Aug 23, 2019

Santa Clara, CA

ONNX Partners Workshop

Edge Working Group

Contributors

- Milan Oljaca – Qualcomm (co-chair)
- Ofer Rosenberg – Qualcomm (co-chair)
- Yedong Liu – Huawei
- Saurabh Tangri - Intel

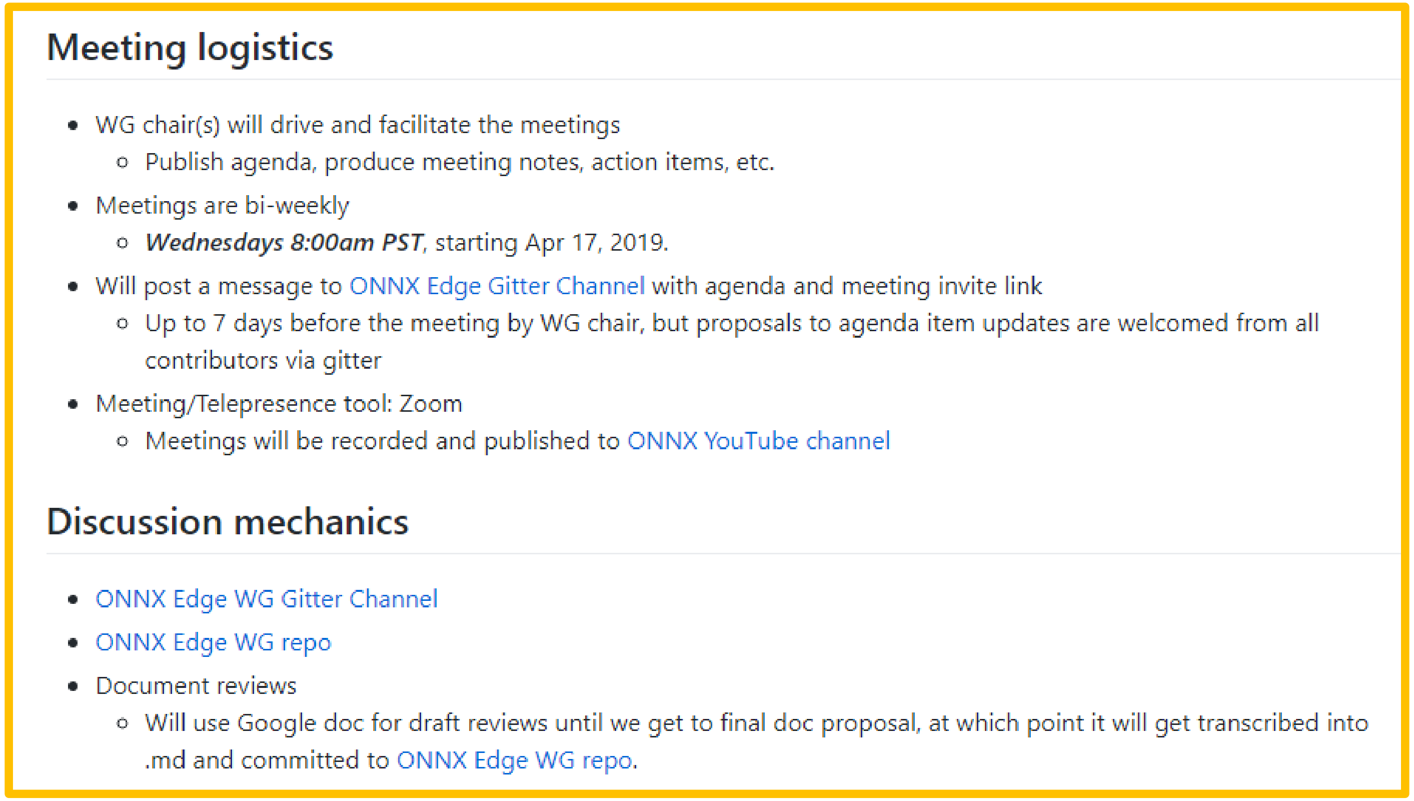
Agenda

- Status update (10 min)
- Breakout session (45 min)

ONNX Edge WG Status Update

Since March 2019 workshop

Edge WG Home



Meeting logistics

- WG chair(s) will drive and facilitate the meetings
  - Publish agenda, produce meeting notes, action items, etc.
- Meetings are bi-weekly
  - ***Wednesdays 8:00am PST***, starting Apr 17, 2019.
- Will post a message to ONNX Edge Gitter Channel with agenda and meeting invite link
  - Up to 7 days before the meeting by WG chair, but proposals to agenda item updates are welcomed from all contributors via gitter
- Meeting/Telepresence tool: Zoom
  - Meetings will be recorded and published to ONNX YouTube channel

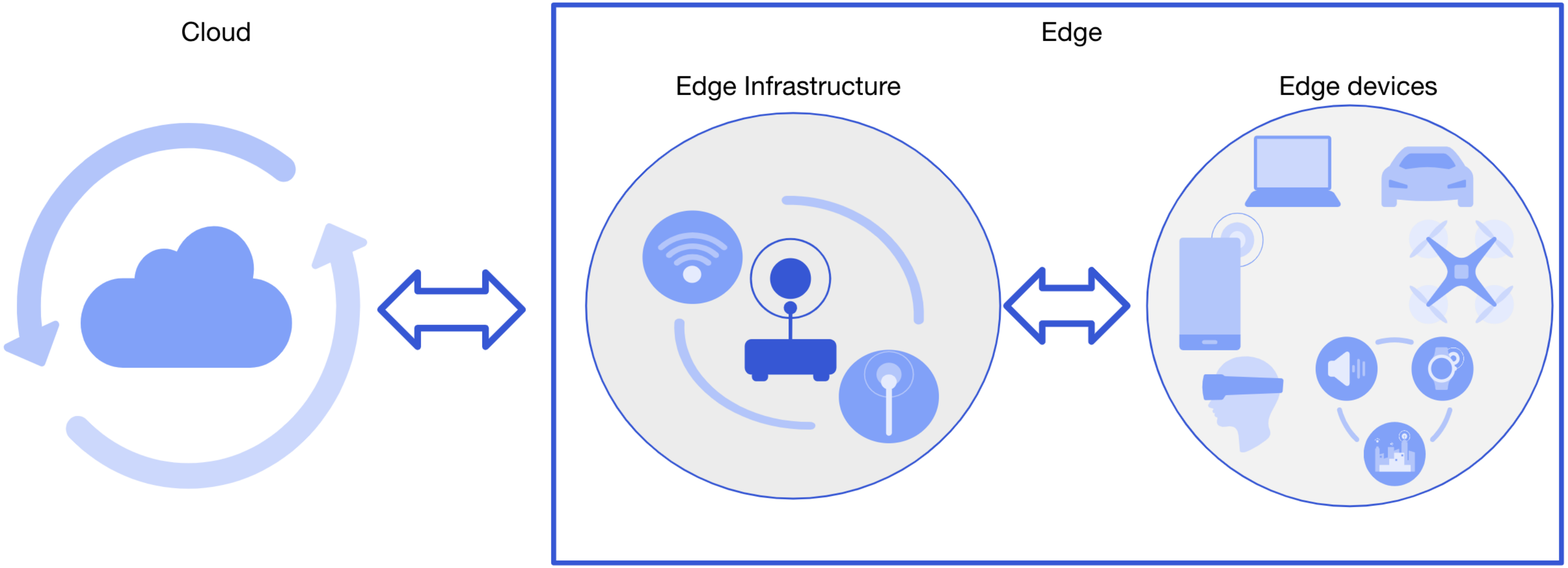
Discussion mechanics

- ONNX Edge WG Gitter Channel
- ONNX Edge WG repo
- Document reviews
  - Will use Google doc for draft reviews until we get to final doc proposal, at which point it will get transcribed into .md and committed to ONNX Edge WG repo.

https://github.com/onnx/working-groups/blob/master/edge/README.md

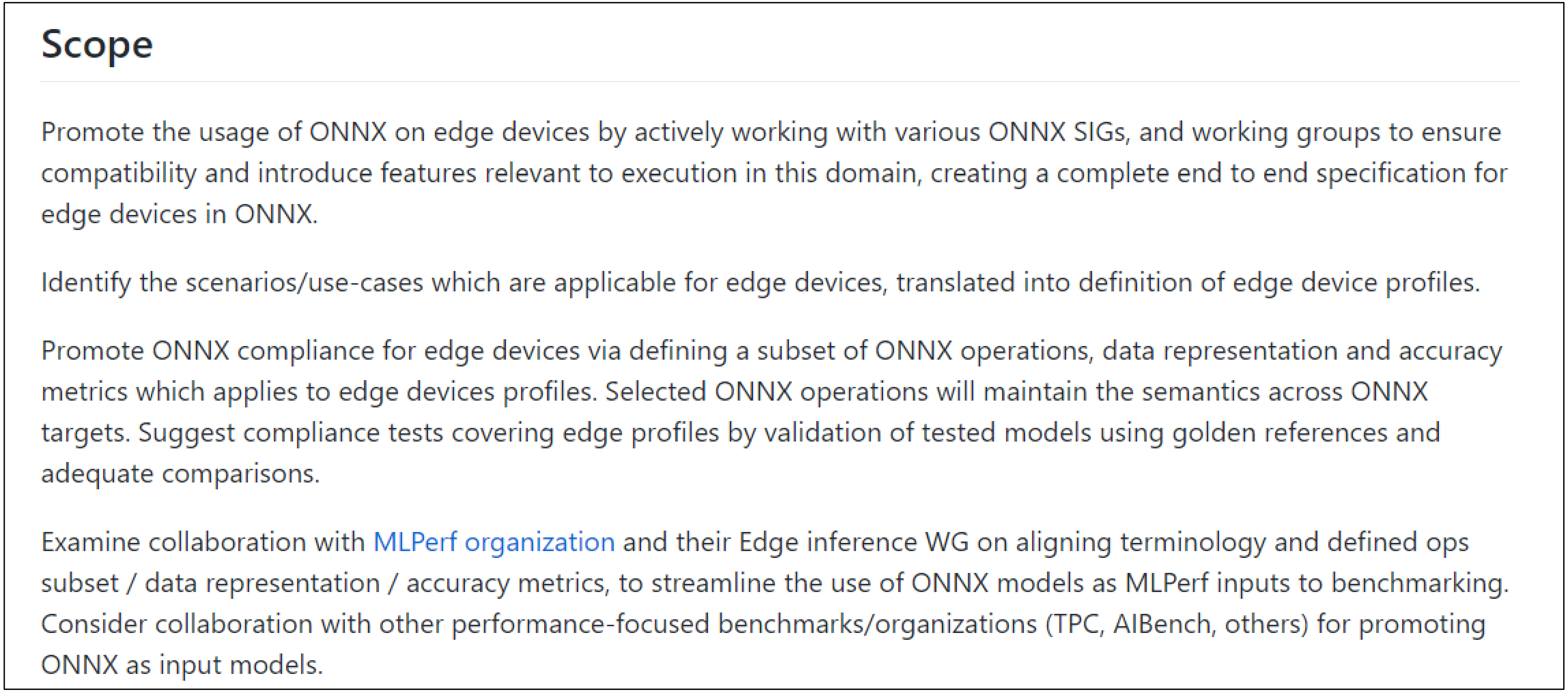
What is “Edge” ?

- The Edge working group defined the scope of “Edge” to contain
  - Edge infrastructure
  - Edge devices



https://github.com/onnx/working-groups/blob/master/edge/artifacts/onnx-edge-scope-and-profile-definition.md

Edge Working group scope (from SoW)



Scope

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

Identify the scenarios/use-cases which are applicable for edge devices, translated into definition of edge device profiles.

Promote ONNX compliance for edge devices via defining a subset of ONNX operations, data representation and accuracy metrics which applies to edge devices profiles. Selected ONNX operations will maintain the semantics across ONNX targets. Suggest compliance tests covering edge profiles by validation of tested models using golden references and adequate comparisons.

Examine collaboration with MLPerf organization and their Edge inference WG on aligning terminology and defined ops subset / data representation / accuracy metrics, to streamline the use of ONNX models as MLPerf inputs to benchmarking. Consider collaboration with other performance-focused benchmarks/organizations (TPC, AIBench, others) for promoting ONNX as input models.

Progress Report

- Working group had 13 meetings so far
  - https://github.com/onnx/working-groups/tree/master/edge/meetings
  - ONNX YouTube channel

- Completed and published 2 documents
  - Statement of Work document
  - Edge scope and profile definition document

- All published documents are in artifacts folder
  - https://github.com/onnx/working-groups/tree/master/edge/artifacts

- Proposals currently under review
  - Prototype profile / profile definition template
  - Profile topology map
  - Profile compliance workflow

Statement of Work (SoW) document

- Defines deliverables and exit criteria
  - Edge profiles and compliance
- Goals and milestones
  - 2019Q2
    - Edge scope and profile definition
  - 2019Q3
    - Definition of specific profiles as identified by WG
  - 2019Q4
    - Definition of profile compliance workflow and content

We are here

…/onnx/working-groups/edge/artifacts/onnx-edge-wg-sow.md

Challenges

- Scope of Edge
  - Everything but cloud
  - Potentially large number of profiles
  - Identifying profiles

- Community participation
  - Modest participation raises a question about relevance of Edge WG
  - The relatively light level of engagement unduly burdens the few participants with quite a bit of work, especially now as edge profiles will need to be established in detail.

Breakout Session

Agenda

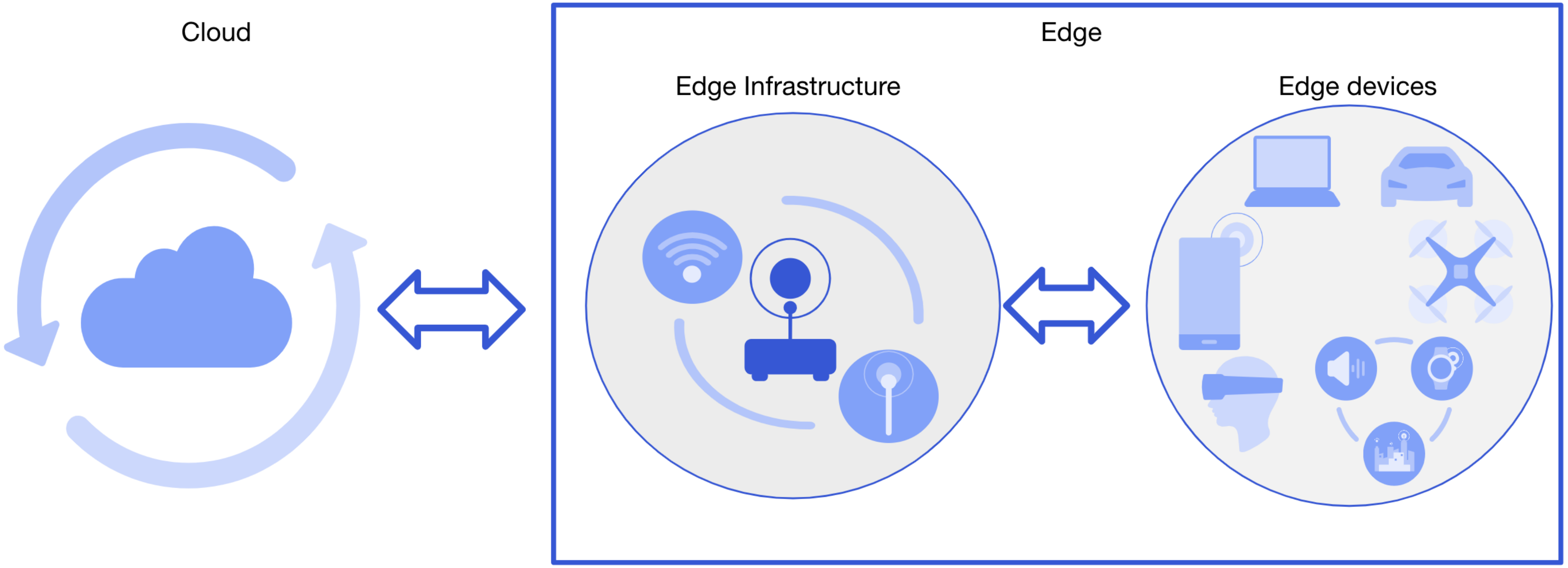
- Edge scope and profile definition - overview
- Proposals under review – discussion
  - Prototype profile / profile definition template (Ofer)
  - Profile topology map (Ofer)
  - Execution Mode (Saurabh)

Edge scope and profile definition document

- An edge profile defines :
  - Attributes
  - Op set
  - Other limitations

- Profile attributes :
  - Accuracy
  - Latency
  - Size
  - Power consumption
  - Data locality

Scope

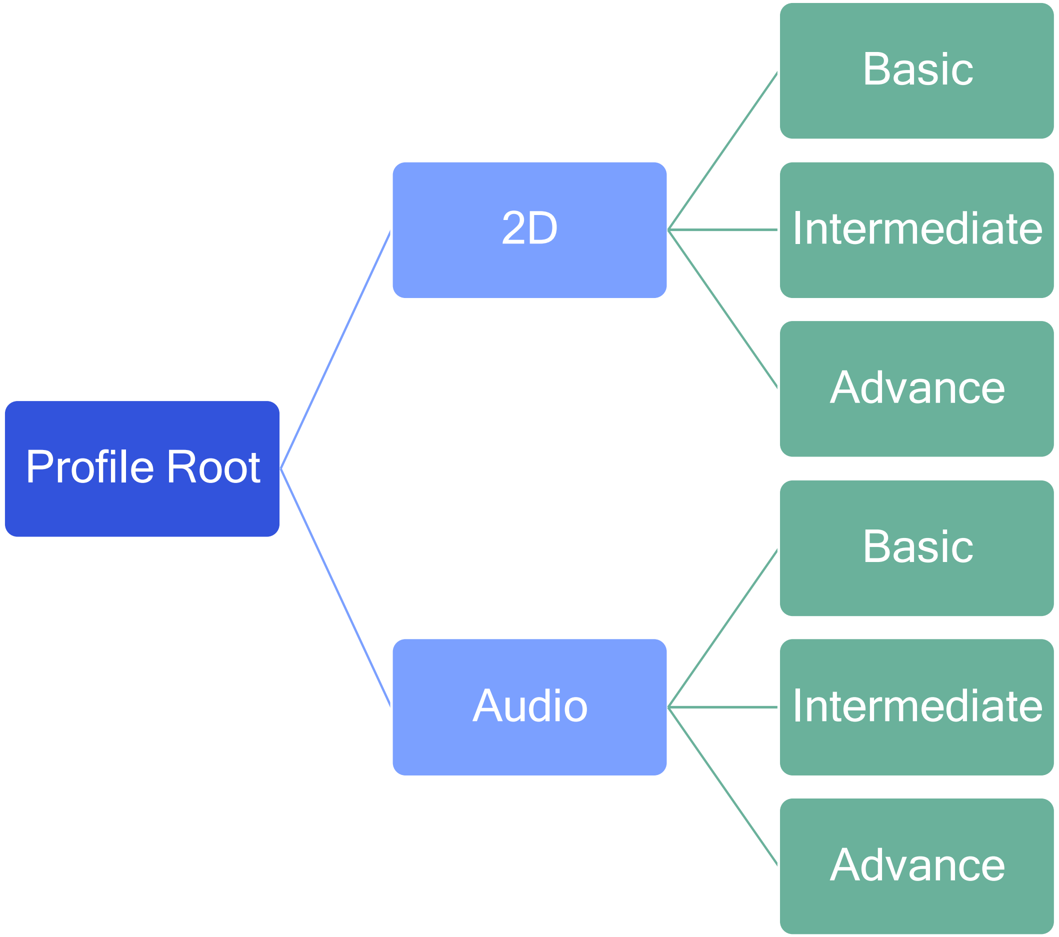


…/onnx/working-groups/edge/artifacts/onnx-edge-scope-and-profile-definition.md

Proposal : Profile Topology Map

- Change the matching of profiles to systems
  - By “system” we refer to edge infra or edge devices

- Original direction : One profile per system
  - For example, “smart speaker profile”
- Proposed direction :
  - Profile defines sensor + level
  - System supports a collection of profiles



Proposal : extend Profiles to describe Execution Mode

- Original purpose/scope of defining profiles is about device compliance/certification
  - Vendor runs “Profile Test Package” from ONNX and certifies the device

- Some devices may support / be certified for a range of profiles
  - Example : PC is certified for 2D_Image Basic, 2D_Image Intermediate & 2D_image Advance

- Proposal : extend profiles to also be used to set “execution mode” for devices
  - One way is to add “execution mode” to the ONNX model as an attribute

- Requires extending the scope of the working group SoW

Readout

ONNX Edge WG - Breakout session readout

- Presented Edge scope and profile definition
  - https://github.com/onnx/working-groups/blob/master/edge/artifacts/onnx-edge-scope-and-profile-definition.md
  - Walk through parts of the document
  - Request feedback on defined Attributes
    - Are these sufficient ?
    - Pointed raised regarding Accuracy – led to follow-up on testing compliance

- Issue clarrifed during the meeting : Profile compliance will be done using test package
  - Package contains inputs and golden outputs for accuracy calculation
  - Edge WG needs to contribute “example packages” for few profiles to the ONNX repo

- New proposal : extend Profiles to define Execution Mode
  - Setting profile as part of model (optionally) – within the scope of ONNX community
  - Query which profiles are supported by the device – requires API (extend the WG work to define it ?)

Thank You!