Robust Resource Allocation Model Using Edge Computing for Delay Sensitive Tasks in Vehicle to Infrastructure (V2I) Networks

Anna Kovalenko[1], Razin Farhan Hussain[2], Mohsen Amini Salehi[3] and Omid Semiari[4]
High Performance Cloud Computing (HPCC) Lab. School of Computing and Informatics, University of Louisiana at Lafayette Louisiana, USA. aok8889@louisiana.edu
High Performance Cloud Computing (HPCC) Lab. School of Computing and Informatics, University of Louisiana at Lafayette Louisiana, USA. razinfarhan.hussain1@louisiana.edu
High Performance Cloud Computing (HPCC) Lab. School of Computing and Informatics, University of Louisiana at Lafayette Louisiana, USA. amini@louisiana.edu
Department of Electrical and Computer Engineering, Georgia Southern University Georgia, USA. osemiari@georgiasouthern.edu

ABSTRACT

Development of autonomous vehicles is one of the most ambitious and promising projects in human history. Such vehicles require agile and reliable services to manage hazardous road situations. Vehicular Networks is the technology that can provide high-quality services for self-driving vehicles. A large percentage of service requests in these networks have an urgent nature (*e.g.*, disaster updates, hazard alerts, etc.) In other words, these requests are delay intolerant and require immediate service. Therefore, Vehicular Networks, and particularly, Vehicle-to-Infrastructure (V2I) systems must provide a consistent real-time response to autonomous vehicles. During increased traffic congestion or even natural disasters, it can be particularly tricky for V2I systems to maintain an optimal performance level. In such situations, a surge of requests arriving at a Base Station (a network edge device with computing capabilities) can drastically decrease V2I system response time. The consequences of even a millisecond delay for an urgent request can be dangerous, sometimes fatal. Hence, the goal of our research is to increase robustness (*i.e.*, ability to maintain optimal performance) of the V2I systems. To achieve this goal, we offer a resource allocation model that can load balance (*i.e.*, dynamically utilize resources from neighboring Base Stations), when the system is oversubscribed (experiencing an unusually dense service requests arrival). We propose an allocation algorithm based on a calculated probability of the arriving request to be served in time on several neighboring Base Stations. We introduce a Load Balancer component which assigns the request to the Base Station with a maximum precomputed probability. After all, we evaluate our model under various oversubscription levels and urgent requests percentages. Simulation results demonstrate that the proposed model decreases overall service miss rate by up to 20 % and urgent requests miss rate by up to 50 %.

Keywords: Vehicular Networks, V2I, Edge Computing, High Performance.

1. INTRODUCTION

Recent advancements in communication and computation technologies have stimulated a rapid development of vehicular networks. Federal Communications Commission (FCC) has reserved 5.850 to 5.925 GHz frequency band for Vehicle-to-Everything (V2X) communications [Ali and Chan, 2011]. Vehicle-to-Infrastructure (V2I) communications is one prominent form of V2X that draws the majority of work to itself. In V2I, infrastructure refers to all edge and core technologies that facilitate communications and computations for vehicular requests.

As shown in Figure 1, autonomous vehicles send their service requests (tasks) to Base Stations while operating on the road. A Base Station is capable of communicating with vehicles and processing vehicular tasks [Bok et al., 2016]. Upon the completion of the processing, the results sent back to the requesting vehicle. Examples of such vehicular tasks can be a Wrong Way Driver warning [Bonte and Owen, 2013], Cooperative Forward Collision warning [ElBatt et al., 2006], and Lane Change warning [Bonte and Owen, 2013]. This type of tasks can only tolerate a short end-to-end delay [Ali and Chan, 2011]. For such delay-sensitive requests, there is no value in executing them after a tolerable delay.

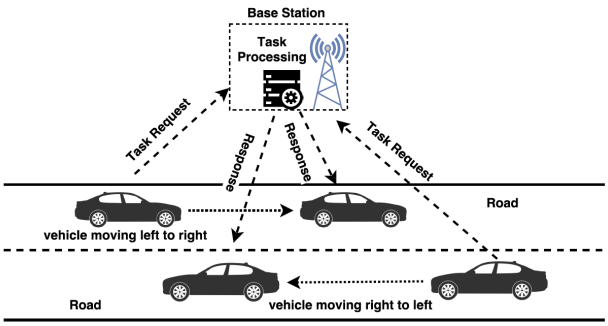

Figure 1: A Vehicle to Infrastructure (V2I) scenario where vehicles send requests to a Base Station and receive the response. A Base Station is a roadside unit with communicational and computational abilities.

Significant problems arise during road emergencies (*e.g.*, road accidents and disasters) when a rapid increase in service requests to Base Stations significantly affects the tasks' service time. In fact, in this situation, Base Station resources become oversubscribed, and it cannot provide enough computational power for all the arriving tasks to meet their deadlines. Accordingly, our goal, in this research, was to design the V2I system to be robust against uncertain task arrival. In the literature, *robustness* is defined as the degree to which a system can maintain a certain level of performance even with given uncertainties [Ali et al., 2004, Smith et al., 2009, Canon and Jeannot, 2010]. In our research, we evaluate robustness of the V2I system according to the number of tasks that can meet their deadlines. The main question we try to answer is how to allocate arriving tasks among the Base Stations so that the system stays robust? Or, in other words, we try to find a way to maximize the number of tasks meeting their deadlines.

Previous research works either discard these uncertainties [Bok et al., 2016] or focus on the uncertainty introduced by communication [Ali and Chan, 2011]. Alternatively, to assure robustness of the V2I system, we propose a probabilistic resource allocation model that copes with uncertainties introduced by both communication and computation. Our proposed model is aware of the connectivity amongst Base Stations (*i.e.*, edge nodes) and their heterogeneity. In the face of oversubscription, we devise a Load Balancer at the Base Station level that can

leverage the computational capabilities of other Base Stations to improve robustness of the V2I system.

According to our model, when the task arrives to a Base Station it enters the Load Balancer (Figure 2). The Load Balancer works in an immediate mode to allocate arriving tasks to the Base Stations. It can allocate the task to the receiving Base Station or to the one-hop distance neighboring Base Station. When the task is allocated, it enters the batch queue of the Base Station to be scheduled for processing.

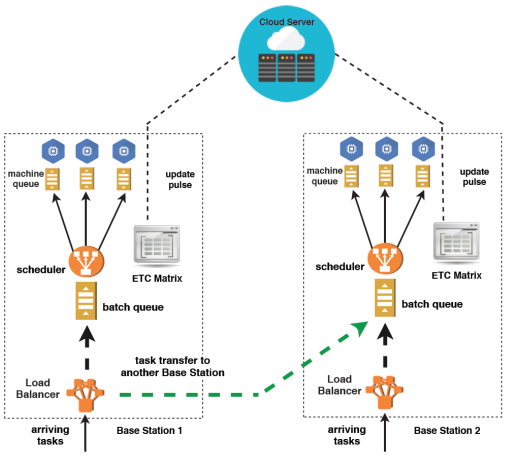

Figure 2: A proposed model where the task gets efficiently allocated by the Load Balancer between receiving and neighboring Base Stations.

Every Base Station stores two matrices (ETC Matrix component in Figure 2) to enable probability computation proposed in our model. One matrix is the Estimated Task Completion time (ETC) matrix [Ali et al., 2000]. Another one is the Estimated Task Transfer time (ETT) matrix. These two matrices contain estimated task completion and task transfer time normal distributions (X$\sim \mathcal{N}(\mu, \sigma^2)$) respectively. These distributions are based on historical execution and transfer times of different task types (delay tolerant and intolerant). Matrices are updated periodically through the cloud.

For each task t_i of task type "i," Load Balancer calculates the probability (P_i^j) of this task to meet its deadline δ_i across the receiving and neighboring Base Stations. For the receiving Base Station "j", the probability can be defined as $P_i^j(\gamma_i^j < \delta_i) = P_i^j(Z < z)$ where "z" is $(\delta_i - \mu_i^j) / \sigma_i^j$. We standardize the distribution with $\mu_i = 0$ and $\sigma_i = 1$. For all of the neighboring Base Stations, to calculate the probability, Load Balancer convolves the ETC distribution with respective ETT distribution. The convolution is necessary to account for the transfer time to a neighboring Base Station. The resulting distribution (W$\sim \mathcal{N}(\mu, \sigma^2)$) is used to calculate the probability of the task meeting the deadline in a specific Base Station. If a neighboring Base Station is "k" and task type is "i" then the probability can be defined as "P_i^k" where $z = (\delta_i - \mu_i^k) / \sigma_i^k$. When the probability of the received task in all of the Base Stations (receiving and neighboring) is calculated, the task gets allocated to the Base Station that offers the highest probability. When the task's probability to meet its deadline is zero (0), the task is dropped (it will not enter any batch queue for scheduling). Task dropping procedure during the oversubscription situation implicitly increases the probability of the other tasks to meet their deadlines.

The results of our research prove that the proposed model offers a better, more robust allocation algorithm. We evaluate our model using EdgeCloudSim simulation [Sonmez et al., 2017]. Our resource allocation model is tested against the Baseline model. The Baseline model

always allocates arriving task to a receiving Base Station. Figure 3 represents simulation results of overall system performance for medium and high system oversubscription levels. Our model provides noticeably better results when the number of vehicles is less than or equal to 100. Otherwise, it's performance decreases as well as the Baseline's performance. Regardless, our system consistently performs 2-5 % better than the Baseline.

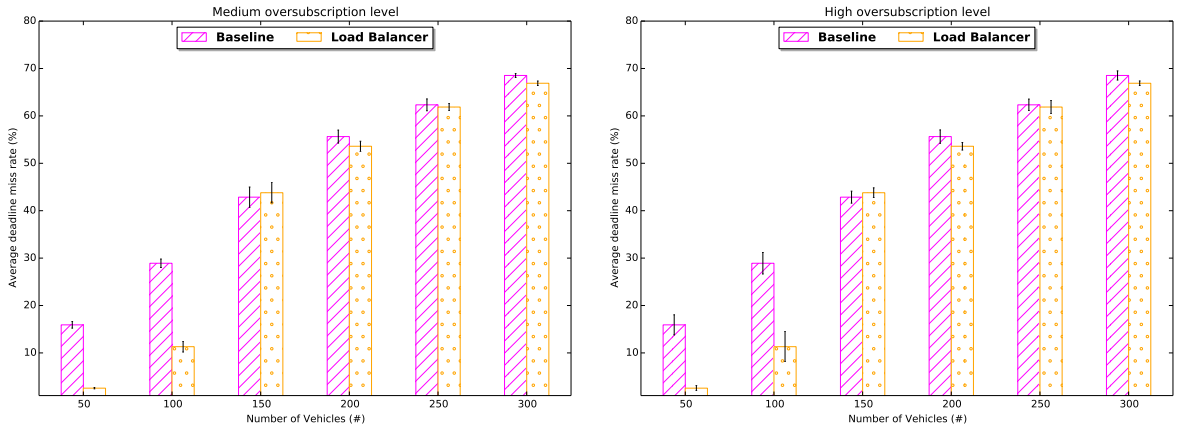

Figure 3: Load Balancer overall service miss rate compared to the Baseline.

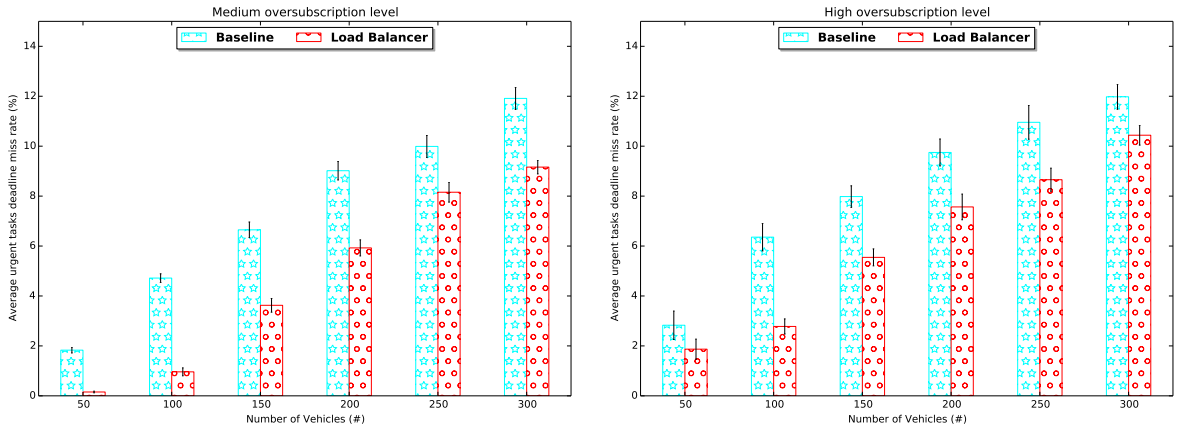

Figure 4: Load Balancer urgent service miss rate compared to the Baseline.

Nevertheless, the most crucial improvement our system presented in Figure 4. Our research aimed to provide a resource allocation model that would be robust against computational and communicational uncertainties, especially for the service requests that have an urgent nature. Figure 4 shows the deadline miss rates for urgent tasks both for our model and the Baseline. The left caption represents the performance for the medium level of oversubscription and the right one for the high. We can notice, that our proposed model consistently performs better than the Baseline. When the number of vehicles is low, Load Balancer allows for up to 80 % improvement. With a more significant amount of vehicles up to 50 % performance improvement is present.

Finally, I can conclude our research results to be successful. We proposed a model that encompasses the uncertainties exist in communication and computation. We developed a load balancing heuristic that increases the robustness of the V2I system. The analysis of the results confirms the success and allows to continue expanding our model, and possibly introducing a real-world implementation in the future.

References

[Ali and Chan, 2011] Ali, G. M. N. and Chan, E. (2011). Co-operative data access in multiple Road Side Units (RSUs)-based Vehicular Ad Hoc Networks (VANETs). In *Telecommunication Networks and Applications Conference (ATNAC), 2011 Australasian*, pages 1–6. IEEE.

[Ali et al., 2004] Ali, S., Maciejewski, A. A., Siegel, H. J., and Kim, J.-K. (2004). Measuring the Robustness of a Resource Allocation. *IEEE Transactions on Parallel and Distributed Systems*, 15(7):630–641.

[Ali et al., 2000] Ali, S., Siegel, H. J., Maheswaran, M., Hensgen, D., and Ali, S. (2000). Representing Task and Machine Heterogeneities for Heterogeneous Computing Systems. *Tamkang Journal of Science and Engineering*, 3(3):195–207.

[Bok et al., 2016] Bok, K., Hong, S., Lim, J., and Yoo, J. (2016). A multiple RSU scheduling for V2I-based data services. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 163–168. IEEE.

[Bonte and Owen, 2013] Bonte, D. and Owen, G. (2013). Allied Business Intelligenc Inc. `https://www.abiresearch.com/market-research/product/1016485-v2v-and-v2i-applications-and-use-cases/`.

[Canon and Jeannot, 2010] Canon, L.-C. and Jeannot, E. (2010). Evaluation and Optimization of the Robustness of DAG Schedules in Heterogeneous Environments. *IEEE Transactions on Parallel and Distributed Systems*, 21(4):532–546.

[ElBatt et al., 2006] ElBatt, T., Goel, S. K., Holland, G., Krishnan, H., and Parikh, J. (2006). Cooperative Collision Warning Using Dedicated Short Range Wireless Communications. In *Proceedings of the 3rd International Workshop on Vehicular Ad Hoc Networks*, VANET '06, pages 1–9, New York, NY, USA. ACM.

[Smith et al., 2009] Smith, J., Shestak, V., Siegel, H. J., Price, S., Teklits, L., and Sugavanam, P. (2009). Robust resource allocation in a cluster based imaging system. *Parallel Computing*, 35(7):389–400.

[Sonmez et al., 2017] Sonmez, C., Ozgovde, A., and Ersoy, C. (2017). EdgeCloudsim: An environment for performance evaluation of Edge Computing systems. *2017 Second International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 39–44.