Machine Learning On Edge Computing

Presented by:

Rohit Singh

Akanksha Raina

Srujana Malisetti

Contents

Motivation

Problem Statement

Goal

Related Work

Design

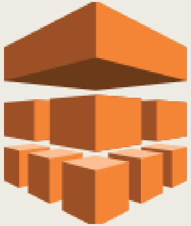
Measurement

Expected Result

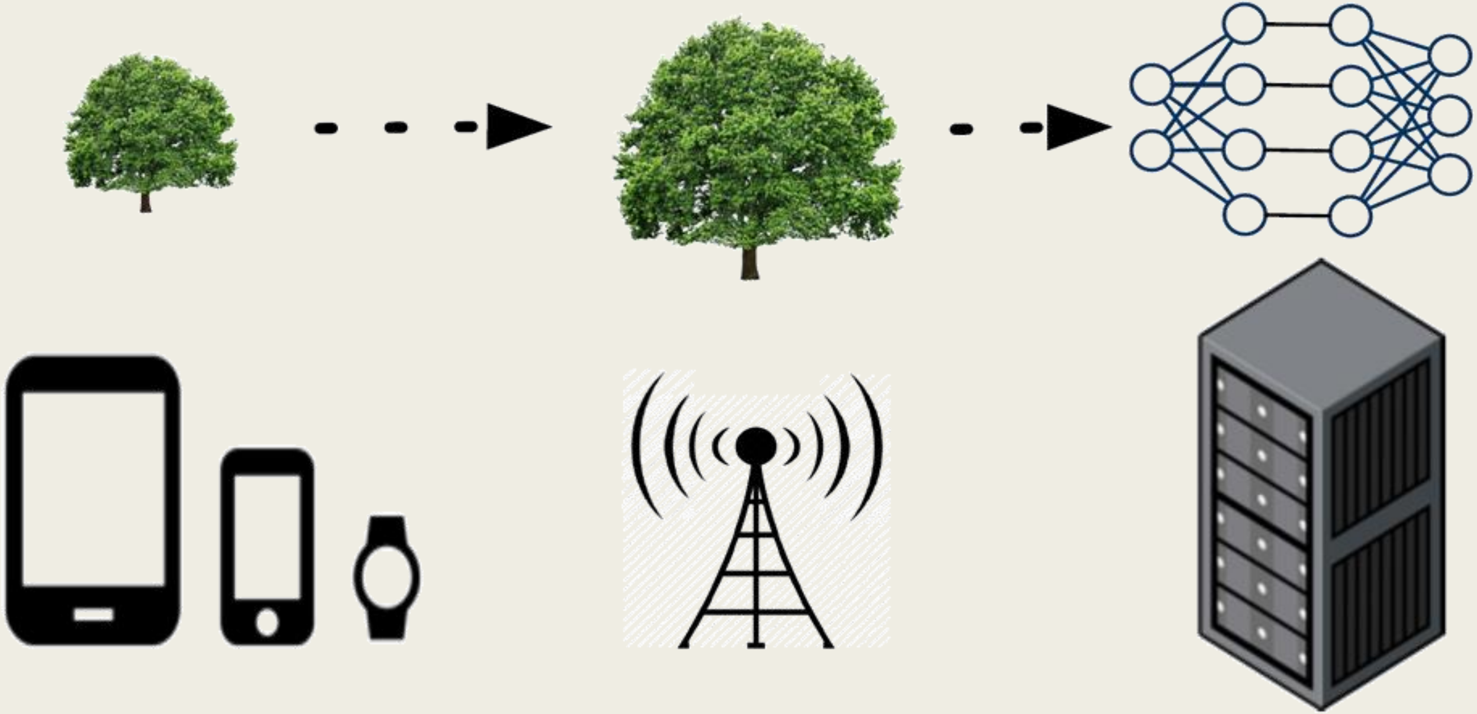
Timeline

Machine Learning Services

- Many Cloud Providers now a days are providing Machine Learning Services.
- Termed as MLaaS.

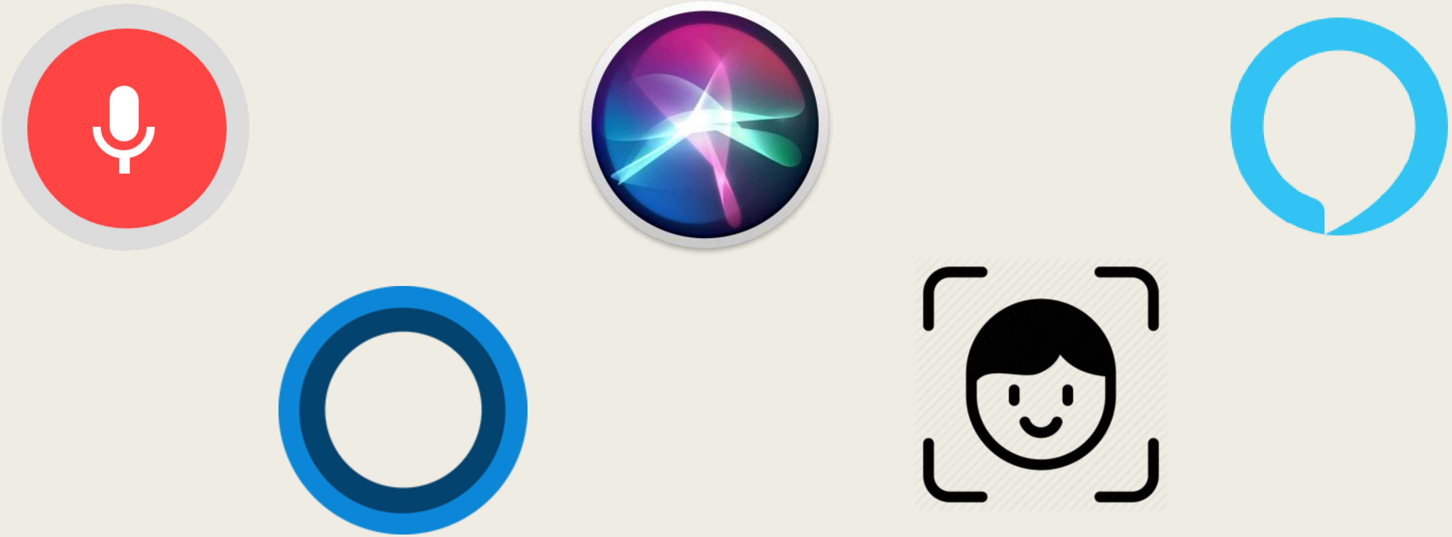


Status Quo Approach

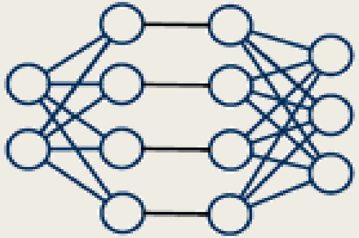


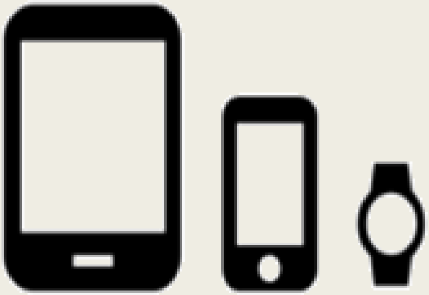
On Other side

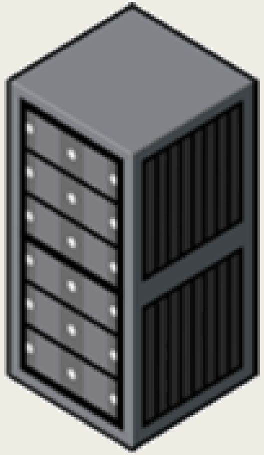
- Intelligent Personal Assistants running on SoC integration devices, have capability to run ML Models efficiently.



How about Edge Computing ?







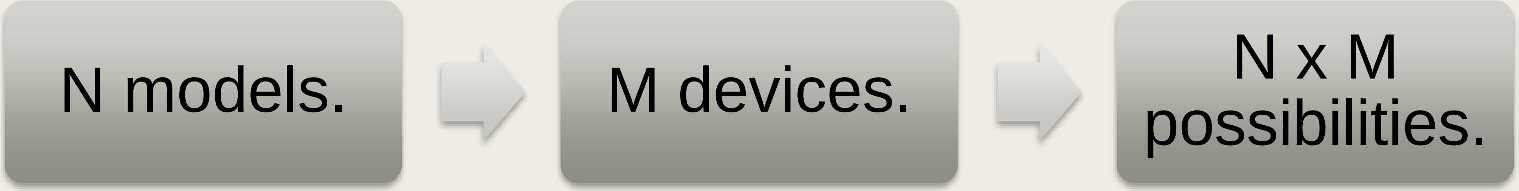
Many Options

AlexNet VGG CaffeNet	Image Classification
DeepFace FaceNet NormFace	Face Recognition
Kaldi DeepSpeech	Speech Recognition
SENNA Tesseract	Text Recognition

Apple Siri
Microsoft Cortana
Google Now
Amazon Alexa
Raspberry Pi
Jetson Nano
Cloud - VM, Container, Functions

N **x** M

Choose Best?



N models.

M devices.

N x M possibilities.

How to choose best devices or models?

Help from!!

- **Complexity v/s Performance : Empirical Analysis of Machine Learning as a Service**

http://people.cs.uchicago.edu/~ravenben/publications/pdf/mlaas-imc17.pdf

- **Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge**

http://web.eecs.umich.edu/~jahausw/publications/kang2017neurosurgeon.pdf

- **Spock: Exploiting Serverless Functions for SLO and Cost Aware Resource Procurement in Public Cloud**

http://www.cse.psu.edu/~pxt176/publications/cloud-spock.pdf

- **Distributed Perception by Collaborative Robots**

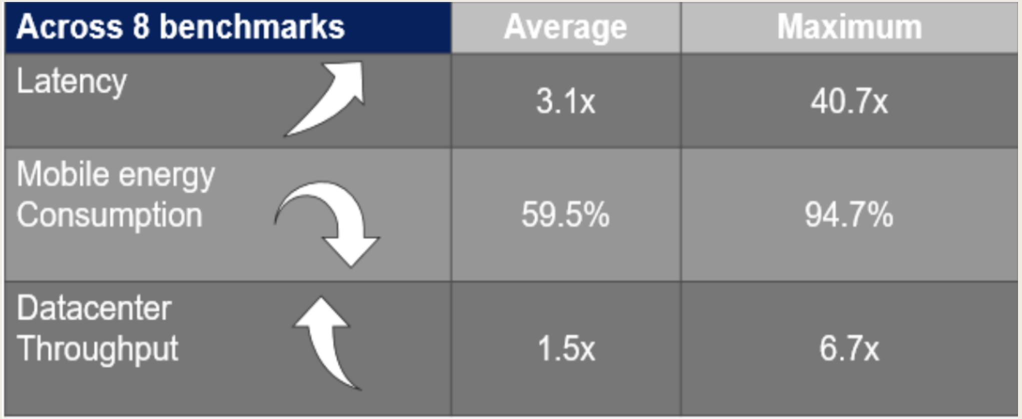
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8411096

Complexity v/s Performance : Empirical Analysis of Machine Learning as a Service

- The paper discusses how MLaaS systems can provide an alternative to standalone ML classifiers.
- The paper provides empirical analysis of MlaaS platforms. Following points were observed during the analysis:
  - With more control comes more potential performance gains as well as greater performance degradation from poor configuration decisions.
  - Fully automated platforms are optimizing classifiers using internal tests.
  - Much of the gains from configuration and tuning come from choosing the right classifier.
  - Experimenting with a small random subset of classifiers is likely to achieve near optimal results.

Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge

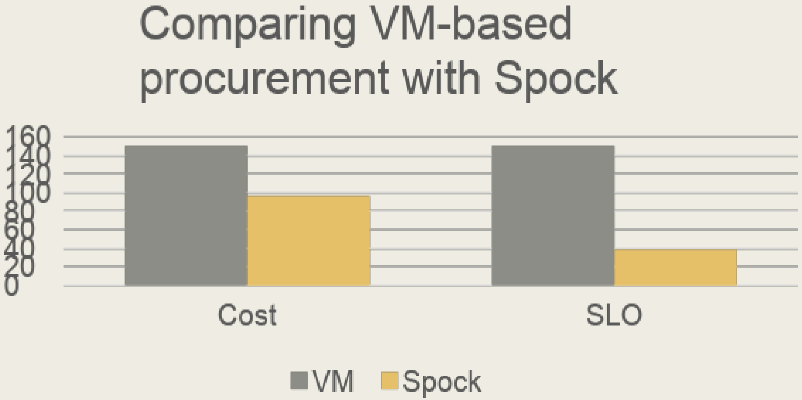
- A system that can automatically partition Deep Neural Networks between mobile devices and the cloud at the granularity of neural network layers.
- Neurosurgeon adapts to dynamic conditions, like server load levels and wireless network connection., while making a decision.
- It chooses partition point for best latency and best mobile energy consumption.



Across 8 benchmarks	Average	Maximum
Latency ↗	3.1x	40.7x
Mobile energy Consumption ↓	59.5%	94.7%
Datacenter Throughput ↑	1.5x	6.7x

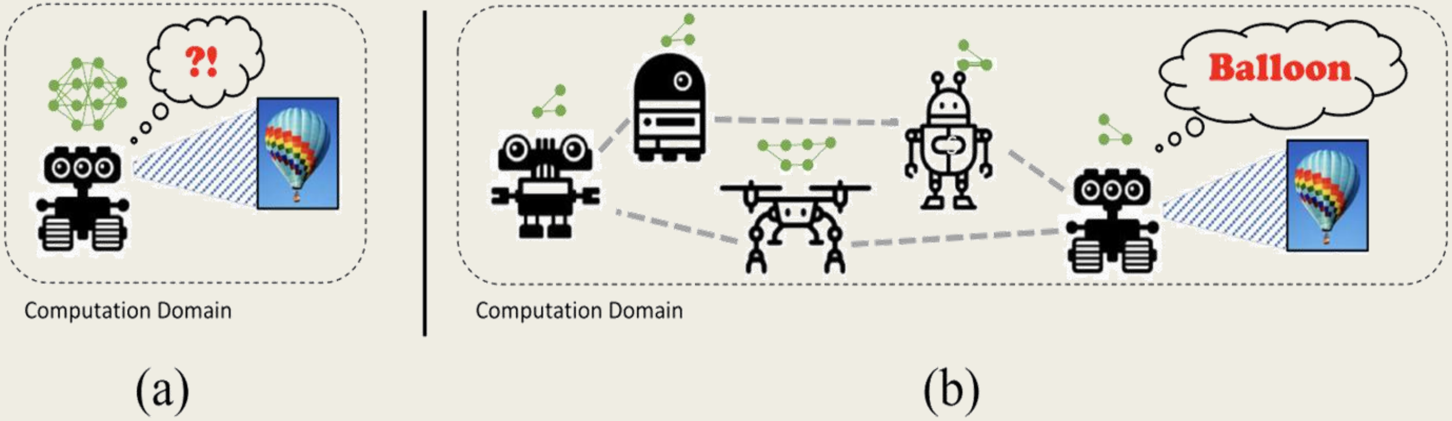
Spock: Exploiting Serverless Functions for SLO and Cost Aware Resource Procurement in Public Cloud

- The paper describes using serverless functions for resource procurement in public cloud of VM –based autoscaling.
- Spock, a new scalable and elastic control system that exploits both VMs and serverless functions to reduce cost and ensure SLO for elastic web services.
- Spock helps in overcoming the shortcomings of VM-based resource procurement.



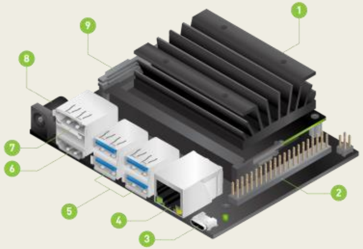
Distributed Perception by Collaborative Robots

- The paper introduces the concept of collaborative approach among robots.
- It enables efficient, dynamic and real time recognition.
- Similar performance results when compared to High Performance machine (HPC) and Jetson TX2

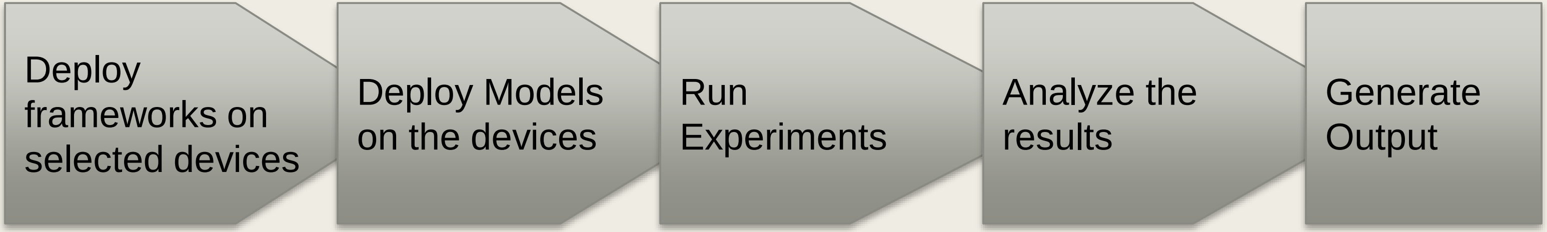


Design

- **List of Devices** :
  - *Raspberry Pi*
  - *Mobile Device (Android)*
  - *AWS Cloud - EC2 Instance, Container, Lambda Functions*
  - *Jetson Nano*
- **List of Machine learning frameworks:**
  - *Caffe*
  - *TensorFlow*
  - *mxNet*
  - *Paddle*
- **List of ML Models or Applications:**
  - *AlexNet*
  - *GoogleNet*
  - *CaffeNet*
  - *DeepFace*
  - *VGG*
  - *SENNA*



Approach



Deploy frameworks on selected devices

Deploy Models on the devices

Run Experiments

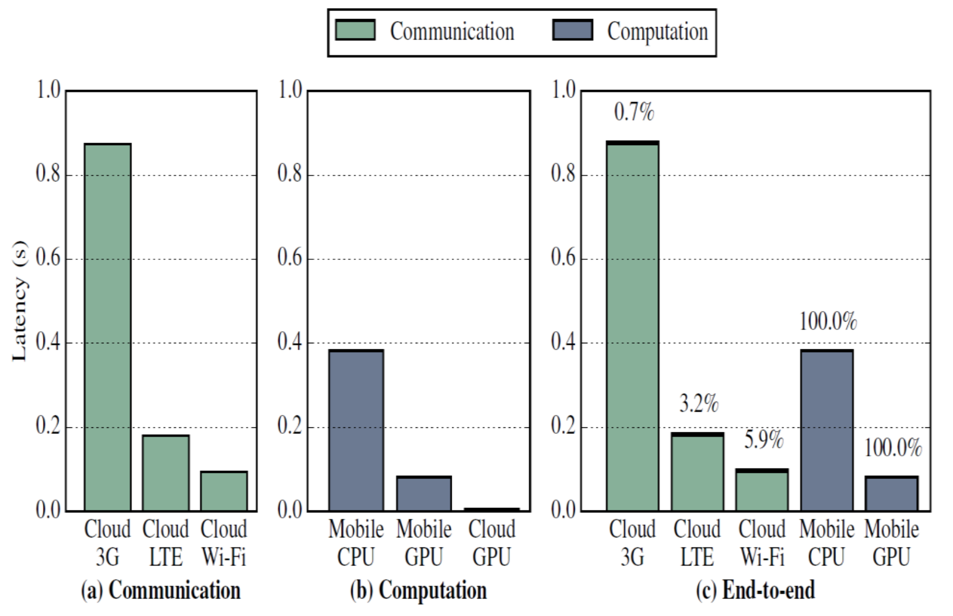
Analyze the results

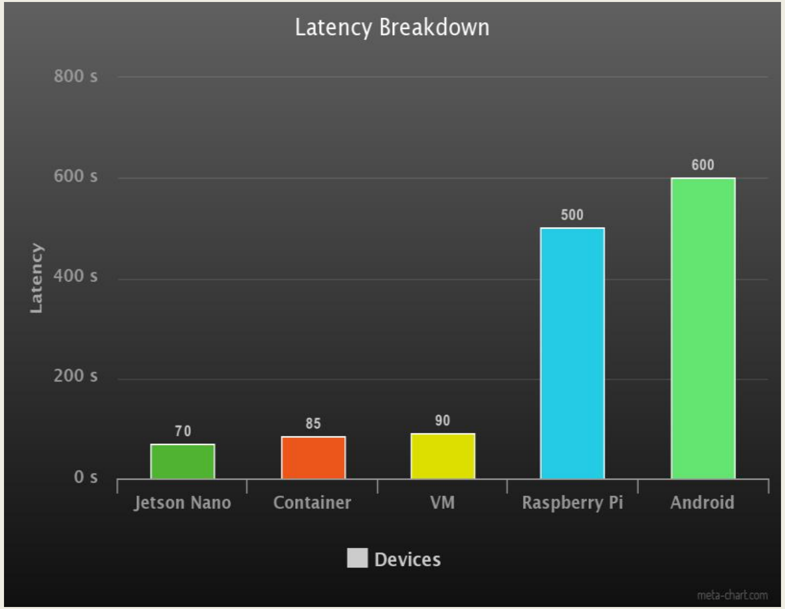
Generate Output

Measurement

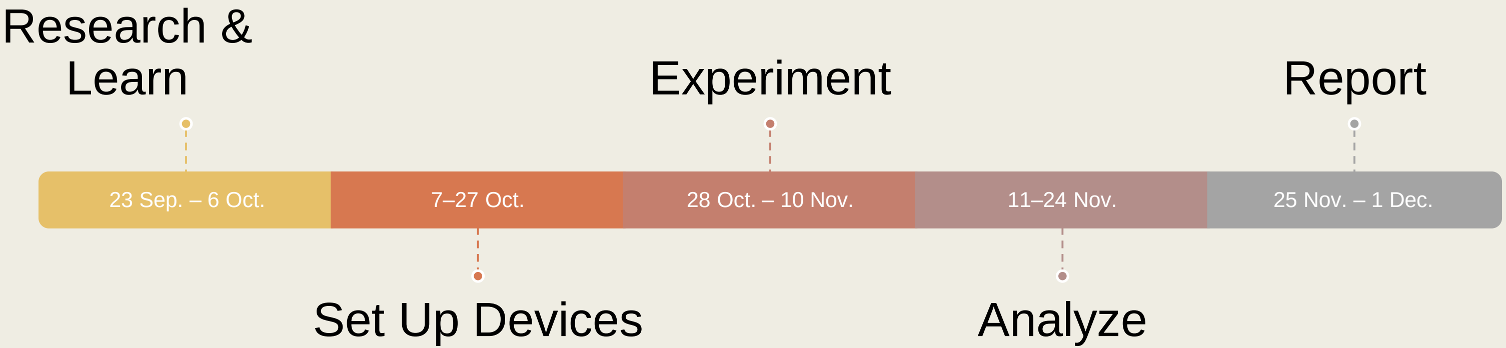
- Baseline - Neurosurgeon
- Performance – Accuracy of calculations (F1-Score, GFLOPS)
- Latency of each model on each device
- Memory consumption on each device
- Battery consumption on each device

Expecting Result





Timeline



Research & Learn

23 Sep. – 6 Oct.

Set Up Devices

7–27 Oct.

Experiment

28 Oct. – 10 Nov.

Analyze

11–24 Nov.

Report

25 Nov. – 1 Dec.

References

- https://skymind.ai/wiki/comparison-frameworks-dl4j-tensorflow-pytorch#ml
- https://medium.com/coinmonks/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification-b93598314160
- https://arxiv.org/pdf/1804.06655.pdf
- https://en.wikipedia.org/wiki/Speech_recognition
- https://github.com/PaddlePaddle/Paddle

Thank You!!

Questions??