Radio Resource Allocation for Bidirectional Offloading in Space-Air-Ground Integrated Vehicular Network

Guangchao Wang[1], Sheng Zhou*[1], Zhisheng Niu[1]

1. Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract: Aerial platforms and edge servers have been recognized as two promising building blocks to improve the quality of service (QoS) in space-air-ground integrated vehicular networks (SAGIN). Communication intensive tasks can be offloaded to aerial platforms via broadcasting, while computation intensive tasks can be offloaded to ground edge servers. However, the key issues including how to allocate radio resources and how to determine the task offloading strategy for the two types of tasks, are yet to be solved. In this paper, the joint optimization of radio resource allocation and bidirectional offloading configuration is investigated. To deal with the non-convex nature of the original problem, we decouple it into a two-step optimization problem. In first step, we optimize the bidirectional offloading configuration in the case of the radio resource allocation is known in advance, which is proved to be a convex optimization problem. In second step, we optimize the radio resource allocation through brute-force search method. We use queuing theories to analyze the average delay of the two tasks with respect to the broadcasting capacity and task arrival rate. The offloading strategies with closed-form expressions of communication intensive tasks are proposed. We then propose a heuristic algorithm which is shown to perform better than interior point algorithm in simulations. The numerical results also demonstrate that the aerial platforms and edge servers can significantly reduce the average delay of the tasks under different network conditions.

Keywords: radio resource allocation, bidirectional offloading, space-air-ground integrated networks

This work is sponsored in part by the Nature Science Foundation of China (No. 91638204, No.61871254, No. 61861136003, No. 61571265, No. 61621091), and Hitachi Ltd.

1 Introduction

In recent years, the intelligent vehicular networks have been proposed to enhance the capabilities of information exchange and data processing for connected and automated vehicles , which is an essential application scenario for future intelligent transportation systems (ITS) [1]. To improve the road safety and traffic efficiency, a wide range of new applications are emerging, such as high-definition (HD) map downloading, collision avoidance, image-aided navigation, and etc. However, many communication-computation intensive tasks generated by these applications face challenges of processing and transmitting the significant amount of data with stringent latency requirements. For communications, the dedicated short range communication (DSRC) and long-term evolution (LTE) are two main supporting technologies in the state of the art of vehicular networks [2], which face the issues of coverage and capacity. For computations, although the aerial platforms and the vehicles are expected to be equipped with on-board processing units, it is still challenging to compute intensive tasks that require real-time processing of huge amounts of sensing data [4].

One feasible complementary solution is under development, i.e. space-air-ground integrated networks (SAGIN), which consists of satellites, aerial platforms and terrestrial networks [6-9]. The aerial platforms can be utilized as a base station to support broadcast and multicast services, the feasibility of which has been proved from the perspective of standardization [10,11]. The aerial platforms are able to not only provide large coverage and seamless connectivity, but also have higher probabilities of line of sight (LoS) connections compared with terrestrial networks [12]. In this paper, we consider that the aerial platforms work at broadcast mode, since the communication intensive tasks of vehicles are usually generated from common interests, such as HD maps and infotainment contents.

Furthermore, mobile edge computing (MEC) has been proposed as a new paradigm shift to enhance vehicular computation services [3,5]. In the MEC architecture, The edge servers are deployed at the edge of the radio access network, which is close to the users and enables fast interactive response for computation task offloading. Thus, aerial platforms and edge servers are two essential building blocks to enhance the quality of services (QoS) of emerging vehicular applications.

In this paper, the differences between the space networks and the air networks are not specified because both segments have same features in our considerations, including large coverage, broadcasting capabilities and scarce computation resources. We consider an aerial platform-aided vehicular cloud network, where the edge servers are deployed at the road side units (RSUs). We focus on handling two types of tasks in the network:

1. ***Communication intensive tasks*:** The tasks that require high data rate transmission, such as HD map downloading and video streaming.

2. ***Computation intensive tasks*:** The tasks with high computational complexity and require high speed processing, such as computer vision-based navigation and self-localization.

The communication intensive tasks of terrestrial networks can be offloaded to the aerial platform via broadcasting, while the computation intensive tasks generated by the aerial platforms or the vehicles can be offloaded to the edge servers. The feasibility and performance gain of this bidirectional offloading has been validated in our previous work [8].

In this paper, two important problems are investigated to improve the bidirectional offloading, (1) how to allocate radio resources of terrestrial networks to the two types of tasks, i.e. radio resource allocation (RRA), since both the files of communication intensive tasks and the feedback of computation results need to be transmitted via unicasting, and (2) how many communication intensive tasks and computation intensive tasks are supposed to be offloaded according to the network status, i.e. bidirectional offloading configuration (BOC). We consider a joint optimization of RRA and BOC for communication-computation intensive tasks, which is shown to be a non-convex optimization problem. Therefore we decouple it into a two-step optimization problem. First, we optimize the bidirectional offloading configuration given the radio resource allocation. The closed-form expressions for task offloading configuration of communication intensive tasks are derived. Then, we prove that the optimal resource allocation can only be obtained at the boundary of the feasible domain, which enables us to optimize the radio resource allocation through brute force search under tolerable complexity. Finally, we proposed a heuristic algorithm based on our analysis and obtain near-optimal solution. Extensive simulations are conducted to validate the performance of the proposed algorithm and demonstrate the performance gain of the bidirectional offloading.

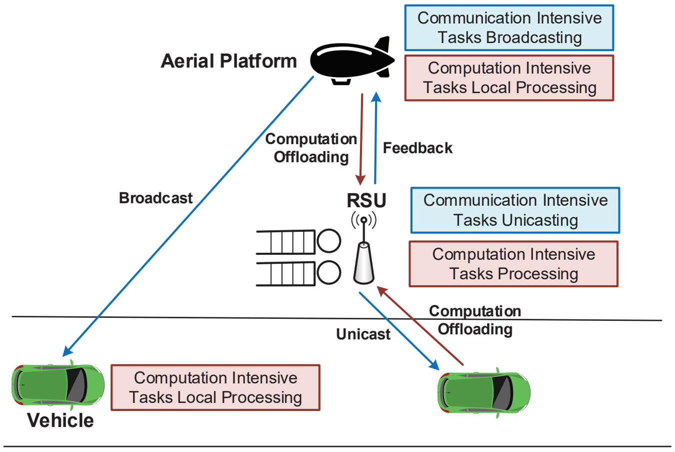

Aerial Platform
Communication Intensive Tasks Broadcasting
Computation Intensive Tasks Local Processing
Computation Offloading
Feedback
Broadcast
RSU
Communication Intensive Tasks Unicasting
Computation Intensive Tasks Processing
Computation Offloading
Unicast
Computation Intensive Tasks Local Processing
Vehicle

Figure 1 Aerial platform-aided vehicular network.

The rest of the paper is organized as follows. In section 2, we provide the system model for the aerial platform-aided vehicular network and queuing model for the two types of tasks. In section 3, the problem formulation is introduced and the analytical results are provided followed by a heuristic algorithm. The numerical results are presented and discussed in section 4. Finally, the paper is concluded in section 5.

2 System Model

As shown in Fig. 1, we consider an aerial platform-aided vehicular cloud network, which is a hybrid broadcast-unicast system. We assume that both communication intensive tasks and computation intensive tasks arrive following Poisson process of rate λ_D and λ_R respectively. The vehicles are assumed to be equipped with dual broadcast-unicast receiver. Thus, the communication intensive tasks can be transmitted either via the broadcast system or the unicast system. The broadcast entity is the aerial platform, which stores the files of the communication intensive tasks over a large coverage area [13]. We assume that the aerial platform forms multiple cells through advanced beamforming technologies [14] and can broadcast the task files to all vehicles within its coverage [10]. The broadcasting capacity is denoted by r_H Mbps. The unicast entity is the RSU, which stores the files of the communication intensive tasks of its own service region. The unicasting capacity is denoted by r_D Mbps. When a vehicle enters the service region of the RSU, it can download the corresponding task files from the RSU via unicasting, or it can download the task files from the aerial platform via broadcasting.

The RSU also acts as an edge server. Therefore, the computation intensive tasks of the vehicles can either be processed locally using on-board computing resources, or be offloaded to the RSU. If the task is offloaded to the RSU, the feedback needs to be transmitted to the vehicles after processing. The transmission capacity of the feedback is denoted by r_R Mbps. We assume that the processing time of both the local on-board processing unit and the edge server follow exponential distributions, with parameters μ_L^{-1} and μ_R^{-1} respectively.

In this model, the RSU plays two roles as the unicast entity for the communication intensive tasks and the edge server for the computation intensive tasks, respectively. Thus, the RSU will manage two queues for the two types of tasks, respectively.

We denote the average delay of the communication intensive tasks by d_D, which consists of the transmission delay and the queuing delay. The communication intensive tasks are generated when the vehicles enter the service region of the RSU, the corresponding files should be downloaded from either the aerial platform or the RSU. The file size is denoted by L_D. If the task is transmitted via broadcasting from the aerial platform, the transmission delay is L_D/r_H. Since the aerial platform forms multiple cells through beamforming, the vehicles can obtain the task file without queuing. If the task file is transmitted via unicasting from the RSU, the service procedure can be modeled as an M/D/1 queue with service rate r_D/L_D. We denote the probability that the communication intensive task is transmitted via broadcasting by x. Thus, the arrival rate of the communication intensive tasks for the RSU is $(1-x)\lambda_D$. We then obtain the average delay of communication intensive tasks by

$$d_D = \frac{L_D x}{r_H} + \frac{L_D(1-x)}{r_D} + \frac{\lambda_D L_D^2 (1-x)^2}{2(r_D^2 - (1-x)\lambda_D L_D r_D)}. \quad (1)$$

The first term is the transmission delay via broadcasting. The sum of the second term and the third term is the sojourn time of the M/D/1 queue [15], where the second term is the transmission delay via unicasting and the last term is the waiting delay.

We denote the average delay of the computation intensive tasks by d_R, which consists of the transmission delay, the processing delay and the queuing delay. If the task is processed locally, then the average processing delay is μ_L^{-1}. If the task is offloaded to the RSU, the average delay is composed of uploading delay of the task file and the service delay in RSU. As the uploading procedure is not our main focus, we assume that the uploading delay is a constant, denoted by d_{up}. The service for computation intensive tasks in RSU consists of the processing of the tasks and the transmission of the feedback, which can be modeled as an M/G/1 queue. We denote the file size of the result feedback by L_R and the probability that the computation intensive task is processed locally by y. Thus, the arrival rate of the computation intensive tasks for the RSU is $(1-y)\lambda_R$. According to Pollaczek-Khinchine formula [15], the average delay of computation intensive tasks is

$$\begin{aligned} d_R = {} & \mu_L^{-1} y + d_{up}(1-y) + \mu_R^{-1}(1-y) + \frac{L_R(1-y)}{r_R} \\ & + \frac{(1-y)^2 \lambda_R (2r_R^2 + 2L_R\mu_R r_R + L_R^2\mu_R^2)}{2\mu_R((\mu_R - (1-y)\lambda_R)r_R^2 - (1-y)\lambda_R L_R \mu_R r_R)}. \end{aligned} \quad (2)$$

The first term is the local processing delay. The second term is the uploading delay of the task file. The remaining part of the equation is the sojourn time of the M/G/1 queue [15], where the third term is the processing delay via offloading, the fourth term is the transmission delay of the feedback and the last term is the waiting delay.

3 Problem Formulation and Solution

Our goal is to minimize the weighted sum of the average delay of the two types of tasks. The radio resource of the RSU is limited, the total capacity of which is denoted by C_R. The radio resource needs to be allocated to the unicast transmission of communication intensive tasks and the feedback transmission of computation intensive tasks, respectively. The offloading probability also needs to be properly configured for the two types of tasks. The problem is how to find the optimal allocation of the radio resource to the two types of tasks as well as the optimal configuration of both communication offloading and computation offloading. Then, the joint optimization problem can be expressed as

$$\begin{aligned} \mathbf{P1}: \min_{x,y,r_D,r_R} \quad & d_D + \omega d_R, \\ s.t. \quad & \mathbf{C1}: r_D + r_R \le C_R, \\ & \mathbf{C2}: (1-x)\lambda_D - \frac{r_D}{L_D} \le 0, \\ & \mathbf{C3}: (1-y)\lambda_R - \frac{\mu_R r_R}{\mu_R L_R + r_R} \le 0, \\ & \mathbf{C4}: x, y \in [0,1], \end{aligned} \quad (3)$$

where ω is the weight indicating the importance of d_R compared with d_D. The constraint **C1** indicates that the sum of the radio resources that allocated to the two types of tasks can not exceed the resource capacity of the RSU. The constraints **C2** and **C3** guarantee the stability of two queues in RSU. **P1** is a non-convex optimization problem due to the non-convexity of the objective function, which is hard to solve even numerically. We consider to decouple this joint optimization problem into a two-step optimization problem. In the first step, we optimize the task offloading configuration under the condition that the allocation of the radio resources is given. In the second step, we consider how to allocate the radio resources and propose a heuristic algorithm to obtain near-optimal solution.

Given r_D and r_R, we can decouple **P1** into **P2** and **P3** as

$$\begin{aligned}
&\mathbf{P2}:\min_{x} \quad d_D,\\
&s.t. \quad (1-x)\lambda_D-\frac{r_D}{L_D}\leq 0,\\
&\qquad x\in[0,1].\\
&\mathbf{P3}:\min_{y} \quad d_R,\\
&s.t. \quad (1-y)\lambda_R-\frac{\mu_R r_R}{\mu_R L_R+r_R}\leq 0,\\
&\qquad y\in[0,1].
\end{aligned} \tag{4}$$

For **P2** and **P3**, we have a lemma as follows:

Lemma 1 **P2** and **P3** are convex optimization problems.

Proof See appendix A.

Then, **P2** and **P3** can be solved in polynomial time using existing algorithms, such as interior point algorithm [16]. Furthermore, by solving **P2**, we can derive the closed-form expression of the communication offloading probability x, as follows:

Theorem 1 When $\frac{\lambda_D L_D}{r_D}>1$, indicating that the terrestrial network is congested. Then the aerial platform is necessary, and we have

$$x=\begin{cases}1, & r_H\geq r_D,\\ x_{\text{on}}, & r_H<r_D.\end{cases} \tag{5}$$

When $\frac{\lambda_D L_D}{r_D}\leq 1$, indicating that the terrestrial network is not congested. Then the aerial platform is used only when the broadcasting capacity is larger than a threshold $r_{H,\text{on}}$, and we have

$$x=\begin{cases}1, & r_H\geq r_D,\\ x_{\text{on}}, & r_{H,\text{on}}<r_H<r_D,\\ 0, & r_H\leq r_{H,\text{on}},\end{cases} \tag{6}$$

where

$$x_{\text{on}}=\frac{(\lambda_D L_D-r_D)(2r_D-r_H)+r_D\sqrt{(2r_D-r_H)r_H}}{\lambda_D L_D(2r_D-r_H)}. \tag{7}$$

The threshold is given by

$$r_{H,\text{on}}=\frac{2r_D(r_D-\lambda_D L_D)^2}{r_D^2+(r_D-\lambda_D L_D)^2}. \tag{8}$$

Proof See appendix B.

In next step, we consider how to allocate the radio resource to the two types of tasks. We have a lemma as follows:

Lemma 2 The average delay for communication intensive tasks d_D is monotonically decreasing with r_D, and the average delay for computation intensive tasks d_R is monotonically decreasing with r_R.

Proof See appendix C.

We denote the the minimum sum average delay of the two types of tasks by d_{opt}. Then, we have a proposition as follows:

Theorem 2 d_{opt} is obtained only when $r_D+r_R=C_R$.

Proof See appendix D.

According to lemma 1 and theorem 1, we can obtain optimal task offloading configuration in polynomial time if the radio resource allocation is given. Theorem 2 indicates that the optimal radio resource allocation can only be obtained at the boundary of the feasible domain. Thus we can optimize the radio resource allocation based on brute-force search under tolerable complexity. Then, we propose a heuristic algorithm based on our analysis to jointly solve the radio resource allocation and task offloading configuration and obtain near-optimal solution as Algorithm 1.

Algorithm 1 Heuristic algorithm for RRA & BOC

```
Initialization: $r_D=\Delta r$ and $r_R=C_R-\Delta r$, set $d_{\text{opt}}$ to a large number
while $r_D<C_R$ do
    obtain $d_{\min}=\min_x d_D+\min_y d_R$ and corresponding $x,y$ by solving P2 and P3
    if $d_{\text{opt}}>d_{\min}$ then
        $d_{\text{opt}}=d_{\min}, x_{\text{opt}}=x, y_{\text{opt}}=y,$
        $r_{D,\text{opt}}=r_D, r_{R,\text{opt}}=r_R$
    end if
    $r_D=r_D+\Delta r$, $r_R=r_R-\Delta r$
end while
return $d_{\text{opt}}, x_{\text{opt}}, y_{\text{opt}}, r_{D,\text{opt}}, r_{R,\text{opt}}$
```

The proposed heuristic algorithm can be divided into two steps. First, we initialize the resource allocation scheme, and the optimal bidirectional offloading configuration is obtained by solving **P2** with Proposition 2 and solving **P3** with interior point algorithm. Then, the optimal radio resource allocation is searched by brute force search. The performance of the algorithm highly depends on the resource searching step Δr. If Δr is small, we can get better results at the expense of longer computing time. If Δr is large, the algorithm will run fast but the results are worse.

4 Numerical Results

In this section, the performance of the proposed heuristic algorithm is evaluated. The weight ω is set as 1 and other

Table 1 Simulation Parameters

Parameter	Value	Description
$\{L_D, L_R\}$	$\{2,0.128\}$ Mbit	File size
$\{\lambda_D, \lambda_R\}$	$\{20,20\}$ tasks/s	Task arrival rate
$\{\mu_L, \mu_R\}$	$\{5,20\}$ tasks/s	Processing rate
d_{up}	10 ms	Uploading delay
C_R	60 Mbps	Radio resource capacity
r_H	20 Mbps	Broadcasting capacity
Δr	0.5 Mbps	Searching step

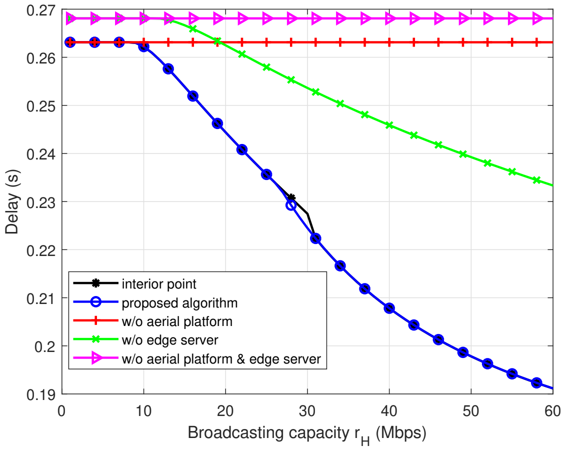


Figure 2 Average delay versus broadcasting capacity.

basic simulation parameters are listed in Table 1.

Fig. 2 shows the delay performance of the proposed algorithm compared with that of the interior point algorithm versus the broadcast capacity. The results of the network without aerial platform, the network without edge server and the network without both aerial platform and edge server are presented as benchmarks. Note that the proposed algorithm achieve better performance than interior point algorithm in some cases. This is because that the interior point algorithm can only obtain local optimal solution for non-convex optimization problems, the performance of which can not be guaranteed due to the uncertainty of the initial point selection. However, the result of our proposed algorithm can gradually approach the optimal solution as Δr decreases. The results also indicate that both aerial platform and edge server are necessary to decrease the average delay of the two types of tasks.

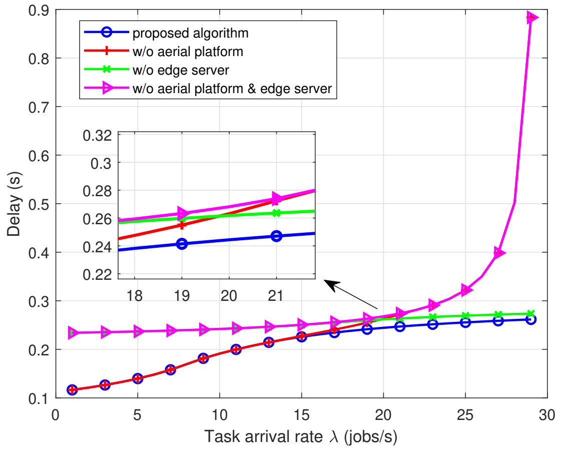


Figure 3 Average delay versus task arrival rate.

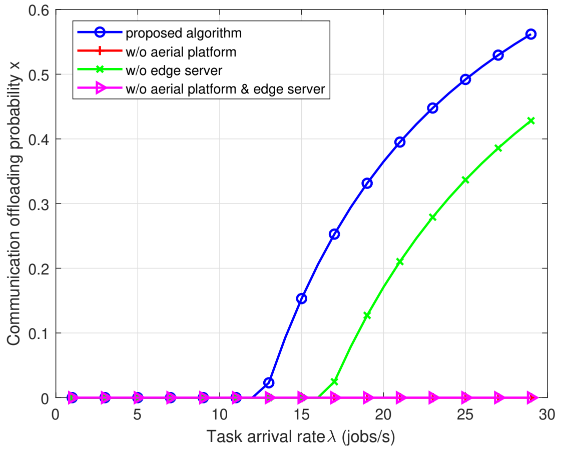


Figure 4 Communication offloading probability versus task arrival rate.

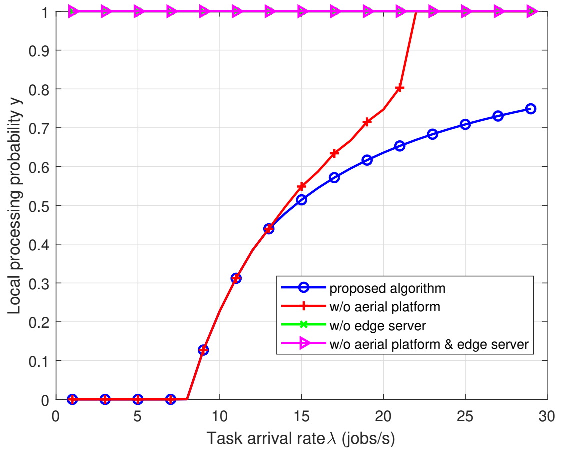


Figure 5 Local processing probability versus task arrival rate.

The contributions of the aerial platform and the edge server under different levels of network congestions are validated in Fig. 3. For simplification, we set λ_D and λ_R to the same value λ. We can see that, the network with both aerial platform and edge server always have best delay performance. When the network is not congested, the network without edge server has very long delay due to higher local processing delay. However, the gap decreases as λ increases. This is because when the network is congested, more radio resources are allocated to the communication intensive tasks to guarantee the queuing

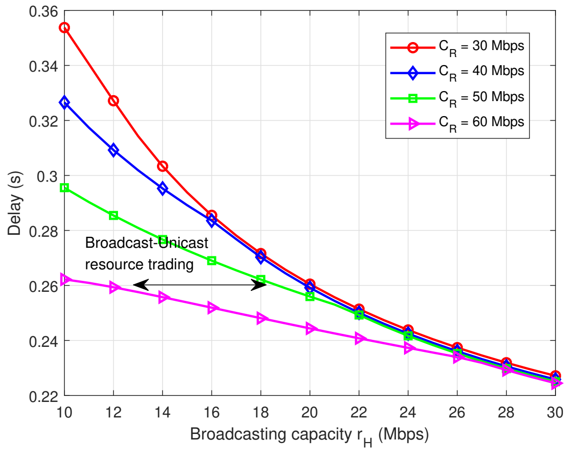


Figure 6 Tradeoff between broadcast resources and unicast resources.

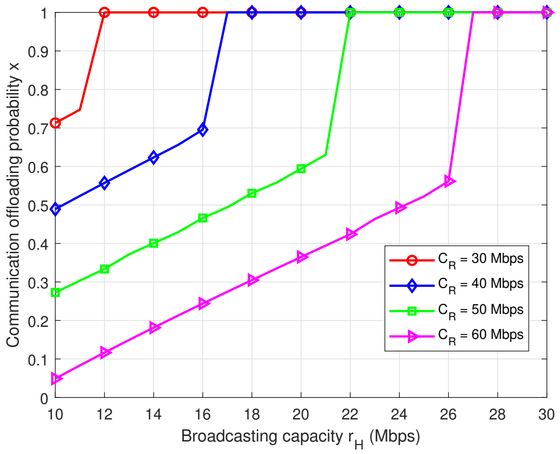


Figure 7 Communication offloading probability versus broadcasting capacity.

stability and more computation intensive tasks have to be processed locally. On the other hand, under non-congested case, the network without aerial platform has good delay performance. This is because the radio resource of the RSU is rich enough to handle the two types of tasks and the aerial platform is not used. However, the gap increases as λ increases, because more communication intensive tasks are offloaded to the aerial platform to relief the burden of the RSU. But for the network without aerial platform, the communication tasks have to be transmitted via unicasting and the queuing delay increases as λ increases. As shown in Fig. 4, the communication offloading probability increases as λ increases. Moreover, the network with both edge server and aerial platform has higher communication offloading probability than the network without edge server. This is because higher communication offloading probability indicates that more radio resources of the RSU are allocated to the computation intensive tasks. As shown in Fig. 5, the local processing probability also increases as λ increases. Furthermore, the network with both edge server and aerial platform has lower local processing probability than the network without aerial platform. This is because more radio resources are left for the computation offloading due to the communication offloading.

Fig. 6 shows the tradeoff between broadcast resources and unicast resources. As can be seen, the delay performance decreases as broadcast capacity increases. Moreover, when the unicast resources are tight or when the delay requirement is high, a small amount of broadcast resources can trade for a large amount of unicast resources. On the other hand, when the unicast resources are rich or when the delay requirement is low, a small amount of unicast resources can trade for a large amount of broadcast resources. Fig. 7 and Fig. 8 show that both communication and computation offloading probabilities increase as broadcast capacity increases. We can observe that communication offloading probability is large when C_R is small, indicating that more communication intensive tasks are tend to be offloaded to aerial platforms if the radio resource of terrestrial network is scarce. When C_R is large, more computation intensive tasks are offloaded to the edge servers if broadcast resources are rich, otherwise the tasks are preferred to be processed locally.

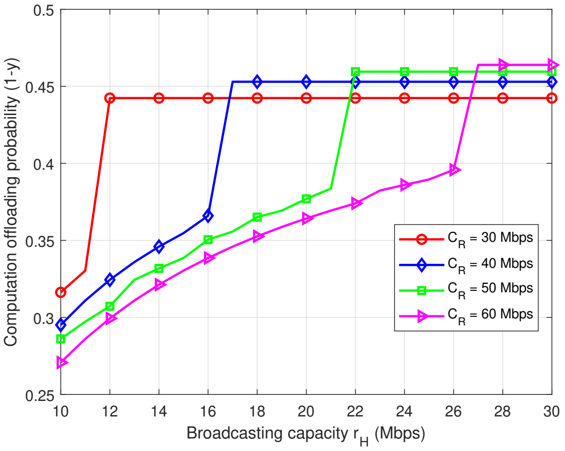


Figure 8 Computation offloading probability versus broadcasting capacity.

5 Conclusion

In this paper, we consider the joint optimization of RRA and BOC for communication-computation intensive tasks in an aerial platform-aided vehicular cloud network. The problem is hard to be solved directly because of the non-convex nature of the objective function. The analytical results indicate that the aerial platform is necessary when the terrestrial network is congested, otherwise aerial platform is helpful only when its broadcast capacity is larger than a threshold. A heuristic algorithm is proposed to obtain near-optimal solution based on the analytical results, which is shown to have good performance. The simulations also validate the performance gain of bidirectional offloading for communication-computation intensive tasks in vehicular networks, and demonstrate the tradeoff between broadcast resources and unicast resources.

Appendix

A) For **P2**, taking the second-order derivative of d_D with respect to x, we have

$$\frac{\partial^2 d_D}{\partial x^2} = \frac{\lambda_D L_D^2 r_D}{[r_D - (1-x)\lambda_D L_D]^3}. \tag{9}$$

According to the constraints of **P2**, we have $(1-x)\lambda_D - \frac{r_D}{L_D} \leq 0$, which is equivalent to

$$r_D - (1-x)\lambda_D L_D \geq 0. \tag{10}$$

Since the numerator of equation (9) is nonnegative, we have $\frac{\partial^2 d_D}{\partial x^2} \geq 0$. According to [16], the objective function of **P2** is a convex function. As all constraints are linear constraints, **P2** is proved to be a convex optimization problem.

Similarly, taking the second-order derivative of d_R with respect to y, we have

$$\frac{\partial^2 d_R}{\partial y^2} = \frac{\lambda_R \mu_R r_R (2r_R^2 + 2L_R \mu_R r_R + L_R^2 \mu_R^2)}{[\mu_R r_R - (1-y)\lambda_R (r_R + L_R \mu_R)]^3}. \tag{11}$$

According to the constraints of **P3**, we have $(1-y)\lambda_R - \frac{\mu_R r_R}{\mu_R L_R + r_R} \leq 0$, which is equivalent to

$$\mu_R r_R - (1-y)\lambda_R (r_R + L_R \mu_R) \geq 0. \tag{12}$$

The numerator of equation (11) is also nonnegative, thus we have $\frac{\partial^2 d_R}{\partial y^2} \geq 0$. Then, the objective function of **P3** is also convex. Since all constraints are linear constraints, **P3** is proved to be a convex optimization problem. Lemma 1 is proved.

B) The Lagrangian function of **P2** is

$$\begin{aligned} L(x,\boldsymbol{\nu}) =& \frac{L_D x}{r_H} + \frac{L_D(1-x)}{r_D} + \frac{\lambda_D L_D^2 (1-x)^2}{2(r_D^2 - (1-x)\lambda_D L_D r_D)} \\ &+ \nu_1 \left[(1-x)\lambda_D - \frac{r_D}{L_D} \right] - \nu_2 x + \nu_3 (x-1), \end{aligned} \tag{13}$$

where $\boldsymbol{\nu} = [\nu_1, \nu_2, \nu_3]$ are the Lagrange multipliers. Since the constraints are all linear inequalities, the Slater's condition holds. Therefore strong duality holds for the prime problem and dual problem. According to lemma 1, let $x^\star, \boldsymbol{\nu}^\star$ are primal and dual optimal points, then the Karush-Kuhn-Tucker (KKT) conditions must be satisfied [16]

$$\begin{aligned} &(1-x^\star)\lambda_D - \frac{r_D}{L_D} \leq 0,\ -x^\star \leq 0,\ x^\star - 1 \leq 0, \\ &\nu_1^\star \geq 0,\ \nu_2^\star \geq 0,\ \nu_3^\star \geq 0, \\ &\nu_1^\star \left[(1-x^\star)\lambda_D - \frac{r_D}{L_D} \right] = 0,\ \nu_2^\star x^\star = 0,\ \nu_3^\star (x^\star - 1), \\ &\frac{\lambda_D L_D^2 (1-x)}{\lambda_D L_D r_D (1-x) - r_D^2} - \frac{\lambda_D^2 L_D^3 r_D (1-x)^2}{2[\lambda_D L_D r_D (1-x) - r_D^2]^2} \\ &+ \frac{L_D}{r_H} - \frac{L_D}{r_D} - \nu_1^\star \lambda_D - \nu_2^\star + \nu_3^\star = 0. \end{aligned} \tag{14}$$

(1) When $\frac{\lambda_D L_D}{r_D} > 1$, we have $1 \geq x^\star \geq 1 - \frac{r_D}{L_D \lambda_D} > 0$. Thus $\nu_2^\star = 0$. Let $x^\star = 1$, we have $\nu_1^\star = 0$ and $\frac{L_D}{r_H} - \frac{L_D}{r_D} + \nu_3^\star = 0$. Then we have $\nu_3^\star = \frac{L_D}{r_D} - \frac{L_D}{r_H} \geq 0$. Therefore $r_H \geq r_D$ must be satisfied. Let $x^\star < 1$, we have $\nu_3^\star = 0$. Then we can eliminate $\nu_1^\star$ and obtain

$$\begin{aligned} (2r_D - r_H) L_D^2 \lambda_D^2 (1-x^\star)^2 - 2L_D \lambda_D r_D (2r_D - r_H)(1-x^\star) \\ -2(r_H - r_D) r_D^2 = 0 \end{aligned} \tag{15}$$

Since we should guarantee that the quadratic equation (15) has feasible point with $x \in (1 - \frac{r_D}{L_D \lambda_D}, 1)$, we can derive that

$$x^\star = \frac{(\lambda_D L_D - r_D)(2r_D - r_H) + r_D \sqrt{(2r_D - r_H) r_H}}{\lambda_D L_D (2r_D - r_H)} \triangleq x_{\text{on}}, \tag{16}$$

when $r_H < r_D$.

(2) When $\frac{\lambda_D L_D}{r_D} \leq 1$, we have $0 \leq x^\star \leq 1$. Let $x^\star = 0$, we have $\nu_1^\star = 0$ and $\nu_3^\star = 0$. Then we have

$$\begin{aligned} \nu_2^\star =& \frac{\lambda_D L_D^2 (1-x)}{\lambda_D L_D r_D (1-x) - r_D^2} - \frac{\lambda_D^2 L_D^3 r_D (1-x)^2}{2[\lambda_D L_D r_D (1-x) - r_D^2]^2} \\ &+ \frac{L_D}{r_H} - \frac{L_D}{r_D} \geq 0 \end{aligned} \tag{17}$$

Therefore we can obtain

$$r_H \leq \frac{2r_D (r_D - \lambda_D L_D)^2}{r_D^2 + (r_D - \lambda_D L_D)^2} \triangleq r_{H,\text{on}}. \tag{18}$$

Let $0 < x^\star < 1$, we have $\nu_2^\star = 0$ and $\nu_3^\star = 0$. By eliminating $\nu_1^\star$ we can also obtain equation (15). We should guarantee that the quadratic equation (15) has feasible point with $x \in (0,1)$. Then we can derive that $x^\star = x_{\text{on}}$, when $r_{H,\text{on}} < r_H < r_D$. Let $x^\star = 1$, we have $\nu_1^\star = 0$ and $\nu_2^\star = 0$. Then we have $\nu_3^\star = \frac{L_D}{r_D} - \frac{L_D}{r_H} \geq 0$. Therefore we obtain $r_H \geq r_D$. Theorem 1 is proved.

C) Taking the derivative of d_D with respect to r_D, we have

$$\frac{\partial d_D}{\partial r_D} = -\frac{L_D(1-x)}{r_D^2} - \frac{\lambda_D L_D^2 (1-x)^2 (2r - \lambda_D L_D (1-x))}{2r_D^2 [r_D - \lambda_D L_D (1-x)]^2}. \tag{19}$$

According to the equation (10), we have $\frac{\partial d_D}{\partial r_D} \leq 0$. Thus d_D is monotonically decreasing with r_D when x is given.

Similarly, taking the derivative of d_R with respect to r_R, we have

$$\begin{aligned}\frac{\partial d_R}{\partial r_R} &= -\frac{L_R(1-y)}{r_R^2} \\ &- \frac{\lambda_R L_R \mu_R (1-y)^2\{2r_R^2 + L_R[\mu_R - (1-y)\lambda_R] + L_R\Phi\}}{2r_R^2\Phi^2},\end{aligned} \tag{20}$$

where

$$\Phi = [\mu_R - (1-y)\lambda_R]r_R - (1-y)\lambda_R L_R \mu_R. \tag{21}$$

According to the constraints **C3** in **P1**, we have

$$(1-y)\lambda_R \leq \frac{\mu_R r_R}{\mu_R L_R + r_R} \leq \mu_R, \tag{22}$$

which is equivalent to $\Phi \geq 0$ and $\mu_R - (1-y)\lambda_R \geq 0$. Thus we have $\frac{\partial d_R}{\partial r_R} \leq 0$, indicating that d_R is monotonically decreasing with r_R when y is given.

Given r_{D1} and r_{D2}, we assume minimum d_D are obtained with x_1 and x_2 respectively. Let $r_{D1} < r_{D2}$, we have

$$d_D(r_{D1}, x_1) \leq d_D(r_{D2}, x_1) < d_D(r_{D2}, x_2). \tag{23}$$

Therefore, d_D is monotonically decreasing with r_D. Similarly, we can prove d_R is monotonically decreasing with r_R. Lemma 2 is proved.

D) Assume that $d_{\text{opt}} = d_D + \omega d_R$ is obtained when $r_D + r_R < C_R$. Then we have $r_D + r_R + \Delta r = C_R$, where $\Delta r > 0$. Let $r_D^* = r_D + \Delta r > r_D$, according to lemma 2, we have $d_D(r_D^*) < d_D(r_D)$. Thus $d_{\text{opt}}^* = d_D(r_D^*) + \omega d_R < d_{\text{opt}}$, which is conflict with the assumption. Therefore d_{opt} is obtained only when $r_D + r_R = C_R$. Theorem 2 is proved.

References

[1] Y. Zhang, H. Zhang, K. Long, Q. Zheng and X. Xie, “Software-Defined and Fog-Computing-Based Next Generation Vehicular Networks,” *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 34-41, Sept. 2018.

[2] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang and Y. Zhou, “Heterogeneous Vehicular Networking: A Survey on Architecture, Challenges, and Solutions,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2377-2396, Fourthquarter 2015.

[3] K. Zhang, Y. Mao, S. Leng, Y. He and Y. ZHANG, “Mobile-Edge Computing for Vehicular Networks: A Promising Network Paradigm with Predictive Off-Loading,” *IEEE Veh. Technol. Mag.*, vol. 12, no. 2, pp. 36-44, June 2017.

[4] J. Wang, D. Feng, S. Zhang, J. Tang and T. Q. S. Quek, “Computation Offloading for Mobile Edge Computing Enabled Vehicular Networks,” *IEEE Access*, vol. 7, pp. 62624-62632, 2019.

[5] X. Huang, R. Yu, J. Kang, Y. He and Y. Zhang, “Exploring Mobile Edge Computing for 5G-Enabled Software Defined Vehicular Networks,” *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 55-63, Dec. 2017.

[6] G. Wang, S. Zhou, Z. Niu, and X. S. Shen, “Service Function Chain Planning with Resource Balancing in Space-Air-Ground Integrated Networks,” *IEEE GLOBECOM 19*, Waikoloa, Dec. 2019.

[7] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang and X. S. Shen, “Software Defined Space-Air-Ground Integrated Vehicular Networks: Challenges and Solutions,” *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101-109, 2017.

[8] S. Zhou, G. Wang, S. Zhang, Z. Niu and X. S. Shen, “Bidirectional Mission Offloading for Agile Space-Air-Ground Integrated Networks,” *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 38-45, April 2019.

[9] X. Cao, P. Yang, M. Alzenad, X. Xi, D. Wu and H. Yanikomeroglu, “Airborne Communication Networks: A Survey,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1907-1926, Sept. 2018.

[10] G. Araniti, A. Iera, and A. Molinaro, “The role of HAPs in supporting multimedia broadcast and multicast services in terrestrial-satellite integrated systems,” *Wireless Pers. Commun.*,vol. 32, no. 3-4, pp. 195-213, 2005.

[11] A. Mohammed, A. Mehmood, F. N. Pavlidou, and M. Mohorcic, “The role of high-altitude platforms (HAPs) in the global wireless connectivity,” *Proc. IEEE*, vol. 99, no. 11, pp. 1939-1953, Nov. 2011.

[12] G. Avdikos, G. Papadakis and N. Dimitriou, “Overview of the application of High Altitude Platform (HAP) systems in future telecommunication networks,” *2008 10th International Workshop on Signal Processing for Space Communications*, Rhodes Island, 2008.

[13] S. Zhang, W. Quan, J. Li, W. Shi, P. Yang and X. Shen, “Air-Ground Integrated Vehicular Network Slicing With Content Pushing and Caching,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2114-2127, Sept. 2018.

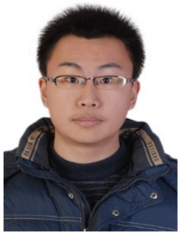
[14] M. Dessouky, H. Sharshar, and Y. Albagory, “Improving the cellular coverage from a high altitude platform by

novel tapered beamforming technique," *J. of Electron. Waves and Appl.*, vol. 21, no. 13, pp. 1721-1731, May 2007.

[15] S. Asmussen, *Applied Probability and Queues.* Springer, 2003.

[16] S. Boyd, and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

About the Authors



Guangchao Wang received the B.S. degree in communications engineering from Beijing Jiaotong University, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree in electronic engineering with Tsinghua University, Beijing, China. Her research interests include space-air-ground integrated network reconfiguration and UAV-aided traffic offloading. (Email: wgc15@mails.tsinghua.edu.cn)

Sheng Zhou [corresponding author] received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2005 and 2011, respectively. From January to June 2010, he was a visiting student at the Wireless System Lab, Department of Electrical Engineering, Stanford University, Stanford, CA, USA. From November 2014 to January 2015, he was a visiting researcher in Central Research Lab of Hitachi Ltd., Japan. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include cross-layer design for multiple antenna systems, mobile edge computing, vehicular networks, and green wireless communications. (Email: sheng.zhou@tsinghua.edu.cn)

Zhisheng Niu graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992-94, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994 joined with Tsinghua University, Beijing, China, where he is now a professor at the Department of Electronic Engineering. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks. Dr. Niu has served as Chair of Emerging Technologies Committee (2014-15), Director for Conference Publications (2010-11), and Director for Asia-Pacific Board (2008-09) in IEEE Communication Society, and currently serving as Director for Online Contents (2018-19) and Area Editor of IEEE Trans. Green Commun. & Networks. He received the Outstanding Young Researcher Award from Natural Science Foundation of China in 2009 and the Best Paper Award from IEEE Communication Society Asia-Pacific Board in 2013. He was also selected as a distinguished lecturer of IEEE Communication Society (2012-15) as well as IEEE Vehicular Technologies Society (2014-18). He is a fellow of both IEEE and IEICE. (Email: niuzhs@tsinghua.edu.cn)