TVM for edge computing platform

NTT Software Innovation Center

Kazutaka Morita
Inference in 5G era

Devices

Offload

Offload

MEC (Mobile edge computing) server

Edge

Internet

Base station

~10 ms latency

Cloud
Benefits of offloading inference

Computing resource

- **Edge**
  - **GPU**: High-end server-spec accelerators are available
  - 5G
- **Device**
  - **CPU**
  - **AI chip**: Edge is one of the targets of AI accelerators

Inference with data

- **Cloud**
  - **Big data**
  - Real-time inference with big data
  - 5G
- **Edge**
  - **data**
- **Device**

AI chip is unavailable for low-end devices

Interaction with other devices
Example – Augmented Reality

Occlusion

Object segmentation inference

Plane detection

Object detection inference can also provide collider from moving real world objects

Object will not collide

Point cloud

Cloud

Point cloud data

Captured images

Inference with big data in the cloud

Many Inference tasks

HYPER-REALITY: https://vimeo.com/166807261
Edge computing platform with TVM

Offload inference if necessary, based on device and communication status

Developer

Developing framework for edge computing

Device SDK

TVM

Device

Edge

Cloud

Distribute runtimes to device, edge, and cloud

Data

Offload

Internet
What are required for TVM?

**Heterogeneous runtime with offloading support**
- Execute on edge via RPC
- Auto tuning support would be also nice

**Dynamic runtime**
- Switch based on device and communication status

**Smart NIC support**
- No overhead of PCIe communication or host memory access

- Edge
  - CPU
  - GPU
  - NIC
- Device
  - Smart NIC
  - FPGA
Thank You!

Email: kazutaka.morita.fp@hco.ntt.co.jp