# DRL-Based V2V Computation Offloading for Blockchain-Enabled Vehicular Networks

Jinming Shi, *Student Member, IEEE,* Jun Du, *Member, IEEE,* Yuan Shen, *Member, IEEE,* Jian Wang, *Senior Member, IEEE,* Jian Yuan, *Member, IEEE* and Zhu Han, *Felllow, IEEE*

**Abstract**—Vehicular edge computing (VEC) is an effective method to increase computational capability of vehicles, where vehicles share their idle computing resource among each other. However, due to high vehicle mobility, it is challenging to design an optimal task allocation policy that adapts to the dynamic vehicular environment. Further, vehicular computation offloading often occurs between unfamiliar vehicles, how to motivate vehicles to share their computing resource while guaranteeing the reliability of resource allocation in task offloading is one main challenge. In this paper, we propose a blockchain-enabled VEC framework to ensure the reliability and efficiency of vehicle-to-vehicle (V2V) task offloading. Specifically, we develop a deep reinforcement learning (DRL)-based computation offloading scheme in smart contract of blockchain, where task vehicles can offload part of computation-intensive tasks to neighboring service vehicles. To ensure the security and reliability in task offloading, we evaluate the reliability of vehicles in resource allocation by blockchain. Moreover, we propose an enhanced consensus algorithm based on practical Byzantine fault tolerance (PBFT), and design a DRL-based consensus nodes selection algorithm to improve the efficiency of blockchain consensus and motivate base station to improve reliability in task allocation. Simulation results validate the effectiveness of our proposed scheme for blockchain-enabled VEC.

**Index Terms**—Vehicular edge computing (VEC), computation offloading, blockchain, deep reinforcement learning (DRL).

## 1 Introduction

With the development of autonomous driving and Internet of Vehicles (IoV) technologies, various vehicular applications have been emerging to improve the vehicle security and the driving experience. However, because of the limited onboard computing and storage resource, some computation-intensive tasks cannot be performed within the deadlines, which need the assistance of central cloud or edge computing [1], [2]. Due to the high transmission delay between vehicles and central cloud, vehicular edge computing (VEC) is commonly exploited to improve the computational capability of vehicles, where vehicles can transmit computational tasks to either base station (BS) or neighboring vehicles. Considering that with the increase of the traffic density in the coverage of a BS, limited computational capability in the BS cannot satisfy all the vehicular computational requirements at the same time. Vehicle-to-vehicle (V2V) computation offloading is becoming a promising mechanism that can alleviate the workload of BS effectively, where task vehicles with limited computing resource can offload some computational tasks to neighboring vehicles with idle computing resource [3]. Moreover, comparing to the fixed-position BS, the link duration between two vehicles driving in the same direction is longer because the relative velocity between the two vehicles is smaller.

However, most studies on VEC aimed at the computing resource allocation and the incentive of sharing computing resource among vehicles, the security and trust issues in VEC were seldom considered. Due to the high vehicle mobility, V2V computation offloading always occurs between unfamiliar vehicles. Service vehicles may not allocate sufficient computing resource for task vehicles, so that the offloading tasks may not be completed within the maximum tolerable delay. Furthermore, in the process of V2V computation offloading, some traditional methods, such as game theory [4] and auction mechanism [5], [6], can only optimize the relationship between resource assignment and service price, the reliability of resource allocation in service vehicles is hardly guaranteed. Therefore, it is necessary for VEC to establish a mechanism that ensures the security and reliability in the process of vehicular computation offloading, and meanwhile motivates vehicles to share more computing resource to reduce the offloading delay and improve the completion ratio of vehicular offloading tasks.

To meet the requirements of security and reliability in computation offloading, some works [7], [8], [9] have applied blockchain in edge computing. In the scenario of VEC, blockchain can effectively improve the security and reliability of V2V computation offloading [10], where transactions in terms of vehicular computation offloading are recorded in a decentralized, transparent and immutable manner [11], [12]. Moreover, blockchain and VEC have the same distributed network framework and the same functions of storage and computation, the integration of blockchain and VEC can play their respective advantages [13], [14]. It is worth noting that in public blockchain, all distributed nodes need to participate in the consensus process, and it takes a long time in block generation and verification, which is not suitable for high dynamic vehicular networks. As a result, some works [15], [16] employed consortium blockchain in

• *J. Shi, J. Du, Y. Shen, J. Wang, and J. Yuan are with the Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China. E-mail: shijinmings@163.com, blgdujun@gmail.com, {shenyuan_ee, jianwang, jyuan}@tsinghua.edu.cn.*

• *Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea. E-mail: zhan2@uh.edu.*

the VEC system, where only a part of BSs are authorized to participate in the consensus process of blockchain. To improve the efficiency of consensus, some works [17], [18] proposed non-compute-intensive algorithms based on practical Byzantine fault tolerance (PBFT), which are suitable for lightweight IoV-oriented blockchains.

Furthermore, there are still some challenges to be solved in blockchain-enabled VEC. First, due to the dynamic vehicular environment, it is challenging to establish an efficient mechanism for vehicular task offloading with the consideration of vehicle mobility and heterogeneous vehicular computational capability. Second, how to motivate vehicles to contribute their onboard computing resource and meanwhile guarantee the reliability of onboard resource allocation is also a problem that should be carefully investigated. Third, the performance of PBFT-based consensus protocol in consortium blockchain depends on the number of consensus nodes and the block size. Therefore, how to design a scheme that jointly optimizes the consensus nodes selection and block size with the aim of improving the performance of blockchain is also a challenge.

Motivated by the challenges mentioned above, we design a VEC framework based on a lightweight blockchain, which makes V2V computation offloading performed in a secure, reliable, and efficient manner. Besides, considering the dynamic vehicular environment, the policy of vehicular task offloading should adapt to the state of vehicular network, we therefore utilize deep reinforcement learning (DRL) in vehicular task allocation, where the policy of task offloading is updated in real time by observing the state of vehicular network. The main contributions of this paper are summarized as follows:

- We propose a blockchain-enabled VEC framework, where BSs are utilized to maintain a consortium blockchain. The blockchain is responsible for evaluating the reliability of vehicles according to historical transactions and further motivate vehicles to make appropriate resource allocation. Moreover, the smart contract in blockchain facilitates the sharing of vehicular computing resource on demand by automatically running the vehicular computation offloading algorithm, and guarantees the security and reliability of task offloading.
- We develop a smart contract-based vehicular task allocation scheme, where dynamic pricing and DRL are utilized to motivate vehicles to contribute their idle computing resource and improve the efficiency and reliability of V2V task offloading. Furthermore, to make the proposed scheme adapt to the dynamic vehicular environment, the vehicular computational capability, the V2V link state, and the reliability of service vehicles are all considered in task allocation.
- Since the reliability of service vehicles is evaluated by previous task offloading events recorded in the transactions, and the accuracy of the estimated reliability depends on the timeliness of the transactions, we design a PBFT-based consensus, and develop a DRL-based algorithm to improve the efficiency of transaction verification by jointly optimizing consensus node selection and block size, and meanwhile motivate BSs to improve reliability in task allocation.

The remainder of the paper is organized as follows. The related works are presented in Section 2. In Section 3, we present the architecture of the blockchain-enabled VEC. Section 4 describes the V2V computation offloading algorithm in smart contract. The consensus node selection in blockchain is detailed in Section 5. Finally, we present the simulation results and conclusion in Section 6 and Section 7, respectively.

## 2 Related Work

### 2.1 Vehicular Edge Computing

Some works have investigated vehicular computation offloading in VEC. In [19], a workflow was introduced to support the autonomous organization of VEC, and the task offloading problem was solved by a modified Ant Colony Optimization algorithm. In [5], the problem of computation offloading in VEC was formulated as an auction-based graph job allocation problem and solved by a structure-preserved matching algorithm. Moreover, by considering the constraints of vehicle mobility and task delay in VEC, a DRL-based resource allocation framework with multi-timescale was proposed in [20]. Besides, authors in [1] designed an imitation learning enabled online task scheduling algorithm for VEC. In [21], computational tasks with inter-dependency were executed in different vehicles to minimize the overall response time, and a cooperative task scheduling scheme that considered the instability and heterogeneity of VEC was developed. In order to motivate vehicles to contribute their idle computing resource in VEC, authors in [22] designed a contract-based incentive mechanism that combined resource contribution and resource utilization, while in [23], a timeliness-aware incentive mechanism was proposed to stimulate the participation of vehicles with the consideration of vehicle's uncertain travel time. In addition, a partial offloading and adaptive task scheduling framework was constructed for vehicular networks in [4].

All of the works mentioned above treat the service providers as trustworthy devices and do not consider the security and privacy of task offloading in VEC. In this work, we exploit blockchain technology to guarantee the security and reliability in vehicular computation offloading.

### 2.2 Blockchain for Vehicular Networks

Recently, some works have investigated the integration of blockchain and vehicular networks, and some methods have been proposed for computation offloading in blockchain-enabled vehicular networks. In [24], blockchain and smart contract were employed to facilitate fair task offloading and mitigate security attacks in VEC. In [15], a lightweight blockchain-based resource sharing scheme was proposed and a DRL-based smart contract scheme was developed to match the supply and demand of computing resource in VEC. Moreover, a secure fog computing paradigm was proposed in [25], where RSUs were used to allocate computational tasks for nearby fog vehicles based on repute scores maintained by a semi-private consortium blockchain. In [26], consortium blockchain was employed in a vehicular computing resource trading system to guarantee transaction security and privacy protection in a distributed manner. Furthermore, a resource transaction architecture based on

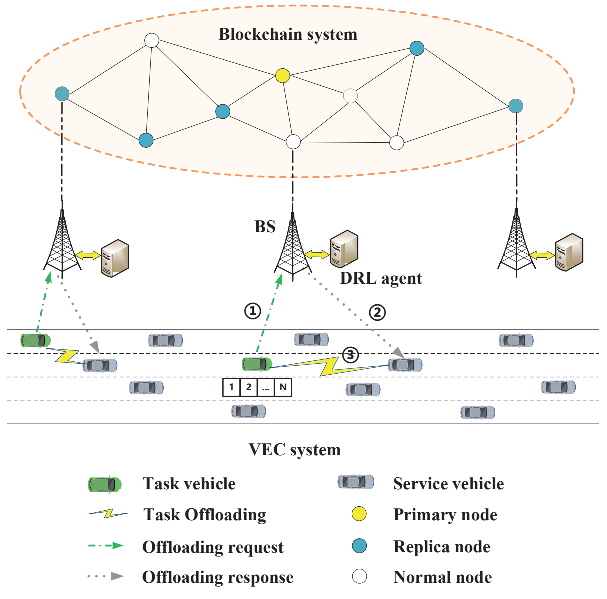


Fig. 1. Blockchain-enabled VEC system architecture.

blockchain was designed in [27], and a double auction mechanism was proposed to maximize the total satisfaction of vehicles and service providers in the form of the smart contract. In addition, authors in [18] exploited Byzantine fault tolerance based Proof-of-Stake consensus to provide fast and deterministic block finalization, and provided a contract-based incentive mechanism to motivate vehicles to share their computing resource.

However, few of the works mentioned above investigated the dynamic trust management for both BSs and vehicles in VEC. In this paper, we propose a blockchain-enabled VEC framework that considers the reliability of both BSs and service vehicles in resource allocation. We further develop a DRL-based algorithm combined with dynamic pricing for vehicular computation offloading, and design an enhanced consensus protocol to improve the performance of blockchain.

## 3 Blockchain-Enabled VEC

In this section, we will demonstrate the architecture of our proposed blockchain-enabled VEC system in Section 3.1 and Section 3.2, and further describe our proposed smart contract for vehicular computation offloading and the consensus scheme for the blockchain-enabled VEC in Section 3.3 and Section 3.4, respectively.

### 3.1 VEC System Model

As shown in Fig. 1, we consider a VEC scenario where multiple BSs are distributed alongside the roads, and each BS is equipped with an edge server that is responsible for computing resource allocation among vehicles. When a vehicle enters the communication range of a BS and is willing to participate in the VEC, it first sends a registration request message that includes the ID, current position, velocity, and available onboard computing resource to the BS. When a vehicle cannot execute all of its local tasks due to the limited onboard computing resource, it can offload its computational tasks to either BSs or neighboring vehicles that drive in the same direction. Considering that with the number of task offloading requests increasing, the BS cannot handle a large number of tasks simultaneously due to the limited computational capability, some computational tasks should be offloaded to neighboring vehicles with idle computing resource. In this work, we will mainly focus on the problem of V2V computation offloading in VEC.

In the VEC system, BSs are responsible for collecting vehicular computing service requests, observing the states of the vehicles inside the communication range, and allocating vehicular computational tasks to vehicles with idle computing resource. If a vehicle needs to offload part of its computational tasks due to the limited onboard computing resource, it first sends a service request to the nearby BS, and we call this vehicle as the task vehicle, and call the vehicles in the communication range of the task vehicle as the service vehicles. Then, the BS performs the vehicular task allocation and selects some of the service vehicles to execute the computational tasks from the task vehicle. Moreover, to motivate service vehicles to contribute their idle computing resource, the task vehicle pays a service price for each offloading task, and the selected service vehicle allocates part of its computing resource for the offloading task according to the service price. Specially, the selection of service vehicles and the service price are dynamically determined by the agent of the BS according to the states of vehicles inside the communication range of the BS.

### 3.2 Blockchain Model

Due to the high dynamic vehicular environment, a task vehicle driving on the road may have different neighboring service vehicles in different time slots, and the task vehicle may not obtain much information from all of the neighboring vehicles in real time, which makes it difficult to evaluate the willingness of neighboring vehicles to contribute their computing resource. In certain cases, a service vehicle agrees to accept the service request from the task vehicle, but the computing resource that the service vehicle allocates for the offloading task cannot ensure the task completed within the deadline, which lead to an offloading failure. Therefore, it is necessary to construct an authentication mechanism for vehicles in VEC. Blockchain is deemed as a promising solution, where the historical transactions of vehicular task offloading are stored and utilized to evaluate the reliability of vehicles in resource allocation. Additionally, considering that public blockchain requires massive computation and time cost in the consensus process, which is not suitable for vehicular networks, we thus employ a lightweight consortium blockchain in the VEC system. Moreover, since the position of BSs are fixed and BSs are equipped with higher computational capability than vehicles, we select BSs as the blockchain nodes.

In the blockchain-enabled VEC, BSs are responsible for maintaining the blockchain, and recording the transactions in terms of vehicular task offloading. The transactions recorded by BSs are grouped and stored in a persistent, immutable and tamper-proof ledger, and are verified by consensus process before attaching to the blockchain. Furthermore, the traceability of V2V computation offloading in blockchain make it possible for guaranteeing the security and reliability of vehicles and BSs. Specifically, when a BS performs the vehicular task allocation, it first check the previous records to evaluate the reliability of services vehicles, and then selects the service vehicle and service price by a DRL-based task allocation algorithm in the smart contract. In addition, to motivate BSs to perform vehicular task allocation appropriately, we also evaluate the reliability of BSs according to the performance in vehicular task allocation, and regard the reliability of BSs as a basis in the consensus node selection.

## 3.3 Smart Contract

In the blockchain-enabled VEC, smart contract is a predefined script [10], which is utilized to perform the computational task allocation among vehicles and record the events of task offloading automatically. In order to incentivize vehicles to share their idle computing resource, a dynamic pricing scheme is exploited in the vehicular task offloading, where service vehicles allocate their idle computing service according to the received service price. Furthermore, a DRL-based vehicular task allocation algorithm is developed to dynamically select the service vehicle and determine the service price for each offloading task from task vehicles. Once a task vehicle sends a task offloading request to the BS, the smart contract embedded in the blockchain is triggered. The dynamic vehicular task allocation is then automatically performed by the BS. After the task completed by the service vehicle, the BS verifies the execution result. If the result is valid and the offloading delay does not exceed the deadline of the task, the task vehicle sends service price to the service vehicle, and a transaction that records the task offloading event is generated in the blockchain. The detailed algorithm of the smart contract-based vehicular task allocation is presented in Section 4.

## 3.4 Consensus Process

After each offloading task completed, the offloading information is recorded as a transaction in the blockchain system, which includes the IDs of task vehicle and service vehicle, the task profile, the offloading delay, the execution result of the task, and the timestamp. In each round of blockchain consensus, newly generated transactions are verified by blockchain nodes and permanently stored in blocks, and the verified transactions can be used for evaluating the reliability of service vehicles and BSs. Considering that the accuracy of the estimated reliability depends on the timeliness of the transactions, the delay of transaction verification should be optimized in the blockchain consensus. We therefore propose a consensus algorithm to improve the efficiency of transaction verification. In the following, we will demonstrate how to implement the consensus in the blockchain system.

Due to the high dynamic of vehicular network, the consensus delay of the blockchain should be short enough so that the information of vehicular computation offloading can be updated in time. Some existing consensus protocols, such as Proof-of-Work (PoW), Proof-of-Stake (PoS), suffer from long time in consensus process, which is not suitable for the VEC scenario. Therefore, in the consensus of the blockchain-enabled VEC, we employ the PBFT protocol [28], [29], which has been widely applied in consortium blockchains and can achieve less consensus latency than PoW and PoS. We then propose an enhanced consensus scheme based on PBFT to optimize the performance of blockchain consensus. In the consensus scheme, all the consensus nodes are selected from the BSs, and the consensus nodes contain one primary node and $N_c - 1$ replica nodes. The primary node and replica node are responsible for block generation and block verification, respectively. Additionally, the blockchain consensus employs signature and message authentication code (MAC) to ensure the integrity and authentication of a transaction, and we assume that signing a block or transaction, verifying a signature, generating a MAC, and verifying a MAC require $a$, $b$, $c$, $c$ CPU cycles, respectively. Moreover, to guarantee the security of the blockchain system, the consensus allows at most $f = \lfloor (N_c - 1)/3 \rfloor$ consensus nodes are faulty [29]. Finally, the main steps of the consensus are shown as follows.

### *3.4.1 Consensus Nodes Selection*

In our blockchain-enabled VEC system, the available computing resource for consensus process in the BS and the reliability of the BS in vehicular task allocation are the two main factors that affect the consensus nodes selection. In other words, the BS with higher reliability and more computing resource will have higher probability of being selected as a consensus node. In each round of consensus, each consensus node will gain a reward for the contribution in the blockchain consensus. Therefore, in order to gain more reward from the blockchain, each BS tends to improve their reliability in vehicular task allocation, which prevents some malicious BSs from giving inappropriate or unfair policy in vehicular task allocation. The detailed algorithm of consensus nodes selection is presented in Section 5.

### *3.4.2 Collect*

At the beginning of a block interval, the primary node first collects unverified transactions from all BSs and sorts the transactions by timestamps. Then, the primary node forms the first $M_t$ transactions as an unvalidated block, and the size of the block is denoted as $S_{bk}$. We assume that the average size of transactions is $\chi$, then we have $M_t = \lfloor S_{bk}/\chi \rfloor$. Then, the primary node verifies $M_t$ signatures and $M_t$ MACs from the $M_t$ transactions. The computation cost of primary node is $M_t(b + c)$.

### *3.4.3 Pre-prepare*

In this step, the primary node generates one signature and one MAC for the unvalidated block, and generates $N_c - 1$ MACs for the pre-prepare message. The pre-prepare message that contains the block is then sent to other replica nodes. Each Replica node verifies one MAC from the block, and $M_t$ MACs and $M_t$ signatures from the transactions in the block. Then, the computation cost of the primary node is $a + N_c c$, and the computation cost of each replica node is $c + M_t(b + c)$. In addition, without loss of generality, we assume that the transmission time of a block is proportional to block size. Then, the time of message delivery is $\tau_{bk} S_{bk}$.

### *3.4.4 Prepare*

Once verifying the pre-prepare message, replica nodes send the prepare message to other consensus nodes. After each replica node collects $2f$ prepare messages matched with the pre-prepare message, the consensus enters the next step. In this step, the primary node needs to verify $2f$ MACs, every replica node needs to generate $N_c - 1$ MACs and verifies $2f$ MACs. Therefore, the computation cost of the primary node is $2fc$, and the computation cost of each replica node is $(N_c - 1 + 2f)c$. The time of message delivery is $\tau_{bk} S_{bk}$.

#### *3.4.5 Commit*

Upon receipt of $2f$ matching prepare messages, each consensus node sends commit message to the other consensus nodes. In this step, the primary node and each replica node verify $2f$ MACs and generate $N_c - 1$ MACs. The computation cost of the primary node and each replica node is $(N_c - 1 + 2f)c$. The time of message delivery is $\tau_{bk} S_{bk}$.

#### *3.4.6 Reply*

Upon receipt of $2f$ matching commit messages, replica node accepts the block as valid and appends the verified block to its local ledger, and sends the reply message containing the validated block to other BSs, then each BS updates its global view from the verified block after receiving $f + 1$ matching reply messages. In this step, the primary node and each replica node generate one MAC for the reply message, the computational cost of each node is $c$. The time of message delivery is $\tau_{bk} S_{bk}$.

Finally, the computation time of the primary node is

$$T_p^V = \frac{M_t(b+c) + a + (2N_c + 4f)c}{G_p}, \tag{1}$$

where $G_p$ is the available computing resource in the primary node. Similarly, the computation time of a replica node is

$$T_r^V = \frac{M_t(b+c) + (2N_c + 4f)c}{G_r}. \tag{2}$$

Similar to [30], we use *time to finality* (TTF) to represent the latency of the consensus process, which is shown as

$$T^F = T^I + T^D + T^V, \tag{3}$$

where $T^I$, $T^D$, and $T^V$ denote the block interval, time of message delivery, and validation time, respectively. According to the above analysis, we have $T^D = 4\tau_{bk} S_{bk}$, and $T^V = \max\{T_p^V, T_r^V\}$. Moreover, in order to prevent new generated blocks suffering from long time of verification, the TTF should satisfy the following constraint:

$$T^F \leq \delta T^I, \tag{4}$$

where $\delta > 1$ is a constant.

### 3.5 Process of Blockchain-Enabled VEC

The main process of our proposed blockchain-enabled VEC system is presented in Fig. 2. In the VEC system, a task vehicle that has a number of offloading tasks first sends the offloading request to the nearby BS, then the smart contract in the BS is triggered. The BS collects the information of the vehicles surrounding the task vehicle and evaluates the reliability of the vehicles according to the transactions stored in the consortium blockchain, and then selects service vehicle for task offloading. After the V2V computation offloading finished, a transaction that records the result of the computation offloading is generated and verified by the blockchain consensus. In each round of consensus, the consensus nodes are selected from the BSs in the VEC system,

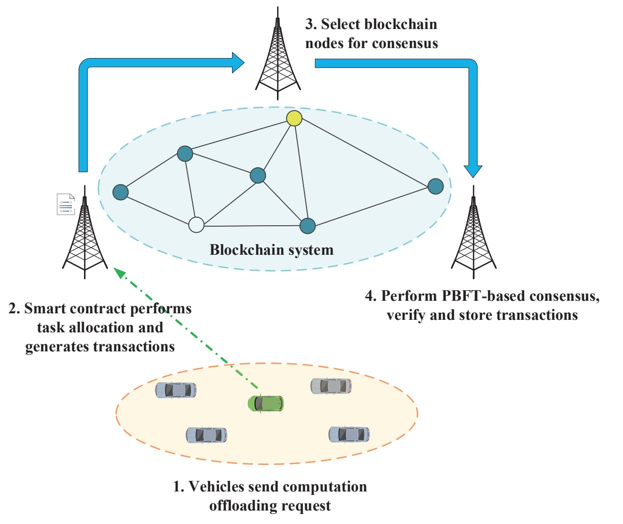

3. Select blockchain nodes for consensus
Blockchain system
2. Smart contract performs task allocation and generates transactions
4. Perform PBFT-based consensus, verify and store transactions
1. Vehicles send computation offloading request

Fig. 2. The main process of the blockchain-enabled VEC system.

and then perform the PBFT-based consensus. Finally, the verified transactions are stored in the blockchain system and can be further used for evaluating the reliability of vehicles and BSs.

## 4 Smart Contract for V2V Computation Offloading

In this section, we will detail our proposed smart contract in terms of vehicular task allocation for the blockchain-enabled VEC system. In the smart contract, the problem of vehicular task allocation in BS is a sequential decision problem, which can be formulated as a Markov decision process (MDP). To solve the problem, we develop a DRL-based vehicular task allocation algorithm to improve the efficiency and the reliability of vehicular computation offloading.

We first consider the VEC scenario illustrated in Fig. 1, where multiple vehicles drive on an one-way road, and several BSs are distributed at the roadside. Similar to [31], We utilize a free flow model to characterize the traffic model in the communication range of a BS, then the relationship between average vehicle velocity $\bar{v}$ and traffic density $\rho$ is

$$\bar{v} = v_{max}\left(1 - \frac{\rho}{\rho_{max}}\right), \tag{5}$$

where $v_{max}$ and $\rho_{max}$ are the maximum vehicle velocity and maximum traffic density, respectively.

### 4.1 Computation Model

We denote the profile of a computational task as $\{D_n, C_n, \tau_n\}$, where $D_n$ is the data size of the task, $C_n$ is the computation size which represents the required CPU cycles for completing the task, $\tau_n$ is the maximum tolerable delay of the task. In many cases, the execution result of a task is much smaller than data size $D_n$, we therefore ignore the transmission time of execution result in the calculation of offloading delay. Then, the offloading delay of a task is

$$t_n = \frac{D_n}{r_{ts}} + \frac{C_n}{f_n}, \tag{6}$$

where $r_{ts}$ is the transmission rate of the V2V link between task vehicle $V_t$ and service vehicle $V_s$, $f_n$ is the allocated computing resource for the task in service vehicle. Similar to [32], we utilize the channel capacity to estimate the V2V transmission rate, which can be given as

$$r_{ts} = W_{ts}\log_2\left(1 + \frac{P_t d_{ts}^{-\alpha} h_{ts}^2}{\sigma_w^2 + I_{ts}}\right), \tag{7}$$

where $W_{ts}$ is the allocated bandwidth of the V2V link between $V_t$ and $V_s$, $P_t$ is the transmission power of $V_t$, $d_{ts}$ is the distance between $V_t$ and $V_s$, $\alpha$ is a constant which represents the path loss factor, $h_{ts}$ is the channel gain of the V2V link, $\sigma_w^2$ represents the power spectrum density of additive white Gaussian noise, and $I_{ts}$ denotes the interference introduced by other V2V transmissions.

Particularly, due to the short V2V link duration and limited vehicular computational capability, we do not consider the queuing time in the task offloading, i.e., if a V2V link interruption occurs in the process of task offloading, the task offloading will be regraded as a failure, and the utility of the task will be negative as a penalty. Then, the task vehicle needs to resubmit a new service request to BS for the task.

### 4.2 Task Vehicle Model

Due to the limited onboard computing resource, some vehicles can not complete all of its computational tasks within the deadlines, they choose to offload some computational tasks to neighboring vehicles, and we regard the vehicles with limited computing resource as the task vehicles. When a task vehicle offloads a task to a service vehicle, it needs to pay the service vehicle the service price, and we assume that the service price of performing a computational task in a service vehicle is proportional to the computation size of the task, which is represented as $p_n C_n$, where $p_n$ denotes the unit price.

For a computational task $\phi_n$, the utility is related to the completion time of the task, which is shown as

$$U_n = \begin{cases} \log(1+\tau_n - t_n), & t_n \le \tau_n, \\ -\Lambda, & t_n > \tau_n, \end{cases} \tag{8}$$

where $-\Lambda < 0$ is a constant that represents the penalty of offloading failure. Then, the utility of $V_t$ for offloading a task can be given as

$$\mathcal{U}_n = U_n - p_n C_n. \tag{9}$$

### 4.3 Service Vehicle Model

In the process of task offloading, once a service vehicle receives a service request and corresponding service price, the service vehicle will contribute part of its computing resource for the offloading task. We consider a service vehicle $V_s$, and assume that the cost of executing a computational task is proportional to the energy consumption, and denote the computing resource allocated for task $\phi_n$ as $f_n$. Similar to [33], the energy consumption of executing task $\phi_n$ is calculated by

$$E_n = \kappa f_n^3 \frac{C_n}{f_n} = \kappa f_n^2 C_n, \tag{10}$$

where $\kappa > 0$ is a constant. Then, the utility of $V_s$ for executing task $\phi_n$ can be given as

$$U_n^s = p_n C_n - \kappa f_n^2 C_n. \tag{11}$$

Since the utility of a service vehicle must be non-negative, the range of $f_n$ is $[0, \min\{F_s, \sqrt{p_n/\kappa}\}]$, where $F_s$ is the maximum available computing resource in $V_s$. Moreover, it can be seen from (11) that the utility of a service vehicle mainly depends on the unit service price and the allocated computing resource for task $\phi_n$. A service vehicle can increase the utility by allocating little computing resource for task $\phi_n$, but it will reduce the utility of task $\phi_n$ or even cannot complete the task within the deadline. In order to prevent some malicious vehicles from allocating insufficient computing resource for offloading tasks, we further establish a reliability model to evaluate the efficiency of task offloading in service vehicle.

We first define a normalized utility for an offloading task that can be completed within the deadline, which is shown as follows:

$$\tilde{U}_n = \frac{\log(1+\tau_n - t_n)}{\log(1+\tau_n)}, \tag{12}$$

where $\log(1+\tau_n)$ represents the maximum utility of task $\phi_n$. After a service vehicle completes an offloading task, the computation efficiency of the service vehicle is updated by

$$\xi_s = (1-\omega_1)\xi_s' + \omega_1 \tilde{U}_n, \quad \omega_1 \in (0,1), \tag{13}$$

where $\xi_s'$ represents the previous computation efficiency of $V_s$. Besides, the completion ratio of a service vehicle is updated by

$$\zeta_s = \frac{N_s * \zeta_s' + \mathbb{1}(t_n \le \tau_n)}{N_s + 1}, \tag{14}$$

where $N_s$ represents the total number of received offloading tasks in $V_s$, $\zeta_s'$ denotes the previous completion ratio of $V_s$, and $\mathbb{1}(\cdot)$ is the indicator function. Then, the reliability of a service vehicle can be defined as

$$\eta_s = \omega_2 \xi_s + (1-\omega_2)\zeta_s, \quad \omega_2 \in (0,1). \tag{15}$$

As a result, if a service vehicle cannot allocate appropriate computing resource for an offloading task according to the service price, the reliability of the service vehicle will be reduced. In our proposed V2V computation offloading algorithm, the reliability of service vehicle is an important factor that affects the selection of service vehicles and service price, a vehicle with low reliability usually has a low probability of being selected as the service vehicle for offloading tasks.

### 4.4 Problem Formulation: Vehicular Task Allocation

In the VEC system, we consider a task vehicle which is denoted as $V_t$. The service vehicles in the communication range of $V_t$ are denoted as $\mathcal{S} = \{V_1, ..., V_s, ..., V_S\}$. When $V_t$ sends a task offloading request to a BS, the smart contract in the BS is automatically triggered. We assume that there are $N$ offloading tasks from $V_t$ in a period, which are denoted as $\mathcal{N} = \{\phi_1, ..., \phi_n, ..., \phi_N\}$. For each offloading task, the BS dynamically chooses the service vehicle and determines corresponding service price by observing the vehicular environment. After the task completed by the service vehicle, the execution result is transmitted to $V_t$ and the BS, and then the BS verifies the result and sends the validity of the result to $V_t$. If the execution result is valid and the offloading delay is within the maximum tolerable delay of the task, $V_t$ will send service price to the service vehicle. Otherwise, the service vehicle will not obtain any payment from $V_t$. Then, the BS generates a transaction that records the task offloading event. In the V2V computation offloading, we formulate the problem of vehicular task allocation in the BS as an optimization problem with the objective of maximizing the utility of $V_t$ with the consideration of the mobility, the computational capability, and the reliability of vehicles. The optimization problem is formulated as follows:

$$\begin{aligned} \max \quad & \frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{S} x_n^s \mathcal{U}_n, \\ \text{s.t.} \quad & C1: x_n^s \in \{0,1\}, \quad \forall n \in \mathcal{N}, s \in \mathcal{S}, \\ & C2: \sum_{s=1}^{S} x_n^s = 1, \quad \forall n \in \mathcal{N}, \\ & C3: x_n^s(l_s - t_n) \geq 0, \quad \forall n \in \mathcal{N}, s \in \mathcal{S}, \\ & C4: x_n^s(\eta_s - \eta_0) \geq 0, \quad \forall n \in \mathcal{N}, s \in \mathcal{S}. \end{aligned} \tag{16}$$

In constraint $C1$, $x_n^s = 1$ means that task $\phi_n$ is offloaded to $V_s$, and $x_n^s = 0$ otherwise. Constraint $C2$ guarantees that a computational task is only offloaded to one service vehicle. Constraint $C3$ ensures that the completion time of an offloading task should be less than the V2V link duration between $V_t$ and $V_s$. Constraint $C4$ ensures that the reliability of the selected service vehicle is higher than threshold $\eta_0$.

In dynamic vehicular network, the V2V link state, the idle computing resource, and the reliability of service vehicles are all time-variant, and the multi-task allocation can be regarded as a sequential decision problem. We reformulate the optimization problem in (16) as an MDP, and the system state space, action space and reward are defined as follows.

#### 4.4.1 *State Space*

The system state at time $t$ can be defined as

$$\begin{aligned} \mathbf{s}_t = [&F_1(t), ..., F_s(t), ..., F_S(t); \gamma_1(t), ..., \gamma_s(t), ..., \gamma_S(t); \\ &l_1(t), ..., l_s(t), ..., l_S(t); \eta_1(t), ..., \eta_s(t), ..., \eta_S(t); \\ &D(t), C(t), \tau(t)], \end{aligned} \tag{17}$$

where $F_s(t)$ represents the available computing resource in $V_s$, $\gamma_s(t)$ is the signal-to-noise ratio (SNR) of the V2V link between $V_t$ and $V_s$, $l_s(t)$ denotes the estimated V2V link duration between $V_t$ and $V_s$, $\eta_s(t)$ is the reliability of $V_s$, $D(t)$, $C(t)$, and $\tau(t)$ are the data size, computation size, and deadline of the offloading task at time $t$, respectively.

#### 4.4.2 *Action Space*

The conducted action at time $t$ can be given as

$$\mathbf{a}_t = [x_n^1(t), ..., x_n^s(t), ..., x_n^S(t); p_n^1(t), ..., p_n^s(t), ..., p_n^S(t)], \tag{18}$$

where $x_n^s(t) = 1$ represents that service vehicle $V_s$ is selected to execute task $\phi_n$, and $x_n^s(t) = 0$ otherwise, and $p_n^s(t)$ denotes the unit service price of $\phi_n$ paid to $V_s$.

#### 4.4.3 *Reward Function*

According to the optimization objective presented in (16), we define the immediate reward after conducting action $\mathbf{a}_t$ at time $t$ as

$$R_t = \sum_{s=1}^{S} x_n^s \mathcal{U}_n, \tag{19}$$

where $\mathcal{U}_n$ is the utility of $V_t$ for offloading a task as defined in (9), which is related to service vehicle and service price.

### 4.5 SAC-Based Computation Offloading Algorithm

In this part, we propose an algorithm based on soft actor-critic (SAC) [34] to solve the optimization problem of vehicular task allocation. SAC is a reinforcement learning algorithm which is based on off-policy actor-critic model. To be specific, the actor network in SAC is used to generate action according to the observed system state and improve the policy, while the critic network is responsible for evaluating the policy provided by the actor network. In the following, we will demonstrate how SAC works in task allocation.

#### 4.5.1 *Policy and Value Function*

The main goal of SAC is to find an optimal policy that maximizes the expected long-term reward, and meanwhile maximizes the entropy of policy to improve the robustness and exploration. The policy is defined as an action-selection strategy which can be either deterministic or stochastic. In SAC, the policy is stochastic and is denoted as $\pi(\mathbf{a}|\mathbf{s})$, which represents a probability distribution over actions given an observed state $\mathbf{s}$. We then define a soft action-value function that represents the expected long-term reward combined with the expected entropy of policy $\pi$ when given an initial state $\mathbf{s}$ and action $\mathbf{a}$, which is shown as

$$Q^\pi(\mathbf{s},\mathbf{a}) = \mathbb{E}\left[\sum_{t=0}^{T-1}\nu^t R_t + \beta\sum_{t=1}^{T-1}\nu^t H(\pi(\cdot|\mathbf{s}_t))|\mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}\right], \tag{20}$$

where $\beta \in [0,1]$ is the temperature parameter that controls the stochasticity of the policy, and $\nu$ is the discount factor. Further, we define a soft state-value function that denotes the expected long-term reward combined with the entropy of policy starting from state $\mathbf{s}$ and taking actions following some policy $\pi$, which is given as follows:

$$V^\pi(\mathbf{s}) = \mathbb{E}\left[\sum_{t=0}^{T-1}\nu^t(R_t + \beta H(\pi(\cdot|\mathbf{s}_t)))|\mathbf{s}_0 = \mathbf{s}\right]. \tag{21}$$

Since both the dimension of state space and action space could be extremely large with the increase of traffic density, we employ neural networks to approximate both the soft action-value function and the policy, which are denoted as $Q_\theta(\mathbf{s},\mathbf{a})$ and $\pi_\psi(\mathbf{a}|\mathbf{s})$, respectively.

#### 4.5.2 *Critic Network Training*

In our algorithm, critic network is in charge of policy evaluation. In each step, the agent of BS samples a batch of experiences from replay buffer $\mathcal{B}$, and then trains the critic network. Specifically, the policy is measured by the mean square error (MSE) between soft action-value function and the target soft action-value function, and the loss function is given as follows:

$$L_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\mathcal{B}}\left[\frac{1}{2}\left(Q_\theta(\mathbf{s}_t,\mathbf{a}_t) - \hat{Q}_{\hat{\theta}}(\mathbf{s}_t,\mathbf{a}_t)\right)^2\right], \tag{22}$$

where the target soft action value function is defined as

$$\hat{Q}_{\hat{\theta}}(\mathbf{s}_t,\mathbf{a}_t) = R_t + \nu\mathbb{E}_{\mathbf{s}_{t+1}\sim\rho_\pi}\left[V_{\hat{\theta}}(\mathbf{s}_{t+1})\right]. \tag{23}$$

Then, the critic network is trained by minimizing the loss function in (22) in each step.

#### 4.5.3 Actor Network Training

The actor network is in charge of generating actions by the observed states. According to [34], the actor network is trained by minimizing the Kullback-Leibler (KL) divergence, which is presented as follows:

$$\pi_t = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}} \left( \pi'(\cdot|\mathbf{s}_t) \parallel \frac{\exp(\frac{1}{\beta} Q^{\pi}(\mathbf{s}_t, \cdot))}{Z^{\pi}(\mathbf{s}_t)} \right), \qquad (24)$$

where $Z^{\pi}(\mathbf{s}_t)$ is the function that normalizes the distribution, which does not have effect on the gradient in terms of the policy. We transform the expression of $\pi_t$ in (24) by multiplying $\beta$ and ignoring the term $Z^{\pi}(\mathbf{s}_t)$, and then obtain the loss function of policy, which is given as follows:

$$L_{\pi}(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{B}} \left[ \mathbb{E}_{\mathbf{a}_t \sim \pi_{\psi}} \left[ \beta \log \pi_{\psi}(\mathbf{a}_t|\mathbf{s}_t) - Q_{\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]. \qquad (25)$$

In each step, the actor network is trained by minimizing the loss function in (25).

#### 4.5.4 Temperature Adjustment

In the above descriptions of training process of actor and critic networks, temperature parameter $\beta$ is regarded as a constant. According to [35], the temperature parameter can be learned in each step by minimizing the loss function

$$L_t(\beta) = \mathbb{E}_{\mathbf{a}_t \sim \pi} \left[ -\beta \log \pi(\mathbf{s}_t, \mathbf{a}_t) - \beta H_0 \right], \qquad (26)$$

where $H_0$ is the entropy threshold.

Moreover, in order to mitigate the positive bias in policy improvement, two critic networks are established in SAC [35], and both of which are trained independently. Then in the calculation of the stochastic gradient of $L_Q(\theta)$ and the policy gradient of $\pi(\mathbf{a}|\mathbf{s})$, we use the minimum soft Q-value of the two critic networks. .

Finally, the detailed vehicular task allocation algorithm is presented in Algorithm 1.

**Algorithm 1** SAC-Based Task Allocation Algorithm

```
Initialize: Initialize critic networks $Q_{\theta_1}(\mathbf{s},\mathbf{a})$ and $Q_{\theta_2}(\mathbf{s},\mathbf{a})$ with weights $\theta_1$ and $\theta_2$, respectively.
Initialize target critic networks $\hat{Q}_{\hat{\theta}_1}(\mathbf{s},\mathbf{a})$ and $\hat{Q}_{\hat{\theta}_2}(\mathbf{s},\mathbf{a})$ with weights $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively.
Initialize actor network $\pi_{\psi}(\mathbf{a}|\mathbf{s})$ with weights $\psi$.
Initialize replay buffer $\mathcal{B} = \emptyset$.
for each period do
  for each task $\phi_n$ do
    The BS collects information of vehicles and evaluate current state $\mathbf{s}_t$.
    Generates action $\mathbf{a}_t$ that determines service vehicle $V_s$ and unit price $p_n$.
    Sends a task allocation message to $V_t$ and $V_s$.
    $V_s$ allocates computing resource for task $\phi_n$.
    The BS calculates the immediate reward $R_t$ and estimates the next state $\mathbf{s}_{t+1}$.
    Store $(\mathbf{s}_t, \mathbf{a}_t, R_t, \mathbf{s}_{t+1})$ into replay buffer $\mathcal{B}$.
    Sample a batch of experiences $\mathcal{E}$ from $\mathcal{B}$.
    Update weights of $Q_{\theta_1}(\mathbf{s},\mathbf{a})$ and $Q_{\theta_2}(\mathbf{s},\mathbf{a})$ by computing the gradient of $L_Q(\theta_i)$ in (22),
        $\theta_i = \theta_i - \lambda_Q \nabla_{\theta_i} \frac{1}{\mathcal{E}} \sum_{\mathcal{E}} L_Q(\theta_i), \quad i = 1, 2.$
    Update weights of $\pi_{\psi}(\mathbf{a}|\mathbf{s})$ by computing the gradient of $L_{\pi}(\psi)$ in (25),
        $\psi = \psi - \lambda_{\pi} \nabla_{\psi} \frac{1}{\mathcal{E}} \sum_{\mathcal{E}} L_{\pi}(\psi).$
    Update $\beta$ by computing the gradient of $L_t(\beta)$ in (26),
        $\beta = \beta - \lambda_{\beta} \nabla_{\beta} \frac{1}{\mathcal{E}} \sum_{\mathcal{E}} L_t(\beta).$
    Update weights of $\hat{Q}_{\hat{\theta}_1}(\mathbf{s},\mathbf{a})$ and $\hat{Q}_{\hat{\theta}_2}(\mathbf{s},\mathbf{a})$ by
        $\hat{\theta}_i = \omega_s \hat{\theta}_i + (1 - \omega_s)\theta_i, \quad i = 1, 2.$
  end for
end for
```

## 5 DRL-Based Consensus Node Selection

In our proposed V2V computation offloading scheme, the reliability of service vehicles in resource allocation is a main factor that affects the decision in vehicular task allocation. In the blockchain-enabled VEC, the reliability of a vehicle in resource allocation is evaluated by the transactions of vehicular computation offloading, and the accuracy of the estimated reliability highly depends on the timeliness of the transactions. Therefore, it is necessary to design a consensus scheme to optimize the efficiency of transaction verification.

In this section, we propose a consensus node selection algorithm for the PBFT-based consensus in the blockchain-enabled VEC system, where the consensus nodes and the block size in each round of blockchain consensus are both determined according to the states of BSs. In order to maximize the efficiency of blockchain consensus and meanwhile motivate BSs to improve their reliability in vehicular task allocation, we model the consensus nodes and incorporate the incentive mechanism into the reward of consensus nodes in Section 5.1. Besides, considering that the available computing resource for consensus in BSs and the reliability of BSs are both time-variant, we employ a DRL-based algorithm in Section 5.3 so that the policy of consensus nodes selection can dynamically adapt to the state of the VEC system.

### 5.1 Consensus Node Model

In the blockchain-enabled VEC system, we assume that there are $B$ BSs distributed in a certain area, which are denoted as $\mathcal{I} = \{I_1, ..., I_b, ..., I_B\}$. In each round of blockchain consensus, all the consensus nodes are selected by the primary node of the previous round of consensus. According to the analysis of blockchain consensus presented in Section 3.4, the performance of consensus mainly depends on the block size and available computing resource in BSs. Considering that with the traffic density in the communication range of a BS increasing, there will be more vehicles sending computing service requests to the BS, and without loss of generality, we assume that the arrival rate of vehicular service requests in a BS is proportional to the traffic density in the communication range of the BS. Then, the available computing resource for blockchain consensus in the BS can be given as

$$G_b = G_b^{total} - (\tilde{G}_b + \mu \rho_b), \qquad (27)$$

where $G_b^{total}$ is the total computational capability of BS $I_b$, $\tilde{G}_b$ represents the computing resource reserved for some non-vehicular computational tasks in $I_b$, $\rho_b$ is the traffic density in the communication range of $I_b$, and $\mu > 0$ is a constant. Following similar assumptions of block reward in [33], we define the reward of a consensus node after completing a round of consensus as

$$R_b(t)=\begin{cases}\left(\varepsilon_1\iota_b+\varepsilon_2\frac{S_{bk}}{\chi}\right)\log(1+\mathcal{T}-T_b^V), & T_b^V\le\mathcal{T},\\ -\Gamma, \quad T_b^V>\mathcal{T},\end{cases} \tag{28}$$

where $\iota_b$ is the reliability of BS $I_b$, $\varepsilon_1$ and $\varepsilon_2$ are the weights of reliability and number of validated transactions, respectively, $T_b^V$ denotes the delay of block verification in $I_b$, $\mathcal{T}=\delta T^I-(T^I+T^D)$ represents the maximum tolerable delay of block verification in $I_b$, and $-\Gamma<0$ is a constant that represents the penalty. It can be seen from (28) that the reward of a consensus node is simultaneously determined by the block size and the reliability, the BS that validates more transactions and has higher reliability will get more reward in the consensus process. Therefore, BSs are motivated to improve their reliability in vehicular task allocation and contribute more computing resource for blockchain consensus. After a round of consensus completed, the average reward obtained by the consensus nodes from blockchain is

$$\mathcal{R}(t)=\frac{1}{N_c}\sum_{b=1}^{N_c}R_b(t). \tag{29}$$

### 5.2 Problem Formulation: Consensus Node Selection

Our aim is to maximize the expected long-term reward of consensus nodes. According to the reward of consensus nodes defined in (29), we formulate the optimization problem as

$$\begin{aligned}\max\quad & \sum_{t'=t}^{T}\epsilon^{t'-t}\mathcal{R}(t'),\\ \text{s.t.}\quad & C1: y_b(t)\in\{0,1\},\quad \forall b\in\mathcal{I},\\ & C2: \sum_{b=1}^{B}y_b(t)=N_c,\\ & C3: y_b(t)(G_b-\mathcal{G})\ge 0,\quad \forall b\in\mathcal{I}\\ & C4: y_b(t)(\iota_b(t)-\mathcal{L})\ge 0,\quad \forall b\in\mathcal{I}\\ & C5: T^F-\delta T^I\le 0,\end{aligned} \tag{30}$$

where $\epsilon\in(0,1]$ is the reward discount factor. In constraint $C1$, $y_b(t)=1$ indicates that BS $I_b$ is selected as a consensus node at time $t$, and $y_b(t)=0$ otherwise. Constraint $C2$ indicates that the total number of consensus nodes is $N_c$. Constraint $C3$ guarantees that the available computing resource in a consensus node should be higher than threshold $\mathcal{G}$. Constraint $C4$ ensures that the reliability of a consensus node should be higher than threshold $\mathcal{L}$. In constraint $C5$, the consensus delay should satisfy the constraint in (4).

In the consensus node selection, the available computing resource for blockchain consensus in each BS depends on the traffic density in the communication range of the BS, the reliability of each BS depends on the performance in vehicular task allocation, and both of which are time-variant. Moreover, the consensus node selection is a joint decision making problem, where the block size and multiple consensus nodes are jointly determined in each round of consensus, which is difficult to be solved by traditional optimization methods. In the following, we will transform the problem as an MDP and demonstrate how to solve the problem by using DRL-based method.

#### 5.2.1 *State Space*

In each period, the system state of all the BSs is denoted as

$$\begin{aligned}\mathbf{s}_t=[&G_1(t),...,G_b(t),...,G_B(t);\\ &\iota_1(t),...,\iota_b(t),...,\iota_B(t);N_{tr}(t)],\end{aligned} \tag{31}$$

where $G_b(t)$ is the available computing resource for blockchain consensus in BS $I_b$ at time $t$, $\iota_b(t)$ represents the reliability of $I_b$ in vehicular task allocation, which is evaluated by the average normalized utility defined in (12) and the completion ratio of the vehicular offloading tasks allocated by the BS, and the calculation of $\iota_b(t)$ is the same as (15). $N_{tr}(t)$ represents the number of transactions generated by BSs in a block interval.

#### 5.2.2 *Action Space*

We represent the action conducted by the agent at time $t$ as

$$\mathbf{a}_t=[y_1(t),...,y_b(t),...,y_B(t);S_{bk}(t)], \tag{32}$$

where $y_b(t)=1$ indicates that BS $I_b$ is selected as a consensus node, and $y_b(t)=0$ otherwise. $S_{bk}(t)$ denotes the size of unverified block, and $S_{bk}(t)\in\{0.2,0.4,...,\dot{S}_{bk}\}$, where $\dot{S}_{bk}$ is the maximum block size in the blockchain. Furthermore, each consensus node has the potential to be selected as the primary node, and the other consensus nodes are the replica nodes. Among the consensus nodes chosen from $\mathbf{a}_t$, we select the consensus node that maximizes $\mathcal{R}(t)$ as the primary node.

#### 5.2.3 *Reward Function*

According to the optimization objective defined in (30), the immediate reward in a round of consensus can be given as

$$R(\mathbf{s}_t,\mathbf{a}_t)=\mathcal{R}(t). \tag{33}$$

### 5.3 Consensus Node Selection Algorithm

Considering that the action in consensus node selection is discrete, we employ double deep Q-network (DDQN) [36] to solve the optimization problem. Since the aim of our algorithm is to find an optimal policy that maximizes the expected long-term reward of the consensus nodes, we first define a Q-value function $Q^\pi(\mathbf{s}_t,\mathbf{a}_t)$ to represent the expected long-term reward after performing action $\mathbf{a}_t$ in state $\mathbf{s}_t$ according to some policy $\pi$, which is shown as

$$Q^\pi(\mathbf{s}_t,\mathbf{a_t})=\mathbb{E}\left[R_t+\nu' Q^\pi(\mathbf{s}_{t+1},\mathbf{a}_{t+1})|\mathbf{s}_t,\mathbf{a}_t\right]. \tag{34}$$

Then, the optimal policy that maximizes the expected long-term reward is represented as

$$\pi(\mathbf{s}_t)=\arg\max_{\mathbf{a}}Q(\mathbf{s}_t,\mathbf{a}). \tag{35}$$

In order to obtain the optimal policy, a value iteration of Q-value function is employed, which can be given as follows:

$$Q(\mathbf{s}_t,\mathbf{a}_t)=Q(\mathbf{s}_t,\mathbf{a}_t)+\beta'\left(R_{t+1}+\nu'\max_{\mathbf{a}}Q(\mathbf{s}_{t+1},\mathbf{a})-Q(\mathbf{s}_t,\mathbf{a}_t)\right). \tag{36}$$

Considering that the dimension of state space and action space could be extremely large, it is difficult to obtain the optimal policy by looking up the table which stores the Q-value of the historical experiences. Therefore, in DDQN, neural network is employed to approximate Q-value function, which is called Q-network and denoted as $Q(\mathbf{s},\mathbf{a};\theta)$, where $\theta$ represents the weights of Q-network. Then, the optimal policy becomes $\pi(\mathbf{s}_t)=\arg\max_{\mathbf{a}} Q(\mathbf{s}_t,\mathbf{a};\theta)$.

In order to mitigate the over estimation in the learning process, two Q-networks are established and trained independently, i.e., the main Q-network and the target Q-network, which are represented as $Q(\mathbf{s},\mathbf{a};\theta)$ and $Q^{-}(\mathbf{s},\mathbf{a};\theta^{-})$, respectively. The main Q-network is responsible for choosing action according to the observed system state, while the target Q-network is used to evaluate the policy obtained from the main Q-network. In the target Q-network, the target Q-value can be calculated as follows:

$$z_t^{-}=R_t+\nu' Q^{-}(\mathbf{s}_{t+1},\arg\max_{\mathbf{a}} Q(\mathbf{s}_{t+1},\mathbf{a};\theta);\theta^{-}). \tag{37}$$

Since our aim is to make the Q-value of the main Q-network close to the target Q-value, the loss function of the main Q-network is defined as follows:

$$J(\theta)=\frac{1}{\mathcal{E}'}\sum_{i=1}^{\mathcal{E}'}\left(z_i^{-}-Q(\mathbf{s}_i,\mathbf{a}_i;\theta)\right)^2. \tag{38}$$

In each period, the agent samples a batch of experiences $\mathcal{E}'$ from the replay buffer, and train the main Q-network by minimizing the loss function in (38). Finally, after every $N^{-}$ steps, the weights of the target Q-network $\theta^{-}$ is updated with an exponential moving average of $\theta$. The detailed algorithm is presented in Algorithm 2.

**Algorithm 2** Consensus Node Selection Algorithm

```
Initialize: Initialize main Q-network $Q(\mathbf{s},\mathbf{a};\theta)$ and target Q-network $Q^{-}(\mathbf{s},\mathbf{a};\theta^{-})$ with random weights $\theta$ and $\theta^{-}$, respectively.
Initialize replay buffer $\mathcal{B}'=\emptyset$.
for each period do
    Primary node evaluates current state $\mathbf{s}_t$ according to the computing resource and reliability of each BS.
    Generate a random probability $p$.
    if $p<\epsilon$, then
        Randomly select action $\mathbf{a}_t$ from the action space.
    else
        Choose action $\mathbf{a}_t=\arg\max_{\mathbf{a}} Q(\mathbf{s}_t,\mathbf{a};\theta)$.
    end if
    Broadcast the message that contains the consensus nodes selection and block size.
    Store $(\mathbf{s}_t,\mathbf{a}_t,R_t,\mathbf{s}_{t+1})$ into replay buffer $\mathcal{B}'$.
    Calculate target Q-value $z_t^{-}$ by (37).
    Update the weights of $Q(\mathbf{s},\mathbf{a};\theta)$ by computing the gradient of $J(\theta)$ in (38), $\theta=\theta-\lambda'\nabla_\theta J(\theta)$.
    Update the weights of $Q^{-}(\mathbf{s},\mathbf{a};\theta^{-})$ every $N^{-}$ steps by $\theta^{-}=\omega_d\theta+(1-\omega_d)\theta^{-}$.
end for
```

## 6 Performance Evaluation

### 6.1 Simulation Setup

We conduct our simulation on a desktop, which has two NVIDIA TITAN Xp GPUs, a 128G RAM and an Intel Xeon CPU. The simulation platform is based on Pytorch with Python 3.7 on Ubuntu 16.04 LTS.

TABLE 1
Simulation Parameters

| Parameter | Value |
|---|---|
| Traffic density $\rho$ (vehicles/km) | $5\sim 40$ |
| Computational capability of vehicle $F_s$ (GHz) | $[3,7]$ |
| Computational capability of BS $G_b^{total}$ (GHz) | $[20,29]$ |
| Communication range of vehicle (m) | 500 |
| Communication range of BS (km) | 3 |
| V2V bandwidth $W_{ts}$ (MHz) | 10 |
| Data size of task $D_n$ (Mbits) | $[0.2,4]$ |
| Computation size of task $C_n$ ($10^9$ cycles) | $[0.2,3.2]$ |
| Maximum delay of task $\tau_n$ (s) | $\{0.5,1,2,4\}$ |
| Average transaction size $\chi$ (KB) | 2 |
| Maximum block size $\dot{S}_{bk}$ (MB) | 8 [37] |
| Cost of signing a block/transaction $a$ (Mcycles) | 2 |
| Cost of verifying a signature $b$ (Mcycles) | 8 [38] |
| Cost of generating/verifying a MAC $c$ (Mcycles) | 0.5 [38] |

In the simulation of V2V computation offloading, we consider an one-way road, where multiple vehicles drive in the same direction, and a BS deployed by the road takes charge of vehicular task allocation. We set the number of offloading tasks of a task vehicle in a period as $32$, and simulate the performance of our proposed algorithm under different traffic densities in the communication range of the BS. Besides, in the simulation of blockchain consensus, we consider an area with $50$ BSs in total, and the traffic density in the communication range of the BSs varies from $5$ vehicles/km to $40$ vehicles/km. We conduct multiple simulations with different number of consensus nodes $N_c$, and the range of $N_c$ is set as $[6,30]$. The other parameters in our simulation are summarized in Table 1.

### 6.2 Performance of V2V Computation Offloading

In the simulation of the vehicular computation offloading algorithm, actor network $\pi_\psi(\mathbf{a}|\mathbf{s})$ and critic networks $Q_{\theta_1}(\mathbf{s},\mathbf{a})$ and $Q_{\theta_2}(\mathbf{s},\mathbf{a})$ are all fully-connected deep neural networks (DNNs) with two hidden layers. We set the size of the hidden layers in $\pi_\psi(\mathbf{a}|\mathbf{s})$ as $(800,500)$, and set the learning rate of $\pi_\psi(\mathbf{a}|\mathbf{s})$ as $\lambda_\pi=8\times10^{-4}$. Similarly, we set the size of the hidden layers in both $Q_{\theta_1}(\mathbf{s},\mathbf{a})$ and $Q_{\theta_2}(\mathbf{s},\mathbf{a})$ as $(800,500)$, and the learning rate of the two networks is set as $\lambda_Q=8\times10^{-4}$. In addition, the batch size is set as $|\mathcal{E}|=256$, and the delay factor is set as $\omega_s=0.005$.

The complexity of the vehicular computation offloading algorithm is commensurate with that of the DNN model. Given a VEC scenario where the number of service vehicles is $S$, then the dimension of the state space is $4S+3$ and the dimension of the action space is $2S$ according to (17) and (18). Then, the complexity of action generation in the algorithm can be given as $O((4K+3)h_1+h_1h_2+2h_2K)$, where $(h1,h2)$ is the size of the hidden layers in both actor network and critic network. In our simulations, the size of hidden layers in actor and critic is fixed, the complexity of action generation can be given as $O(K)$. Besides, in the training process of the algorithm, the complexity in each step is $O(K)$ as well.

To verify the performance of our proposed algorithm, we introduce the following algorithms for comparison:

- Greedy task offloading (GTO): For each offloading task, the BS always chooses the service vehicle with

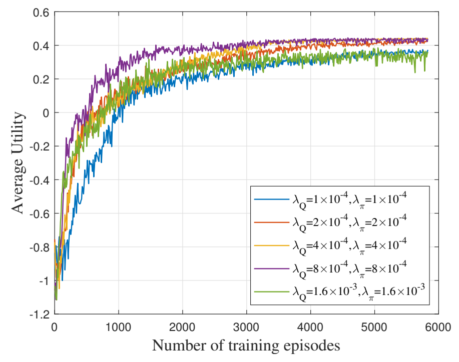


Fig. 3. The average utility of task vehicle for offloading a task with different learning rates.

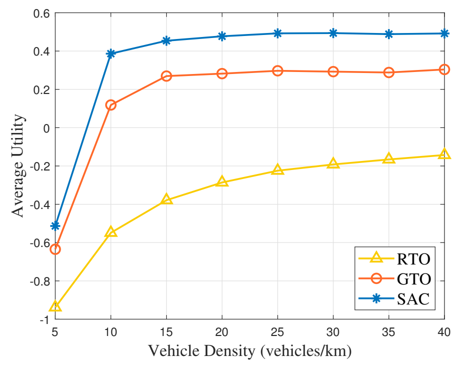


Fig. 4. The average utility of task vehicle for offloading a task under different traffic densities.

the maximum idle computing resource, and gives the service price that maximizes the utility of the task.

- Random task offloading (RTO): For each offloading task, the BS randomly chooses a service vehicle, and randomly gives a service price in the price domain.

Fig. 3 shows the average utility of task vehicle for offloading a task with different learning rates in our proposed algorithm. It can be seen that the average utility reaches convergence around 3,000 training episodes, and the algorithm with learning rates $\lambda_Q = 8 \times 10^{-4}, \lambda_\pi = 8 \times 10^{-4}$ achieves the best performance comparing with the algorithm with other learning rates.

As shown in Fig. 4, the average utility of task vehicle for offloading a task in our proposed algorithm is higher than that in greedy algorithm and random algorithm, because the optimization objective of our proposed algorithm is to maximize the expected long-term reward, and the reward is related to the utility of task vehicle, while in greedy algorithm, the agent always chooses the service vehicle with the maximum idle computing resource in each step, which may not obtain the maximum average utility of offloading tasks. Moreover, greedy algorithm does not consider the reliability of service vehicles in task offloading, some tasks may be offloaded to service vehicles with low reliability and obtain less computing resource, which leads to lower average utility of task vehicle.

In Fig. 5, the completion ratio of offloading tasks by applying the three algorithms is presented. It can be seen that the completion ratio of offloading tasks by applying

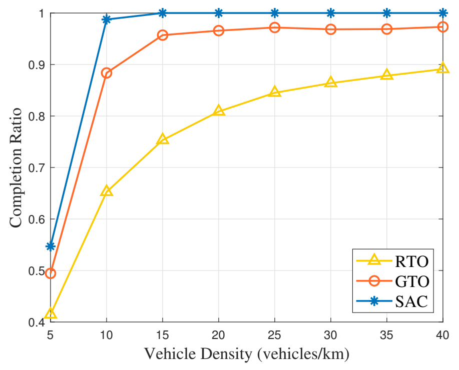
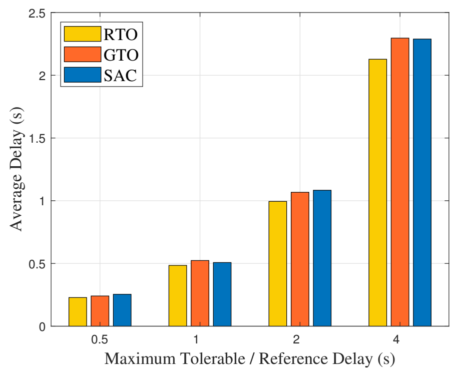
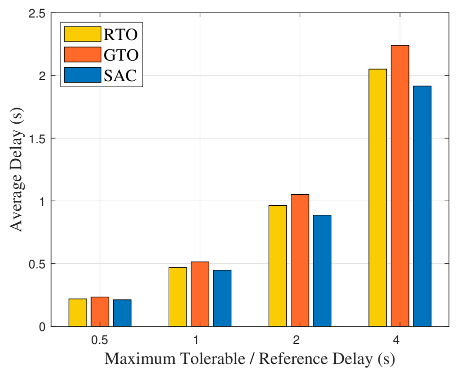


Fig. 5. The completion ratio of offloading tasks under different traffic densities.



(a) The average delay of offloading tasks when the traffic density is 10 vehicles/km.



(b) The average delay of offloading tasks when the traffic density is 30 vehicles/km.

Fig. 6. The average delay of offloading tasks with different maximum tolerable delays.

our proposed algorithm is higher than by applying greedy algorithm and random algorithm. Furthermore, with the traffic density increasing, there are more service vehicles available for task offloading, thus offloading tasks may obtain more computing resource from service vehicles, which increases both the average utility and the completion ratio of offloading tasks.

Fig. 6(a) and Fig. 6(b) present the average delay of the offloading tasks that have been successfully completed under the traffic density of 10 vehicles/km and 30 vehicles/km, respectively. We can observe that under low traffic density, the average delays of offloading tasks with $\tau_n \in \{0.5, 1, 2, 4\}$ in our algorithm are close to that in greedy algorithm,

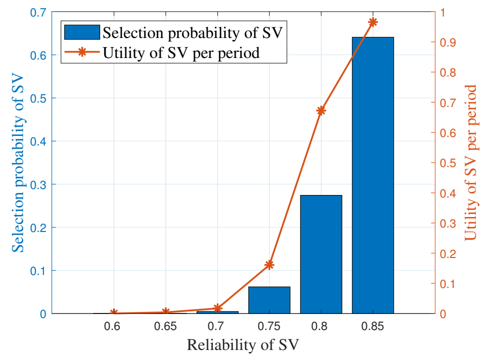


Fig. 7. The average utility per period and the selection probability of service vehicles (SVs) with different reliabilities when the traffic density is 30 vehicles/km.

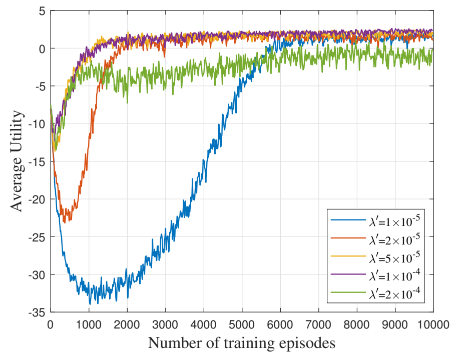


Fig. 8. The average utility of consensus nodes in a round of consensus with different learning rates.

but the completion ratio of greedy algorithm is lower than our proposed algorithm as illustrated in Fig. 5. Moreover, under high traffic density, it can be seen that the average delays of offloading tasks with different deadlines by applying our proposed algorithm are all lower than the other two algorithms. This is because our proposed algorithm considers the reliability of service vehicles and meanwhile maximizes the expected long-term reward which is related to the offloading delay, while the other two algorithms do not consider these aspects in vehicular task allocation.

In addition, Fig. 7 presents the probability of being selected as a service vehicle for an offloading task and the average utility per period of service vehicles with different reliabilities by applying our proposed algorithm. It can be seen that vehicles with higher reliability are more likely to be chosen as the service vehicles for offloading tasks. Meanwhile, service vehicles with higher reliability can obtain higher utility in a period, this is because service vehicles with higher reliability can receive more offloading tasks in vehicular task offloading.

### 6.3 Performance of Consensus Process

In the simulation of the consensus node selection algorithm, we utilize fully-connected DNNs to represent the Q-networks. Both the main Q-network $Q(\mathbf{s},\mathbf{a};\theta)$ and target Q-network $Q^{-}(\mathbf{s},\mathbf{a};\theta^{-})$ have two hidden layers with size $(1000, 1000)$, and the learning rate of the algorithm is set as $\lambda' = 5\times10^{-5}$. Besides, we set the batch size $|\mathcal{E}'| = 256$, and set the delay factor $\omega_d = 0.01$.

The complexity of the consensus node selection algorithm is commensurate with that of the DNN model as well. Similar to the complexity analysis of Algorithm 1, the complexity of action generation and training in each step is $O(N_B)$, where $N_B$ is number of the BSs in the VEC system.

To validate the performance of the DRL-based consensus node selection algorithm, we provide two benchmark algorithms for comparison:

- Greedy node selection (GNS): In each round of consensus, GNS always selects $N_c$ BSs with the most available computing resource as the consensus nodes, and chooses the block size that maximizes the average reward of the consensus nodes.
- Random node selection (RNS): In each round of consensus, RNS selects $N_c$ BSs in a random manner, and chooses the block size that maximizes the average reward of the consensus nodes.

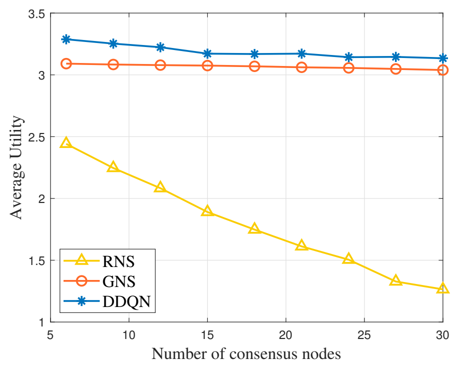


Fig. 9. The average utility of consensus nodes in a round of consensus with different number of consensus nodes.

Fig. 8 shows the average utility of consensus nodes in a round of consensus with different learning rates in our proposed algorithm. We can notice that our proposed algorithm with learning rate $\lambda' = 5\times10^{-5}$ or $\lambda' = 1\times10^{-4}$ reaches convergence around 2,000 training episodes. Besides, it can be seen from Fig. 8 that when the learning rate becomes smaller, the proposed algorithm needs to take more episodes to achieve convergence. When the learning rate becomes large enough, the proposed algorithm falls into a local optimum and cannot achieve the best performance.

As shown in Fig. 9, the average utility of consensus nodes in our proposed algorithm is higher than that in the benchmark algorithms, because our proposed algorithm considers both the available computing resource and the reliability of BSs, and jointly optimizes the consensus node selection and block size, while benchmark algorithms do not consider the reliability of the BSs in the selection of consensus nodes.

Fig. 10 presents the throughput of blockchain consensus with different number of consensus nodes by applying our proposed algorithm and benchmark algorithms. We can notice that greedy node selection achieves higher throughput than other algorithms, because greedy-based algorithm chooses the BSs with the maximum available computing resource and do not consider the reliability of BSs, which makes the block verification completed in the least time. Besides, we can notice that the throughput of consensus nodes

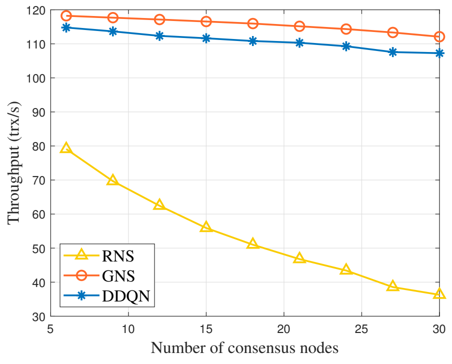


Fig. 10. The throughput of blockchain consensus with different number of consensus nodes.

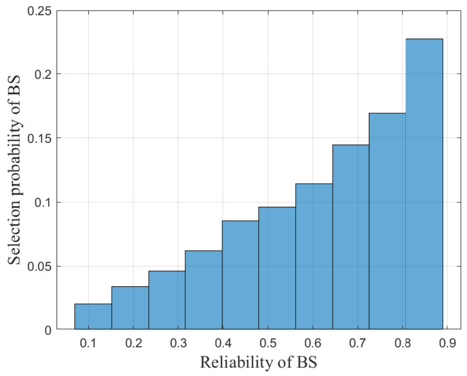


Fig. 11. The selection probability of BSs with different reliabilities in consensus node selection.

decreases with the number of consensus nodes increasing, this is because with the increase of consensus nodes, the BSs with less available computing resource will have more opportunity to participate in the blockchain consensus, and the time of block verification in a round of consensus depends on the minimum available computing resource of the consensus nodes.

In addition, Fig. 11 shows that the BS with a higher reliability has higher probability of being selected as a consensus node in our proposed algorithm. As presented in (28), the utility of a consensus node not only depends on the consensus delay and the number of validated transactions, but also depends on the reliability of the BS. Since the objective of our proposed algorithm is to maximize the average utility of consensus nodes, the agent tends to select BSs with high reliability as the consensus nodes.

## 7 Conclusion

In this paper, we have investigated the integration of blockchain and VEC, and proposed a vehicular task allocation scheme to ensure secure and reliable computation offloading among vehicles in the smart contract. To make the policy of computation offloading adapt to the dynamic vehicular environment, we formulated the vehicular task allocation problem as a sequential decision problem, and developed a DRL-based algorithm to solve the problem. Furthermore, in order to improve the efficiency of blockchain consensus and motivate BSs to improve the reliability in vehicular task allocation, we proposed an enhanced consensus protocol based on PBFT, and then designed a DDQN-based consensus node selection algorithm. Finally, simulation results revealed that the proposed algorithm can effectively improve the performance of V2V computation offloading and blockchain consensus, and meanwhile motivate BSs to improve their reliability in vehicular resource allocation.

## References

[1] X. Wang, Z. Ning, S. Guo, and L. Wang, "Imitation learning enabled task scheduling for online vehicular edge computing," *IEEE Trans. Mobile Comput.*, to be published, doi:10.1109/TMC.2020.3012509.
[2] J. Du, C. Jiang, J. Wang, Y. Ren, and M. Debbah, "Machine learning for 6G wireless networks: Carry-forward-enhanced bandwidth, massive access, and ultrareliable/low latency," *IEEE Veh. Technol. Mag*, vol. 15, no. 4, pp. 123–134, Dec. 2020.
[3] X. Peng, K. Ota, and M. Dong, "Multiattribute-based double auction toward resource allocation in vehicular fog computing," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3094–3103, Apr. 2020.
[4] Z. Ning, P. Dong, X. Wang, X. Hu, J. Liu, L. Guo, B. Hu, R. Kwok, and V. C. M. Leung, "Partial computation offloading and adaptive task scheduling for 5G-enabled vehicular networks," *IEEE Trans. Mobile Comput.*, to be published, doi:10.1109/TMC.2020.3025116.
[5] Z. Gao, M. Liwang, S. Hosseinalipour, H. Dai, and X. Wang, "A truthful auction for graph job allocation in vehicular cloud-assisted networks," *IEEE Trans. Mobile Comput.*, to be published, doi:10.1109/TMC.2021.3059803.
[6] J. Du, C. Jiang, H. Zhang, Y. Ren, and M. Guizani, "Auction design and analysis for SDN-based traffic offloading in hybrid satellite-terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2202–2217, Oct. 2018.
[7] H. Xu, W. Huang, Y. Zhou, D. Yang, M. Li, and Z. Han, "Edge computing resource allocation for unmanned aerial vehicle assisted mobile network with blockchain applications," *IEEE Trans. Wireless Commun.*, to be published, doi:10.1109/TWC.2020.3047496.
[8] Z. Xiong, Y. Zhang, D. Niyato, P. Wang, and Z. Han, "When mobile blockchain meets edge computing," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 33–39, Aug. 2018.
[9] Z. Xiong, S. Feng, W. Wang, D. Niyato, P. Wang, and Z. Han, "Cloud/fog computing resource management and pricing for blockchain networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4585–4600, Jun. 2019.
[10] M. B. Mollah, J. Zhao, D. Niyato, Y. L. Guan, C. Yuen, S. Sun, K. Y. Lam, and L. H. Koh, "Blockchain for the internet of vehicles towards intelligent transportation systems: A survey," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4157–4185, Mar. 2021.
[11] J. Feng, F. R. Yu, Q. Pei, J. Du, and L. Zhu, "Joint optimization of radio and computational resources allocation in blockchain-enabled mobile edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4321–4334, Jun. 2020.
[12] F. Guo, F. R. Yu, H. Zhang, H. Ji, M. Liu, and V. C. M. Leung, "Adaptive resource allocation in future wireless networks with blockchain and mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1689–1703, Mar. 2020.
[13] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1508–1532, 2nd Quart. 2019.
[14] Z. Zhuang, J. Wang, Q. Qi, J. Liao, and Z. Han, "Adaptive and robust routing with lyapunov-based deep RL in MEC networks enabled by blockchains," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2208–2225, Feb. 2021.
[15] H. Chai, S. Leng, K. Zhang, and S. Mao, "Proof-of-reputation based-consortium blockchain for trust resource sharing in internet of vehicles," *IEEE Access*, vol. 7, pp. 175 744–175 757, 2019.
[16] Y. Dai, D. Xu, K. Zhang, S. Maharjan, and Y. Zhang, "Deep reinforcement learning and permissioned blockchain for content caching in vehicular edge computing and networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4312–4324, Apr. 2020.
[17] X. Zhang and X. Chen, "Data security sharing and storage based on a consortium blockchain in a vehicular ad-hoc network," *IEEE Access*, vol. 7, pp. 58 241–58 254, 2019.
[18] S. Wang, D. Ye, X. Huang, R. Yu, Y. Wang, and Y. Zhang, "Consortium blockchain for secure resource sharing in vehicular edge computing: A contract-based approach," *IEEE Trans. Netw. Sci. Eng.*, to be published, doi:10.1109/TNSE.2020.3004475.

[19] J. Feng, Z. Liu, C. Wu, and Y. Ji, "AVE: Autonomous vehicular edge computing framework with ACO-based scheduling," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10 660–10 675, Dec. 2017.

[20] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10 190–10 203, Nov. 2018.

[21] F. Sun, F. Hou, N. Cheng, M. Wang, H. Zhou, L. Gui, and X. Shen, "Cooperative task scheduling for computation offloading in vehicular cloud," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11 049–11 061, Nov. 2018.

[22] J. Zhao, M. Kong, Q. Li, and X. Sun, "Contract-based computing resource management via deep reinforcement learning in vehicular fog computing," *IEEE Access*, vol. 8, pp. 3319–3329, 2020.

[23] X. Chen, L. Zhang, Y. Pang, B. Lin, and Y. Fang, "Timeliness-aware incentive mechanism for vehicular crowdsourcing in smart cities," *IEEE Trans. Mobile Comput.*, to be published, doi:10.1109/TMC.2021.3052963.

[24] H. Liao, Y. Mu, Z. Zhou, M. Sun, Z. Wang, and C. Pan, "Blockchain and learning-based secure and intelligent task offloading for vehicular fog computing," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi:10.1109/TITS.2020.3007770.

[25] S. Iqbal, A. W. Malik, A. U. Rahman, and R. M. Noor, "Blockchain-based reputation management for task offloading in micro-level vehicular fog network," *IEEE Access*, vol. 8, pp. 52 968–52 980, 2020.

[26] X. Lin, J. Wu, S. Mumtaz, S. Garg, J. Li, and M. Guizani, "Blockchain-based on-demand computing resource trading in IoV-assisted smart city," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi:10.1109/TETC.2020.2971831.

[27] K. Xiao, W. Shi, Z. Gao, C. Yao, and X. Qiu, "DAER: A resource preallocation algorithm of edge computing server by using blockchain in intelligent driving," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9291–9302, Oct. 2020.

[28] C. Qiu, H. Yao, F. R. Yu, C. Jiang, and S. Guo, "A service-oriented permissioned blockchain for the internet of things," *IEEE Trans. Services Comput.*, vol. 13, no. 2, pp. 203–215, Mar./Apr. 2020.

[29] J. Luo, Q. Chen, F. R. Yu, and L. Tang, "Blockchain-enabled software-defined industrial internet of things with deep reinforcement learning," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5466–5480, Jun. 2020.

[30] M. Liu, F. R. Yu, Y. Teng, V. C. M. Leung, and M. Song, "Performance optimization for blockchain-enabled industrial internet of things (IIoT) systems: A deep reinforcement learning approach," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3559–3570, Jun. 2019.

[31] W. L. Tan, W. C. Lau, O. Yue, and T. H. Hui, "Analytical models and performance evaluation of drive-thru internet systems," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 207–222, Jan. 2011.

[32] J. Du, E. Gelenbe, C. Jiang, H. Zhang, and Y. Ren, "Contract design for traffic offloading and resource allocation in heterogeneous ultra-dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2457–2467, Nov. 2017.

[33] M. Liu, F. R. Yu, Y. Teng, V. C. M. Leung, and M. Song, "Distributed resource allocation in blockchain-based video streaming systems with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 695–708, Jan. 2019.

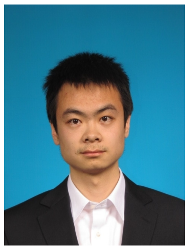
[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 80, Jul. 2018, pp. 1861–1870.

[35] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," *arXiv:1812.05905*, 2018.

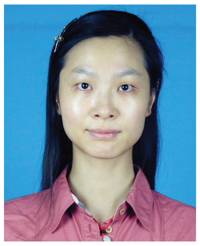
[36] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 13th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 2094–2100.

[37] J. Feng, F. Richard Yu, Q. Pei, X. Chu, J. Du, and L. Zhu, "Cooperative computation offloading and resource allocation for blockchain-enabled mobile-edge computing: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6214–6228, Jul. 2020.

[38] C. Qiu, F. R. Yu, H. Yao, C. Jiang, F. Xu, and C. Zhao, "Blockchain-based software-defined industrial internet of things: A dueling deep $Q$ -learning approach," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4627–4639, Jun. 2019.



**Jinming Shi** received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015. He is currently working toward the Ph.D degree in electronic engineering with Tsinghua University, Beijing, China. His research interests include Internet of Vehicles, blockchain, vehicular fog computing, and deep reinforcement learning.



**Jun Du** received her B.S. in information and communication engineering from Beijing Institute of Technology, in 2009, and her M.S. and Ph.D. in information and communication engineering from Tsinghua University, Beijing, in 2014 and 2018, respectively. She currently holds a postdoctoral position with the Department of Electrical Engineering, Tsinghua University. Her research interests are mainly in resource allocation and system security of heterogeneous networks and space-based information networks.

**Yuan Shen** received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the S.M. degree in computer science and the Ph.D. degree in electrical engineering from MIT, Cambridge, MA, USA, in 2008 and 2014, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include network localization and navigation, inference techniques, resource allocation, and cooperative networks.

**Jian Wang** (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006. In 2006, he joined the faculty of Tsinghua University, where he is currently a Professor with the Department of Electronic Engineering. His research interests include application of statistical theories, optimization, machine learning to communication, networking, navigation, and resource allocation problems.

**Jian Yuan** received his Ph.D. degree in electrical engineering from the University of Electronic Science and Technology of China, in 1998. He is currently a professor in the Department of Electronic Engineering at Tsinghua University, Beijing, China. His main research interest is in complex dynamics of networked systems.

**Zhu Han** (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, in 2003.,He is a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as with the Computer Science Department, University of Houston, Houston, TX, USA. He is also a Chair Professor with National Chiao Tung University, Hsinchu, Taiwan.