One Fish, Two Fish: Choosing Optimal Edge Topologies for Real-Time Autonomous Fish Surveys

Jonathan Tsen
dudujonathantsen@gmail.com
FATEC Shunji Nishimura
Pompéia, São Paulo, Brazil

Jason Anderson (advisor)
jasonanderson@uchicago.edu
The University of Chicago
Chicago, Illinois, USA

Kate Keahey (advisor)
keahey@anl.gov
Argonne National Laboratory
Lemont, Illinois, USA

Leonardo Bobadilla (advisor)
bobadilla@cs.fiu.edu
Florida International University
Miami, Florida, USA

ABSTRACT

The oceans make up more than seventy percent of the Earth's surface and the marine ecosystems are central to many global challenges. However, monitoring the environment and ensuring the sustainable use of marine resources presents unique challenges which span socio-economic balances, technology development, and dynamic ecological processes that span spatial and temporal scales. Technological advances in computer vision technology, artificial intelligence, and cloud computing allow marine scientists to collect and process data in greater volumes than ever before and offer promise for solutions that may improve the efficiency of real-time data acquisition and analysis to address complex biological and ecological questions.

KEYWORDS

replicability, open testbeds, notebooks, YOLO, IoT, CHI@Edge, GPU

1 INTRODUCTION

The Institute of Environment at Florida International University (FIU) has developed an interdisciplinary autonomous vehicle (AV) initiative consisting of several instrumented vessels and buoys and aims to use the AVs to characterize the health of the ecosystem with particular focus on quantifying the spatial and temporal patterns of fish in Biscayne Bay, Florida. We aim to develop indices of fish abundance and map distributions to important habitats in the bay in real-time and leverage the capabilities of the AVs to improve survey efficiency to better understand the Biscayne Bay ecosystem. To accomplish this goal, we need to ask two questions.

1. What is the best strategy for collecting and analyzing data from autonomous vehicles and can we leverage cloud computing resources to improve access to data products in real-time?

2. How does the resolution of video data and quality of network connection influence which strategy is best?

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

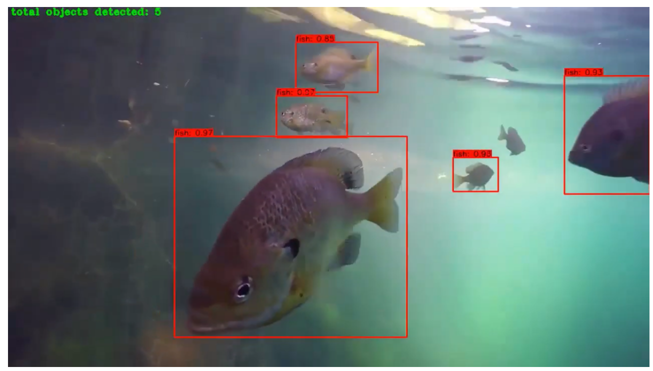

Figure 1: YoloV4 results on Bare Metal.

In this work, we compare two configurations of the AV: one with an intelligent edge device capable of performing local analysis, and one with a "dumb" edge device that delegates analysis to cloud computing resources.

2 APPROACH

For this experiment, we used Chameleon's CHI@Edge platform, to provide an edge device of the same type that will be used on the AV, as well as provision cloud resources for our experiment. We were able to connect the devices to GPU nodes on Chameleon's Cloud [5] and thus answer the questions relevant to the introduction. The experiments were implemented via a Jupyter notebook that provisioned the resources and ran the experiments.

For the fish counting application, analysis involves counting the number of fish captured by a camera (shown in Figure 1). Our implementation uses the YOLO object detection model [1], which we updated to output the number of objects (fish) detected in a scene. We evaluate two versions of YOLO: YOLOv4 (deployed on the cloud resources) and Tiny-YOLOv4 [4] (used on the edge device), a smaller version of the YOLO model suitable for deployment on resource-constrained hardware.

2.1 Response time in different configurations.

We measured response time in each configuration, defined as how long from the time of video frame capture until the result can be

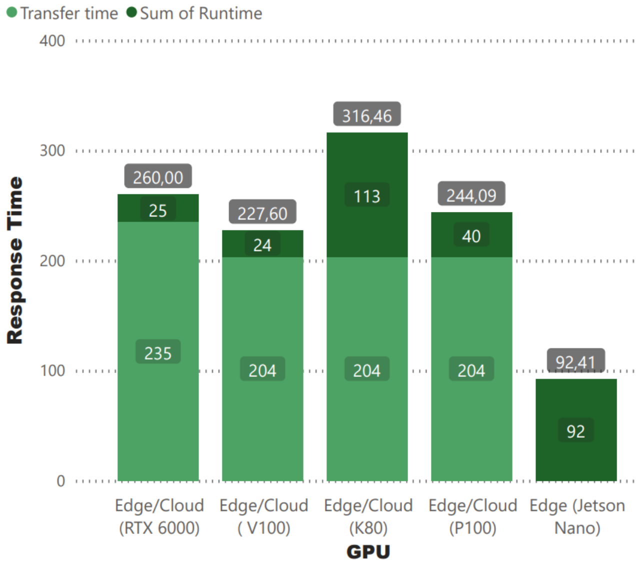

Figure 2: Response time in different configurations.

received from the AI model. We found that in the cloud configuration, runtime defined as the time it takes for the model to run in a frame was dominated by transfer time [2] that is the time it takes for video to be uploaded and returned from IoT to Chameleon Cloud (for example, RTX 6000 could evaluate at 25 ms but it took 235 ms to transfer the data, for a total response time of 260 ms) (shown in Figure 2).

2.2 Price - Performance of GPUs

Besides measuring response time, we also measured the value/ performance ratio by multiplying the GPU cost by the FPS (frames per second), which is the reciprocal of response time. The prices of each GPU in dollars used to calculate the (price / performance) was collected in August 2021, where FPS is the reciprocal of response time: how many frames can be processed in a given second (shown in Figure 3) [3]. The Jetson Nano is almost 19 times cheaper per FPS than Edge/Cloud (V100). The cheaper RTX 6000 outperforms the more expensive V100 with this perspective.

GPU	Location	Price
RTX 6000	CHI@UC	$4,000.00
Jetson Nano	CHI@Edge	$99.00
K80	CHI@TACC	$398.00
V100	CHI@TACC	$7,179.00
P100	CHI@TACC	$2,429.00

3 CONCLUSIONS

A system was developed to monitor fish populations in the regional, national, and even global ecosystems. To optimize the system, tests were performed to evaluate cloud vs local GPUs. From the tests we reached the following conclusions:

1. When you factor in the transfer time, the edge has a higher performance.

2. The Jetson Nano also has the best price/performance ratio.

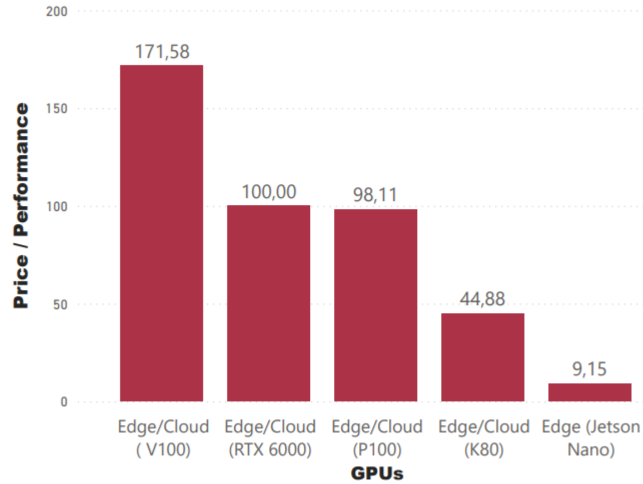

Figure 3: Price-Performance of GPU

3. The network condition has to significantly improve for the edge cloud to outperform the Jetson Nano.

4. In terms of Response time and Value-Performance, we see that the RTX 6000 has a much better cost-benefit ratio than the V100.

5. This experiment is reproducible through a Jupyter Notebook using the Python language in the Chameleon Cloud.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] PTS Borman, RHN Tijssen, C Bos, CTW Moonen, BW Raaymakers, and M Glitzner. Characterization of imaging latency for real-time mri-guided radiotherapy. *Physics in Medicine & Biology*, 63(15):155023, 2018.

[3] Abhranil Das. Process time comparison between gpu and cpu. *Hamburger Sternwarte*, 2011.

[4] Zicong Jiang, Liquan Zhao, Shuaiyang Li, and Yanfei Jia. Real-time object detection method based on improved yolov4-tiny. *arXiv preprint arXiv:2011.04244*, 2020.

[5] Wuyang Zhang, Sugang Li, Luyang Liu, Zhenhua Jia, Yanyong Zhang, and Dipankar Raychaudhuri. Hetero-edge: Orchestration of real-time vision applications on heterogeneous edge clouds. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1270–1278. IEEE, 2019.