AN EDGE FRAMEWORK FOR FACIAL RECOGNITION

RUOYU YING, QI JIA “JOY” LIU AND SHANE WANG

INTEL
AGENDA

• Why and what is Edge Computing
• One Use Scenario
• Framework for Cloud Stack and Edge Stack
• Summary
## EDGE COMPUTING

<table>
<thead>
<tr>
<th>NFV Edge Infrastructure</th>
<th>Wireless (vRAN,vEPC)</th>
<th>Wireline (PON)</th>
<th>uCPE (SD-WAN)</th>
<th>IP Enterprise Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomous Devices</td>
<td>Drones</td>
<td>Autonomous Vehicles</td>
<td>Industry Robots</td>
<td>Medical</td>
</tr>
<tr>
<td>Immersive Experiences</td>
<td>Virtual Reality</td>
<td>Augmented Reality</td>
<td>360 Video</td>
<td>Wearable Cognitive Assistance</td>
</tr>
<tr>
<td>IoT &amp; Analytics</td>
<td>Industrial Sensors</td>
<td>Home Devices</td>
<td>Retail</td>
<td>Healthcare</td>
</tr>
</tbody>
</table>

- **On-Demand NFV**
- **Hardware Acceleration**
- **A.I.**
- **Microservices**
- **5G**
FROM CLOUD TO EDGE

Clients

- Flexibility / -Capacity / -Latency / + Locality

Base Station
Fiber
Cable

Last mile/access

Local CO

Regional CO

CORE

Cloud / OTT

+ Flexibility / +Capacity / + Latency / - Locality

Clients

Edge

- StarlingX (OSF Edge)
- Akraino (LF Edge)
- OpenNESS
- EdgeXFoundry
- ...

NFV

- OpenStack
- OPNFV
- ONAP
- ...

NFV
ONE USE SCENARIO

- Leveraging 5G Core network, the system separates and forwards surveillance data to edge computing platform, and then adopt AI technology for facial recognition and ID confirmation, after training personal data on the cloud.
- The solution on edge and 5G could adopt at the spots like airports, railway stations, ferries, tourist attractions, etc. for public safety, rescue, onsite checking and commanding, people density warning, etc., and reduce staff workloads.
CLOUD STACK: OPENSTACK AND KUBERNETES

**APPLICATIONS AND SERVICES**

- **DEEP LEARNING (INFEERENCE)**
  - OpenVINO™
    - Open Visual Inference & Neural Network Optimization toolkit for inference deployment on CPU/GPU/FPGA for TensorFlow®, Caffe® & MXNet®
  - Intel® Movidius™ SDK
    - Optimized inference deployment on Intel VPUs for TensorFlow® & Caffe®

- **REASONING**
  - Intel® Saffron™ AI
    - Cognitive solutions on CPU for anti-money laundering, predictive maintenance, more

- **DEEP LEARNING FRAMEWORKS**
  - Now optimized for CPU
    - Optimizing
      - TensorFlow®, MXNet®, Caffe®, BigDL (Spark)®, Caffe2®, PyTorch®, CNTK®, PaddlePaddle®,...

**MACHINE LEARNING LIBRARIES**

- **Python**
  - Scikit-learn
  - Pandas
  - NumPy
- **R**
  - Cart
  - Random Forest
  - e1071
- **Distributed**
  - MLlib (on Spark)
  - Mahout

**ANALYTICS, MACHINE & DEEP LEARNING PRIMITIVES**

- **Python**
  - Intel distribution optimized for machine learning
- **DAAL**
  - Intel® Data Analytics Acceleration Library (incl. machine learning)
- **MKL-DNN**
  - Open-source deep neural network functions for CPU / integrated graphics
- **cIDNN**

**DEEP LEARNING GRAPH COMPILER**

- Intel® nGraph™ Compiler
  - Open-sourced compiler for deep learning model computations optimized for multiple devices from multiple frameworks
## EDGE STACK: INTEGRATED CLOUD NATIVE

<table>
<thead>
<tr>
<th>AI applications (Inference)</th>
<th>Tools to developers to optimize and convert for Edges (e.g DPDK, OneAPI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC Orchestration (ONAP4K8S)</td>
<td>MC Security Controller</td>
</tr>
<tr>
<td>Multi-tenant Kubernetes</td>
<td>MC Network Controller</td>
</tr>
<tr>
<td>QoS based Admission control</td>
<td>Augment K8S for all deployment types</td>
</tr>
<tr>
<td>Geo Distributed DBaaS</td>
<td>Monitoring Platform (fluentd, Elasticsearch, Kiana)</td>
</tr>
<tr>
<td>Geo Distributed HSM</td>
<td>Optimized Service mesh ISTIO/Envoy</td>
</tr>
<tr>
<td>Accelerator Plugins QAT, SRIOV</td>
<td>Telemetry Collection Collectd</td>
</tr>
<tr>
<td>Telemetry Distribution Prometheus</td>
<td>Distributed Security &amp; Network functions</td>
</tr>
<tr>
<td>NFD</td>
<td>Distributed Data &amp; Analytics platform</td>
</tr>
<tr>
<td>OVN4K8S</td>
<td>Monitoring Platform (fluentd, Elasticsearch, Kiana)</td>
</tr>
<tr>
<td>Docker</td>
<td>Optimized Service mesh ISTIO/Envoy</td>
</tr>
<tr>
<td>Kata Containers</td>
<td>Virtualization of • Local accelerators</td>
</tr>
<tr>
<td>Virtlet</td>
<td>• Remote Accelerators</td>
</tr>
<tr>
<td>Ubuntu</td>
<td>Virtualization of • Local accelerators</td>
</tr>
<tr>
<td>Clear</td>
<td>• Remote Accelerators</td>
</tr>
<tr>
<td>SmartNIC</td>
<td>Virtualization of • Local accelerators</td>
</tr>
<tr>
<td>Inline Crypto Acceleration</td>
<td>• Remote Accelerators</td>
</tr>
<tr>
<td>Autonomous Media Acceleration</td>
<td>Virtualization of • Local accelerators</td>
</tr>
<tr>
<td>Elastic Power Management</td>
<td>• Remote Accelerators</td>
</tr>
<tr>
<td>RDT</td>
<td>Virtualization of • Local accelerators</td>
</tr>
</tbody>
</table>

**Cloud Stack**
- Apps, VNFs, CNFs
- Service Orchestrator
- Resource Orchestrator
- Edge Common Services
- Platform Services
- Virtualization & Container Run time
- Operating system
- Hardware platform
SUMMARY

- Edge computing will widely adopted for many new use scenarios, such as facial recognition, autonomous vehicles etc. as AI is adopted and 5G emerges.

- We propose an edge framework with the edge stack running on Integrated Cloud Native (ICN) and OpenNESS, and the cloud stack running on OpenStack or Kubernetes, where AI workloads run in VMs and containers.

- Edge and AI workloads are fully optimized on Intel platforms with acceleration technologies, e.g., SR-IOV, TSN, DPDK, GPU, FPGA and others.
Thanks for Listening
CLEARING TOOLKIPv5

Cross-Platform Tool to Accelerate Computer Vision & Deep Learning Inference Performance

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR = Intermediate Representation file

Inference Engine
Optimized Inference

Code Samples & 10 Pre-trained Models

Traditional Computer Vision Tools & Libraries

Optimized Libraries

OpenCV*

OpenVX*

Photography Vision

Code Samples

For Intel® CPU & CPU with integrated graphics

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes
For CPU with integrated graphics

Optimize Intel® FPGA

FPGA Run Time Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams
FPGA – Linux* only

Intel® Architecture-Based Platforms Support

software.intel.com/openvino-toolkit