Intel Software

OpenVINO™

OPTIMIZATION BEST PRACTICES TO MAXIMIZE APPS WITH THE INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

WEBINAR

Zoe Cayetano
Technical Product Manager, Intel

Anna Belova
DL Software Engineer, Intel

Dmitry Temnov
Software Technical Consulting Engineer, Intel

WELCOME TO THE WEBINAR SERIES!

APRIL 2020

Deep Learning Streamer

Missed this session? You can still watch it on-demand!

MAY 2020

Microsoft Azure & ONNX Runtime for OpenVINO™ toolkit

Missed this session? You can still watch it on-demand!

TODAY

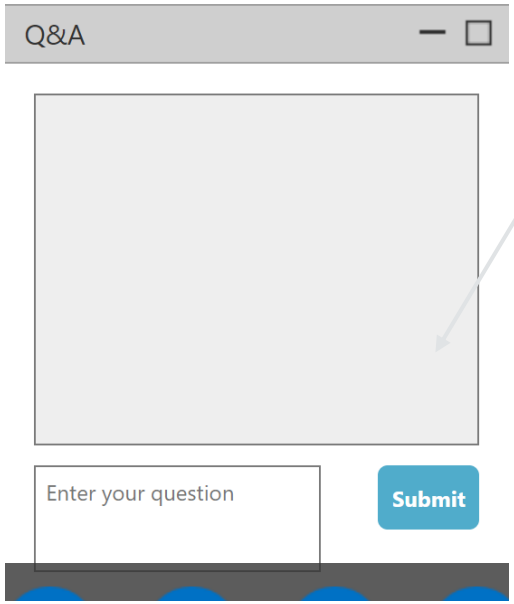
Optimization Best Practices in Deep Learning

https://software.seek.intel.com/openvino-webinar-series

Copyright © 2020, Intel Corporation. All rights reserved.

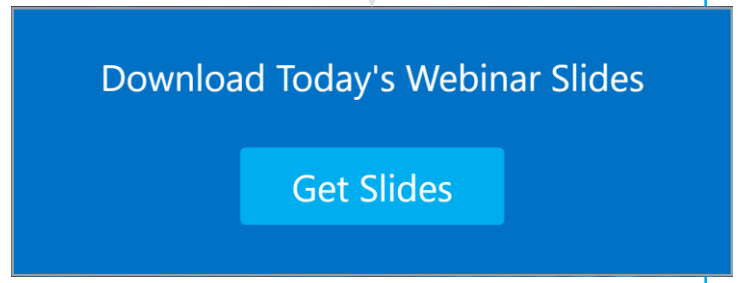
FAMILIARIZE WITH THE WEBINAR CONSOLE

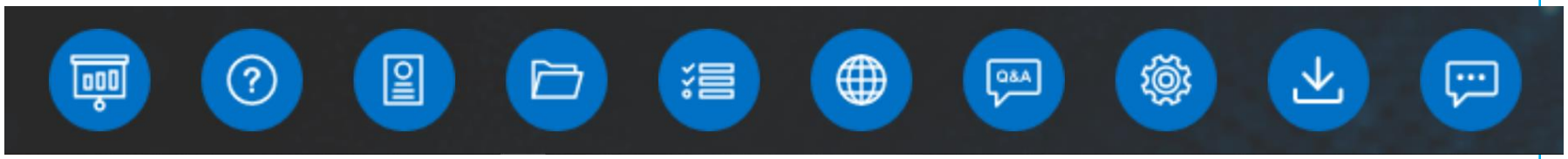
Try it yourself!
Download OpenVINO™ toolkit



Ask a question anytime
Type in the box and hit “submit”

Get today’s presentation
Follow along, take notes or review later





Resources | Community Forum | Survey

Copyright © 2020, Intel Corporation. All rights reserved.

OpenVINO

IN THIS WEBINAR, YOU WILL:

- Get an understanding of the different optimization metrics that are important in deep learning
- Learn about the available tools and techniques to optimize your deep learning applications within Intel® Distribution of OpenVINO™ toolkit
- See a demo of how to use the DL Workbench to visually fine-tune your models
- Explore benchmarking resources to test on your own applications

Copyright © 2020, Intel Corporation. All rights reserved.

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

Tool Suite for High-Performance, Deep Learning Inference

Faster, more accurate real-world results using high-performance, AI and computer vision inference deployed into production across Intel® architecture from edge to cloud

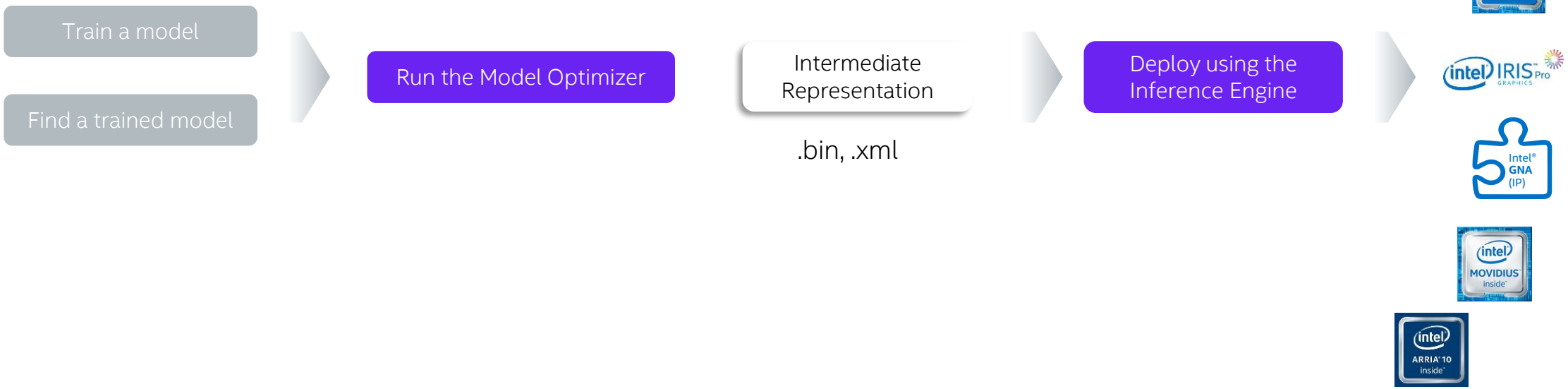
High-Performance,
Deep Learning Inference

Streamlined Development,
Ease of Use

Write Once,
Deploy Anywhere

Copyright © 2020, Intel Corporation. All rights reserved.

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT GETTING STARTED



Copyright © 2020, Intel Corporation. All rights reserved.

DEEP LEARNING IN REAL-WORLD APPLICATIONS

Powered by the Intel® Distribution of OpenVINO™ toolkit

JLK INSPECTION **Healthcare Access and Quality** Solution Brief

Reduced average inference time on Intel® NUC (with no GPU) from 4.23 seconds to just 2.81 seconds, which helps medical professionals reach more people, accelerate screening and help improve quality of care.

dc water is life® **Sewer pipe inspection analysis** Solution Brief

Inference time was improved with a reduction of up to 80% using Intel Xeon processors with the OpenVINO toolkit, while not producing significant loss in model precision or accuracy.

GitHub Repo

DEVELOPER SPOTLIGHT: Chicken Coop Protector

This application is designed to identify potential predators around chicken coops using AI, computer vision and a classification model. Once a predator is identified the application should trigger a deterrent noise to scare off the predator.

ZEROFOX **Security Against Social and Digital Attacks** Solution Brief

Performance improvements of up to 2.3x faster, reducing latency by up to 50 percent for threat detection and remediation to protect businesses against targeted social and digital attacks.

Retail Business Services **Frictionless retail checkout** Solution Brief

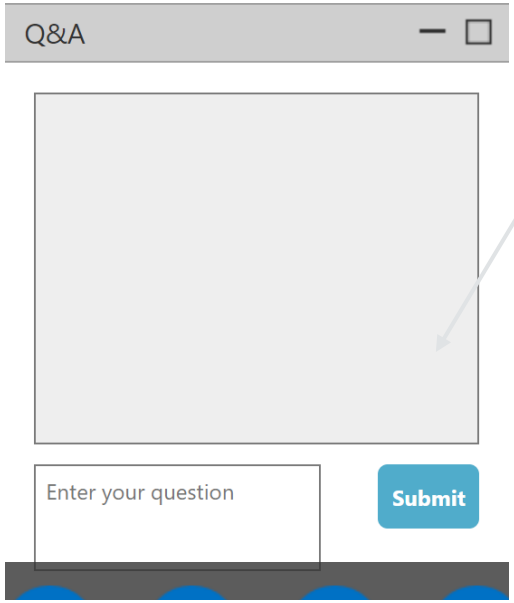
Using existing Intel-based point-of-sale systems, automatic inventory and shopper tracking with cashier-less checkout at a physical retail store was deployed at Quincy, Massachusetts.

Success Stories ▸ https://intel.com/openvino-success-stories

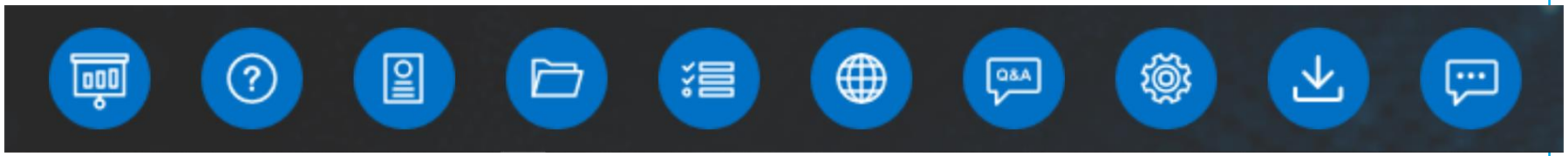
intel Copyright © 2020, Intel Corporation. All rights reserved.

BE FEATURED ON OUR NEXT WEBINAR

Share a link to your project in the Q&A box, webinar survey or tag us on your favorite social channels #OpenVINO



Ask a question anytime
Type in the box and hit “submit”



Resources | Community Forum | Survey

Copyright © 2020, Intel Corporation. All rights reserved.

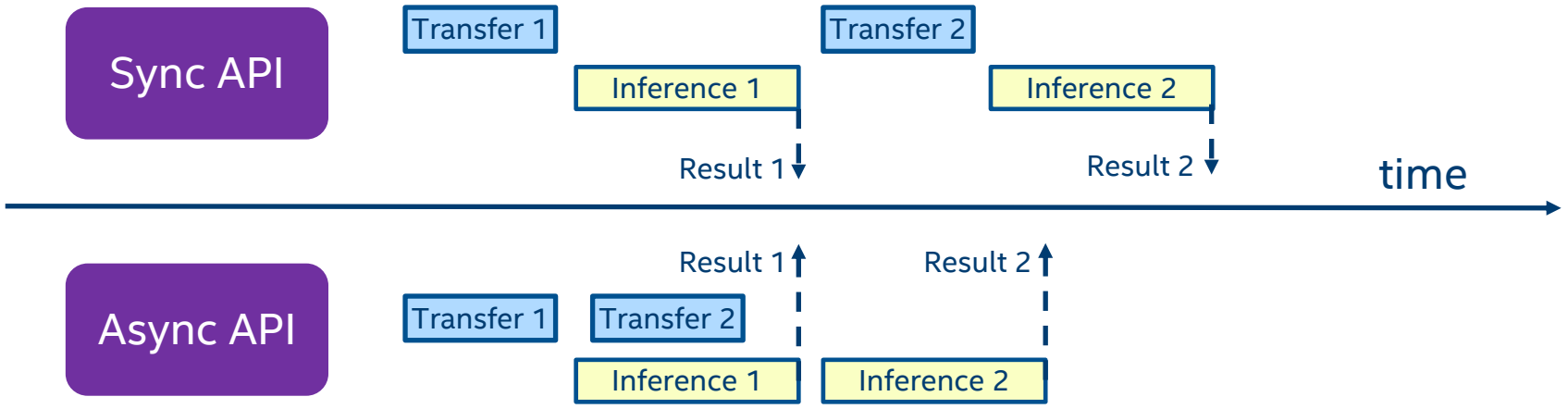
OPTIMIZATION
BEST PRACTICES

INFERENCE ENGINE

Synchronous vs Asynchronous Execution

In IE API model can be executed by **Infer Request** which can be:

- **Synchronous** - blocks until inference is completed.
- **Asynchronous** – checks the execution status with the wait, or specify a completion callback *(recommended way)*.



Copyright © 2020, Intel Corporation. All rights reserved.

INFERENCE ENGINE

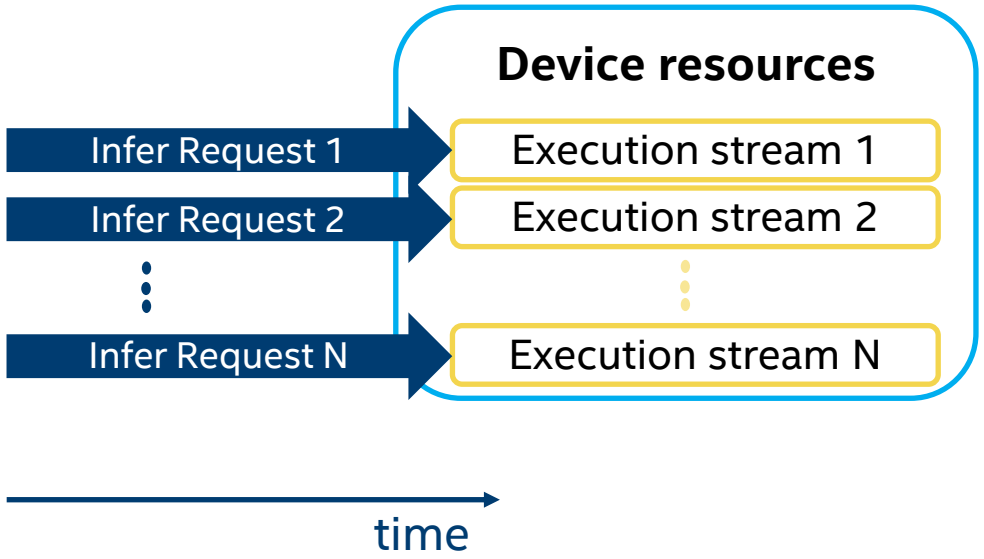
Throughput Mode for CPU, iGPU and VPU

Latency – inference time of 1 frame (ms).

Throughput – overall amount of frames inferred per 1 second (FPS)

“Throughput" mode allows the Inference Engine to efficiently run multiple infer requests simultaneously, greatly improving the overall throughput.

Device resources are divided into execution “**streams**” – parts which runs infer requests in parallel



Copyright © 2020, Intel Corporation. All rights reserved.

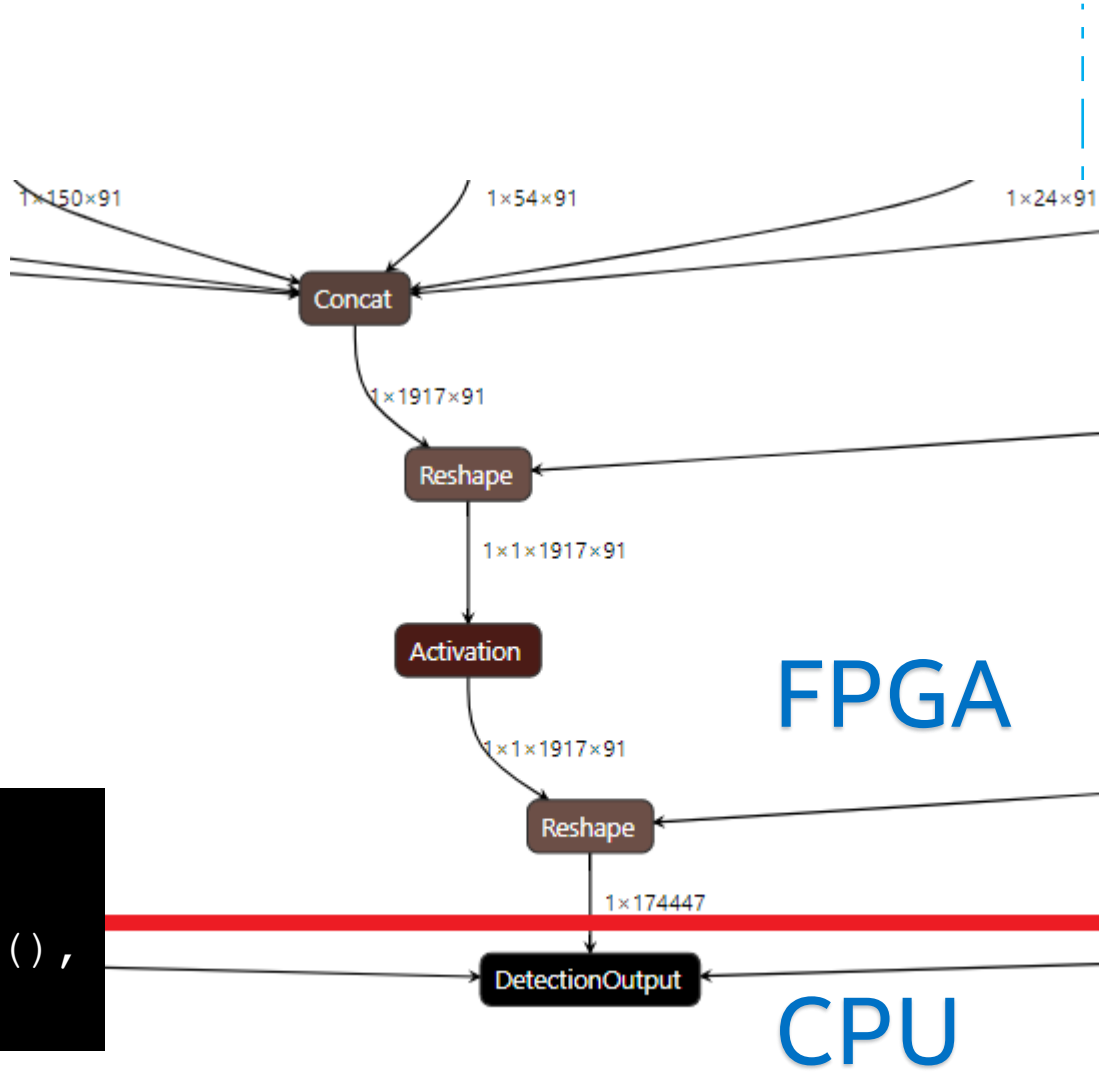
OpenVINO

WRITE ONCE, DEPLOY ANYWHERE

Heterogeneous Support

- You can execute different layers on different HW units
- Offload unsupported layers on fallback devices:
  - Default affinity policy
  - Setting affinity manually (`CNNLayer::affinity`)
- All device combinations are supported (CPU, GPU, FPGA, MYRIAD, HDDL)
- Samples/demos usage "`-d HETERO:FPGA,CPU`"

```
InferenceEngine::Core core;
        auto executable_network =
        core.LoadNetwork(reader.getNetwork(),
        "HETERO:FPGA,CPU");
```



Copyright © 2020, Intel Corporation. All rights reserved.

OpenVINO

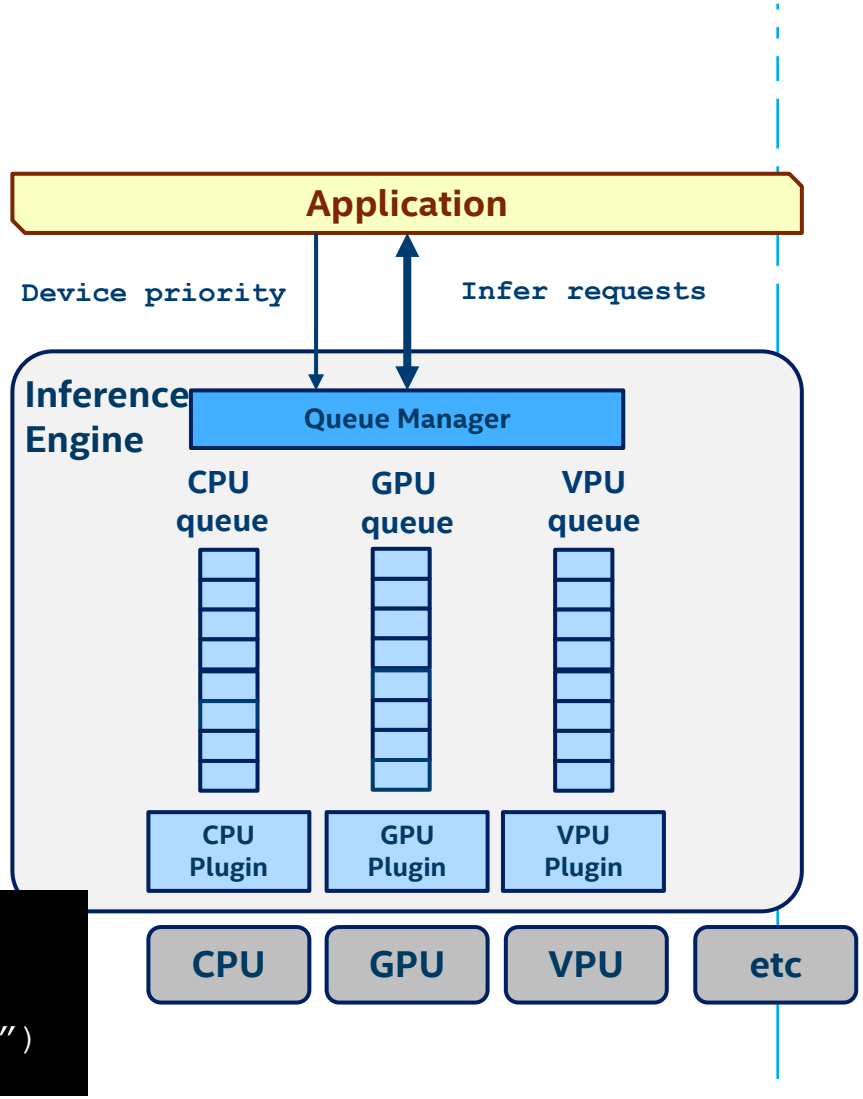
WRITE ONCE, DEPLOY ANYWHERE

Multi-device Support

Automatic load-balancing between devices (inference requests level) for full system utilization

- Any combinations of the following devices are supported (CPU, iGPU, VPU, HDDL)
- As easy as “-d MULTI:CPU,GPU” for cmd-line option of your favorite sample/demo
- C++ example (Python is similar)

```
Core ie;
ExecutableNetwork exec =
ie.LoadNetwork(network,{{“DEVICE_PRIORITIES”, “CPU,GPU”}}, “MULTI”)
```



Copyright © 2020, Intel Corporation. All rights reserved.

OPTIMIZATION TECHNIQUE

Low Precision

1. Low-precision runtime for CPU (VNNI-powered Cascade Lake or Ice Lake)
2. Post-training Optimization Tool (POT)
3. Neural Network Compression Framework (NNCF):

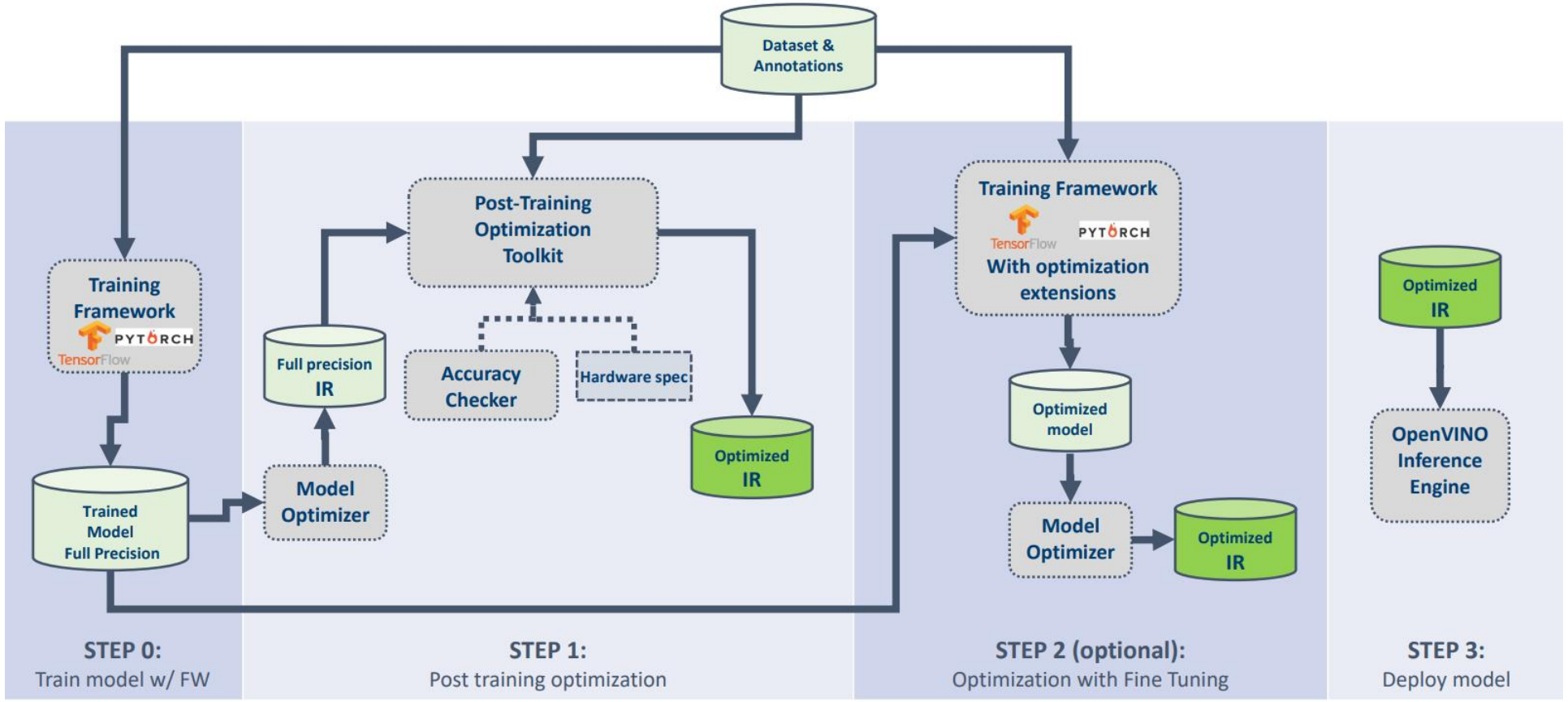
- https://github.com/openvinotoolkit/nncf_pytorch

Tip: If you have the model in framework format which is already INT8 (Example: models trained with Tensorflow FakeQuantized technology), just use Model Optimizer to convert it to IR and run with Inference Engine

Copyright © 2020, Intel Corporation. All rights reserved.

OPTIMIZATION TECHNIQUE

Low Precision Workflow



Copyright © 2020, Intel Corporation. All rights reserved.

OpenVINO

POST-TRAINING OPTIMIZATION TOOL (POT)

Converts models into a more hardware-friendly representation by applying specific methods that do not require re-training

- Int8 quantization
- Sparsity

Available as

- CLI Python 3-based tool
- API
- inside DL Workbench

Low Precision Optimization Guide ▸ docs.openvinotoolkit.org/latest/_docs_LowPrecisionOptimizationGuide.html

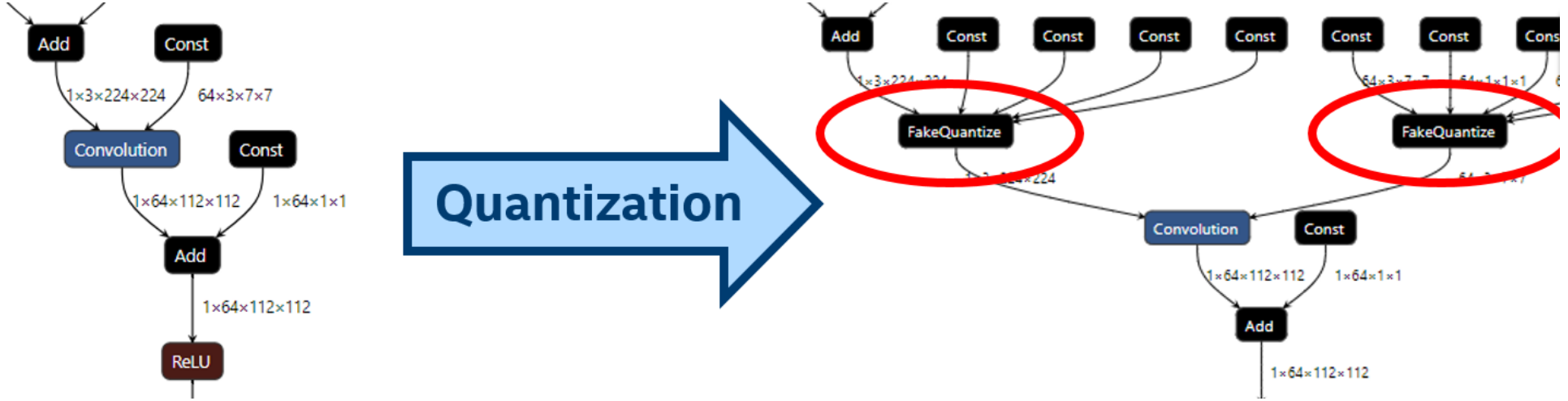
intel Copyright © 2020, Intel Corporation. All rights reserved.

POT.QUANTIZATION ALGORITHMS

Low Precision

Quantization to low precision (INT8) inserting FakeQuantize layers

- DefaultQuantization
- AccuracyAwareQuantization



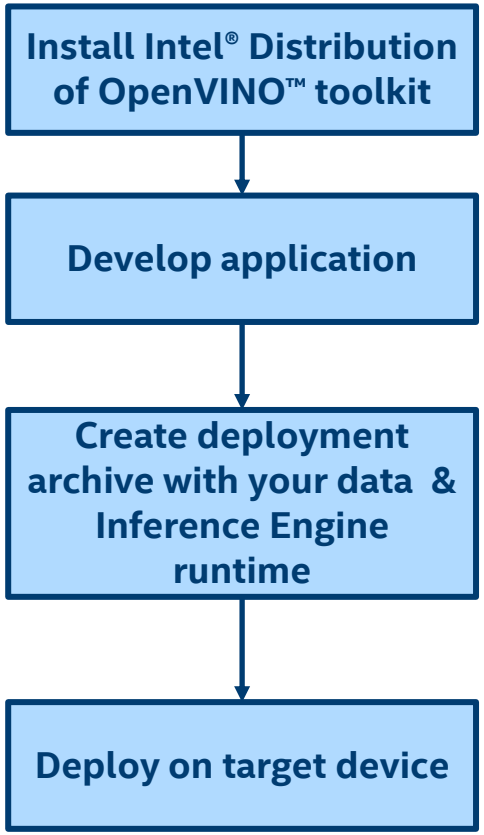
Low Precision Optimization Guide ▸ docs.openvinotoolkit.org/latest/_docs_LowPrecisionOptimizationGuide.html

Copyright © 2020, Intel Corporation. All rights reserved.

DEPLOYMENT MANAGER

- Generate an optimal, minimized runtime package for selected target device.
- Deploy Inference Engine with pre-compiled application-specific data such as models, config, and a subset of required hardware plugins.
- Achieve deployment footprint to be several times smaller than the development footprint.

For more details, see Introduction to CLI Deployment Manager



Target	Size, MB
CPU only	65
GPU only	26
Myriad only	22
HDDL only	27
GNA only	15

Measurements for deployment archives based on 2019 R3

Copyright © 2020, Intel Corporation. All rights reserved.

BENCHMARKING
RESOURCES

OpenVINO

HOW DO YOU MEASURE PERFORMANCE?

- Benchmark app – command line application (C++ and Python) which calculates the latency (ms) and throughput (fps). Recommended way to estimate Inference Engine performance.
- Distributed in binary tool and as part of IE samples
- Usage:

```
$ benchmark_app -m <model> -i <input> -d <device>
```

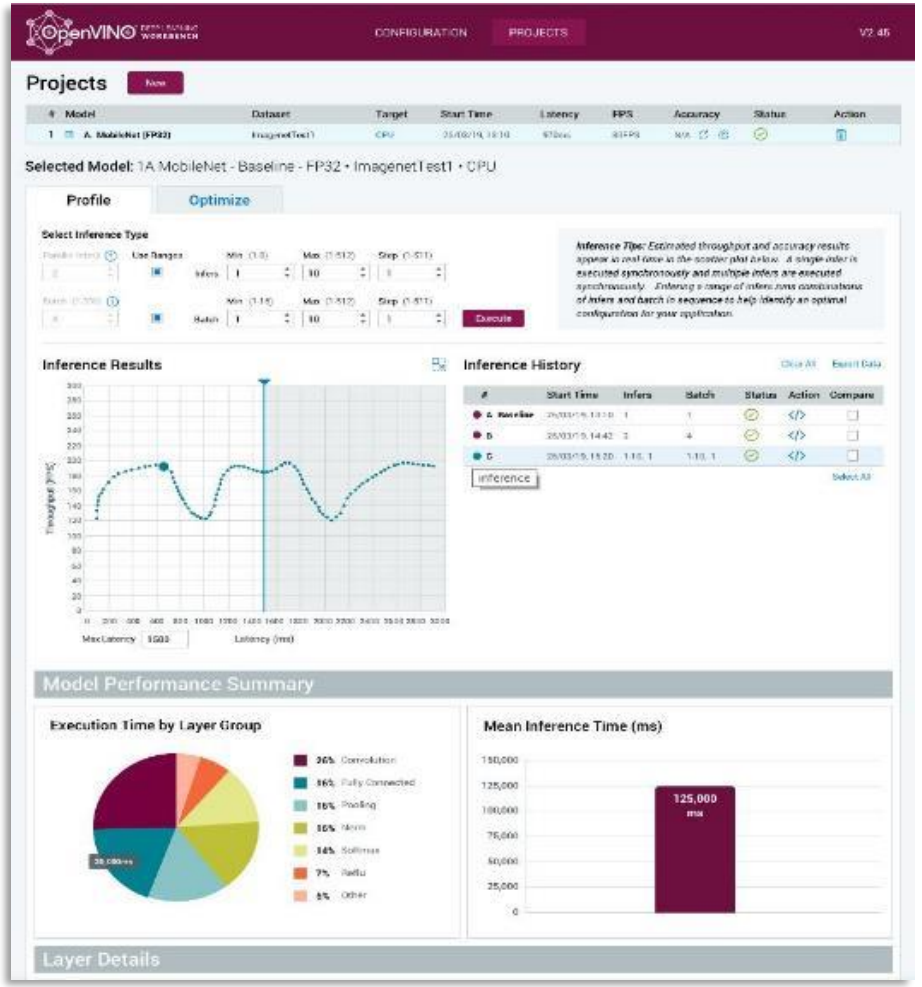
- Automatically queries and sets optimal values of the following options to saturate the device:
  - **"-nireq"** – number of Infer requests
  - **"-nstreams"** (CPU, GPU, VPU) – number of execution "streams"

Performance Benchmarks ▸ docs.openvinotoolkit.org/latest/_docs_performance_benchmarks.html

intel Copyright © 2020, Intel Corporation. All rights reserved.

DEEP LEARNING WORKBENCH

- Web-based, **UI extension tool** of the Intel® Distribution of OpenVINO™ toolkit
- **Visualizes performance data** for topologies and layers to aid in model analysis
- **Automates analysis** for optimal performance configuration (streams, batches, latency)
- **Experiment** with int8 or Winograd calibration for optimal tuning
- Provide **accuracy information** through accuracy checker
- **Direct access** to models from public set of Open Model Zoo



Development Guide
https://docs.openvinotoolkit.org/latest/_docs_Workbench_DG_Introduction.html

Copyright © 2020, Intel Corporation. All rights reserved.

DEMO
USING THE DL WORKBENCH

KEY TAKEAWAYS & RESOURCES

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

OpenVINO

IN THIS WEBINAR, YOU:

- Got an understanding of the different optimization metrics that are important in deep learning
- Learned about the available tools and techniques to optimize your deep learning applications within Intel® Distribution of OpenVINO™ toolkit
- Saw a demo of how to use the DL Workbench to visually fine-tune your models
- Explored benchmarking resources to test on your own applications

intel Copyright © 2020, Intel Corporation. All rights reserved.

GETTING STARTED WITH INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

QUALIFY

- Use a trained model and check if framework is supported

  – *or* –

- Take advantage of a pre-trained model from the Open Model Zoo

INSTALLATION

- Download the Intel® OpenVINO™ toolkit package from Intel® Developer Zone, or by YUM or APT repositories
- Utilize the Getting Started Guide

PREPARE

- Understand sample demos and tools included
- Understand performance
- Choose hardware option with Performance Benchmarks
- Build, test and remotely run workloads on the Intel® DevCloud for the Edge before buying hardware

HANDS ON

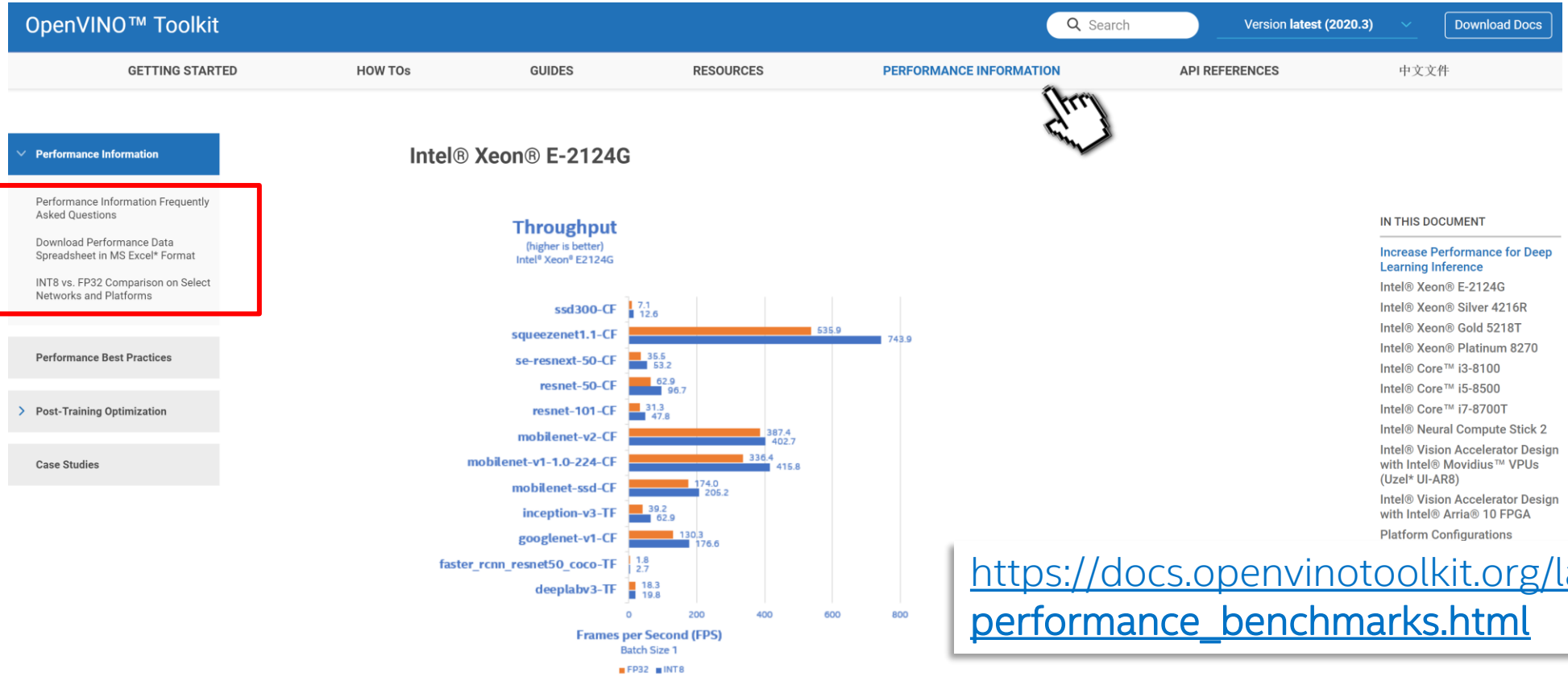
- Visualize metrics with the Deep Learning Workbench
- Utilize prebuilt, Reference Implementations to become familiar with capabilities
- Optimize workloads with these performance best practices
- Use the Deployment Manager to minimize deployment package

SUPPORT

- Ask questions and share information with others through the Community Forum
- Engage using #OpenVINO on Stack Overflow
- Visit documentation site for guides, how to's, and resources
- Attend training and get certified

Copyright © 2020, Intel Corporation. All rights reserved.

PERFORMANCE BENCHMARKS



https://docs.openvinotoolkit.org/latest/_docs_performance_benchmarks.html

Copyright © 2020, Intel Corporation. All rights reserved.

JUMPSTART DEEP LEARNING TODAY!

Download Free ▸
Intel® Distribution of OpenVINO™ toolkit

Also available from
Docker | YUM | APT | [NEW] Anaconda Cloud

Try out Intel® Distribution of OpenVINO™ toolkit and **join the conversation** in the community forum or by tagging **#AskOpenVINO** online

Missed the past webinars? You can still watch the replay at any time!
https://software.seek.intel.com/openvino-webinar-series

Copyright © 2020, Intel Corporation. All rights reserved.

LIVE Q&A

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

intel