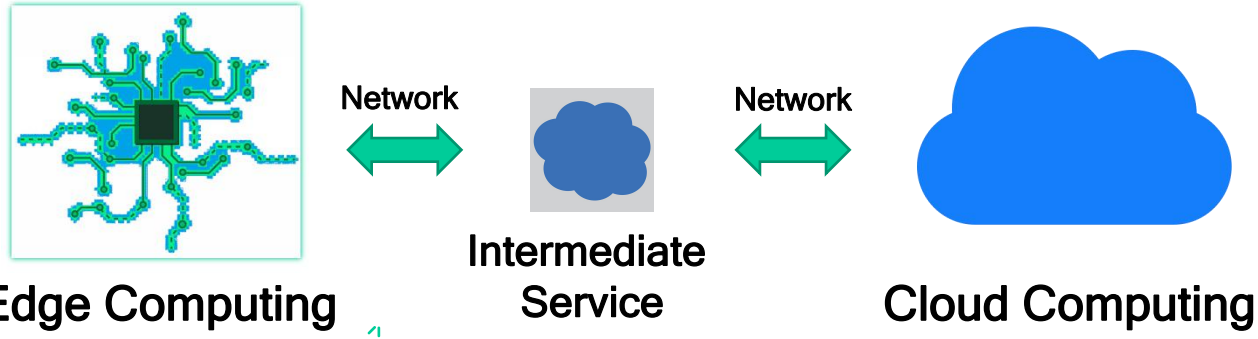
Work-In-Progress: Making Machine Learning (ML) Real-Time Predictable

Hang Xu, Frank Mueller

Motivation



Why does ML move?

- **Real-Time features of ML API on edge**
  - Shorter average execution time
  - Tighter worst case execution time (WCET)
- Large streaming data inputs
- Data privacy concerns
- Lower latencies

What ML Tasks On Embedded System

ML Tasks	Training	Inference
Resource Demand	High	Low
Location	Cloud	Edge
Time/Power Cost	High	Low

Unsupervised learning - intrinsically a training task

ML Libraries On Edge

1. Keras (Tensorflow backended)
   - Interpreter-based language
   - No real-time control of dynamic memory management

2. Caffe
   - Native C++ language
   - Real-time control of dynamic memory management

Caffe

3. Enhanced Caffe
   - Remove third party library invocation functions in source code
   - Remove multi-core support

RT Performance Comparison

Keras vs. Original Caffe

Average execution time

- 4:1

Stadard deviation of execution time

- less varying : much more varying

RT-Enhanced Caffe vs. original Caffe

Average execution time

- 1: 6

Standard deviation of execution time

- 1:25 (comparison between the minimum values)