Ultra-Reliable Low-Latency Communications in Autonomous Vehicular Networks

Xiaohu Ge, Senior Member, IEEE

Abstract—Autonomous vehicles are expected to emerge as a main trend in vehicle development over the next decade. To support autonomous vehicles, ultra-reliable low-latency communications (URLLC) is required between autonomous vehicles and infrastructure networks, e.g., fifth generation (5G) cellular networks. Hence, reliability and latency must be jointly investigated in 5G autonomous vehicular networks. Utilizing the Euclidean norm theory, we first propose a reliability and latency joint function to evaluate the joint impact of reliability and latency in 5G autonomous vehicular networks. The interactions between reliability and latency are illustrated via Monte-Carlo (MC) simulations of 5G autonomous vehicular networks. To improve both the reliability and latency performance and implement URLLC in 5G autonomous vehicular networks, resource slicing to service and function slicing is presented for 5G autonomous vehicular networks. The simulation results indicate that the proposed network slicing solution can improve both the reliability and latency performance and ensure URLLC in 5G autonomous vehicular networks.

Index Terms—Reliability, latency, autonomous vehicle, vehicular network, network slicing.

I. INTRODUCTION

Vehicular networks are emerging as a key application scenario for fifth generation (5G) mobile communication systems [1], and in the next decade, autonomous vehicles will represent one of the main transmitters/receivers of 5G vehicular networks [2]. Compared with traditionally manned vehicles, autonomous vehicles are extremely dependent on ultra-reliable low-latency communication (URLLC) in 5G vehicular networks. Previous studies have indicated that conflicts can occur between reliability and latency performance in vehicular networks [3], [4]. Hence, many studies have focused on optimizing either the reliability or latency performance in vehicular networks [5]–[7]. Considering the URLLC requirement, a joint model of reliability and latency must be investigated for 5G autonomous vehicular networks. Moreover, developing a solution for improving both reliability and latency in 5G autonomous vehicular networks, i.e., implementing URLLC is a considerable challenge.

Studies have investigated methods of improving the latency or reliability of vehicular networks [8]–[17]. A mobile-edge computing (MEC) architecture was proposed for cellular vehicular networks in which an inter-cell handover mechanism was developed for vehicles to enhance the latency performance of mobility management [8]. Based on software-defined networks (SDNs) and OpenFlow technologies, a SDN-enabled network architecture assisted by MEC was proposed to reduce the latency in vehicular networks [9], [10]. A multicast transmit beamforming technology was developed for vehicle to everything (V2X) communications that employs long-term evolution (LTE) multimedia broadcast single-frequency network (MBSFN) capabilities [11]. To reduce the latency in multi-hop vehicular networks, a new scheme was proposed to optimize the one hop transmission range based on a genetic algorithm [12]. Considering a limited number of roadside units (RSUs) along roads, a routing scheme was proposed for broadcast-based safety applications with ultra-reliability in vehicular ad hoc networks (VANETs) [13]. The connectivity probability was analyzed for platoon-based VANETs in which vehicles in the network have a Poisson distribution considering different traffic densities [14]. The simulation results showed that the connectivity probabilities in VANETs based on platoons are larger than those in VANETs without platoons in the V2X communication scenarios. To improve the network capacity and system computing capability, a matrix game approach was developed to manage the cloudlet resources of vehicular networks [15]. A three-stage radio resource management algorithm was developed to optimize resource sharing among vehicular users and cellular users in V2X applications [16]. By formulating the virtual resource allocation and caching strategies as a joint optimization problem, a new framework with information-centric wireless virtualization and device-to-device communications was proposed to enable content caching not only in the air but also in mobile devices [17]. However, enhancing both the reliability and latency of autonomous vehicular networks remains a key issue. To resolve this key issue, the network slicing technology is emerging as a potential solution for URLLC in autonomous vehicular networks.

Numerous network slicing concepts and methods have been proposed in the literatures on wireless networks [18]–[28]. The basic principles of network slicing that underly the mapping of dedicated and shared slices were discussed in [18] considering the necessary flexibility and scalability associated with 5G network implementation. A common framework was presented to integrate and discussing the latest developments in 5G network slicing, and the identified gaps were evaluated [19]. Network slicing was introduced as an integral approach to wireless network virtualization and shown to promote the programmable and configurable characteristics of network services [20]. A mobile-oriented Open-Flow protocol (MOFP) was proposed for the sharing of resources at adjacent base stations (BSs) and the implementation of service slicing for 5G cellular networks [21]. In [22], network slicing technology was proposed to satisfy different network slicing requirements and content by dynamically scheduling the edges and

Xiaohu Ge is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, China (Email: xhge@mail.hust.edu.cn).
central clouds in the studied wireless network. Based on network slicing technology, the virtual resource allocation of full-duplex relaying (FDR) networks was formulated as an optimization problem, and an efficient alternating direction method of multipliers (ADMM)-based distributed virtual resource allocation algorithm was developed to solve this problem [23]. The virtual network functions (VNFs) placement problem was formalized for radio access networks, and a slice scheduling mechanism was proposed to ensure resource and performance isolation among different slices [24]. The concept of a 5G network slice broker was introduced for 5G networks, and it enables mobile virtual network operators, over-the-top providers and industry vertical market players dynamically to request and lease resources from infrastructure providers via signaling [25]. In [26], the bandwidth of wireless networks was formed as a type of virtual slice by wireless network virtualization functions. Furthermore, [27] presented a framework that enables wireless virtualization and discussed a number of challenges that must be addressed for the deployment of wireless virtualization in the next generation of mobile cellular networks. Network slicing technology has been introduced into vehicular networks by utilizing resource sharing schemes. Based on SDN and fog computing technologies, a new vehicular network architecture was proposed to improve the coverage probability of the network slicing technology in 5G vehicular networks [28]. However, investigations of URLLC in autonomous vehicular networks are surprisingly rare in the open literatures. Additionally, the impact of network slicing technology on the reliability and latency performance of vehicular networks has been limited to simple scenarios, and methods of improving both the reliability and latency based on network slicing have not been investigated in 5G autonomous vehicular networks.

Motivated by the above research gaps, a reliability and latency joint function is proposed in this paper to evaluate the joint impact of reliability and latency on 5G autonomous vehicular networks. Moreover, a new network slicing solution with service, function and resource slicing is proposed to improve both the reliability and latency of 5G autonomous vehicular networks. The contributions and novelties of this paper are summarized as follows.

1) Based on the Euclidean norm theory, a reliability and latency joint function is proposed to evaluate the joint impact of reliability and latency on 5G autonomous vehicular networks. Moreover, the interactions between reliability and latency are analyzed for 5G autonomous vehicular networks, and a maximum of reliability and latency joint function with respect to the vehicle density is validated via Monte-Carlo (MC) simulations.

2) To improve both the reliability and latency performance of 5G autonomous vehicular networks, i.e., to implement URLLC, a new network slicing solution that extends from resource slicing to service and function slicing is presented. Moreover, a new network slicing algorithm is developed to improve both the reliability and latency performance of 5G autonomous vehicular networks.

3) The simulation results indicate that the proposed network slicing solution can improve both of the reliability and latency performance of 5G autonomous vehicular networks, i.e., by implementing URLLC. Moreover, the optimization of the proposed network slicing solution depends on the vehicle density in the 5G autonomous vehicular networks.

The remainder of this paper is outlined as follows. Section II describes the system model of autonomous vehicular networks. Section III investigates the coupled relationship between the reliability and latency in 5G autonomous vehicular networks based on the proposed reliability and latency joint function. Additionally, the reliability and latency performance are analyzed via MC simulations for 5G autonomous vehicular networks. Section IV presents the proposed network slicing solution for improving the reliability and latency in 5G autonomous vehicular networks, in which network slicing is extended from network resource slicing to service and function slicing. Section V compares and analyzes the reliability and latency performance of 5G autonomous vehicular networks with and without network slicing. Finally, Section VI presents the conclusions.

II. SYSTEM MODEL

Without loss of generality, an urban outdoor environment is projected in a plane $\mathbb{R}^2$, in which the roads are modeled by the Manhattan Poisson line process (MPLP) [29], [30]. In Fig. 1, the road distribution model features two unit-density homogeneous Poisson point processes (PPPs) $\Psi_x, \Psi_y \subset \mathbb{R}^2$ along the $x$-axis and $y$-axis, respectively. At each point in the processes, an avenue (west-east direction) and a street (north-south direction) grows infinitely along the $x$-axis and $y$-axis, respectively. The resulting line process is denoted by $\tilde{\Omega}$, which divides the plane into an infinite number of blocks. Based on the MPLP road model, the RSUs are assumed to be uniformly distributed along the roads in Fig. 1.

The infrastructure providers (InPs) $\mathcal{I} = \{Inp_{i_1}, Inp_{i_2}, ..., Inp_{i_n}\}$ in the plane $\mathbb{R}^2$ are assumed to be governed by an independent PPP distribution. The plane $\mathbb{R}^2$ is split into irregular polygons that correspond to different cell coverage areas. The split method is based on the Delaunay triangulation method where the perpendicular bisector lines are connected by each pair of InPs [31]. This stochastic and irregular topology forms a so-called Poisson-Voronoi tessellation (PVT) [32]. In each cell coverage area, an InP is located in the center of the cell and connects with all RSUs inside this cell coverage area. The InP gathers information from the RSUs and provides a virtualized network platform (VNP) to share network resources with the RSUs in the PVT cell.

Based on the Palm theory [32], [33], the analytical results for a typical PVT cell, e.g., $\text{Area}_c$ in Fig. 1, can be extended to the entire PVT network. Hence, $\text{Area}_c$ is magnified in Fig. 1 to show the distribution details in a typical PVT cell. In $\text{Area}_c$, the RSUs have network resources that can be used to communicate with autonomous vehicles on the road by millimeter wave wireless transmissions. Moreover, the beamforming technology is adopted for wireless links between...
the RSUs and vehicles to avoid the interference from adjacent RSUs. For each RSU, a RSU to infrastructure provider (R2I) link is configured to provide communications between the RSU and InP.

III. RELIABILITY AND LATENCY JOINT FUNCTION IN 5G AUTONOMOUS VEHICULAR NETWORKS

Most studies have separately modeled and analyzed the reliability and latency performances of vehicular networks [34]; however, these parameters are coupled in vehicular networks. When an algorithm is applied to optimize the latency or reliability performance of vehicular networks, the other performance parameters in the vehicular network will be affected, and performance deterioration can occur [35]. Considering the URLLC requirement in 5G autonomous vehicular networks, the interactions between reliability and latency must be modeled and analyzed. In this section, a new reliability and latency joint function based on the Euclidean norm theory is proposed for 5G autonomous vehicular networks. Furthermore, the coupled relationship between reliability and latency is analyzed for 5G autonomous vehicular networks.

A. Reliability and Latency Joint Function

Based on the Euclidean norm theory, the reliability and latency joint function of 5G autonomous vehicular networks $S(L_t, L_p)$ is defined as follows:

$$S(L_t, L_p) = \left( |L_t|^2 + |\omega \cdot L_p|^2 \right)^{1/2}$$  \hspace{1cm} (1)

where $L_p$ and $L_t$ are the reliability and latency utilities in autonomous vehicular networks, respectively; $\omega$ is the weight factor that balances the effects of reliability and latency on $S(L_t, L_p)$; $I_p(P_{req})$ is the reliability utility function with the reliability constraint $P_{req}$; and $I_t(T_{req})$ the latency utility function with the latency constraint $T_{req}$, $|\cdot|$ is the norm operation. Based on the reliability and latency joint function, both the reliability and latency performance can be evaluated by the same metric, i.e., the norm length in Euclidean space. Therefore, the URLLC can be investigated based on a uniform model of the reliability and latency joint function in 5G autonomous vehicular networks.

To evaluate how reliability and latency are satisfied in autonomous vehicular networks, the reliability and latency utility functions are extended as follows:

$$I_p(P_{req}) = e^{a_p \frac{(P_{req} - P)}{T_{req}}}, \hspace{1cm} (2)$$

$$I_t(T_{req}) = e^{a_t \frac{(T_{req} - T)}{P_{req}}}, \hspace{1cm} (3)$$

where $P$ and $T$ are the reliability and latency values in 5G autonomous vehicular networks, respectively; $P_{req}$ and $T_{req}$ are the required reliability and latency thresholds for the desired services in 5G autonomous vehicular networks, respectively; and $a_p$ and $a_t$ are the weighted factors for the reliability and latency utility functions, respectively. When the values of reliability and latency are less than the required thresholds in autonomous vehicular networks, the reliability and latency utility function values are always larger than 0 and less than 1, i.e., $0 < I_t(T_{req}) < 1$ and $0 < I_p(P_{req}) < 1$. When the reliability and latency requirements are satisfied in autonomous vehicular networks, i.e., the reliability and latency values are larger than or equal to the required thresholds, the reliability and latency utility function values are always larger than or equal to 1, i.e., $I_t(T_{req}) \geq 1$ and $I_p(P_{req}) \geq 1$.  

Fig. 1. System model. Coverage areas of different hot spots are shown as circles with different radius, $s_{ij}$ is the total number of AEs in circles marked as the purple regions.
Moreover, the reliability utility function value increases as the reliability increases, and the latency utility function value increases as the latency in autonomous vehicular networks decreases. In this case, the value of $S(L_t, L_p)$ increases as the reliability and latency improve in 5G autonomous vehicular networks. When the value of $S(L_t, L_p)$ is less than 1, either the reliability or latency requirement has not been satisfied in the 5G autonomous vehicular networks.

B. Reliability Model

The unreliability of messages in 5G autonomous vehicular networks is mainly caused by transmission errors in wireless links between the vehicles and RSUs. In this case, the reliability of autonomous vehicular networks is denoted by the probability of the message being successfully transmitted between the vehicle and the RSU. The wireless link between the vehicle and RSU is denoted by $L$, and the number of hops is $N_L$. The distance between two nodes in the wireless link $L$ is denoted by $D_{i,j}^L$, where $i$ and $j$ are positive integers, $D_{i,j}^L$ denotes the distance between two vehicles and $D_{i,j}^R$ denotes the distance between the vehicle and the RSU. If the vehicle node is denoted as $D^V_i$ and the RSU node is denoted as $D^R_j$, then the wireless link $L$ between the vehicle $D^V_i$ and RSU $D^R_j$ is expressed as a set as follows:

$$L = \{D_{i,1}^V, D_{i,2}^V, \ldots, D_{i,n}^V \}$$

where $n$ is a positive integer.

In this paper, millimeter wave transmissions are adopted for the wireless links in 5G autonomous vehicular networks, and the frequency of the millimeter wave transmissions is 75 GHz [36, 37]. Without loss of generality, every RSU is assumed to cover a road with a length of $L$ meters in an urban environment. The density of RSUs in an urban area is denoted as $\rho_{RSU} = \frac{1}{L}$. The vehicle density on the road is configured as follows:

$$\rho_i = \frac{1}{L}$$

The path loss $PL[dB](\cdot)$ between the vehicle $V_i$ and the RSU $RSU_i$ can be expressed as follows:

$$PL[dB](\cdot) = 69.6 + 20.9 \log(d) + \xi, \xi \sim (0, \sigma^2)$$

where $\xi$ is additive white Gaussian noise (AWGN) with a mean of 0 and variance $\sigma^2$.

In this paper, beamforming technology is adopted for millimeter wave transmissions. Considering the directionality and fast fading features of millimeter wave transmissions, the interference among vehicles and RSUs is ignored in 5G autonomous vehicular networks. The threshold of the signal-to-noise ratio (SNR) at the receivers is configured as $\theta$ in one hop in a 5G autonomous vehicular network. When the SNR at the receivers is larger than or equal to the threshold $\theta$, the signal can be successfully accepted at the receivers [39, 40]. The one hop successful transmission probability $P_{hop}$ that the message can be successfully transmitted in a time slot is as follows [41]:

$$P_{hop} = P(L \leq P_{tx}(dB) - \theta(dB) - N_0 W_{mmWave}(dB))$$

where $P_{tx}$ is the transmission power, $N_0$ is the power spectrum density of AWGN, and $W_{mmWave}$ is the bandwidth of the millimeter wave transmissions. By substituting (5) into (6), and the one hop successful transmission probability $P_{hop}$ can be expressed as follows:

$$P_{hop} = P(\xi \leq P_{tx}(dB) - \theta(dB))$$

$$= \frac{1}{2} \left(1 + erf \left(\frac{\psi(d)}{\sqrt{2}\sigma}\right)\right)$$

where $erf(\cdot)$ is the deviation function and $\psi(d) = P_{tx}(dB) - \theta(dB) - N_0 W_{mmWave}(dB) - 69.6 - 20.9 \log_{10}d$.

C. Latency Model

The messages between vehicles and adjacent RSUs are transmitted by vehicle to infrastructure (V2I) links. The RSU normalizes all received messages and forms a message queue to manage all vehicle messages in the coverage area. Without loss of generality, the total latency between the vehicles and RSUs can be divided into propagation latency and handling latency. Hence, the total latency of message $T$ can be expressed as follows:

$$T = T_{p}^M + T_{h}^M$$

where $T_{p}^M$ is the propagation latency in the wireless links and $T_{h}^M$ is the handling latency in the queues of RSUs.

When the wireless link between the vehicle and the RSU comprises one wireless hop, the propagation latency is expressed as follows:

$$T_{hop}(d) = t_{slot} / P_{hop} = \frac{2t_{slot}}{1 + erf \left(\frac{\psi(d)}{\sqrt{2}\sigma}\right)}$$

where $t_{slot}$ is a constant time slot used for transmitting messages.

When the wireless link $L$ between the vehicle and the RSU is composed of multiple hops, i.e., a message from the vehicle is relayed via multiple vehicles to reach the desired RSU, the propagation latency of multi-hop $T_{p}^M$ in the wireless link $L$ is expressed as follows:

$$T_{p}^M = T_{trans} + T_{proc}$$

where $T_{trans}$ is the transmission latency in the wireless link $L$ and $T_{proc}$ is the total relaying latency, which includes the processing latency $t_{proc}$ of every relaying vehicle in the wireless link $L$. 

C. Latency Model

The messages between vehicles and adjacent RSUs are transmitted by vehicle to infrastructure (V2I) links. The RSU normalizes all received messages and forms a message queue to manage all vehicle messages in the coverage area. Without loss of generality, the total latency between the vehicles and RSUs can be divided into propagation latency and handling latency. Hence, the total latency of message $T$ can be expressed as follows:

$$T = T_{p}^M + T_{h}^M$$

where $T_{p}^M$ is the propagation latency in the wireless links and $T_{h}^M$ is the handling latency in the queues of RSUs.

When the wireless link between the vehicle and the RSU comprises one wireless hop, the propagation latency is expressed as follows:

$$T_{hop}(d) = d_{slot} / P_{hop} = \frac{2d_{slot}}{1 + erf \left(\frac{\psi(d)}{\sqrt{2}\sigma}\right)}$$

where $d_{slot}$ is a constant time slot used for transmitting messages.

When the wireless link $L$ between the vehicle and the RSU is composed of multiple hops, i.e., a message from the vehicle is relayed via multiple vehicles to reach the desired RSU, the propagation latency of multi-hop $T_{p}^M$ in the wireless link $L$ is expressed as follows:

$$T_{p}^M = T_{trans} + T_{proc}$$

where $T_{trans}$ is the transmission latency in the wireless link $L$ and $T_{proc}$ is the total relaying latency, which includes the processing latency $t_{proc}$ of every relaying vehicle in the wireless link $L$. 

C. Latency Model

The messages between vehicles and adjacent RSUs are transmitted by vehicle to infrastructure (V2I) links. The RSU normalizes all received messages and forms a message queue to manage all vehicle messages in the coverage area. Without loss of generality, the total latency between the vehicles and RSUs can be divided into propagation latency and handling latency. Hence, the total latency of message $T$ can be expressed as follows:

$$T = T_{p}^M + T_{h}^M$$

where $T_{p}^M$ is the propagation latency in the wireless links and $T_{h}^M$ is the handling latency in the queues of RSUs.

When the wireless link between the vehicle and the RSU comprises one wireless hop, the propagation latency is expressed as follows:

$$T_{hop}(d) = d_{slot} / P_{hop} = \frac{2d_{slot}}{1 + erf \left(\frac{\psi(d)}{\sqrt{2}\sigma}\right)}$$

where $d_{slot}$ is a constant time slot used for transmitting messages.

When the wireless link $L$ between the vehicle and the RSU is composed of multiple hops, i.e., a message from the vehicle is relayed via multiple vehicles to reach the desired RSU, the propagation latency of multi-hop $T_{p}^M$ in the wireless link $L$ is expressed as follows:

$$T_{p}^M = T_{trans} + T_{proc}$$

where $T_{trans}$ is the transmission latency in the wireless link $L$ and $T_{proc}$ is the total relaying latency, which includes the processing latency $t_{proc}$ of every relaying vehicle in the wireless link $L$. 

C. Latency Model

The messages between vehicles and adjacent RSUs are transmitted by vehicle to infrastructure (V2I) links. The RSU normalizes all received messages and forms a message queue to manage all vehicle messages in the coverage area. Without loss of generality, the total latency between the vehicles and RSUs can be divided into propagation latency and handling latency. Hence, the total latency of message $T$ can be expressed as follows:

$$T = T_{p}^M + T_{h}^M$$

where $T_{p}^M$ is the propagation latency in the wireless links and $T_{h}^M$ is the handling latency in the queues of RSUs.

When the wireless link between the vehicle and the RSU comprises one wireless hop, the propagation latency is expressed as follows:

$$T_{hop}(d) = d_{slot} / P_{hop} = \frac{2d_{slot}}{1 + erf \left(\frac{\psi(d)}{\sqrt{2}\sigma}\right)}$$

where $d_{slot}$ is a constant time slot used for transmitting messages.
Every RSU is configured with the same number of local resource blocks $RB_{rsu}$. Considering the random mobility of vehicles on the roads, the number of vehicles in the coverage area of every RSU is different, as is the arrival rate of messages at every RSU. In this case, the message handling process at the RSU is assumed as a $GI/M/1/\infty$ queuing system. In this $GI/M/1/\infty$ queuing system, it is assumed that $\tau_0 = 0$ is the arrival epoch of the first task; the inter-arrival times $\{\tau_i, i \geq 1\}$ are independent and identically distributed with a general distribution function denoted by $F(t)$, $t \geq 0$; and the mean is $0 < \frac{1}{\mu} = \int_{0}^{\infty} t df(t), (\lambda > 0)$, where $\lambda$ is the arrival rate. The service times $\{\chi_i, i \geq 1\}$ during a service period are exponentially distributed at the rate of $\mu (\mu > 0)$; therefore, the distribution function of service times is denoted by $G(t) = 1 - e^{-\mu t}, t \geq 0$. Considering the $GI/M/1/\infty$ queuing system, the average service time of a message at RSUs is dependent on the number of resource blocks in the RSU $[42]$. To simplify the analysis, $\frac{1}{\rho_{rsu}(i) \mu_{0}}$ denotes the average service time when a message is served by a resource block, and $\frac{1}{\rho_{rsu}(i) \mu_{0}}$ is the average service time when a message is served by RSU, with $RB_{rsu}(i)$ resource blocks. In the $GI/M/1/\infty$ queuing system, the handling latency of a message at an RSU corresponds to the following theorems.

**Theorem 1:** Without loss of generality, $RSU_i (i = 1, 2, \ldots, m)$ covers a road with a length of $L$ meters. The average density of vehicles on the road is denoted as $\rho_i$, and the transmission rate of messages generated by a vehicle is $\lambda_v$. The cumulative distribution function (CDF) of the message interval arrival time at $RSU_i$ is expressed as follows:

$$F(t) = 1 - e^{-\rho_i L (1 - e^{-\lambda_v t})} - e^{-\rho_i L}.$$  

The proof for **Theorem 1** is provided in Appendix A.

**Theorem 2:** When the message handling process at an RSU is assumed to occur based on a $GI/M/1/\infty$ queuing system, the distribution of the message dwelling time at the RSU is governed by the following equation:

$$W(t) = 1 - e^{-\mu t} + \frac{K^*}{1 - \delta} \left[1 - e^{-c\mu(1-\delta)t}\right],$$  

where

$$K^* = \left[\frac{1}{1 - \delta} + \sum_{k=1}^{c} \left(\frac{e}{k} D_k (1 - \varepsilon_k) \right) \cdot \frac{c(1 - \varepsilon_k) - k}{e(1 - \delta) - k}\right]^{-1}.$$  

$$D_k = \left\{\begin{array}{ll} 1 & k = 0 \\ \prod_{l=1}^{k} \frac{1}{\varepsilon_l} & k = 1, 2, \ldots, c \end{array}\right.$$  

$$\varepsilon_l = \int_{0}^{\infty} e^{-\mu t} dF(t), \quad l = 1, 2, \ldots.$$  

The proof for **Theorem 2** is provided in Appendix B.

Furthermore, the average dwelling time of a message in the $GI/M/1/\infty$ queuing system is derived as follows:

$$T_{rsu}^q = E [W(t)] = \frac{1}{\mu} + \frac{K^*}{c\mu(1-\delta)^2}. \quad (17)$$

Considering the $GI/M/1/\infty$ queuing system at the RSU, the parameter $c$ is set as a constant, i.e., $c = 1$. Hence, the average dwelling time of a message is derived as follows:

$$T_{rsu}^q = E [W(t)] = \frac{1}{\mu (1 - \delta)}. \quad (18)$$

When the value of $\delta$ is limited as $\delta \in (0, 1)$, $\delta$ can be solved based on the following relation:

$$\sum_{k=0}^{\infty} \delta^k \int_{0}^{\infty} e^{-\mu t} \left(\frac{t}{k!}\right)^k df(t) = \int_{0}^{\infty} e^{-\mu(1-\delta)t} df(t) = \delta.$$

When the CDF of message interval time in (12) is substituted into (19), (19) can be extended as follows:

$$\int_{0}^{\infty} e^{-\mu(1-\delta)t} \lambda_v \rho_i L e^{-\rho_i L} dt = \delta. \quad (20)$$

**D. Performance Analysis of Reliability and Latency**

To analyze the reliability and latency performance of 5G autonomous vehicular networks, the default simulation parameters are configured as follows: the weight factor $\omega$ is set to 10, the road length $L$ covered by the RSU is 100 meters [43], [44], the vehicle density on the road $\rho_i$ is 0.2 vehicle per meter (veh/m) [45], the transmission rate of messages generated from a vehicle $\lambda_v$ is 50 messages per second, the average service time $1/\mu_0$ is 5 milliseconds [46], the transmission power of the vehicle $P_t$ is 30 dBm [47], the noise power density $N_0$ is -174 dBm/Hz [47], the duration of a slot $t_{slot}$ is 50 microseconds [48], the SNR threshold $\theta$ is 5 dB [47], and the number of resource blocks in a RSU $RB_{rsu}$ is 10. Without loss of generality, the wireless link between the vehicle and RSU is composed of random multiple hops, which are governed by a uniform distribution. Moreover, MC simulations are used for the performance analysis in this paper.

Fig. 2 describes the propagation latency with respect to the vehicle density based on different RSU densities.
Fig. 3. Handling latency with respect to the vehicle density considering different RSU densities.

RSU density is fixed, the handling latency increases as the vehicle density increases. When the vehicle density is fixed, the handling latency decreases as the RSU density increases.

Fig. 4 illustrates the total latency with respect to the density of vehicle considering different RSU densities. When the RSU density is fixed, the total latency increases as the vehicle density increases. When the vehicle density is fixed, the total latency decreases as the RSU density increases.

Fig. 5 depicts the reliability with respect to the density of vehicle considering different RSU densities. When the RSU density is fixed, the reliability increases as the vehicle density increases. When the vehicle density is fixed, the reliability increases as the RSU density increases.

Fig. 6 indicates the relationship between the reliability and latency utility functions in 5G autonomous vehicular networks. When the RSU density is fixed, the reliability utility function increases and the latency utility function decreases as the vehicle density increases. Hence, a conflict occurs between the reliability and latency performance with respect to the vehicle density in 5G autonomous vehicular networks. When the vehicle density is fixed, the reliability and latency utility functions increase as the RSU density increases.

Fig. 7 analyzes the reliability and latency joint function with respect to the vehicle density considering different RSU densities. When the RSU density is fixed and the vehicle density is less than 0.19 veh/m, the values of the reliability and latency joint function initially increase as the vehicle density increases. When the RSU density is fixed and the vehicle density is larger than or equal to 0.19 veh/m, the values of the reliability and latency joint function decrease as the vehicle density increases. Hence, a maximum of reliability and latency joint function value exists with respect to the vehicle density. When the vehicle density is fixed, the values of the reliability and latency joint function value increase as the RSU density increases.
IV. NETWORK SLICING SOLUTION

Based on the results in Fig. 6, a conflict is observed regarding the reliability and latency performance in 5G autonomous vehicular networks. To implement URLLC, both the reliability and latency performance in 5G autonomous vehicular networks must be improved. Considering the reliability and latency gains associated with using network slicing technology, network slicing has emerged as an attractive solution for 5G autonomous vehicular networks. In conventional network slicing technology, only network resources are sliced to improve the resource utilization in wireless networks. However, satisfying the URLLC requirement in 5G autonomous vehicular networks based on network resource slicing alone is difficult. In this case, we extend network slicing from resource slicing to service and function slicing to improve both the reliability and latency of 5G autonomous vehicular networks. The relevant methods are described in detail in this section.

A. Service Slicing

Different types of vehicular services are available in 5G autonomous vehicular networks. To improve the service access efficiency in 5G autonomous vehicular networks, we normalize different types of vehicular services into three types of service slices.

- State-Report Service Slice
  State-report service slices (SRSSs) are used to provide the state of the vehicle, such as the speed and location. SRSSs have the following features: the arrival rate is high; the holding time of the state is short; and the SRSSs have similar contexts. SRSSs are used for vehicle collision avoidance, vehicle lane changes and vehicle deceleration. Hence, SRSSs must be quickly accessed. Moreover, SRSSs should be processed at RSU locations to reduce the response time.

- Event-Driven Service Slice
  Event-driven service slices (EDSSs) are generated by specific events, such as the emergency broadcasts and the road information updates. Although the EDSS arrival rate is low, EDSSs usually involve with security issues. Hence, EDSSs must be accessed with a low latency and high reliability in 5G autonomous vehicular networks.

- Entertainment-Application Service Slice
  Entertainment-application service slices (EASSs) are generated from vehicle users. The latency requirement of EASSs depends on the type of service applications. Different types of service applications have different latency requirements. Hence, the EASS requirements change for different types of service applications in 5G autonomous vehicular networks.

Based on the analysis of the three types of service slices, different types of service slices can be distinguished by the arrival rate, handling time and latency requirement. To simplify the analysis of service slices in 5G autonomous vehicular networks, the service slice of a 5G autonomous vehicular network is denoted as \( S \{ \lambda_a, \mu_s, T_{req} \} \), where \( \lambda_a \) is the arrival rate of the service slice, \( \mu_s \) is the handling time of service slice and \( T_{req} \) is the latency requirement of the service slice.

B. Function Slicing and Resource Slicing

To improve both the reliability and latency, the protocol functions at the RSUs and InPs must be sliced and restructured by SDN technologies to support service and resource slicing in 5G autonomous vehicular networks. Therefore, we propose a new 5G autonomous vehicular network architecture to implement service, resource and function slicing, as shown in Fig. 8. In this paper, function slicing is implemented by dynamically scheduling and assembling function modules in the logical layers of the proposed autonomous vehicular network architecture. Based on the proposed autonomous vehicular network architecture, the protocol structure of an InP is composed of two logical layers, i.e., the InP control layer and InP shared resource layer. The InP control layer includes two function modules: the monitoring module and the allocation module. The monitoring module monitors the number of occupied resource blocks at the RSUs in an InP coverage area. Based on the information from the monitoring module, such as the number of occupied resource blocks that reach the available number of resource blocks and the number of resource blocks that is still required by the service slices at the RSU, the allocation module will schedule the unoccupied resource blocks from adjacent RSUs to support the requirements of the specified RSUs. The InP shared resource layer is a type of virtual network resource pool that includes all unoccupied resource blocks in the coverage of the InP. The resource slicing is performed by the InP shared resource layer to improve the resource utilization efficiency in 5G autonomous vehicular networks. In the real world, resource blocks are stored at the RSUs. According to the logic relationships in the proposed autonomous vehicular network architecture, all unoccupied resource blocks in the coverage of the InP can be regarded as stored in the virtual network resource pool in the InP shared resource layer. Furthermore, all resource blocks stored in the virtual network resource pool can be sliced to effectively satisfy the resource requirements of service slices at the RSUs.

The protocol structure of an RSU is divided into three logical layers: the RSU interface layer, RSU virtualization layer and RSU resource layer. The RSU interface layer provides different interfaces corresponding to different types of service slices. The main function of the RSU interface layer is to separate RSU resource requirements with different types of service slices. Based on the RSU interface layer, the relationships among different types of service slices are normalized by separating the resource requirements.

The function of the RSU virtualization layer is to allocate locational RSU resource blocks to satisfy the requirements of the service slices from the RSU interface layer. Moreover, the RSU virtualization layer can optimize the handling process to promptly respond to the service slices with low latency constraint.

The RSU resource layer manages the resource blocks at the RSU location. The RSU resource blocks were previously allocated for the location service slices. Thus, the RSU virtualization layer can require resource blocks from the associated InP control layer when the RSU resource layer cannot
satisfy the requirements of the location service slices. When the resource blocks from the InP shared resource layer are scheduled in the RSU virtualization layer, the RSU resource layer will allocate the additional resource blocks from the RSU virtualization layer to meet the requirements of the location service slices.

C. Network Slicing Algorithm

Assuming that $N_{RSU}$ RSUs exist in a coverage area of an InP, the average number of $N_{RSU}$ is derived in Appendix C. A $N_{RSU} \times N_{RSU}$ matrix $A_{N_{RSU} \times N_{RSU}}$ is used to describe the multiplexing of resource blocks at an RSU when the network slicing technology is adopted for 5G autonomous vehicular networks. Notably, $a_{kg} \in [0, 1)$, $1 \leq k \leq N_{RSU}$ and $1 \leq g \leq N_{RSU}$, is the element of the matrix $A_{N_{RSU} \times N_{RSU}}$. When $k \neq g$, $a_{kg}$ is the percentage of resource blocks at the $g$th RSU that is scheduled for the $k$th RSU. When $k = g$, $a_{kg}$ is the percentage of resource blocks at the $g$th RSU used for local service slices. To ensure that local service slice handling occurs at the local RSU, an upper bound ratio $a_{kg}^{Max}$ is established to restrict the percentage of local RSU resource blocks used for other RSUs, i.e., $0 \leq a_{kg} \leq a_{kg}^{Max} < 1$ when $k = g$. In the proposed 5G autonomous vehicular network architecture, the RSU virtualization layer of $RSU_k$ can schedule the resource blocks from adjacent RSUs for local service slicing and reduce the handling latency of the service slice. The handling latency of a service slice at $Inp_k$ is denoted as $T_s^{k}$ ($k = 1, 2, ..., N_{RSU}$) and the handling latency of service slice at the local RSU $RSU_k$ is denoted by $T_s^{m,k}$. When network slicing is adopted for 5G autonomous vehicular networks, the handling latency $T_s^{m,k}$ in the queue of the RSU is calculated based on the total handling latency of a service slice at the RSU $RSU_k$, i.e., $T_s^{m,k} = T_s^{M,k}$. Moreover, the total handling latency of a service slice at $RSU_k$ is derived as follows:

$$T_s^{M,k} = \frac{a_{kk}}{N_{RSU}} T_s^{k} + \sum_{g=1}^{g=1} a_{kg} \frac{a_{kg}}{N_{RSU}} T_s^{q}.$$

Based on the proposed autonomous vehicular network architecture in Fig. 8, the resource blocks in a 5G autonomous vehicular network can be multiplexed by the following process. When a service slice arrives at $RSU_k$, the associated $Inp_k$ determines whether the local resource blocks of $RSU_k$ can satisfy the latency requirement of service slice $T_s^{eq}$. Additional, when the local resource blocks can satisfy the latency requirement of the service slice, i.e., $T < T_s^{eq}$, $Inp_k$ decreases the number of local resource blocks that can be scheduled by $RSU_k$ until $T = T_s^{eq}$. In this case, the number of occupied resource blocks is $RB_{rsu}(k)'$, which satisfies $T = T_s^{eq}$. The remaining resource blocks $RB_{rem}(k) = RB_{rsu}(k) - RB_{rsu}(k)'$ at $RSU_k$ are added to the virtual network resource pool which can be scheduled by $Inp_k$. When the local resource blocks can not satisfy the latency
Algorithm 1 Network Slicing Algorithm

Begin:

1: Initialize $A_{N_{RSU} \times N_{RSU}}$, $N_{RSU}$, $RB_{rsu}(k)$, $RB_{rem}$, $RB_{req}$, $T_{S}^{req}$, $a_{kg}^{Max}$;
2: for $k = 1$ to $N_{RSU}$ do
3:     $T = T_{P}^{M} + T_{M}^{M}$;
4:     while $T > T_{S}^{req}$ do
5:         update $RB_{rsu}(k)$ by decreasing $RB_{rsu}(k)$;
6:         calculate $T$ with $RB_{rsu}(k)$ by (11);
7:         $RB_{rem}(k) = RB_{rsu}(k) - RB_{rsu}(k)$;
8:     while $T > T_{S}^{req}$ do
9:         update $RB_{rsu}(k)$ by increasing $RB_{rsu}(k)$;
10:        calculate $T_{S,M,k}$ with $RB_{rsu}(k)$ by (21);
11: $RB_{req}(k) = RB_{rsu}(k) - RB_{rsu}(k)$;
12: $RB_{req} = RB_{req} + RB_{req}(k)$;
13: while $RB_{rem} \& RB_{req} \neq 0$ do
14:     schedule resource blocks from the virtual network resource pool to the RSU, which requires the resource blocks to reduce the handling latency of the service slice;
15:     update the matrix $A_{N_{RSU} \times N_{RSU}}$;
16:     for $k = 1$ to $N_{RSU}$ do
17:     calculate $T_{S,M,k}$ by (21);

The requirement of a service slice, i.e., $T > T_{S}^{req}$, $Inp_{k}$ increases the additional number of resource blocks scheduled by $RSU_{k}$ until $T = T_{S}^{req}$. In this case, $Inp_{k}$ schedules the number of resource blocks $RB_{req}(k) = RB_{rsu}(k) - RB_{rsu}(k)$ from the virtual network resource pool to support the service slice handling at $RSU_{k}$. The algorithm is described in detail in the Network Slicing Algorithm.

V. SIMULATION RESULTS AND DISCUSSION

To analyze the performance of the network slicing algorithm for 5G autonomous vehicular networks, the default simulation parameters in Section III are used in the following simulations. When service slicing is adopted in 5G autonomous vehicular networks, all messages are classified into three types: SRSS, EDSS and EASS. Different types of service slices have different packet sizes. Different data packet sizes have different transmission times that correspond to different slot values $t_{slot}$. 

To simplify the simulation analysis, the values of $t_{slot}$ are set to 25, 50 and 100 milliseconds, which correspond to the SRSS, EDSS and EASS in 5G autonomous vehicular networks with network slicing, respectively. Because SRSSs are regularly transmitted by vehicles, the number of SRSSs is obviously larger than the numbers of EDSSs and EASSs. Without loss of generality, the proportion of the numbers of SRSS, EDSS and EASS is configured as 8:1:1. The network slicing solution is adopted for 5G autonomous vehicular networks, and the reliability and latency are then analyzed and compared based on MC simulations in this section.

Fig. 9 compares the reliability and latency utility functions in 5G autonomous vehicular networks with and without network slicing. When network slicing is adopted in 5G autonomous vehicular networks, the reliability utility function increases and the latency utility function decreases as the vehicle density increases. A conflict still occurs between the reliability and latency utility functions. Compared with the values of reliability and latency utility functions without network slicing, the reliability and latency utility function values are both improved by adopting the network slicing algorithm in 5G autonomous vehicular networks.

Fig. 10 analyzes the reliability and latency joint function with respect to the density of vehicle considering different upper bound ratios $a_{kg}^{Max}$ in 5G autonomous vehicular networks with and without network slicing. When the upper bound ratio is fixed, the values of the reliability and latency joint function with network slicing are larger than those without network slicing. When the network slicing solution is adopted and the vehicle density is less than 0.12 veh/m, the reliability and latency joint function values for $a_{kg}^{Max} = 0.8$ are larger than those for $a_{kg}^{Max} = 0.9$. If a high value of upper bound ratio, e.g., $a_{kg}^{Max} = 0.9$ is configured for low vehicle density scenarios, i.e., the vehicle density is less than 0.12 veh/m, overmuch resource blocks will be scheduled by adjacent RSUs and then the reliability and latency joint function of local RSU has to be obviously reduced. As a consequence, the reliability and latency joint function of autonomous vehicular networks is decreased. When the network slicing solution is adopted, and the vehicle density is larger than or equal to 0.12 veh/m, the reliability and latency joint function values for $a_{kg}^{Max} = 0.8$ are less than or equal to those for $a_{kg}^{Max} = 0.9$.

If a low value of upper bound ratio, e.g., $a_{kg}^{Max} = 0.8$ is configured for the high vehicle density scenarios, i.e., the vehicle density is larger than or equal to 0.12 veh/m, a few of resource blocks can be scheduled by adjacent RSUs and then the reliability and latency joint function of adjacent RSUs has to be suppressed. As a result, the reliability and latency joint function of autonomous vehicular networks is repressed.

VI. CONCLUSION

In this paper a reliability and latency joint function is proposed to evaluate the joint impact of reliability and latency on autonomous vehicular networks.
5G autonomous vehicular networks. Moreover, the interactions between reliability and latency are quantitatively analyzed based on MC simulations. The simulation results indicate that a maximum reliability and latency joint function value occurs with respect to the vehicle density in 5G autonomous vehicular networks. To improve both the reliability and latency in 5G autonomous vehicular networks, i.e., to implement URLLC, a network slicing solution is proposed in this paper. Moreover, we propose a new vehicular network architecture in which network slicing is extended from network resource slicing to service and function slicing. Furthermore, a new network slicing algorithm is developed to implement URLLC in 5G autonomous vehicular networks. The simulation results show that the proposed network slicing algorithm can improve both the reliability and latency of 5G autonomous vehicular networks.

In future work, ultra-reliable low-latency performance must be optimized in future autonomous vehicular networks, and the following potential topics should be addressed: 1) optimize the service slicing scheme to match the type and number of messages in future autonomous vehicular networks; 2) optimize the network resource slicing scheme to account for the service slicing results; and 3) optimize the function slicing scheme to support the service and network resource slicing schemes in future autonomous vehicular networks.

APPENDIX A

The message interval time generated from a vehicle is denoted as $T_0$. For the time duration $t \geq 0, s \geq 0$, we have the following result can be obtained:

$$P \{ T_0 > t + s | T_0 > s \} = P \{ T_0 > t \} = e^{-\lambda s t}. \quad (22)$$

Assuming that there are $n$ vehicles in the coverage of a RSU, when the message arrival time interval at the RSU is denoted by $T^s_{RSU}$, the probability that $T^s_{RSU}$ is less than or equal to the time duration $t$ is derived as follows:

$$P \{ T^s_{RSU} \leq t \} = 1 - P \{ T^s_{RSU} > t \} = 1 - P \{ \text{There are no messages generated in the coverage area of the RSU within the time slot (0, t]} \} = 1 - \sum_{i=1}^{\infty} \frac{(e^{-\lambda t})^i}{1 - e^{-\lambda t}}. \quad (23)$$

Hence, the CDF of the message arrival time interval at the RSU is derived as follows:

$$F(t) = 1 - \frac{e^{-\rho s t}(1 - e^{-\lambda t}) - e^{-\rho s t}}{1 - e^{-\rho s t}}. \quad (24)$$

Additional the PDF of the message arrival time interval at the RSU is derived as follows:

$$f(t) = \lambda s \rho i L e^{-\rho s t + \rho s t e^{-\lambda t} - \lambda t} \frac{1}{1 - e^{-\rho s t}}. \quad (25)$$

Thus, Theorem 1 is proven.

APPENDIX B

Assume that the message handling process at the RSU is a $GI/M/1/\infty$ queuing system. Thus, $p_j^-$ is the probability that there are $j$ messages in the queue of the RSU. Based on the results in [49], $\{ p_j^- = 0, j \geq 0 \}$ is not a stable distribution in the $GI/M/1/\infty$ queuing system when $\sum \frac{\lambda}{c \mu} \geq 1$. $\{ p_j^- \neq 0, j \geq 0 \}$ is a stable distribution in the $GI/M/1/\infty$ queuing system when $\sum \frac{\lambda}{c \mu} < 1$. Considering $\sum \frac{\lambda}{c \mu} < 1$, the probability $p_j^-$ is expressed as follows:

$$p_j^- = \sum_{i=0}^{\infty} p_i^- \cdot p_{ij} \cdot (1), \quad j = 0, 1, 2, \cdots, \quad (26)$$

where $p_{ij} \cdot (1)$ is the one step transition probability when the number of queuing messages changes from $i$ to $j$. Considering $\sum_{j=0}^{\infty} p_j^- = 1$ and $j \geq c$, we obtain the following result:

$$p_{ij} \cdot (1) = \left\{ \begin{array}{ll}
\int_0^{\infty} e^{-cjt} (\frac{cjt}{(i-j+1)!})^k dF(t), & i \geq j - 1 \\
0, & i < j - 1
\end{array} \right. \quad (27)$$

When (27) is substituted into (26) and $j \geq c$, the probability $p_j^-$ is derived as follows:

$$p_j^- = \sum_{i=j-1}^{\infty} p_i^- \cdot \int_0^{\infty} e^{-cjt} (\frac{cjt}{(i-j+1)!})^k dF(t), \quad j = c, c + 1, c + 2, \cdots \quad (28)$$

Based on the results in [49], the following relation can be derived:

$$\sum_{k=0}^{\infty} k \cdot \int_0^{\infty} e^{-cjt} (\frac{cjt}{k!})^k dF(t) = \frac{c \mu}{\lambda} > 0. \quad (29)$$

Additionally, based on (29), the following equation can be derived:
\[
\sum_{k=0}^{\infty} \delta^k \cdot \int_0^{\infty} e^{-c\mu t} \frac{(c\mu t)^k}{k!} \, dF(t) = f(c\mu(1-\delta)) = \delta. \quad (30)
\]

Based on the results in [49], (30) can be solved, and the solution is exclusively in the range of \((0, 1)\). When the solution of (30) is denoted as \(\delta(0 < \delta < 1)\), the probability \(p_j^{-}\) is calculated as follows:

\[
p_j^- = K^* \cdot \delta^{j-j}, \quad j = c-1, c, c+1, \cdots, \quad (31)
\]

where \(K^*\) is a constant. When \(j = c-1\), \(K^*\) is calculated as follows:

\[
K^* = \delta \cdot p_{c-1}^-.
\]

Considering \(c = 1\), i.e., the GI/M/1/\(\infty\) queuing system is implemented, \(K^*\) can be simply denoted by \(K^* = \delta \cdot p_0^-\). Furthermore, the probability \(p_j^-\) is derived as follows:

\[
p_j^- = p_0^- \cdot \delta^j, \quad j = 0, 1, 2, \cdots, \quad (33)
\]

Considering \(\sum_{j=0}^{\infty} p_j^- = 1\), the probability of an empty queue is calculated by \(p_0^- = 1 - \delta\). By substituting the value of \(p_0^-\) into (33), the probability \(p_j^-\) can be calculated as follows:

\[
p_j^- = (1-\delta) \cdot \delta^j, \quad j = 0, 1, 2, \cdots. \quad (34)
\]

When \(c\) is configured as \(c = 1\), \(K^*\) is calculated as follows:

\[
K^* = \delta \cdot (1-\delta). \quad (35)
\]

In this paper the message is assumed to be serviced by the first-in-first-out (FIFO) scheme in the GI/M/c/\(\infty\) queuing system. The waiting time of the \(m - th\) message is denoted by \(W_{qm}\) in the GI/M/c/\(\infty\) queuing system. When \(\frac{\lambda}{c\mu} < 1\), the distribution of \(W_{qm}\) is expressed as \(P\{W_{qm} \leq t\} = W_{qm}(t)\). When the queuing length is \(N^-_m = j\) and \(j < c\), the \(m - th\) message can be directly handled without waiting. When \(j \geq c\), the \(m - th\) message must wait until the services for \(j-c+1\) messages have been completed in the queuing system. Based on the total probability theorem, the waiting time of the \(m - th\) message is expressed as follows:

\[
W_{qm}(t) = \sum_{j=0}^{c-1} P\{N^-_m = j\} \cdot 0
+ \sum_{j=c}^{\infty} P\{N^-_m = j\} \cdot \int_0^t e^{-c\mu x} \frac{(c\mu x)^j}{(j-c)!} \, dx.
\]

Considering \(\frac{\lambda}{c\mu} < 1\), the limitation of the waiting time of the \(m - th\) message is derived as follows:

\[
\lim_{m \to \infty} W_{qm}(t) = \sum_{j=0}^{c-1} p_j^- \cdot 0
+ \sum_{j=c}^{\infty} p_j^- \cdot \int_0^t e^{-c\mu x} \frac{(c\mu x)^j}{(j-c)!} \, dx
= K^* \sum_{j=c}^{\infty} \delta^{j-c} \cdot \int_0^t e^{-c\mu x} \frac{(c\mu x)^j}{(j-c)!} \, dx
= K^* \frac{1}{\delta} \left[ 1 - e^{-c\mu(1-\delta)t} \right]. \quad (37)
\]

The service time of the \(m - th\) message is denoted by \(\chi_m\) and the distribution of \(\chi_m\) is denoted by \(\chi_m(t)\). Hence, the distribution of the message dwelling time is expressed as

\[
W(t) = \lim_{m \to \infty} (\chi_m(t) + W_{qm}(t)).
\]

Because \(W_{qm}\) and \(\chi_m\) are independent, the distribution of the message dwelling time is derived as follows:

\[
W(t) = 1 - e^{-\mu t} + \frac{K^*}{1-\delta} \left[ 1 - e^{-c\mu(1-\delta)t} \right]. \quad (38)
\]

Thus, Theorem 2 is proven.

APPENDIX C

Given that the density of InP is \(\lambda_{Inp}\), a typical PVT cell area denoted by \(R_{Inp}\), follows a Gamma distribution [47]. The PDF of the InP distribution is expressed as follows:

\[
f_R(\in_p) = \frac{(b\lambda_{Inp})^a}{\Gamma(a)} \cdot e^{-bx\lambda_{Inp}}, \quad (39)
\]

where \(\Gamma(x) = \int_0^\infty \Gamma(x-1)e^{-t}\, dt\) is a Gamma function, \(a\) is the shape parameter, \(b\lambda_{Inp}\) is the inverse scale parameter of the Gamma distribution. The default values of \(a\) and \(b\) are configured as \(a = 3.61\) and \(b = 3.57\) in the simulation analysis, respectively. The road length in a coverage area of an InP is derived as follows:

\[
L_{R_{Inp}}(x) = \int_0^\infty \rho_{road} \frac{(b\lambda_{Inp})^a}{\Gamma(a)} \cdot e^{-bx\lambda_{Inp}} \, dx, \quad (40)
\]

where \(\rho_{road}\) is the road density in an urban environment. Considering the uniform distribution of RSUs along the road, the expected \(N_{RSU}\) is derived as follows:

\[
E(N_{RSU}) = \frac{\int_0^\infty \rho_{road} \frac{(b\lambda_{Inp})^a}{\Gamma(a)} \cdot e^{-bx\lambda_{Inp}} \, dx}{\rho_{RSU}}, \quad (41)
\]

REFERENCES

[1] NGMN Alliance, “5G White Paper,” version 1, Feb. 2015.
[2] K. Yang, S. Ou, H-H Chen, “On effective offloading services for resource-constrained mobile devices running heavier mobile Internet applications,” IEEE Communications Magazine, vol. 46, no. 1, Jan. 2008.
[3] X. Ge, Z. Li, S. Li, “5G Software Defined Vehicular Networks,” IEEE Communications Magazine, vol. 55, no. 7, pp. 87–93, July 2017.
[4] K. Liu, J. Ng, V. Lee, et al., “Cooperative Data Scheduling in Hybrid Vehicular Ad Hoc Networks: VANET as a Software Defined Network,” IEEE/ACM Transactions on Networks, vol. 24, no. 3, June 2016, pp. 1759–1773.
[5] Zaki Y, “LTE mobile network virtualization”, Mobile Networks & Applications, vol. 16, no. 4, pp. 424–432, Aug 2011.
I. Safiulin, S. Schwarz and M. Rupp, “Multicast Beamforming Capabilities in the Physical Layer,” IEEE Transactions on Vehicular Technology, vol. 65, no. 5, pp. 3844–3846, Oct. 2016.

T. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, “Cloudlets for Network and Service Management,” vol. 13, no. 2, pp. 65–77, Feb. 2016.

R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed and T. Ahmed, “Scheduling Wireless Virtual Networks Functions,” IEEE Transactions on Network and Service Management, vol. 13, no. 2, pp. 240–252, June 2016.

K. Samdanis, X. Costa-Perez and V. Sciancalepore, “From network slicing to multi-tenancy: The 5G network slice,” IEEE Communications Magazine, vol. 54, no. 5, pp. 94–100, May 2017.

R. Zan, S. Zhang, P. Yang, O. Alhussein, W. Zhuang and X. S. Shen, “Software Defined Space-Air-Ground Integrated Vehicular Networks: Challenges and Solutions,” IEEE Communications Magazine, vol. 55, no. 7, pp. 101–109, 2017.

Valkeila E., “An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure, 2nd Edition by Daryl J. Daley, David Vere-Jones”. International Statistical Review, vol. 76, no. 2, pp. 316–316, 2018.
Xiaohu Ge (M'09-SM'11) received the Ph.D. degree in communication and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003. Since 2005, he has been with HUST, where he is currently a Full Professor with the School of Electronic Information and Communications. He is an Adjunct Professor with Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He is the Director of the China International Joint Research Center of Green Communications and Networking.