Chasing Clouds with Donkeycar: Holistic Exploration of Edge and Cloud Inferencing Trade-Offs in E2E Self-Driving Cars

Kyle Zheng
zhengkyle811@gmail.om
Modesto Junior College
Modesto, California, USA

Alicia Esquivel (Advisor)
ace6qv@mail.missouri.edu
University of Missouri
Columbia, Missouri, USA

Katarzyna Keahey (Advisor)
keahey@mcs.anl.gov
Argonne National Laboratory
Chicago, Illinois, USA

ABSTRACT

In autonomous driving, there is a growing need for computational resources due to the burden of inference models. One possibility is offloading inference to the cloud [1]. However, this begs the question of whether cloud-aided real-time inference is viable given the latency between the car and the data center. To investigate this problem, we present a modular Cloud-Aided Real-time Inferencing Framework that integrates with Donkeycar, a self-driving car platform, and strategically distributes computational load across cloud and edge [2]. The comparisons of performance metrics between cloud-aided and edge-only models provide a demonstration of advantages offered by a Cloud-Aided framework, particularly in regards to RNN performance which changed from 0% autonomy to 90% autonomy, outperforming convolutional neural networks when the domain shifted. We also comprehensively analyze the strengths, limitations, and potential of leveraging cloud resources in real-time edge environments with autonomy scores and latency trade-offs [3].

ACM Reference Format:
Kyle Zheng, Alicia Esquivel (Advisor), and Katarzyna Keahey (Advisor). 2023. Chasing Clouds with Donkeycar: Holistic Exploration of Edge and Cloud Inferencing Trade-Offs in E2E Self-Driving Cars. In *Proceedings of SC 2023 (SC).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Edge inferencing, while offering increased reliability due to its independence from cloud-based systems, is inherently constrained by the available resources (Figure 1). This constraint can create a bottleneck in the actuation of autonomous vehicles. As the cloud infrastructure continues to evolve at an accelerated pace, with numerous enterprises advancing the development of cloud-connected vehicular technology, a cloud-assisted, modular framework could provide valuable insights. Our proposed framework is characterized by its modularity and the incorporation of gRPC bidirectional streaming, optimized through the use of a NVIDIA Triton Inference Server hosted on a Chameleon Cloud node (Figure 2). These features enhance the framework's adaptability, making it a robust platform for future investigations into cloud-based solutions for autonomous vehicles.

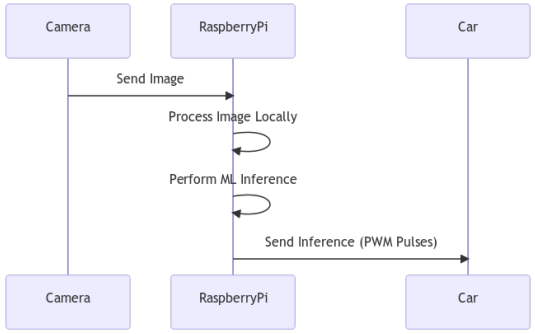

Figure 1: Figure 1: Data Flow in During Pure Edge Inferencing

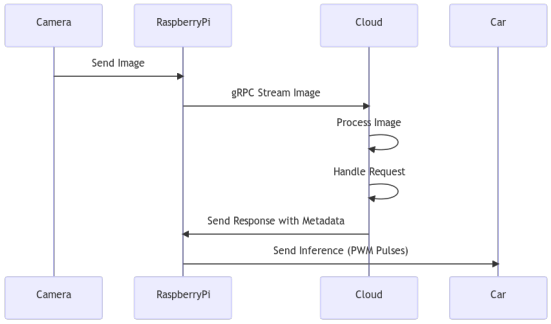

Figure 2: Figure 2: Data Flow with Cloud-Aided Framework

2 METHODOLOGY

In our study, we use a scaled car navigating a track filled with a changing array of obstacles. This setup serves as the training environment for the inference models we deploy on both the edge and the cloud.

For edge inferencing, we run the models on a Raspberry Pi 4 (RPi4). These models produce a steering angle based on pixel mapping. During this process, we keep track of the resource usage and the speed of inference, measured in frames inferred per second (Figure 3).

To create a Cloud-Aided Real-time Inferencing Framework, we launch an NVIDIA Triton Inference Server utilizing a Docker Container. This server holds our models. At the same time, we set up a

SC, November 2023, Denver, CO

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

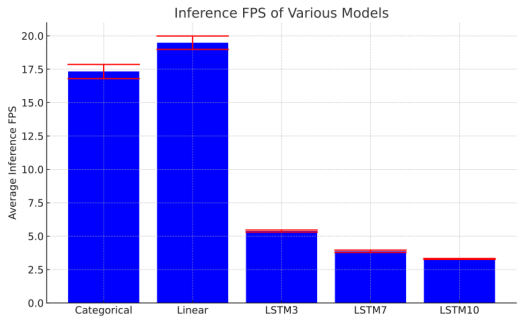

Figure 3: Figure 3: Inference FPS at the Edge

client on the car that intercepts the data flow from the Donkeycar platform. This arrangement allows images to be redirected via a gRPC bidirectional asynchronous stream to the server at port 8001.

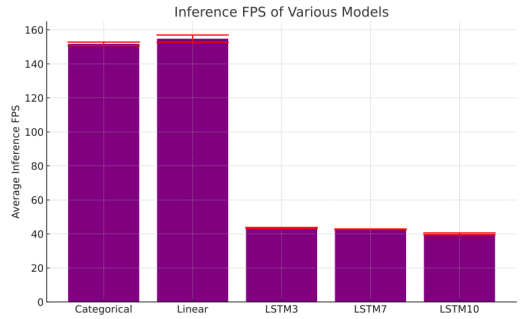

Figure 4: Figure 4: Inference FPS with Cloud Support

The server processes the images and sends back the outputs. The client processes these outputs and records the latency. We assess performance based on several factors: the average speed of the car, the computational resources used, the latency, the speed of inference, and the overall autonomy score calculated with

$$\text{autonomy} = \left(1 - \frac{\text{interventions} \times 4.5\,\text{seconds}}{\text{total time}}\right) \times 100\%$$

This multi-faceted approach allows us to thoroughly evaluate the system's performance, which has much faster inference time and does not bottleneck the system (Figure 4). The Long Short-Term memory of sequence length 7 in the cloud is able to process 42 frames per second, which is about 11 times faster than the model optimally runs on the edge, eliminating the bottleneck of the vehicle loop.

3 EVALUATION

Our modifications in offloading inference to the cloud resulted in substantial improvements in the vehicle's responsiveness. Comparisons of the autonomy scores between edge-only and cloud-aided revealed a significant performance advantage for the latter. This was particularly evident in the case of recurrent network models (Table 2).

Under the constraints of edge computing, slow inference speeds made it challenging to accurately gauge RNN capabilities (Table 1). However, when sufficient resources were provided through cloud offloading, the RNNs demonstrated performance on par with, or even superior to, traditional CNNs.

Table 1: Autonomy Scores of Edge Models

Speed	Linear	Categorical	LSTM3	LSTM7	LSTM10
0.25m/s	80%	62.5%	30%	0%	0%
0.20m/s	50%	80%	0%	0%	0%
0.27m/s	30%	0%	0%	0%	0%

Table 2: Autonomy Scores of Cloud Models

Speed	Linear	Categorical	LSTM3	LSTM7	LSTM10
0.25m/s	93.75%	93.75%	95%	90%	90%
0.20m/s	90%	93.75%	95%	80%	63.8%
0.27m/s	80%	78.8%	81.3%	80%	30%

The cloud offloading approach effectively alleviated the computational burden of intensive inference calculations on the vehicle loop, which is capped at 50 ms per loop. This resulted in an unhindered vehicle loop improving overall system performance.

In comparison to the edge-only framework, the cloud-aided framework demonstrates a significant performance boost, particularly with the use of RNNs.

4 CONCLUSION AND FUTURE WORK

This paper provides a comprehensive examination of the possibility of cloud computing in end-to-end autonomous vehicles and underscores the significant potential of integrating autonomous driving with cloud resources. Our findings demonstrate a marked improvement in performance when proximal cloud resources are optimally utilized. Specifically, Long Short-Term Memory (LSTM) networks of various sequence lengths exhibited substantial enhancements due to the availability of sufficient resources for timely analysis. Furthermore, the modular framework allows for additional capabilities to be built on top.

Looking ahead, our future work will focus on integrating Liquid Time Constant Neural Network models, a type of RNN, into the Donkeycar platform [4]. This will enhance domain adaptability and allow us to fully leverage cloud capabilities.

5 ACKNOWLEDGEMENTS

Many thanks to Alicia Esquivel and Michael Sherman for their help. This work was supported by Chameleon Cloud and NSF REU 1757964 BigDataX: From theory to practice in Big Data computing at eXtreme scales.

REFERENCES

[1] Y. Maalej and E. Balti, "Integration of vehicular clouds and autonomous driving: Survey and future perspectives," 2022.

[2] M. G. Bechtel, E. McEllhiney, and H. Yun, "Deeppicar: A low-cost deep neural network-based autonomous car," *CoRR*, vol. abs/1712.08644, 2017. [Online]. Available: http://arxiv.org/abs/1712.08644

[3] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016. [Online]. Available: http://arxiv.org/abs/1604.07316

[4] R. M. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, "Liquid time-constant networks," *CoRR*, vol. abs/2006.04439, 2020. [Online]. Available: https://arxiv.org/abs/2006.04439