Twin-Timescale Radio Resource Management for Ultra-Reliable and Low-Latency Vehicular Networks

Haojun Yang, *Student Member, IEEE*, Kan Zheng, *Senior Member, IEEE*, Long Zhao, *Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

Abstract—To efficiently support safety-related vehicular applications, the ultra-reliable and low-latency communication (URLLC) concept has become an indispensable component of vehicular networks (VNETs). Due to the high mobility of VNETs, exchanging near-instantaneous channel state information (CSI) and making reliable resource allocation decisions based on such short-term CSI evaluations are not practical. In this paper, we consider the downlink of a vehicle-to-infrastructure (V2I) system conceived for URLLC based on idealized perfect and realistic imperfect CSI. By exploiting the benefits of the massive MIMO concept, a two-stage radio resource allocation problem is formulated based on a novel twin-timescale perspective for avoiding the frequent exchange of near-instantaneous CSI. Specifically, based on the prevalent road-traffic density, Stage 1 is constructed for minimizing the worst-case transmission latency on a long-term timescale. In Stage 2, the base station allocates the total power at a short-term timescale according to the large-scale fading CSI encountered for minimizing the maximum transmission latency across all vehicular users. Then, a primary algorithm and a secondary algorithm are conceived for our V2I URLLC system to find the optimal solution of the twin-timescale resource allocation problem, with special emphasis on the complexity imposed. Finally, our simulation results show that the proposed resource allocation scheme significantly reduces the maximum transmission latency, and it is not sensitive to the fluctuation of road-traffic density.

Index Terms—Vehicular networks (VNET), ultra-reliable and low-latency communications (URLLC), resource management, finite blocklength theory, massive MIMO.

I. INTRODUCTION

GIVEN the rapid development of the Internet of things (IoT), the design of radio access networks is undergoing a paradigm shift from promoting spectral-efficiency and energy-efficiency toward connecting everything [1]. As a representative application scenario of IoT, vehicular networks (VNETs) support various vehicular applications for reducing the probability of traffic accidents and for improving both

the driving and infotainment experience [2], [3]. Traditional safety-related vehicular applications, such as platooning and road safety generally require a transmission latency within a few milliseconds and a reliability in terms of error probability down to 10^{-5} (or 10^{-6}). Furthermore, in order to achieve the ultimate goal of autonomous driving on the road, future autonomous vehicles need sophisticated sensors to support safety-related applications. Ultra-reliable and low-latency communications (URLLCs) exchanging near-real-time information are also required to offer assistance to control systems [4], [5]. To this end, it is paramount to conceive URLLC techniques for future VNETs. In the conception of URLLC techniques, radio resource management plays a key role in improving the performance of VNETs, hence ground-breaking research is required.

The latency encountered may be classified according to the hierarchical architecture of networks. The physical (PHY) layer aims for reducing the transmission latency, while the media access control (MAC) layer aspires to reduce the queueing latency. The theoretical analysis of PHY layer radio resource management conceived for device-to-device-based vehicle-to-vehicle communications can be found in [6], where both the transmission latency and reliability are considered simultaneously. However, the study of vehicular mobility was not considered in [6]. Moreover, neither the ergodic capacity nor the outage capacity are suitable for characterizing the tradeoff among the achievable rate, transmission latency and reliability in URLLCs [7]–[9]. Hence, the recent advances in studying an attractive tradeoff among the rate, latency and reliability of multiple-antenna aided channels are illustrated in [10] utilizing finite blocklength theory. However, there is still a paucity of literature on addressing the URLLC problems of the PHY layer in VNETs under the consideration of vehicular mobility models.

With respect to the MAC layer, sophisticated techniques have been used for analyzing and optimizing the queueing latency, such as Markov decision processes [11], [12], stochastic network calculus [13], [14] and queueing theory [15]. Based on the Markov decision process framework, a minimum queueing delay-based radio resource allocation scheme is proposed in [16] for vehicle-to-vehicle communications, where the successful packet reception ratio is considered as the reliability metric. However, it is not practical to inform the base station of the global channel state information (CSI) including the links which are not connected to the base station. Moreover, the analysis in [16] has not been provided for URLLCs, since it was not based on radical advances

in finite blocklength theory. Relying on finite blocklength theory and stochastic network calculus, a novel probabilistic delay bound is conceived in [17] for low-latency machine-to-machine applications. Furthermore, a URLLC cross-layer optimization framework is proposed in [18] by jointly considering the transmission latency and queueing latency. However, the advances mentioned above are not applicable to VNETs due to the lack of vehicular mobility model.

In addition to latency, reliability is another important performance metric for VNETs. In the MAC layer, reliability is generally characterized either by the stability of the queue [19] or by the probability of the queueing latency exceeding the tolerance threshold [20], [15]. As for the PHY layer, typically the latency vs reliability (packet error rate) tradeoff is quantified [21], where one of the pivotal parameters of ensuring a high reliability is deemed to be the multi-antenna diversity gain. Hence, large-scale MIMO schemes constitute a promising solution of reducing latency and enhancing reliability in VNETs. On a similar note, massive MIMO schemes can also be adopted for formulating large-scale fading-based optimization problems by exploiting the channel hardening phenomenon [22]–[24].

Against this background, in this paper we conceive radio resource management for URLLC VNETs based on a realistic vehicular mobility model. To conceive URLLC VNETs, the following pair of crucial characteristics has to be considered:

1) The near-instantaneous CSI rapidly becomes outdated and its more frequent update imposes a high overhead, hence reducing the communication efficiency; and
2) Compared to the time-scale of CSI fluctuation on the millisecond level, that of the road-traffic fluctuation is higher (second level). Requiring resource allocation updates on the order of milliseconds would impose an excessive implementation complexity.

In tackling the above-mentioned challenges in the downlink of a massive MIMO vehicle-to-infrastructure (V2I) system conceived for URLLC, our contributions are:

- We optimize the transmission latency based on a twin-timescale perspective for reducing the signaling overhead, where a macroscopic road-traffic model is adopted for charactering the relationship between the vehicular velocity and road-traffic density.
- We first derive the transmission latency based on finite blocklength theory for the matched filter (MF) precoder and zero-forcing (ZF) precoder both under perfect and imperfect CSI.
- Then, a two-stage radio resource allocation problem is formulated for our V2I URLLC system. In particular, based on the long-term timescale of road-traffic density, Stage 1 aims for optimizing the worst-case transmission latency by appropriately setting the system's bandwidth. For Stage 2, the base station (BS) allocates the total power based on the large-scale fading CSI for minimizing the maximum transmission latency guaranteeing fairness among all vehicular users (VUEs). Finally, the proposed primary and secondary radio resource allocation algorithms are invoked for optimally solving the above two

problems at a reasonable complexity.

- Our simulation results show that the proposed resource allocation scheme is capable of significantly reducing the maximum transmission latency, while flexibly accommodating the road-traffic fluctuation. Hence, the radio resources can be allocated at a larger time interval, which means the proposed algorithms can get the effect of semi-persistent scheduling. Additionally, the ZF precoder is shown to constitute a compelling choice for our V2I URLLC system.

The remainder of this paper is organized as follows. First of all, finite blocklength theory is introduced in Section II. Section III describes the V2I URLLC system model, while Section IV formulates the twin-timescale resource allocation problem. Then, the proposed resource allocation algorithms are discussed in Section V. Finally, Section VI illustrates the simulation results, while our conclusions are offered in Section VII.

Notations: Uppercase boldface letters and lowercase boldface letters denote matrices and vectors, respectively, while $\text{diag}\{\mathbf{a}\}$ is a diagonal square matrix whose main diagonal is formed by vector $\mathbf{a}$. Furthermore, $(\cdot)^{\text{T}}$, $(\cdot)^{*}$, $(\cdot)^{\text{H}}$ and $(\cdot)^{-1}$ represent the transpose, conjugate, conjugate transpose and pseudo-inverse of a matrix/vector, respectively, while $\|\cdot\|$ denotes the Euclidian norm, and $[\cdot]_{kk}$ denotes the k-th diagonal element of a square matrix. Finally, $\mathbb{E}(\cdot)$ represents the mathematical expectation, while $\mathbb{CN}(\mu, \sigma^2)$ is the complex Gaussian distribution with mean μ and real/imaginary component variance $\sigma^2/2$.

II. PRELIMINARIES: FINITE BLOCKLENGTH THEORY

The Shannon formula quantifies the error-free capacity at which information can be transmitted over a band-limited channel in the presence of noise and interferences,

$$C(\gamma) = \mathbb{E}\left[B \log_2(1 + \gamma)\right], \quad (1)$$

where B is the bandwidth and γ is the signal-to-interference-plus-noise ratio (SINR). This capacity can only be approached at the cost of excessive coding latency and complexity, i.e.,

$$C(\gamma) = \lim_{\epsilon \to 0} C_\epsilon(\gamma, \epsilon) = \lim_{\epsilon \to 0} \lim_{n \to \infty} R(\gamma, n, \epsilon), \quad (2)$$

where $C_\epsilon(\gamma, \epsilon)$ is the so-called outage capacity at the error probability ϵ. For the operational wireless systems, both the ergodic and outage capacity are reasonable performance metrics, because the packet size is typically large.

However, the assumption of large packet size does not meet the requirements of URLLCs. Thus, a more refined analysis of $R(\gamma, n, \epsilon)$ is needed. Fortunately, during the last few years, significant progress has been made for satisfactorily addressing the problem of approximating $R(\gamma, n, \epsilon)$ [7], [8], [25]:

$$R(\gamma, L, \epsilon) \approx \mathbb{E}\left\{B\left[\log_2(1 + \gamma) - \sqrt{\frac{V}{LB}}Q^{-1}(\epsilon)\right]\right\}, \quad (3)$$
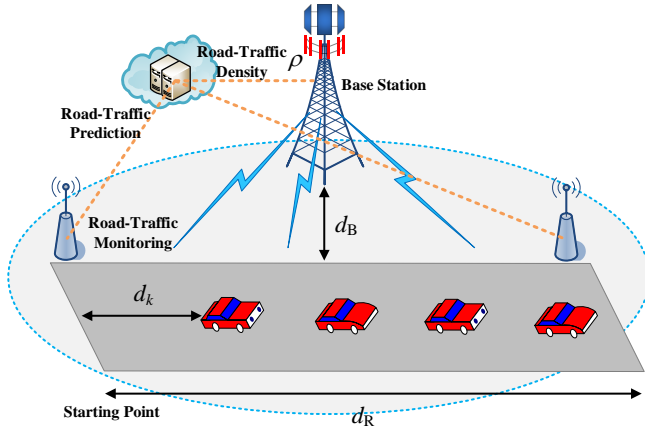
Fig. 1. Illustration of system model.

where $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q-function and V is the so-called channel dispersion. For a complex channel, the channel dispersion is given by [7], [8]:

$$V = \left(1 - \frac{1}{(1+\gamma)^2}\right)(\log_2 e)^2. \tag{4}$$

Furthermore, L is the transmission latency[1], while LB, which is also referred to as the blocklength of channel coding, represents the number of transmitted symbols. When LB is high enough, the approximation (3) approaches the ergodic capacity. Compared to the ergodic capacity, the approximation (3) also implies that the rate reduction is proportional to $1/\sqrt{LB}$, when aiming for meeting a specific error probability at a given packet size. Finite blocklength theory constitutes a powerful technique of dealing with the URLLC-related optimization problems. In addition to the contributions mentioned above, some new advances based on (3) analyze and optimize the URLLC performance of 5G and IoT networks [26]–[28].

III. V2I URLLC SYSTEM MODEL

As shown in Fig. 1, for the downlink of the V2I URLLC system, we consider a single roadside BS and a highway road segment of length d_R. The BS is d_B meters away from the road. Furthermore, the BS employs M antennas and simultaneously sends information to K single-antenna aided VUEs (massive MIMO with $M \gg K$). The system operates in the time-division duplex (TDD) mode.

A. Road-Traffic Model

According to the classic theory, the road-traffic model is divided into macroscopic and microscopic scales. Specially, the macroscopic model describes the average behavior of a certain number of vehicles at specific locations and instances, treating the road-traffic similarly to fluid dynamics. By contrast, the microscopic model describes the specific behavior of

---

[1]In general, the one-way latency of the PHY layer entails both the propagation latency of electromagnetic wave and the coding latency. For example, for a 300 m cell radius, the propagation delay of 1 μs is typically negligible. Hence, the coding latency dominates the one-way delay of the PHY layer's transmission latency.

each individual entity (such as vehicle or pedestrian), hence it is more sophisticated than the macroscopic model. The open-source simulation of urban mobility (SUMO) mainly focuses on generating the microscopic model, namely the trajectory of vehicles. However, our proposed algorithms depend both on the road-traffic density and on the location information of all VUEs at a certain moment. Based on the macroscopic model, the above two types of information can be predicted and periodically reported by the traffic center [29], [30]. To this end, we adopt the one-dimensional macroscopic model in this paper, which is also known as the Lighthill-Whitham-Richards (LWR) model [31]. Given the road-traffic density ρ and the road-traffic velocity v, the road-traffic flow rate (flux) f is generally given by

$$f = \rho v. \tag{5}$$

Numerous models have been proposed for characterizing how the vehicular velocity depends on the road-traffic density. In this paper, the Underwood model is employed, and its speed-density function can be written as [31]:

$$v(\rho) = v_F \exp\left(\frac{-\rho}{\rho_m}\right), \tag{6}$$

where v_F is the free-flowing velocity and ρ_m is the maximum density. Moreover, the road-traffic density can be used for modeling the number of VUEs, according to $K = \rho d_R$.

B. Channel Model

All channels experience independent flat block-fading, i.e., they remain constant during a a coherence block, but change independently from one block to another. Let $\mathbf{G} = \mathbf{D}^{1/2}\mathbf{H} = [\mathbf{g}_1, \cdots, \mathbf{g}_K]^T \in \mathbb{C}^{K \times M}$ denote the system's CSI, where the diagonal matrix $\mathbf{D} = \text{diag}\{\beta_1, \cdots, \beta_K\} \in \mathbb{R}^{K \times K}$ and $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]^T \in \mathbb{C}^{K \times M}$ represent the large-scale fading and fast fading. The channel vector spanning from the BS to the k-th VUE is given by $\mathbf{g}_k = \beta_k^{1/2}\mathbf{h}_k \in \mathbb{C}^{M \times 1}$. Furthermore, imperfect channel estimation is considered [32]–[34], i.e., $\mathbf{g}_k = \beta_k^{1/2}\mathbf{h}_k = \beta_k^{1/2}\sqrt{\chi_k}\hat{\mathbf{h}}_k + \beta_k^{1/2}\sqrt{1-\chi_k}\mathbf{e}_k$, where $\hat{\mathbf{h}}_k$, $\mathbf{e}_k$ and $\chi_k \in \mathbb{R}$ are the estimate, error and estimation accuracy of $\mathbf{h}_k$, respectively. If $\chi_k = 1$, $\mathbf{g}_k$ corresponds to perfect channel estimation. Finally, all random variables are independent and identically distributed (i.i.d.) complex Gaussian random variables with mean 0 and variance 1, namely $\hat{h}_{km}, e_{km} \overset{\text{i.i.d.}}{\sim} \mathbb{CN}(0,1)$.

C. Transceiver Model

1) Transmitter: It is widely exploited that low-complexity linear transmit precoders are capable of attaining asymptotically optimal performance in massive MIMO. Therefore we consider the MF precoder and ZF precoder in this paper. Let $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_K] \in \mathbb{C}^{M \times K}$ and $\mathbf{v}_k \in \mathbb{C}^{M \times 1}$ denote the precoder matrix and vector, where

$$\mathbf{V} = \begin{cases} \hat{\mathbf{G}}^H, & \text{for MF,} \\ \hat{\mathbf{G}}^H\left(\hat{\mathbf{G}}\hat{\mathbf{G}}^H\right)^{-1}, & \text{for ZF,} \end{cases} \tag{7}$$

and $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \cdots, \hat{\mathbf{g}}_K]^{\mathrm{T}}$ with $\hat{\mathbf{g}}_k = \beta_k^{1/2}\hat{\mathbf{h}}_k$. The normalized form of $\mathbf{V}$ is given by $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K] \in \mathbb{C}^{M \times K}$ with

$$\mathbf{w}_k = \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|}, \forall k. \tag{8}$$

As a result, given the symbol vector $\mathbf{s} = [s_1, \cdots, s_K]^{\mathrm{T}} \in \mathbb{C}^{K \times 1}$ with $\mathbb{E}\left[\mathbf{s}\mathbf{s}^{\mathrm{H}}\right] = \mathbf{I}_K$, the transmitted signal vector of all VUEs can be written as

$$\mathbf{x} = \mathbf{W}\mathbf{P}^{1/2}\mathbf{s}, \tag{9}$$

where $\mathbf{P} = \mathrm{diag}\left\{\frac{\mathbf{P}}{M}\right\} = \frac{1}{M}\mathrm{diag}\left\{p_1, \cdots, p_K\right\} \in \mathbb{R}^{K \times K}$ is the power allocation matrix and the total power at the BS is P_{B}.

2) Receiver: The received signal vector is given by

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{n} = \mathbf{G}\mathbf{W}\mathbf{P}^{1/2}\mathbf{s} + \mathbf{n}, \tag{10}$$

where $\mathbf{n} = [n_1, \cdots, n_K]^{\mathrm{T}} \in \mathbb{C}^{K \times 1}$ is the i.i.d. additive white Gaussian noise with $n_k \overset{\text{i.i.d.}}{\sim} \mathbb{CN}(0, \sigma^2)$. The received signal of the k-th VUE can be written as

$$y_k = \sqrt{\frac{p_k}{M}}\sqrt{\chi_k}\hat{\mathbf{g}}_k^{\mathrm{T}}\mathbf{w}_k s_k + \sum_{i=1, i \neq k}^{K} \sqrt{\frac{p_i}{M}}\sqrt{\chi_k}\hat{\mathbf{g}}_k^{\mathrm{T}}\mathbf{w}_i s_i$$
$$+ \sum_{j=1}^{K} \sqrt{\frac{p_j}{M}}\sqrt{1 - \chi_k}\beta_k^{1/2}\mathbf{e}_k^{\mathrm{T}}\mathbf{w}_j s_j + n_k, \forall k. \tag{11}$$

Hence, the SINR of the k-th VUE is given by

$$\gamma_k = \frac{\frac{p_k}{M}\chi_k\left|\hat{\mathbf{g}}_k^{\mathrm{T}}\mathbf{w}_k\right|^2}{\sum\limits_{i=1, i \neq k}^{K}\frac{p_i}{M}\chi_k\left|\hat{\mathbf{g}}_k^{\mathrm{T}}\mathbf{w}_i\right|^2 + \frac{P_{\mathrm{B}}}{M}\beta_k(1 - \chi_k) + \sigma^2}$$

$$= \begin{cases} \dfrac{\frac{p_k}{M}\chi_k\left|\hat{\mathbf{g}}_k^{\mathrm{T}}\frac{\hat{\mathbf{g}}_k^*}{\|\hat{\mathbf{g}}_k^*\|}\right|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K}\frac{p_i}{M}\chi_k\left|\hat{\mathbf{g}}_k^{\mathrm{T}}\frac{\hat{\mathbf{g}}_i^*}{\|\hat{\mathbf{g}}_i^*\|}\right|^2 + \frac{P_{\mathrm{B}}}{M}\beta_k(1-\chi_k) + \sigma^2}, & \text{for MF,} \\[8pt] \dfrac{1}{\left[\left(\hat{\mathbf{G}}\hat{\mathbf{G}}^{\mathrm{H}}\right)^{-1}\right]_{kk}}\dfrac{\frac{p_k}{M}}{\frac{P_{\mathrm{B}}}{M}\beta_k(1-\chi_k) + \sigma^2}, & \text{for ZF.} \end{cases} \tag{12}$$

Recall from $\hat{\mathbf{g}}_k = \beta_k^{1/2}\hat{\mathbf{h}}_k$ that the instantaneous γ_k^{-1} can be rewritten as

$$\gamma_k^{-1} = \begin{cases} \sum\limits_{\substack{i=1 \\ i \neq k}}^{K}\frac{p_i}{p_k}\left|\frac{\hat{\mathbf{h}}_k^{\mathrm{T}}}{\|\hat{\mathbf{h}}_k^{\mathrm{T}}\|}\frac{\hat{\mathbf{h}}_i^*}{\|\hat{\mathbf{h}}_i^*\|}\right|^2 + \frac{P_{\mathrm{B}}\beta_k(1-\chi_k)+M\sigma^2}{p_k\chi_k\beta_k\|\hat{\mathbf{h}}_k\|^2}, & \text{for MF,} \\[8pt] \frac{P_{\mathrm{B}}\beta_k(1-\chi_k)+M\sigma^2}{p_k\chi_k\beta_k}\left[\left(\hat{\mathbf{H}}\hat{\mathbf{H}}^{\mathrm{H}}\right)^{-1}\right]_{kk}, & \text{for ZF.} \end{cases} \tag{13}$$
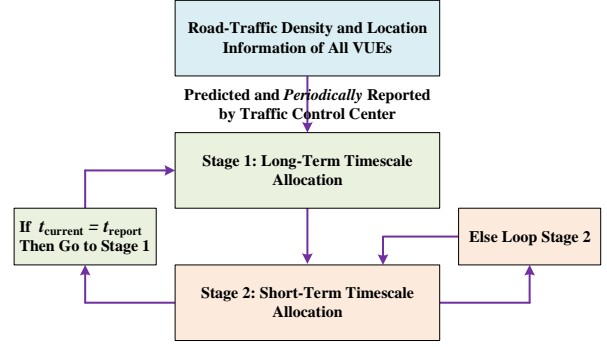

Fig. 2. Illustration of twin-timescale radio resource management.

Based on [35], we have the following distributions, i.e.,

$$\Omega_{\mathrm{B}}^{\mathrm{MF}} \triangleq \left|\frac{\hat{\mathbf{h}}_k^{\mathrm{T}}}{\|\hat{\mathbf{h}}_k^{\mathrm{T}}\|}\frac{\hat{\mathbf{h}}_i^*}{\|\hat{\mathbf{h}}_i^*\|}\right|^2 \sim \mathrm{Beta}\left(1, M-1\right), \forall i \neq k, \tag{14}$$

$$\Omega_{\mathrm{G}}^{\mathrm{MF}} \triangleq \frac{1}{\|\hat{\mathbf{h}}_k\|^2} \sim \mathrm{Gamma}^{-1}\left(M, 1\right), \forall k, \tag{15}$$

$$\Omega_{\mathrm{G}}^{\mathrm{ZF}} \triangleq \left[\left(\hat{\mathbf{H}}\hat{\mathbf{H}}^{\mathrm{H}}\right)^{-1}\right]_{kk} \sim \mathrm{Gamma}^{-1}\left(M-K+1, 1\right), \forall k. \tag{16}$$

Then γ_k^{-1} can be finally expressed as

$$\gamma_k^{-1} = \begin{cases} \frac{P_{\mathrm{B}}-p_k}{p_k}\Omega_{\mathrm{B}}^{\mathrm{MF}} + \frac{P_{\mathrm{B}}\beta_k(1-\chi_k)+M\sigma^2}{p_k\chi_k\beta_k}\Omega_{\mathrm{G}}^{\mathrm{MF}}, & \text{for MF,} \\[8pt] \frac{P_{\mathrm{B}}\beta_k(1-\chi_k)+M\sigma^2}{p_k\chi_k\beta_k}\Omega_{\mathrm{G}}^{\mathrm{ZF}}, & \text{for ZF.} \end{cases} \tag{17}$$

IV. PROBLEM FORMULATION OF TWIN-TIMESCALE RADIO RESOURCE MANAGEMENT

In this treatise, we optimize the transmission latency of V2I communications based on the road-traffic density and location information, avoiding the frequent exchange of instantaneous global CSI. As the speed of mobility gradually increases, the fluctuation of the small-scale channel fading envelope generally becomes more rapid. However, the impact of small-scale fading can be mitigated by exploiting the channel hardening and asymptotic orthogonality of massive MIMO [22], [35], [23]. As a result, the mobility of vehicles (the change of location information) only impacts the large-scale channel fading. Furthermore, compared to the time-scale of CSI fluctuation at the millisecond scale, that of the road-traffic (including the density and location information) fluctuation is on the scale of say 90 km/h (2.5 cm/ms). Therefore, considering the above facts, we formulate a twin-timescale large-scale fading-based radio resource allocation problem in order to reduce the signaling overhead imposed. As show in Fig. 2, according to the road-traffic density, the first stage is constructed for optimizing the worst-case latency by setting the system's bandwidth from a long-term timescale. As for the second stage of the short-term timescale, the BS will allocate the total power based on the large-scale fading CSI (equivalent to location information) to minimize the maximum transmission latency for ensuring fairness amongst all VUEs.

A. Large-Scale Fading-Based Transmission Latency

In high-speed vehicular environments, small-scale channel fading tends to fluctuate rapidly, hence the instantaneous V2I CSI is often outdated. Furthermore, exchanging the instantaneous V2I CSI with the BS is not practical. As a result, large-scale fading-based allocation problems are generally considered by exploiting the channel hardening and asymptotic orthogonality of massive MIMO [36], [22], [35], [23]. The theorem for quantifying the maximum achievable rate is first formulated as follows.

Theorem 1: When the MF or ZF precoder is employed in the downlink of our V2I URLLC system, the maximum achievable ergodic rate of the k-th VUE can be approximated ($M \gg 1$) by [7], [8], [25]:

$$R_k(\Gamma_k, L_k, \epsilon_k) = B \log_2 (1 + \Gamma_k)$$
$$- \log_2 e \sqrt{\frac{B}{L_k}} Q^{-1}(\epsilon_k) \sqrt{1 - (1 + \Gamma_k)^{-2}}, \forall k, \quad (18)$$

where

$$\Gamma_k = \begin{cases} \dfrac{M p_k}{P_B - p_k + \varphi_k}, & \text{for MF,} \\ p_k \varphi_k, & \text{for ZF,} \end{cases} \quad (19)$$

and

$$\varphi_k = \begin{cases} \dfrac{P_B \beta_k (1 - \chi_k) + M \sigma^2}{\chi_k \beta_k} \dfrac{M}{M - 1}, & \text{for MF,} \\ \dfrac{\chi_k \beta_k (M - K)}{P_B \beta_k (1 - \chi_k) + M \sigma^2}, & \text{for ZF.} \end{cases} \quad (20)$$

Proof: See Appendix A. ∎

According to Theorem 1, we can obtain the following corollary for illustrating the transmission latency.

Corollary 1: Given the desired rate $R_k(\Gamma_k, L_k, \epsilon_k)$ and the reliability ϵ_k, the transmission latency of the k-th VUE can be expressed as

$$L_k = \left\lceil \frac{\sqrt{B} Q^{-1}(\epsilon_k) \log_2 e \sqrt{1 - (1 + \Gamma_k)^{-2}}}{B \log_2 (1 + \Gamma_k) - R_k(\Gamma_k, L_k, \epsilon_k)} \right\rceil^2, \forall k. \quad (21)$$

With respect to the reliability, $Q^{-1}(\epsilon)$ with $\epsilon \in [0, 1]$ obeys the following properties,

$$Q^{-1}(\epsilon) \begin{cases} < 0, & \epsilon > 0.5, \\ = 0, & \epsilon = 0.5, \\ > 0, & \epsilon < 0.5, \end{cases} \quad (22)$$

and $\lim_{\epsilon \to 0} Q^{-1}(\epsilon) = +\infty$. Thus, a higher reliability (also known as $1 - \epsilon$) leads to a higher latency, and vice versa.

B. Stage 1: Long-Term Timescale Allocation

In somewhat simplistic, but plausible terms one could argue that doubling the system bandwidth allows doubling the bitrate, halving the framelength and hence halving the transmission latency. This plausible concept is also adopted by the 3GPP [37] for reducing the transmission slot-duration. Another influential factor is the coding delay, because the ultimate Shannon capacity limit is only achievable, if the codeword-length tends to infinity. Broadly speaking, the system bandwidth can be appropriately adjusted for reducing the transmission latency, subject to the ubiquitous bandwidth constraints. Hence the system bandwidth determined in Stage 1 can also be viewed as part of the system design. Based on the long-term timescale of road-traffic density, the objective of Stage 1 is to optimize the worst-case transmission latency. Notice that the road-traffic density is valid for a limited road segment within about 50 m to 200 m [29], [30]. Furthermore, our system model is based on a highway segment rather than on an urban scenario. The BS is deployed very close to the road segment, and thus generally there are no signal obstructions. To this end, the impact of shadow fading is ignored in this paper [38], [39]. Without loss of generality, the worst situation is encountered in the scenario of assuming a largest distance from the BS in case of equal power allocation. The path loss at the largest distance from the BS is given by

$$\beta_W = \theta \left[d_B^2 + \left(\frac{d_R}{2} \right)^2 \right]^{-\frac{\alpha}{2}}, \quad (23)$$

where θ is a constant related both to the antenna gain and to the carrier frequency, while α is the path loss exponent. Then, with $p_k = \frac{P_B}{K} = \frac{P_0 B}{\rho d_R}$, the worst-case SINR can be expressed as

$$\Gamma_W = \begin{cases} \dfrac{M P_0}{P_0 (\rho d_R - 1) + \frac{P_0 \beta_W (1 - \chi_{th}) + M N_0}{\chi_{th} \beta_W} \frac{\rho d_R M}{M - 1}}, & \text{for MF,} \\ \dfrac{P_0}{\rho d_R} \dfrac{\chi_{th} \beta_W (M - \rho d_R)}{P_0 \beta_W (1 - \chi_{th}) + M N_0}, & \text{for ZF,} \end{cases} \quad (24)$$

where P_0 and N_0 are the power spectral density of the transmitted signal and noise, respectively, while χ_{th} is the minimum requested estimation accuracy, namely $\chi_{th} = \min_k \{\chi_k\}$. Therefore, the worst-case transmission latency L_W is finally given by

$$L_W = \left\lceil \frac{\sqrt{B} Q^{-1}(\epsilon_W) \log_2 e \sqrt{1 - (1 + \Gamma_W)^{-2}}}{B \log_2 (1 + \Gamma_W) - R_W} \right\rceil^2, \quad (25)$$

where $\epsilon_W = \min_k \{\epsilon_k\}$ and $R_W = \max_k \{R_k(\Gamma_k, L_k, \epsilon_k)\}$.

On the other hand, with (6), the maximum Doppler frequency is given by

$$f_{MD} = \frac{v(\rho)}{\lambda} = \frac{f_C v_F \exp(-\rho/\rho_m)}{c}, \quad (26)$$

where f_C is the carrier frequency, and c is the speed of light. Hence, the typical coherence time can be calculated as [40]:

$$T_C = \sqrt{\frac{9}{16 \pi f_{MD}^2}} = \sqrt{\frac{9 c^2}{16 \pi f_C^2 v_F^2 [\exp(-\rho/\rho_m)]^2}}. \quad (27)$$

The design of the existing wireless systems follows a rule that CSI remains constant during the coherence time [41], [37]. Thus, motivated by this rule, we assume that the worst-case transmission latency is lower than the coherence time, which is similar with the design of the LTE-based systems. On the basis of the above derivations, the following optimization problem is formulated.

Problem 1 (Long-Term Timescale Bandwidth Allocation): Let δ denote a very small threshold constant. Given the road-traffic density ρ, the long-term timescale bandwidth allocation problem of Stage 1 is formulated as

$$\textbf{P1:} \quad L_{\text{W}}(B, \rho) \leqslant \delta T_{\text{C}}(\rho). \tag{28}$$

Typically, δ is equal to $\frac{1}{20}$ for the LTE-based systems [41]. Hence, in our simulations, we let δ be equal to $\frac{1}{20}$ or $\frac{1}{40}$.

C. Stage 2: Short-Term Timescale Allocation

By solving Problem 1, the total power at the BS can be obtained, namely $P_{\text{B}} = P_0 B^*(\rho)$. In Stage 2, the BS allocates the total power based on the short-term timescale large-scale fading CSI of all VUEs for minimizing the maximum transmission latency. Then the following optimization problem is formulated.

Problem 2 (Short-Term Timescale Power Allocation): Given the desired rate $R_k(\Gamma_k, L_k, \epsilon_k)$, reliability ϵ_k, total power P_{B} and the large-scale fading CSI of all VUEs, the optimization objective of Stage 2 is to minimize the maximum transmission latency of all VUEs ensuring fairness amongst all VUEs, i.e.,

$$\textbf{P2:} \quad \min_{\mathbf{p}} \ \max_{k} \{L_k\} \tag{29a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} p_k = P_{\text{B}}, \tag{29b}$$

$$p_k \geqslant 0, \forall k. \tag{29c}$$

V. TWIN-TIMESCALE RESOURCE ALLOCATION ALGORITHMS FOR V2I URLLC SCYSTEM

In this section, the solutions to the above pair of problems are first studied, and then efficient twin-timescale allocation algorithms are proposed for our V2I URLLC system.

A. Solution to Problem 1

According to Problem 1, we have

$$\frac{\sqrt{B}Q^{-1}(\epsilon_{\text{W}})\log_2 e\sqrt{1 - (1+\Gamma_{\text{W}})^{-2}}}{B\log_2(1+\Gamma_{\text{W}}) - R_{\text{W}}(\gamma, L, \epsilon)} \leqslant \sqrt{\delta T_{\text{C}}}, \tag{30}$$

which shows that

$$B\sqrt{\delta T_{\text{C}}}\log_2(1+\Gamma_{\text{W}}) - \sqrt{\delta T_{\text{C}}}R_{\text{W}}(\gamma, L, \epsilon)$$
$$- \sqrt{B}Q^{-1}(\epsilon_{\text{W}})\log_2 e\sqrt{1 - (1+\Gamma_{\text{W}})^{-2}} \geqslant 0. \tag{31}$$

Upon defining

$$\Delta = \left[Q^{-1}(\epsilon_{\text{W}})\right]^2 \log_2^2 e\left[1 - (1+\Gamma_{\text{W}})^{-2}\right]$$
$$+ 4\delta T_{\text{C}}\log_2(1+\Gamma_{\text{W}})R_{\text{W}}(\gamma, L, \epsilon) > 0, \tag{32}$$

the optimal system bandwidth B^* can be finally expressed as

$$B^* = \left[\frac{Q^{-1}(\epsilon_{\text{W}})\log_2 e\sqrt{1 - (1+\Gamma_{\text{W}})^{-2}} + \sqrt{\Delta}}{2\sqrt{\delta T_{\text{C}}}\log_2(1+\Gamma_{\text{W}})}\right]^2, \tag{33}$$

where Γ_{W} can be found in (24). Given the system bandwidth, the total power at the BS is $P_{\text{B}} = P_0 B^*$.

B. Solution to Problem 2

1) Problem Transformation: Problem 2 can be transformed into the following form:

$$\textbf{P2:} \quad \max_{\mathbf{p}} \ \min_{k} \left\{-\sqrt{L_k}\right\} \tag{34a}$$

$$\text{s.t.} \ (29b) - (29c). \tag{34b}$$

2) Property of Objective Function: Recall from (21) that $-\sqrt{L_k}$ can be expressed as

$$\frac{-\sqrt{B^*}Q^{-1}(\epsilon_k)\log_2 e\sqrt{1 - (1+\Gamma_k)^{-2}}}{B^*\log_2(1+\Gamma_k) - R_k(\Gamma_k, L_k, \epsilon_k)} \triangleq \frac{f_k(\mathbf{p})}{g_k(\mathbf{p})}, \forall k. \tag{35}$$

Theorem 2: When the MF or ZF precoder is employed in the downlink of our V2I URLLC system, $-\sqrt{L_k}$ is a quasi-concave function of p_k.

Proof: See Appendix B. ∎

Given the definition of quasi-concavity, $\min_k\left\{-\sqrt{L_k}\right\}$ is also a quasi-concave function. By contrast, even if $-\sqrt{L_k}, \forall k$ is pseudo-concave, $\min_k\left\{-\sqrt{L_k}\right\}$ still cannot be pseudo-concave, since the function $\min(\cdot)$ is not differentiable.

3) Optimal Solution: Since the objective function of (34) is QC, the Karush-Kuhn-Tucker (KKT) condition cannot be adopted. However, we will show that Problem 2 can still be solved optimally by an efficient algorithm.

Based on (35), the following auxiliary function $F(\eta)$ is constructed,

$$F(\eta) = \max_{\mathbf{p}} \min_{k}\left\{f_k(\mathbf{p}) - \eta g_k(\mathbf{p})\right\}. \tag{36}$$

Intuitively, $F(\eta)$ is strictly monotonically decreasing and $F(\eta) = 0$ has a unique root.

Lemma 1: For any $\tilde{\mathbf{p}} \in \mathcal{P}$, we have $F(\eta_{\tilde{\mathbf{p}}}) \geqslant 0$ with $\eta_{\tilde{\mathbf{p}}} = \min_k\{f_k(\tilde{\mathbf{p}})/g_k(\tilde{\mathbf{p}})\}$, and the equality holds when $\tilde{\mathbf{p}} = \arg\max_{\mathbf{p}} \min_k\{f_k(\mathbf{p}) - \eta_{\tilde{\mathbf{p}}}g_k(\mathbf{p})\}$.

Proof: See Appendix C. ∎

Based on Lemma 1, the following theorem describing the optimal allocation is formulated.

Theorem 3: For $\mathbf{p}^* \in \mathcal{P}$ and $\eta^* = \min_k\{f_k(\mathbf{p}^*)/g_k(\mathbf{p}^*)\}$, $\mathbf{p}^*$ is the optimal solution of (34) if and only if

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \min_{k}\left\{f_k(\mathbf{p}) - \eta^* g_k(\mathbf{p})\right\}. \tag{37}$$

Proof: It relies on proving both sufficiency and necessity. Firstly, let $\mathbf{p}^*$ be the optimal power allocation of (34),

$$\eta^* = \min_{k}\left\{\frac{f_k(\mathbf{p}^*)}{g_k(\mathbf{p}^*)}\right\} \geqslant \min_{k}\left\{\frac{f_k(\mathbf{p})}{g_k(\mathbf{p})}\right\}, \forall \mathbf{p} \in \mathcal{P}. \tag{38}$$

Therefore,

$$\min_{k}\left\{f_k(\mathbf{p}) - \eta^* g_k(\mathbf{p})\right\} \leqslant 0, \forall \mathbf{p} \in \mathcal{P}, \tag{39}$$

$$\min_{k}\left\{f_k(\mathbf{p}^*) - \eta^* g_k(\mathbf{p}^*)\right\} = 0, \tag{40}$$

which yields (37) and $F(\eta^*) = 0$.

Conversely, let $\mathbf{p}^* = \arg\max_{\mathbf{p}} \min_k\{f_k(\mathbf{p}) - \eta^* g_k(\mathbf{p})\}$, i.e.,

$$\min_{k}\left\{f_k(\mathbf{p}) - \eta^* g_k(\mathbf{p})\right\} \leqslant \min_{k}\left\{f_k(\mathbf{p}^*) - \eta^* g_k(\mathbf{p}^*)\right\}$$
$$= F(\eta^*) \stackrel{\text{(a)}}{=} 0, \forall \mathbf{p} \in \mathcal{P}, \tag{41}$$

Algorithm 1. Road-traffic-aware twin-timescale resource allocation algorithm for V2I URLLC system

---

Initialization:
- Periodically reported road-traffic density ρ.
- The location information of all VUEs.
- The rate requirement $R_k(\Gamma_k, L_k, \epsilon_k)$ and reliability requirement ϵ_k of each VUE.
- Iterative index $j = 0$, maximum iterative tolerance $\zeta_P > 0$ and initial objective value $\eta_0 < 0$.

Stage 1: Long-term timescale resource allocation

5: **1)** Calculate the optimal system bandwidth B^* based on (33).

    2) Calculate the total power at the BS with $P_B = P_0 B^*$.

Stage 2: Short-term timescale resource allocation

  loop (j)

    3) Based on Algorithm 2, solve the power allocation $\mathbf{p}_j$ of Sub-problem 1 with the given η_j.

    4) Calculate $F_j = \min_k \{f_k(\mathbf{p}_j) - \eta_j g_k(\mathbf{p}_j)\}$.

10:     **if** $F_j \leqslant \zeta_P$ **then**

        Update $\mathbf{p}^* = \mathbf{p}_j$ and $\eta^* = \min_k \left\{ \dfrac{f_k(\mathbf{p}^*)}{g_k(\mathbf{p}^*)} \right\}$.

        Set $j = j + 1$ and **break**.

    else *continue*.

    end if

15:     **5)** Update $\eta_{j+1} = \min_k \{f_k(\mathbf{p}_j)/g_k(\mathbf{p}_j)\}$.

    6) Set $j = j + 1$.

    7) If the current slot is the moment of reporting periodic road-traffic density information, **Then** *go to Stage 1* and **continue**; **Else** *loop Stage 2*.

  end loop

---

where (a) follows from Lemma 1. Then, (41) shows that

$$\eta^* \geqslant \min_k \left\{ \frac{f_k(\mathbf{p})}{g_k(\mathbf{p})} \right\}, \forall \mathbf{p} \in \mathcal{P}, \tag{42}$$

$$\eta^* = \min_k \left\{ \frac{f_k(\mathbf{p}^*)}{g_k(\mathbf{p}^*)} \right\}. \tag{43}$$

Thus $\mathbf{p}^*$ is the optimal power allocation. ∎

Corollary 2: Based on the monotonicity of $F(\eta)$ and Theorem 3, for any η, we have

$$F(\eta) \begin{cases} < 0, & \eta > \eta^*, \\ = 0, & \eta = \eta^*, \\ > 0, & \eta < \eta^*, \end{cases} \tag{44}$$

where η^* is the optimal objective function value of (34).

According to Theorem 3 and Corollary 2, solving Problem 2 is equivalent to finding the unique root of $F(\eta) = 0$.

C. Twin-Timescale Resource Allocation Algorithms for V2I URLLC System

By taking into account the above solutions, the following primary algorithm and secondary algorithm are proposed for the optimal V2I URLLC resource allocation.

1) Primary Algorithm: As shown in [42], there are two classic techniques of solving fractional optimization problems, namely Charnes-Cooper transform and Dinkelbach's method. In simple terms, Dinkelbach's method can be interpreted as being reminiscent of Newton's method applied to the convex function $F(\eta)$ [43]. Compared to Charnes-Cooper transform, Dinkelbach's method has the advantage of requiring no extra constraints [42]. Therefore, we opt for Dinkelbach's method in this paper. Based on the Dinkelbach's method [43], a twin-timescale iterative algorithm is put forward for jointly solving Problem 1 and 2, which is described in Algorithm 1. The following theorem illustrates the convergence and optimality of Algorithm 1.

Theorem 4: Algorithm 1 must converge to the optimal power allocation.

Proof: According to Step 5 in Algorithm 1, we have

$$\begin{aligned} F_j &= \min_k \left\{ f_k(\mathbf{p}_j) - \eta_j g_k(\mathbf{p}_j) \right\} \\ &\overset{(a)}{=} \min_k \left\{ (\eta_{j+1} - \eta_j) g_k(\mathbf{p}_j) \right\} \\ &\overset{(b)}{\geqslant} 0. \end{aligned} \tag{45}$$

(a) follows from

$$\frac{f_k(\mathbf{p}_j)}{g_k(\mathbf{p}_j)} \geqslant \eta_{j+1}, \forall k \Rightarrow f_k(\mathbf{p}_j) \geqslant \eta_{j+1} g_k(\mathbf{p}_j), \tag{46}$$

where the equality holds when $k = \arg\min_k \{f_k(\mathbf{p}_j)/g_k(\mathbf{p}_j)\}$. (b) follows from Lemma 1. This implies that when the algorithm does not achieve convergence, we have $\eta_{j+1} > \eta_j$. Therefore, as the iterations proceed, η is gradually increased and $F(\eta)$ is gradually decreased, hence Algorithm 1 converges.

The optimality can be proved by the method of contradiction. Let $\eta_{\tilde{\mathbf{p}}} = \min_k \{f_k(\tilde{\mathbf{p}})/g_k(\tilde{\mathbf{p}})\} < \eta^*$ be an improved, but sub-optimal value. Given the monotonicity and convergence of $F(\eta)$, we have $F(\eta_{\tilde{\mathbf{p}}}) > F(\eta^*)$ and $F(\eta_{\tilde{\mathbf{p}}}) = 0$. On the other hand, based on Theorem 3, we have $F(\eta^*) = 0$, which leads to a contradiction. ∎

2) Secondary Algorithm: As illustrated in Algorithm 1, Step 3 raises the following optimization problem.

Sub-Problem 1 (Max-Min Power Allocation): Given $\eta_j < 0$, the max-min power allocation problem is formulated as

$$\textbf{SP1:} \quad \max_{\mathbf{p}} \ \min_k \{f_k(\mathbf{p}) - \eta_j g_k(\mathbf{p})\} \tag{47a}$$

$$\text{s. t. } (29b) - (29c). \tag{47b}$$

To solve Sub-problem 1, the monotonicity of $h_k(\mathbf{p}) = f_k(\mathbf{p}) - \eta_j g_k(\mathbf{p})$ is first studied. The proof of the ZF is similar to that of the MF, hence only the case of the MF is shown here. The derivative $\partial h_k^{MF}(\mathbf{p})/\partial p_k = \partial f_k^{MF}(\mathbf{p})/\partial p_k - \eta_j \partial g_k^{MF}(\mathbf{p})/\partial p_k$ is given by (48). Since the order of magnitude for the bandwidth[2] is generally of kHz, we can obtain

$$- Q^{-1}(\epsilon_k) - \eta_j \sqrt{B^*} \sqrt{1 - (1 + \Gamma_k^{MF})^{-2}} \left(1 + \Gamma_k^{MF}\right)^2$$

---

[2]In the operational wireless systems, the MIMO-related techniques are generally used at the narrowband (subcarriers), hence the order of magnitude on basic scheduling resource is of kHz such as the resource block (180 kHz) in the LTE-related systems.

Algorithm 2. Max-min resource allocation algorithm for Sub-problem 1

Initialization:
- The location information of all VUEs.
- The rate requirement $R_k(\Gamma_k, L_k, \epsilon_k)$ and reliability requirement ϵ_k of each VUE.
- Total power P_B and initial power allocation $\mathbf{p}_0 = [p_1^0, \cdots, p_K^0]^T$ with $p_k^0 = P_B/K, \forall k$.
- Iterative index $i = 0$, maximum iterative tolerance $\zeta_S > 0$ and adjustable allocation step length $\mu_0 \in (0, P_B)$.

Power allocation for Sub-problem 1:

5: **1)** Calculate initial $\mathbf{H}_0 = [h_1^0(\mathbf{p}_0), \cdots, h_K^0(\mathbf{p}_0)]^T$ with $h_k^0(\mathbf{p}_0) = f_k(\mathbf{p}_0) - \eta_j g_k(\mathbf{p}_0)$.

loop (i)

    2) Set $i = i + 1$.

    3) Calculate $k_i^m = \arg\min_k \{\mathbf{H}_{i-1}\}$ and $k_i^M = \arg\max_k \{\mathbf{H}_{i-1}\}$.

    4) Update $\mathbf{p}_i$ with $p_{k_i^m}^i = p_{k_i^m}^{i-1} + \mu_{i-1}$, $p_{k_i^M}^i = p_{k_i^M}^{i-1} - \mu_{i-1}$ and $p_k^i = p_k^{i-1}(\forall k \neq k_i^m, k_i^M)$.

10:     **5)** Calculate $\mathbf{H}_i$ with $h_k^i(\mathbf{p}_i) = f_k(\mathbf{p}_i) - \eta_j g_k(\mathbf{p}_i)$.

    if $\max_k \{\mathbf{H}_i\} - \min_k \{\mathbf{H}_i\} \leqslant \zeta_S$ **then**

        Update $\mathbf{p}_j = \mathbf{p}_i$ and ***break***.

    else

        if $h_{k_i^m}^i(\mathbf{p}_i) \leqslant h_{k_i^m}^{i-1}(\mathbf{p}_{i-1})$ or $h_{k_i^M}^i(\mathbf{p}_i) \geqslant h_{k_i^M}^{i-1}(\mathbf{p}_{i-1})$ or $p_{k_i^M}^i \leqslant 0$ or $\max_k \{\mathbf{H}_{i-1}\} - \min_k \{\mathbf{H}_{i-1}\} \leqslant \max_k \{\mathbf{H}_i\} - \min_k \{\mathbf{H}_i\}$ **then**

15:             Update $\mu_i = \mu_{i-1}/2$, $\mathbf{p}_i = \mathbf{p}_{i-1}$ and $\mathbf{H}_i = \mathbf{H}_{i-1}$.

        else

            Update $\mu_i = \mu_{i-1}$ and ***continue***.

        end if

    end if

20: **end loop**

$$\geqslant -Q^{-1}(\epsilon_k) - \eta_j \sqrt{B^*} \frac{\Gamma_k^{MF}}{1 + \Gamma_k^{MF}}$$
$$= \frac{-Q^{-1}(\epsilon_k)\left(1 + \Gamma_k^{MF}\right) - \eta_j \sqrt{B^*}\Gamma_k^{MF}}{1 + \Gamma_k^{MF}} > 0, \quad (49)$$

which implies that the function $h_k^{MF}(\mathbf{p}) = f_k^{MF}(\mathbf{p}) - \eta_j g_k^{MF}(\mathbf{p})$ is increasing. The theorem capable of achieving the optimal max-min power allocation is presented as follows [33].

Theorem 5: In order to maximize the minimum objective function value of (47), all VUEs should have the same objective value, hence we have

$$h_1(\mathbf{p}) = h_2(\mathbf{p}) = \cdots = h_K(\mathbf{p}) \triangleq H_{\mathbf{p}}. \quad (50)$$

Proof: (50) can be obtained by the method of contradiction. Let $\mathbf{p}^* = [p_1^*, \cdots, p_K^*]^T$ be the optimal power allocation of Sub-problem 1, which does not satisfy (50). Without loss of generality, let $h_1(p_1^*) = \max_k \{h_k(p_k^*)\}$ and $h_K(p_K^*) = \min_k \{h_k(p_k^*)\}$. As $h_k(\mathbf{p})$ is monotonically increasing, we should have $\mu \in (0, p_1^*)$ satisfying

$$h_1(p_1^*) > h_1(p_1^* - \mu) = h_K(p_K^* + \mu) > h_K(p_K^*). \quad (51)$$

Hence, a new power allocation $\mathbf{p}_N^* = [p_1^* - \mu, \cdots, p_K^* + \mu]^T$ satisfying (47b) is constructed. In other words, the minimum objective function value has been increased, which contradicts to the max-min criterion. Hence the theorem is proved. ∎

Based on Theorem 5, an iterative procedure described by Algorithm 2 is proposed for solving Sub-problem 1. According to $\lim_{i \to \infty}(|h_{k_i^m}^i(\mathbf{p}_i) - h_{k_i^M}^i(\mathbf{p}_i)|) = 0$, we can intuitively find that Algorithm 2 is capable of converging to the same objective function value $H_{\mathbf{p}}$. Meanwhile, the optimality can be guaranteed by Theorem 5. In the initialization phase of Algorithm 2, the objective values of all VUEs are calculated for the equal power allocation. As the iterations proceed, Algorithm 2 will accordingly increase the minimum power and reduce the maximum power every time until we have $\max_k \{\mathbf{H}_i\} - \min_k \{\mathbf{H}_i\} \leqslant \zeta_S$. To avoid the ping-pong phenomenon between k_i^m and k_i^M, four judgement conditions are listed in Algorithm 2 to adjust the allocation step length μ_i.

3) Algorithm Complexity: First of all, the optimal solution of Problem 1 is directly achieved by the quadratic formula. Furthermore, since this optimal solution is independent of the number of VUEs and it is a closed-form solution, the complexity of solving Problem 1 can be ignored. With respect to solving Problem 2, the convergence and optimality are guaranteed by Theorem 4 and 5, while the complexity is determined by Stage 2 of Algorithm 1, and Algorithm 2. Based on the maximum iterative tolerance ζ_P of Algorithm 1, the associated precision is given by $Z = \log_{10}(\zeta_P^{-1})$-digit. Thus, the complexity order of Algorithm 1 is $\mathcal{O}(\log_2 Z)$ [44], [43]. On the other hand, because $|h_k(p_k) - H_{\mathbf{p}}| \leqslant \frac{\zeta_S}{2}$ is a sufficient condition for satisfying the termination condition $\max_k \{\mathbf{H}_i\} - \min_k \{\mathbf{H}_i\} \leqslant \zeta_S$, the complexity of Algorithm 2 is on the order of $\mathcal{O}(\sum_{k=1}^K \log_2[h_k^{-1}(H_{\mathbf{p}} + \frac{\zeta_S}{2}) - h_k^{-1}(H_{\mathbf{p}} - \frac{\zeta_S}{2})]^{-1})$ [24], [33]. Hence, the total complexity order is $\mathcal{O}(\log_2 Z \sum_{k=1}^K \log_2[h_k^{-1}(H_{\mathbf{p}} + \frac{\zeta_S}{2}) - h_k^{-1}(H_{\mathbf{p}} - \frac{\zeta_S}{2})]^{-1})$ [45].

Finally, due to the handover in high mobility VNETs, we discuss the scalability of the proposed algorithms.

Remark 1 (Extension of Algorithms): The proposed algorithms depend on the road-traffic density and the location information of all VUEs. In general, the traffic control center can use roadside monitoring sensors or real-world traffic datasets to predict and periodically report the vehicular density and location information [29], [30]. For multi-cell scenarios, both of the density and location information can be shared via wired links between the traffic control center and all the base stations. Therefore, as long as the density and location information are available, our proposed algorithms can be readily extended to the multi-cell scenarios. In addition, due to the wired links, the sharing of the density and location information do not bring the extra signal overhead for radio access networks.

VI. SIMULATION RESULTS

A. Simulation Setup

In the simulations, the large-scale fading of the k-th VUE is given by $\beta_k = \theta[(d_k - d_R/2)^2 + d_B^2]^{-\alpha/2}$, where θ is a constant related to the antenna gain and carrier frequency, α is the path

$$\frac{\partial h_k^{\text{MF}}(\mathbf{p})}{\partial p_k} = \frac{\sqrt{B^*}\left(\frac{\Gamma_k^{\text{MF}}}{p_k} + \frac{(\Gamma_k^{\text{MF}})^2}{Mp_k}\right)}{\sqrt{1 - (1 + \Gamma_k^{\text{MF}})^{-2}}\,(1 + \Gamma_k^{\text{MF}})^3 \ln 2}\left[-Q^{-1}(\epsilon_k) - \eta_j\sqrt{B^*}\sqrt{1 - (1 + \Gamma_k^{\text{MF}})^{-2}}\,(1 + \Gamma_k^{\text{MF}})^2\right]. \qquad (48)$$

TABLE I
BASIC SIMULATION PARAMETERS

Parameter	Value
BS distance d_{B}	20 m
Road length d_{R}	200 m
Maximum road-traffic density ρ_{m}	0.15
Free flow speed v_{F}	80 km/h
Estimation accuracy $\chi_k, \forall k$	0.8
Carrier frequency f_{C}	2 GHz
Constant θ	10^{-3}
Path loss exponent α	3.8
Noise power spectral density N_0	-130 dBm/Hz
Transmitted signal power spectral density P_0	-10 dBm/Hz

loss exponent, and d_k is the distance between the k-th VUE and the starting point of the road. The noise power is calculated by $\sigma^2 = N_0 B$, where N_0 and B are the noise power spectral density and system bandwidth, respectively. Furthermore, all VUEs are uniformly distributed on the one-way road, namely $d_k \sim \text{unif}(0, d_{\text{R}})$. Since the average length of a vehicle is 6.5 m (already plus 50% of additional safety distance), the maximum road-traffic density is set to $\rho_{\text{m}} = 0.15$. In the following simulation results, PCSI and IPCSI represent the perfect and imperfect channel state information, respectively. Finally, all other basic simulation parameters are listed in Table I.

B. Simulation Results for the Maximum Achievable Rate and Transmission Latency

1) Tradeoff: Upon using the equal power allocation (EPA) and MF precoder, Fig. 3 illustrates the tradeoff of a VUE among the maximum achievable rate, transmission latency and reliability for the case of IPCSI. As shown in Fig. 3, regardless of what the latency and reliability are, the ergodic capacity remains a constant. By contrast, Theorem 1 indicates that some rate regions are not achievable (negative) because of the ultra-high reliability and ultra-low latency. Moreover, the maximum achievable rate given by Theorem 1 is lower than the ergodic capacity, which confirms that V2I URLLCs are indeed possible, but only at the cost of a reduced rate. Finally, Fig. 3 shows that the transmission latency can only be optimized within a reasonable range. The results of the ZF are similar to those of the MF, thus they are omitted here due to space limitations.

2) Tightness: Upon adopting the EPA schemes, Fig. 4 and 5 illustrate the maximum achievable rate and transmission latency versus the number of antennas, respectively. The reliability threshold ϵ_k of these two figures is set to 10^{-6}. The asymptotic results ($M \to \infty$) are calculated from (18) and (21) with $\Gamma_k^{\infty} = p_k \chi_k \beta_k / \sigma^2$ ($\chi_k = 1$ for PCSI). For the PCSI and IPCSI, Fig. 4 and 5 indicate that upon increasing the
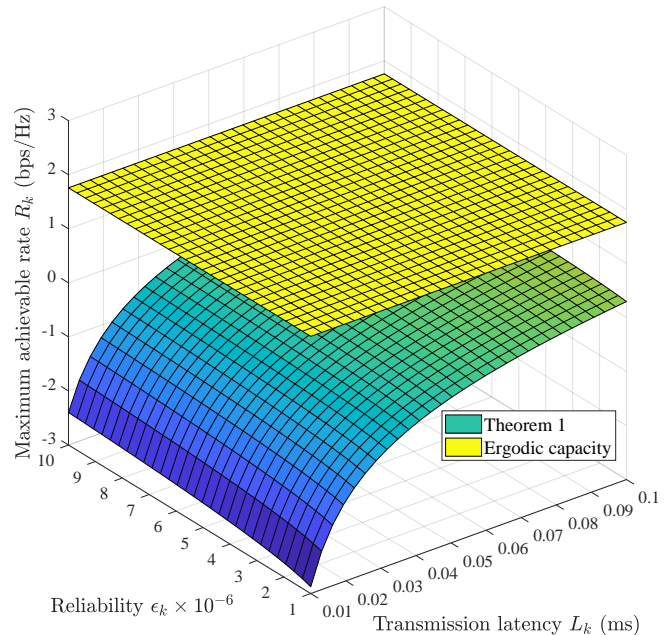

Fig. 3. Tradeoff among maximum achievable rate, transmission latency and reliability in (1) and (18) based on the MF precoder, with $\rho = 0.05$, $M = 300$, $B = 200$ kHz and $P_{\text{B}} = 10$ dBW.
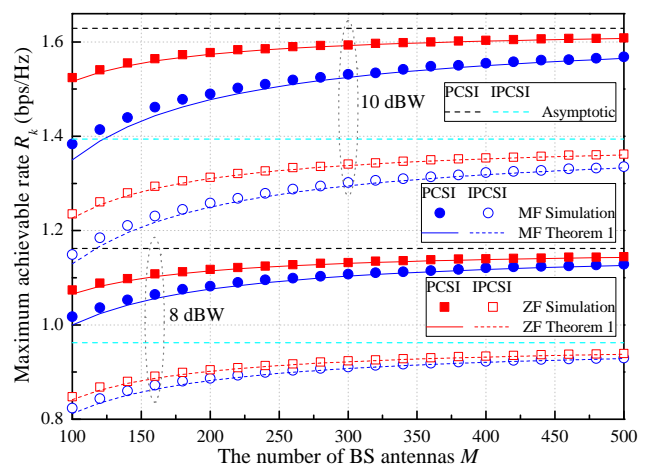

Fig. 4. Maximum achievable rate versus the number of antennas in (18), with $\rho = 0.05$, $B = 200$ kHz and $L_k = 1$ ms.

number of antennas, the maximum achievable rate increases, while the latency of a VUE decreases. The approximate results of Theorem 1 and Corollary 1 are close to the simulation results, and both of them tend to the asymptotic results as the number of antennas increases. Because of the channel estimation error, the case of IPCSI is always worse than the case of PCSI. Furthermore, both the rate and latency of the ZF are better than those of the MF in the case of PCSI or IPCSI, due to the higher SINR Γ_k. Finally, these two figures suggest
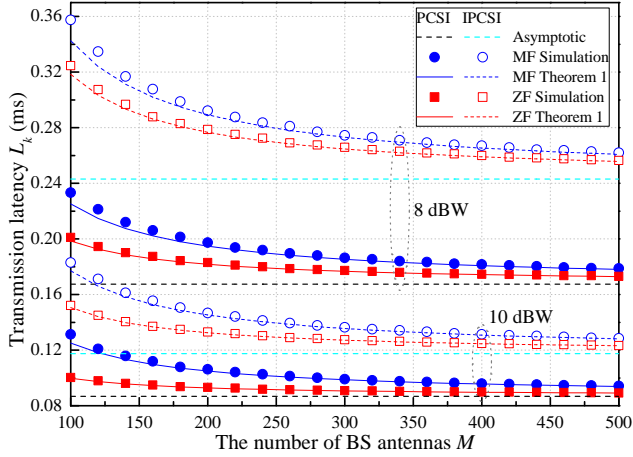
Fig. 5. Transmission latency versus the number of antennas in (21), with $\rho = 0.05$, $B = 200$ kHz and $R_k = 100$ kbps.
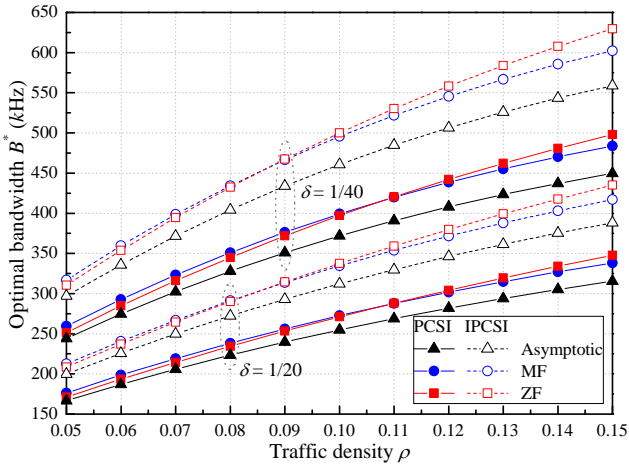

Fig. 7. Optimal system bandwidth versus minimum reliability in (33), with $M = 300$, $R_W = 100$ kbps and $\delta = 1/20$.

Fig. 6. Optimal system bandwidth versus road-traffic density in (33), with $M = 300$, $R_W = 100$ kbps and $\epsilon_W = 10^{-6}$.

of antennas tends to infinity, the bandwidth is only reduced by at most 75 kHz (IPCSI at $\delta = 1/40$ with $\rho = 0.15$), which implies that the V2I URLLC system does not require an excessive numbers of antennas.

Fig. 7 illustrates the optimal system bandwidth versus the reliability for the two road-traffic densities considered. As shown in Fig. 7, upon relaxing the minimum reliability requirement, the optimal system bandwidth decreases, regardless of the CSI. The gap between the MF results and ZF results is in line with the trend seen in Fig. 6 under these two types of road-traffic densities. Additionally, by comparing the results of $\epsilon_W = 10^{-6}$ to those of 10^{-9}, the optimal bandwidth is only increased by at most 75 kHz (IPCSI at $\rho = 0.1$), which suggests that the reliability can be significantly improved by only modestly increasing the bandwidth.

D. Simulation Results for Problem 2

1) Convergence: Based on a specific channel realization in the simulations, Fig. 8(a) and 8(b) illustrate the convergence of the two proposed algorithms, respectively. Fig. 8(a) indicates that as the number of iterations increases, the auxiliary function $F(\eta)$ converges to zero in Algorithm 1, as stated by Theorem 4. On the other hand, as shown in Fig. 8(b), as the number of iterations increases, the maximum objective function value decreases and the minimum objective function value increases. All objective function values finally converge to a constant, which characterizes the convergence of Algorithm 2.

2) Transmission Latency: Fig. 9 illustrates the maximum transmission latency among all VUEs versus the road-traffic density. Observe that when the road-traffic density increases, the maximum transmission latency of both the proposed scheme and of the EPA scheme increase accordingly. Upon comparing the MF precoder to the ZF precoder, regardless of having either PCSI or IPCSI, the performance of the ZF is better than that of the MF both for the proposed scheme and for the EPA scheme. Fig. 9 also indicates that the proposed scheme performs better than the EPA scheme both for the MF precoder and for the ZF precoder in the case of either PCSI or

that increasing the total transmitted power improve both the rate and latency. Based on the tightness of Corollary 1, the simulation results of Problem 1 and 2 will be discussed in the following subsections.

C. Simulation Results for Problem 1

Based on the PCSI and IPCSI, Fig. 6 illustrates the optimal system bandwidth versus the road-traffic density, where the MF, ZF and asymptotic results are calculated by substituting (24) and $\Gamma_W^\infty = P_0 \chi_{th} \beta_W / (N_0 \rho d_R)$ into (33). In order to keep the worst-case latency below a certain threshold, Fig. 6 shows that with the road-traffic density increasing, the optimal system bandwidth also increases. This means that the more VUEs are supported, the more bandwidth is required. In the region of low road-traffic density, the optimal bandwidth of the ZF is lower than that of the MF, whereas in the region of high road-traffic density, the required bandwidth of the ZF is higher than that of the MF due to the lower SINR. Moreover, the increase of the threshold δ requires more bandwidth. Compared to the case of PCSI, the case of IPCSI requires more bandwidth to satisfy the worst-case latency requirement. Finally, when the number
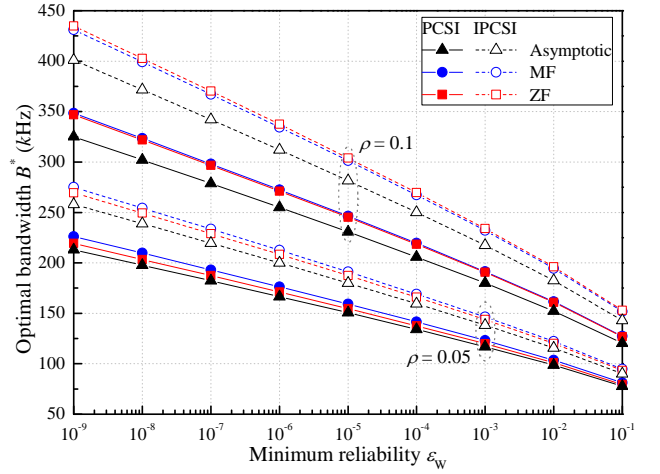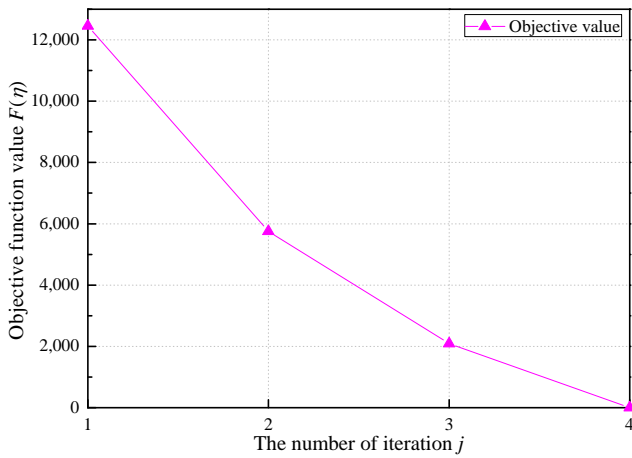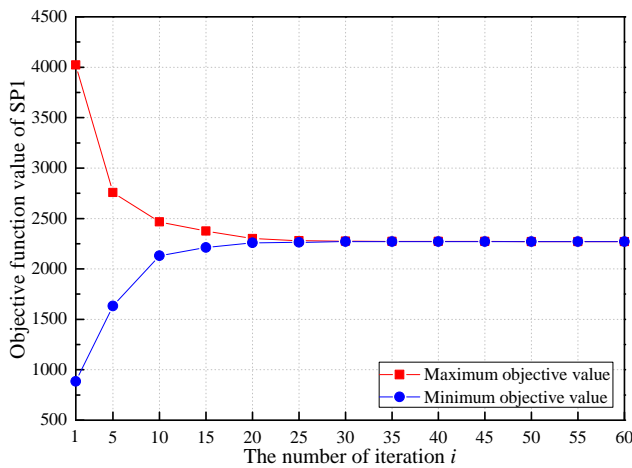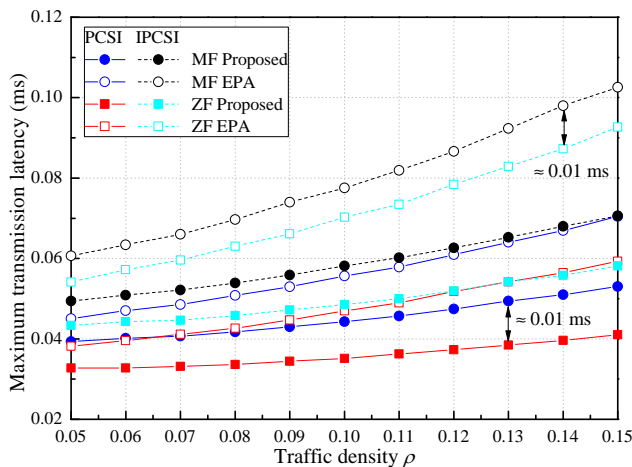
(a) Convergence of Algorithm 1 in (44) and Theorem 4.

(b) Convergence of Algorithm 2 in Theorem 5.

Fig. 8. Convergence of the two proposed algorithms with $\zeta_P = \zeta_S = 10^{-2}$, $\eta_0 = -3 \times 10^{-2}$ and $\mu_0 = P_B/(2\rho d_R)$.

Fig. 9. Maximum transmission latency versus road-traffic density solved by Algorithm 1 and 2, with $M = 300$, $\delta = 1/20$, $R_k = 100$ kbps and $\epsilon_k = 10^{-6}$ for all VUEs.

IPCSI. Furthermore, the performance gap between the MF and ZF is always about 0.01 ms. Naturally, the channel estimation error increases latency. Nevertheless, the IPCSI results of the
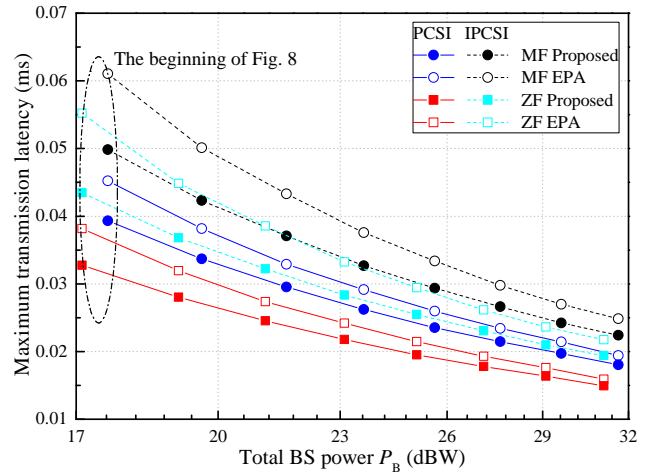
Fig. 10. Maximum transmission latency versus total transmitted power solved by Algorithm 1 and 2, with $M = 300$, $\delta = 1/20$, $\rho = 0.05$, $R_k = 100$ kbps and $\epsilon_k = 10^{-6}$ for all VUEs.

proposed scheme are close to the PCSI results of the EPA scheme both in the case of the MF and the ZF precoder. Finally, the slope of the proposed scheme is lower for both the PCSI and IPCSI, which implies that the proposed scheme is not sensitive to the road-traffic density. Hence, the proposed algorithms can get the effect of semi-persistent scheduling for the V2I URLLC system.

Given $\rho = 0.05$ and increase the total power based on $P_0 B^*$, Fig. 10 quantifies the maximum transmission latency among all VUEs versus the total transmitted power. As shown in Fig. 10, the maximum transmission latency decreases for both the proposed scheme and for the EPA scheme, as the total power increases. The performance of the ZF is better than that of the MF both for the proposed scheme and for the EPA scheme in the case of either PCSI or IPCSI. Additionally, the gap between the ZF relying on IPCSI and the MF using PCSI is relatively small for the proposed scheme. Hence, the ZF precoder is the better choice for V2I URLLCs, as evidenced both by Fig. 9 and Fig. 10.

VII. CONCLUSIONS

In order to reduce the signaling overhead, this paper studied the problem of minimizing the transmission latency in the downlink of the V2I URLLC system based on the road-traffic density and large-scale fading CSI. The expression of transmission latency was first derived for the MF precoder and ZF precoder based on both perfect and imperfect CSI. Then, a two-stage radio resource allocation problem was formulated relying on our twin-timescale perspective. More particularly, the first stage optimized the worst-case transmission latency by adjusting the system's bandwidth based on a long-term timescale, while the second stage minimized the maximum transmission latency among all VUEs by judiciously allocating the total power on a short-term timescale basis. For optimally solving our two-stage problem formulated, a moderate-complexity twin-timescale radio resource allocation algorithm was conceived. Our simulation results showed that the proposed resource allocation scheme significantly reduced the

maximum transmission latency and that the ZF precoder was the better choice for the V2I URLLC system considered.

APPENDIX A
PROOF OF THEOREM 1

Recall from (3) that the ergodic rate of the k-th VUE is given by

$$R_k(\gamma_k, L_k, \epsilon_k) \approx \mathbb{E}\left\{ B\left[\log_2(1+\gamma_k) - \sqrt{\frac{V_k}{L_k B}} Q^{-1}(\epsilon_k) \right] \right\}$$

$$= \mathbb{E}\left[B\log_2\left(1 + \frac{1}{\gamma_k^{-1}}\right) \right] - \mathbb{E}\left[\sqrt{\frac{BV_k}{L_k}} Q^{-1}(\epsilon_k) \right]. \quad (52)$$

Let $u_k = \log_2(1 + 1/\gamma_k^{-1})$ denote the spectral efficiency of the k-th VUE, then V_k can be rewritten as

$$V_k = \left(1 - \frac{1}{(1+\gamma_k)^2}\right)(\log_2 e)^2 = \left(1 - 2^{-2u_k}\right)(\log_2 e)^2, \quad (53)$$

and we have

$$R_k(\gamma_k, L_k, \epsilon_k) \approx B\mathbb{E}\left[\log_2\left(1 + \frac{1}{\gamma_k^{-1}}\right) \right]$$
$$- \log_2 e \sqrt{\frac{B}{L_k}} Q^{-1}(\epsilon_k)\mathbb{E}\left[\sqrt{1 - 2^{-2u_k}} \right]. \quad (54)$$

Since the function $a(x) = \log_2(1+1/x), \forall x > 0$ is convex, Jensen's inequality gives

$$U_k \triangleq \mathbb{E}[u_k] \geqslant U_k^L \triangleq \log_2\left(1 + \frac{1}{\mathbb{E}[\gamma_k^{-1}]}\right). \quad (55)$$

Substituting $\mathbb{E}\left[\Omega_B^{MF}\right] = \frac{1}{M}$, $\mathbb{E}\left[\Omega_G^{MF}\right] = \frac{1}{M-1}$ and $\mathbb{E}\left[\Omega_G^{ZF}\right] = \frac{1}{M-K}$ into (17), we obtain

$$U_k^L = \begin{cases} \log_2\left(1 + \dfrac{Mp_k}{P_B - p_k + \varphi_k}\right), & \text{for MF}, \\ \log_2\left(1 + p_k\varphi_k\right), & \text{for ZF}, \end{cases} \quad (56)$$

where

$$\varphi_k = \begin{cases} \dfrac{P_B\beta_k(1-\chi_k) + M\sigma^2}{\chi_k\beta_k} \dfrac{M}{M-1}, & \text{for MF}, \\ \dfrac{\chi_k\beta_k(M-K)}{P_B\beta_k(1-\chi_k) + M\sigma^2}, & \text{for ZF}. \end{cases} \quad (57)$$

Note that for $M \to \infty$, we have $U_k^L = \log_2(1 + \frac{p_k\chi_k\beta_k}{\sigma^2})$. On the other hand, due to the channel hardening phenomenon, we arrive at $\lim_{M\to\infty} U_k = \log_2(1 + \frac{p_k\chi_k\beta_k}{\sigma^2})$, which indicates that U_k^L is tight for U_k when M is large. Therefore, U_k^L can be used for approximating the first term of (54), i.e.,

$$\mathbb{E}\left[\log_2\left(1 + \frac{1}{\gamma_k^{-1}}\right) \right] \approx U_k^L, M \gg 1. \quad (58)$$

Similarly, since the function $b(x) = \sqrt{1 - 2^{-2x}}, \forall x > 0$ is concave, with Jensen's inequality,

$$\mathbb{E}\left[\sqrt{1 - 2^{-2u_k}} \right] \leqslant \sqrt{1 - 2^{-2\mathbb{E}[u_k]}}. \quad (59)$$

Based on the channel hardening phenomenon, we arrive at

$$\lim_{M\to\infty} \mathbb{E}\left[\sqrt{1 - 2^{-2u_k}} \right] = \sqrt{1 - \left(1 + \frac{p_k\chi_k\beta_k}{\sigma^2}\right)^{-2}}$$
$$= \lim_{M\to\infty} \sqrt{1 - 2^{-2\mathbb{E}[u_k]}}, \quad (60)$$

which implies that $\sqrt{1 - 2^{-2\mathbb{E}[u_k]}}$ is also tight for $\mathbb{E}\left[\sqrt{1 - 2^{-2u_k}}\right]$ when M is large, namely $\mathbb{E}\left[\sqrt{1 - 2^{-2u_k}}\right] \approx \sqrt{1 - 2^{-2\mathbb{E}[u_k]}}, M \gg 1$.

Recalling that $U_k \approx U_k^L$, we finally have

$$\mathbb{E}\left[\sqrt{1 - 2^{-2u_k}} \right] \approx \sqrt{1 - 2^{-2U_k^L}}, M \gg 1. \quad (61)$$

Substituting (58) and (61) into (54) yields (18).

APPENDIX B
PROOF OF THEOREM 2

We prove it for the case of the MF, stating that the proof of the ZF is similar to that of the MF, hence it is omitted here. For the MF precoder, we have

$$f_k^{MF}(\mathbf{p}) = -\sqrt{B^*} Q^{-1}(\epsilon_k) \log_2 e \sqrt{1 - (1 + \Gamma_k^{MF})^{-2}}, \quad (62)$$

$$\frac{\partial f_k^{MF}(\mathbf{p})}{\partial p_k} = -\frac{\sqrt{B^*} Q^{-1}(\epsilon_k) \log_2 e \left(\dfrac{\Gamma_k^{MF}}{p_k} + \dfrac{(\Gamma_k^{MF})^2}{Mp_k}\right)}{\sqrt{1 - (1 + \Gamma_k^{MF})^{-2}}(1 + \Gamma_k^{MF})^3}, \quad (63)$$

Since $M \gg 1 \Rightarrow 3M - 2 > 0$, then $f_k^{MF}(\mathbf{p}) < 0$, $\partial f_k^{MF}(\mathbf{p})/\partial p_k < 0$ and $\partial^2 f_k^{MF}(\mathbf{p})/\partial p_k^2 > 0$, namely $f_k^{MF}(\mathbf{p})$ is negative, differentiable, decreasing and convex.

On the other hand, we have

$$g_k^{MF}(\mathbf{p}) = B^* \log_2\left(1 + \Gamma_k^{MF}\right) - R_k(\Gamma_k, L_k, \epsilon_k), \quad (65)$$

$$\frac{\partial g_k^{MF}(\mathbf{p})}{\partial p_k} = \frac{B^*\left(\dfrac{\Gamma_k^{MF}}{p_k} + \dfrac{(\Gamma_k^{MF})^2}{Mp_k}\right)}{(1 + \Gamma_k^{MF})\ln 2}, \quad (66)$$

$$\frac{\partial^2 g_k^{MF}(\mathbf{p})}{\partial p_k^2} = \frac{B^*\left(M + \Gamma_k^{MF}\right)\left(2 + \Gamma_k^{MF} - M\right)}{(P_B - p_k + \varphi_k)^2(1 + \Gamma_k^{MF})^2 \ln 2}. \quad (67)$$

Since $M \gg 1 \Rightarrow 2 + \Gamma_k^{MF} - M < 0$, then $g_k^{MF}(\mathbf{p}) > 0$, $\partial g_k^{MF}(\mathbf{p})/\partial p_k > 0$ and $\partial^2 g_k^{MF}(\mathbf{p})/\partial p_k^2 < 0$, namely $g_k^{MF}(\mathbf{p})$ is non-negative, differentiable, increasing and concave.

Lemma 2: Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex set and $r : \mathcal{C} \to \mathbb{R}$. Then, r is QC if and only if its superlevel set $\mathcal{S}_t = \{\mathbf{x} \in \mathcal{C} : r(\mathbf{x}) \geqslant t\}$ is convex for all $t \in \mathbb{R}$.

Proof: The omitted proof can be found in [46], [43]. □

According to Lemma 2, the superlevel set $\mathcal{S}_t = \{\mathbf{p} \in \mathcal{C} : f_k^{MF}(\mathbf{p})/g_k^{MF}(\mathbf{p}) \geqslant t\}$ is taken into account. $\mathcal{S}_t$ is the empty set for all $t > 0$ when $f_k^{MF}(\mathbf{p}) < 0$ and $g_k^{MF}(\mathbf{p}) > 0$, thus only $t \leqslant 0$ has to be considered. The equivalent form of $\mathcal{S}_t$ is given by $\mathcal{S}_t = \{\mathbf{p} \in \mathcal{C} : f_k^{MF}(\mathbf{p}) - tg_k^{MF}(\mathbf{p}) \geqslant 0\}$. Let $h_k^{MF}(\mathbf{p}) = f_k^{MF}(\mathbf{p}) - tg_k^{MF}(\mathbf{p})$, then $\partial^2 h_k^{MF}(\mathbf{p})/\partial p_k^2 = \partial^2 f_k^{MF}(\mathbf{p})/\partial p_k^2 - t\partial^2 g_k^{MF}(\mathbf{p})/\partial p_k^2$ is given by (68). Because

$$\frac{\partial^2 f_k^{\text{MF}}(\mathbf{p})}{\partial p_k^2} = \frac{\sqrt{B^*}Q^{-1}(\epsilon_k)\log_2 \mathrm{e}\left(M+\Gamma_k^{\text{MF}}\right)\left(3M-2+\Gamma_k^{\text{MF}}\right)}{\left(P_{\text{B}}-p_k+\varphi_k\right)^2\sqrt{1-\left(1+\Gamma_k^{\text{MF}}\right)^{-2}\left(1+\Gamma_k^{\text{MF}}\right)^4}}. \tag{64}$$

$$\frac{\partial^2 h_k^{\text{MF}}(\mathbf{p})}{\partial p_k^2} = -\frac{\sqrt{B^*}\left(M+\Gamma_k^{\text{MF}}\right)}{\left(P_{\text{B}}-p_k+\varphi_k\right)^2\sqrt{1-\left(1+\Gamma_k^{\text{MF}}\right)^{-2}\left(1+\Gamma_k^{\text{MF}}\right)^4}\ln 2}$$
$$\times\left[Q^{-1}(\epsilon_k)\left(-3M+2-\Gamma_k^{\text{MF}}\right)-t\sqrt{B^*}\sqrt{1-\left(1+\Gamma_k^{\text{MF}}\right)^{-2}}\left(1+\Gamma_k^{\text{MF}}\right)^2\left(M-2-\Gamma_k^{\text{MF}}\right)\right]. \tag{68}$$

$M \gg 1$ and the order of magnitude on bandwidth is generally of kHz, we obtain

$$Q^{-1}(\epsilon_k)\left(-3M+2-\Gamma_k^{\text{MF}}\right)$$
$$-t\sqrt{B^*}\sqrt{1-\left(1+\Gamma_k^{\text{MF}}\right)^{-2}}\left(1+\Gamma_k^{\text{MF}}\right)^2\left(M-2-\Gamma_k^{\text{MF}}\right)$$
$$\geqslant Q^{-1}(\epsilon_k)\left(-3M+2-\Gamma_k^{\text{MF}}\right)$$
$$-t\sqrt{B^*}\frac{\Gamma_k^{\text{MF}}}{1+\Gamma_k^{\text{MF}}}\left(M-2-\Gamma_k^{\text{MF}}\right) > 0, \tag{69}$$

which states that the function $h_k^{\text{MF}}(\mathbf{p}) = f_k^{\text{MF}}(\mathbf{p}) - tg_k^{\text{MF}}(\mathbf{p})$ is concave, namely the superlevel set $\mathcal{S}_t$ is convex, and $-\sqrt{L_k}$ is a QC function of p_k.

APPENDIX C
PROOF OF LEMMA 1

Recall from (36) that

$$F(\eta_{\tilde{\mathbf{p}}}) = \max_{\mathbf{p}}\min_k\left\{f_k(\mathbf{p})-\min_k\left\{\frac{f_k(\tilde{\mathbf{p}})}{g_k(\tilde{\mathbf{p}})}\right\}g_k(\mathbf{p})\right\}, \forall\mathbf{p}\in\mathcal{P}$$
$$\overset{(a)}{\geqslant}\min_k\left\{f_k(\tilde{\mathbf{p}})-\min_k\left\{\frac{f_k(\tilde{\mathbf{p}})}{g_k(\tilde{\mathbf{p}})}\right\}g_k(\tilde{\mathbf{p}})\right\}$$
$$\overset{(b)}{=}f_n(\tilde{\mathbf{p}})-\frac{f_n(\tilde{\mathbf{p}})}{g_n(\tilde{\mathbf{p}})}g_n(\tilde{\mathbf{p}})=0, \tag{70}$$

where the equality of (a) holds when $\tilde{\mathbf{p}} = \arg\max_{\mathbf{p}}\min_k\left\{f_k(\mathbf{p})-\eta_{\tilde{\mathbf{p}}}g_k(\mathbf{p})\right\}$, and (b) follows from

$$\frac{f_n(\tilde{\mathbf{p}})}{g_n(\tilde{\mathbf{p}})} < \frac{f_k(\tilde{\mathbf{p}})}{g_k(\tilde{\mathbf{p}})}, \forall k\neq n, n=\arg\min_k\left\{\frac{f_k(\tilde{\mathbf{p}})}{g_k(\tilde{\mathbf{p}})}\right\}, \tag{71}$$

$$\Rightarrow f_k(\tilde{\mathbf{p}})-\frac{f_n(\tilde{\mathbf{p}})}{g_n(\tilde{\mathbf{p}})}g_k(\tilde{\mathbf{p}}) > 0. \tag{72}$$

REFERENCES

[1] T. Qiu, N. Chen, K. Li, M. Atiquzzaman, and W. Zhao, "How can heterogeneous Internet of things build our future: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2011–2027, Thirdquarter 2018.

[2] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2377–2396, Fourthquarter 2015.

[3] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1858–1877, Thirdquarter 2018.

[4] K. Zheng, Q. Zheng, H. Yang, L. Zhao, L. Hou, and P. Chatzimisios, "Reliable and efficient autonomous driving: The need for heterogeneous vehicular networks," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 72–79, Dec. 2015.

[5] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1243–1274, Secondquarter 2019.

[6] W. Sun, E. G. Ström, F. Brännström, K. C. Sou, and Y. Sui, "Radio resource management for D2D-based V2V communication," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6636–6650, Aug. 2016.

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

[8] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[9] 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, Sep. 2016.

[10] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[11] Q. Zheng, K. Zheng, H. Zhang, and V. C. M. Leung, "Delay-optimal virtualized radio resource scheduling in software-defined vehicular networks via stochastic learning," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7857–7867, Oct. 2016.

[12] K. Zheng, H. Meng, P. Chatzimisios, L. Lei, and X. Shen, "An SMDP-based resource allocation in vehicular cloud computing systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7920–7928, Dec. 2015.

[13] G. Yang, M. Xiao, and H. V. Poor, "Low-latency millimeter-wave communications: Traffic dispersion or network densification?" *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3526–3539, Aug. 2018.

[14] H. Forssell, R. Thobaben, H. Al-Zubaidy, and J. Gross, "Physical layer authentication in mission-critical mtc networks: A security and delay performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 795–808, Apr. 2019.

[15] J. Choi, "An effective capacity based approach to multi-channel low-latency wireless communications," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2476–2486, Mar. 2019.

[16] H. Yang, L. Zhao, L. Lei, and K. Zheng, "A two-stage allocation scheme for delay-sensitive services in dense vehicular networks," in *Proc. IEEE International Conference on Communications Workshops (ICC Workshops)*, Paris, France, May 2017, pp. 1358–1363.

[17] S. Sebastian, G. James, and A.-Z. Hussein, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, Cancun, Mexico, Nov. 2015, pp. 13–22.

[18] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[19] Y. Cui, V. K. N. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems–large deviation theory, stochastic Lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1677–1701, Mar. 2012.

[20] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.

[21] N. A. Johansson, Y. . E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. IEEE International Conference on Communications Workshops (ICC Workshops)*, London, UK, Jun. 2015, pp. 1184–1189.

[22] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[23] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang, and L. Hanzo, "Survey of large-scale MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1738–1760, Thirdquarter 2015.

[24] L. Zhao and X. Wang, "Massive MIMO downlink for wireless information and energy transfer with energy harvesting receivers," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3309–3322, May 2019.

[25] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[26] Y. Hu, Y. Zhu, M. C. Gursoy, and A. Schmeink, "SWIPT-enabled relaying in IoT networks operating with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 74–88, Jan. 2019.

[27] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.

[28] L. Zhou, A. Wolf, and M. Motani, "On lossy multi-connectivity: Finite blocklength performance and second-order asymptotics," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 735–748, Apr. 2019.

[29] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[30] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.

[31] P. Kachroo and S. Sastry. Traffic flow theory–Mathematical framework. [Online]. Available: http://faculty.unlv.edu/pushkin/TrafficFlowTheory.pdf

[32] S. Kashyap, E. Björnson, and E. G. Larsson, "On the feasibility of wireless energy transfer using massive antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3466–3480, May 2016.

[33] L. Zhao, T. Riihonen, W. Xiang, Y. Kuang, and K. Zheng, "Resource optimization of wireless information and energy supply control systems with massive MIMO," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2734–2737, Dec. 2017.

[34] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On low-resolution ADCs in practical 5G millimeter-wave massive MIMO systems," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 205–211, Jul. 2018.

[35] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[36] X. Liu, Y. Li, L. Xiao, and J. Wang, "Performance analysis and power control for multi-antenna V2V underlay massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4374–4387, Jul. 2018.

[37] "NR physical channels and modulation," 3GPP, Tech. Rep. TS 38.211 V15.6.0, Jun. 2019.

[38] F. Abbas, P. Fan, and Z. Khan, "A novel low-latency V2V resource allocation scheme based on cellular V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2185–2197, Jun. 2019.

[39] Y. Park, T. Kim, and D. Hong, "Resource size control for reliability improvement in cellular-based V2V communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 379–392, Jan. 2019.

[40] B. H. Fleury, "An uncertainty relation for WSS processes and its application to WSSUS systems," *IEEE Commun. Mag.*, vol. 44, no. 12, pp. 1632–1634, Dec. 1996.

[41] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Cambridge, Massachusetts: Academic Press, 2013.

[42] K. Shen and W. Yu, "Fractional programming for communication systems-Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[43] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 3-4, pp. 185–396, Jun. 2015.

[44] J. M. Borwein and P. B. Borwein, *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*. Hoboken, New Jersey: John Wiley & Sons, 1987.

[45] K. T. K. Cheung, S. Yang, and L. Hanzo, "Achieving maximum energy-efficiency in multi-relay OFDMA cellular networks: A fractional programming approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2746–2757, Jul. 2013.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, England: Cambridge University Press, 2004.

Haojun Yang (S'18) received the B.S. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014. Currently, he is working towards his Ph.D. degree in information and communication engineering at Intelligent Computing and Communications (IC^2) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, BUPT. His research interests include wireless communications and vehicular networks.

Kan Zheng (SM'09) received the B.S., M.S., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China, in 1996, 2000, and 2005, respectively. He is currently a full professor with BUPT. He has rich experience in research and standardization of new emerging technologies. He has authored over 200 journal articles and conference papers in the field of wireless communications, vehicular networks, IoT, security, and so on. He holds editorial board positions with several journals. He has organized several special issues in the journals including IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE COMMUNICATIONS MAGAZINE, and IEEE SYSTEMS JOURNAL. He has also served in the organizing/TPC committees for more than ten conferences.

Long Zhao (M'17) received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015, where he is currently an associate professor. From April 2014 to March 2015, he was a Visiting Scholar at the Department of Electrical Engineering, Columbia University. His research interests include wireless communications and signal processing.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk) FREng, FIEEE, FIET, Fellow of EURASIP, DSc received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 120+ PhD students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10000 pages, published 1850+ research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing an academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE ComSoc and VTS. During 2008-2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk.