intel Software

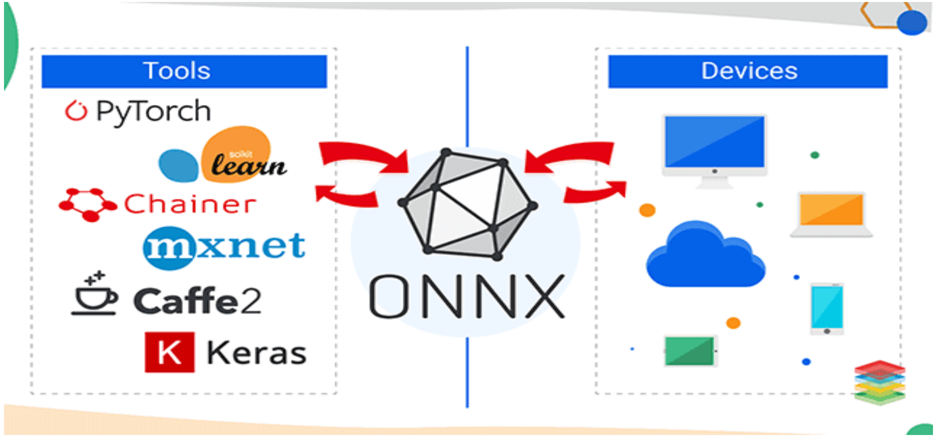
ACCELERATE AI INFERENCE FROM CLOUD TO EDGE WITH ONNX RUNTIME AND INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

October 2019

ONNX Runtime integration

ONNX Introduction

- Open Neural Network Exchange Format
- Framework interoperability: Train in one framework and run inference in another framework



[Source: XenonStack]

ONNX Runtime

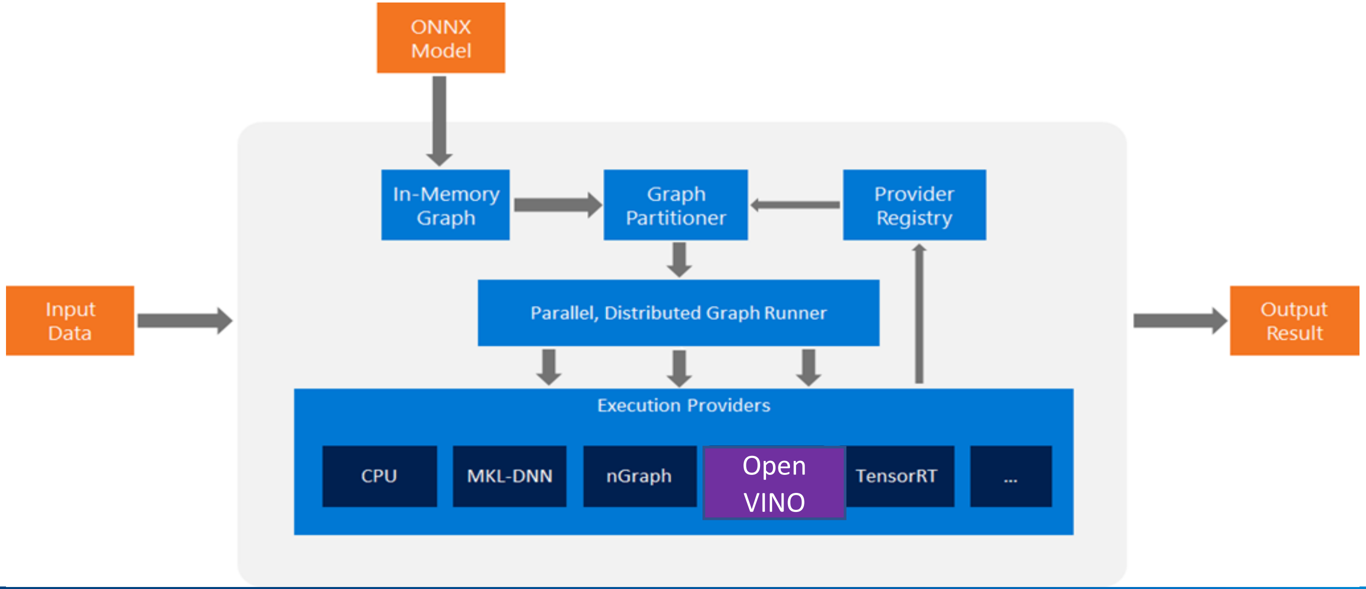
- ML/DL Inferencing framework by Microsoft
- Built specifically for ONNX format models
- Supports execution on multiple hardware backends
- Completely open-source development on Github

Simple API

```
import onnxruntime


session = onnxruntime.InferenceSession(“model.onnx”)

x = GetInputData()

y = session.run([session.get_outputs()[0].name],

                {session.get_inputs()[0].name : x})
```

ONNX Runtime Architecture

- Modular architecture : can plug-in multiple hardware backends
- Each hardware backend managed by its ‘Execution Provider’ (EP)
- Partitions graph into subgraphs that can be scheduled on different EPs.
- Current Intel EPs: MKLDNN, nGraph & OpenVINO



ONNX Model

In-Memory Graph

Graph Partitioner

Provider Registry

Input Data

Parallel, Distributed Graph Runner

Output Result

Execution Providers

CPU

MKL-DNN

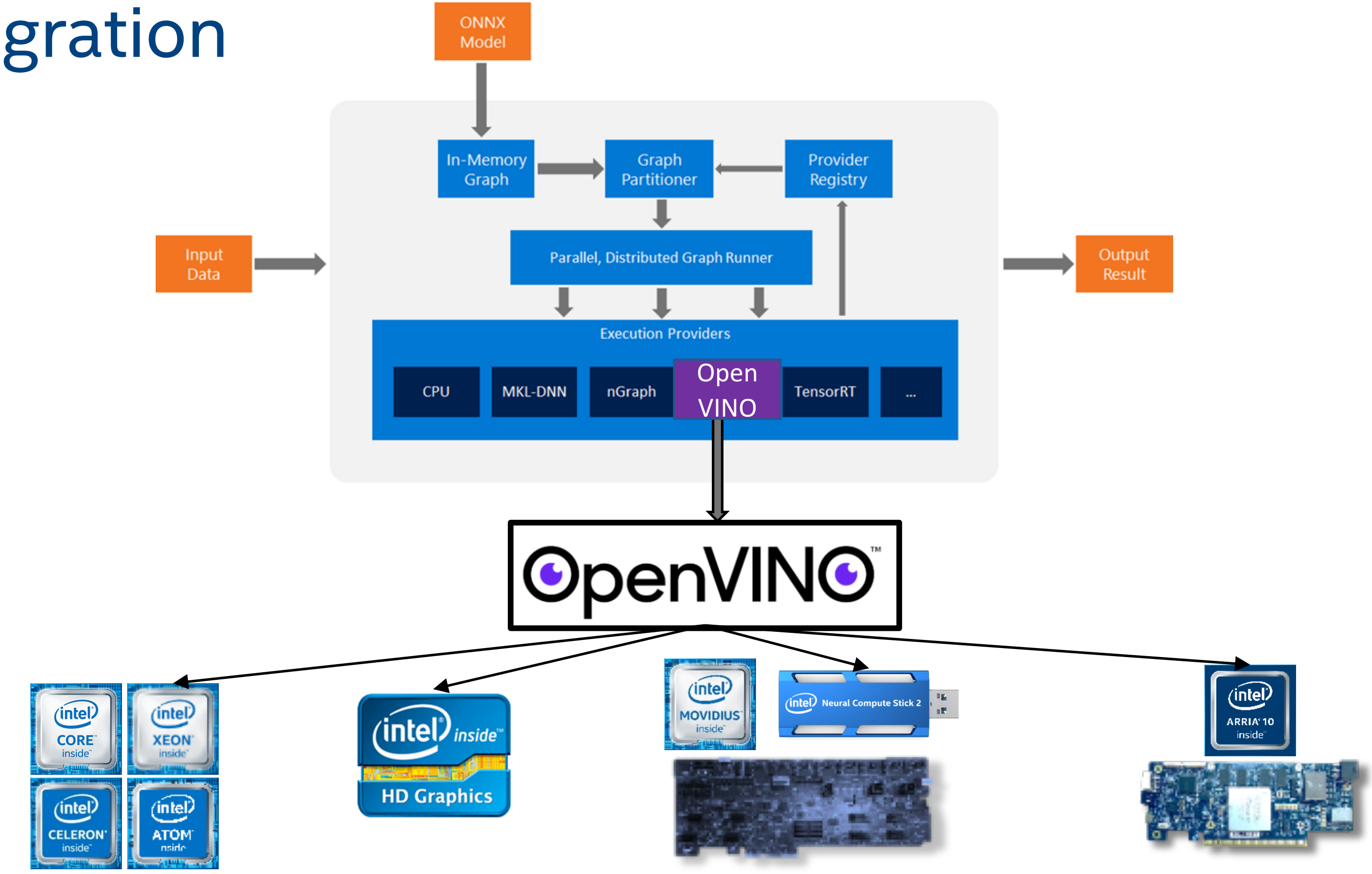
nGraph

Open VINO

TensorRT

...

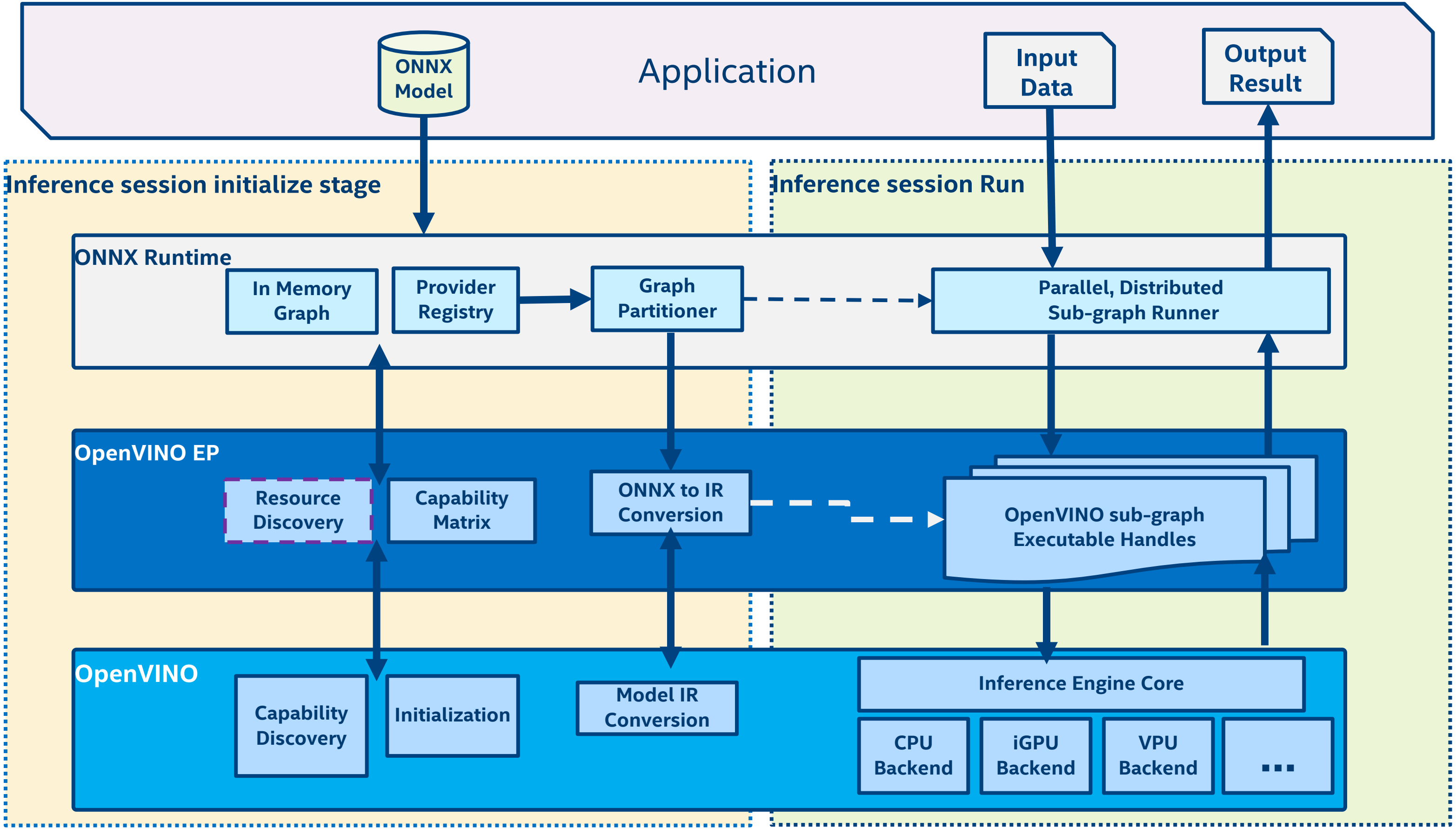
Integration



Optimization Notice

CPU iGPU VPU FPGA

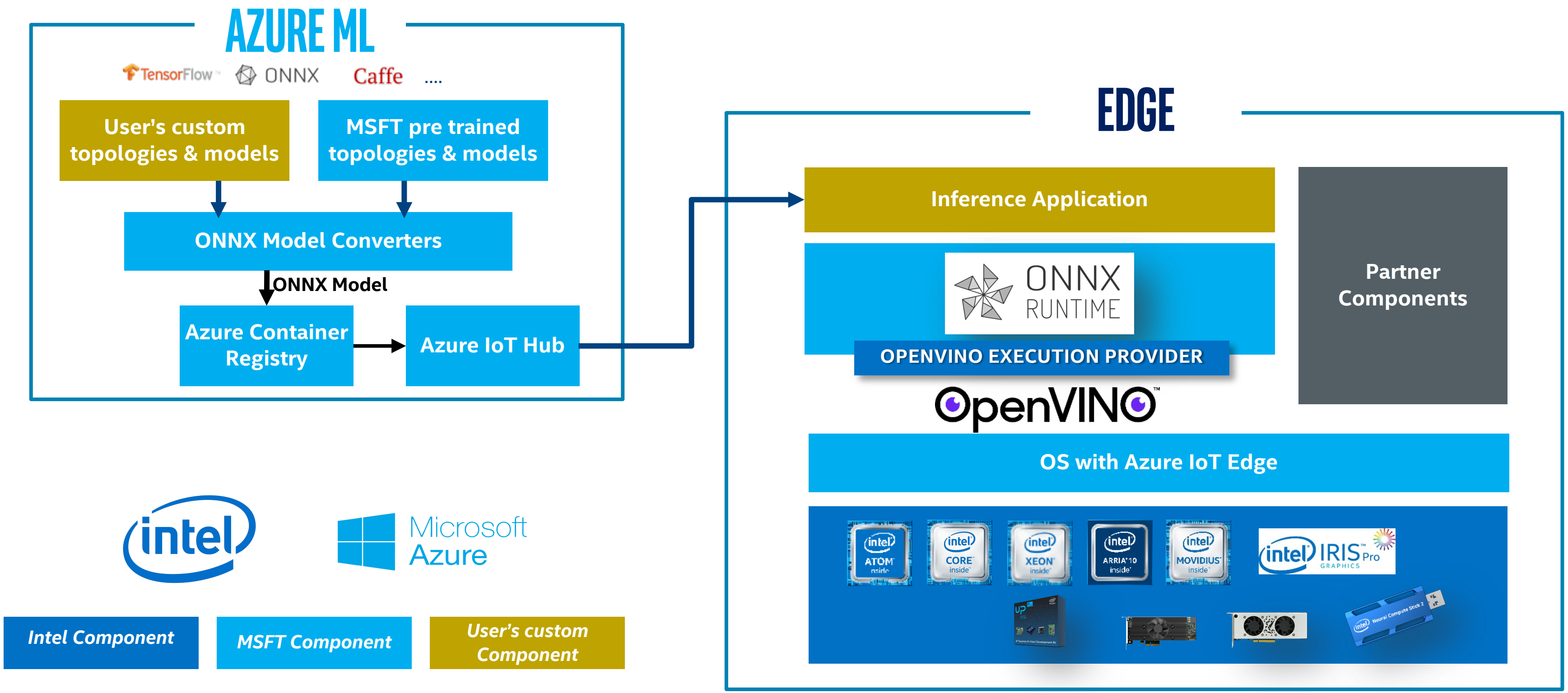
ONNX Runtime: OpenVINO Execution Provider



Optimization Notice

Azure ML integration

SEAMLESS WORKFLOW FOR AZURE ML DEVELOPERS @ EDGE



AZURE ML

TensorFlow ONNX Caffe ....

User's custom topologies & models

MSFT pre trained topologies & models

ONNX Model Converters

ONNX Model

Azure Container Registry

Azure IoT Hub

EDGE

Inference Application

ONNX RUNTIME

OPENVINO EXECUTION PROVIDER

OpenVINO™

Partner Components

OS with Azure IoT Edge

Intel Component

MSFT Component

User's custom Component

Optimization Notice

Existing feature set

- Accelerator support: CPU, iGPU, MyriadX VPU (USB and embedded), Vision Accelerator Design VPU (2x, 4x & 8x MyriadX VPU), Vision Accelerator Design with Arria 10 FPGA
- Quantization support: Full precision (32 bit) and Half precision (16 bit) floating point
- Operator coverage: majority models from ONNX Model Zoo github.com/onnx/models
- OS Support: Linux and Windows
- Docker container support: Linux only
- Azure ML integration: Train model on Azure ML and deploy on connected edge devices

Feature roadmap

- Addl. Quantization formats: 8-bit Int support
- New Features: Hetero and multi-device plugin
- OpenVINO version support : Support for major OpenVINO releases (recurring)
- Latest ONNX operator coverage: support for updated ONNX operators (recurring)
- Docker container support: Windows OS
- Auto resource discovery: detect hardware accelerators on platform
- Latest hardware accelerator coverage (recurring)

Resources

Pre-Requisites:

Please make sure you download the Intel Distribution of OpenVINO (IDOO) installer(tgz) before building the Docker images.*
Build ONNX RT pre-requisites for OpenVINO: https://github.com/microsoft/onnxruntime/blob/master/BUILD.md

- ONNX Runtime with OpenVINO EP along with Azure IoT edge dependencies Docker versions. (use this with Azure IoT Edge). OpenVINO R1.1: (Pre-Req IDOO*)
  Please refer to README and Docker files at: https://github.com/intel/Edge-Analytics-FaaS/tree/R1_2019/Azure-IoT-Edge/OnnxRuntime

- ONNX Runtime with OpenVINO Execution Provider(EP) and OpenVINO R1.1: (Pre-Req IDOO)*)
  (This is the native installation with OpenVINO EP. Use this with your own orchestration framework)
  Please check the following link for README.MD and Dockerfile.openvino:
  https://github.com/microsoft/onnxruntime/blob/master/dockerfiles/Dockerfile.openvino

- To build from source: https://github.com/microsoft/onnxruntime
  (needed only if you want to re-validate or rebuild. Still need Intel Distribution of OpenVINO downloaded)

- Cloud to Edge Deployment flow using Azure ML and Azure IoT Edge. Link to JuPyter Notebook:
  https://aka.ms/onnxruntime-openvino & https://aka.ms/onnx-openvino

- Documentation on Execution Providers: https://github.com/microsoft/onnxruntime/tree/master/docs/execution_providers
  Other Links: Azure IoT Edge: https://azure.microsoft.com/en-us/services/iot-edge/
  ONNX Runtime : https://azure.microsoft.com/en-us/blog/onnx-runtime-is-now-open-source/

Optimization Notice

Support

Process to request ONNX Runtime + OpenVINO EP support:

- All software issues related to ONNX Runtime with OpenVINO EP code should be logged at: "Issues" Tab @https://github.com/Microsoft/onnxruntime with [OpenVINO-EP] tag.
- All Hardware related issues should be routed towards Intel FAEs or the hardware ODM.
- Link to supported models on OpenVINO EP: https://github.com/microsoft/onnxruntime/blob/master/docs/execution_providers/OpenVINO-ExecutionProvider.md
  All issues related to these models should be routed towards Intel FAEs.
- OpenVINO issues should be reported through OpenVINO Forum
- All other model issues should be logged at:
  "Issues" Tab @https://github.com/Microsoft/onnxruntime

intel®

experience
what's inside™