ACCELERATE AI INFERENCE FROM CLOUD TO EDGE

WITH ONNX* RUNTIME + OPENVINO™ TOOLKIT

DATE: 01/26/20 – PUBLIC, OK FOR NON-NDA DISCLOSURE

NOTICES AND DISCLAIMER

AGENDA

- Challenges in cloud to edge AI deployment today
- Benefits of cloud computing and AI cloud computing
- ONNX* Exchange & ONNX* Runtime Value and Benefits
- Getting Started
- Developer Kits, Use Cases, & Case Studies
- Support and Resources

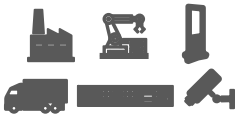
BENEFITS OF CLOUD DEVELOPMENT

CLOUD

Benefits of Cloud Scale Development

- Large scale parallel computing for training
- Many frameworks supported
- Training development made easy
  - No software downloads
  - No configuration
  - No Installations

BENEFITS OF EDGE INFERENCE



EDGE DEVICE/THINGS

Benefits of Edge Inference

- Near real time decision making, close to or at the edge
- Avoid significant data transfer costs to cloud
- Heterogeneous hardware architectures
  - Existing edge device compute
  - Many new hardware platforms with new features

CHALLENGE

DEPLOYING MODELS DEVELOPED IN THE CLOUD TO THE EDGE

BENEFIT OF CLOUD TO EDGE DEPLOYED AI

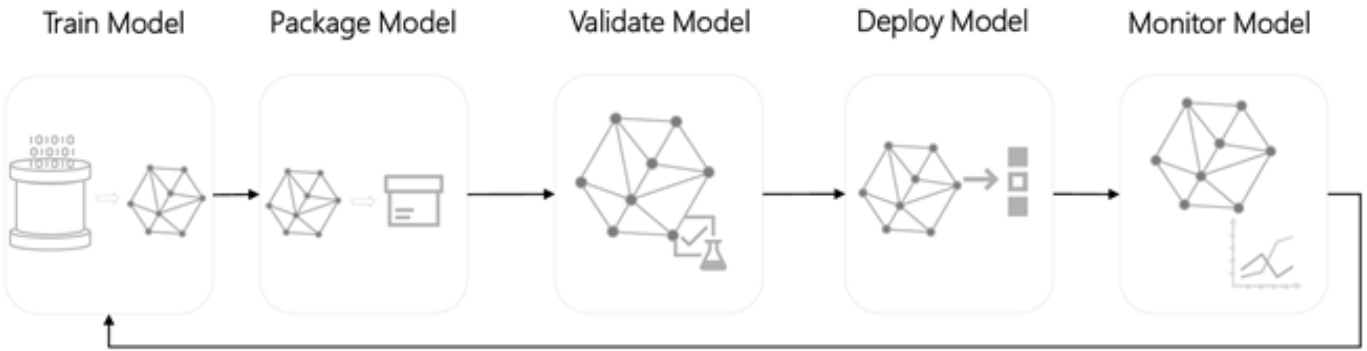
Develop and Train Model
with reusable machine learning pipelines

Package Model
using containers to capture runtime dependencies for inference

Validate Model
behavior responsiveness and regulatory compliance

Deploy Model
to edge target for real-time, streaming, or batch processing

Monitor Model
behavior and business value, replace/deprecate when model is stale



Source: Microsoft

ONNX* RUNTIME + OPENVINO™ TOOLKIT VALUE AND BENEFITS

ONNX* EXCHANGE

ONNX

What it is

ONNX is an open ecosystem that empowers AI developers to choose the right tools as their project evolves. ONNX provides an open source format for AI models, both deep learning and traditional ML. It defines an extensible computation graph model, as well as definitions of built-in operators and standard data types. ONNX is currently focused on the capabilities needed for inferencing (scoring).

Target audience

- Computer vision, machine learning and deep learning software developers
- Data scientists
- OEMs, ISVs, System Integrators

Usages

AI for robotics, retail, healthcare, security surveillance, office automation, transportation, non-vision use cases (speech, NLP, Audio, text) & more.

AI FRAMEWORK INTEROPERABILITY – COMMON FORMAT

TOOLS TO CONVERT MODEL FORMATS TO ONNX

MODEL CATALOG THROUGH ONNX MODEL ZOO

STREAMLINING PATH FROM PROTOTYPE TO PRODUCTION

Homepage ▸ onnx.ai

Github ▸ github.com/onnx/onnx

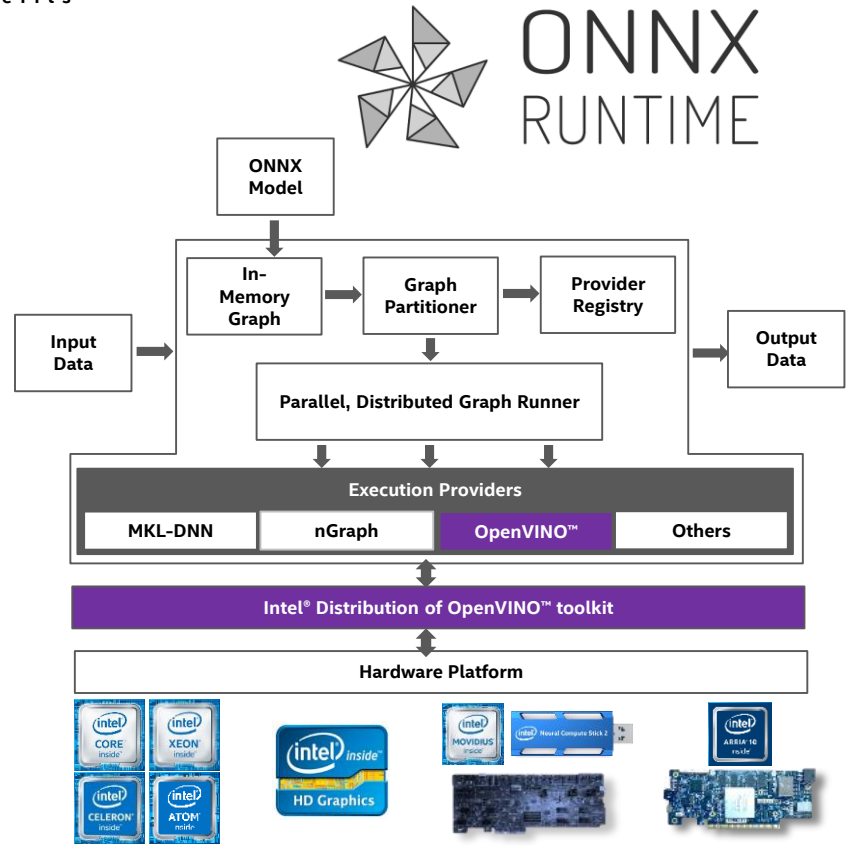
ONNX Model Zoo Github ▸ https://github.com/onnx/models

ONNX* RUNTIME

What it is

ONNX Runtime is a performance-focused complete scoring engine for Open Neural Network Exchange (ONNX) models, with an open extensible architecture to continually address the latest developments in AI and Deep Learning. ONNX Runtime stays up to date with the ONNX standard and supports all operators from the ONNX v1.2+ spec with both forwards and backwards compatibility. Execution Provider plugin allows the support of ONNX RT for Intel® Distribution of OpenVINO™ toolkit.

- BUILT SPECIFICALLY FOR ONNX FORMAT MODELS
- SUPPORTS EXECUTION ON MANY TYPES OF HARDWARE
- COMPLETELY OPEN SOURCED ON GITHUB



Homepage ▸ onnx.ai

Github ▸ github.com/Microsoft/onnxruntime

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

What it is

A toolkit to accelerate development of **high performance computer vision** & **deep learning inference into vision/AI applications** used from edge to cloud. It enables deep learning on hardware accelerators and easy deployment across multiple types of Intel® platforms.

Target audience

- Computer vision, deep learning software developers
- Data scientists
- OEMs, ISVs, System Integrators

Usages

AI for robotics, retail, healthcare, security surveillance, office automation, transportation, non-vision use cases (speech, NLP, Audio, text) & more.

OpenVINO™

HIGH PERFORMANCE, PERFORM AI AT THE EDGE

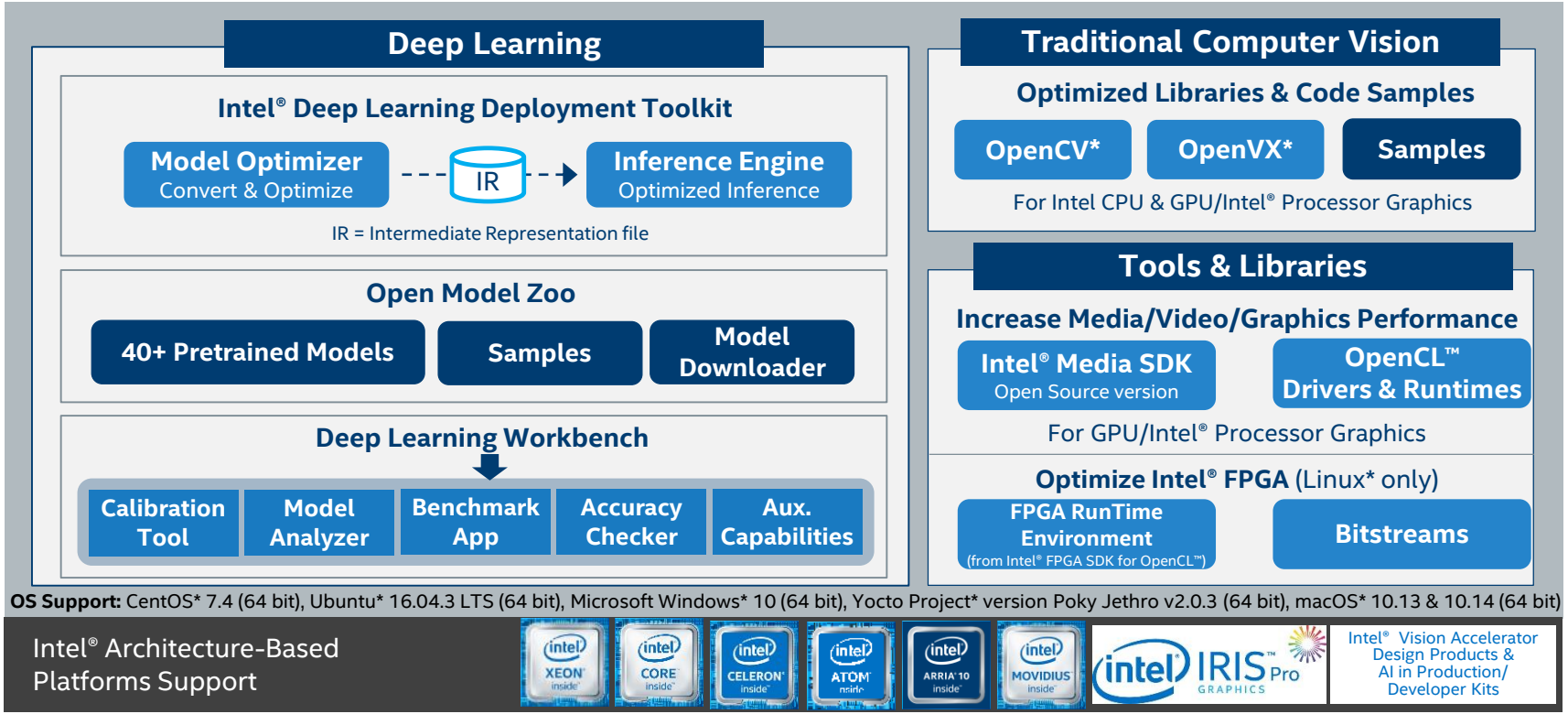
STREAMLINED & OPTIMIZED DEEP LEARNING INFERENCE

HETEROGENEOUS, CROSS-PLATFORM FLEXIBILITY

Free Download ▸ software.intel.com/openvino-toolkit

Open Source version ▸ 01.org/openvinotoolkit

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT



An open source version is available at 01.org/openvinotoolkit (deep learning functions support for Intel CPU/GPU/NCS/GNA).

VALUE PROPOSITION: ONNX RT + OPENVINO™ TOOLKIT INTEGRATION

COMPLETE AZURE ML (CLOUD) TO EDGE INTEGRATION

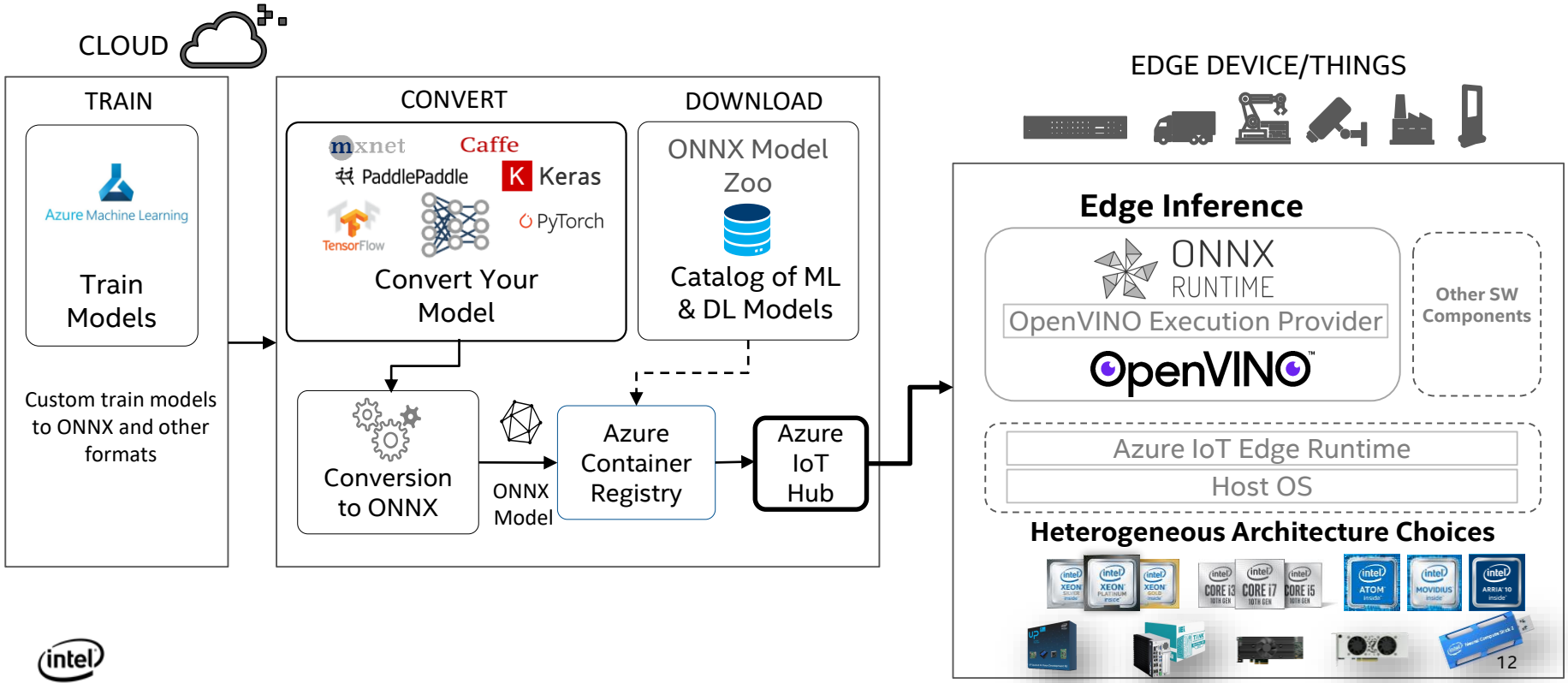
POSITIONED FOR THE EDGE WITH COMBINED BENEFITS OF ONNX* RT & OPENVINO™

HETEROGENEOUS, CROSS-PLATFORM FLEXIBILITY

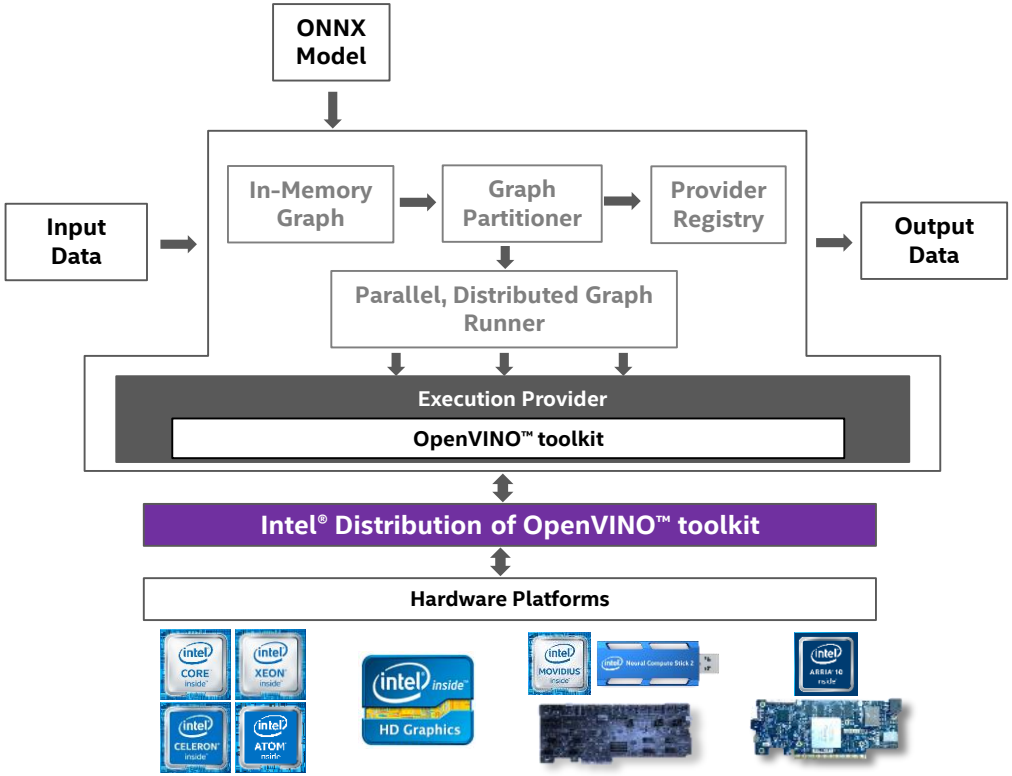
DEVELOPER FRIENDLY:

- ✓ TRAIN IN CLOUD AND DEPLOY ON EDGE BY CONTAINERS
- ✓ PRE-VALIDATED WITH ONNX MODEL ZOO
- ✓ SELECT FRAMEWORKS FOR SPECIFIC WORKLOADS
- ✓ IMPROVE SCORING LATENCY & EFFICIENCY OF MODELS WITH AZURE SERVICES

CLOUD TO EDGE: AZURE ML, ONNX RT & INTEL OPENVINO ™



AT THE EDGE



ONNX* ECOSYSTEMS

Frameworks	Caffe2, Chainer, mxnet, Microsoft Cognitive Toolkit, ML.NET, PaddlePaddle, PyTorch, MATLAB
Converters	ML, LibSVM, Keras, TensorFlow, scikit learn, dmlc XGBoost, XL
Runtimes	ONNX RUNTIME, OpenVINO™, And others…
Visualizers	NETRON, Visual DL

GET STARTED

3 SETUP OPTIONS IN GITHUB

BUILD FROM SOURCE

- (ONNX RT + OV) *
- RUN NATIVELY, COMPILE FROM SCRATCH - PROVIDES MAXIMUM FLEXIBILITY
- DEPLOY NATIVELY AT THE EDGE
- Github Readme ▸ https://github.com/microsoft/onnxruntime/blob/master/docs/execution_providers/OpenVINO-ExecutionProvider.md

DEPLOY CONTAINERS FROM AZURE IOT EDGE

- [ONNX RT + OV + AZURE IOT EDGE) *
- PROVIDES FLEXIBILITY AS WELL AS CONVENIENCE THROUGH CONTAINER SUPPORT
- DEPLOY CUSTOM APPLICATIONS IN CONTAINERS FROM AZURE IOT EDGE
- Github Azure IoT Hub Instructions ▸ https://github.com/intel/Edge-Analytics-FaaS/tree/master/Azure-IoT-Edge/OnnxRuntime

DEPLOY CONTAINERS FROM AZURE ML

- [ONNX RT + OV + AZURE IOT EDGE]
- MORE AUTOMATED, AZURE ML CONSTRUCTS THE CONTAINER FROM PRE-DEFINED AZURE ML FORMAT APPLICATIONS
- DEPLOY AZURE ML APPLICATIONS IN CONTAINERS FROM AZURE ML SERVICES
- Github Azure ML Container Dockerfiles ▸ https://github.com/microsoft/onnxruntime/tree/master/dockerfiles

*Note: Download the Intel® Distribution of OpenVINO™ toolkit installer(tgz) before building the above Docker image.

Additional Github Resource: Azure ML Instructions ▸

Cloud to Edge Deployment flow using Azure ML and Azure IoT Edge

Using Azure ML to deploy Azure ML container applications

https://github.com/Azure-Samples/onnxruntime-iot-edge/tree/master/AzureML-OpenVINO

HOW IT WORKS (RUNTIME)

```
import onnxruntime

session = onnxruntime.InferenceSession("model.onnx")

x = GetInputData()

y = session.run([session.get_outputs()[0].name],
                {session.get_inputs()[0].name : x})
```

Simple runtime call pointing to model location

ONNX* MODEL ZOO

Models

Read the Usage section below for more details on the file formats in the ONNX Model Zoo
Python code for validating your ONNX model using test data.

Vision

- Image Classification
- Object Detection & Image Segmentation
- Body, Face & Gesture Analysis
- Image Manipulation

Language

- Machine Comprehension
- Machine Translation
- Language Modelling

Other

- Visual Question Answering & Dialog
- Speech & Audio Processing
- Other interesting models

Object Detection & Image Segmentation

Object detection models detect the presence of multiple objects in an image and segment out areas of the image where the objects are detected. Semantic segmentation models partition an input image by labeling each pixel into a set of pre-defined categories.

Model Class	Reference	Description
Tiny YOLOv2	Redmon et al.	A real-time CNN for object detection that detects 20 different classes. A smaller version of the more complex full YOLOv2 network.
SSD	Liu et al.	
Faster-RCNN	Ren et al.	
Mask-RCNN	He et al.	

Machine Comprehension

This subset of natural language processing models that answer questions about a given context paragraph.

Model Class	Reference	Description
Bidirectional Attention Flow	Seo et al.	A model that answers a query about a given context paragraph.
BERT-Squad	Devlin et al.	This model answers questions based on the context of the given input paragraph.

➢ Github ▸ https://github.com/onnx/models

Sources: Microsoft

ONNX* TUTORIALS

Get started with ONNX and tutorials
Docker image for ONNX and Caffe2/PyTorch
Docker image for ONNX, ONNX Runtime, and various converters

- **Getting ONNX models –** ONNX Model Zoo
- **Services –** Output ONNX models customized for your data
  Azure Custom Vision service
  Azure Machine Learning automated ML
- **Converting to ONNX format**
- **Scoring ONNX models –** Score accuracy

➢ Github ▸ https://github.com/onnx/tutorials

Sources: Microsoft

FEATURES SET & ROADMAP

Current Feature Set

- **Compute and Accelerator support:**
  - Intel® CPU, integrated GPU
  - Intel® Movidius™ Myriad™ X VPU (USB and embedded)
  - Intel® Vision Accelerator Design Products with Intel® Movidius™ Myriad™ X VPU (2x, 4x & 8x)
  - Intel® Vision Accelerator Design Products with Intel® 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 Win10*
- **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
- **Intel® Distribution of OpenVINO™ toolkit version support :** Support for major releases (recurring)
- **Latest ONNX operator coverage:** Support for updated ONNX operators (recurring)
- **Docker container support:** Win10*
- **Auto resource discovery:** Detect hardware accelerators on platform
- **Latest hardware accelerator coverage (recurring)**

ONNX Runtime* Release Notes ▸
https://github.com/microsoft/onnxruntime/releases/

DEVELOPER KITS, USE CASES, & CASE STUDIES

DEVELOPER KITS

INTEL® NEURAL COMPUTE STICK 2

link

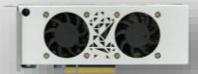
Powered by the
Intel® Movidius™ Myriad™ X VPU

IEI TANK* AIOT DEVELOPER KIT

link

Intel® Vision Accelerator Design Product Choices

Powered by Intel® Movidius™ VPU (link)



Powered by Intel® Arria® 10 FPGA (link)
In Preview

UP SQUARED* AI VISION X DEV KIT

link

Intel® Vision Accelerator Design Product

Powered by Intel® Movidius™ VPU (link)

EQUIPMENT MAKER OFFERS

- IEI* TANK AIoT Developer Kit
  Intel® Core® i7/i5/i3 Processor & Intel® Xeon® Processor
  Use Case: Industrial
- IEI* FLEX-BX200
  Intel® Core® i3/5/i7 Processor
  Use Cases: Public Safety, Parking Mgmt., License Plate Detection

- UP* Squared; UP* Core Plus
  Intel® Atom™ Processor ; Intel® Core® i7/i5/i3 Processor
  Use Cases: Retail, DSS
- Aaeon* BOXER-6841M
  Intel® Core® i7/i5/i3 Processor
  Use Cases: Industrial, Smart Retail and Smart City

- Advantech* ARK-1124 + VEGA-320
  Intel® Atom™ Processor + Intel® Movidius™ Myriad™ X VPU
  Use Cases: Age & Gender Recognition

SUPPORT & RESOURCES

SUPPORT

Software Issues
Software issues related to ONNX Runtime with OpenVINO Execution Provider code should be logged at: “Issues” Tab https://github.com/Microsoft/onnxruntime with [OpenVINO-EP] tag.

Hardware Issues
Hardware issues should be routed towards your equipment maker suppliers, your Intel Representative, or Intel Premier Support

Supported Models
Link to supported models for the ONNX Runtime with OpenVINO Execution Provider https://github.com/microsoft/onnxruntime/blob/master/docs/execution_providers/OpenVINO-ExecutionProvider.md.
All issues related to these models should be routed towards your Intel Representative

Intel® Distribution of OpenVINO™ toolkit Support
OpenVINO issues should be reported through the OpenVINO “Computer Vision” Forum https://software.intel.com/en-us/forums/computer-vision

ONNX Support
All other ONNX model issues should be logged at “Issues” tab https://github.com/Microsoft/onnxruntime

INTEL® IOT RFP READY KITS

Check RFP Ready Kit Playbook for details on each kit

Retail

- Kiosk & digital signage
- POS & mobile POS
- Inventory Management

Industrial

- Manufacturing
- Building Management
- Agriculture
- Energy

Transportation

- Fleet Management
- Logistics

Smart Cities

- Security surveillance
- Smart lighting
- Connected Transportation
- Air quality management

Healthcare

- Medical (in-hospital)
- Remote health management

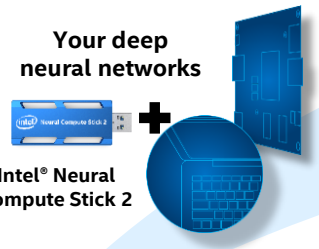
Horizontal

- Security Surveillance Video
- Connectivity

ACCELERATE PROTOTYPE TO PRODUCTION & SOLUTION DEPLOYMENT

DEVELOP ON HOST SYSTEM

Your deep neural networks



Intel® Neural Compute Stick 2

USE VISION ACCELERATOR KITS

Intel® Distribution of OpenVINO toolkit

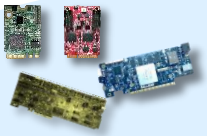
AAEON UP Squared AI Vision Developer Kit

IEI Tank AIoT Developer Kit

OpenVINO™

INCREASE PERFORMANCE

Intel® Vision Accelerator Design Products



Intel® Movidius™ Myriad™ X VPU
Intel® Arria® 10 FPGA

DEVELOP USE CASE SPECIFIC OFFERS

Developer optimization & use case specific applications

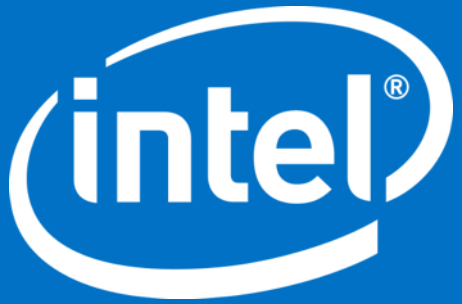
Intel® RFP Ready Kits

SCALE

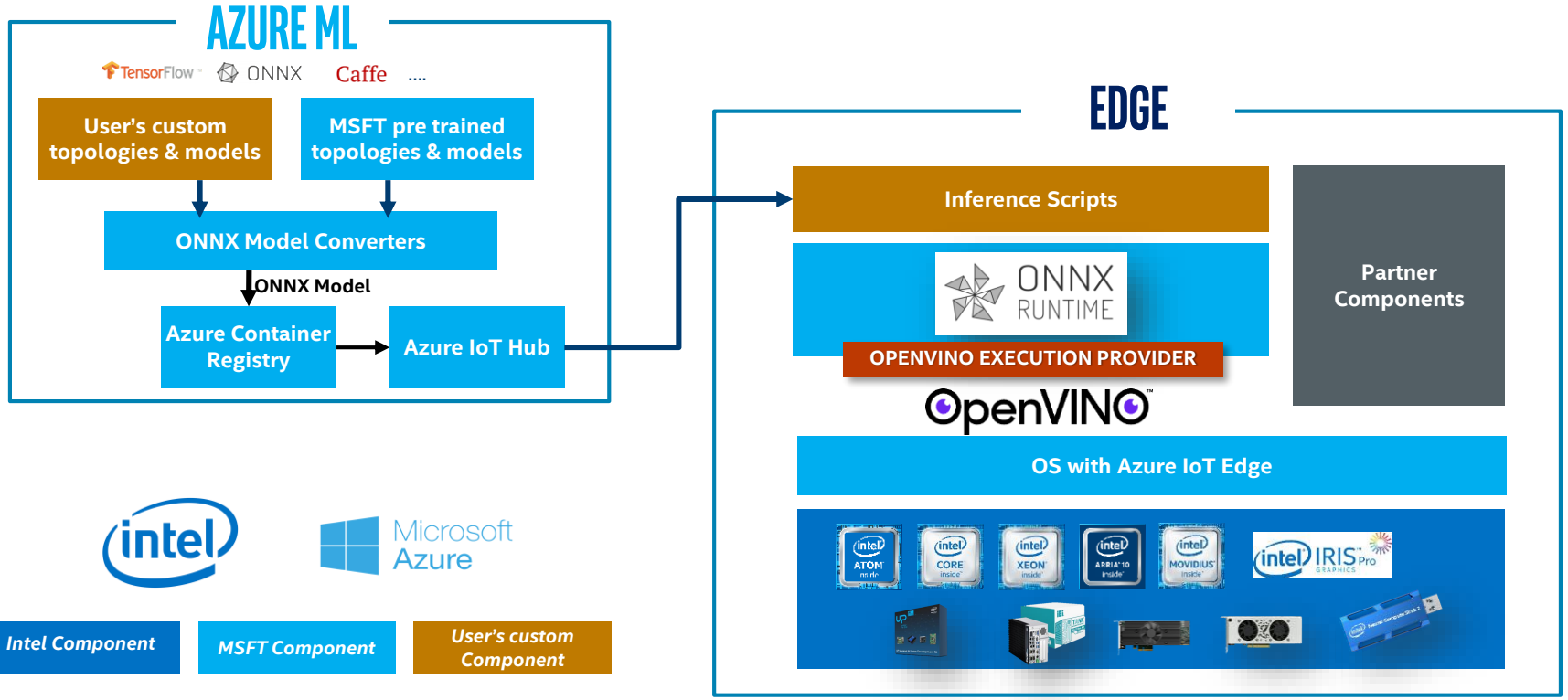
Deploy solution & solve business problems, and scale with Intel® IoT Solution Aggregators & Ecosystem

Intel® AI: In Production

https://software.intel.com/ai-in-production


intel®

SEAMLESS WORKFLOW FOR AZURE ML DEVELOPERS @ EDGE



Benefits of Intel® Distribution of OpenVINO™ toolkit

Maximize the Power of Intel® Processors: CPU, GPU/Intel® Processor Graphics, FPGA,VPU

ACCELERATE PERFORMANCE

Access Intel computer vision accelerators.
Speed code performance.
Supports heterogeneous execution.

INTEGRATE DEEP LEARNING

Unleash CNN-based deep learning inference using a common API, 40+ pretrained models, & computer vision algorithms. Validated on more than 100 public/custom models.

SPEED DEVELOPMENT

Reduce time using a library of optimized OpenCV* & OpenVX* functions, & 15+ samples.

Develop once, optimize and deploy for current & future Intel-based devices.

INNOVATE & CUSTOMIZE

Use OpenCL™ kernels/tools to add your own unique code. Customize layers without the overhead of frameworks.

Deep learning revenue is estimated to grow from $655M in 2016 to **$35B** by 2025[1].

[1]Tractica 2Q 2017