Real-time, robust, and reliable (R^3) machine learning across wireless networks

By Akshar Vedantham, Kirthana Ram, Varun Kota

Advisors - Prof. Anand Sarwate and Prof. Waheed Bajwa

Project Introduction

- Phones, cars, and other devices will all want to start using ML/AI applications
- Leverage the cloud to help them with this
- Issue: Latency, security

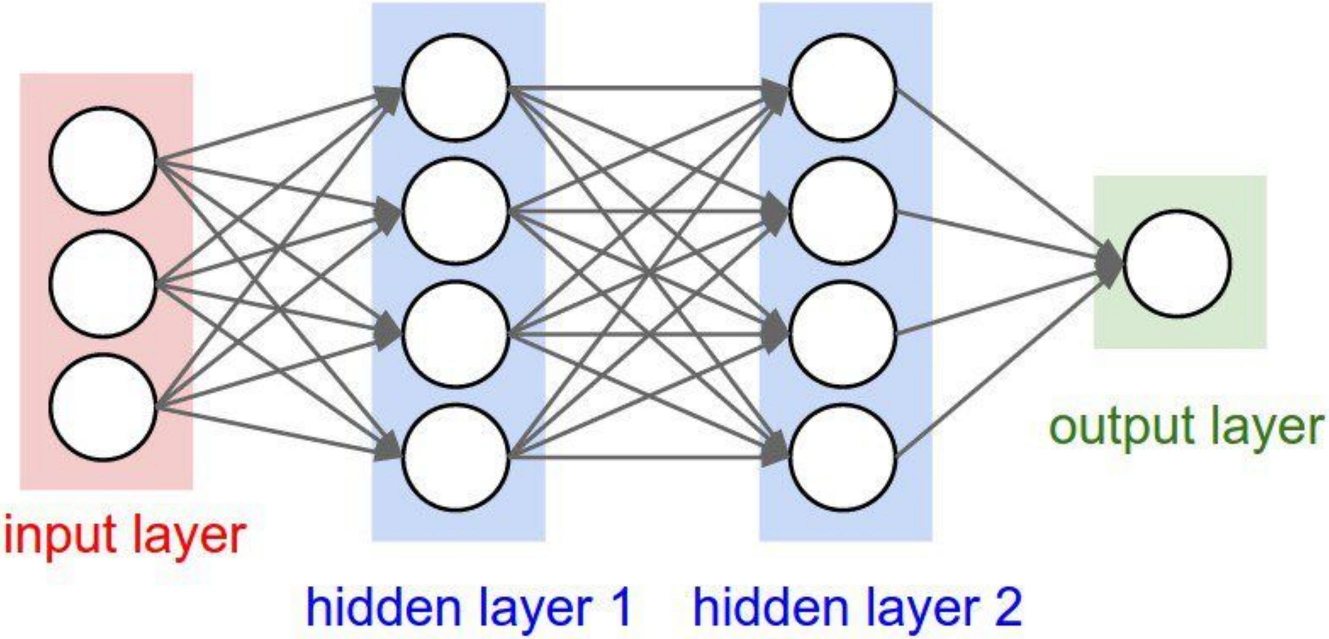
Example Scenario: Security System

Problem: **When** should phones offload to the cloud and ask for help?

1.) AI Model
2.) Strategies to offload
3.) Network conditions
4.) Evaluate

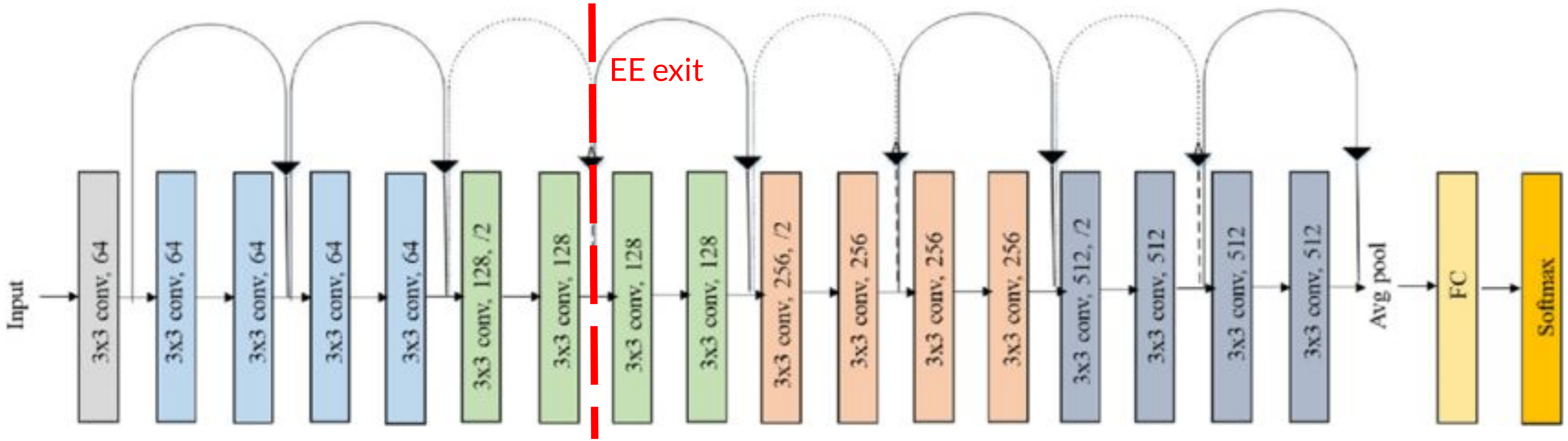
The AI Model

What is a Neural Network?

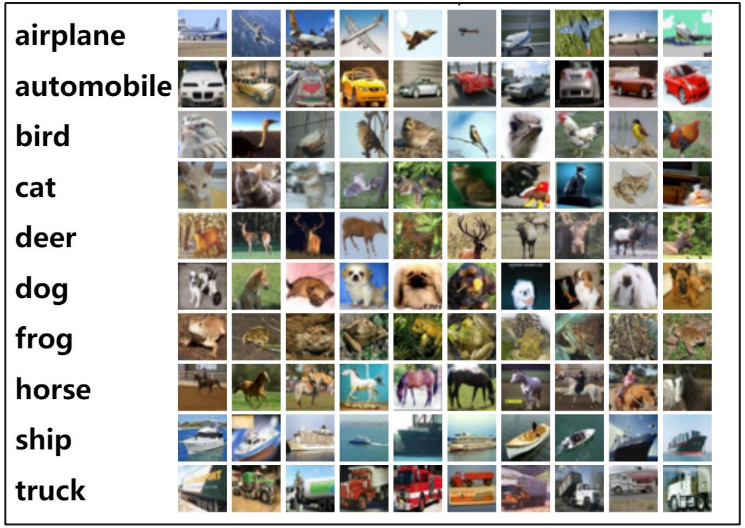


Choice of Model - ResNet18

- Popular model available
- Accuracy of 92.42%



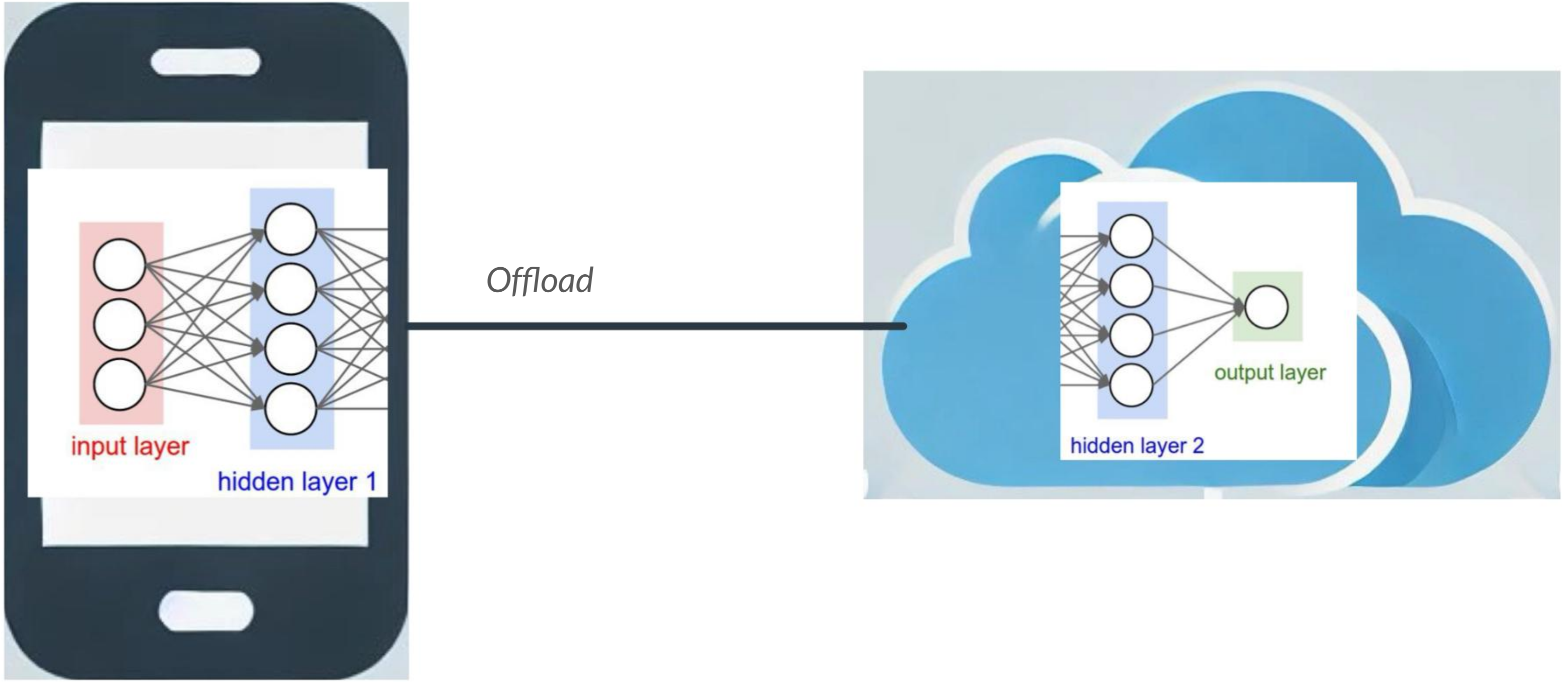
CIFAR-10 Dataset



- Commonly used for machine learning models
- 60,000 images

Offloading Strategies

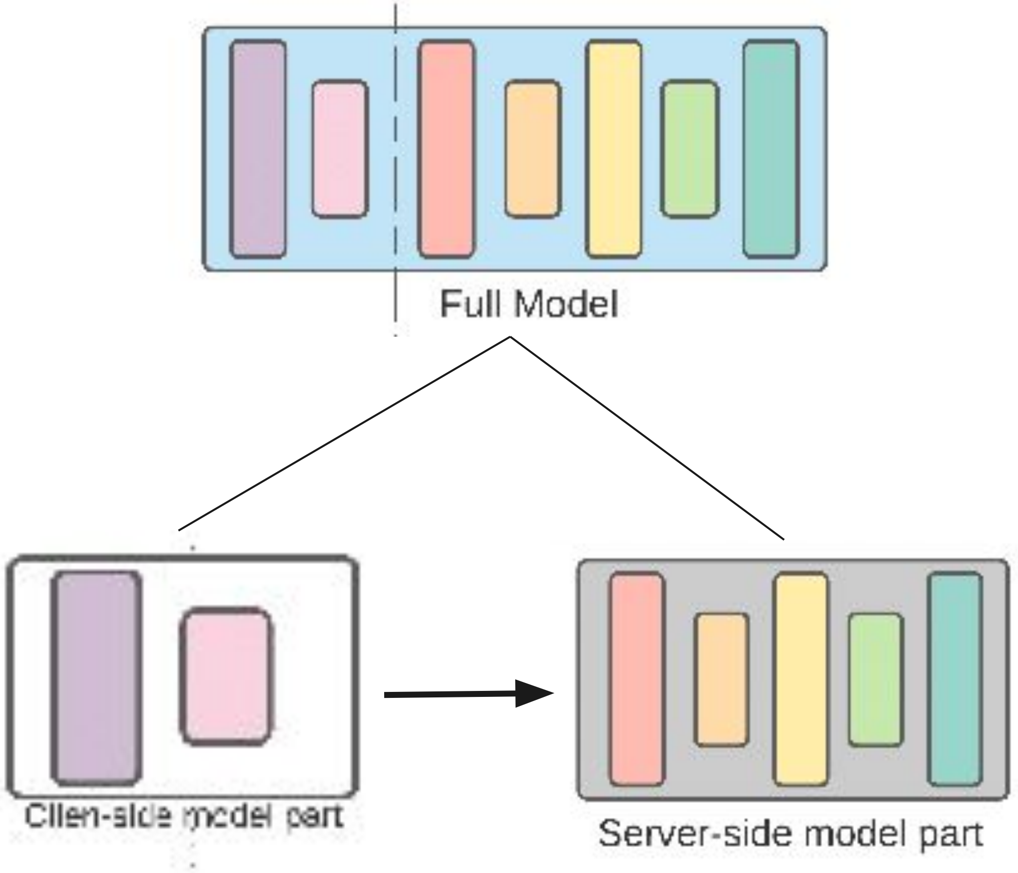
Offloading Explained



Offload

Early Exiting

- Take full model, split into a small model and larger one
- Client device gets smaller one, tries to make prediction



Networking Conditions

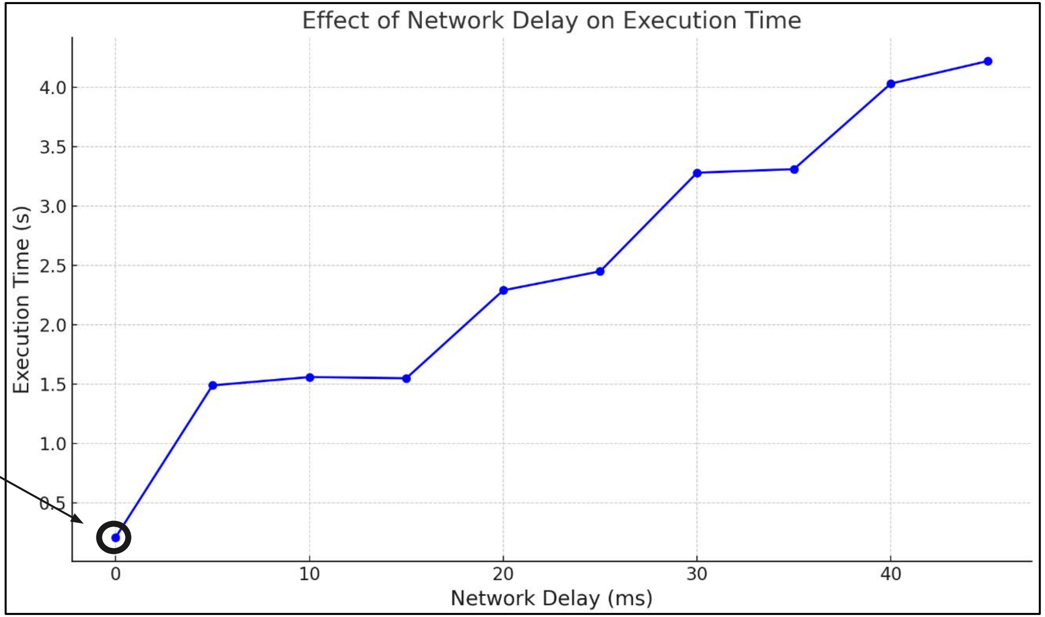
Speed

- Python’s Fast API to communicate with server and device
- Sending tensors as buffer streams instead of JSON

Processing 500 images/second over an ethernet connection

LOONEY TUNES and all related characters and elements © & ™ Warner Bros. Entertainment Inc. (s21)

Effects of Network Speed

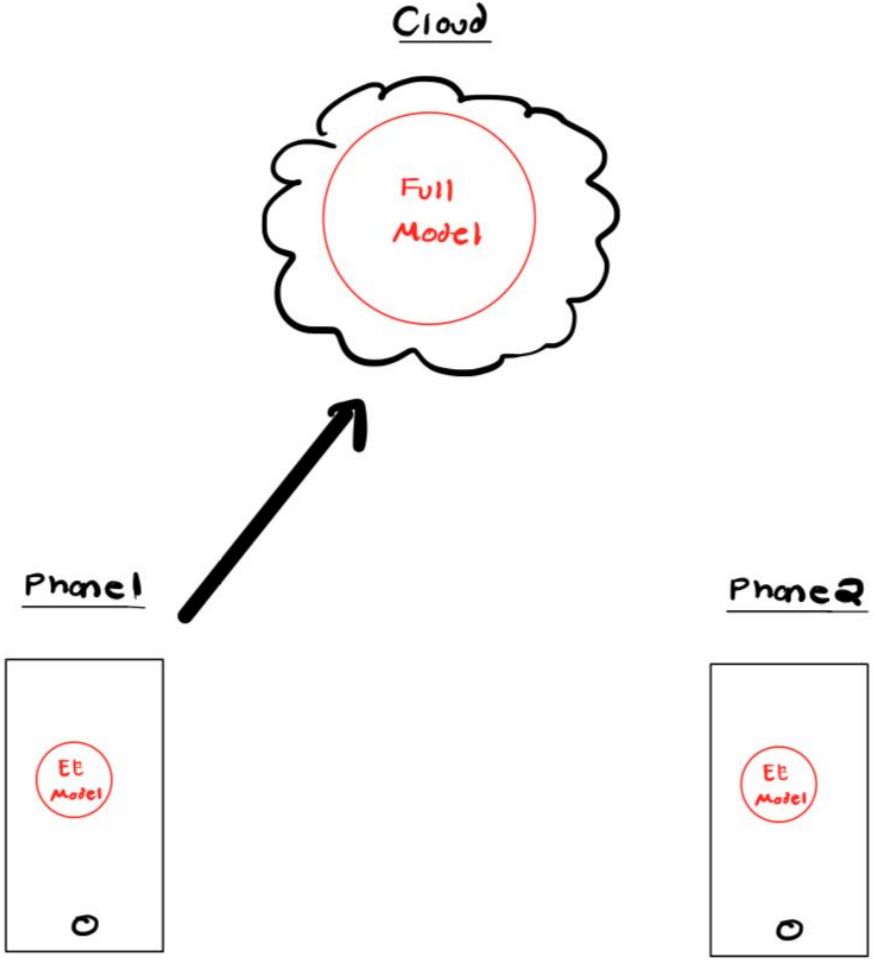


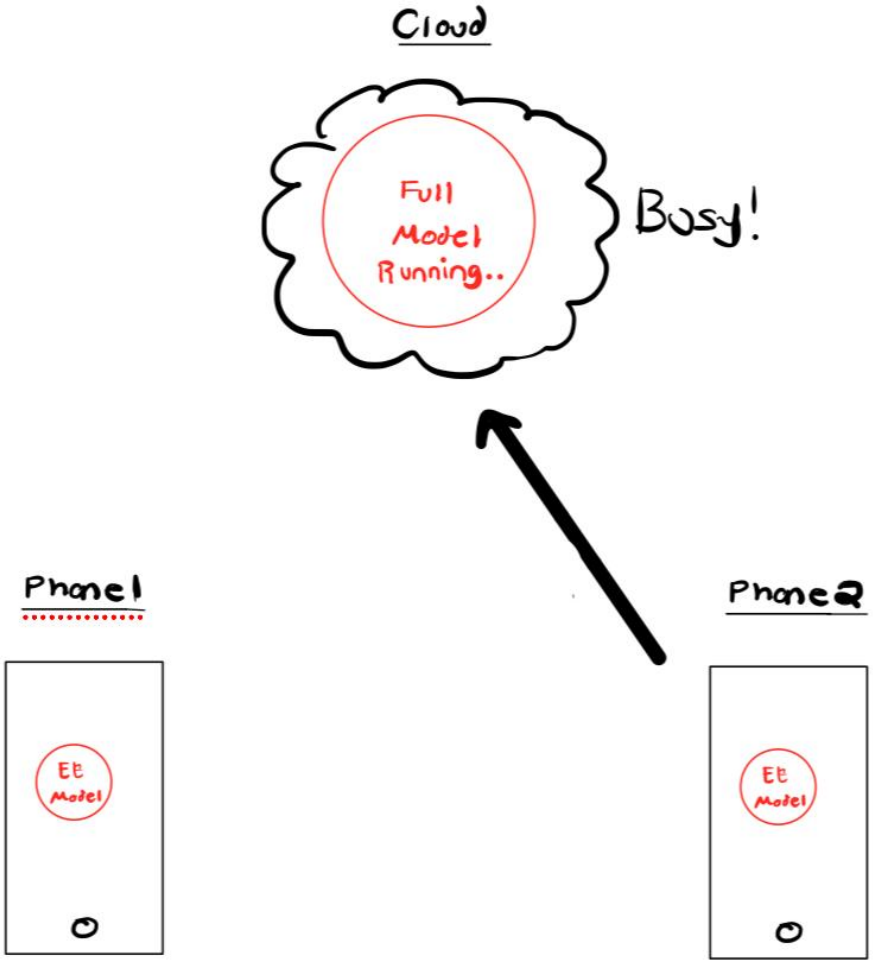
Ethernet
(Control)

Evaluation

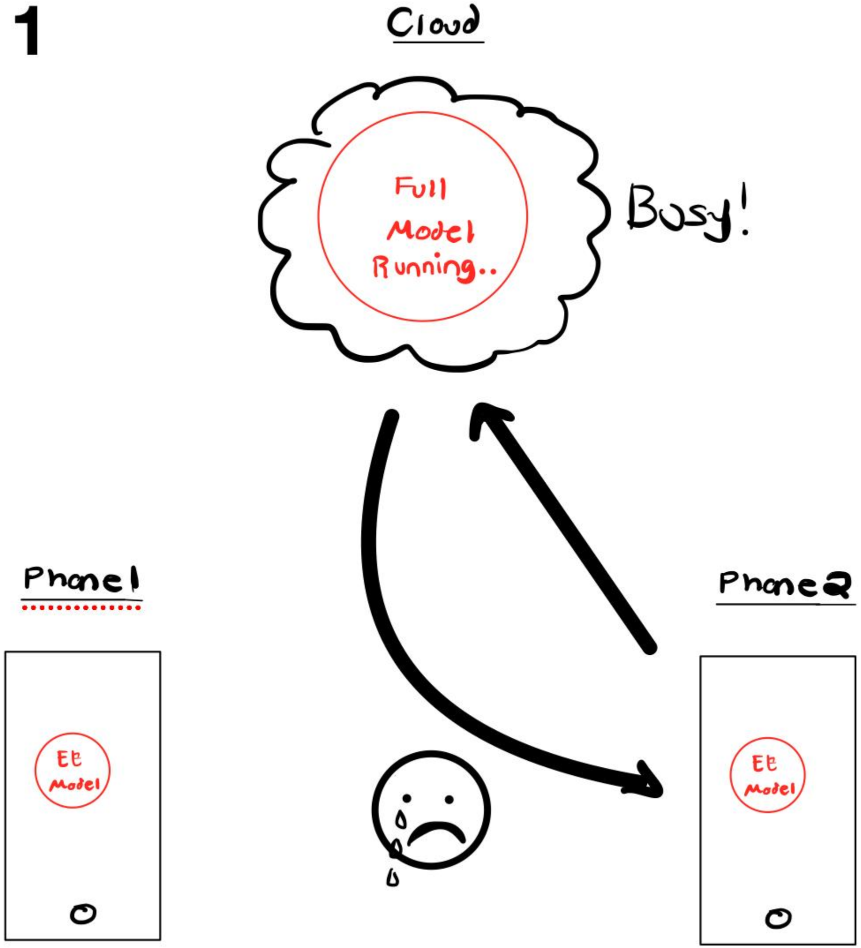
Experiment Objective

Simulated a busy server and evaluated the different strategies client devices could use to obtain optimal results.

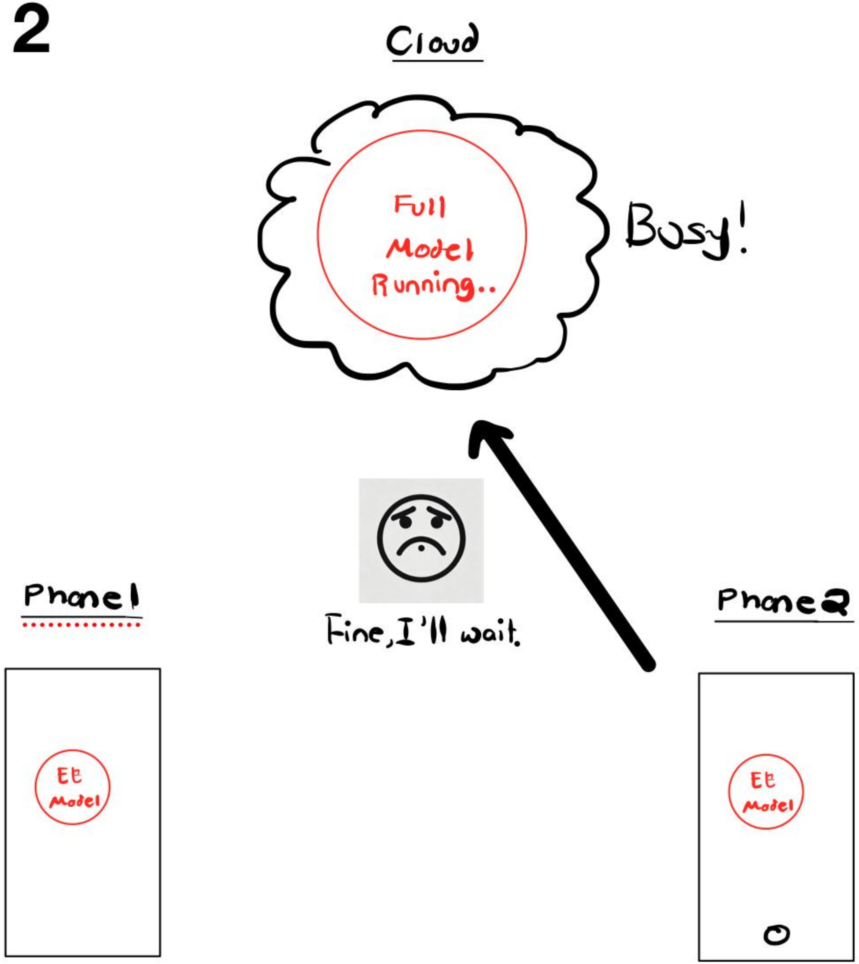

Cloud
Full Model
Phone1
Phone2
EE Model
EE Model


Cloud
Full Model Running..
Busy!
Phone1
EE Model
Phone2
EE Model

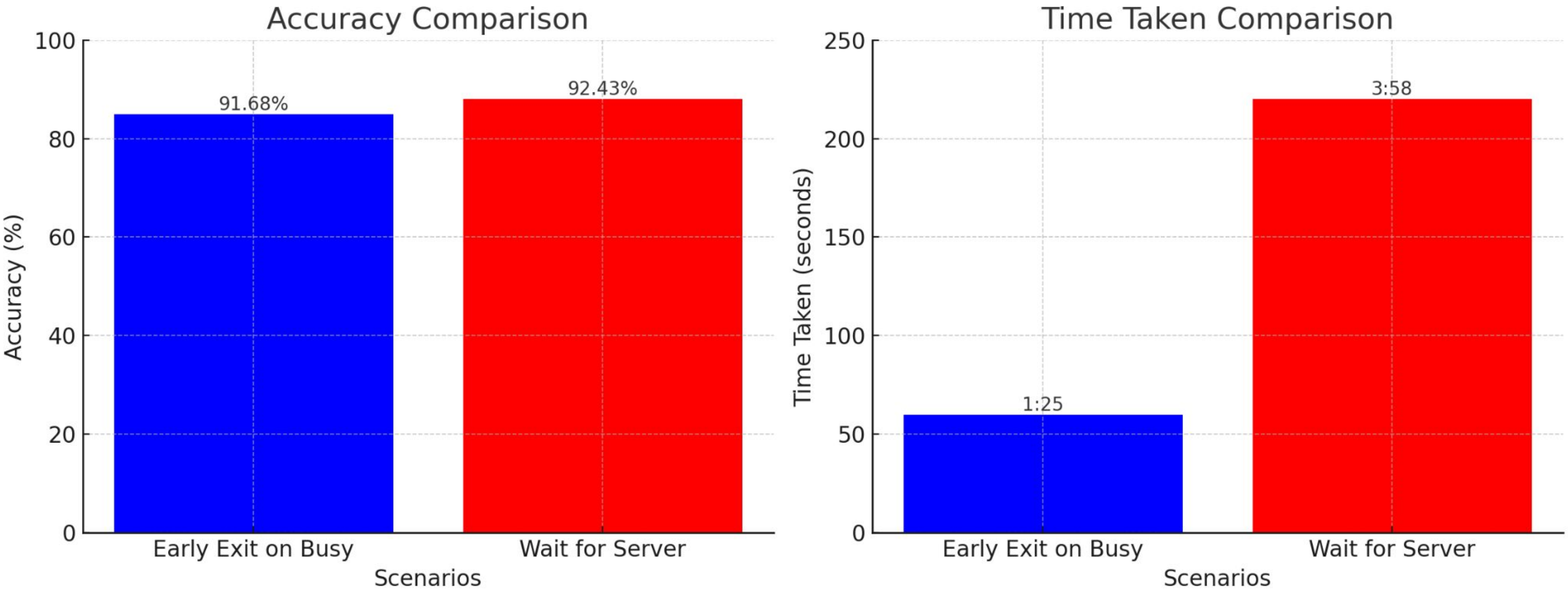
Scenario 1

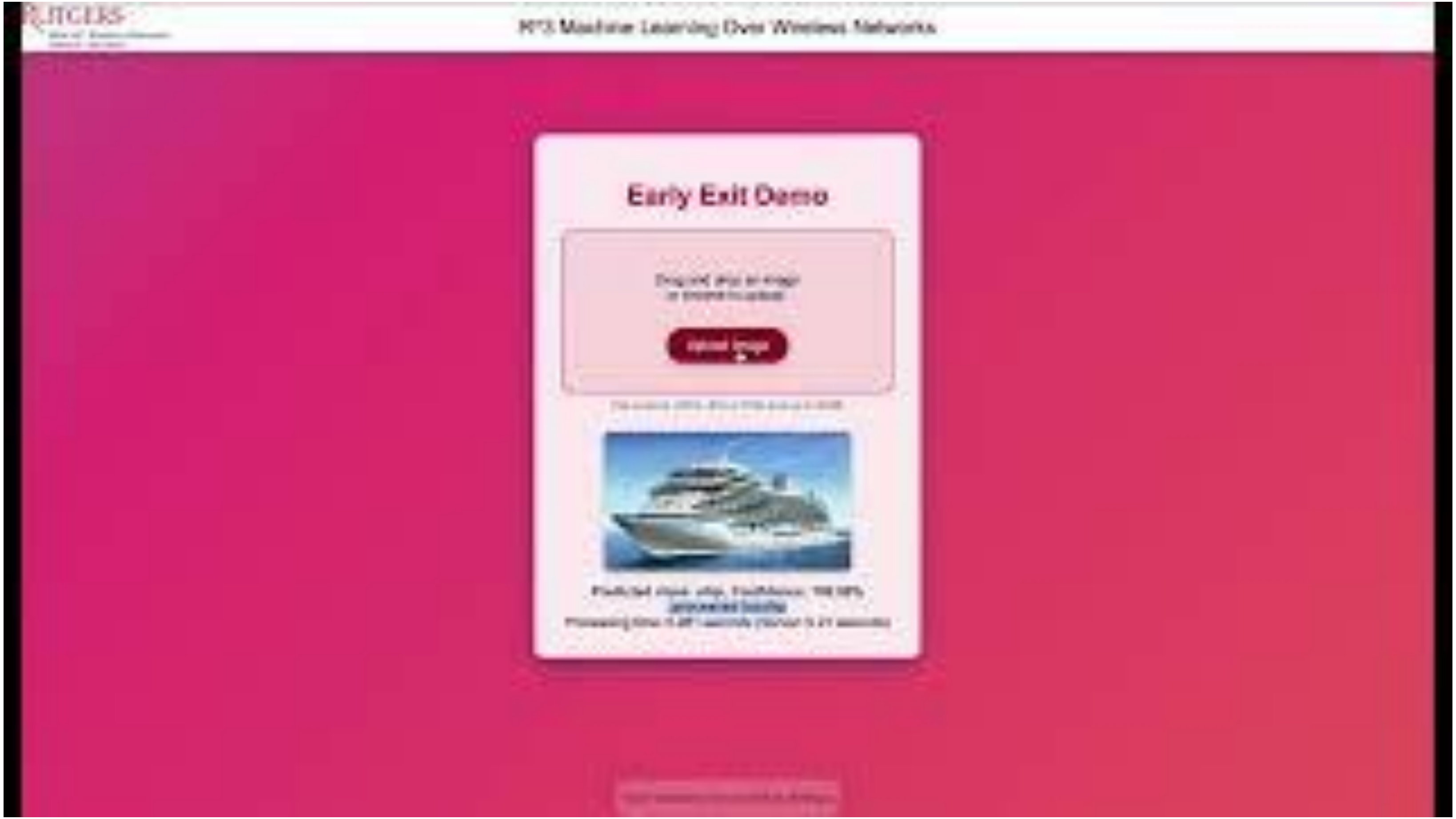

Cloud
Full Model Running..
Busy!
Phone1
EE Model
Phone2
EE Model

Scenario 2


Cloud
Full Model Running..
Busy!
Fine, I'll wait.
Phone1
Phone2
EE Model
EE Model

Comparison of Scenarios: Early Exit vs. Waiting for Server





RT3 Machine Learning Over Wireless Networks

Early Exit Demo

Drag and drop an image
or browse to upload

[illegible]

Predicted class: ship, Confidence: 100.00%
[illegible]
Processing time: 0.47 seconds (Server: 0.21 seconds)

Conclusions

- For a small ↓ in accuracy, we can get ↑ in speed by offloading when needed
- Applications may prioritize accuracy or latency
- Challenges:
    - Managing multiple clients in real life
    - Model portability for micro devices
    - Performance in weak wireless connections

Future Work

- Compress and optimize models
- Explore dynamic thresholds
- Train the model with diverse datasets
- Investigate other types of models

Thank You!