ONNX Edge WG Update

- Milan Oljaca
- Ofer Rosenberg
- Yedong Liu
- Saurabh Tangri
- Manash Goswami

Progress Report

- No significant progress was made since last workshop
- Several proposals still in review
  - SoW update
  - Execution mode compliance
  - Defining profiles: 3 different ways. Working on a survey to decide on method to use.
- Challenge (“crossroads”)
  - Compliance to static edge profile vs. execution mode compliance
- Next steps
  - Resolve definition of “compliance to the edge profile”
  - Seek guidance from Steering Committee about execution mode compliance and mechanism to query supported profiles
    - These are generally applicable concepts, not strictly Edge specific

Learnings from WG discussions

Goal of Edge Profiles : ONNX to publish a suite of tests that generate a certificate which shows compliance to a predefined execution profile. This is necessary for applications to know the device capabilities.

Profile = {Accuracy, Memory, Latency, Power, Data Locality}

Question : Can a general purpose computing platform which could run AI many ways remain compliant to static definition of profiles all the time?

Question : Can a general purpose computing platform which could run AI many ways remain compliant to static definition of profiles all the time?

General Purpose Computing Platform: This could be edge device such as Phone or a PC which runs many functions simultaneously. Such devices have schedulers in HW or SW managing resources for best user experience. Such devices can also run a model in hetero mode and run subgraphs on different IPs

Accuracy :

A model running on CPU can generate different results than another IP(GPU, DSP, AI Accelerator) because of different numerical and design choices.

On same hardware a runtime may choose a different algorithm after considering operating conditions. Winograd convolutions are faster, bloat memory but impact precision and accuracy.

Memory, Power & Latency:

Depending on available memory and Power a platform may pick a threading policy which impacts both Memory and Latency.

HW can autonomously throttle frequency or power gate the IP running AI. Such thermal and other environment conditions are dynamic and can make the device non-compliant to a static profile.

Question : Can a general purpose computing platform which could run AI many ways remain compliant to static definition of profiles all the time?

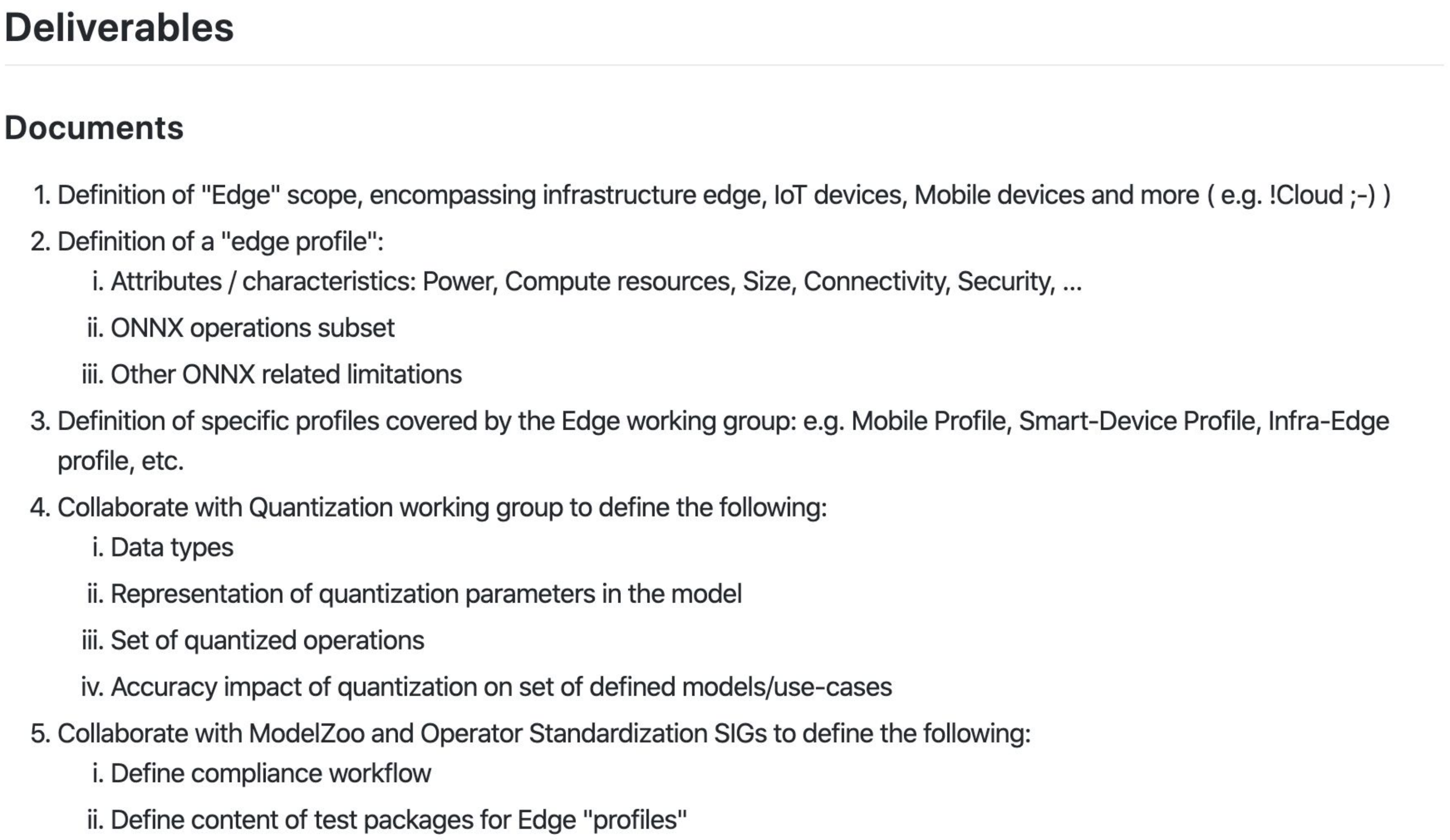
Discussion :

1. Is compliance possible for systems that are dynamically resource rebalancing?
2. Is it ok to have compliance to static profiles be defined within compliance test setup constraints?
3. What is meaning of compliance?
4. Would strict compliance to static profiles discourage adoption or innovation within ONNX ecosystem?

Recommendation:

System needs a hint about the operating environment prior to running a model in a specific profile. WG seeks a guidance on an API or model attribute that can be used to specify it.

Edge WG SoW update



Deliverables

Documents

1. Definition of "Edge" scope, encompassing infrastructure edge, IoT devices, Mobile devices and more (e.g. !Cloud ;-))
2. Definition of a "edge profile":
   i. Attributes / characteristics: Power, Compute resources, Size, Connectivity, Security, ...
   ii. ONNX operations subset
   iii. Other ONNX related limitations
3. Definition of specific profiles covered by the Edge working group: e.g. Mobile Profile, Smart-Device Profile, Infra-Edge profile, etc.
4. Collaborate with Quantization working group to define the following:
   i. Data types
   ii. Representation of quantization parameters in the model
   iii. Set of quantized operations
   iv. Accuracy impact of quantization on set of defined models/use-cases
5. Collaborate with ModelZoo and Operator Standardization SIGs to define the following:
   i. Define compliance workflow
   ii. Define content of test packages for Edge "profiles"

We are here

Edge WG SoW update

- Identified the following issues in deliveries :
  - “Definition of specific profiles covered by the Edge working group” needs to be updated to reflect the new approach, of multiple profiles per device
  - “Collaborate with Quantization working group to define the following” needs to change to "collaborate with Operator SIG", as there's no quantization WG
  - “Collaborate with ModelZoo and Operator Standardization SIGs to define the following” - currently no active ModelZoo SIG. Put on hold
  - Add a new delivery : creating a document (white paper) describing the different options for sets of profiles, with initial suggested set per option. See more details below.

- Adjust goals as follows :
  - 2019Q3 - write the white paper described above, do a survey to select between options, and decide on an initial set

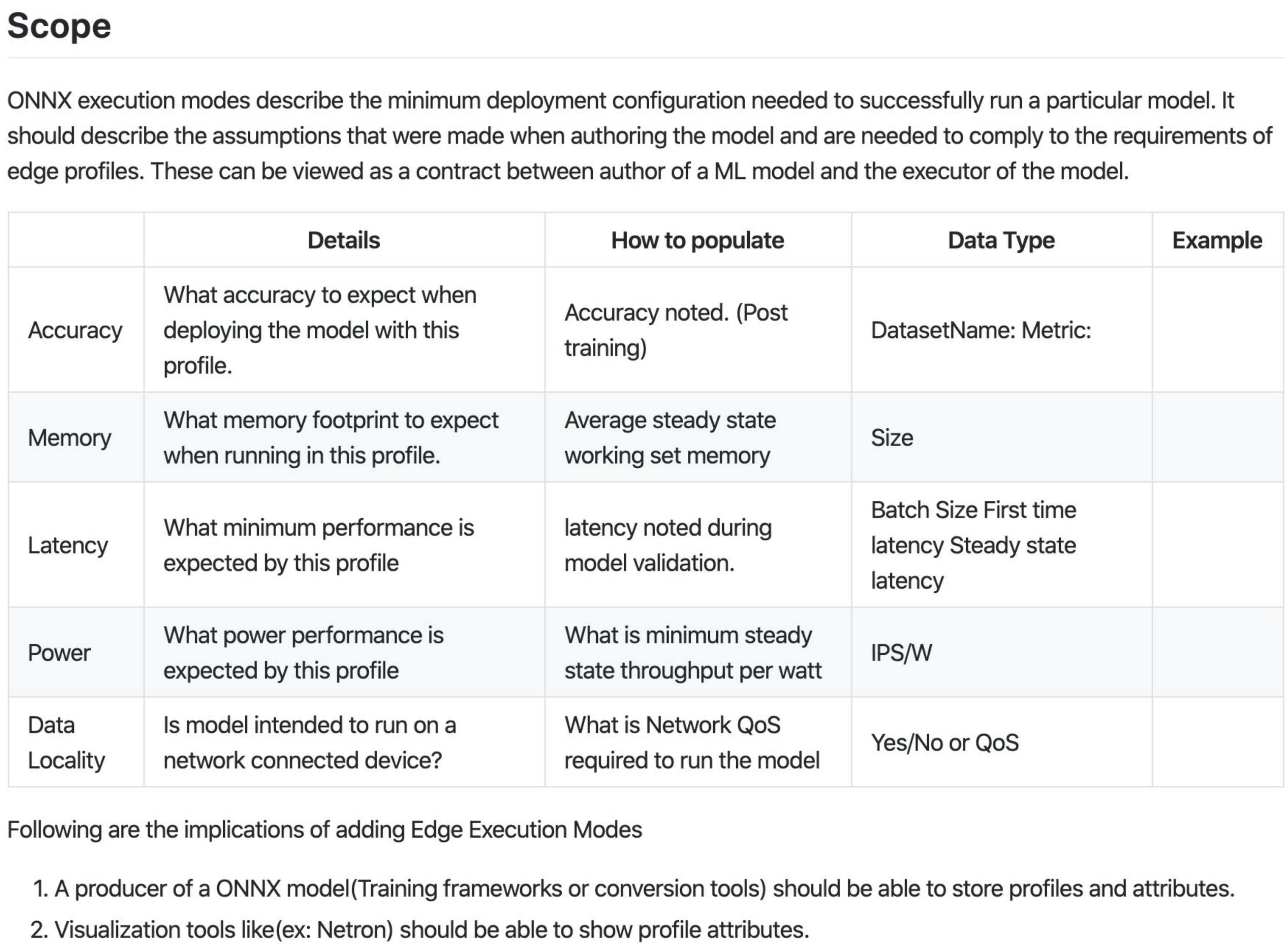
ONNX Edge Execution Mode Proposal

Background

- Edge devices can be compliant to single profile, or to a few profiles . A fixed function device such as a smart speaker may be compliant to a single edge profile. Other multi-functional edge devices (for example ,a PC) may be compliant to a range of edge profiles. In this case, the meaning of edge profile can be extended and viewed as defining execution mode for a given loaded model.

- At times, when such a device operates multiple ONNX models simultaneously, where each model was loaded with a defined execution mode, the execution mode attributes provide hints to the platform. This establishes an execution contract between the application and the underlying platform which guarantees Quality of Service.

ONNX Edge Execution Mode Proposal



Scope

ONNX execution modes describe the minimum deployment configuration needed to successfully run a particular model. It should describe the assumptions that were made when authoring the model and are needed to comply to the requirements of edge profiles. These can be viewed as a contract between author of a ML model and the executor of the model.

	Details	How to populate	Data Type	Example
Accuracy	What accuracy to expect when deploying the model with this profile.	Accuracy noted. (Post training)	DatasetName: Metric:	
Memory	What memory footprint to expect when running in this profile.	Average steady state working set memory	Size	
Latency	What minimum performance is expected by this profile	latency noted during model validation.	Batch Size First time latency Steady state latency	
Power	What power performance is expected by this profile	What is minimum steady state throughput per watt	IPS/W	
Data Locality	Is model intended to run on a network connected device?	What is Network QoS required to run the model	Yes/No or QoS	

Following are the implications of adding Edge Execution Modes

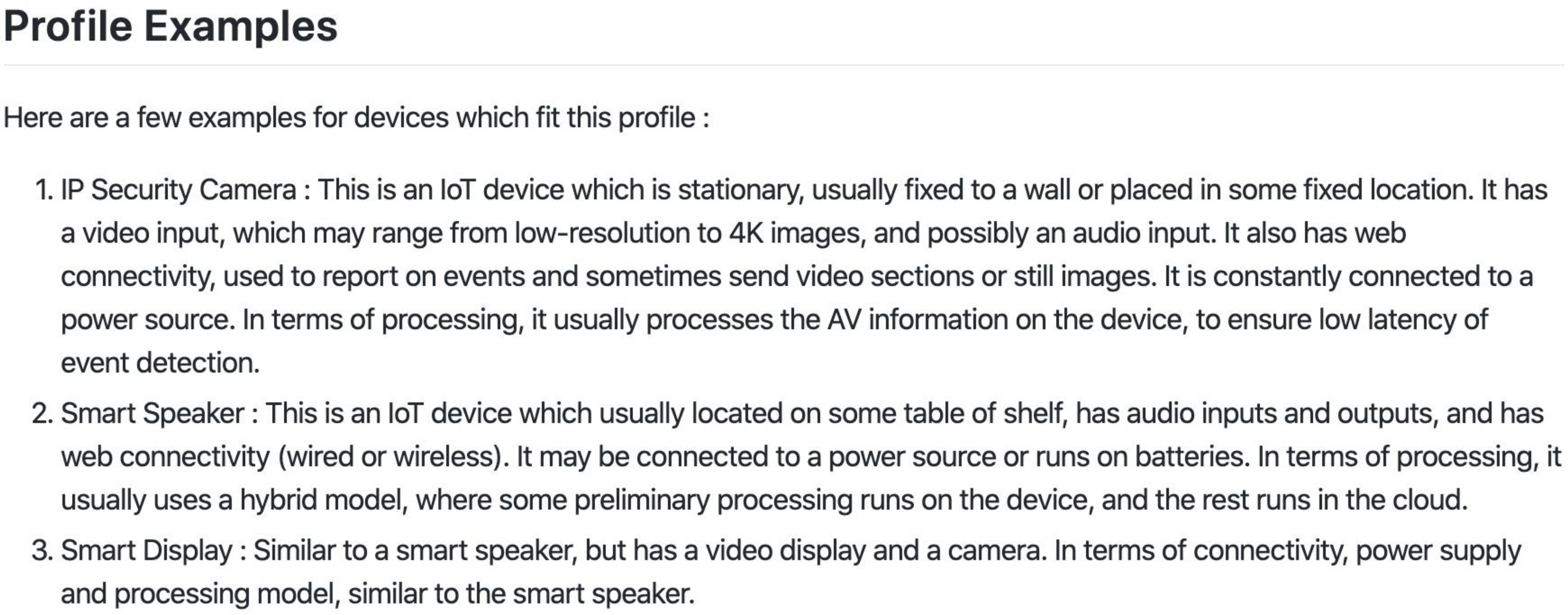
1. A producer of a ONNX model(Training frameworks or conversion tools) should be able to store profiles and attributes.
2. Visualization tools like(ex: Netron) should be able to show profile attributes.

ONNX Stationary IoT device with AV profile prototype

Introduction:

- The profile described in this document is of a stationary IoT device with Audio/Video interface (input and output). The family of edge devices covered by this profile are devices which remain in the same location (fixed to a wall, lying on a table, etc), constantly connected to the web (wired or wireless) and have audio and video interfaces - inputs, outputs or both. The next section provides a few examples for devices which fit into this profile.

ONNX Stationary IoT device with AV profile prototype



Profile Examples

Here are a few examples for devices which fit this profile :

1. IP Security Camera : This is an IoT device which is stationary, usually fixed to a wall or placed in some fixed location. It has a video input, which may range from low-resolution to 4K images, and possibly an audio input. It also has web connectivity, used to report on events and sometimes send video sections or still images. It is constantly connected to a power source. In terms of processing, it usually processes the AV information on the device, to ensure low latency of event detection.
2. Smart Speaker : This is an IoT device which usually located on some table of shelf, has audio inputs and outputs, and has web connectivity (wired or wireless). It may be connected to a power source or runs on batteries. In terms of processing, it usually uses a hybrid model, where some preliminary processing runs on the device, and the rest runs in the cloud.
3. Smart Display : Similar to a smart speaker, but has a video display and a camera. In terms of connectivity, power supply and processing model, similar to the smart speaker.

Defining profiles

A few different ways to define profiles :

- Initial direction : one profile per device (like "smart speaker") - suggestion is to deprecate this approach
- Based on "sensor" type (Video, Audio, Speech) and level (Low, Mid, High)
- Based on Network type (Classification, Segmentation, NLP) and level

Applying/using the Profiles

Profiles can be provided to the runtime/backends in few ways :

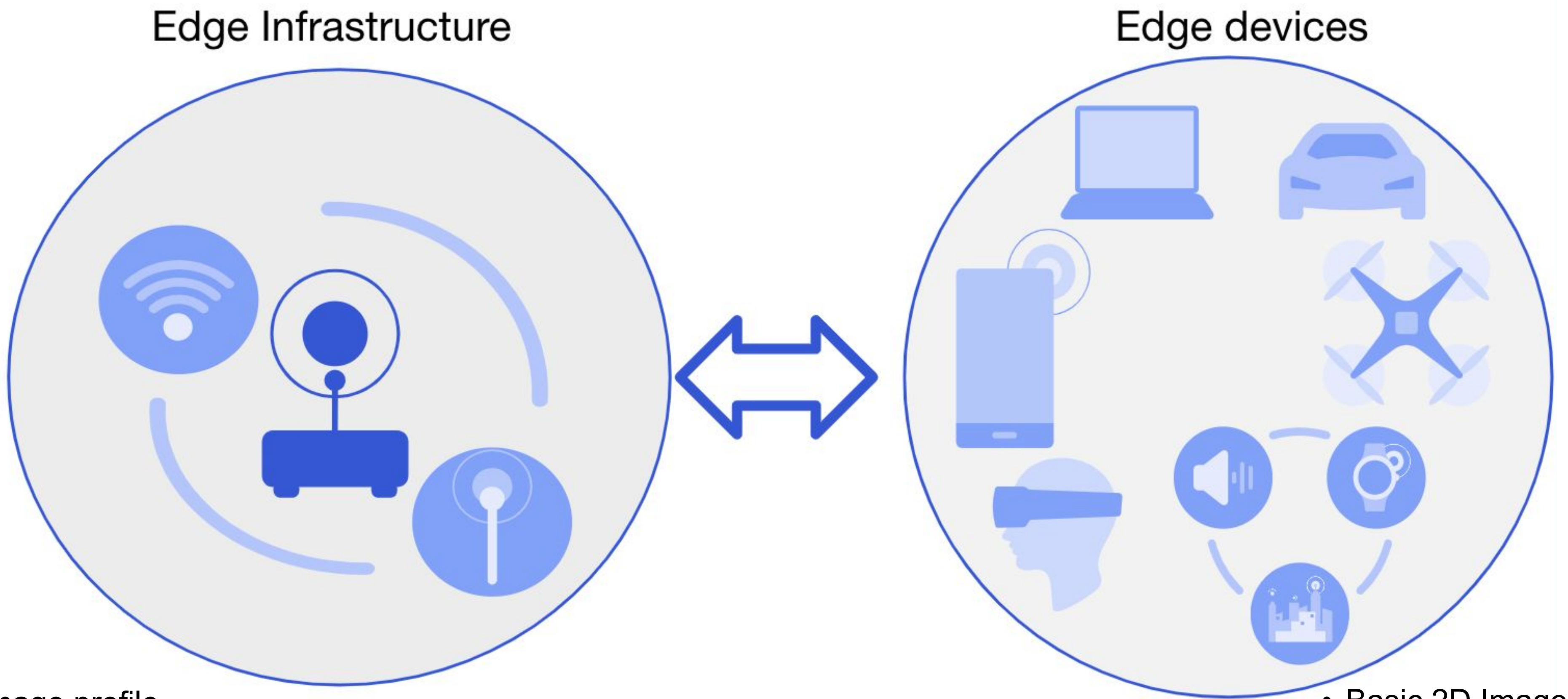
- A new optional field in the ONNX container.
  - Not mandatory, as it may be hard for converter to support it initially - the App/Runtime may come and "edit" the container to add a profile before sending it to the backend
  - To implement it, converters will need to know what each profile means to examine the model and fill the value
- A sideband API to specify the profile

ONNX Profile Topology Map

- Introduction:

This document defines a hierarchical classification of ONNX edge profiles, i.e. a profile topology map. Intent with such profile breakdown is to establish categorization of edge profiles based on use cases and profile attributes identified in onnx edge scope and profile definition.md document.

ONNX Profile Topology Map



Edge Infrastructure

Edge devices

- Basic 2D Image profile
  - Computer Vision
- Advanced 2D Image Profile
  - Computer Vision: high resolution, low latency
- Basic Audio Profile
  - Speech Recognition
- Advanced Audio Profile
  - NLP, Translation

- Basic 2D Image profile
  - Computer Vision
- Advanced 2D Image Profile
  - Computer Vision: data locality
- Basic Audio Profile
  - Speech Recognition
- Advanced Audio Profile
  - NLP, Translation

Future Work

- Update SoW based on discussion
- Propose initial draft for profile sets white paper, and work out survey
- etc.

Thank you!