Mar 22, 2019

Mountain View, CA

ONNX Partners Workshop

ONNX Edge WG Session

Milan Oljaca
Principal Engineer/Mgr.
Qualcomm Technologies Inc.

Ofer Rosenberg
Senior Staff Engineer
Qualcomm Technologies Inc.

Agenda

- Intro
- Logistics
- Goals and Discussion

Intro

Presenters

Background

- Milan Oljaca – Principal Engineer/Mgr @ Qualcomm Technologies Inc.
  - AI Software Team – San Diego site lead
  - Working in AI/ML domain space for last 6+ years
- Ofer Rosenberg – Senior Staff Engineer @ Qualcomm Technologies Inc.
  - AI Software Team Architect
  - Working in AI/ML domain space for last 4 Years
  - Participant in a few Khronos working groups / specifications (OpenCL since 1.0, NNEF, Vulkan)
- QTI AI Software team responsible for AI software products across our SoC line
  - Qualcomm Neural Processing SDK
  - Android NN HAL
  - Lower level accelerator libraries for Qualcomm® Adreno™ GPU, Qualcomm® Hexagon™ DSP
- Experience and expertise in AI/ML domain space
  - On-device inference
  - Exposure to a variety of use cases through support of many customers

Qualcomm® Neural Processing SDK, Qualcomm® Hexagon™, Qualcomm® Adreno™, Qualcomm® Kryo™ are products of Qualcomm Technologies Inc and/or its subsidiaries.

Qualcomm and ONNX

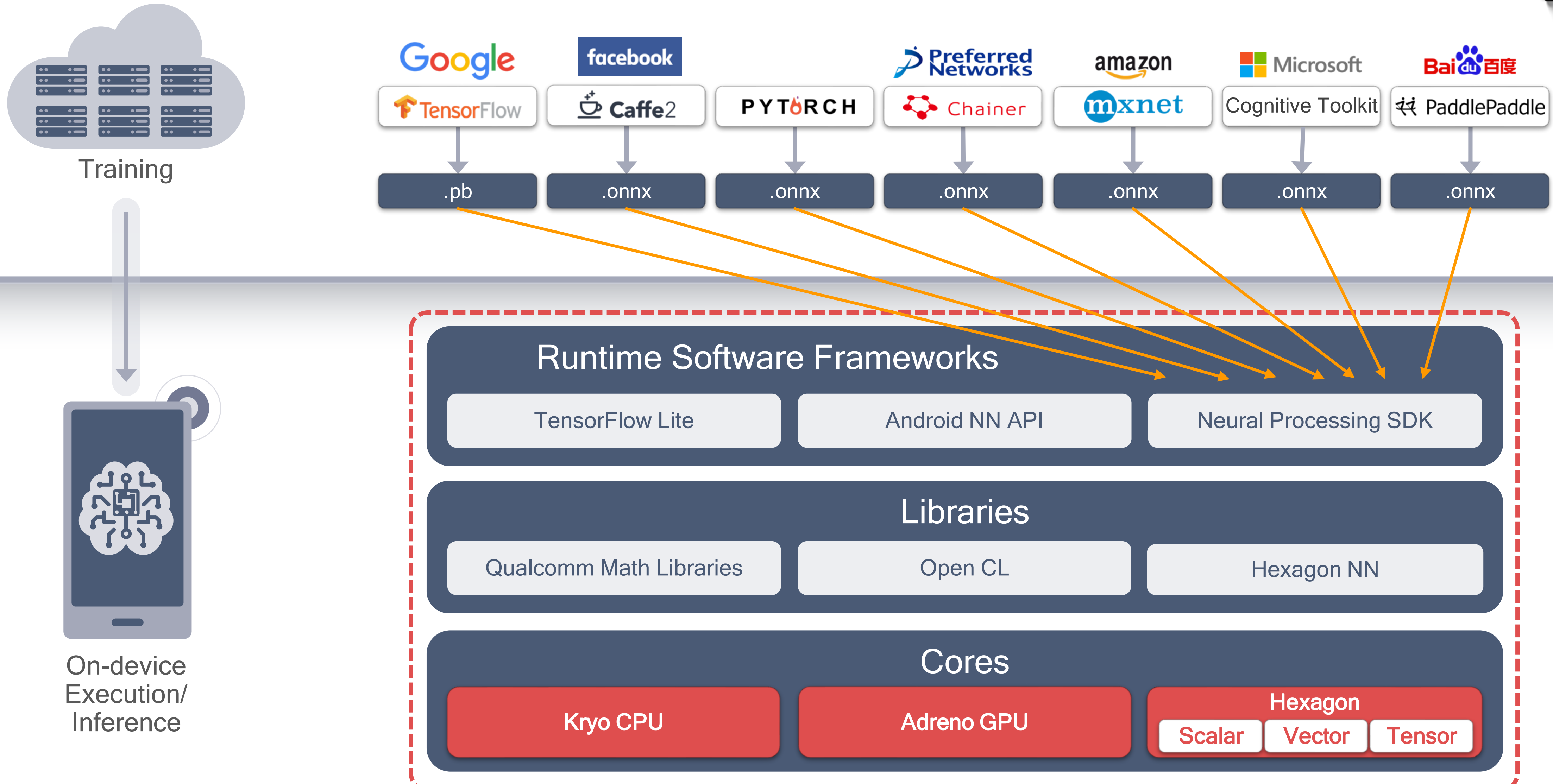
- Qualcomm is a leading supplier of chipsets and solutions for mobile/phone devices
  - Our history in AI goes back 10+ years
  - On device became a reality 4 years ago - Qualcomm® Snapdragon™ 820 - our 1st mobile AI platform
- Early ONNX partner and first to offer ONNX support for edge devices
- Qualcomm Neural Processing SDK
  - https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk
  - Supporting ONNX network conversion since Mar 2018

Qualcomm® Snapdragon™ is a product of Qualcomm Technologies Inc and/or its subsidiaries.

Qualcomm AI SW on Snapdragon

2015 2016 2017 2018 2019

Snapdragon 855



Training

Google — TensorFlow → .pb

facebook — Caffe2 → .onnx

PYTORCH → .onnx

Preferred Networks — Chainer → .onnx

amazon — mxnet → .onnx

Microsoft — Cognitive Toolkit → .onnx

Baidu 百度 — PaddlePaddle → .onnx

On-device Execution/ Inference

Runtime Software Frameworks

TensorFlow Lite | Android NN API | Neural Processing SDK

Libraries

Qualcomm Math Libraries | Open CL | Hexagon NN

Cores

Kryo CPU | Adreno GPU | Hexagon: Scalar, Vector, Tensor

Qualcomm® Hexagon™, Qualcomm® Adreno™ ,Qualcomm® Kryo™, Qualcomm Artificial Intelligence Engine are products of Qualcomm Technologies Inc and/or its subsidiaries.

AI on the Edge Challenges

The challenge of AI workloads

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies
- Always-on
- Real-time

Power and thermal efficiency are essential for on-device AI

Constrained operating environment

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage / Memory bandwidth limitations
- Software support is scarce, unoptimized and fragmented

Session Goals

ONNX Workshop - Edge Session Goals

Build on discussion output from Beijing workshop

- Edge scenarios/requirements/models analysis, compliance …

Agenda for the next 40 min

- Review and agree on WG logistics (10 min)
- Huawei - ONNX/MLPerf discussion (10 min)
- Present/discuss updated proposal for scope and goals of Edge WG

ONNX *edge*

WG Logistics

Edge WG Chairs

- WG Chair will drive Edge WG through regular meetings
  - Produce meeting notes, action items, etc.

- Edge WG Chairs
  - Milan Oljaca
    - https://github.com/moljacq
  - Ofer Rosenberg
    - https://github.com/OferRosenberg

Principal Engineer/Mgr

QTI AI Software Team

Senior Staff Engineer

QTI AI Software Team

Meetings

- Bi-weekly
  - First meeting proposal: ***Thursday Apr 4, 2019, 8:00 am PST***
  - Will vote on subsequent meeting schedule
- Moderated by chairs
  - Publish clear agendas, meeting minutes, action items, agreements, etc.
- Invites
  - Will post a message to https://gitter.im/onnx/edge with agenda and meeting invite link, ***7 days*** before the meeting
- ONNX Edge WG meeting invite signup: https://github.com/onnx/working-groups/issues/1#issue-424084433
- Meeting/Telepresence tool
  - Zoom

Discussion mechanics

- Communication channel
  - https://gitter.im/onnx/edge

- New github repo for ONNX Edge WG to capture artifacts
  - https://github.com/onnx/working-groups/edge
  - Contribution areas
    - Meeting minutes
    - Documents: recommendations, agreements, etc.

ONNX/MLPerf discussion

Huawei

Edge WG Discussion

ONNX Workshop - Edge Session Discussion topics

- Proposal for scope and goals of Edge WG
  - WG SoW
  - WG action steps
  - Discuss Edge definition

Proposed Edge WG Statement of Work

ONNX governance requires each WG to establish an SoW

Proposed SoW Charter

Promote the usage of ONNX on Edge devices by actively working with various ONNX SIGs, to ensure compatibility and introduce features relevant to execution in this domain (such as quantization), creating a complete end to end specification for edge devices in ONNX.

Propose "ONNX compliance for edge devices" via a subset of ONNX which applies to edge devices, maintaining the semantics of ONNX operators across ONNX targets while introducing a defined subsets of the full ONNX specification applicable to edge devices.

Proposed Approach for ONNX on Edge Devices

- We propose defining ONNX Edge “profile(s)”
  - Strict subsets of the operator space which apply to edge devices
  - Avoid complex operations imposed by limitations on edge devices
    - High computational complexity
    - Use-cases not applicable to the edge (e.g. large scale training scenarios)

- We propose that a set of representative use-cases be identified
  - Ensure that a sufficient core set of operations in edge profile(s) is covered
  - Provide a basis of compliance to edge profile(s)
  - The set will NOT be an exhaustive set of possible edge use-cases

Edge WG Proposed action items

- Define what are the characteristics of an “edge” device

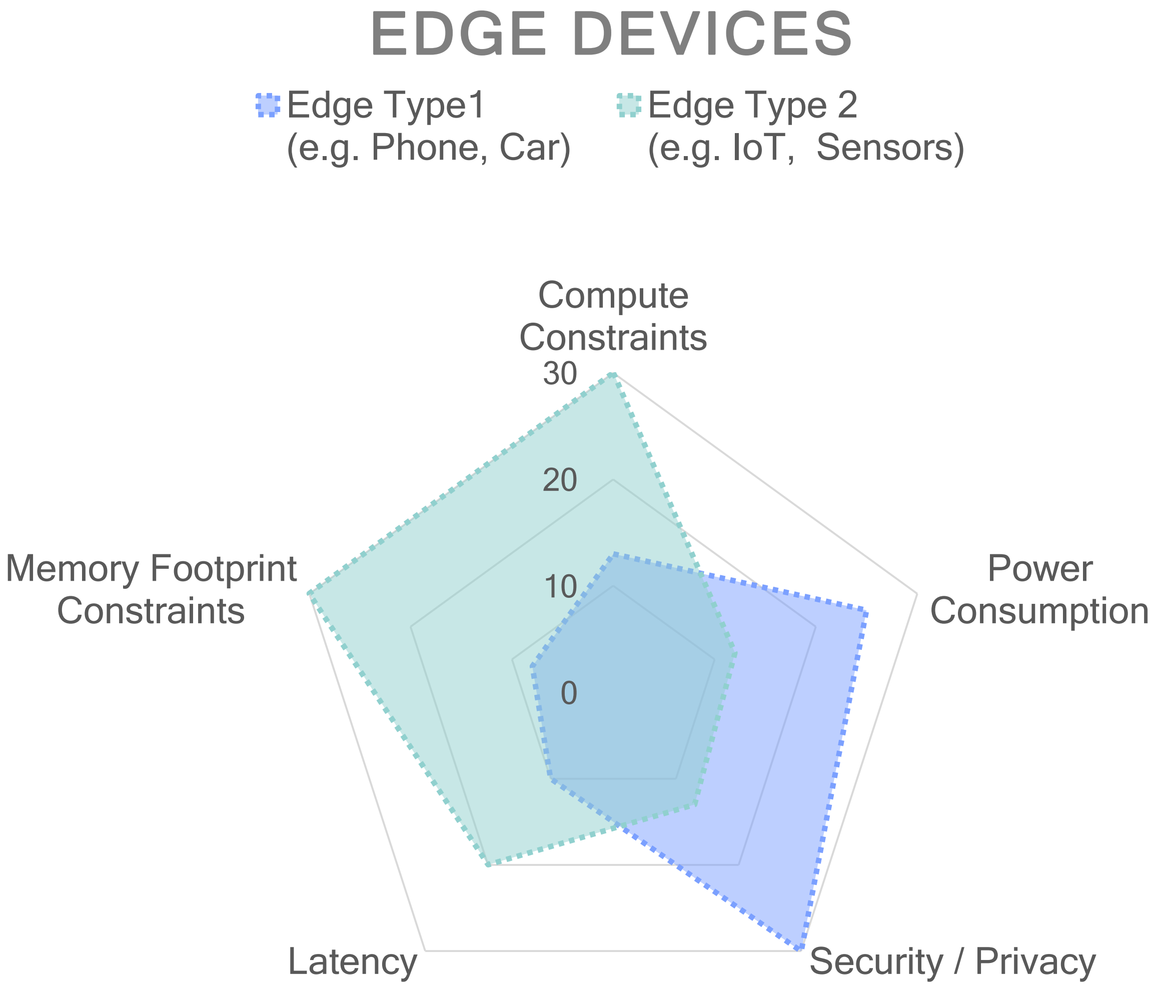
- Define core use-cases for edge devices to benchmark our definition and use as the basis for test-cases to cover compliance

- Define Edge Operation Profile(s)
  - Subset of ONNX operations for Edge use cases

- Define cross-WG dependencies and requirements
  - Compliance
  - Quantization

What is Edge ?

- Compute resource constrains
- Memory footprint constraints
- Power (battery or wall powered)
- Inference latency requirements
- Security/privacy (data locality)
- Specific use cases
- Connected or not



Identifying Edge “key attributes” is important to define requirements for WGs and SIGs

Qualcomm

Thank you!

Follow us on: f 🐦 in

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.