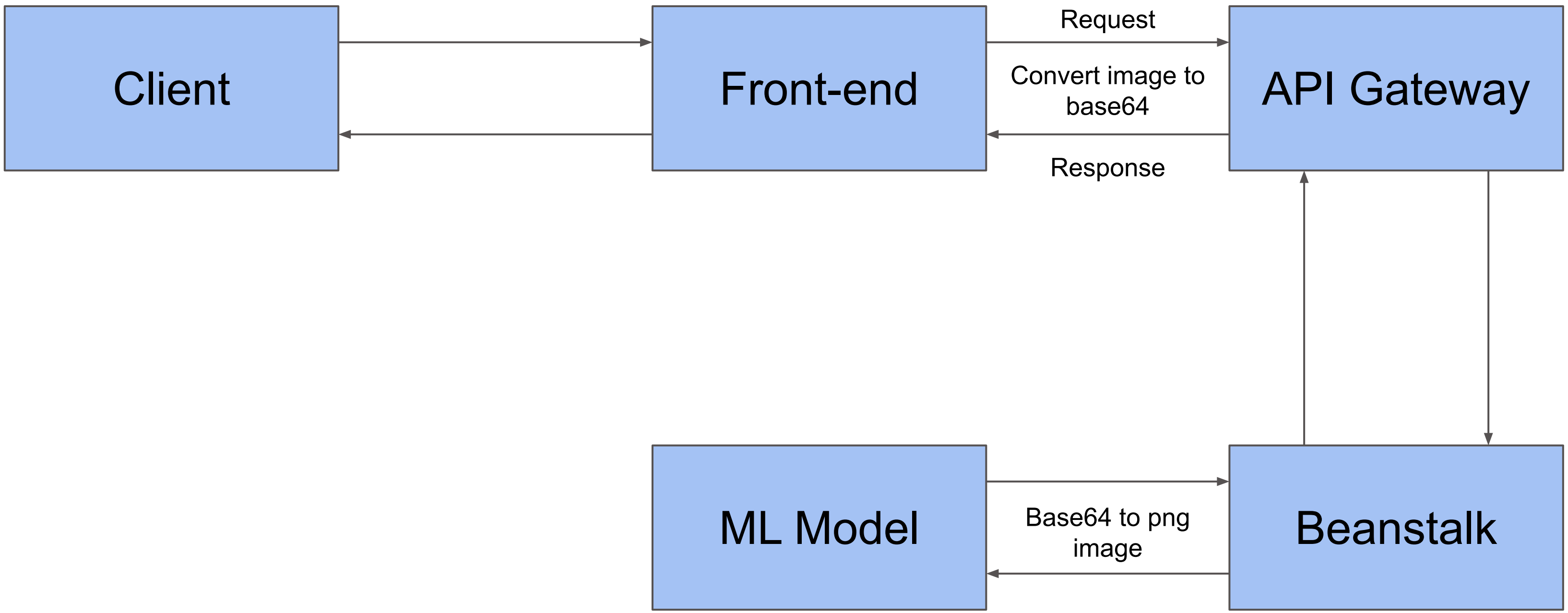
# Distributed Machine Learning

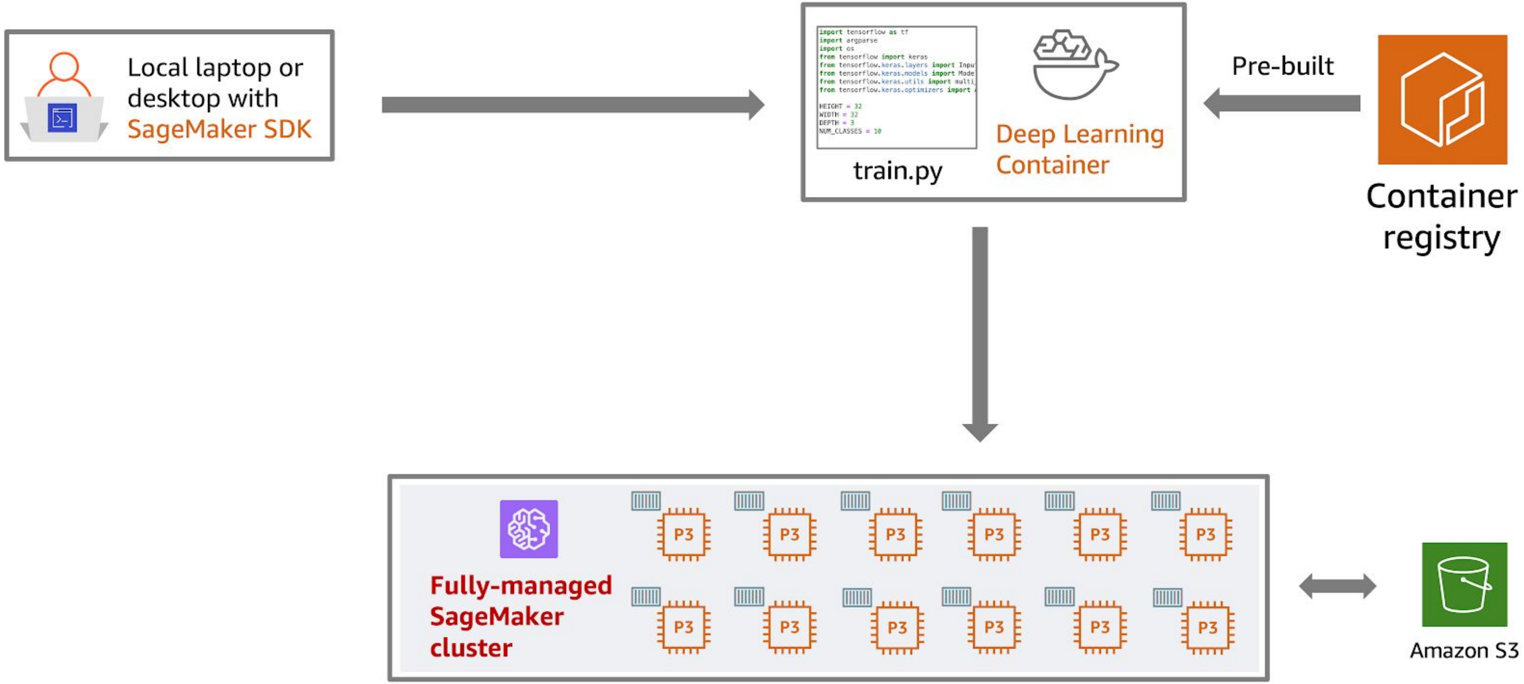
# Idea

Perform distributed training with TensorFlow and Horovod on Amazon Sagemaker for image classification of fruit dataset from Kaggle

# Overall Architecture



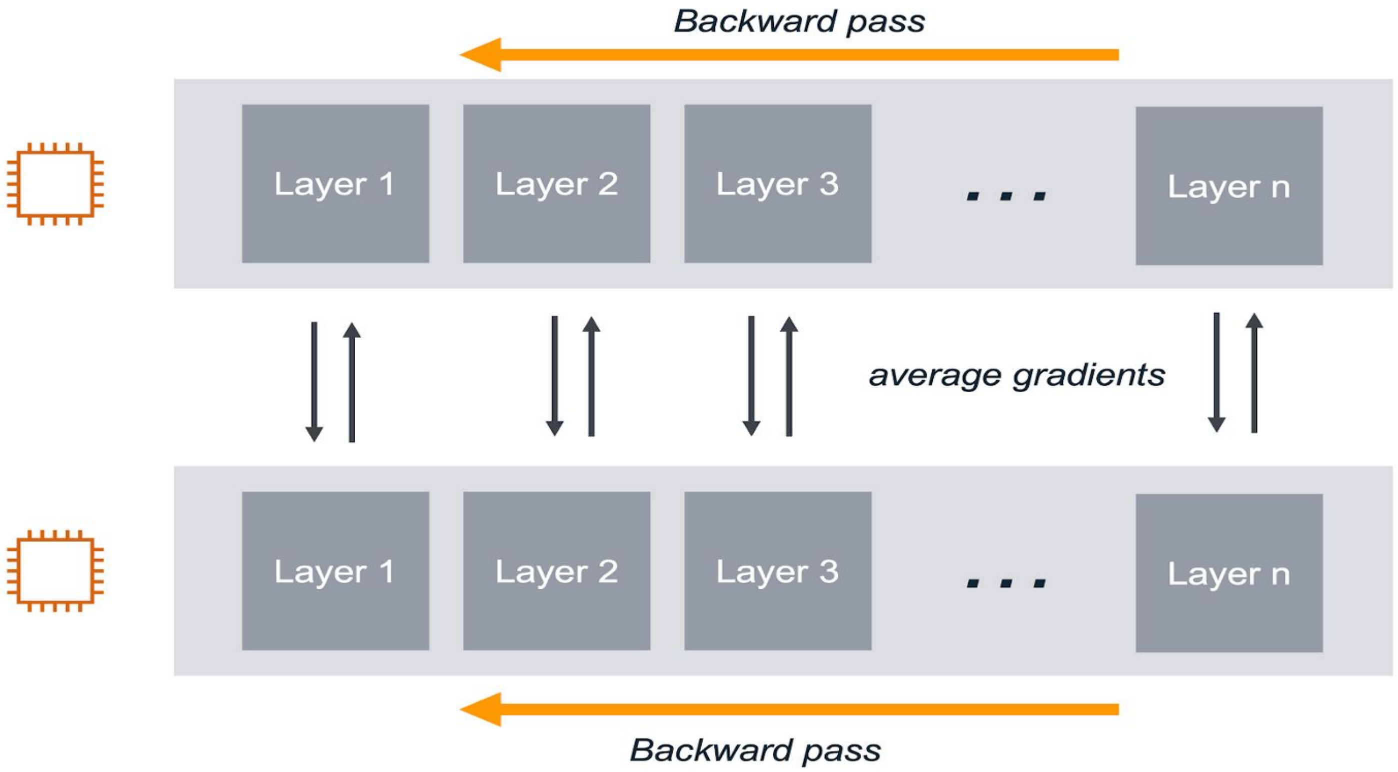
# ML Architecture Design



Local laptop or desktop with SageMaker SDK
Deep Learning Container
train.py
Pre-built
Container registry
Fully-managed SageMaker cluster
P3
Amazon S3

```
import tensorflow as tf
import argparse
import os
from tensorflow import keras
from tensorflow.keras.layers import Inpu
from tensorflow.keras.models import Mode
from tensorflow.keras.utils import multi_
from tensorflow.keras.optimizers import

HEIGHT = 32
WIDTH = 32
DEPTH = 3
NUM_CLASSES = 10
```


Backward pass
Layer 1
Layer 2
Layer 3
. . .
Layer n
average gradients
Layer 1
Layer 2
Layer 3
. . .
Layer n
Backward pass

# Implementation

1. Modified training script to be distributed ready using Horovod library
2. Used Horovod library to ensure the training script successfully scales to train across many GPUs in parallel, i.e. the training script is GPU count agnostic
    i.

# Dataset

Dataset: Subset of Fruits 360 from Kaggle - apples, kiwi and guava

Training images: 9600

Validation images: 1200

Testing images: 1200

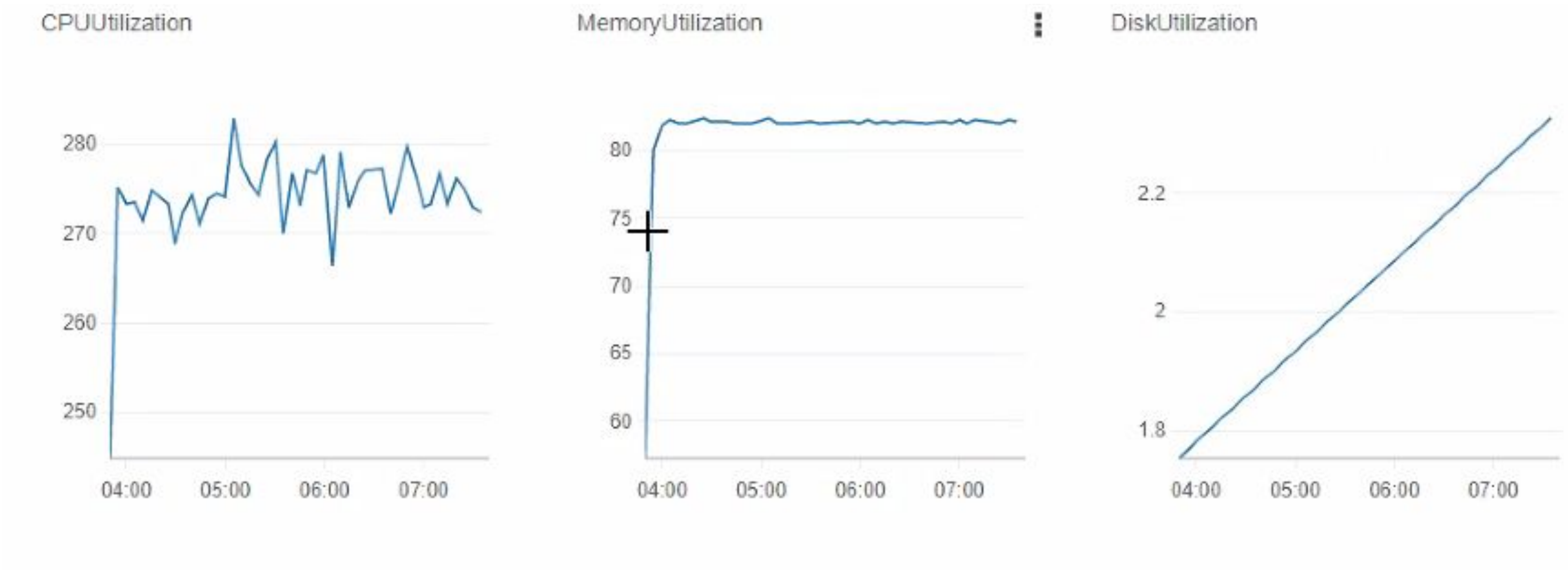
# Distributed training using Amazon Sagemaker

Used Amazon Sagemaker to run the training set on scale, with the ability to scale up and down GPU’s as desired

- Launched Amazon Sagemaker NB instance
- Defined the estimator with training script, location to save trained models, type of GPUs, number of GPUs per instance, TensorFlow version, MPI distribution type
- Specified paths to training, validation and test datasets in Amazon S3, passed those parameters to estimators fit function

# Monitoring

We monitored the progress through Amazon Cloud Watch



# User Interface

Web application to serve the users to use the machine learning model.

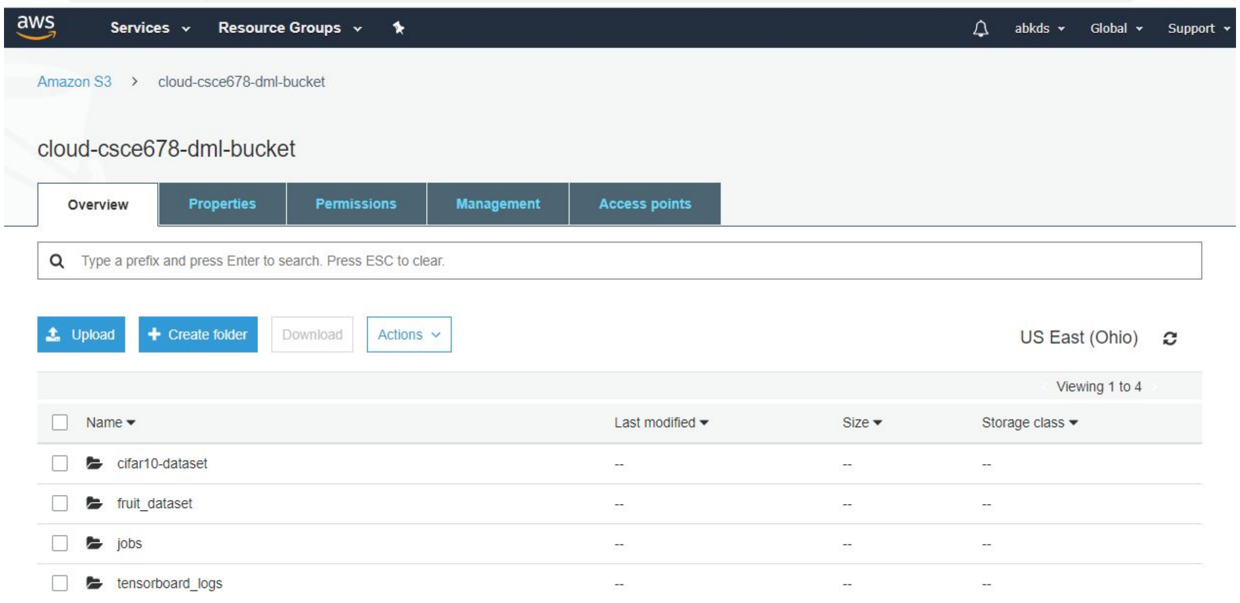
Frontend (React and Angular.js)

https://github.com/abkds/dml-frontend

Backend: EC2 Beanstalk

# Results

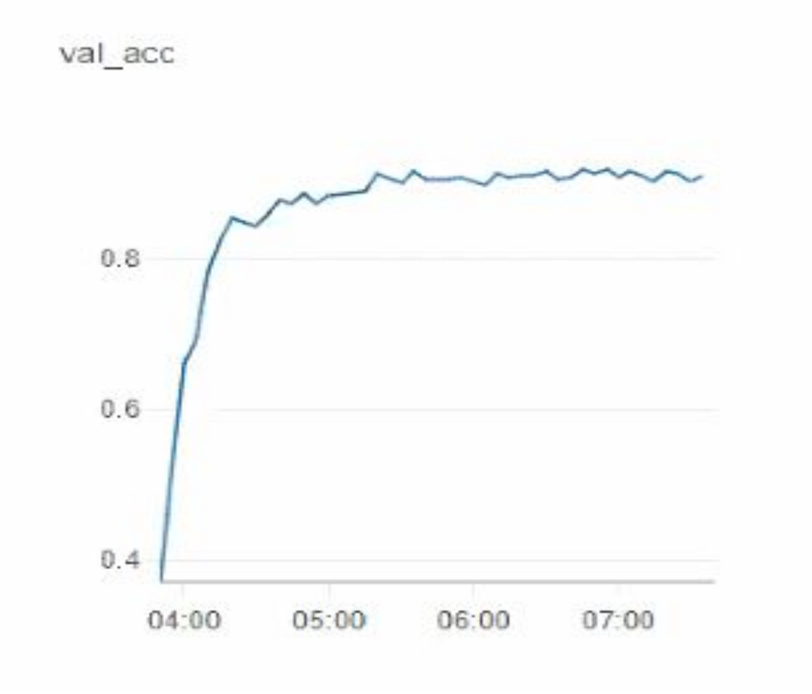
Once training was complete, Sagemaker automatically uploaded training artifacts such as trained nodes, checkpoints and tensorboard logs into our S3 bucket.



cloud-csce678-dml-bucket

| | Name | Last modified | Size | Storage class |
|---|---|---|---|---|
| □ | cifar10-dataset | -- | -- | -- |
| □ | fruit_dataset | -- | -- | -- |
| □ | jobs | -- | -- | -- |
| □ | tensorboard_logs | -- | -- | -- |

# Test accuracy : 90.72



# Thank You!