Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8837751, 12 pages
https://doi.org/10.1155/2020/8837751

Research Article

Joint V2V-Assisted Clustering, Caching, and Multicast Beamforming in Vehicular Edge Networks

Kan Wang[1], **Ruijie Wang**[1], **Junhuai Li**[1] and **Meng Li**[2]

[1]*School of Computer and Science Engineering, Xi'an University of Technology, Xi'an 710048, China*
[2]*Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*

Correspondence should be addressed to Meng Li; limeng720@bjut.edu.cn

Received 28 August 2020; Revised 28 September 2020; Accepted 15 October 2020; Published 19 November 2020

Academic Editor: Hongwei Wang

As an emerging type of Internet of Things (IoT), Internet of Vehicles (IoV) denotes the vehicle network capable of supporting diverse types of intelligent services and has attracted great attention in the 5G era. In this study, we consider the multimedia content caching with multicast beamforming in IoV-based vehicular edge networks. First, we formulate a joint vehicle-to-vehicle- (V2V-) assisted clustering, caching, and multicasting optimization problem, to minimize the weighted sum of flow cost and power cost, subject to the quality-of-service (QoS) constraints for each multicast group. Then, with the two-timescale setup, the intractable and stochastic original problem is decoupled at separate timescales. More precisely, at the large timescale, we leverage the sample average approximation (SAA) technique to solve the joint V2V-assisted clustering and caching problem and then demonstrate the equivalence of optimal solutions between the original problem and its relaxed linear programming (LP) counterpart; and at the small timescale, we leverage the successive convex approximation (SCA) method to solve the nonconvex multicast beamforming problem, whereby a series of convex subproblems can be acquired, with the convergence also assured. Finally, simulations are conducted with different system parameters to show the effectiveness of the proposed algorithm, revealing that the network performance can benefit from not only the power saving from wireless multicast beamforming in vehicular networks but also the content caching among vehicles.

1. Introduction

Smart devices equipped with the capability to interact with the physical environment and the function to communicate with each other are prompting the Internet towards the so-called Internet of Things (IoT) [1]. IoT holds the promise to improve our lives and the way we interact with devices, such as actuators, sensors, cell phones, and home automation. With the ever-increasing proliferation of IoT devices, the worldwide multimedia traffic is anticipated to experience a rapid growth in the 5G era [2]. Different from the legacy IoT with information exchange at the byte level, image, audio, video, and other traffic in the 5G era are typically with a large volume of information, thereby bringing forth the new terminology, namely, multimedia IoT (MIoT). As an emerging type of IoT, MIoT denotes the IoT with multimedia traffic as outputs and inputs and has been extensively used in healthcare, smart homes, communication-based train control system (CBTC), and Internet of Vehicles (IoV) [3–6].

The introduction of IoV-based vehicular edge networks and the increasing demand for IoV services have imposed more stringent constraints on the quality-of-service (QoS) requirements of image and video, particularly for high-quality real-time multimedia services in fast-moving vehicles. These constraints indeed pose challenges in exploiting the full potential of vehicle-to-vehicle (V2V) communications, e.g., resource allocation, multiple-input-multiple-output- (MIMO-) based beamforming, network architecture, multicast routing, and dynamic controls [7–9].

To better cope with these constrains imposed on multimedia content retrieval in IoV-based vehicular edge networks, wireless content caching begins to emerge as promising solutions [10]. By bringing multimedia contents closer to vehicles via prefetching them from the Internet or core networks, the traffic loads and network costs can be

significantly reduced. In the meantime, the QoS for IoV applications could also be improved, since lower access delay and higher data rate are enabled by caching in edge nodes (e.g., base stations (BSs) and roadside units (RSUs)). Therefore, wireless content caching contributes to the agile multimedia content distribution with a higher data rate and enhanced broadband connectivity. On the other hand, as the extension of unicast transmission and by exploiting the broadcast nature of wireless channels, the point-to-multipoint multicast transmission provides a more efficient capacity-offloading approach, to deliver the identical content to multiple vehicles on the same frequency band. Thus, the integration of caching with multicasting becomes the key enabler in IoV-based edge networks, which could not only reduce the flow (traffic) cost in core or transport networks but also improve the spectrum efficiency in vehicular edge networks [11].

Besides, there have been some works related to caching-based multicast beamforming design [12–15]. Nevertheless, the integration of caching with multicasting in vehicular edge networks is not smooth and still faces several challenges, due to the following observations. First, the caching decision needs to operate at the large timescale (e.g., several minutes), and the multicast beamforming has to be tailored at the small timescale (e.g., several seconds) to exploit the mobility. That is, it is a prerequisite to decouple the joint caching and multicast beamforming to different timescales. Second, when the edge network is confronted with IoV in the 5G era, numerous vehicles emerge to assist the V2V communications between each other. Thus, although with smaller storage size compared to the BS, caching profits in vehicles are nonnegligible, which yet has not been studied thoroughly, especially in the caching-based multicast beamforming. Third, the cooperative multicasting necessitates the clustering of vehicles (namely, deciding on which vehicle should be picked into the cluster to serve the same receiver), which needs to be optimized at the large timescale as well.

In this paper, we study the joint V2V-assisted clustering, caching, and wireless multicast beamforming in vehicular edge networks. The distinctive features of this work are as follows:

The distinctive features of this work are as follows:

(1) We present a two timescale-based V2V-assisted clustering, caching, and multicast beamforming problem, with the objective of minimizing the total network cost (involving both flow and power costs), subject to the QoS constraint for each multicast group, and the caching memory limitation for each vehicle. With the two-timescale setup, different types of variables are decoupled at separate timescales

(2) At the large timescale, we leverage the sample average approximation (SAA) technique to resolve the joint clustering and caching problem. We first reformulate and relax the stochastic original problem as a deterministic integer linear programming (ILP) one and then exhibit the equivalence of optimal solutions between the original one and its relaxed linear programming (LP) counterpart, thereby simplifying the computation substantially

(3) We leverage the successive convex approximation (SCA) method to solve the multicast beamforming problem at the small timescale. As such, the original nonconvex problem can be transformed into a series of convex subproblems, with the convergence assured

(4) Simulations are executed with different system parameters to show the effectiveness of the proposed algorithm, as well as the convergence of the proposed two-timescale setup. Simulation results reveal that the system performance could benefit from both the power saving from multicasting and the content caching among vehicles

The remainder of this work is organized as follows. In Section 2, we list works related to caching-based multicasting and V2V-assisted caching, respectively. In Section 3, we present the system model and formulate a joint clustering, caching, and multicast beamforming problem, to minimize the total network cost subject to QoS constraints. In Section 4, we leverage the SAA and SCA methods to solve the large- and small-timescale problem, respectively. In Section 5, we present and discuss the simulation results. We conclude this paper in Section 6.

2. Related Work

Some recent studies related to caching-based multicast beamforming and V2V-assisted caching in edge networks are presented. We believe these works can motivate more achievements in the academia and industrial domain.

2.1. Caching-Based Multicasting. In the literature, multicast beamforming has emerged as an effective approach to mitigate the interference in cellular systems, e.g., in smart-grid powered cellular networks [16] or in full-duplex cellular networks [17]. Furthermore, the integration of caching and multicasting could multicast the identical content to multiple receivers in the same frequency band, and thus attracts significant interests.

In particular, in the cache-enabled cloud radio access network (RAN), Tao et al. [12] investigated the joint design of content-centric BS clustering and multicast beamforming for wireless content delivery. In particular, the problem is formulated as one mixed-integer nonlinear programming (MINLP) problem, and a sparse multicast beamforming algorithm is proposed based on the L_0 norm approximation. In the two-tier heterogeneous networks, Cui et al. [13] considered a random caching and multicasting mechanism with caching distributions as design parameters, to enable the efficient content dissemination. First, the successful transmission probabilities in the general and high signal-to-noise ratio region are derived via stochastic geometry; then, a nonconvex joint caching and

multicasting problem is formulated to maximize the successful transmission probability in the asymptotic region.

Also, in cache-enabled ultradense cellular networks, Nguyen et al. [14] proposed a cooperative multicast beamforming approach to improve the cost-efficiency. Besides, Zhou et al. [15] minimized the energy cost in a multicell multigroup setup, involving caching, computing, and communication resources.

2.2. V2V-Assisted Caching. As an analogous terminology to V2V, device-to-device (D2D) caching refers to the caching of popular contents in mobile devices. D2D could directly deliver the content to adjacent devices, thereby offloading the BS's traffic. Malak et al. [18] leveraged the stochastic geometry theory to derive the outage probability in the presence of interference and noise, designed the distributed caching strategy, and maximized the density of successful receptions. Jiang et al. [19] modelled the local caching as a backpack problem and characterized the D2D pairing as the maximum weight matching problem in the bipartite graph, to maximize the BS traffic offloading. In addition, Cacciapuoti et al. [20] claimed that in D2D networks, caching performance is closely related to devices' interests and preferences in multimedia contents. In particular, content popularity focuses on community interests, while device preference reflects the difference of request probability between different individuals. Therefore, Zhang et al. [21] designed a unified D2D-caching utility function, not only considering the D2D communication distance but also incorporating the similarity of preferences among adjacent devices.

With recent advents in the IoV paradigm, Su et al. [7] developed an edge caching mechanism in RSUs. Specifically, involving the content access pattern, vehicle velocity, and traffic density, the features of content requests are analysed, followed by a model to decide whether and where to cache the content. By proposing an analytical model, Tan et al. [22] quantified the effects of velocity, traffic density, and service rate on the content hit ratio. Also focusing on the edge caching in RSUs (termed as edge nodes therein), Zhao et al. [23] proposed one cost-efficient method to minimize the time-averaged service response delay, by jointly studied caching, request routing, and wireless resource allocation over the long run in an online manner. To solve the dynamics in IoV, Lyapunov optimization is leveraged to make near-optimal decisions, with a stable system assured. In addition, Xiao et al. [24] proposed one adaptive and dynamic user-centric virtual cell scheme to facilitate the multicasting of vehicular-to-everything (V2X) message, followed by a max-min fair problem formulation. Furthermore, focusing on 802.11p protocols rather than 3GPP standards, Wu et al. [25] proposed one semi-Markov decision process- (SMDP-) based formulation to maximize the long-term reward in vehicular edge networks, by jointly considering the transmission delay, computing delay, and task diversity.

2.3. Discussions. Edge computing and edge caching have attracted lots of attention [26–28]. The aforementioned works typically solve the clustering and multicast beamforming optimization at the same timescale. Nevertheless, it is often the case that the clustering and caching execute at a much larger timescale, while the multicast beamforming is tailored per small timescale to exploit the fast fading of wireless channels. Although Qiao et al. [28] decoupled content placement and content delivery at different timescales in vehicular edge networks, the multicast transmission was ignored, which possibly incurs low-spectrum efficiency. That is, it is a prerequisite to propose a joint clustering, caching, and multicast beamforming algorithm in vehicular edge networks on the basis of a two-timescale setup.

3. System Model and Problem Formulation

Vehicular edge network with one BS and multiple IoV devices is considered. The BS is equipped with L antennas, and each vehicle has one single antenna. As shown in Figure 1, in the context of V2V communications, the content requesting vehicle (CRV) could not only attach to the BS via cellular links but also associate with content caching vehicles (CCVs) directly. The BS is indexed by $j = 0$, while CCVs are denoted by $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$. Furthermore, $\mathcal{J}_0 = \mathcal{J} \cup \{0\}$ denotes as the set of all transmitters. Meanwhile, the CRVs with the identical content request are categorized into one multicast group, and all multicast groups are denoted by $\mathcal{S} = \{1, \cdots, s, \cdots, S\}$. For each multicast group $s \in \mathcal{S}$, its all associated CRVs are denoted by $\mathcal{J}_s$, and $\cup_{s\in\mathcal{S}} \mathcal{J}_s = \mathcal{J}$ holds, where $\mathcal{J}$ is the set of all CRVs. In addition, we follow the equal content size for all contents [12], namely, each of which is chunked and normalized to the size of one.

3.1. Two-Timescale Setup. Mixed-timescale setup in [29, 30] is incorporated in this study. Each large timescale accommodates multiple successive small timescale slots, e.g., a total of T slots, denoted by $\mathcal{T} = \{1, \cdots, T\}$. In addition, in each slot, the beamformer is designed for each multicast group, to exploit the fast fading and mobility experienced by CRVs. As revealed in Figure 2, at the end of each large timescale, the network operator has to decide the CCV clustering (i.e., the CCVs serving for the same multicast group) and the content caching for transmitters. Then, following this decision unchanged throughout the next large timescale, the picked content is cached in the local memory or storage in both BS and clustered CCVs.

In this work, we assume that all multicast groups can be admitted and satisfied with their content requests. The case also occurs that insufficient resources cannot accommodate all multicast groups (e.g., in [31]), which is yet beyond the scope of this work.

Afterward, following the two-timescale setup, two types of variables can be specified as:

(i) Large-timescale variables: $x_{s,j} \in \{0, 1\}$ and $z_{s,j} \in \{0, 1\}$, indicating whether or not that transmitter j belongs to the serving cluster of group s, and whether or not that transmitter j caches the content requested by group s, respectively. That is, $x_{s,j} \in \{0, 1\}$ indicates a clustering variable: if $x_{s,j} = 1$, then transmitter j is within the cluster of s; and $x_{s,j} = 0$ vice versa

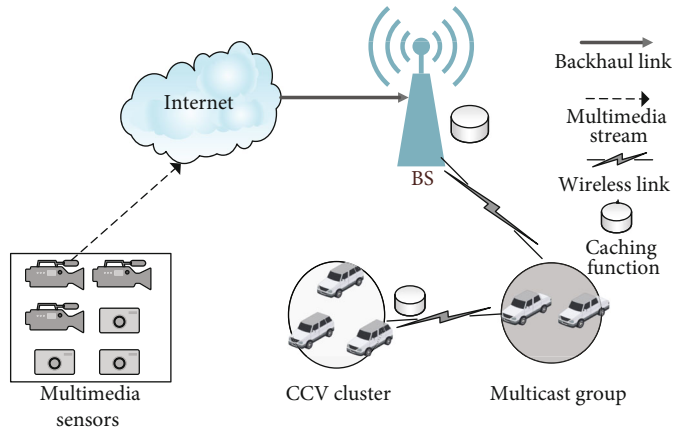

Figure 1: System model of vehicular edge networks.

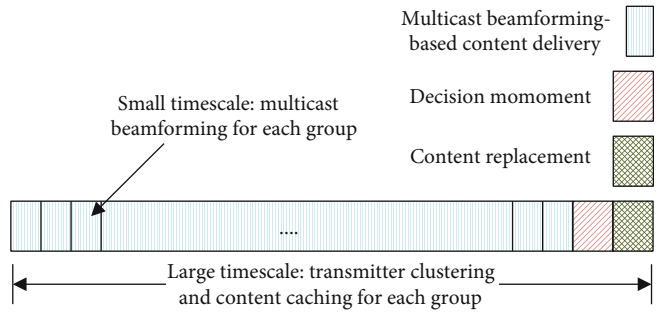

Figure 2: Two-timescale setup.

(ii) Small-timescale variables: $\mathbf{w}_{s,0}(t) \in \mathbb{C}^{L\times 1}$ (or $w_{s,j}(t) \in \mathbb{C}, j \in \mathcal{J}$), denoting the per-slot multicast beamforming vector of the BS (or CCV j) for group s

The two-timescale setup along with its two types of variables necessitates a network cost involving both the flow and power costs, which will be introduced in the following, respectively.

3.2. Caching Model. At the end of large timescale, both the clustering variable $\{x_{s,j}\}$ and caching variable $\{z_{s,j}\}$ have to be designed for the next large timescale. It is straightforward that if the content is not cached in the network, then it must be retrieved via the backhaul link from the Internet, as shown in Figure 1, thus incurring flow cost in the backhaul link or transport networks. The data rate of fetching content from core networks should be as least as the transmission rate of its associated multicast group. Following the fixed transmission rate R_s in [12] for each multicast group s, the total flow passing through the backhaul link on the BS can be represented as $\sum_{s\in\mathcal{S}}(1 - z_{s,0})R_s$. In particular, when $z_{s,0} = 1$, the immediate transmission from the BS to CRV i does not incur any flow cost on the backhaul link, and $z_{s,0} = 0$ vice versa. Besides, different from the BS, CCVs could not directly retrieve the content through the backhaul link from the Internet or core networks. Without caching, CCVs have to resort to the BS for the content retrieval, which would claim a dedicated frequency band to avert the interference to CRVs, thereby contributing to the flow cost as well. As such, the flow cost should involve both the backhaul cost and frequency bandwidth cost, and the latter one could be equivalently translated to the former one in terms of fixed transmission rate R_s. Till now, the total flow cost (on the backhaul link and dedicated frequency band) can be represented as

$$C_F = \sum_{s\in\mathcal{S}} \sum_{j\in\mathcal{J}_0} (1 - z_{s,j}) R_s. \tag{1}$$

Then, for the clustering variable $x_{s,j}$, it is irrational to cache the requested content in transmitter j when it is not picked to join the cluster of multicast group s. That is, $z_{s,j}$ could only take value zero when $x_{s,j} = 0$ holds, which can be formulated as

$$z_{s,j} \le x_{s,j}, \forall s, j. \tag{2}$$

Next, also revealed in Equation (2), $z_{s,j}$ tends to take value one to save the flow cost, given $x_{s,j} = 1$. Yet, finite storage size in CCVs constrains the cached content number, namely, only parts of contents can be accommodated in each transmitter j, written as

$$\sum_{s\in\mathcal{S}} z_{s,j} \le C_j, \forall j, \tag{3}$$

which indicates the caching capacity constraint C_j with a normalized content size of one.

Finally, although the caching function does also incur costs for both BS and CCVs, we only focus on the trade-off between backhaul cost and power cost, and the study on caching cost is beyond the scope of this work.

3.3. Multicast Model. The multicast beamforming vector is tailored per slot. For simplicity, let $\mathbf{w}_s(t) = [\mathbf{w}_{s,0}^H(t), \mathrm{w}_{s,1}^H(t), \cdots, \mathrm{w}_{s,J}^H(t)]^H$ be the network-wide beamforming vector for group s from all transmitters [32]. Furthermore, for each CRV $i \in \mathcal{I}_s$, $\mathbf{h}_i(t) \in \mathbb{C}^{(L+J)\times 1}$ denotes as the aggregate channel vector from all transmitters to CRV i. As such, the received signal-to-noise-ratio (SINR) for CRV i at slot t can be written as

$$\mathrm{SINR}_i(t) = \frac{\left|\mathbf{h}_i^H(t)\mathbf{w}_s(t)\right|^2}{\sum_{s'\neq s}\left|\mathbf{h}_i^H(t)\mathbf{w}_{s'}(t)\right|^2 + \sigma_i^2}, \forall i \in \mathcal{I}_s, s, \quad (4)$$

where σ_i^2 is the additive white Gaussian noise power per CRV and $\sum_{s'\neq s}|\mathbf{h}_i^H(t)\mathbf{w}_{s'}(t)|^2$ is the intergroup interference from any other multicast group $s' \neq s$. For each group s, it is ensured that the achievable data rate of any CRV i is no smaller than its group's fixed transmission rate R_s, namely,

$$B\log_2(1 + \mathrm{SINR}_i(t)) \geq R_s, \forall i \in \mathcal{I}_s, \quad (5)$$

which can be further recast as

$$\mathrm{SINR}_i(t) \geq 2^{R_s/B} - 1, \forall i \in \mathcal{I}_s, \mathrm{s}, \quad (6)$$

where B denotes the frequency bandwidth.

Furthermore, considering any pair of transmitter j and multicast group s, if j does not belong to the serving clustering of s, then it is irrational to design a nonzero beamforming vector for s. That is to say, if $x_{s,j} = 0$ holds, then we have $\mathbf{w}_{s,0}(t) = \mathbf{0}$ (or $\mathrm{w}_{s,j}(t) = 0$) at any time slot t, which can be represented as

$$\begin{aligned}(1 - x_{s,0})\mathbf{w}_{s,0}(t) &= 0, \forall s, t\\ \left(1 - x_{\mathrm{s},j}\right)\mathrm{w}_{s,j}(t) &= 0, \forall s, j \in \mathcal{J}, t.\end{aligned} \quad (7)$$

Moreover, since each transmitter is with the maximum transmit power threshold, the peak power budget E_j per transmitter should be imposed as

$$\begin{aligned}\sum_{s\in\mathcal{S}}\left|\left|\mathbf{w}_{s,0}(t)\right|\right|_2^2 &\leq E_0, \forall t,\\ \sum_{s\in\mathcal{S}}\left|\left|\mathrm{w}_{s,j}(t)\right|\right|_2^2 &\leq E_j, \forall j \in \mathcal{J}, t.\end{aligned} \quad (8)$$

In addition, summing over the transmit power of all network-wide beamforming vectors, the total power cost can be computed as

$$C_P = \frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}||\mathbf{w}_s(t)||_2^2, \quad (9)$$

which is averaged over all time slots in one large timescale. In particular, C_P is exactly defined as the time average of small-timescale transmit power cost $\sum_{s\in\mathcal{S}}||\mathbf{w}_s(t)||_2^2$.

3.4. Problem Formulation. Finally, involving both the flow cost and power cost, the overall cost minimization problem can be formulated as

$$\begin{aligned}\mathscr{P}_0 : \min\ & C_F + \eta C_P\\ \text{s.t.}\ & (2), (3), (6), (7), (8)\\ \text{var}\ & x_{s,j}, z_{s,j} \in \{0, 1\}, \forall s, j,\\ \text{var}\ & \mathbf{w}_s(t), \forall s, t,\end{aligned} \quad (10)$$

where η is a coefficient to balance the trade-off between flow and power costs and could be regulated artificially in line with the price of backhaul link and transmit power.

4. Joint Clustering, Caching, and Multicast Beamforming Algorithm

In this section, we first present the challenges to solve problem $\mathscr{P}_0$ and then leverage the SAA technique to decouple the two-timescale problem into one large-timescale problem and a series of independent small-timescale subproblems.

4.1. Algorithm Design Challenges. We desire to solve $\mathscr{P}_0$ at the end of each large timescale for the next one. Nevertheless, it poses challenges to solve $\mathscr{P}_0$ due to the following observations:

(i) The channel vectors $\{\mathbf{h}_i(t)\}_{\forall t\in\mathcal{T}}$ are unknown for the operator, since all channel vectors are future ones in the next large timescale

(ii) Both $\mathbf{x} = \{x_{s,j}\}$ and $\mathbf{z} = \{z_{s,j}\}$ are binary variables, rending $\mathscr{P}_0$ an MINLP problem

(iii) Even though $\mathbf{x}$ and $\mathbf{z}$ are relaxed, both Equations (6) and (7) are still nonconvex constraints

Thus, to make $\mathscr{P}_0$ tractable, an approximation approach is a prerequisite to solve it.

4.2. SAA-Based Cost Minimization. The channel vectors $\{\mathbf{h}_i(t)\}_{\forall t\in\mathcal{T}}$ are exactly stochastic at each decision moment, since they are scattered in the next large timescale and thus unpredictable. As in [29, 30], the SAA technique is leveraged to approximate the random variables $\{\mathbf{h}_i(t)\}_{\forall t\in\mathcal{T}}$, with the basic principle assuming that $\{\mathbf{h}_i(t)\}_{\forall t\in\mathcal{T}}$ is drawn from a certain distribution. As such, Equation (9) can be deemed as the time average of T random variables, and its expectation can be approximately computed as

$$E_{\hat{\mathbf{h}}}\left(\sum_{s\in\mathcal{S}}||\widehat{\mathbf{w}}_s||_2^2\right) \approx \frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}||\mathbf{w}_s(t)||_2^2, \quad (11)$$

where $\widehat{\mathbf{h}}$ is the stochastic channel vector space, and $\widehat{\mathbf{w}}_s$ denotes the $\widehat{\mathbf{h}}$-based beamforming vector. Then, by leverage the SAA, a series of samples are generated, and the expectation (11) can be approximated by its sample average. To distinguish the sample from real channel vectors per time slot, v is utilized to denote the sample index, and a total of samples are produced. Therefore, by substituting t with v at the decision moment of each large timescale, an approximate problem of $\mathscr{P}_0$ can be recast as

$$\begin{aligned}\mathscr{P}_1 : \min\ & C_F + \frac{\eta}{V}\sum_{v\in\mathscr{V}}\sum_{s\in\mathcal{S}}||\mathbf{w}_s(v)||_2^2 \\ & \text{s.t. } (2), (3), (6), (7), (8) \\ & \text{var } x_{s,j}, z_{s,j} \in \{0,1\}, \forall s, j, \\ & \text{var } \mathbf{w}_s(v), \forall s, v,\end{aligned} \tag{12}$$

where $\mathbf{h}_i(v)$ is the v-th sample from a certain distribution, and $\mathbf{w}_s(v)$ is the $\mathbf{h}_i(v)$-based beamforming vector.

From Equation (12), it follows that $\mathscr{P}_1$ turns out to be a deterministic problem rather than a stochastic one. Nevertheless, solving $\mathscr{P}_1$ still poses challenges, since both Equations (2) and (6) are nonconvex constraints. Furthermore, small-timescale variable $\mathbf{w}_s(v)$ and large-timescale variable $x_{s,j}$ are still tied in Equation (7). For this problem with two types of coupled variables, an intuitive approach is to decouple them firstly and then optimize them separately. Thus, we would propose an iterative algorithm, with the outline as follows:

(i) First, given any feasible $\mathbf{w}$, search for the optimal $\{\mathbf{x}^*, \mathbf{z}^*\}$ in $\mathscr{P}_1$

(ii) Then, given $\{\mathbf{x}^*, \mathbf{z}^*\}$, search for the optimal $\mathbf{w}^*$ with the sample $\{\mathbf{h}_i(\mathrm{v})\}_{\forall i,v}$

(iii) Repeat aforementioned procedures until convergence

(iv) With the acquired $\{\mathbf{x}^*, \mathbf{z}^*\}$, the optimal $\{\mathbf{w}_s(t)\}_{\forall s,t}$ for problem $\mathscr{P}_0$ can be obtained by solving the following multicast beamforming problem per slot as follows:

$$\begin{aligned}\mathscr{P}_2 :\ & \min\ \frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{s\in\mathcal{S}}||\mathbf{w}_s(t)||_2^2 \\ & \text{s.t. } (6), (7), (8) \\ & \text{var } \mathbf{w}_s(t), \forall s, t.\end{aligned} \tag{13}$$

By comparing $\mathscr{P}_1$ and $\mathscr{P}_2$, it reaches the conclusion that $\mathscr{P}_1$ is a large-timescale problem with sample channel vectors, while $\mathscr{P}_2$ reduces to a small-timescale one with actual channel vectors. In the next, we would separately introduce approaches to solve them.

4.3. Joint Clustering and Caching Algorithm. Given any feasible $\mathbf{w}$, any group-transmitter pair with nonzero beamforming vector can be determined. Define the set $\mathscr{K}_1 = \{(s,j) \mid \mathbf{w}_{s,0}(v) \neq \mathbf{0} \text{ or } \mathrm{w}_{s,j}(v) \neq 0, \exists \mathrm{v} \in \mathscr{V}\}$. Following Equation (7), if $\mathbf{w}_{s,0}(v) \neq \mathbf{0}$ or $\mathrm{w}_{s,j}(v) \neq 0$, $\exists v \in \mathscr{V}$, then $x_{s,j} = 1$ must hold. On the contrary, although $\mathbf{w}_{s,0}(v) = \mathbf{0}$ or $\mathrm{w}_{s,j}(v) = 0, \forall v$ cannot straightforwardly reach the conclusion with $x_{s,j} = 0$ from Equation (7), we can claim that $x_{s,j} = 0$ holds, namely, j does not need to belong to the serving cluster of s when $\mathbf{w}_{s,0}(v) = \mathbf{0}$ or $\mathrm{w}_{s,j}(v) = 0$, $\forall v$. Likewise, define the set $\mathscr{K}_2 = \{(s,j) \mid \mathbf{w}_{s,0}(v) = \mathbf{0} \text{ or } \mathrm{w}_{s,j}(v) = 0, \forall v \in \mathscr{V}\}$.

As such, the clustering variable $\{x_{s,j}\}$ is fixed given any feasible $\mathbf{w}$, and $\mathscr{P}_1$ reduces to the following large-timescale problem as

$$\begin{aligned}P_{1-I} : \min\ & C_F \\ & \text{s.t.} (2), (3), \\ & x_{s,j} = 1, \forall (s,j) \in K_1, \\ & x_{s,j} = 1, \forall (s,j) \in K_2, \\ & \text{var } x_{s,j}, z_{s,j} \in \{0,1\} \forall s, j,\end{aligned} \tag{14}$$

which exactly involves only the caching variable $\{z_{s,j}\}$, since $\{x_{s,j}\}$ is determined.

It follows that $\mathscr{P}_{1-I}$ is a typical 0-1 ILP problem. On one hand, the celebrated cutting-plane or branch-and-bound methods have been extensively utilized to solve it, yet with the computational complexity scaled with S and J. Thus, solving it is prohibitively complicated in a larger-sized network. On the other hand, there are also extensive works (e.g., in [2]) to leverage heuristic algorithms to solve the ILP problem. Nevertheless, the optimality cannot be ensured in this case, and it poses challenges to analyse the gap between the heuristic and optimality as well.

Motivated by our previous works in [33, 34], we resort to the LP counterpart of $\mathscr{P}_{1-I}$. If the equivalence of optimal solutions between $\mathscr{P}_{1-I}$ and its LP counterpart can be established, then we claim that the optimal solution to LP counterpart is also integer-valued and optimal to $\mathscr{P}_{1-I}$ as well.

First, with relaxed $z_{s,j}$, $\mathscr{P}_{1-I}$ can be recast as

$$\begin{aligned}P_{1-R} : \min\ & C_F \\ & \text{s.t.} (2), (3), \\ & x_{s,j} = 1, \forall (s,j) \in K_1, \\ & x_{s,j} = 0, \forall (s,j) \in K_2, \\ & \text{var } 0 \le z_{s,j} \le 1, \forall s, j,\end{aligned} \tag{15}$$

where the clustering variable $\{x_{s,j}\}$ is eliminated on the basis of $\mathscr{K}_1$ and $\mathscr{K}_2$ for conciseness.

It is widely known that the optimal solution of LP must locate on the vertex of the polyhedron for a feasible set. Thus, what we only need to do next is to establish that any vertex of the polyhedron of $\mathscr{P}_{1-R}$ is integer-valued. As such, the

optimal solution to $\mathscr{P}_{1-R}$ must be integer-valued and is also optimal to $\mathscr{P}_{1-I}$. Henceforth, we further provide a sufficient condition, under which the relaxed $\mathscr{P}_{1-R}$ has the integer-valued optimal solution.

Theorem 1. *If each transmitter's storage capacity is the integer multiplier of content size, i.e., $C_j \in \mathbb{Z}, \forall j \in \mathscr{J}_0$, then the optimal solution to $\mathscr{P}_{1-R}$ is integer-valued.*

Proof. From [35], it follows that if the constraint matrix is totally unimodular, then the optimal solution is integer-valued and locates at the vertex of the polyhedron. Therefore, we prove this theorem by establishing the total unimodularity of the constraint matrix of $\mathscr{P}_{1-R}$. By introducing a caching variable vector $\mathbf{Y} = \{y_{1,0}, \cdots, y_{S,0}, y_{1,1}, \cdots, y_{S,1}, \cdots, \cdots, y_{1,J}, \cdots, y_{S,J}\}^T$ and a storage capacity vector $\mathbf{C} = \{C_0, C_1, \cdots, C_j\}^T$, Equation (3) can be recast as $\mathbf{WY} \leq \mathbf{C}$, and $\mathbf{W}$ turns out to be as

$$\begin{aligned}\mathbf{W} &= \left[\mathbf{W}_0\, \mathbf{W}_1\, \mathbf{W}_2\, \cdots\, \mathbf{W}_j\right] \\ &= \begin{pmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\ * & * & \vdots & * & * & * & \vdots & * & \ddots & * & * & \vdots & * \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix}\end{aligned} \tag{16}$$

where each $\mathbf{W}_j$ is with the dimension of $(J+1) \times S$. In particular, the $(j+1)$-th row of $\mathbf{W}_j$ is all one and all zero for the other rows. Since each column of $\mathbf{W}_j$ contains only one nonzero entry belonging to $\{0,+1,-1\}$, we claim that the constraint matrix $\mathbf{W}$ meets the four subconditions of total unimodularity in [35]. As such, $\mathbf{W}$ is totally unimodular, and thus the optimal solution of $\mathscr{P}_{1-R}$ is integer-valued.

The proof is completed.

With Theorem 1, we can effortlessly solve $\mathscr{P}_{1-I}$ via classical simplex method or ellipsoid algorithm, due to the equivalence of optimal solutions between $\mathscr{P}_{1-R}$ and $\mathscr{P}_{1-I}$, thus significantly reducing the computational complexity. Although the classical cutting-plane or branch-and-bound algorithm can also work for the ILP, they are only applicable to small- and moderate-sized vehicular edge networks. When it comes to a large-sized network, Theorem 1 exhibits more effectiveness than legacy ones.

4.4. SCA-Based Multicast Beamforming. Once the aforementioned $\{\mathbf{x}^*, \mathbf{z}^*\}$ is given, $\mathscr{P}_1$ reduces to the following one as

$$\begin{aligned}\mathscr{P}_{1-S}: \ &\min \ \frac{1}{V}\sum_{v\in\mathscr{V}}\sum_{s\in\mathscr{S}}||\mathbf{w}_s(v)||_2^2 \\ &\text{s.t. } (6), (7), (8) \\ &\text{var } \mathbf{w}_s(v), \forall s, v,\end{aligned} \tag{17}$$

which is exactly the same as $\mathscr{P}_2$ with the only difference in substituting the index t in $\mathscr{P}_2$ with v in $\mathscr{P}_{1-S}$. Notice that, $\mathscr{P}_2$ is the small-timescale multicast beamforming problem with actual channel vectors $\{\mathbf{h}_i(t)\}_{\forall t\in\mathscr{T}}$, while $\mathscr{P}_{1-S}$ is solved at the large timescale with channel samples $\{\mathbf{h}_i(v)\}_{\forall v\in\mathscr{V}}$, just to proceed with the algorithm iteration in solving $\mathscr{P}_{1-S}$. Thus, in the following, by putting $\mathscr{P}_2$ and $\mathscr{P}_{1-S}$ together, we take $\mathscr{P}_2$ as a goal to find its approximation algorithm, and do not distinguish them strictly.

Although the objective as well as constraints 7 and 8 turn into convex ones, the nonconvex constraint 6 renders $\mathscr{P}_2$ a nonconvex one. Thus, by resorting to the SCA technique in [12], we first recast Equation (6) as

$$\begin{aligned}&\left(2^{R_s/B}-1\right)\left(\sum_{s'\neq s}\left|\mathbf{h}_i^H(t)\mathbf{w}_s{}'(t)\right|^2+\sigma_i^2\right)-\left|\mathbf{h}_i^H(t)\mathbf{w}_s(t)\right|^2 \\ &\quad\leq 0, \forall i\in\mathscr{I}_s, s, t.\end{aligned} \tag{18}$$

It is evident that the left-hand side of Equation (18) is the difference of two convex functions. Next, by substituting the second term with its first-order Taylor expansion, at the $(o+1)$-th iteration, we have

$$\begin{aligned}&2Rs/B-1\sum_{s'\neq s}\left|\mathbf{h}_i^H(t)\,\mathbf{w}_s{}'(t)\right|^2+\sigma_i^2 \\ &\quad-2\mathscr{R}\left\{\left(\mathbf{w}_s^o(t)\right)^{\mathrm{H}}\mathbf{h}_i(t)\,\mathbf{h}_i^H(t)\,\mathbf{w}_s(t)\right\}-\left|\mathbf{h}_i^H(t)\,\mathbf{w}_s^o(t)\right|^2 \\ &\leq 0, \forall i\in\mathscr{I}s, s, t,\end{aligned} \tag{19}$$

which becomes a convex constraint, with $\mathbf{w}_s^o(t)$ acting as the initial feasible solution obtained from the o-th iteration.

Till now, at the $(o+1)$-th iteration, $\mathscr{P}_2$ can be transformed to the following one as

$$\begin{aligned}\mathscr{P}_{2-C}: \ &\min \ \frac{1}{T}\sum_{t\in\mathscr{T}}\sum_{s\in\mathscr{S}}||\mathbf{w}_s(t)||_2^2 \\ &\text{s.t. } (7), (8), (19) \\ &\text{ar } \mathbf{w}_s(t), \forall s, t,\end{aligned} \tag{20}$$

which is a convex quadratically constrained quadratic programming (QCQP) problem and can be solved via many mature convex algorithms effortlessly, e.g., interior-point method [36]. At each iteration, a convex QCQP problem needs to be solved until convergence. From Equation (19), it follows that the obtained $\mathbf{w}_s^o(t)$ from the o-th iteration is always feasible to the subproblem at the $(o+1)$-th iteration. Besides, since the objective of $\mathscr{P}_{2-C}$ is the minimization of power cost, $\mathbf{w}_s^{o+1}(t)$ must be a better solution than $\mathbf{w}_s^o(t)$, revealing the monotonicity of objective.

4.5. Algorithm Outline and Computational Complexity. At the end of a large timescale, the operator has to iteratively resolve problems $\mathscr{P}_{1-S}$ and $\mathscr{P}_{1-I}$, until the termination

Initialization: $x_{s,j} = 1,\ \forall s \in \mathcal{S}, \forall j \in \mathcal{J}_0$ (Full cooperative multicasting is as the initial solution).
Repeat
Solve problem $\mathcal{P}_{1-S}$ via SCA to acquire $\mathbf{w}^*$;
Obtain $\mathcal{K}_1$ and $\mathcal{K}_2$ from Equation (14);
Solve problem $\mathcal{P}_{1-I}$ via Theorem 1 to acquire $\{\mathbf{x}^*, \mathbf{z}^*\}$.
Until the error tolerance is met or maximum iterations are reached.
for $t = 1, 2 \cdots, T$
Solve problem $\mathcal{P}_2$ via SCA with the available $\{\mathbf{x}^*, \mathbf{z}^*\}$.
end for

Algorithm 1: Joint V2V-assisted clustering, caching, and multicasting algorithm.

criteria are satisfied, namely, one convergent solution is acquired or the maximum iteration number is reached, as described in Algorithm 1. Then, with the available $\{\mathbf{x}^*, \mathbf{z}^*\}$ from large timescale, the operator proceeds to tailor the multicast beamforming per slot, to solve $\mathcal{P}_2$ via SCA for the optimal $\mathbf{w}_s^o(t)$.

From Algorithm 1, it follows that the SCA is a prerequisite at both timescales. Assume the maximum iteration numbers for joint V2V clustering and caching algorithm, and SCA method reach $\varrho_{\max}$ and $o_{\max}$, respectively. From [37], it follows that the interior point method to solve each per-iteration subproblem for $\mathcal{P}_{1-S}$ or $\mathcal{P}_2$ is with the computational complexity of $\Phi = \mathcal{O}((J+L)^3 S^3 V^3)$ or $\Phi = \mathcal{O}((J+L)^3 S^3 T^3)$. Thus, the overall computational complexity reaches $\mathcal{O}((\varrho_{\max}+1) o_{\max} \Phi) = \mathcal{O}(\varrho_{\max} o_{\max} \Phi)$.

5. Simulation Results and Discussions

In this section, we demonstrate the performance of proposed joint V2V-assisted clustering, caching, and multicast beamforming algorithm in vehicular edge networks via simulation results. In particular, the impact of the following two parameters is studied: (1) the number of CCVs; and (2) the average data rate requirement per CRV. Meanwhile, two costs are leveraged as performance metrics: (1) total flow cost; and (2) total transmit power cost. In addition, for performance comparison, three benchmark schemes are also evaluated, listed as follows:

(i) Multicast without (w.o.) V2V caching refers to the multicast beamforming scheme proposed in [11], which only involves the BS caching but neglecting the V2V caching. For fair comparison, the two-timescale setup is also used in this scheme

(ii) Unicast with (w.) V2V caching refers to the unicast transmission scheme proposed in [23], where both V2V and BS caching are involved. In this scheme, each CRV is assigned an individual beamforming vector, regardless of its requested content

(iii) Unicast w.o. V2V caching. In this scheme, each CRV has to access the BS to fetch its requested content

5.1. Simulation Parameters. We investigate a time-slotted wireless network consisting of one BS and 20 CRVs, with a radius of 500 m for the cell coverage. The BS is equipped with 20 antennas, and each CRV and CCV is identically equipped with one antenna. Overall, 20 types of multimedia contents exist in the system. The BS has a storage size of 6 contents, while each CCV could accommodate at most 1 content. For simplicity, we do not employ the well-known Zipf distribution in [2]; instead, each CRV equally requests any content with a probability of 5%.

The system bandwidth is 5 MHz, and the additive white Gaussian noise power spectral is -174 dBm/Hz. The transmit power budgets for the BS and CCV are 46 dBm and 24 dBm, respectively. The path loss model is $35.3 + 37.6 \log_2(d(m))$, the log-normal shadowing parameter is 8 dB, and the multipath channel model with Rayleigh fading is assumed [20]. Overall, 200 samples are produced to simulate stochastic channels vectors. Besides, the Monte Carlo approach is utilized, both CRVs and CCVs are uniformly distributed and dropped in the cell coverage, and all simulation results are averaged over 1000 droppings. In addition, the vehicle speed follows a truncated normal distribution ranging from 15 m/s and 31 m/s.

5.2. Convergence Performance. Figure 3 illustrates the convergence performance of joint V2V clustering and caching algorithms. In this setup, the number of CCVs is 20, and the data rate requirement per group is 2 Mbps. To overcome the impact of η with too large or small values on the objective of $\mathcal{P}_0$ and for fair comparison, only the system power cost is evaluated. First, it can be observed from Figure 3 that the joint V2V clustering and caching algorithm tends to saturate within 30 iterations. In spite of moderate convergence, the result is acceptable since it is operated at a large timescale and does not need to work per small timescale slot. Second, with the increase of η, the total power cost gradually reduces. This is because the objective of $\mathcal{P}_0$ is the minimization of system cost; a larger η will increase the weight of power cost, thus reducing the power proportion in the optimal value.

Figure 4 reveals the convergence of the SCA-based multicast beamforming algorithm. Likewise, the number of CCVs is 20, and the data rate requirement per group is 2 Mbps. We leverage the L_2 norm on the difference of successive optimal solutions, i.e., $||\mathbf{w}_s^{o+1}(t) - \mathbf{w}_s^o(t)||_2$ as the metric. As shown in Figure 4, all settings under different values of η have good convergence, saturating within 10 steps.

5.3. Performance Comparison. Figure 5 shows the power-flow cost tradeoff curve of the proposed algorithm under different

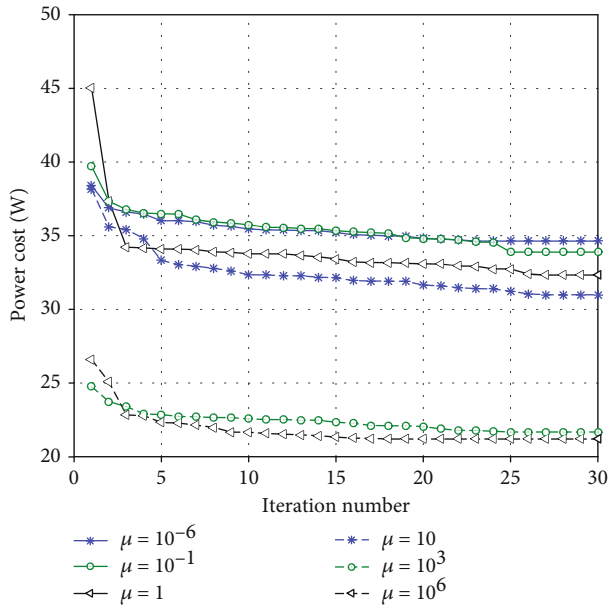

Figure 3: The convergence of the proposed algorithm.

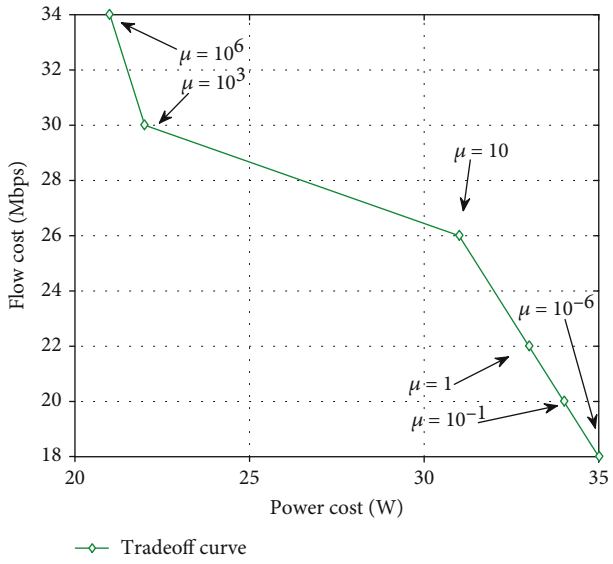

Figure 5: Flow-power cost tradeoff.

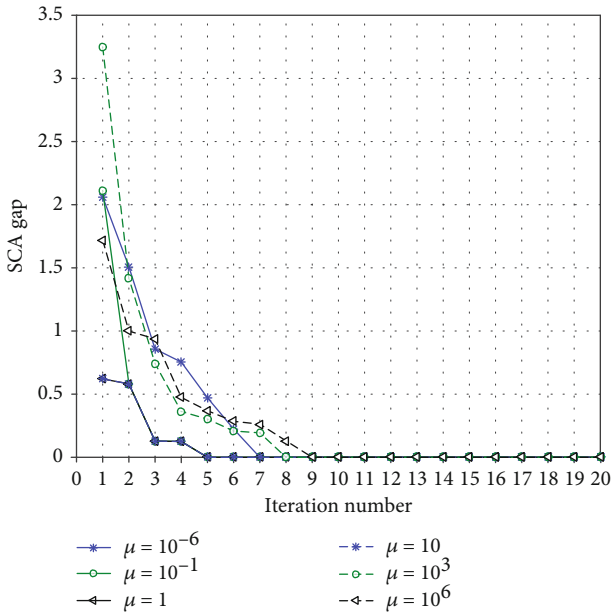

Figure 4: The convergence of the SCA method.

values of η. Similar to the result in Figure 3, when η approaches infinity, the proportion of power cost is getting larger, and thus the power cost decreases in the minimization of the weighted sum. In particular, the flow cost reduces to 21 W while the flow cost reaches 34 Mbps, given $\eta = 10^6$. On the contrary, when η approaches zeros, the proportion of power cost gets smaller, and more weights are imposed on the flow cost. In particular, the flow cost increases to 35 W while the flow cost reduces to 18 Mbps, given $\eta = 10^{-6}$.

Figure 6 shows the impact of CCV number on the flow cost. In this scenario, the average data rate requirement is 2 Mbps, $\eta = 10^3$, and the CRV number ranges from 15 to 30. As shown in Figure 6, the curve is broken line-shaped, since the flow cost must be the integer multiplier of the data rate requirement. The flow cost reduces with the increase of CCV number. This is because the wireless network incorporating more CCVs would introduce multiuser diversity gain and a more flexible clustering combination. In the unicast w.o. V2V caching scheme, the caching functions are only available in the BS, and each CCV provided a requested content individually, thus resulting in the largest flow cost. In addition, unicast w. V2V caching outperforms multicast w.o. V2V caching in terms of flow cost, indicating that the V2V caching gains overwhelm the multicasting gains in this setting, especially when the CCV number is relatively large.

Figure 7 shows the impact of CCV number on the power cost, with identical settings as in Figure 6. Likewise, unicast w.o. V2V caching still gets the largest power cost compared to the other three schemes. Meanwhile, unicast w. V2V caching is also with worse performance compared to multicast w.o. V2V caching. This is because multicast w.o. V2V caching tends to save more transmit power than unicast w. V2V caching, while the latter one must allocate an individual beamformers to each CCV. The performance gap also indicates the multicast gains over unicast. Furthermore, the proposed algorithm gets close performance with multicast w.o. V2V caching, only with a slightly small gap, since in this setting, multicast gains overwhelm V2V caching gains.

In Figure 8, we compare the performance with different data rate requirements R_s, ranging from 2 Mbps to 3.8 Mbps. In this scenario, there are 20 CCVs, and $\eta = 10^3$. Figure 8 shows the flow cost is almost proportional to the data rate requirement. In the meantime, there exists a significant gap between the proposed algorithm and the other three schemes (approximately 30 Mbps on average compared to multicast w.o. V2V caching), which can be explained as follows: on one hand, the unicast transmission would result in more flow costs, since each CRV is equipped with a CCV cluster and a content flow; on the other hand, the multicast w.o. V2V

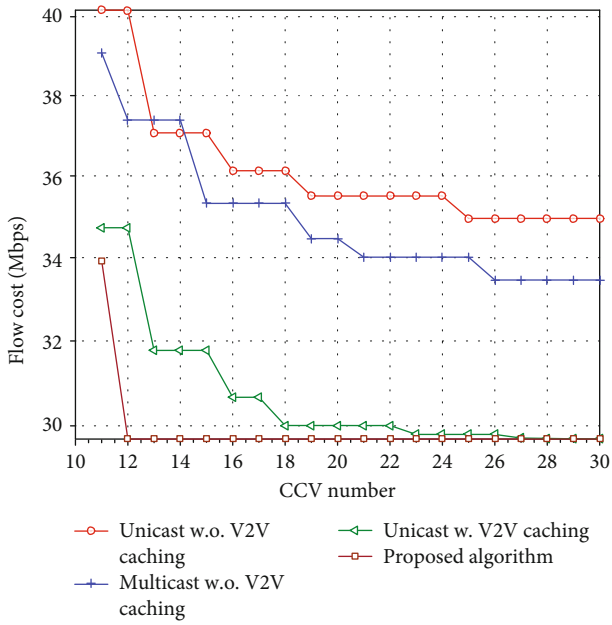

Figure 6: Flow cost versus the CCV number.

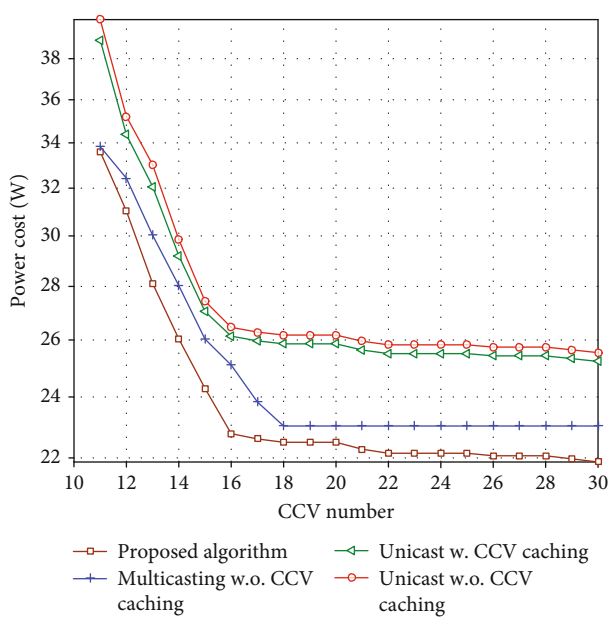

Figure 7: Power cost versus the CCV number.

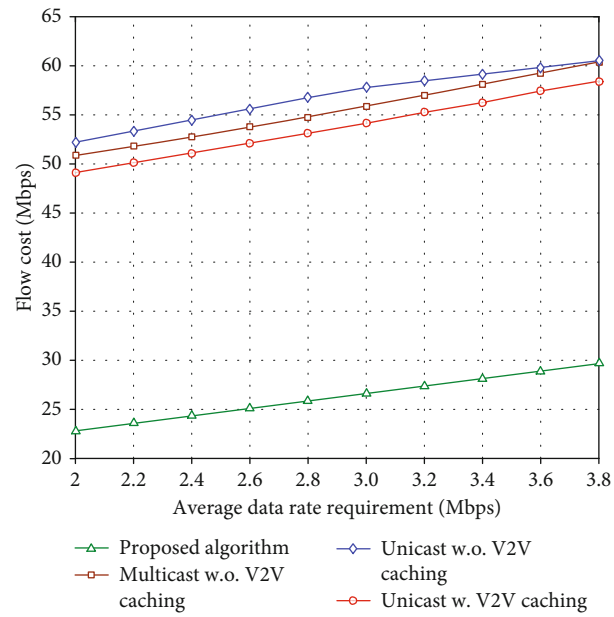

Figure 8: Flow cost versus the average data rate requirement.

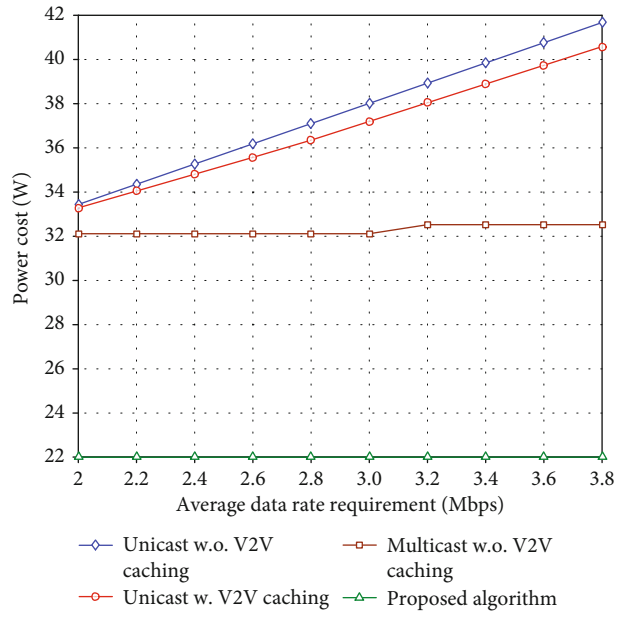

Figure 9: Power cost versus the average data rate requirement.

caching neglects the caching functions in CCVs, thus incurring the traffic congestion.

Figure 9 shows the impact of data rate requirement on power cost, with identical settings as in Figure 8. Different from the results in Figure 8, both the proposed algorithm and multicast w.o. V2V almost remain unchanged with the increase of data rate requirements. This is because in this setting, sufficient power is available to support a larger data rate, and thus multicast transmissions could save more power than unicast. Meanwhile, the performance gap between the proposed algorithm and multicast w.o. V2V caching reveals that, the lack of V2V caching would incur that all CRVs tend to access the BS, the power resource of CCVs is wasted, and thus a worse feasible solution is produced. In addition, unicast would allocate an individual beamformer to each CRV, thus getting the worst performance.

6. Conclusions

In this study, wireless content caching along with multicast beamforming was studied in vehicular edge networks. First, a two-timescale setup was proposed to decouple the joint clustering, caching, and multicast beamforming problem at separate timescales. Next, at the large timescale, with the

SAA technique, a theorem was verified to reveal the equivalence of optimal solutions between the original ILP problem and its relaxed LP counterpart. Then, with the SCA method, the multicast beamforming-based power minimization problem was solved per slot. Simulation results revealed that the network performance could benefit from not only the power saving from wireless multicast beamforming but also the content caching and sharing among vehicles.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61801379, 61901011, and 61971347, in part by the Open Research Fund from the State Key Laboratory of Integrated Services Networks (Xidian University) under Grant ISN21-08, and the Youth Innovation Team of Shaanxi Universities.

References

[1] S. A. Alvi, B. Afzal, G. A. Shah, L. Atzori, and W. Mahmood, "Internet of multimedia things: vision and challenges," *Ad Hoc Networks*, vol. 33, pp. 87–111, 2015.

[2] C. Liang, Y. He, F. R. Yu, and N. Zhao, "Enhancing video rate adaptation with mobile edge computing and caching in software-defined mobile networks,," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 7013–7026, 2018.

[3] Y. Mei, F. Li, L. He, and L. Wang, "Joint source and channel rate allocation over noisy channels in a vehicle tracking multimedia internet of things system," *Sensors*, vol. 18, no. 9, article 2858, 2018.

[4] J. Zheng and Q. Wu, "Performance modelling and analysis of the IEEE 802.11p EDCA mechanism for VANET," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2673–2687, 2016.

[5] L. Zhu, Y. Li, F. R. Yu, B. Ning, T. Tang, and X. Wang, "Cross-layer defense methods for jamming-resistant CBTC systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2020.

[6] L. Zhu, Y. He, F. R. Yu, B. Ning, T. Tang, and N. Zhao, "Communication-based train control system performance optimization using deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 10705–10717, 2017.

[7] Z. Su, Y. Hui, Q. Xu, T. Yang, J. Liu, and Y. Jia, "An edge caching scheme to distribute content in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5346–5356, 2018.

[8] M. Fallgren, T. Abbas, S. Allio et al., "Multicast and broadcast enablers for high-performing cellular V2X systems," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 454–463, 2019.

[9] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 719–728, 2012.

[10] G. Xylomenos, C. N. Ververidis, V. A. Siris et al., "A survey of information-centric networking research," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 1024–1049, 2014.

[11] X. Duan, Y. Liu, and X. Wang, "SDN enabled 5G-VANET: adaptive vehicle clustering and beamformed transmission for aggregated traffic," *IEEE Communications Magazine*, vol. 55, no. 7, pp. 120–127, 2017.

[12] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast Beamforming for cache-enabled cloud RAN," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6118–6131, 2016.

[13] Y. Cui, Z. Wang, Y. Yang, F. Yang, L. Ding, and L. Qian, "Joint and competitive caching designs in large-scale multi-tier wireless multicasting networks," *IEEE Transactions on Communications*, vol. 66, no. 7, pp. 3108–3121, 2018.

[14] H. T. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and W. J. Hwang, "Collaborative multicast beamforming for content delivery by cache-enabled ultra dense networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3396–3406, 2019.

[15] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Cache-aware multicast beamforming design for multicell multigroup multicast," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11681–11693, 2018.

[16] Y. Dong, M. J. Hossain, J. Cheng, and V. C. M. Leung, "Cross-layer scheduling and beamforming in smart-grid powered cellular networks with heterogeneous energy coordination," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 2711–2725, 2020.

[17] Y. Dong, A. el Shafie, M. J. Hossain, J. Cheng, N. al-Dhahir, and V. C. M. Leung, "Secure beamforming in full-duplex MISO-SWIPT systems with multiple eavesdroppers," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6559–6574, 2018.

[18] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4365–4380, 2016.

[19] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, 2017.

[20] A. S. Cacciapuoti, M. Caleffi, M. Ji, J. Llorca, and A. M. Tulino, "Speeding up future video distribution via channel-aware caching-aided coded multicast,," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2207–2218, 2016.

[21] T. Zhang, H. Fan, J. Loo, and D. Liu, "User preference aware caching deployment for device-to-device caching networks," *IEEE Systems Journal*, vol. 13, no. 1, pp. 226–237, 2019.

[22] W. L. Tan, W. C. Lau, O. Yue, and T. H. Hui, "Analytical models and performance evaluation of drive-thru internet systems," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 207–222, 2011.

[23] J. Zhao, X. Sun, Q. Li, and X. Ma, "Edge caching and computation management for real-time internet of vehicles: an online and distributed approach," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–15, 2020.

[24] H. Xiao, X. Zhang, A. T. Chronopoulos, Z. Zhang, H. Liu, and S. Ouyang, "Resource management for multi-user-centric V2X communication in dynamic Virtual-Cell-Based ultra-dense networks," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6346–6358, 2020.

[25] Q. Wu, H. Liu, R. Wang, P. Fan, Q. Fan, and Z. Li, "Delay-sensitive task offloading in the 802.11p-based vehicular fog computing systems," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 773–785, 2020.

[26] Y. Dong, M. Z. Hassan, J. Cheng, M. J. Hossain, and V. C. M. Leung, "An edge computing empowered radio access network with UAV-mounted FSO fronthaul and backhaul: key challenges and approaches," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 154–160, 2018.

[27] H. Zhang, Y. Dong, J. Cheng, M. J. Hossain, and V. C. M. Leung, "Fronthauling for 5G LTE-U ultra dense cloud small cell networks," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 48–53, 2016.

[28] G. Qiao, S. Leng, S. Maharjan, Y. Zhang, and N. Ansari, "Deep reinforcement learning for cooperative content caching in vehicular edge computing and networks," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 247–257, 2020.

[29] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 881–895, 2019.

[30] J. Tang, T. Q. S. Quek, T. Chang, and B. Shim, "Systematic resource allocation in cloud RAN with caching as a service under two timescales," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7755–7770, 2019.

[31] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, 2018.

[32] S. Gong, S. X. Wu, A. M. So, and X. Huang, "Distributionally robust collaborative beamforming in D2D relay networks with interference constraints," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5048–5060, 2017.

[33] K. Wang, F. R. Yu, L. Wang et al., "Interference alignment with adaptive power allocation in full-duplex-enabled small cell networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3010–3015, 2019.

[34] K. Wang, W. Ji, J. Li, H. Wang, and T. Cao, "Wireless content caching in sliced cellular networks with multicast beamforming," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Xi'an, China, 2019.

[35] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, and J. Xu, "Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1751–1767, 2018.

[36] S. Boyd and L. Vandenberghe, "Interior-point methods," in *Convex Optimization*, pp. 561–623, Cambridge University Press, Cambridge, UK, 2004.

[37] M. El-Absi, M. Shaat, F. Bader, and T. Kaiser, "Interference alignment with frequency-clustering for efficient resource allocation in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 7070–7082, 2015.