Semi-persistent V2X Resource Allocation with Traffic Prediction in Two-tier Cellular Networks

Ping Chu[1,2], J. Andrew Zhang[2], Xiaoxiang Wang[1], Gengfa Fang[2], Dongyu Wang[1]
[1]Key Laboratory of Universal Wireless Communications, Beijing University of Posts and Telecommunications, China
[2]University of Technology Sydney, Global Big Data Technologies Centre (GBDTC), Australia
ping.chu@student.uts.edu.au; {Andrew.Zhang; Gengfa.Fang}@uts.edu.au; {cpwang; dy_wang}@bupt.edu.cn

***Abstract*—In a dense urban area, conventional cellular V2X communications require frequent and heavy resource allocation, which can lead to processing congestion and large delay. In this paper, we propose a semi-persistent resource allocation scheme using least minimum mean square error (LMMSE) traffic prediction in a two-tier network. The two-tier network architecture includes a central macro base station (MBS) and multiple roadside units (RSU). In the proposed scheme, the MBS pre-allocates persistent resource to RSUs based on predicted traffic, and then allocate dynamic resource upon real-time requests from vehicles through RSUs. We formulate an optimization problem for minimizing the total bandwidth under latency constraints, and provide optimal solution to the problem. Simulation is conducted for both artificially generated and real-world data, and the results validate the effectiveness of the proposed semi-persistent scheme.**

***Index Terms*—V2X communications, resource allocation, LMMSE, heterogeneous networks**

I. Introduction

Vehicular communications have been widely studied in recent years [1]–[4]. The 3rd Generation Partnership Project (3GPP) group has completed initial Cellular Vehicle-to-Everything (V2X) standard [5]. V2X communications aim to make transportation safer, comfortable and more efficient, and also pave the way for autonomous driving. There are mainly two categories of vehicular services, namely non-safety and safety services. Targeting at decreasing traffic accidents, safety service has more stringent requirements on delay and reliability. System latency includes both signal transmission time and the delay caused by associated signalling process, e.g., resource request and allocation.

For cellular V2X communications, resource allocation and interference management are important issues to be addressed due to limited spectrum and resource reuse between traditional cellular users and vehicles. A number of recent works have emerged focusing on V2X resource allocation. A theoretical analysis for resource allocation for D2D-based V2X communications is given in [6], where optimization for capacity without considering delay was investigated. In [7], the authors proposed interference graph-based resource sharing schemes for resource allocation in order to enhance the network throughput. In [8], a resource allocation scheme with latency constraints for LTE V2V communications was proposed, where the scenario of cellular and vehicular users coexistence was studied under a one-tier cellular network. There are also quite a few studies on investigating heterogeneous networks for vehicular communication [9]–[12], with major focus on the cooperation between different vehicular networks for V2X communications. Resource allocation for V2X networks using delay as the priority is still very limited.

In this paper, we propose a semi-persistent resource allocation scheme based on a two-tier heterogeneous cellular network. This scheme can offload the signalling requests that could potentially cause network traffic congestion, and significantly reduce the signalling delay. In the two-tier cellular architecture, a macro base station (MBS) centrally controls multiple micro base stations, which we call as RSUs. The MBS provides centralized control over resource allocation for RSUs, and each RSU provides direct short-range communications to vehicles in its coverage. This architecture can efficiently offload the centralized traffic and reduce processing latency, as well as improving the sum-rate capacity.

Considering the predictability of vehicular network traffic, we combine the prediction into resource allocation for V2X communications, and propose a semi-persistent resource allocation scheme for V2X communications with strict latency requirements. We introduce a non-model-based *linear minimum mean square error (LMMSE)* predictor for predicting vehicular network traffic. Based on predicted traffic, the MBS pre-allocates persistent resource to each RSU and then provides additional real-time dynamic resource allocation to those RSUs with insufficiently allocated resource during pre-allocation. With pre-allocated resource, each RSU can directly provide initial resource allocation to vehicles when receiving their requests, without having to wait for the resource being assigned by the MBS. This pre-allocation can significantly reduce the signaling latency and mitigate congestion. A cost function for the total bandwidth for the area of interest is developed under latency constraints for resource allocation.

The remainder of this paper is organized as follows. Section II formulates the research problem and introduces the semi-persistent resource allocation scheme. Section III presents the *LMMSE* predictor for network traffic prediction and the optimization of cost function. Finally, simulation results are provided in Section IV, and Section V concludes the paper.

II. Semi-persistent Resource Allocation

In this section, we present the system model, propose the semi-persistent resource allocation scheme, and then formulate

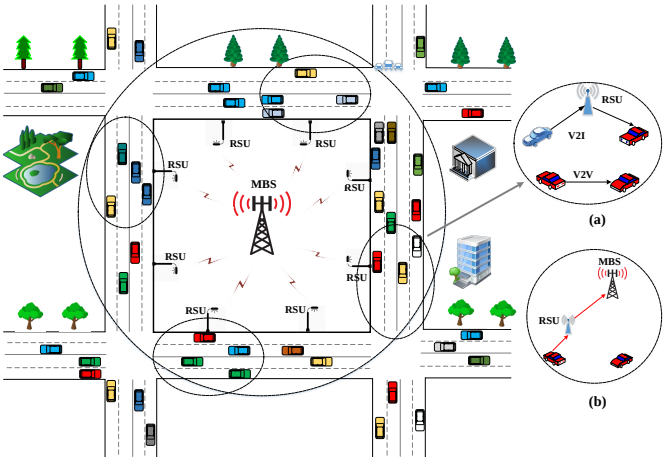

Fig. 1. Two-tier cellular architecture and the segmented road model.

the cost function of the total bandwidth.

A. System Model

As shown in Fig.1, we consider an urban cellular V2X network with a two-tier architecture where a macro base station (MBS) is in the first tier and several roadside units (RSUs) are in the second tier. We divide the area of interest into N segments, where each RSU supports the communication of vehicles in each segment. Denote the segment set as $\mathcal{N} = \{1, 2, ..., N\}$. Assume that MBS provides centralized control over network resource for RSUs, and each RSU is directly responsible for providing access to vehicles in the V2I communication mode, or do local resource allocation for vehicles in the mode of direct V2V communications within its segment as shown in the subfigure Fig. 1 (a).

In this paper, we mainly investigate bandwidth allocation across segments in the area in the two-tier heterogeneous network. Assume that there is no frequency reuse and orthogonal frequency channels are allocated across segments.

B. Semi-persistent Resource Allocation

In the two-tier heterogeneous network, purely dynamic resource allocation will incur delay due to resource requests and confirmation from all vehicular users. As shown in Fig. 1 (b), each vehicle user who wants to communicate with other needs to send a request to MBS via RSU. This is also impractical due to the instantaneous very-high loading resulted from the fast-varying traffic flow in the urban area.

Our semi-persistent scheme combines pre-allocation based on prediction and dynamic allocation, which can achieve an excellent balance between improving bandwidth efficiency and reducing latency. In the proposed semi-persistent scheme, the MBS pre-allocates resources to each RSU based on network traffic prediction in advance, which will be detailed in Section III. In real-time, the MBS then further allocates resources dynamically to the vehicle users who need additional resources based on their delay requirements.

Note that our prediction is only for the mean traffic for one time period of T_s, and hence the resource pre-allocation is applied every T_s seconds. Real-time traffic request and dynamic allocation then happens many times during the T_s seconds. Assume that we are at time tT_s, and are now processing the resource allocation problem for the next time period from tT_s to $(t+1)T_s$. For simplicity, we use $t+1$ to represent this period. We now define some symbols and rules as follows:

- The total allocated persistent and dynamic bandwidth are denoted as B^P, B^D respectively.
- The pre-allocated persistent bandwidth for n-th segment at time $t+1$ is $B_n^p(t+1), n \in \mathcal{N}$. Therefore $B^P = \sum_{n=1}^{N} B_n^p(t+1)$.
- The dynamic bandwidth allocated to n-th segment for $t+1$ is $B_n^d(t+1)$, where $B^D = \sum_{n=1}^{N} B_n^d(t+1)$.
- The actual bandwidth requirement at $t+1$ is denoted as $B_n^{req}(t+1)$.
- When $B_n^p(t+1) < B_n^{\text{req}}(t+1)$, some vehicle users in the n-th segment will request dynamic resources from the MBS via RSU.
- The transmission latency requirements of vehicular users in the n-th segment at time $t+1$ is t_n^r and a signalling latency t_n^s will be introduced for vehicular users who need dynamic allocation.

C. Formulation of the Cost Function

Now, we formulate a cost function for the total required bandwidth in the area of interest. Our main objective is to minimize the total bandwidth under the transmission latency constraints of vehicular users. Based on the proposed semi-persistent scheme, the total required bandwidth can be expressed as

$$B(t+1) = \sum_{n=1}^{N} [B_n^p(t+1) + B_n^d(t+1)]. \tag{1}$$

Therefore, the cost funtion can be defined by

$$\begin{aligned} \min_{\{t_n\}_{n=1}^N} \quad & B(t+1) \\ \text{subject to} \quad & 0 < t_n^r \leq T_n^r. \end{aligned} \tag{2}$$

Where T_n^r is the threshold of latency for vehicular communications in the n-th segment.

III. LMMSE Predictor for Network Traffic and Optimization of Cost Function

A. LMMSE Predictor for Persistent Allocation

Network traffic exhibits high correlation in short timescales and long-range dependence over segments, and such correlation and dependence can be well exploited for traffic prediction. In this paper, we introduce an LMMSE predictor which is widely used for channel estimation and traffic prediction. [13], [14].

Let $b(t)$ be the average network traffic in the $[(t-1)T_s, tT_s]$ time period, where T_s is the interval of observations. We

propose the following M-coefficients linear predictor

$$\hat{b}(t+1) = \sum_{k=0}^{M-1} a(k)b(t-k) \tag{3}$$

for predicting the mean network traffic one sample ahead of the current one, where $a_0, a_1, ..., a_{M-1}$ are the LMMSE coefficients. These coefficients can be obtained as

$$[a(0)\ a(1)\ ...\ a(M-1)] = [R(1)\ R(2)\ ...\ R(L)]\ \mathbf{R}_t^{\dagger},$$
$$\mathbf{R}_t = \begin{bmatrix} R(0) & R(1) & ... & R(L-1) \\ R(1) & R(0) & ... & R(L-2) \\ ... & ... & ... & ... \\ R(M-1) & R(M-2) & ... & R(L-M) \end{bmatrix}, \tag{4}$$

where $L \geq M$, the superscript $\dagger$ denotes the pseudo-inverse of a matrix, $\mathbf{R}_t$ is the autocorrelation matrix at time t with

$$R(i) = \frac{1}{M} \sum_{t=i+1}^{M+i} b(t)b(t-i), 0 \leq i \leq M-1. \tag{5}$$

According to the predicted network traffic, the MBS allocates persistent bandwidth to each segment. Here, we introduce a scaling factor θ instead of directly using the predicted value to allocate the bandwidth to segments, and the pre-allocated persistent bandwidth for the n-the segment is given by

$$B_n^p(t+1) = \theta \hat{b}_n(t+1). \tag{6}$$

The reason we introduce θ here is that there will be gaps between the predicted and real values. Therefore, an appropriate θ is helpful for minimizing the total bandwidth for the area of interest. In Section IV, we will investigate the optimal θ numerically.

B. Minimization of Total Bandwidth

After allocating the persistent bandwidth, we now need to complete dynamic allocation in real time. With a given θ, our goal is to find the optimal dynamic allocation $B_n^d(t+1)$ based on the actually requested bandwidth $B_n^{\text{req}}(t+1)$ and the latency threshold. Note that in the case of $B_n^p(t+1) \geq B_n^{\text{req}}(t+1)$, the allocated persistent bandwidth can already meet the latency requirement of vehicular communications. Let $\mathcal{V}$ denote the index sets of the segments where $B_n^p(t+1) < B_n^{\text{req}}(t+1)$. The size of the sets is denoted as V. Then we can rewrite (1) as

$$B(t+1) = \theta \sum_{n=1}^{N} \hat{b}_n(t+1) + \sum_{v=1}^{V} B_v^d(t+1) \tag{7}$$

According to the Shannon's capacity formula, the transmission rate of a vehicular communication link is $r = B\log(1+\text{SNR})$, where B denotes the channel bandwidth. Assume that b^{req} bits need to be transmitted over one link. Thus, the transmission delay can be expressed by $t = \frac{b^{\text{req}}}{B\log(1+SNR)}$. It is obvious that the transmission delay is inversely proportional to the bandwidth. Therefore, we define the relationship between transmission delay and bandwidth as $t_n = \frac{c_n}{B_n}$. Thus, we can obtain

$$t_v^r = \frac{c_v}{B_v^{req}(t+1) - B_v^p(t+1)}, \forall v \in V. \tag{8}$$

For the v-th segment, the transmission time needs to meet the following relationship based on the constraints of the latency

$$t_v^r - t_v^s = \frac{c_v}{B_v^d(t+1)}. \tag{9}$$

By substituting (8) into (9), the allocated dynamic bandwidth can be expressed by $B_v^d(t+1) = \frac{t_v^r}{t_v^r - t_v^s}[B_v^{req}(t+1) - B_v^p(t+1)]$. We then rewrite (7) as

$$B(t+1) = \theta \sum_{n=1}^{N} \hat{b}_n(t+1) + \sum_{v=1}^{V} \frac{t_v^r}{t_v^r - t_v^s}[B_v^{\text{req}}(t+1) - B_v^p(t+1)]. \tag{10}$$

With a given θ and the constraint $0 < t_v^r \leq T_v^r$, $B_v^d(t+1)$ obtains its minimum value when $t_v^r = T_v^r$. The minimum is given by

$$B_{op}(t+1) = \theta \sum_{n=1}^{N} \hat{b}_n(t+1) + \sum_{v=1}^{V} \frac{T_v^r}{T_v^r - t_v^s}[B_v^{req}(t+1) - B_v^p(t+1)]. \tag{11}$$

IV. Simulation Results

In this section, we present simulation results for the proposed semi-persistent resource allocation scheme (SPRAS) and demonstrate its performance based on two types of data sets. One data set is generated from simulated AR process, and the other is from real measured traffic flow data.

For comparison, we also provide results for the conventional purely dynamic resource allocation scheme (DRAS) which allocates bandwidth based on the latency requirements in real time. The minimum total bandwidth for DRAS is given by

$$B_{op}^D(t+1) = \sum_{n=1}^{N} \frac{T_n^r}{T_n^r - t_n^s} B_n^{req}(t+1). \tag{12}$$

In order to evaluate the performance of proposed prediction model LMMSE, we define and use Mean Absolute Percent Error (MAPE) [15] as a performance metric

$$\text{MAPE} = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{\hat{b}_j - b_j}{b_j} \right|. \tag{13}$$

Larger MAPE means worse prediction performance.

A. Simulation Results based on Autoregressive (AR) Model

Firstly, we generate the average bandwidth needs using an AR model with coefficients $a_i = \exp(r * i), i = 1, 2, 3$ and Gaussian noise with variance 10^{-4}. Let the prediction interval be $T_s = 5$s. Denote the generated average bandwidth for $[tT_s, (t+1)T_s]$ as $\overline{B}_n^{req}(t+1)$. We generate 8 sets of data (i.e., $N = 8$) based on different values of r

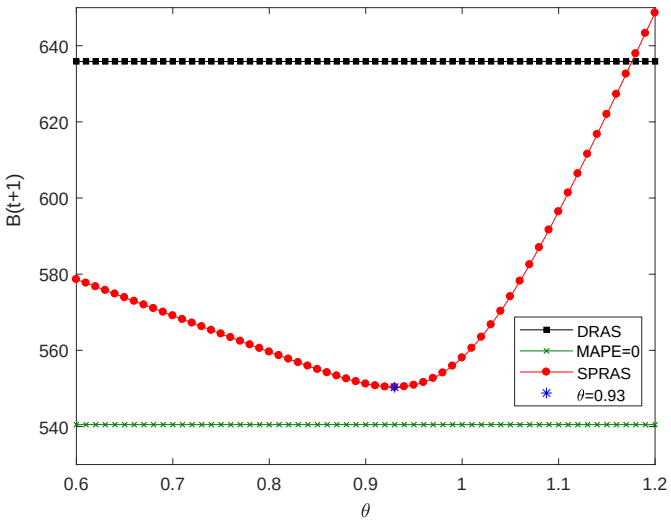

Fig. 2. Minimum total bandwidth versus θ for dynamic and semi-persistent schemes; MAPE=5.63%; $t^s = 15$ms.

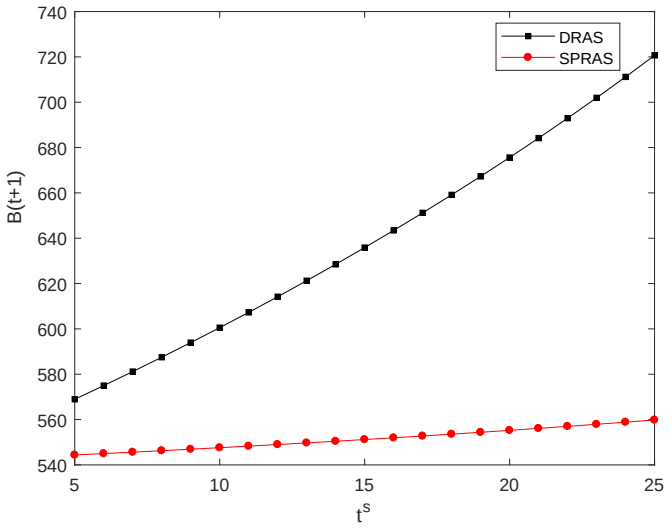

Fig. 3. Achieved total bandwidth versus signalling latency t^s for DRAS and SPRAS.

$(r = -0.95, -1, -1.05, -1.1, -1.15, -1.2, -1.25, 1.3)$. Each set includes 500 data samples. Using the LMMSE predictor, we can obtain the predicted bandwidth requirement for each segment, i.e., $\hat{b}(t+1)$. Based on the predicted average traffic, persistent bandwidth will be requested and pre-allocated in advance.

Assume that the actual bandwidth requirement in each 2s period follows the Gaussian distribution with mean $\overline{B}_n^{req}(t+1)$ and variance $0.2\overline{B}_n^{req}(t+1)$. In this paper, each $5s$ interval is divided into 20 timeslots. In each timeslot, dynamic bandwidth is then requested and allocated.

According to the requirement specified by 3GPP TR 36.885 [5], we adopt 100ms as the latency threshold of vehicular communication in the simulation, i.e., $T_n^r = 100ms$ and the signalling delay t_n^s is set to vary from $5ms$ to $25ms$.

We first study whether there is an optimal θ for the proposed scheme. Fig. 2 presents the optimal total bandwidth for SPRAS and DRAS, respectively, with $t^s = 15$ms. Note that the curve for the dynamic scheme is a level straight line as it is unrelated to θ. We can see that SPRAS provides smaller total bandwidth compared with DRAS for all of the θ values. Moreover, there exists an optimal θ ($\theta = 0.93$ in this example). The green curve is for the ideal situation (denoted as $B_{ideal}(t+1)$) where the accuracy of traffic prediction equals to 100%. When $\theta = 0.93$, the gap between $B_{op}(t+1)$ and $B_{ideal}(t+1)$ reaches the minimum value.

In Fig. 3, we show how the total bandwidth varies with the signalling latency t^s, where $\theta = 0.93$. As expected, the total bandwidth increases linearly with t^s for DRAS. Across the simulated range of t^s, the proposed SPRAS grows slowly with t^s increasing, and always achieves the lower total bandwidth.

In Fig. 4, we present the optimal values of θ for different signalling latencies for SPRAS. The optimal θ increases with t^s growing. The range of the optimal θ is $[0.894\ 0.952]$.

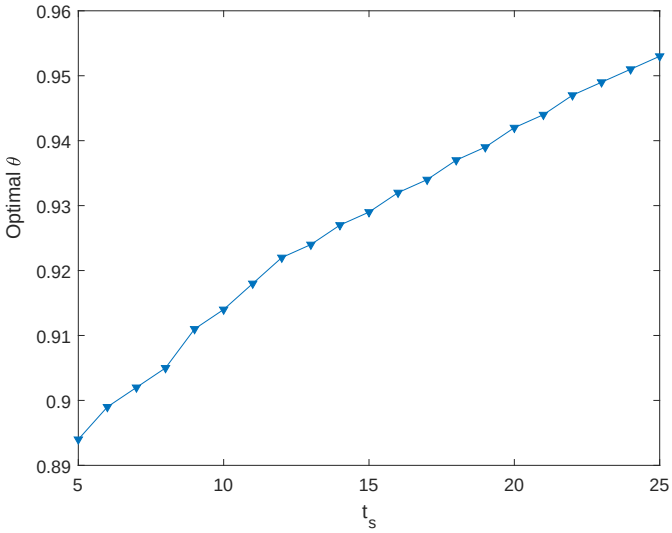

Fig. 4. Optimal θ with varying signalling latency t^s; $\theta = 0.93$

B. Simulation Results based on Real Traffic Flow Data

Assume there is a direct mapping between vehicle density and the network bandwidth requirements, we test the performance of the proposed scheme using real vehicle density data collected from a section of Interstate 80 (I-80) freeway located in Emeryville, California [16]. There are six traffic stations collecting the data, resembling six RSUs (i.e., $N = 6$). The record is for 10 days and the traffic flow data from the eighth day (8:00am-12am) is selected for testing. The time interval of the collected data is 30s, and hence $T_s = 30$s. Similarly, we divide T_s into 100 timeslots, and real-time bandwidth requirement in each timeslot is simulated using the same Gaussian distribution with that in Section IV-A.

Firstly, for the prediction correlation matrix $\mathbf{R}_t$, we fix the product of the number of its rows M and the number of

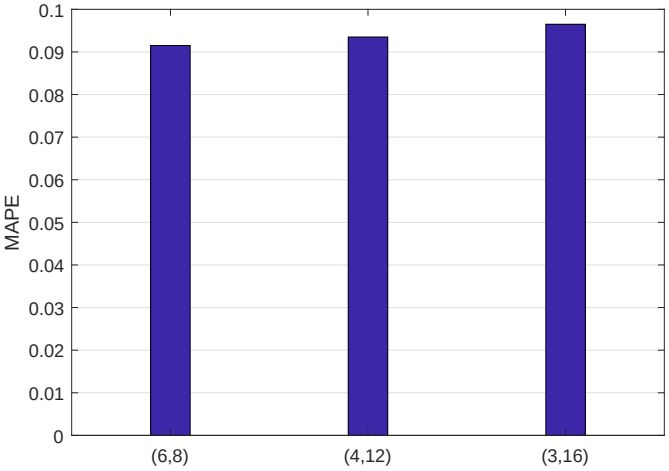

Fig. 5. Prediction results with different M and L values in segment 5.

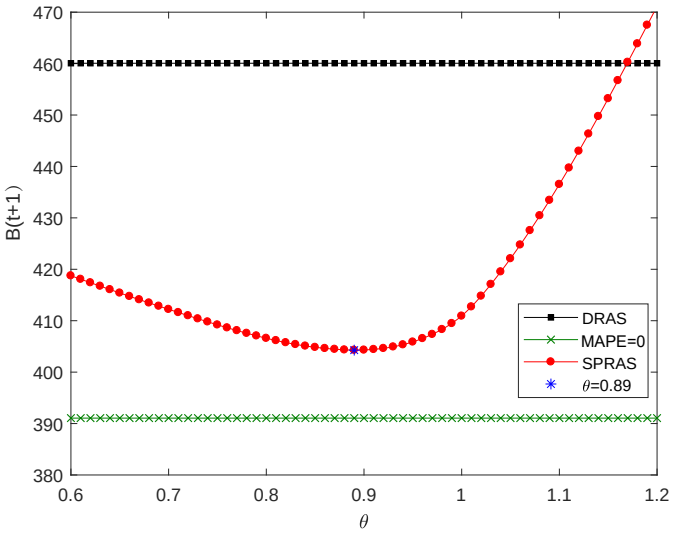

Fig. 6. Total bandwidth versus θ for DRAS and SPRAS using real vehicle density data; MAPE=9.15%; $t^s = 15$ms.

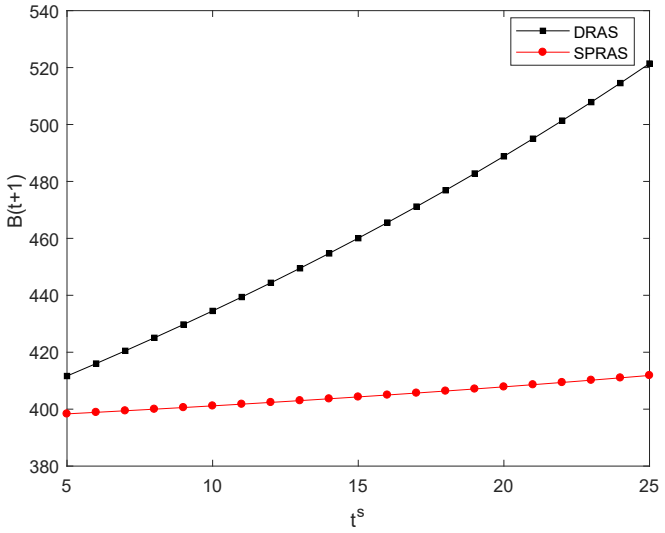

Fig. 7. Total bandwidth versus signalling latency t^s for DRAS and SPRAS with the real vehicle density data; $\theta = 0.89$.

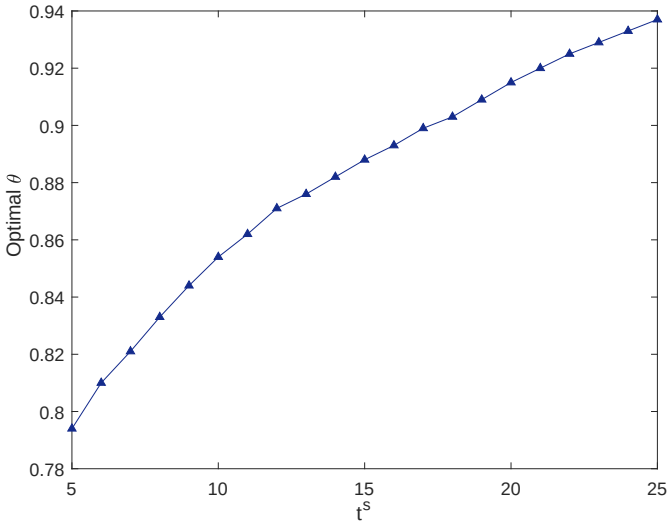

Fig. 8. Optimal θ versus signalling latency t^s.

columns L ($M * L = 48$) to be 48, and testing the prediction performance of the LMMSE predictor. In Fig. 5, we compare the MAPE values for different pairs of (M, L). The algorithm is found to achieve the best accuracy when $M = 6$ and $L = 8$.

Fig. 6, Fig. 7 and Fig. 8 show the simulation results for the real traffic data, in parallel to those for the artificially generated ones in Fig. 2, 3 and 4. Both of these results demonstrate the effectiveness of the proposed semi-persistent resource allocation scheme in minimizing the bandwidth usage under given time delay constraints.

V. Conclusions

We have presented a novel semi-persistent resource allocation scheme based on a two-tier heterogeneous cellular network for V2X communication. Using the predicted average network traffic for a small time interval of T_s seconds, the MBS pre-allocates persistent (bandwidth) resource to each RSU, which is responsible for V2X communications of the vehicles in its segment. During each timeslot in the T_s period, real-time dynamic bandwidth is further requested and allocated by MBS through RSU when needed. The total bandwidth is minimized by formulating an optimization problem under latency constraints. Simulation results demonstrate that our proposed scheme can achieve significant saving on the total allocated bandwidth when meeting the delay constraint, compared to conventional fully real-time resource allocation.

References

[1] X. Cheng, C. Chen, W. Zhang, and Y. Yang, "5g-enabled cooperative intelligent vehicular (5genciv) framework: When benz meets marconi," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 53–59, 2017.

[2] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions," *IEEE communications surveys & tutorials*, vol. 13, no. 4, pp. 584–616, 2011.

[3] B. T. Sharef, R. A. Alsaqour, and M. Ismail, "Vehicular communication ad hoc routing protocols: A survey," *Journal of network and computer applications*, vol. 40, pp. 363–396, 2014.

[4] S. Mumtaz, K. M. S. Huq, M. I. Ashraf, J. Rodriguez, V. Monteiro, and C. Politis, "Cognitive vehicular communication for 5g," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 109–117, 2015.

[5] 3rd Generation Partnership Project, "Study on LTE-based V2X services," Technical Specification Group Radio Access Network, Tech. Rep. 3GPP TR 36.885 V2.0.0, 2016.

[6] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for d2d-enabled vehicular communications," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 3186–3197, 2017.

[7] R. Zhang, X. Cheng, L. Yang, and B. Jiao, "Interference graph-based resource allocation (ingra) for d2d communications underlaying cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 8, pp. 3844–3850, 2015.

[8] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for lte v2v communication systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3850–3860, 2018.

[9] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE communications surveys & tutorials*, vol. 17, no. 4, pp. 2377–2396, 2015.

[10] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of dsrc and cellular network technologies for v2x communications: A survey," *IEEE transactions on vehicular technology*, vol. 65, no. 12, pp. 9457–9470, 2016.

[11] Z. He, J. Cao, and X. Liu, "Sdvn: enabling rapid network innovation for heterogeneous vehicular communication," *IEEE network*, vol. 30, no. 4, pp. 10–15, 2016.

[12] K. Zheng, L. Hou, H. Meng, Q. Zheng, N. Lu, and L. Lei, "Soft-defined heterogeneous vehicular network: architecture and challenges," *IEEE Network*, vol. 30, no. 4, pp. 72–80, 2016.

[13] J. Zhang, D. Smith, and Z. Chen, "Linear finite state markov chain predictor for channel prediction," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2012 IEEE 23rd International Symposium on*. IEEE, 2012, pp. 2085–2089.

[14] Y. Gao, G. He, and J. C. Hou, "On exploiting traffic predictability in active queue management," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 2002, pp. 1630–1639.

[15] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *International journal of forecasting*, vol. 8, no. 1, pp. 69–80, 1992.

[16] NGSIM, "Next generation simulation," *http://ngsim-community.org/*, 2010.