# Cost Optimization for On-Demand Content Streaming in IoV Networks With Two Service Tiers

Xuemin Hong, *Member, IEEE*, Jiping Jiao, Ao Peng, Jianghong Shi, and Cheng-Xiang Wang, *Fellow, IEEE*

***Abstract*—On-demand streaming of high-quality video content is a widely anticipated vehicular infotainment service. How to reduce the cost of content streaming is a primary concern of the service providers, but is still an underinvestigated subject in the literature. This paper proposes an integrated mobile streaming and caching scheme that jointly leverages two communication service tiers and on-board caching resource for cost reduction. Algorithms are presented to achieve optimal buffering at the session level and optimal caching at the device level. An analytical framework is established to characterize the average cost as a function of the streaming rate in a large scale network. Numerical results demonstrate how the "cost-streaming rate" function changes with vehicle density, network congestion level, content length, and average packet transmission time. We learn an important insight that there is a minimum cost threshold even when the streaming rate approaches zero. We also show that the proposed protocol can effectively reduce the overall cost when the network is not congested. Our findings can provide useful guidelines for the business planning and operation of vehicular content streaming services.**

***Index Terms*—Content streaming, Internet-of-Vehicle (IoV), service tier, video-on-demand (VoD).**

## I. Introduction

AS ONE of the major applications envisioned for the Internet of Vehicles (IoV) [1], vehicular infotainment service [2] is gaining momentum from the rapid advancements of self-driving [3] and IoV technologies. On one hand, self-driving technologies allow on-board passengers to shift their attentions from driving to infotainment services. On the other hand, the IoV not only provides accessible mobile broadband connections for vehicles but also enables an ecosystem where diverse information (e.g., user preference, transportation condition, content popularity, etc.) can be jointly processed to make the infotainment service more personalized and attractive [1]. As a result, it is widely anticipated that the vehicular infotainment service will become a ubiquitous service in the near future.

The IoV network underpinning the vehicular infotainment service has a heterogeneous nature and is expected to offer a multitude of service tiers [1]. The tiers are differentiated by their price and quality-of-service metrics such as transmission rate and delay [4]. Two heterogeneous and complementary networking paradigms have been proposed for IoV: 1) vehicle-to-vehicle (V2V) [5]–[9] and 2) vehicle-to-infrastructure (V2I) [10]–[13] communications. In V2V communications, vehicles communicate with each other directly in an *ad-hoc* fashion. It offers a communication service that is opportunistic in nature. In V2I communications, vehicles can communicate with nearby infrastructure, which may include special road-side units or cellular base stations (BSs). A V2I system can itself be designed to offer different service tiers. For example, IoV is a targeted application scenario of the fifth generation (5G) cellular communication system [13]–[15]. Using network virtualization and slicing technologies [16], the 5G-enabled IoV can offer multiple service tiers ranging from ultrareliable service for mission critical applications [17] to ultraelastic services for delay-tolerant applications [18].

On-demand content/video streaming is a major application in vehicular infotainment services. It offers a highly personal and attractive service by allowing passengers to browse a recommended content list and request the favorite contents on demand. One popular way to deliver a requested content is to download the entire content file. This is usually called content dissemination [8]–[11], [21]–[23] in the literature. Another way to deliver a content is to transmit the content as a continuous stream of data while the content is played in real time. This is called content streaming. Measurements show that mobile users tend to abort more frequently than PC users during the content viewing process [19], [20]. As a result, compared with content dissemination, content streaming can offer higher flexibility and resource efficiency. The upmost technical challenge of mobile content streaming is to deal with the versatile wireless channel [24]–[27], which may cause playback freeze and bad user experience [28]. To this end, a "Fast Start" phase is commonly implemented to rapidly fill the playback buffer when a streaming session begins. After this phase, a variety of streaming schemes can be used [28]. These schemes are different variations of a fundamental scheme called encoding rate streaming, in which the streaming rate is matched to the encoding rate (i.e., playing rate). Without loss of generality,

Manuscript received April 20, 2018; revised August 20, 2018; accepted September 20, 2018. Date of publication October 1, 2018; date of current version February 25, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61571378, in part by the National Key Research and Development Program of China under Grant 2018YFB0505202, and in part by the EU H2020 RISE TESTBED Project under Grant 734325. *(Corresponding author: Xuemin Hong.)*

X. Hong and J. Shi are with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education of China, Xiamen University, Xiamen 361005, China (e-mail: xuemin.hong@xmu.edu.cn; shijh@xmu.edu.cn).

J. Jiao and A. Peng are with the Department of Communications Engineering, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: jjp@xmu.edu.cn; pa@xmu.edu.cn).

C.-X. Wang is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: chxwang@seu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2018.2873085

this paper focuses on the fundamental scheme of encoding rate streaming.

The problem of mobile video streaming has been extensively investigated in the context of mobile phones [28]–[32]. For vehicular networks, the content streaming problem has also been widely studied [5], [33]–[36]. The main focus of research had been placed on the buffer/pre-fetching algorithms and rate adaptation algorithms. Despite these efforts, we can still identify three important research gaps.

1) Few research addressed the commercial perspective and investigated the cost issue of content streaming in the context of multiservice-tier networks. There is a bulk of literature that investigated how to aggregate streaming bandwidth from heterogeneous networks. The resulted designs include concurrent multipath transfer [37]–[39] and multihomed video streaming [40]–[42]. However, cost was not a factor of concern in these studies. Wu *et al.* [43] and He *et al.* [44] raised the issue of cost-effectiveness in mobile video streaming, but the cost is interpreted as energy and bandwidth efficiencies instead of monetary service cost.
2) Few research addressed the problem of joint buffering and caching design. Compared with mobile phone networks, a major difference in IoV is that vehicles are less restricted on cache space and energy consumption. Thus, vehicles can potentially cache a large volume of popular contents to ease the burden of online streaming. In this case, the performance of buffering and caching are inherently related. This brings new challenges in joint buffering and caching protocol design. Such a challenge has attracted some recent attention in cellular networks [45], [46], but has not been studied in the context of vehicle-based content streaming and caching.
3) Few research addressed the scaling performance of mobile streaming, i.e., how does the performance scale in large scale networks with multiple users? The scaling performance had been studied in the context of content dissemination in vehicular ad-hoc networks [47], [48]. However, the analytical frameworks therein cannot provide much insight into the distribution of packet delay, which is critical to the performance evaluation of delay-sensitive streaming service.

This paper aims to partly address the above research gaps. Taking the perspective from a vehicular infotainment service provider, we focus on how to reduce the operational expenses (OPEX) of content streaming service in an IoV network with two service tiers. Our main contributions are as follows. First, we propose a streaming scheme that jointly leverages the communication and caching resources for cost reduction. Optimal buffering and caching polices are derived for the proposed scheme. Second, using a novel analytical framework that integrates stochastic geometry models [49] and queueing models [50], the performance of the proposed scheme is analyzed in a large scale network. We demonstrate how the service cost changes with streaming rate, vehicle density, content length, and network congestion level. The benefits and fundamental limits of two-tier content streaming are also revealed.

The remainder of this paper is organized as follows. Section II introduces the application scenario and proposes a streaming protocol. Sections III and IV derive the optimal buffering and caching policies, respectively. Section V evaluates the performance in large scale networks. Finally, the conclusions are drawn in Section VI.

## II. Application Scenario and Proposed Protocol

This paper envisions a novel infotainment service called personalized vehicular media. Using interactive screens installed on vehicles, the service can recommend a personalized list of contents to on-board passengers, wait for passengers to browse the list and request interested contents, and deliver the on-demand content to users through the IoV. This is an attractive value-added service to transportation operators such as taxi companies, limousine companies, or peer-to-peer vehicle sharing companies like Uber. We further envision a commercial entity called vehicular media operator (VMO). The primary concern of the VMO is to maximize the revenue (by advertising for example) and reduce the OPEX. It is expected that the cost of wireless mobile communications will become a major source of OPEX due to the heavy mobile data consumption on a daily basis. Hence, taking a commercial perspective, this paper aims to reduce the communication cost for VMO.

We propose a novel content streaming protocol that jointly leverages the communication and caching resources for overall cost reduction. It is assumed that the underlying IoV network provides two classes of communication services: 1) a primary service and 2) a secondary service. The primary access has a high cost, low delay, and sufficiently large instant data rate. The secondary access, on the other hand, is an opportunistic service that has a low cost at the expense of unreliable delay and data rate. It is further assumed that the vehicles can cache a large volume of popular contents. Our protocol runs on a vehicular device and includes the following steps.

Step 1 *(Streaming Rate Adaptation):* Upon a user request for a content, choose a proper streaming rate based on a rate adaption policy that takes into account the cost budget and overall network traffic conditions, etc.

Step 2 *(Initial Cache Loading):* Check whether part of the requested content has been cached locally. If yes, load the cached data to buffer.

Step 3 *(Instant Buffering):* Calculate the optimal buffer length for the content. If the optimal buffer length is not reached in step 2, invoke the Fast Start procedure [28] to fill the buffer to its optimal length using the primary access link.

Step 4 *(Secondary Streaming):* Start content playback and begin encoding rate streaming [28] via the secondary access link.

Step 5 *(Complementary Primary Access):* Constantly monitors the buffer. If a packet approaching the deadline is not yet received, invoke the primary access to fetch the packet instantaneously.

The basic rationale of our protocol is to quickly buffer a portion of the content before playback. A longer buffer length implies a higher initial cost, but is compensated by a lower cost

in the latter streaming phase because the likelihood of using the primary access reduces (due to a larger delay allowance). In our protocol, it is assumed that the streaming rate is decided at the beginning of a session according to the conditions of the network. The streaming rate remains unchanged throughout a session. Our protocol can support fast streaming rate adaptation by dividing a content file into smaller segments and transmitting each segment with a new session. In this sense, our protocol is compatible with the widely used Dynamic Adaptive Streaming over HTTP protocol, which allows a content to be divided into multiple segments, while each segment can be encoded with different rates. In the remainder of this paper, we will systematically analyze the performance of the proposed protocol at the session level, device level, and network level.

## III. Optimal Buffering Policy at the Session Level

### A. Problem Formulation

This section considers the problem of optimal buffering in a single session, which deals with one user request for a single content file. We denote $\varepsilon_p$ as the cost of primary access, $\varepsilon_s$ as the cost of secondary access, $T$ as the total length (i.e., playing time) of the requested content, $T_c$ as the length of precached content, and $F_D(t)$ as the cumulative distribution function (CDF) of the random delay in the secondary access channel. These parameters are treated as known when a session starts. We note that throughout this paper, we use the term "length" to denote the duration of time. The primary access typically have a much larger maximum data rate and lower delay than the secondary access. For simplicity, we consider an ideal case where the primary access has infinite data rate and zero delay, noting that this will not change the nature of our analysis. In practice, $F_D(t)$ can be obtained by data-driven measurements and predictive modeling. In this paper, we assume that $F_D(t)$ takes the form of a Weibull distribution, which is widely used to model the delay statistics of real-world traffic [51]. The CDF of the Weibull distribution is

$$F_D(t) = 1 - e^{-\left(\frac{t}{\gamma}\right)^k} \tag{1}$$

which is characterized by a shape parameter $k$ and scale parameter $\gamma$.

Let us further define $T_W$ as the buffer length, which is our optimization variable. According to the proposed streaming protocol, the expected cost of delivering a content with length $T$ is

$$y(T_W; T_c, T) = (T_W - T_c)\varepsilon_p + (T - T_W)\left[\varepsilon_s p_\chi + \varepsilon_p\left(1 - p_\chi\right)\right] \tag{2}$$

where

$$p_\chi = F_D(T_W) \tag{3}$$

is the probability of delay outage during secondary streaming. We note that (2) reflects the cost of content delivery incurred by the proposed protocol in Fig. 1. The term $(T_W - T_c)\varepsilon_p$ is the cost incurred in the Fast Start phase, which fills the buffer length to $T_W$ via primary access. The second term

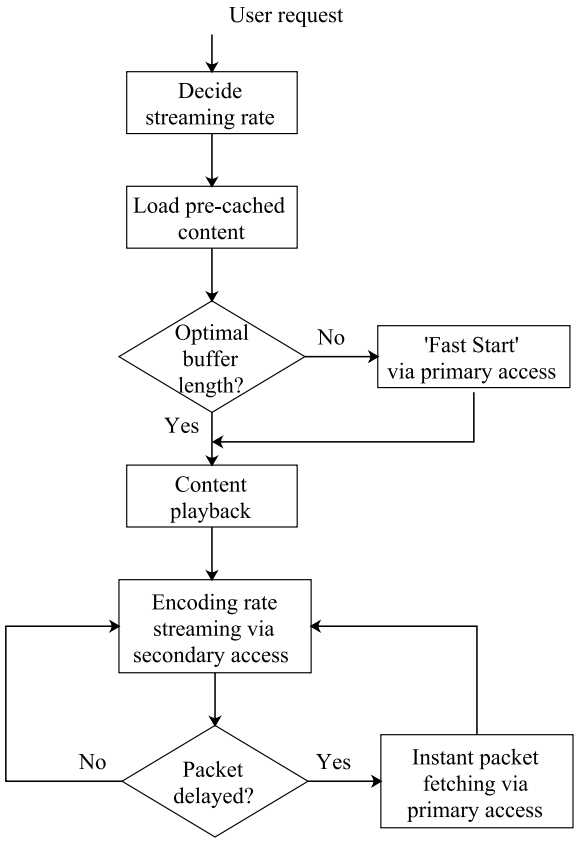



Fig. 1. Proposed content streaming protocol using two service tiers.

$(T - T_W)[\varepsilon_s p_\chi + \varepsilon_p(1 - p_\chi)]$ is the cost incurred in the "streaming phase," during which the probability of using the primary and secondary access are $1 - p_\chi$ and $p_\chi$, respectively. Equation (2) shows that the cost of content delivery is jointly determined by buffering (i.e., $T_W$), caching (i.e., $T_c$), and streaming rate (which is related to $p_\chi$). For clarity, the cost defined in (2) is normalized to the streaming rate, which is fixed at the beginning of a session and unrelated to the optimal buffer length. The problem of optimal buffering can be formulated as (P1)

$$y^*(T_c; T) = \min_{T_W} y(T_W; T_c, T) \quad \text{s.t.} \quad T_c \le T_W \le T. \tag{4}$$

Fig. 2 shows the cost per unit time (i.e., $y(T_W; T_c, T)/T$) as a function of the buffer length $T_W$ with varying settings of cache length $T_c$ and delay distribution $F_D(x)$. We can see that $T_c$ determines the vertical positioning of the cost curve, while $F_D(x)$ determines the shape of the curve. We note that a conventional streaming protocol that uses only one service tier (the primary access) has a fixed cost at $\varepsilon_p = 10$. Fig. 2 shows that our protocol can effectively reduce the cost by leveraging two service tiers. We will subsequently discuss the optimal buffering policy to minimize the cost.

### B. Optimal Buffering Policy

To solve the general problem in (P1), we should first solve a special case at $T_c = 0$ (i.e., the case of no caching). Let us

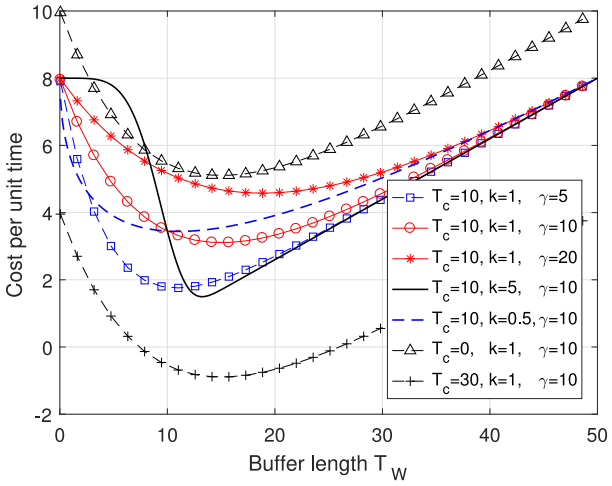



Fig. 2. Cost per unit time as a function of buffer length $T_W$ ($\varepsilon_p = 10$, $\varepsilon_s = 1$, and $T = 50$).

define $T_0$ as the optimal buffer length at $T_c = 0$, i.e.,

$$T_0 := \arg\min_{T_W} y(T_W; 0, T) \quad \text{s.t.} \quad 0 \le T_W \le T. \tag{5}$$

The derivative of $y(T_W; 0, T)$ with respect to $T_W$ is

$$\begin{aligned} &y'(T_W; 0, T) \\ &:= \frac{\partial}{\partial T_W} y(T_W; 0, T) \\ &= (\varepsilon_s - \varepsilon_p)\big[(T - T_W) f_D(T_W) - F_D(T_W)\big] \\ &= (\varepsilon_s - \varepsilon_p)\left[-1 + \left(\frac{k(T - T_W)}{\gamma}\left(\frac{T_W}{\gamma}\right)^{k-1} + 1\right) e^{-\left(\frac{T_W}{\gamma}\right)^k}\right] \end{aligned} \tag{6}$$

where $f_D(T_W)$ is the probability density function (PDF) of the Weibull distribution given by

$$f_D(T_W) = \frac{k}{\gamma}\left(\frac{T_W}{\gamma}\right)^{k-1} e^{-\left(\frac{T_W}{\gamma}\right)^k}. \tag{7}$$

It can be shown that given $\varepsilon_s < \varepsilon_p$, we have

$$y'(0^+; 0, T) = (\varepsilon_s - \varepsilon_p) T f_D(0^+) < 0 \tag{8}$$

and

$$y'(T; 0, T) = -(\varepsilon_s - \varepsilon_p) p_\chi > 0. \tag{9}$$

Hence, the equation

$$y'(T_W; 0, T) = 0 \tag{10}$$

has at least one root on $(0, T)$. We can further prove the following proposition.

*Proposition 1:* When the secondary access delay follows a Weibull distribution, $T_0$ has a unique value on $(0, T)$ given by the root of the equation $y'(T_0; 0, T) = 0$.

*Proof:* First, the second derivative of $y(T_W; 0, T)$ is

$$\begin{aligned} y''(T_W; 0, T) &:= \frac{\partial^2}{\partial T_W^2} y(T_W; 0, T) \\ &= (\varepsilon_s - \varepsilon_p)\big[(T - T_W) f'_D(T_W) - 2 f_D(T_W)\big] \\ &= \frac{(\varepsilon_s - \varepsilon_p) f_D(T_W)}{T_W} g(T_W) \end{aligned} \tag{11}$$

where

$$g(T_W) = \left[(T - T_W)\left(k - 1 - k\left(\frac{T_W}{\gamma}\right)^k\right) - 2T_W\right] \tag{12}$$

and

$$\begin{aligned} f'_D(T_W) &= \frac{\partial}{\partial T_W} f_D(T_W) \\ &= \left(\frac{k-1}{\gamma} \cdot \frac{\gamma}{T_W} - \frac{k}{\gamma}\left(\frac{T_W}{\gamma}\right)^{k-1}\right) f_D(T_W). \end{aligned} \tag{13}$$

When $k \le 1$, we have $g(T_W) < 0$ and $y''(T_W; 0, T) > 0$ on $(0, T)$. It follows that $y'(T_W; 0, T)$ is monotonically increasing. Hence, (10) has only one root.

When $k > 1$, we need to discuss the solution of

$$y''(T_W; 0, T) = 0. \tag{14}$$

Because $([(\varepsilon_s - \varepsilon_p) f_D(T_W)]/h\gamma) < 0$ on $(0, T)$, (14) is equivalent to

$$g(T_W) = 0. \tag{15}$$

Taking the first and second derivatives of $g(T_W)$, we have

$$g'(T_W) = -k - 1 + k(1 + k)\left(\frac{T_W}{\gamma}\right)^k - k^2 \frac{T}{\gamma}\left(\frac{T_W}{\gamma}\right)^{k-1} \tag{16}$$

and

$$g''(T_W) = \frac{1}{\gamma} k^2 \left(\frac{T_W}{\gamma}\right)^{k-2}\left[\frac{(1 + k)T_W}{\gamma} - \frac{(k - 1)T}{\gamma}\right] \tag{17}$$

respectively. The unique solution of $g''(T_W) = 0$ is

$$\tilde{T} = \frac{k - 1}{k+1} T. \tag{18}$$

Substituting $\tilde{T}$ into (16), we obtain the maximal value of $g'(T_W)$ given by

$$\max g'(T_W) = -k - 1 - k\left(\frac{k - 1}{k+1}\right)^{k-1}\left(\frac{T}{\gamma}\right)^k. \tag{19}$$

Because we have $\max g'(T_W) < 0$, $g(T_W)$ is a strictly decreasing function on $(0, T)$. Moreover, because $g(0^+) = (k - 1)T > 0$ and $g(T) = -2T < 0$, (15) has a unique solution on $(0, T)$, which is denoted by $\breve{T}$. When $T_W < \breve{T}$, we have $y''(T_W; 0, T) < 0$. When $T_W > \breve{T}$, we have $y''(T_W; 0, T) > 0$. Therefore, $y'(T_W; 0, T)$ reaches the minimal value at $T_W = \breve{T}$. Recall that the minimum value of $y'(T_W; 0, T)$ has a negative value according to (8). Further considering the fact that $y'(T_W; 0, T)$ is monotonically increasing on $(\breve{T}, T)$, we can conclude that $y'(T_W; 0, T) = 0$ has a unique solution on $(\breve{T}, T)$. It is easy to see that this unique solution is $T_0$ that minimize the cost function. ■

The proof of Proposition 1 shows that $y(T_W; 0, T)$ is a strictly decreasing function of $T_W$ on $[T_0, T)$. Hence, when $T_c \le T_0$, $y(T_W; T_c, T)$ reaches the minimum at $T_W =$

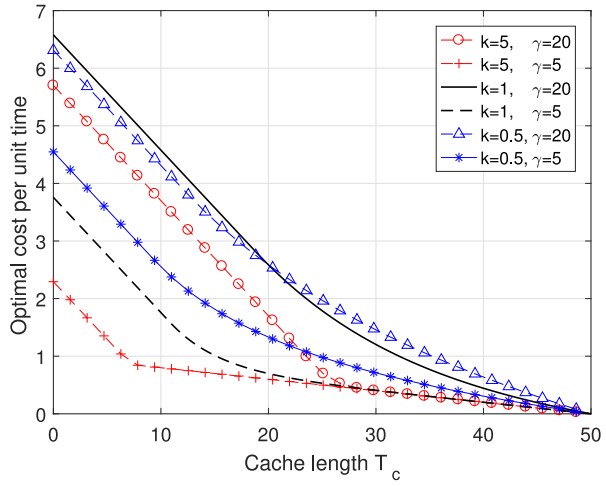



Fig. 3. Optimal cost per unit time $y^*/T$ as a function of cache length $T_c$ ($\varepsilon_p = 10$, $\varepsilon_s = 1$, and $T = 50$).

$T_0$; When $T_c > T_0$, $y(T_W; T_c, T)$ reaches the minimum at $T_W = T_c$. The optimal buffer length is, therefore,

$$T_W^* = \max\{T_c, T_0\}. \tag{20}$$

We can see that the optimal buffer length $T_W^*$ and the minimum cost $y^*(T_c; T) := y^*(T_W; T_c^*, T)$ are both functions of $T_c$. This means the buffering policy is inherently coupled with the caching policy. Fig. 2 shows the optimal cost per unit time $y^*(T_c; T)/T$ as a function of $T_c$ with varying delay distribution $F_D(x)$. We see that the cost increases with increasing $\gamma$, which suggests that a larger mean delay leads to a higher cost. Similarly, smaller values of $k$ tends to boost the cost because the delay distributions have longer-tails. Considering $y^*(T_c; T)$ as a function of $T_c$, we can further prove the following property of $y^*(T_c; T)$.

*Proposition 2:* Given optimal buffering, the optimal cost $y^*(T_c; T)$ is a convex function of $T_c$.

*Proof:* For convenience, we can rewrite $y^*(T_c; T)$ as

$$y^*(T_c; T) = -T_c\varepsilon_p + \tilde{y}^*(T_c; T) \tag{21}$$

where

$$\tilde{y}^*(T_c; T) = T_W^*\varepsilon_p + \big(T - T_W^*\big)\big[\varepsilon_s p_\chi + \varepsilon_p\big(1 - p_\chi\big)\big]. \tag{22}$$

We note that $p_\chi$ is a function of $T_W^*$. On one hand, if $T_c \le T_0$, then $T_W^* = T_0$, so that $T_W^*$ is not related to $T_c$. That is,

$$\frac{\partial}{\partial T_c}\tilde{y}^*(T_c; T) = 0 \tag{23}$$

and

$$\frac{\partial}{\partial T_c}y^*(T_c; T) = -\varepsilon_p. \tag{24}$$

On the other hand, according to (6) and (22) we have

$$\frac{\partial}{\partial T_c}\tilde{y}^*(T_c; T) = y'(T_c; 0, T) \tag{25}$$

for $T_c > T_0$, given that $T_W^* = T_c$ in this case. By Proposition 1, the second derivative of $y(T_c; 0, T)$ with $T_c$, is always positive once $T_c > \tilde{T}$. Because $T_c > T_0$ and $T_0 > \tilde{T}$, it is clear that $T_c > \tilde{T}$. Therefore, the positive second derivative guarantees $\tilde{y}^*(T_c; T)$ and $y^*(T_c; T)$ are convex, and the proof is complete. ■

We can further derive the bounds of $y^*(T_c; T)$ by expanding (25) as

$$\frac{\partial}{\partial T_c}\tilde{y}^*(T_c; T) = \big(\varepsilon_s - \varepsilon_p\big)\big[(T - T_c)f_D(T_c) - p_\chi(T_c)\big]. \tag{26}$$

Because

$$\begin{aligned}(T - T_c)f_D(T_c) - p_\chi(T_c) &\ge -p_\chi(T_c)\\ &\ge -1.\end{aligned} \tag{27}$$

We have

$$0 < \frac{\partial}{\partial T_c}\tilde{y}^*(T_c; T) \le -\big(\varepsilon_s - \varepsilon_p\big) \tag{28}$$

and

$$-\varepsilon_p < \frac{\partial}{\partial T_c}y^*(T_c; T) \le -\varepsilon_s. \tag{29}$$

Combining (24) and (29) yields the range of $(\partial/\partial T_c)y^*(T_c; T)$ as $[-\varepsilon_p, -\varepsilon_s]$. We note that the term $(\partial/\partial T_c)y*(T_c; T)$ has a physical meaning: it shows when $T_c$ length of playback time is already cached for a file, how effective can additional caching help to reduce the cost. The bounds in (29) shows that the effectiveness of additional caching ranges from $-\varepsilon_p$ to $-\varepsilon_s$. This theoretical result collaborate the intuition that caching essentially helps to reduce the needs of primary access (at a cost of $\varepsilon_p$) or secondary access (at a cost of $\varepsilon_s$).

## IV. Optimal Caching Policy at the Device Level

### A. Problem Formulation

This section addresses the problem of optimal caching on a vehicular cache device, which can store multiple content files. One content can have different encoding rates and will result in different files. Let $\boldsymbol{F}$ be the set of all files, $F = \|\boldsymbol{F}\|$ be the total number of files, and $f(1 \le f \le F)$ be the file index. The length of the $f$th file is $T_f$. The cache length of the $f$th file is $T_{c,f}$ $(0 \le T_{c,f} \le T_f)$. Each file is assigned with a weight, which is the product of the encoding rate and popularity (i.e., request number or probability). We note that both the encoding rate and popularity are linearly proportional to the total number of bits transmitted when streaming a file. As the average cost of transmitting a bit can be regarded as a constant, the cost of streaming a file is also linearly proportional to the encoding rate and popularity. The weight of the $f$th file is denoted as $h_f$. The total cache space is denoted as $S$. We consider a continuous caching scheme, which means we can flexibly cache an arbitrary portion of a file.

The caching problem can be formulated as the following optimization problem (P2):

$$\begin{aligned} Y^* := \min &\sum_{f\in\boldsymbol{F}} h_f y^*\big(T_{c,f}; T_f\big)\\ \text{s.t.} &\begin{cases}\sum_{f\in\boldsymbol{F}} T_{c,f} \le S\\ 0 \le T_{c,f} \le T_f\end{cases}\end{aligned} \tag{30}$$

where $y^*(T_{c,f};T_f) = y(T_W^*;T_{c,f},T_f)$ is the minimum cost of the $f$th file written as a function of the file length conditioning the given $T_f, f = 1,2,\ldots,F$.

According to Proposition 2, the cost of each file $y^*(T_{c,f};T_f)$ is a convex function of $T_{c,f}$. Moreover, the cost functions of different files are not directly related. It is easy to see that the feasible domain of (30) is also convex, so that (P2) is a convex optimization problem.

### B. Optimal Caching Policy

Because (P2) is a convex optimization problem, we can solve it by the Lagrangian method. For convenience of expression, let us define

$$H_f(T_{c,f}) \coloneqq H(T_{c,f};h_f,T_f) = h_f y^*(T_{c,f};T_f). \tag{31}$$

The Lagrangian formulation is given by

$$L(\theta,T_{c,1},\ldots,T_{c,\|\boldsymbol{F}\|}) = \sum_{f\in\boldsymbol{F}} H_f(T_{c,f}) - \theta\sum_{f\in\boldsymbol{F}} T_{c,f} \tag{32}$$

where $\theta$ is the Lagrange multiplier. The Kuhn–Tucker condition for the optimality of a solution is

$$\frac{\partial L}{\partial T_{c,f}} = H'_f(T_{c,f}) - \theta\begin{cases} = 0, & 0 < T_{c,f} < T_f \\ \geq 0, & T_{c,f} = 0 \\ \leq 0, & T_{c,f} = T_f \end{cases} \tag{33}$$

where

$$H'_f(T_{c,f}) = \frac{\partial}{\partial T_{c,f}} H_f(T_{c,f}) = h_f \frac{\partial}{\partial T_{c,f}} y^*(T_{c,f};T_f). \tag{34}$$

The optimal cache allocation scheme is given by

$$T_{c,f}^* = \begin{cases} 0, & \theta < H'_f(0) \\ T_c(\theta;T_f,h_f), & H'_f(0) \leq \theta \leq H'_f(T_f) \\ T_f, & \theta > H'_f(T_f) \end{cases} \tag{35}$$

where the level $\theta$ is chosen to satisfy the constraint

$$\sum_{f\in\boldsymbol{F}} T_{c,f} = S. \tag{36}$$

We can see that the structure of (P2) is similar to the classic problem of optimal power allocation in parallel communication channels [52]. The solution should therefore has a similar rationale to the classic water-filling algorithm [52]. However, as illustrated in Fig. 4, our problem is more complicated. Fig. 4 shows the derivative of the weighted cost $H'(T_{c,f})$ as a function of the cache length $T_c$ for three files. When the cache size $T_c$ is relatively small, the derivatives of the cost-$T_c$ functions have constant negative values. This means the cost of delivering a file reduces at a constant speed with increasing $T_c$. This collaborates our previous illustration in Fig. 3, where the cost-$T_c$ curves are shown to be straight lines with a constant slope of $-\varepsilon_p$ when $T_c$ is small. In Fig. 4, the slope is further weighted by the encoding rate and popularity of each file, so that different files have different weighted derivatives. After $T_c$ passes a threshold, the derivative functions increase monotonically. This means increasing the cache size gradually becomes less effective for cost reduction. In Fig 4, there is a horizontal line that represents the "water-level" $\theta$. Depending on how

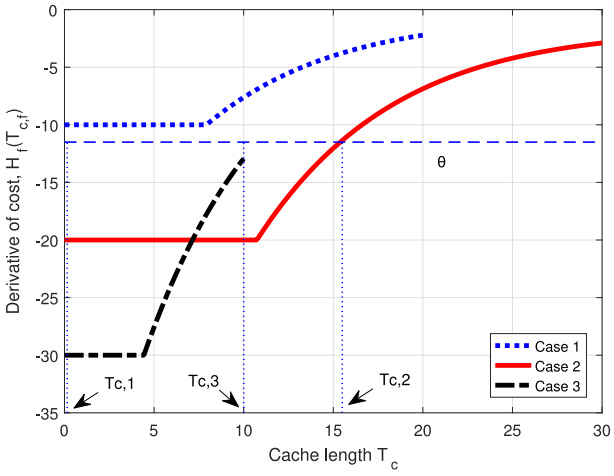



Fig. 4. Derivative of the weighted cost $H'(T_{c,f})$ as a function of the cache length $T_c$, intersecting with the water-level $\theta$ ( $\varepsilon_p = 10$, $\varepsilon_s = 1$, $k = 1$, $\gamma = 10$, $[T_1,T_2,T_3] = [20,30,10]$, and $[h_1,h_2,h_3] = [1,2,3]$).

the water-level intersects with the derivative functions, we can distinguish the three cases below.

Case 1: There is no intersection because $\theta$ is smaller than the minimum value of $H'(T_{c,1})$. In this case no cache should be assigned, i.e., $T_{c,1} = 0$.

Case 2: There is an intersection point. The optimal cache value is then given by the $x$-axis value of the intersection point.

Case 3: There is no intersection because $\theta$ is larger than the maximum value of $H'(T_{c,f})$. In this case, the file should be cached with full length. In Fig. 4, we have $T_{c,3} = T_3$.

The optimal solution formulated in (35) reflects these three cases.

We can see that once the water-level $\theta$ is known, the optimal caching length can be easily determined. Unfortunately, there is no closed-form solution for (36), so that we have to resort to numerical methods. We note that numerical methods used in the classic water-filling algorithm cannot not be directly applied to our case because the derivative function has flat intervals (i.e., intervals with zero slope). As a result, the Ping-Pong phenomenon could occur when we iterate the value of $\theta$ numerically. The Ping-Pong phenomenon means the value of $\theta$ bounces up and down around a flat interval as shown in Fig. 4. It happens when there are still some cache space to be allocated, but the remaining cache space is smaller than the length of the flat interval. To this end, a modified water-filling algorithm is proposed. The algorithm includes the following steps.

Step 1 *(Case Classification):* For a given value of $\theta$, identify three sets of files corresponding to the three cases explained above. Mathematically, we have

$$\boldsymbol{F}_{\text{case1}}(\theta) = \{f|\theta < H'_f(0), f\in\boldsymbol{F}\} \tag{37}$$

$$\boldsymbol{F}_{\text{case3}}(\theta) = \{f|\theta > H'_f(T_f), f\in\boldsymbol{F}\} \tag{38}$$

and

$$\boldsymbol{F}_{\text{case2}}(\theta) = \boldsymbol{F} - \boldsymbol{F}_{\text{case1}}(\theta) - \boldsymbol{F}_{\text{case3}}(\theta). \tag{39}$$

**Algorithm 1** Iterative Level-Resetting Algorithm for (P2)

1: //——Parameter Definition——
2: $\theta_U$ upper boundary of $\theta$ with $\Delta_S(\theta_U) < 0$
3: $\theta_L$ lower boundary of $\theta$ with $\Delta_S(\theta_L) > 0$
4: $\delta_\theta$ iteration-terminating threshold with $\delta_\theta \to 0^+$
5: $i$ iteration indicator
6: $\theta^*$ level determined by this algorithm
7: //——Iterative Processing——
8: $b \leftarrow \theta_U$
9: $a \leftarrow \theta_L$
10: $i \leftarrow 0$
11: **repeat**
12: $i \leftarrow i + 1$
13: $\theta^{(i)} = \frac{1}{2}(a + b)$
14: $\Delta_S^{(i)} = \Delta_S\left(\theta^{(i)}\right)$
15: **if** $\Delta_S^{(i)} = 0$ **then**
16: **go to** Line 23
17: **else**
18: **if** $\Delta_S^{(i)} < 0$ **then**
19: $b \leftarrow \theta^{(i)}$
20: **else**
21: $a \leftarrow \theta^{(i)}$
22: **until** $(b - a) < \delta_\theta$
23: **if** $\Delta_S^{(j)}, j = 1, \ldots, i$ tends to 0 **then**
24: $\theta^* \leftarrow \theta^{(i)}$,
25: $\boldsymbol{F}_{\text{overlapped}}(\theta^*) = \varnothing$;
26: **else**
27: $\theta^*$ is equal to some $-\varepsilon_p H_f$ that is nearest to $\theta^{(i)}$,
28: $\boldsymbol{F}_{\text{overlapped}}$ is not empty.

Step 2 *(Cache Space Comparison):* Calculate the total required cache space $S'$. If $S' < S$, $\theta$ will increase in the next iteration, otherwise $\theta$ will decrease.

Step 3 *(Iteration of $\theta$):* Apply classic iteration methods (e.g., the Newton's method) on $\theta$ till $|S' - S|$ is smaller than a predefined threshold.

Step 4 *(Detection of Ping-Pong Phenomenon and Remainder Cache Allocation):* Track the difference $\Delta\theta$ between two consecutive iterations of $\theta$. If $\Delta\theta$ is smaller than a predefined (small) threshold while $|S' - S|$ approaches a nonzero constant, it means $\theta$ is bouncing up and down around a "flat interval" as shown in Fig. 4. Denote $\Delta_S(\theta)$ as the remaining cache resource, then

$$\Delta_S(\theta) = S - S' = S - \sum_{f \in F_{\text{full}}(\theta)} T_f - \sum_{f \in F_{\text{intersected}}(\theta)} z\left(\frac{\theta}{h_f}; T_f\right) \quad (40)$$

where $z([\partial/\partial T_{c,f}]y^*(T_{c,f}; T_f); T_f)$ is the inverse function of $[\partial/\partial T_{c,f}]y^*(T_{c,f}; T_f)$. The remaining cache resource will be allocated to the file on which $\theta$ converges to.

The pseudo-code of the above iteration procedure is given in Algorithm 1. Fig. 5 shows the average weighted cost per unit time as a function of the cache space, which is normalized to the total size of all files in $\boldsymbol{F}$. The performance of the optimal caching algorithm is compared with two other heuristic benchmark algorithms. One of the heuristic algorithm allocates cache length proportional to the file length, while the other allocates cache length equally for all files. It is assumed that the file length is uniformly distributed in [5] and [50], while the weighted function $h_f$ is uniformly distributed in [1]. We can see that the optimal algorithm consistently outperforms both heuristic algorithms.

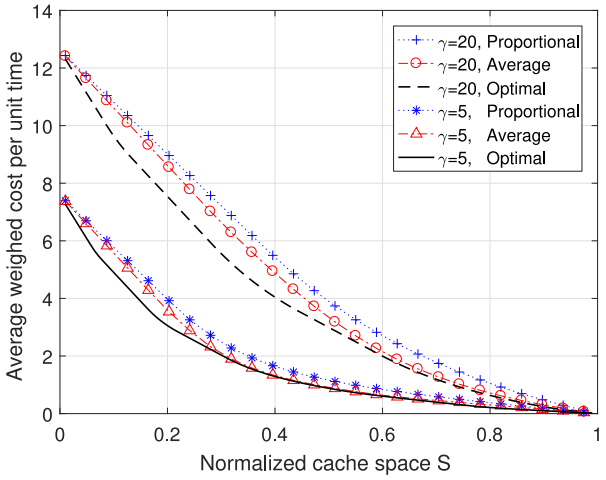



Fig. 5. Averaged weighted cost per unit time as a function of the normalized cache space $S$, with three different caching policies [$\varepsilon_p = 10$, $\varepsilon_s = 1$, $k = 1$, $T \sim U(5, 50)$, and $H \sim U(1, 100)$].

## V. Streaming Rate Adaptation at the Network Level

### A. Problem Formulation

This section analyzes the performance of the proposed protocol in a large scale IoV network. In particular, we are interested in the problem of streaming rate adaptation, i.e., how should the (average) content streaming rate adapt to the varying network conditions given a predefined cost budget. We consider a scenario where the IoV use two service tiers from a cellular network for content streaming. This is a practical scenario in 5G-enabled IoV as the cellular network is evolving to support customizable service tiers [16]. Moreover, we assume an ultraelastic secondary access scheme proposed in [18], where the secondary access can only access a BS when the BS is vacant. In other words, the secondary access traffic has the lowest priority among all types of traffic. This corresponds to the worst-case scenario for secondary access in cellular networks. Hence, our analysis in this section can serve as a performance lower bound for other secondary access schemes.

For analytical tractability, we consider a homogeneous network with identical BSs and vehicular users. The BSs and vehicles are assumed to be distributed on a plane following two independent stationary poisson point processes with intensities $\lambda_b$ and $\lambda_v$, respectively. A vehicle is associated with the nearest BS for communication. This implies a Poisson–Voronoi cell partition. Each BS has identical transmit power denoted as $P$. We define $\lambda = \lambda_v/\lambda_b$ as the vehicle/BS density ratio. It

indicates the average number of vehicles in a cell. The BSs are assumed to have the same vacant probability denoted as $\Omega$. This parameter indicates the congestion level of the network.

The secondary access is assumed to comply with the following access procedure [18].

1) A secondary user periodically evaluates whether the BS is vacant and whether the link to the user has a good quality, i.e., the signal-to-interference-and-noise ratio (SINR) is above a threshold $\phi_{\text{th}}$. The probability that a secondary user has a good link quality is called "coverage probability" and denoted by $p_c$.
2) If both evaluations are positive, the user is admitted into the secondary service and proceeds to the next stage of multiuser access.
3) Multiple contending secondary users have equal opportunities to access the vacant BS through time sharing. The probability for a secondary user to have successful secondary access (having good coverage and successful channel contention) is called "access probability" and denoted by $p_a$. The overall probability for a secondary user to have secondary service is called "service probability" and denoted by $p_s$. We have

$$p_s = p_a \Omega. \tag{41}$$

We are interested in the delay performance of the secondary communication service. According to [18], a two-level M/G/1 priority queuing model with preemptive-resume policy [50] is used to characterize the secondary traffic dynamic, where the transmission of secondary traffic may be preempted (i.e., immediately interrupted) by outages. An outage can be caused by multiple factors such as primary traffic interruption, bad coverage, and failure in multiuser contention. The outage event takes a higher priority in the queue. The random arrival interval of the outage events and secondary traffic are denoted as $\alpha_o$ and $\alpha_s$, respectively. The random durations of an outage event and a secondary packet transmission are denoted as $\beta_o$ and $\beta_s$, respectively. The four random variables $\alpha_o$, $\alpha_s$, $\beta_o$, and $\beta_s$ fully characterize the priority queueing model. Their mean values are denoted by $\bar{\alpha}_o$, $\bar{\alpha}_s$, $\bar{\beta}_o$, and $\bar{\beta}_s$, respectively. The intensities of the outage process and secondary traffic process are characterized by $\rho_o = \bar{\beta}_o/\bar{\alpha}_o$ and $\rho_s = \bar{\beta}_s/\bar{\alpha}_s$, respectively. A stable queue requires $\rho_o + \rho_s < 1$. The single-user queuing model in the temporal domain is related to the multiuser network model in the spatial domain by a simple equation

$$\rho_o = 1 - p_s. \tag{42}$$

This equation means that the probability of secondary service outage evaluated in the spatial–temporal domains should be the same.

According to the above queueing model, the average streaming rate of the secondary traffic is given by $\bar{L}/\bar{\alpha}_s = \bar{L}\rho_s/\bar{\beta}_s$, where $L$ is the random size (in bits or bytes) of a secondary traffic packet and $\bar{L}$ is its mean. Moreover, we have $\beta_s = L/R$, where $R$ is the random transmission rate related to the SINR of a secondary radio link. The packet size $L$, which is mainly related to the application layer protocols, has a predefined distribution in practice. Thus, $\bar{L}$ can be regarded as a fixed parameter. Moreover, the SINR at a random user also converges to a fixed distribution in the interference-limited region regardless of the BS density and transmission power [53]. Consequently, in the interference limited regime, $\beta_s$ can be assumed to have a predefined distribution and its mean $\bar{\beta}_s$ is a fixed parameter. It follows that the streaming rate is linearly proportional to $\rho_s$. Therefore, to avoiding introducing other unnecessary parameters, we will use $\rho_s$ as a measure of the secondary traffic streaming rate.

This section aims to solve the following problem: given a network condition and a cost limit/budget, what is the maximum streaming rate for secondary service? This requires us to establish the relationship between $\rho_s$ and the cost. To this end, we need to carry out analyses in both the spatial–temporal domains.

### *B. Spatial Domain Analysis*

The spatial domain analysis aims to establish $p_s$ defined in (41) as a function of network conditions. Considering a representative case with Rayleigh channel fading and a path loss exponent of 4, the complementary CDF of the SINR $\phi$ perceived by a typical user in the network is given by [53]

$$F_\phi(x) = \frac{\pi^{\frac{3}{2}}\lambda_b}{\sqrt{x/P}} e^{\frac{a^2}{\sqrt{2b}}} \mathrm{Q}\left(\frac{a}{\sqrt{2x/P}}\right) \tag{43}$$

where $\mathrm{Q}(\cdot)$ denotes the $Q$-function and

$$a = \lambda_b \pi \left[1 + \sqrt{x}\arctan(\sqrt{x})\right]. \tag{44}$$

If the system is interference limited, which implies that $P$ is sufficiently large so that the noise is negligible, (43) can be further simplified to [53]

$$F_\phi^{\lim}(x) = \frac{1}{1 + \sqrt{x}\arctan(\sqrt{x})}. \tag{45}$$

A user is in coverage of the secondary service if $\phi > \phi_{th}$. The coverage probability is therefore given by

$$p_c = F_\phi(\phi_{th}). \tag{46}$$

Once $p_c$ is known, the secondary access probability $p_a$ can be evaluated according to the following proposition.

*Proposition 3:* The access probability of secondary service is given by

$$p_a = \frac{3.5}{2.5\lambda}\left[1 - \left(1 + \frac{p_c\lambda}{3.5}\right)^{-2.5}\right] \tag{47}$$

where $\lambda = \lambda_v/\lambda_b$ is the vehicle-BS density ratio.

*Proof:* Let us consider the arrival of a secondary user into a typical Voronoi cell. The number of existing users in the cell is a random variable $N$. The PMF of $N$ is given by

$$f_N(n) = \int_0^\infty \frac{(\lambda x)^n}{n!} e^{-\lambda x} f_V(x) dx \tag{48}$$

where $f_V(x)$ is the PDF of the random variable $V$ that denotes the size of a Voronoi cell normalized by $1/\lambda$. We have [54]

$$f_V(x) = \frac{3.5^{3.5}}{\Gamma(3.5)} x^{2.5} e^{-3.5x}. \tag{49}$$

Let $X(0 \le X \le N+1)$ be a random integer denoting the number of in-coverage secondary users in the cell. If $X = 0$, the chance to have secondary access is zero. In all other cases of $X \neq 0$, the chance for a random user to have secondary access is $1/(N+1)$. The probability for $X = 0$ is $(1-p_c)^{N+1}$. The secondary access probability conditioned on $N$ is then

$$p_a(N) = \frac{1-(1-p_c)^{N+1}}{N+1}. \tag{50}$$

The overall secondary access probability average over $N$ is

$$p_a = \mathrm{E}[p_a(N)] = \sum_{n=0}^{\inf} \frac{1-(1-p_c)^{n+1}}{n+1} f_N(n). \tag{51}$$

Substituting (48) into (51) and applying the equation that

$$\sum_{n=0}^{\inf} \frac{(\lambda x)^{n+1}}{(n+1)!} = e^{\lambda x} - 1 \tag{52}$$

we can get

$$p_a = \int_0^{\infty} \frac{f_V(x)}{\lambda x}(1 - e^{-p_c \lambda x})dx. \tag{53}$$

According to the definition of $f_V(x)$ in (49), we have

$$\int_0^{\infty} \frac{1}{x} f_V(x) e^{-ax})dx = \frac{3.5^{2.5}}{2.5}(3.5+a)^{-2.5}. \tag{54}$$

Substituting (54) into (53) yields (47). ■

Once $p_a$ is known, $p_s$ can be evaluated according to (41).

### C. Temporal Domain Analysis

The temporal domain analysis aims to characterize the distribution of the secondary traffic according to the queueing model. The random delay $D$ of a packet is the sum of the waiting time $D_w$ and transmission time $D_t$. The PDFs of $D$, $D_w$, and $D_t$ are denoted by $f_D(x)$, $f_{D_t}(x)$, and $f_{D_w}(x)$, respectively. For an M/G/1 two-level priority queue with preemptive-resume policy, the Laplace transform of $f_{D_t}(x)$ and $f_{D_w}(x)$ are tractable in some special cases [18], [50]. Hence, the original PDFs can be numerically obtained by the inverse Laplace transform. Such numerical method, however, is computationally demanding due to the long-tail nature of $D$. To facilitate our performance evaluation, we will subsequently propose a closed-form approximation to the PDF of $D$.

Our approximation exploits two facts. First, the means of $D_w$ and $D_t$ are known to be [50]

$$\bar{D}_w = \frac{\bar{\beta}_s \rho_s + \bar{\beta}_o \rho_o}{(1-\rho_o)(1-\rho_o-\rho_s)} \tag{55}$$

$$\bar{D}_t = \frac{\bar{\beta}_s}{1-\rho_o} \tag{56}$$

respectively. Second, $f_{D_w}(x)$ has an impulse at the origin and can be written in the following form:

$$f_{D_w}(x) = (1-\rho_o-\rho_s)\delta(0) + (\rho_o+\rho_s)\tilde{f}_{D_w}(x) \tag{57}$$

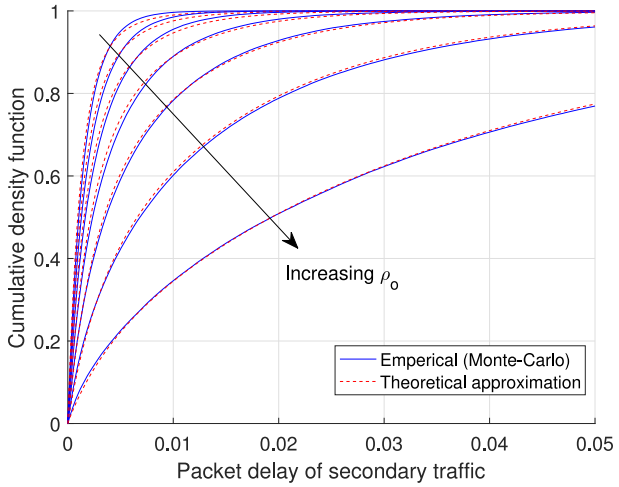



Fig. 6. Empirical CDF and theoretical approximation of the packet delays of the secondary traffic (M/M/1 priority queue, $\rho_s = 0.2$, $\beta_o = \beta_s = 10^{-3}$, $k_1 = 1$, and $k_2 = 0.9$).

where $\delta(x)$ is the impulse response function and $\tilde{f}_{D_w}(x)$ is an unknown distribution. The PDF of $D$ can then be written as

$$f_D(x) = f_{D_t}(x) * f_{D_w}(x) = (1-\rho_o-\rho_s) f_{D_t}(x) + (\rho_o+\rho_s)[f_{D_t}(x) * \tilde{f}_{D_w}(x)] \tag{58}$$

where "$*$" means convolution. Our approximation assumes that $f_{D_t}(x)$ and $f_{D_t}(x) * \tilde{f}_{D_w}(x)$ can be approximated by Weibull distributions with shape parameters $k_1$ and $k_2$, respectively. It follows that the CDF of $D$ can be written as

$$F_D(x) = (1-\rho_o-\rho_s)e^{-(x/\gamma_1)^{k_1}} + (\rho_o+\rho_s)e^{-(x/\gamma_2)^{k_2}}. \tag{59}$$

The scale parameters $\gamma_1$ and $\gamma_2$ can be calculated by fitting the mean values. We have

$$\gamma_1 = \bar{D}_t / \Gamma(1+1/k_1) \tag{60}$$

$$\gamma_2 = \frac{\bar{D}_t + \bar{D}_w/(\rho_o+\rho_s)}{\Gamma(1+1/k_1)}. \tag{61}$$

Fig. 6 demonstrates the accuracy of the proposed approximation. Monte-Carlo simulations are performed for an M/M/1 two-level priority queue with preemptive-resume policy to obtain the empirical CDFs of $D$. We can see that by setting $k_1 = 1$ and $k_2 = 0.9$, a fairly good fit can be achieved with varying $\rho_o$ when $\bar{\beta}_o = \bar{\beta}_s = 10^{-3}$. Trials with different parameter settings show that the mean square error of the approximation is about $5 * 10^{-5}$ for a wide range of practical queueing parameters.

### D. Performance Evaluation

Based on (41), (42), (47), and (59), we are able to evaluate how the cost changes with different parameters. For benchmark convenience, we consider an interference limited network (i.e., $P \to \infty$) and zero SINR threshold (i.e., $\phi_{\text{th}} = 0$). These two extreme conditions give an upper bound on the performance. The content length $T_f$ is assumed to follow a uniform distribution in $[1, T_{\max}]$ (minutes). In addition, results are only shown for $T_c = 0$ due to page limits. Except otherwise mentioned, default parameter values are set to be $\lambda = 10$, $\Omega = 0.5$, $\varepsilon_p = 10$, $\varepsilon_s = 1$, $\beta_o = \beta_s = 10^{-3}$, and $T_{\max} = 50$.

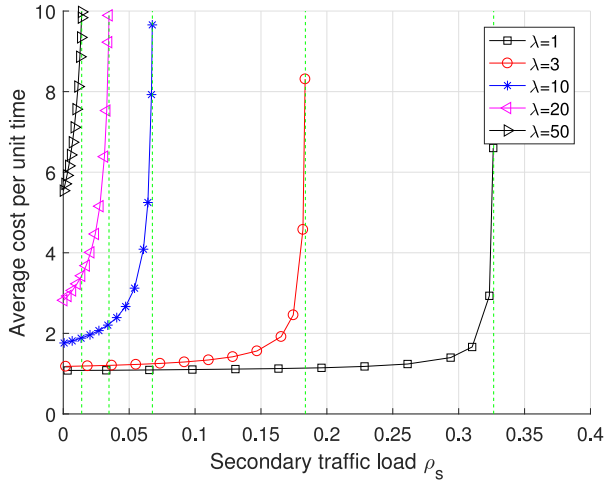



Fig. 7. Average cost per unit time as a function of secondary traffic load $\rho_s$ with varying values of $\lambda$ ($\Omega = 0.5$, $\varepsilon_p = 10$, $\varepsilon_s = 1$, $T_{\max} = 50$, and $\beta_o = \beta_s = 10^{-3}$).

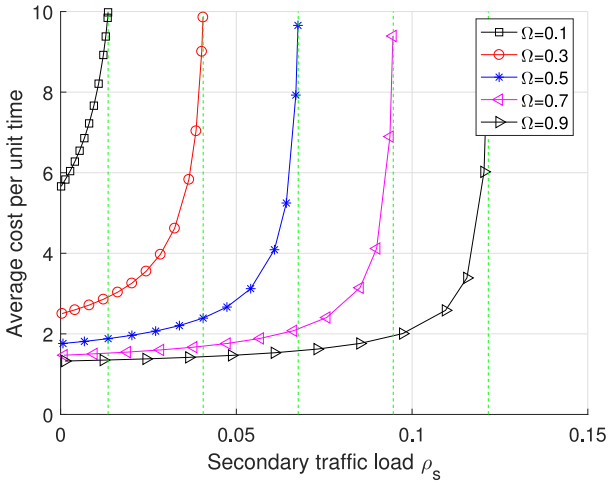



Fig. 8. Average cost per unit time as a function of secondary traffic load $\rho_s$ with varying values of $\Omega$ ($\lambda = 10$, $\varepsilon_p = 10$, $\varepsilon_s = 1$, $T_{\max} = 50$, and $\beta_o = \beta_s = 10^{-3}$).

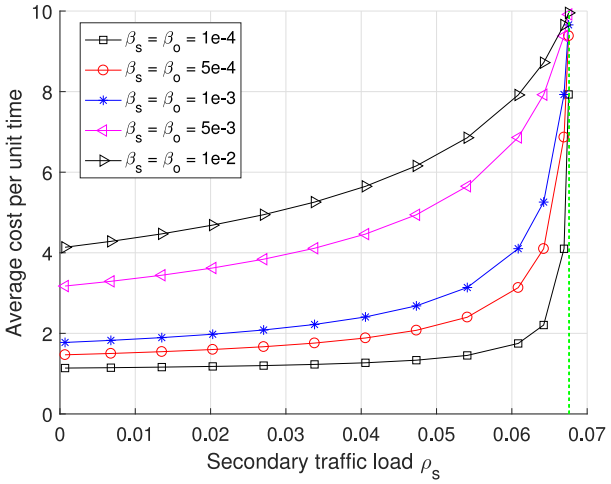



Fig. 9. Average cost per unit time as a function of secondary traffic load $\rho_s$ with varying values of $\beta_o$ and $\beta_s$ ($\lambda = 10$, $\Omega = 0.5$, $\varepsilon_p = 10$, $\varepsilon_s = 1$, and $T_{\max} = 50$).

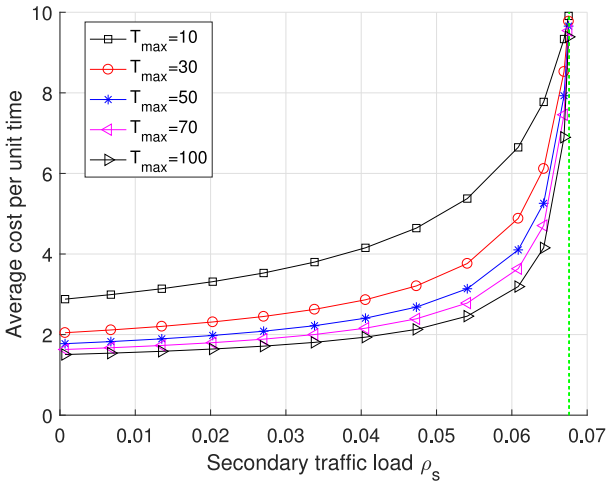



Fig. 10. Average cost per unit time as a function of secondary traffic load $\rho_s$ with varying values of $T_{\max}$ ($\lambda = 10$, $\Omega = 0.5$, $\varepsilon_p = 10$, $\varepsilon_s = 1$, and $\beta_o = \beta_s = 10^{-3}$).

It should be noted that this paper use minute as the time unit. The configuration of $\beta_o = \beta_s = 10^{-3}$ implies that the average transmission time of a primary or secondary packet is 60 ms, which is a realistic value according to measurements [51]. Once the delay distribution is obtained, the optimal cost per file can be calculated according to Section III. Finally, the cost per unit time is averaged over random values of $T_f$.

Fig. 7 shows the average cost per unit time as a function of the secondary traffic load $\rho_s$ with varying vehicle-BS density ratio $\lambda$. It is observed that the cost increases from $\varepsilon_s$ (i.e., the cost of secondary access) to $\varepsilon_p$ (i.e., the cost of primary access) with increasing $\rho_s$ and $\lambda$. Each density $\lambda$ corresponds to a maximum streaming rate $\rho_s$. It is interesting to observe that the minimum cost at $\rho_s = 0$ scales linearly with $\lambda$. This means that given a fixed cost budget, there is a maximum vehicle density that the system can tolerate even when the streaming rate is very small. Moreover, when $\lambda$ is relatively small, $\rho_s$ can reach 80%–90% of its maximum value with a small increase on the minimum cost. This means when the vehicle density is relatively small, most of the network capacity can be utilized for high speed streaming at a low cost.

Fig. 8 shows the cost function with varying BS vacant probability $\Omega$, which is an indicator of the congestion level of the network. It is observed that $\Omega$ has similar impacts on the cost function as $\lambda$. Therefore, the previous discussions on $\lambda$ also suits $\Omega$.

Finally, Figs. 9 and 10 show the cost functions with varying $\beta_s/\beta_o$ and $T_{\max}$, respectively. Parameters $\beta_s$ and $\beta_o$ reflect the packet transmission time, while $T_{\max}$ reflects the duration of contents. We can see that these parameters have no impact on the maximum streaming rate, but can still affect the minimum cost. Generally speaking, a lower cost can be achieved when we have a smaller packet transmission time and longer content length.

## VI. Conclusion

In this paper, we have investigated the issue of cost optimization for vehicle-oriented content streaming services.

This paper has assumed a context in which the vehicular communication network has two service tiers. We have proposed a cost-optimal streaming protocol to jointly optimize the tasks of content streaming and content caching. An analytical framework has been established to reveal how the service cost scales in large-size networks. We have found that the maximum streaming rate is fundamentally limited by the vehicle density and network congestion level. Moreover, given network conditions, there is a minimum cost even when the streaming rate approaches zero. The proposed protocol is shown to be effective in reducing the overall cost of vehicular content streaming services.

We further note that the analytical framework established in this paper can be extended to study general cases of multiple service tiers. Each service tier can be characterized by its bandwidth limit, delay distribution, and cost. An extended protocol is then required to handover among different service tiers and minimize the total cost of streaming service. Such an extension will be pursued in our future work.

## REFERENCES


[1] O. Kaiwartya *et al.*, "Internet of Vehicles: Motivation, layered architecture, network model, challenges, and future aspects," *IEEE Access*, vol. 4, pp. 5356–5373, 2016.

[2] M. Amadeo, C. Campolo, and A. Molinaro. "Enhancing IEEE 802.11p/WAVE to provide infotainment applications in VANETs," *Ad Hoc Netw.*, vol. 10, no. 2, pp. 253–269, Mar. 2012.

[3] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 33–55, Mar. 2016.

[4] W. Dai and S. Jordan, "ISP service tier design," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1434–1447, Jun. 2016.

[5] M. Asefi, J. W. Mark, and X. S. Shen, "A mobility-aware and quality-driven retransmission limit adaptation scheme for video streaming over VANETs," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1817–1827, May 2012.

[6] X. Cheng, L. Yang, and X. Shen, "D2D for intelligent transportation systems: A feasibility study," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1784–1793, Aug. 2015.

[7] B. Hu, L. Fang, X. Cheng, and L. Yang, "Vehicle-to-vehicle distributed storage in vehicular networks," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

[8] X. Shen, X. Cheng, L. Yang, R. Zhang, and B. Jiao, "Data dissemination in VANETs: A scheduling approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2213–2223, Oct. 2014.

[9] F. Zeng, R. Zhang, X. Cheng, and L. Yang, "Channel prediction based scheduling for data dissemination in VANETs," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1409–1412, Jun. 2017.

[10] H. Liang and W. Zhuang, "Cooperative data dissemination via roadside WLANs," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 68–74, Apr. 2012.

[11] T. H. Luan, L. X. Cai, J. Chen, X. S. Shen, and F. Bai, "Engineering a distributed infrastructure for large-scale cost-effective content dissemination over urban vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1419–1435, Mar. 2014.

[12] R. Zhang, X. Cheng, L. Yang, X. Shen, and B. Jiao, "A novel centralized TDMA-based scheduling protocol for vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 411–416, Feb. 2015.

[13] X. Cheng, C. Chen, W. Zhang, and Y. Yang, "5G-enabled cooperative intelligent vehicular (5GenCIV) framework: When Benz meets Marconi," *IEEE Intell. Syst.*, vol. 32, no. 3, pp. 53–59, May/Jun. 2017.

[14] R. Yu *et al.*, "Optimal resource sharing in 5G-enabled vehicular networks: A matrix game approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7844–7856, Oct. 2016.

[15] C.-X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[16] X. Li *et al.*, "5G-crosshaul network slicing: Enabling multi-tenancy in mobile transport networks," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 128–137, Aug. 2017.

[17] H. Shariatmadari *et al.*, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

[18] L. Chen *et al.*, "Capacity and delay tradeoff of secondary cellular networks with spectrum aggregation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3974–3987, Jun. 2018.

[19] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "YouTube everywhere: Impact of device and infrastructure synergies on user experience," in *Proc. Internet Meas. Conf. (IMC)*, Berlin, Germany, 2011, pp. 345–360.

[20] A. Rao *et al.*, "Network characteristics of video streaming traffic," in *Proc. CoNEXT*, Tokyo, Japan, 2011, p. 25.

[21] M. Li, Z. Yang, and W. Lou, "CodeOn: Cooperative popular content distribution for vehicular networks using symbol level network coding," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 223–235, Jan. 2011.

[22] S. H. Ahmed *et al.*, "CODIE: Controlled data and interest evaluation in vehicular named data networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 6, pp. 3954–3963, Jun. 2016.

[23] Z. Zhou *et al.*, "Social big-data-based content dissemination in Internet of Vehicles," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 768–777, Feb. 2018.

[24] X. Cheng *et al.*, "Wideband channel modeling and intercarrier interference cancellation for vehicle-to-vehicle communication systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 434–448, Sep. 2013.

[25] X. Cheng, C.-X. Wang, B. Ai, and H. Aggoune, "Envelope level crossing rate and average fade duration of nonisotropic vehicle-to-vehicle Ricean fading channels," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 62–72, Feb. 2014.

[26] S. Wu, C.-X. Wang, E.-H. M. Aggoune, M. M. Alwakeel, and X. You, "A general 3-D non-stationary 5G wireless channel model," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3065–3078, Jul. 2018.

[27] A. Ghazal *et al.*, "A non-stationary IMT-advanced MIMO channel model for high-mobility wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2057–2068, Apr. 2017.

[28] M. Siekkinen, M. A. Hoque, and J. K. Nurminen, "Using viewing statistics to control energy and traffic overhead in mobile video streaming," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1489–1503, Jun. 2016.

[29] F. M. Tabrizi, J. Peters, and M. Hefeeda, "Dynamic control of receiver buffers in mobile video streaming systems," *IEEE Trans. Mobile Comput.*, vol. 12, no. 5, pp. 995–1008, May 2013.

[30] M. Usman *et al.*, "Frame interpolation for cloud-based mobile video streaming," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 831–839, May 2016.

[31] Y. Liu and J. Y. B. Lee, "Post-streaming rate analysis—A new approach to mobile video streaming with predictable performance," *IEEE Trans. Mobile Comput.*, vol. 16, no. 12, pp. 3488–3501, Dec. 2017.

[32] M. Li, "Queueing analysis of unicast IPTV with adaptive modulation and coding in wireless cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9241–9253, Oct. 2017.

[33] M.-K. Jiau, S.-C. Huang, J.-N. Hwang, and A. V. Vasilakos, "Multimedia services in cloud-based vehicular networks," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 3, pp. 62–79, Jul. 2015.

[34] M. B. Brahim, Z. H. Mir, W. Znaidi, F. Filali, and N. Hamdi, "QoS-aware video transmission over hybrid wireless network for connected vehicles," *IEEE Access*, vol. 5, pp. 8313–8323, Mar. 2017.

[35] H. He, H. Shan, A. Huang, and L. Sun, "Resource allocation for video streaming in heterogeneous cognitive vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7917–7930, Oct. 2016.

[36] F. Soldo, C. Casetti, C.-F. Chiasserini, and P. A. Chaparro, "Video streaming distribution in VANETs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 7, pp. 1085–1091, Jul. 2011.

[37] J. Wu, B. Cheng, C. Yuen, Y. Shang, and J. Chen, "Distortion-aware concurrent multipath transfer for mobile video streaming in heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 4, pp. 688–701, Apr. 2015.

[38] T. Dreibholz *et al.*, "Stream control transmission protocol: Past, current, and future standardization activities," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 82–88, Apr. 2011.

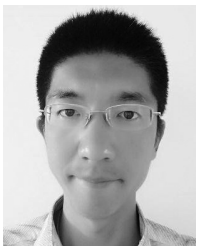
[39] C. Xu, E. Fallon, Y. Qiao, L. Zhong, and G.-M. Muntean, "Performance evaluation of multimedia content distribution over multi-homed wireless networks," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 204–215, Jun. 2011.

[40] N. M. Freris, C.-H. Hsu, J. P. Singh, and X. Zhu, "Distortion-aware scalable video streaming to multinetwork clients," *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 469–481, Apr. 2013.

[41] S. Han, H. Joo, D. Lee, and H. Song, "An end-to-end virtual path construction system for stable live video streaming over heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 5, pp. 1032–1041, May 2011.
[42] J. Lee and S. Bahk, "On the MDP-based cost minimization for video-on-demand services in a heterogeneous wireless network with multihomed terminals," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, pp. 1737–1749, Sep. 2013.
[43] D. Wu, J. Huang, J. He, M. Chen, and G. Zhang, "Toward cost-effective mobile video streaming via smart cache with adaptive thresholding," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 639–650, Dec. 2015.
[44] J. He, Z. Xue, D. Wu, D. O. Wu, and Y. Wen, "CBM: Online strategies on cost-aware buffer management for mobile video streaming," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 242–252, Jan. 2014.
[45] G. Ma *et al.*, "Understanding performance of edge content caching for mobile video streaming," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1076–1089, May 2017.
[46] L. Xiang *et al.*, "Cross-layer optimization of fast video delivery in cache- and buffer-enabled relaying networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11366–11382, Dec. 2017.
[47] T. Kosch, C. J. Adler, S. Eichler, C. Schroth, and M. Strassberger, "The scalability problem of vehicular ad hoc networks and how to solve it," *IEEE Wireless Commun.*, vol. 13, no. 5, pp. 22–28, Oct. 2006.
[48] M. Xing, J. He, and L. Cai, "Utility maximization for multimedia data dissemination in large-scale VANETs," *IEEE Trans. Mobile Comput.*, vol. 16, no. 4, pp. 1188–1198, Apr. 2017.
[49] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 996–1019, 3rd Quart., 2013.
[50] J. W. Cohen, *The Single Server Queue*. Amsterdam, The Netherlands: North-Holland, 1982.
[51] K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, and C. Diot, "Measurement and analysis of single-hop delay on an IP backbone network," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 908–921, Aug. 2003.
[52] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[53] J. G. Andrews, F. Bacccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[54] F. Jarai-Szabo and Z. Neda, "On the size-distribution of Poisson Voronoi cells," *Physica A Stat. Mech. Appl.*, vol. 385, no. 2, pp. 518–526, 2007.

**Xuemin Hong** (S'05–M'12) received the Ph.D. degree from Heriot-Watt University, Edinburgh, U.K., in 2008.

He is currently a Professor with the School of Information Science and Engineering, Xiamen University, Xiamen, China. He has authored or co-authored 1 book chapter and over 60 papers in refereed journals and conference proceedings. His current research interests include cognitive radio networks, content centric networking, and fifth-generation mobile communications.



**Jiping Jiao** received the B.S. degree in electronic information engineering from Hainan University, Haikou, China, in 2008. He is currently pursuing the Ph.D. degree at the School of Information Science and Engineering, Xiamen University, Xiamen, China.

His current research interests include cognitive radio networks and content-centric networking.

**Ao Peng** received the M.Sc. and Ph.D. degrees in communication and information system from Xiamen University, Xiamen, China, in 2011 and 2014, respectively.

In 2015, he joined the School of Information Science and Engineering, Xiamen University, where he is currently an Assistant Professor. His current research interests include digital signal processing and parameter estimation of dynamic systems.

**Jianghong Shi** received the Ph.D. degree from Xiamen University, Xiamen, China, in 2002.

He is currently a Professor with the School of Information Science and Engineering, Xiamen University. He is also the Director of the West Straits Communications Engineering Center, Zhangzhou, China. His current research interests include wireless communication networks and satellite navigation systems.

**Cheng-Xiang Wang** (S'01–M'05–SM'08–F'17) received the B.Sc. and M.Eng. degrees in communication and information systems from Shandong University, Jinan, China, in 1997 and 2000, respectively, and the Ph.D. degree in wireless communications from Aalborg University, Aalborg, Denmark, in 2004.

He was a Research Assistant with the Hamburg University of Technology, Hamburg, Germany, from 2000 to 2001, a Visiting Researcher with Siemens AG Mobile Phones, Munich, Germany, in 2004, and a Research Fellow with the University of Agder, Grimstad, Norway, from 2001 to 2005. He has been with Heriot-Watt University, Edinburgh, U.K., since 2005, where he became a Professor in 2011. In 2018, he joined Southeast University, Nanjing, China, as a Professor and a Thousand Talent Plan Expert. He has authored 2 books, 1 book chapter, and over 340 papers in refereed journals and conference proceedings. His current research interests include wireless channel modeling and (B)5G wireless communication networks, including green communications, cognitive radio networks, high mobility communication networks, massive MIMO, millimeter wave communications, and visible light communications.

Dr. Wang was a recipient of the Highly Cited Researcher Award by Web of Science in 2017 and nine Best Paper Awards from IEEE Globecom 2010, IEEE ICCT 2011, ITST 2012, the IEEE VTC 2013, IWCMC 2015, IWCMC 2016, the IEEE/CIC ICCC 2016, and WPMC 2016. He served or is currently serving as an Editor for nine international journals, including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY since 2011, the IEEE TRANSACTIONS ON COMMUNICATIONS since 2015, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2009. He was the leading Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS "Special Issue on Vehicular Communications and Networks." He is also a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, "Special Issue on Spectrum and Energy Efficient Design of Wireless Communication Networks" and "Special Issue on Airborne Communication Networks," and the IEEE TRANSACTIONS ON BIG DATA "Special Issue on Wireless Big Data." He served or is serving as a TPC member, the TPC Chair, and the General Chair of over 80 international conferences. He is a Fellow of the IET and HEA.