the 2022 Humboldt Research Award from Alexander von Humboldt Foundation. He is one of the highly-cited authors across all areas of Electrical Engineering, recognized from Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the eight consecutive years 2015–2022.

Zixiao Zhao (zzx67120787@stu.xjtu.edu.cn) received the B.S. degree in information engineering from Xi'an Jiaotong University, China, in 2021. She is currently persuing the M.S. degree in information and communications engineering from Xi'an Jiaotong University, China. Her current research interests include ultra-reliable low-latency communications, intrusion detection, and resource allocation.