

Distribution-aware cache replication for cooperative road side units in VANETs

Fei Chen¹  · Detian Zhang¹ · Jian Zhang² · Xiaofeng Wang³ · Lifang Chen¹ · Yuan Liu¹ · Jiangchuan Liu⁴

Received: 19 January 2017 / Accepted: 15 June 2017 / Published online: 5 July 2017
© Springer Science+Business Media, LLC 2017

Abstract Increased storage capacity, computing and communications power, coupled with advances in wireless networking technology, bring a potential to enable new applications for vehicular ad hoc network (VANETs), in which a large number of roadside units (RSUs) are deployed to facilitate the service for drivers and passengers in vehicles. In this

paper, we focus on a cache replication strategy design for distributed RSUs allocated in a sequence. By exploring the relationship between the RSU allocation and content replication, we found that not only the local traffic flow but also the replication status of neighboring RSUs would affect the content replication efficiency of each RSU. The naive replication of most popular demand items may not always the best solution especially when the RSUs are aggregated in a small area. Accordingly, a distribution-aware replication cooperation (DRC) strategy is developed with the consideration of content replication efficiency and RSU allocation. The results have demonstrated the superiority of our proposed solution, as well as the scalability in the various scenarios with diverse request demands.

This article is part of the Topical Collection: *Special Issue on Big Data Networking*
Guest Editors: Xiaofei Liao, Song Guo, Deze Zeng, and Kun Wang

✉ Fei Chen
feicarticle@outlook.com
Detian Zhang
tianzdt@mail.ustc.edu.cn
Jian Zhang
jianzhang_neu@163.com
Xiaofeng Wang
wangxf@jiangnan.edu.cn
Lifang Chen
may7366@163.com
Yuan Liu
lyuan1800@sina.com
Jiangchuan Liu
csljc@ieee.org

Keywords Cache replication · VANET · Distributed systems

1 Introduction

Increased storage capacity, computing and communications power, coupled with advances in wireless networking technology, bring a potential to enable new applications for drivers and passengers in the vehicles [4]. Vehicular ad hoc networks (VANETs) are designed to provide information exchange via Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications [1, 2]. Due to the flexible and elastic resource provisioning capability, the cloud platform can accommodate various request demands in VANETs [3, 6, 9]. Furthermore, the distributed RSUs can act as "the last mile" to facilitate the content delivery, connecting the vehicles and remote cloud servers [4, 10]. As RSUs represent gateways to the Internet, the vehicles can transmit their real-time information and Internet

- ¹ School of Digital Media, Jiangnan University, Wuxi, China
- ² School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China
- ³ School of Internet of Things Engineering, Jiangnan University, Wuxi, China
- ⁴ College of Natural Resources and Environment, South China Agricultural University, Guangzhou, Guangdong, China

access requests to RSUs. RSUs send responses to the Internet queries and road information to vehicles. Generally, it is difficult, in terms of infrastructure cost, to cover roads with a large number of RSUs so that every vehicle on road can be connected to at least one nearby RSU all the time [4]. There have been extensive previous work to explore the RSU allocation, content replication, and interrupt connection issues in RSU services [6, 7, 9].

Comparing with traditional geo-distributed content delivery services, e.g., CDNs, the coverage of each RSU is relatively small, and can hardly provide seamless roaming for vehicles. Therefore, it is necessary to allocate great amounts of RSUs in the urban area for a higher cache hit ratio or a larger service range. Meanwhile, as the road direction is usually confirmed, the vehicles driving along the road would pass several RSUs in a sequence. Obviously, the cooperations between multiple RSUs have the potential opportunity to improve the system performance, especially for the RSUs allocated in the same road segment, which are highly correlated with each other during the service of the traffic flow. Intuitively, the service of an RSU is influenced by the content replication of the RSU allocated in previous position. However, it is sophisticated to qualify the influence, as the correlation between neighboring RSUs is opportunistic in the VANET environment, considering the vehicles may enter or leave the road arbitrarily. Furthermore, the RSUs may be allocated non-uniformly along the road, and the traffic flow may experience dynamic traffic conditions under time-varying request demands.

In this paper, we focus on the cooperative strategy design for the content replication of RSUs in a sequence. A distributed solution is presented for real world implementations. Our proposed strategy considers both of the service efficiency and the impact of RSU allocation from the viewpoint of RSU and vehicles, respectively. The remainder of this paper proceeds as follows. In Section II, we investigate the system model and formulate the replication problem by walking through a toy example. Section III explores the impact of critical parameters in the single-RSU scenario and multiple-RSU scenario, respectively. In Section IV, we develop a distribution-aware replication strategy with the cooperation of neighboring RSUs. The numerical results are presented in Section V to evaluate the performance under different strategies. Finally, Section VI concludes this paper with discussions of some potential future directions.

2 Related work

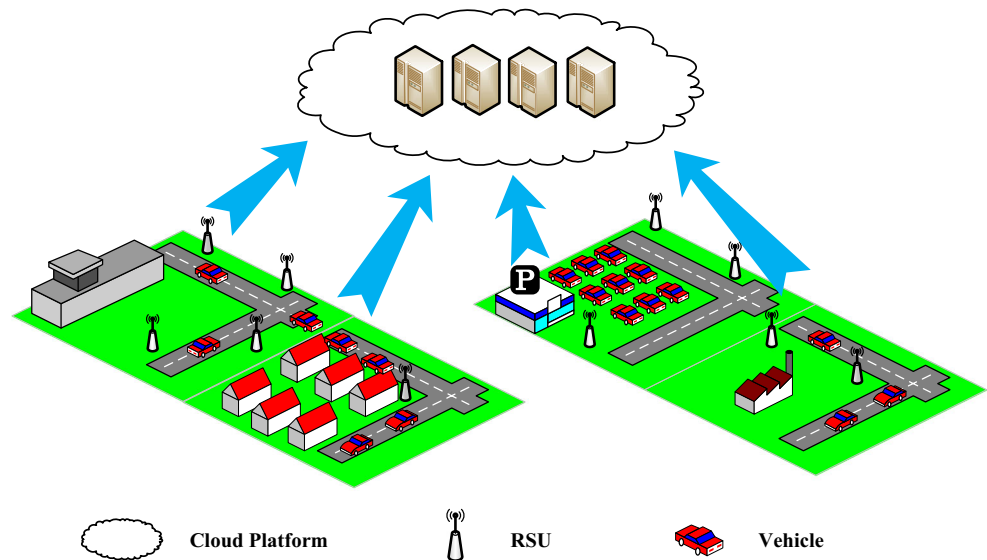
Urban vehicular networks are recognized as a significant component of the future intelligent transportation systems,

and they support various mobile services ranging from content-sharing applications to the information-spreading services [1, 4]. To solve the communication and computational capacity problem, some existing methods were proposed, including 3/4G cellular networks [11, 12], vehicle-to-vehicle communications [2], and vehicular cloud computing [5, 6]. As an extending paradigm of vehicular cloud computing, fog computing has been presented to process certain workloads and services locally on edge devices or edge servers, for mobile applications or real-time low latency services [8]. In VANETs environment, these edge devices are usually considered as the road side units (RSUs) allocated along the roads in urban areas [4, 6]. With the on-board computing facility and pre-cached contents, the RSUs can independently provide pre-defined service applications to passing by vehicles without going through the remote cloud [18, 19].

The content allocation strategy in distributed RSUs has been greatly explored in recent years [13–19]. Generally, these works can be divided into two areas, including the *RSU placement* and *cache replication*. Specifically, RSU placement refers to the position selection according to the vehicle distribution. For example, Wu et al. [14] propose a cost-effective strategy for RSU placement in VANETs with the formulation of vehicle population distribution and the vehicle speed. In paper [15], the optimal RSU allocation strategy is presented to cover all the road parts that have the highest care density using a limited number of RSUs. In paper [16], an RSU placement strategy is proposed to improve the location impacts on the efficiency of vehicular network with the minimal cost provisioning. In addition, Abdrabou et al. [17] present an analytical RSU placement framework, which takes into account the randomness of vehicle data traffic and the statistical variation of the disrupted communication channel. As to the cache replication, it refers to replica allocation in the fixed RSUs with limited capacity according to various request demands. For example, in paper [18], a novel cache invalidation algorithm is proposed to take advantage of the underlying location management scheme to reduce the number of broadcast operations and corresponding query delay. Furthermore, Ding et al. [19] propose an optimal, a sub-optimal, and a greedy strategy for RSU content replication to minimize the average request delay. However, none of these works has taken into account both the RSU allocation and content popularity during the content replication.

In this paper, we will unveil an interesting relationship among the cache capacity, connectivity, and traffic flow mobility, and we also find out the characteristics about the content replication efficiency of distributed RSUs, which are the benefits from the understanding of neighbor RSU distribution.

Fig. 1 An overview of content distribution in VANETs



3 System model and problem formulation

3.1 System overview

In Fig. 1, we consider a vehicular network including vehicles in dynamic traffic flow, RSUs distributed in urban areas, and cloud platform to support in a large scale. As RSUs representing gateways to the Internet, vehicles transmit their real-time information and Internet access requests to RSUs. The RSUs cache the file items which can be downloaded by the vehicles nearby. There are a large set of files in total, which are all stored in the cloud platform as backups. Meanwhile the RSU storage capacity is limited, and cache items need to be updated according to the replication strategy. We assume that there is only one item request for each vehicle, and global demands follow a popularity distribution.

As the RSU distribution is constant, the cache replication strategy refers to the file item management in the local storage of RSUs in the system. Each RSU receives the requests from the traffic flow in its coverage area, and feeds back the request demand according to its local replication. From the viewpoint of a specific vehicle, its request demand can be fulfilled by the RSUs passing by, otherwise it will resort to the support of cloud platform or vehicle-to-vehicle communications. Obviously, the performance of RSU service has been greatly influenced by the efficiency of cache replication strategy.

3.2 A case study

A toy example is presented in Fig. 2. Consider a road segment with three RSUs with the service times as τ_1 , τ_2 and τ_3 respectively, according to their local traffic conditions. These RSUs are allocated in a sequence along the road.

Thus, we can divide the vehicles as Flow 1, Flow 2 and Flow 3 namely, according to the positions they enter the road segment. Traffic Flow 1 enters at the beginning of the road, and will pass by all the RSUs in a sequence. Traffic Flow 2 enters between the RSU 1 and RSU 2, and will pass by the rest RSUs in the following journey. Traffic Flow 3 enters between the RSU 2 and RSU 3, and can only send the request demands toward RSU 3.

Suppose that there are two request items as item a and item b with deadline d_a and d_b , and item b has a less popularity and a more urgent deadline constraint. We assume the current time slot is t_0 and the deadline constraint is $t_0 + \Delta t_1 < d_b < t_0 + \Delta t_2 < t_0 + \Delta t_1 + \Delta t_2 < d_a$. Here we consider that each RSU can only replicate one item for request demands. As RSU 1 can only serve the traffic flow A, its replication strategy can be dependent to the popularity distribution of item 1 and item 2 in traffic flow 1. Due to the deadline constraint $t_0 + \Delta t_1 < d_b < d_a$, RSU 2 can serve traffic flow 1 and flow 2 for both item a and item b. Thus, the content replication of RSU 2 can be determined by the popularity distribution of item a and item b in flow 2, as well as the popularity distribution in flow 1 after the service of RSU 1. Due to the deadline constraint $d_b < t_0 + \Delta t_2$ and $d_a > t_0 + \Delta t_1 + \Delta t_2$, RSU 3 can only serve the traffic flow 1 and flow 2 for item a, besides the

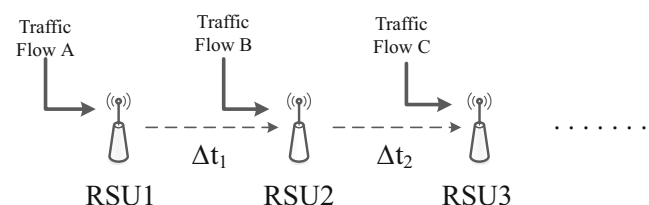


Fig. 2 An illustration of a toy example

traffic flow 3. Then the content replication of RSU 3 can be determined by the popularity distribution of item a and item b in flow 3, as well as the number of vehicles for item a from flow 1 after the service of RSU 1 and the number of vehicles for item a from flow 2 after the service of RSU 1 and RSU 2.

Above all, the content replication strategy for RSUs in a sequence can be influenced by the following five factors: (1) the joint traffic flows in different positions (2) the popularity distribution of request demands in these flows (3) the deadline constraints of the request demands (4) the content replication of RSUs in a sequence (5) the distance of these neighboring RSUs. Specifically, the factors (4) and (5) present the features of RSU allocation in a sequence, which are prevalent in the modern traffic environment and are not sufficiently explored in previous work. Together with the factors (1)–(3) under time-varying environment, they elevate the challenges to develop an efficient cache replication solution for RSUs allocated in a large scale.

3.3 Problem formulation

Without loss of generality, we assume that the set of RSUs $\mathbb{U} = \{u_1, u_2, \dots, u_M\}$ with uniform capacity c is distributed all over the urban areas. All the file items are of unit size and stored at the cloud platform as backups. Each RSU can replicate c items at most. We also assume that the connection between any vehicle and RSU is uniform in the coverage area, and the bandwidth capacity of RSUs is usually not considered as the constraint in VANETs according to previous researches [17, 19]. Therefore, we can denote δ as the entering time period for each moving vehicle to download its requested item passing by neighboring RSUs. Otherwise, this vehicle should resort to the multi-hop vehicle-to-vehicle connection or cloud platform through 3G/LTE transmission. According to the distribution of RSUs, the coverage area can be divided into M different regions as $A = \{a_1, a_2, \dots, a_M\}$. Although there can be overlap coverage areas between neighboring RSUs, we consider that each vehicle would always choose the nearest RSU. Denote the vehicular distribution in each region as $N = \{n_1(t), n_2(t), \dots, n_M(t)\}$ at time t . To further present the traffic condition, let $\{\tau_1(t), \tau_2(t), \dots, \tau_M(t)\}$ be the duration time for vehicles to pass the coverage regions of RSUs at time t .

Specifically, we consider a vehicle $v^k \in \mathbb{N}$ with the time slice between T_{start}^k and T_{end}^k as the travel period. Denote $A^k \subset \mathbb{A}$ as the RSU coverage areas passing by during the period $|T_{end}^k - T_{start}^k|$. For example, $A^k = \{a_1, a_2, a_4\}$ means that vehicle v^k would go through the coverage areas a_1 , a_2 and a_4 at time t_1^k , t_2^k and t_4^k respectively, and request to RSUs u_1 , u_2 and u_4 for file downloading. Let $I(v, u, t)$ be the

indicator to present the cache status of RSU u to serve the request from vehicle v passing by the coverage area at time t . For example, $I(v^k, u_1, t) = 1$ and $I(v^k, u_2, t) = 1$ indicate the requested file content is replicated in RSUs u_1 and u_2 . Oppositely, $I(v^k, u_4, t) = 0$ represents that there is no content replication for vehicle v^k when it passes by the coverage area of RSU u_4 . Accordingly, the service time $T_{service}^k$ from RSUs for vehicle v^k can be computed as follows:

$$T_{service}^k = \int_{T_{start}^k}^{T_{end}^k} I(v^k, u, t) \cdot \tau^k(t) dt \quad (1)$$

Further, we can define the relative service ratio from RSUs to vehicle v^k as follows:

$$R^k = \begin{cases} 1 & \text{if } T_{service}^k \geq \delta \\ \frac{T_{service}^k}{\delta} & \text{otherwise} \end{cases} \quad (2)$$

The optimization problem can be further formulated as:

$$\text{Max.} \quad \sum_{v^k \in \mathbb{N}} R^k \quad (3)$$

$$\text{s.t.} \quad \forall u \in \mathbb{U} \quad \sum_{v^k \in \mathbb{N}} I(v^k, u, t) \leq c$$

where the capacity constraint limits the service ability of each RSU for the time-varying traffic flows. It can be transferred into combinational optimization as the equivalent problem, and solved through mixed integral liner programming. However, the computation complexity grows exponentially as the number of RSU increases, and it is not practical to implement a centralized algorithm in a large scale under dynamic traffic flows.

4 Parameter analysis

Considering the practical implementation issues for RSUs content replication in VANETs, we will explore the distributed strategy through the critical parameter analysis from the viewpoints of both RSUs and vehicles. From the viewpoint of RSUs, replication efficiency is to evaluate the possibility of content replication to be requested in RSUs. From the viewpoint of vehicles, service efficiency is to present the estimation of service time from RSUs during the travel. We will first discuss the single-RSU scenario, then the multiple-RSU scenarios.

4.1 Single-RSU scenario

Different from the traditional Internet applications, the service in VANETs may experience both the request demand variation and dynamic traffic flow at the same time. To analyze the replication efficiency, we will present the definitions to qualify the impact value from replica and location, respectively.

Definition 1 Given a single replica in RSU with a constant cache capacity, the **Request Value (RV)** represents the possibility to be requested during a normal request.

It is obvious that RV refers to the cache hit ratio in a traditional cache policy, which is influenced by the popularity of the file item under the time-varying user demand. Thus, given an RSU $u_i \in \mathbb{U}$ with constant cache capacity c , its RV can be computed through $\sum_{j=1}^c h_{i,j}(t)$, where $h_{i,j}(t)$ is the cache hit ratio for replica j in region a_i at time t .

Definition 2 Given a single RSU with constant range of coverage area, the **Location Value (LV)** represents the service period per unit time under the current traffic flow.

Different from RV, we define LV to reflect the influence of dynamic traffic flow. Given an RSU $u_i \in \mathbb{U}$ with constant coverage range as ω , the traffic condition in a_i at time t can be qualified according to the vehicle density $\frac{n_i(t)}{\omega}$ and traffic speed $\frac{\omega}{\tau_i(t)}$. Therefore, we can further have $\frac{n_i(t)}{\tau_i(t)}$, and $i \in \{1, \dots, M\}$ to indicate the vehicular flow throughput in region a_i at time t . For example, $\frac{n_a(t)}{\tau_a(t)} > \frac{n_b(t)}{\tau_b(t)}$ means that region a has a higher vehicular flow throughput, while in region b the vehicles may experience a traffic congestion (i.e. $\tau_b(t) > \tau_a(t)$) or low vehicle density (i.e. $n_b(t) < n_a(t)$).

Combining RV and LV, we can formulate the replication efficiency of RSU u_i in the single-RSU scenario as follows:

$$\begin{cases} n_i(t) \cdot \sum_{j=1}^c h_{i,j}(t) & \text{if } \delta \geq \tau_i \\ \delta \cdot \frac{n_i(t)}{\tau_i(t)} \cdot \sum_{j=1}^c h_{i,j}(t) & \text{otherwise} \end{cases} \quad (4)$$

From the viewpoint of vehicles, it is straightforward to qualify the service time, as there is no other RSUs to provide content delivery service. Therefore, it can be formulated as 0-1 binomial distribution with probability p for its requested item to be replicated in the cache.

4.2 Multiple-RSU scenario

As an extension from the single-RSU scenario, vehicles may pass several regions covered by different RSUs to download the requested content. In the multiple-RSU scenario, the request from vehicles are affected by the RSUs they passing by. Therefore, we can have the similar replication efficiency formulation with that in the single-RSU scenario. The only revision of Eq. 4 is that δ needs to be replaced by the rest time of file downloading. Furthermore, we have considered the multiple-RSU with a uniform distribution as well as a non-uniform distribution.

4.2.1 Uniform RSU distribution

In the uniform RSU distribution, the number of RSUs passing by is determined by the length of travel, and the service period δ can be divided into several discrete time slices under the coverage of these RSUs. Therefore, we can consider the Independent Reference Model (IRM) as the arrival model, which describes the encounter of RSUs in the way as a sequence of independent and identically distributed random variables. Accordingly, from the viewpoint of a single vehicle the service time duration from multiple RSUs can follow a multinomial distribution:

$$P(T, M) = M! \prod_{t_i \in T} \frac{p_{t_i}^{m_{t_i}}}{m_{t_i}!} \quad (5)$$

where t_i is the service time slice of a single RSU, T is the set of discrete service time slices, and $M = \sum_{t_i \in T} m_{t_i}$ represents the number of arrival RSUs on the way. The service time slices t_i refers to the time duration τ_i of the coverage area passing by, which can follow a poisson distribution. Therefore, the service time expectation of a single vehicle can be estimated in the multiple-RSUs scenario with a uniform distribution.

4.2.2 Non-uniform RSU distribution

In the non-uniform RSU distribution, the number of RSUs in any specific region is not a constant due to the distribution of RSUs. For example, the RSUs are aggregated in Region A, and loosely distributed in Region B. The efficiency of RSU can be influenced by the neighboring RSUs placement, as the demand traffic can be served by the RSUs in previous locations. Consider an extreme case that the distance between neighboring RSUs is infinite. Then each RSU can be regarded as a single-RSU scenario with the same replication efficiency according to Eq. 4, as long as the vehicle can travel long distance enough. Meanwhile, consider another extreme case that the distance between neighboring RSUs is minor. In this case, suppose that there are RSUs, and these RSU can be regarded as a single RSU with $n \cdot c$ capacity. Thus the replication efficiency of these RSUs can be calculated as follows:

$$\begin{cases} n_i(t) \cdot \sum_{j=1}^{n \cdot c} h_{i,j}(t) & \text{if } \delta \geq \tau_i \\ \delta \cdot \frac{n_i(t)}{\tau_i(t)} \cdot \sum_{j=1}^{n \cdot c} h_{i,j}(t) & \text{otherwise} \end{cases} \quad (6)$$

These RSUs are aggregated with short distance tend to be an entity in the replication efficiency analysis, while the RSUs separated by long distance tend to be independent as the replication efficiency analysis in the single-RSU

scenario. Further, we observe that the RSU distribution with infinite separation distance can maximize the replication efficiency, even though the tradeoff is the long tolerant delay. As the separation distance is reduced, the volume of replicating the same content in the neighboring RSUs is reduced, especially for $\delta \leq \tau_i$.

5 Distributed replication cooperation

In this paper we are motivated to explore a distributed content replication solution, which can be implemented in each RSU. As the RSUs are allocated in a sequence, they can have a broader view to improve content replication efficiency through neighboring communications.

In principle, we consider the content replication efficiency in a multiple-RSU scenario with the combination of the single-RSU replication efficiency and the multiple-RSU influence. For each RSU u_i , the single-RSU replication efficiency can be calculated according to Eq. 4, and we denote $Q_{i,j} = \sum_{u < i} [I(j, u, t - T(u, i)) \cdot (\frac{1}{c})^{T(u,i)}]$ to qualify the influence from content replication in previous RSUs $u < i$ for replica j . Specifically, $T(u, i)$ is the time period driving from RSU u to RSU i , and $I(j, u, t - T(u, i))$ is the indicator for content replication of j in RSU u at time $t - T(u, i)$. In addition, $(\frac{1}{c})^{T(u,i)}$ denotes the weight of influence with storage capacity c and driving time $T(u, i)$. Thus, we can calculate the weight of replica j in RSU i at time t , as follows:

$$\omega_{i,j}(t) = \begin{cases} h_{i,j}(t) \cdot \frac{n_i(t)}{\tau_i(t)} - \eta \cdot Q_{i,j} & \delta \geq \tau_i \\ \delta \cdot h_{i,j}(t) \cdot \frac{n_i(t)}{\tau_i(t)} - \eta \cdot Q_{i,j} & \text{otherwise} \end{cases} \quad (7)$$

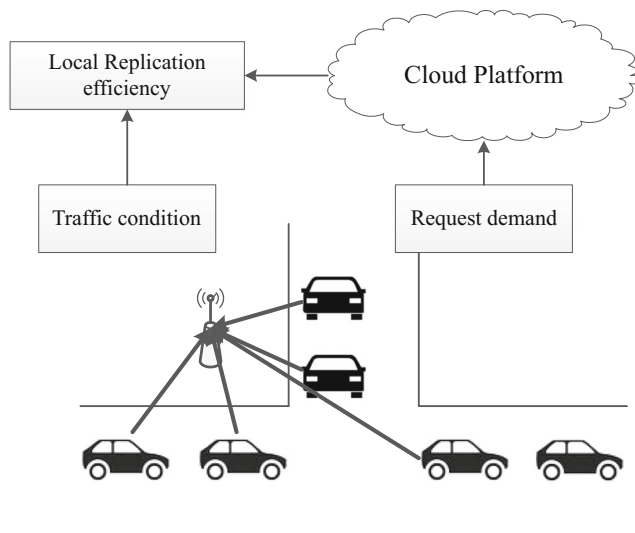


Fig. 3 Local replication efficiency

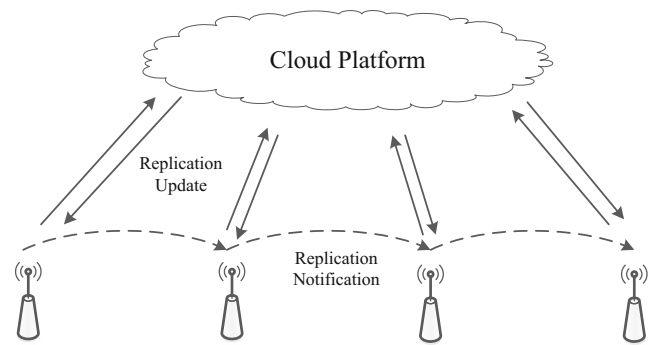


Fig. 4 Cooperative communication structure

in which, coefficient $\eta \in [0, 1]$ is set up to linearly combine the single-RSU replication efficiency and the multiple-RSU influence.

The distribution-aware replication cooperation (DRC) strategy can be implemented among distributed RSUs as the following three steps.

5.1 Local replication efficiency

In the first step, the RSUs need to record the traffic condition and request demand distribution in its local coverage area, and calculate the local replication efficiency correspondingly. The details of the implementation is illustrated in Fig. 3. The request demands distribution in the local coverage area of an RSU is dependent on the newly entering traffic flow, as well as the traffic flow served by the previous RSUs. Specifically, the request demand of the newly entering traffic flow can be reflected by the average popularity distribution, which can be achieved by the cloud platform through the popularity records from all RSUs. In addition, the traffic conditions includes the number of vehicles and driving speed, which determines the service time in the

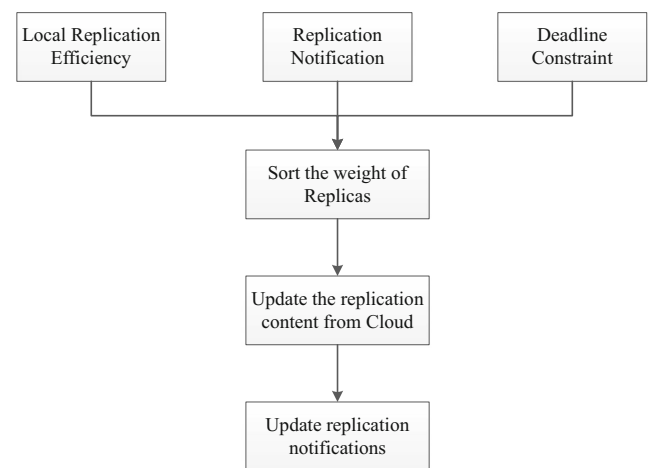


Fig. 5 Replication content update

coverage area. With the average popularity distribution and local traffic condition, we can obtain the local replication efficiency in this area.

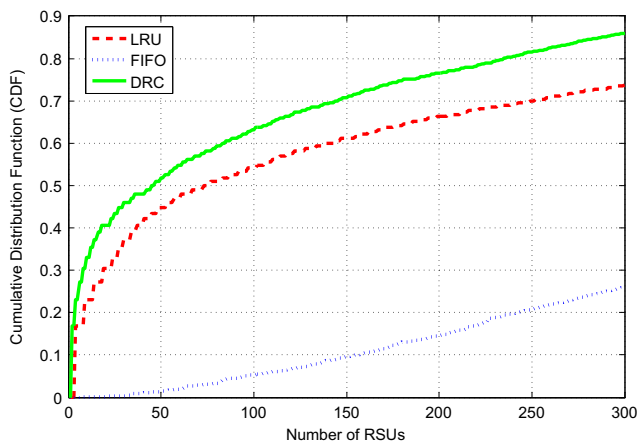
5.2 Cooperative communication structure

In the second step, we will present the cooperative communication structure in Fig. 4. To explore the traffic flow served by previous RSUs, we set up the replication notifications sent among RSUs along the direction of the road. The replication notifications include the information of the replication list and service time slots, starting from the beginning of the road toward the end of the road. When an RSU in the path receives a replication notification, it will add its local information into the notification, and transmits it to the next RSU along the road. According to the replication notifications, the replication statistics and the driving time between neighbor RSUs can be achieved by the following RSUs in the sequence. Furthermore, the distributed RSUs will

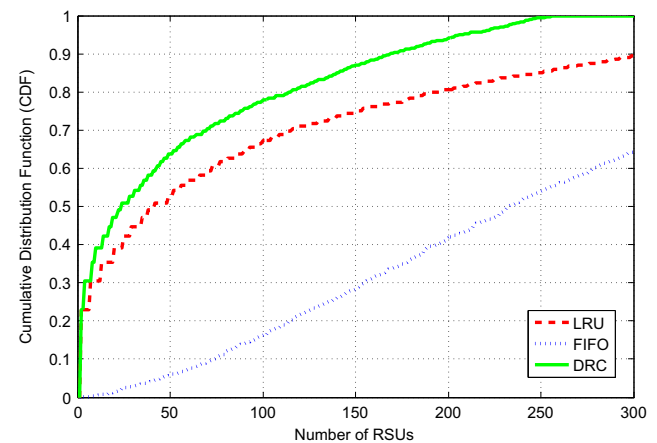
send the popularity record to the cloud platform and update the replication content from cloud in this communication structure.

5.3 Replication content update

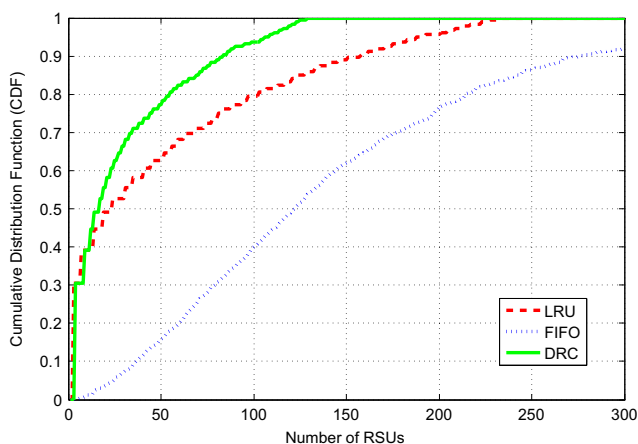
In the last step, we will present the replication content update of each distributed RSU in Fig. 5. For each RSU, the content replication update refers to the following three factors, deadline constraint for each replica, local replication efficiency, and received replication notifications. According to the deadline constraint, we will first remove the outdated replication content. Then we will combine the local replication efficiency and neighboring RSU influence from replication notifications to calculate the weight for each request item according to Eq. 7, and sort the replicas. At last, we will update the content replication from cloud in an online manner, and record the updated replication list in the replication notifications.



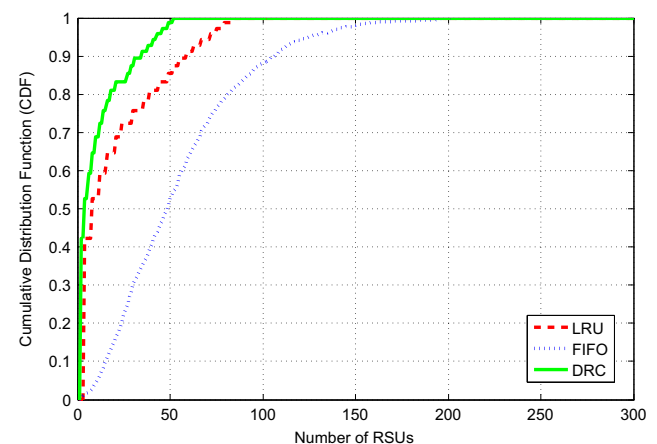
(a) Cache Capacity $C=5$



(b) Cache Capacity $C=10$



(c) Cache Capacity $C=20$



(d) Cache Capacity $C=50$

Fig. 6 Cumulative distribution functions with different cache capacity c

Note that as the impactor factor $(\frac{1}{c})^{T(u,i)}$ is reduced exponentially with the growth of distance between RSU u and i , there is a less impact factor with a larger storage capacity. In another word, if $c_1 > c_2$, the replication content with c_1 would have a less impact range than that with c_2 . In addition, as the storage capacity c increases, the number of valid RSU in impact range would be reduced. Therefore, the complexity of this proposed strategy is convergent and appropriate for the online implementation under the dynamic environment.

6 Numerical results

In this section, we will first explore the content replication of RSUs under heterogenous request demands, and then evaluate the system performance with dynamic traffic flow. We will present the experimental results to compare with different strategies in the two scenarios, respectively.

6.1 Heterogenous request demand

We consider the time-varying environment including the dynamic request demand and traffic condition in different coverage regions. For each vehicle, we assume that there is only one request demand of an item at time t . Generally, the request demand popularity follows a Zipf distribution $P(j) = \frac{j^{-\theta}}{\sum_{k=1}^M k^{-\theta}}$, where $M = 1000$ items in the system and $P(j)$ represents the popularity of the j th item. Furthermore, we assume that traffic conditions follow Poisson distribution $\frac{\lambda^k}{k!} e^{-\lambda}$, including the traffic flow as $\lambda_f(t)$, time durations ratio in coverage areas as $\lambda_t(t)$, and number of passing by RSUs as $\lambda_n(t)$ at each time slot t , respectively. Specifically, we consider the ratio $\lambda_t(t) \in (0, 1]$ as a fraction $\frac{\tau_i(t)}{\delta}$ for each region a_i .

In this experiment, we have compared our proposed replication strategy DRC with FIFO (First In First Out) and LRU (Least Recently Utilized), which are prevalent in the modern cache replication infrastructure [18, 19].

6.1.1 Cache capacity

To explore the impact of RSU cache storage capacity, the experiments are implemented with an assumption of that there is no deadline constraint for each vehicle. We will analyze the system performance variation under different replication strategies with cache capacity $c = 5$, $c = 10$, $c = 20$, and $c = 50$, respectively.

Figure 6 presents the Cumulative Distribution Functions (CDF) of RSU service ratio under different replication strategies with cache capacity c variation. Without the deadline constraint, the CDF of service time can converge to

1 with the increase of the number of RSUs passing by. Faster CDF convergence speed represents higher RSU replication efficiency. Generally, the LRU and our proposed strategy DRC have better performance comparing with the FIFO strategy, especially with a small number of RSUs. It indicates that the content replication in LRU and DRC would be more efficient with a limited replication capacity, whereas the FIFO strategy can only achieve a proportional CDF improvement with the growth of RSU number. As the increase of cache capacity, the distinction between these three strategies becomes minor. On the other hand, the performance gap between LRU and DRC would vary with the change of cache capacity. When cache capacity $c = 5$, the performance gain of DRC is about 10% over that of LRU. It can further reach about 15% improvement when $c = 10$ and $c = 20$. However it returns to 10% as the cache capacity increases to 50. Thus we can infer that the DRC strategy has a convex replication efficiency improvement with the RSU capacity. And the experiment results demonstrate the conjecture that the influence from neighboring RSUs would be reduced as the storage capacity grows continuously.

6.1.2 Correlation coefficient η

To further explore the impact of correlation coefficient η , the experiment is implemented with a deadline constraint, which is considered as the limited number of RSUs passing by. We assume the number of RSUs that a vehicle would pass by follows a Poisson distribution with $\lambda_n(t)$. Generally, given limited travel time, a higher $\lambda_n(t)$ indicates a denser RSU distribution with more neighboring RSUs in a limited range. Then we will evaluate the system performance with different coefficient η under the impact of RSU distribution density.

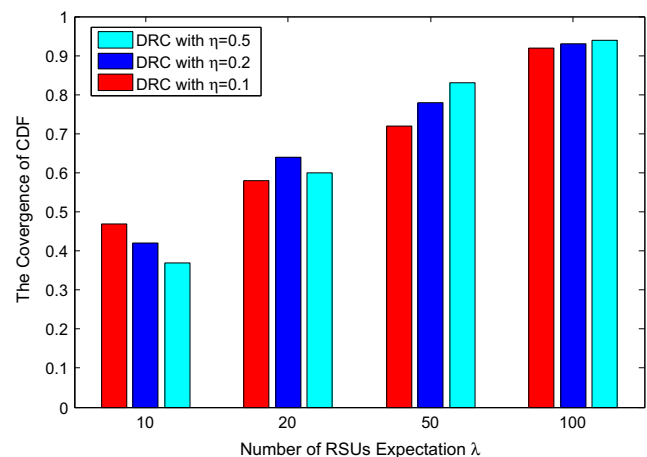


Fig. 7 The convergence of CDF with different correlation coefficient η

Figure 7 presents the CDF convergence of RSU service time under different replication strategies with correlation coefficient η variation. Different from the previous scenario, in which the service from RSUs can converge to 100% eventually. The time constraint T_{end}^k limits the convergence to certain degrees according to the efficiency of the replication strategy. As we can see, when the number of passing by RSUs expectation λ_n is low, the RSUs are loosely distributed, and DRC with $\eta = 0.1$ has a higher service ratio convergence. This is consistent with the extreme case that in the single-RSU scenario, the optimal replication can be simply achieved with a maximal cache hit ratio according to service efficiency definition in Eq. 4. As the number of RSUs expectation increases, the DRC with a moderate coefficient $\eta = 0.2$ would have a higher convergence. It can be explained that the system performance needs to balance both the replication efficiency and the service efficiency at the same time. As the distribution of RSUs become more aggregated, DRC with $\eta = 0.5$ excels as the weight of neighboring RSU influence is increased. As the number continues to grow, these strategies can finally reach a similar convergence of service ratio. It can be explained that without deadline constraint all the replication strategies can converge to 100% service ratio if given enough RSUs.

6.2 Dynamic traffic flow

We will further present the system performance with dynamic traffic flow. As vehicles do not change their velocity or direction frequently, we can use a fluid flow model [20] to capture the mobility of traffic flow in VANET. We have set up 20 RSUs for the service of the traffic flow along the road, and the vehicles can enter or leave the road at an arbitrary location. The popularity of request demand follows a Zipf distribution, and the deadline constraint is between 10 to 90 minutes.

We will evaluate our proposed strategy with other three popular solutions, i.e. popularity-first [19], latency-first, and LMC [18]. The reasons of selecting these schemes are listed as follows: (1) Popularity-first always choose the replication content with the most popularity. (2) Latency-first always choose the replication content with the most urgent

deadline. (3) LMC is the local management based cache strategy, which is dependent on the vehicle velocity and density in the local area.

6.2.1 Cost and latency

The system performance is evaluated through the calculation of total cost saving and average latency of served vehicles. Without content replication in RSU, the request demands of vehicles will be served by cloud platform directly. We take this cost as the benchmark and calculate the relevant cost saving for these four RSU replication strategies, respectively. In addition, we calculate the average latency for the served vehicles. The system performance is evaluated in two scenarios. In one scenario, the average driving time is 10 minutes, which means that the average time period between entering and leaving the road. In the other scenario, the average driving time is 40 minutes. From Table 1, we can see that the local-aware strategies, i.e., DRC and LMC, generally have a higher cost saving and a longer average latency, comparing with the two greedy strategies, i.e., popularity-first, and latency-first. Furthermore, when the average driving time is 40 minutes, the cost saving and average latency would both increase. This can be explained that DRC strategy can distribute the replicas uniformly in a wider range, and results in a higher cost saving.

6.2.2 Service ratio

We further present the RSU service ratio with the variation of driving time in Fig. 8. The service ratio represents the request demand served by the RSUs. Generally, we can see that the service ratio grows when the vehicles have a longer driving time along the road. Specularly, Popularity-first has the slowest growth, as it always selects the hottest replication content and loses the opportunity to serve the request demands with urgent deadline. Oppositely, Latency-first strategy has a continuous growth with a slow start. Meanwhile, the two local-aware solutions can have a fast growth at the beginning, and then converge at 0.84 and 0.75, respectively. We can see that DRC can have a better performance when the vehicles have driven for a longer time.

Table 1 Cost save and average latency comparison

Cache replication strategy	10 minutes		40 minutes	
	Cost Save	Latency	Cost Save	Latency
Popularity-first	15.2%	5.2 minutes	38.7%	10.8 minutes
Latency-first	10.1%	4.7 minutes	68.8%	12.5 minutes
DRC	51.1%	7.1 minutes	82.4%	32.3 minutes
LMC	59.7%	6.8 minutes	71.5%	30.7 minutes

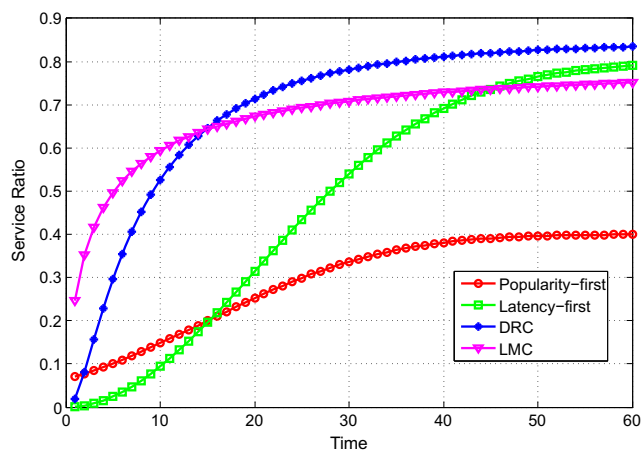


Fig. 8 RSU service ratio with variation of driving time

7 Conclusion

In this paper, we focus on the cache replication strategy design with limited RSU storage capacities under dynamic traffic conditions and RSU distributions. The impact of replication efficiency and service efficiency is analyzed with the single-RSU scenario and multiple-RSU scenario, respectively. A distributed replication cooperation strategy for practical implementation is further presented to accommodate the variation of traffic conditions and skew popularity of request demands. Finally, we have performed a set of numerical experiments and the results have demonstrated the superiority and usability of our proposed solution in different scenarios.

Acknowledgments This research is supported by the following grants: National Natural Science Foundation of China (Grant No. 61602214, 61672264), Natural Science Foundation of Jiangsu Province in China (Grant No. BK20160191), National Key Research and Development Program of China (Grant No. 2016YFB0800801), National Science and Technology Support Program of China (Grant No. 2015BAH54F01).

Compliance with Ethical Standards

Conflict of interests The authors declare that there is no conflict of interest regarding the publication of this article.

References

- Zheng K, Zheng Q, Chatzimisios P, Xiang W, Zhou Y (2015) Heterogeneous vehicular networking: a survey on architecture, challenges, and solutions. *IEEE Communication Surveys & Tutorials* 17(4):2377–2396
- Li Y, Jin D, Wang Z, Hui P, Zeng L, Chen S (2014) A markov jump process model for urban vehicular mobility: Modeling and applications. *IEEE Trans Mob Comput* 13(9):1911–1926
- Gong H, Yu L, Liu N, Zhang X (2016) Mobile content distribution with vehicular cloud in urban VANETs. *China Communications* 13(8):84–96
- Vegni AM, Loscri V (2015) A survey on vehicular social networks. *IEEE Communications Surveys & Tutorials* 17(4):2397–2419
- Salahuddin MA, Al-Fuqaha A, Guizani M (2015) Software-Defined Networking for RSU clouds in support of the internet of vehicles. *IEEE Internet of Things Journal* 2(2):133–144
- Lee E, Lee E, Gerla M, Oh SY (2014) Vehicular cloud networking: architecture and design principles. *IEEE Commun Mag* 52(2):148–155
- Hou X, Li Y, Chen M, Wu D, Jin D, Chen S (2016) Vehicular fog computing: a viewpoint of vehicles as the infrastructures. *IEEE Trans Veh Technol* 65(6):3860–3873
- Jalali F, Hinton K, Ayre R, Alpean T, Tucker RS (2016) Fog computing may help to save energy in cloud computing. *IEEE Journal on Selected Areas in Communications* 34(5):1728–1739
- Kim R, Lim H, Krishnamachari B (2016) Prefetching-Based Data dissemination in vehicular cloud systems. *IEEE Trans Veh Technol* 65(1):292–306
- Xie S, Wang Y (2014) Construction of tree network with limited delivery latency in homogeneous wireless sensor. *Wirel Pers Commun* 78(1):231–246
- Feteiha MF, Hassanein HS (2015) Enabling cooperative relaying VANET clouds over LTE-a networks. *IEEE Trans Veh Technol* 64(4):1468–1479
- Liu B, Jia D, Wang J, Lu K, Wu L (2017) Cloud-assisted Safety Message Dissemination in VANET-cellular Heterogeneous Wireless Network. *IEEE Syst J* 11(1):128–139
- Li P, Liu Q, Huang C, Wang J, Jia X (2015) Delay-bounded minimal cost placement of roadside units in vehicular ad hoc networks. In: *Proceedings of IEEE international conference on communications*
- Wu TJ, Liao W, Chang CJ (2012) A Cost-Effective strategy for Road-Side unit placement in vehicular networks. *IEEE Trans Commun* 60(8):2295–2303
- Farsi A, Szczechowiak P (2014) Optimal deployment of road side units in urban environments. In: *Proceedings of IEEE international conference on connected vehicles & expo*
- Balouchzahi NM, Fathy M, Akbari A (2015) Optimal road side units placement model based on binary integer programming for efficient traffic information advertisement and discovery in vehicular environment. *IEEE IET Intell Transp Syst* 9(9):851–861
- Abdrabou A, Zhuang W (2011) Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity. *IEEE Journal of Selected Areas in Communications* 29(1):129–139
- Lim S, Chae SH, Yu C, Das CR (2008) On Cache Invalidation for Internet-based Vehicular Ad Hoc Networks. In: *Proceedings of IEEE international conference on mobile adhoc and sensor systems*
- Ding R, Wang T, Song L, Han Z, Wu J (2015) Roadside-unit caching in vehicular ad hoc networks for efficient popular content delivery. In: *Proceedings of IEEE wireless communications and networking conference*
- Xu K, Tipper D, Qian Y, Krishnamurthy P (2016) Time-dependent performance analysis of IEEE 802.11p vehicular networks. *IEEE Trans Veh Technol* 65(7):5637–5651