ACCELERATE DEEP LEARNING INFERENCE USING INTEL TECHNOLOGIES

INTRODUCTION: SMART VIDEO

February 2020

SMART VIDEO WORKSHOP OVERVIEW

INTRODUCTION

1. Introduction to Intel technologies for deep learning inference
2. Hardware acceleration techniques

Each module contains a hands-on lab exercise that introduces various Intel technologies to accelerate computer vision application with hardware heterogeneity.

INTEL® DISTRIBUTION OF OPENVINO™ 101	2. Basic End-to-End Object Detection Example
HARDWARE ACCELERATION ON LAPTOP AND DEVCLOUD	3./4./5. Hardware Acceleration with CPU, Integrated GPU, Intel® Movidius™ NCS, FPGA
OPTIMIZATION	6. Optimization Tools and Techniques
APPLICATION	7. Advanced Video Analytics
CUSTOM LAYERS	8. Custom layers

OPTIMIZATION NOTICE

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness or any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.

LEGAL NOTICES AND DISCLAIMERS (1 OF 2)

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services, and processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document.

Arduino* 101 and the Arduino infinity logo are trademarks or registered trademarks of Arduino, LLC.

Altera, Arria, the Arria logo, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Xeon Phi, Movidius, Saffron, and Xeon are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

LEGAL NOTICES AND DISCLAIMERS (2 OF 2)

This document contains information on products, services, and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications, and roadmaps. Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/performance.

Cost-reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future are forward-looking statements that involve a number of risks and uncertainties.

A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors, known as *errata*, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon, Movidius, Saffron, and others are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

AI IS CHANGING EVERY MARKET

EMERGENCY RESPONSE

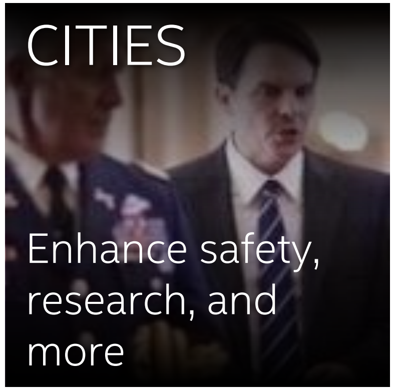
Real-time emergency and crime response

ENERGY

Maximize production and uptime

EDUCATION

Transform the learning experience

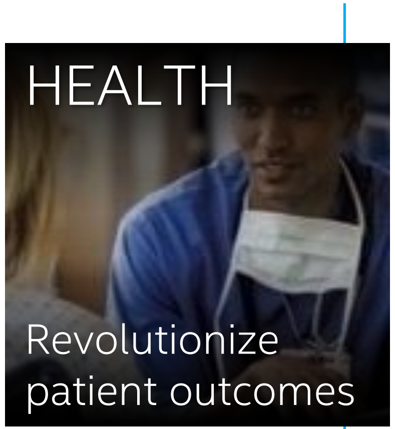


CITIES

Enhance safety, research, and more

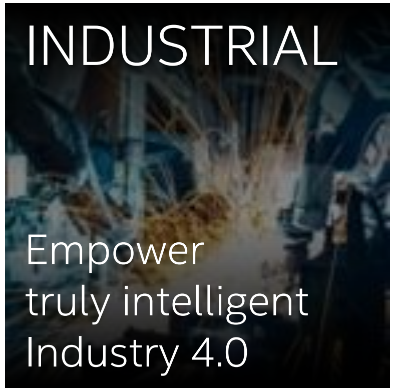
FINANCE

Turn data into valuable intelligence



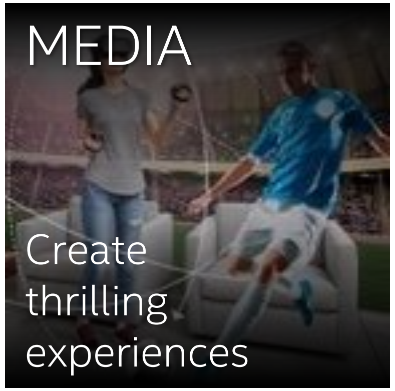
HEALTH

Revolutionize patient outcomes



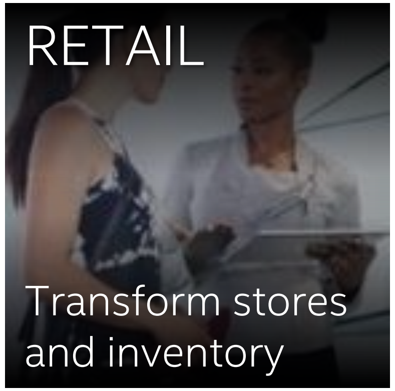
INDUSTRIAL

Empower truly intelligent Industry 4.0



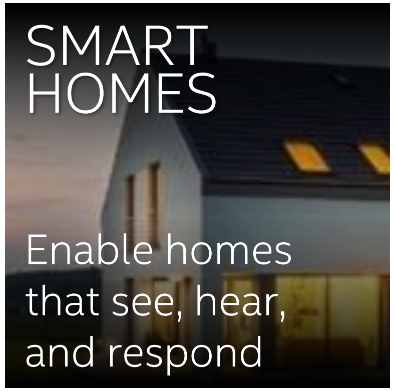
MEDIA

Create thrilling experiences



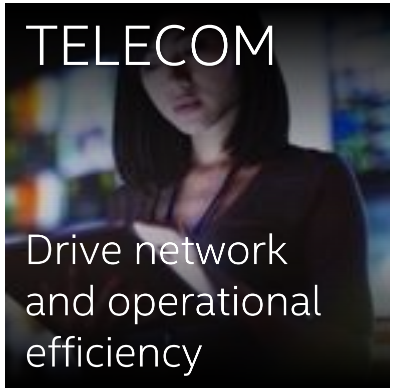
RETAIL

Transform stores and inventory



SMART HOMES

Enable homes that see, hear, and respond



TELECOM

Drive network and operational efficiency

SMART CITIES

Efficient and robust traffic systems

EMERGENCY RESPONSE

FINANCIAL SERVICES

MACHINE VISION

CITIES/TRANSPORTATION

VIDEO: THE “EYE OF IOT”

USE OF VIDEO, COMPUTER VISION AND DEEP LEARNING IS GROWING RAPIDLY

AUTONOMOUS VEHICLES

RESPONSIVE RETAIL

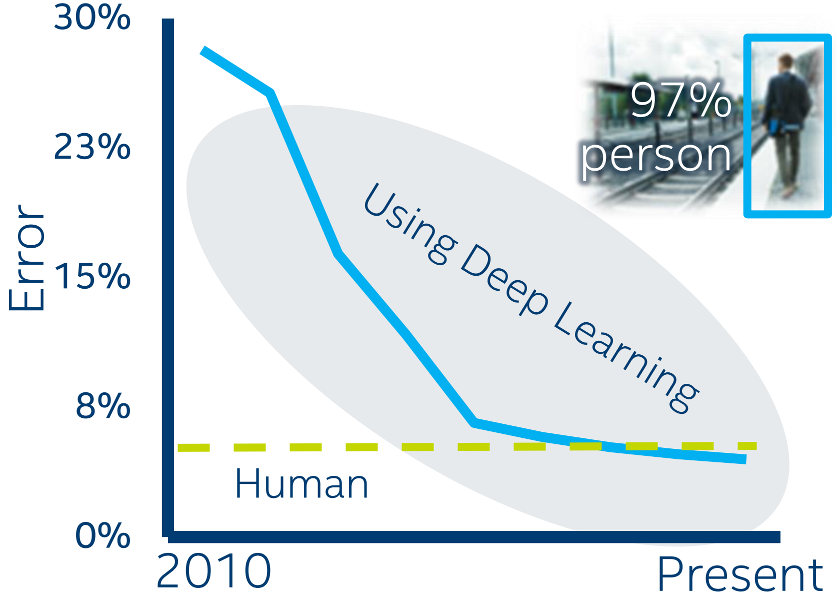
MANUFACTURING

PUBLIC SECTOR

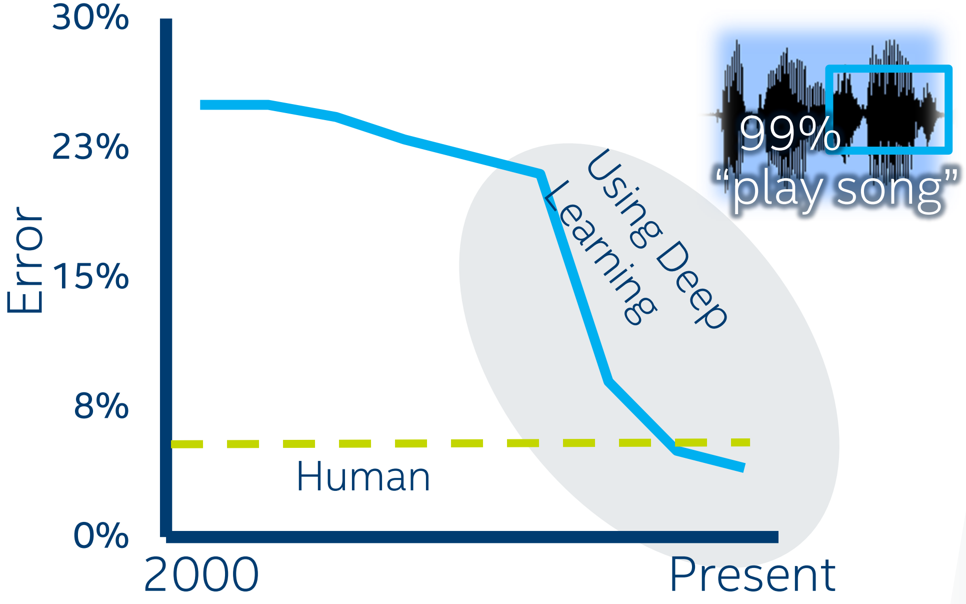
DEEP LEARNING BREAKTHROUGHS AND OPPORTUNITIES

Machines able to meet or exceed human image and speech recognition

Image Recognition



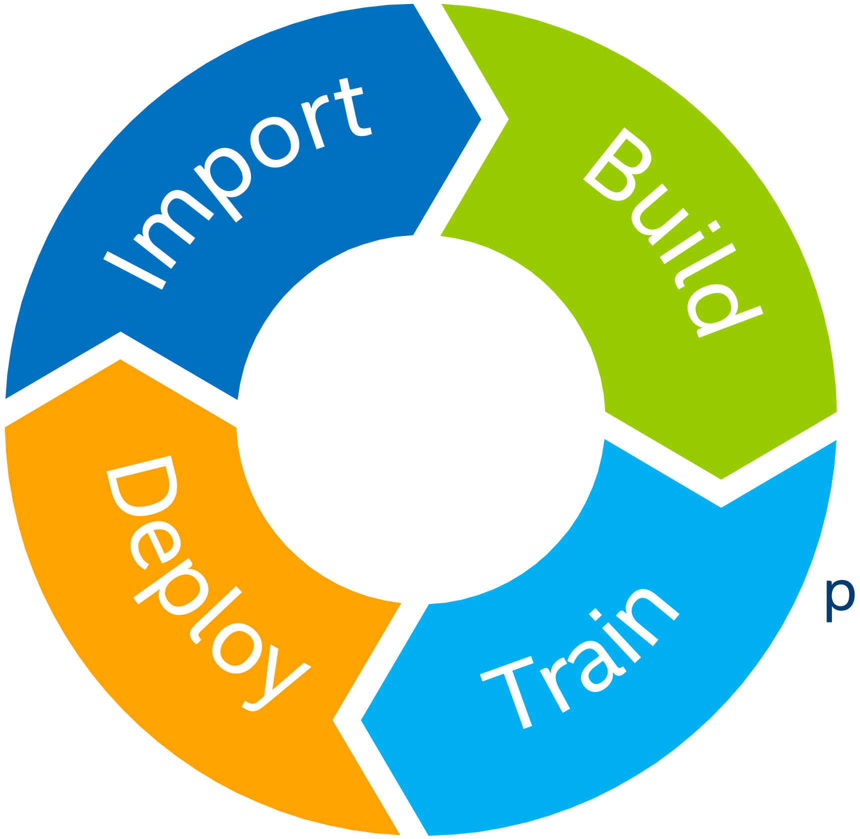
Speech Recognition



ADDITIONAL ECONOMIC IMPACT DRIVEN BY AI $13 TRILLION IN 2030

Source: ILSVRC ImageNet winning entry classification error rate each year 2010-2016 (Left), https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/ (Right)
Source: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

DEEP LEARNING DEVELOPMENT CYCLE



Data acquisition and organization

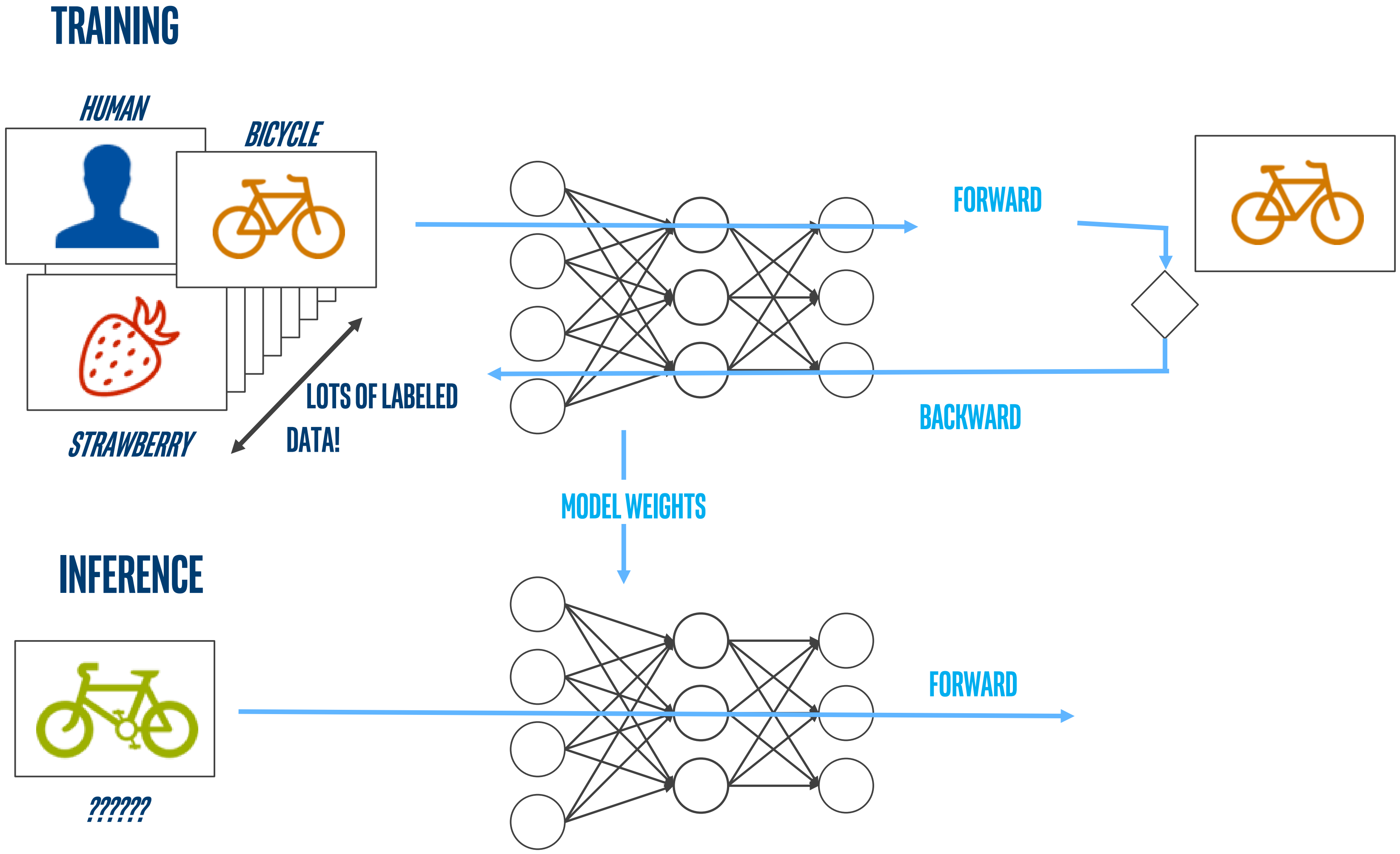
Create models

Adjust models to meet performance and accuracy objectives

Integrate trained models with application code

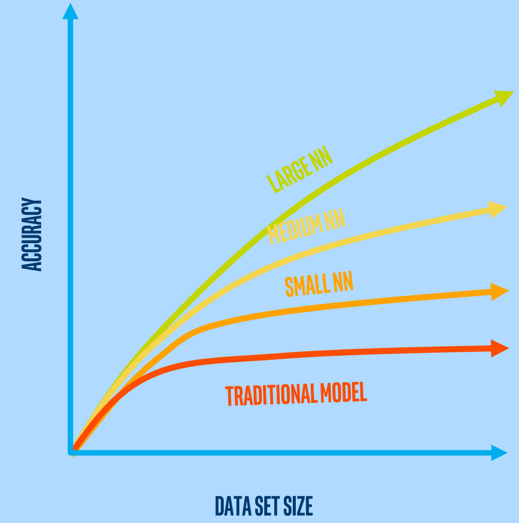
Intel® Distribution OpenVINO™ Toolkit Provides Deployment from Intel® Edge to Cloud

DEEP LEARNING: TRAINING VS. INFERENCE



DID YOU KNOW?

Training requires a very large data set and deep neural network (many layers) to achieve the highest accuracy in most cases



THE CHALLENGES IN DEPLOYING DEEP LEARNING

Unique Inference Needs

Gap in performance and accuracy between trained and deployed models

Low performing, lower accuracy models deployed

Integration Challenges

No streamlined way for end-to-end development workflow

Slow time-to-solution and time–to-market

No One Size Fits All

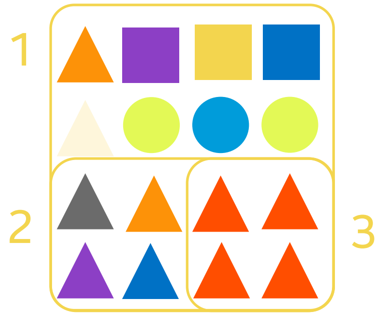
Diverse requirements for myriad use cases require unique approaches

Inability to meet use-case specific requirements

AI COMPUTE CONSIDERATIONS

How do you determine the right computing for your AI needs?

WORKLOADS



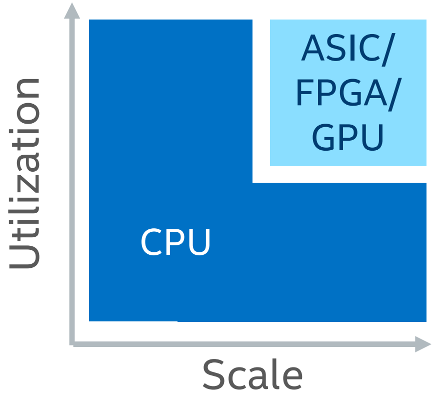
What is my workload profile?

REQUIREMENTS

RESPONSE TIME
SECURITY
DATA
VERSATILITY
SIZE
SCALABILITY
COST
POWER
RELIABILITY
THROUGHPUT
EFFICIENCY
MANAGEABILITY
FLEXIBILITY
ACCURACY

What are my use case requirements?

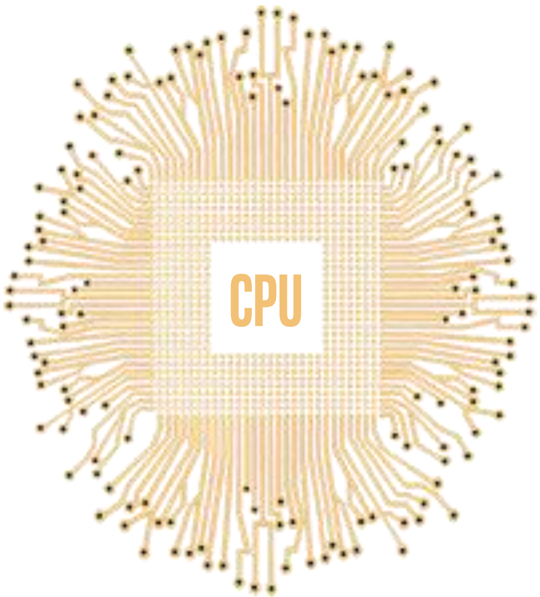
DEMAND



How prevalent is AI in my environment?

WHY INTEL AI COMPUTE?

MAXIMIZE



Get the most out of the foundation for AI from the CPU leader

OPTIMIZE

CPU

FPGA

GPU

ASIC

Choose the right compute for you from the one with all the options

SIMPLIFY

OPTIMIZED SW

DATA PIPELINE

ANALYTICS & AI

SUPPORT

MOVE/STORE

Reduce “moving parts” by building on an optimized AI platform

LEAD

Lead your industry by aligning with the builder of next-gen AI solutions

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

Tool Suite for High-Performance, Deep Learning Inference

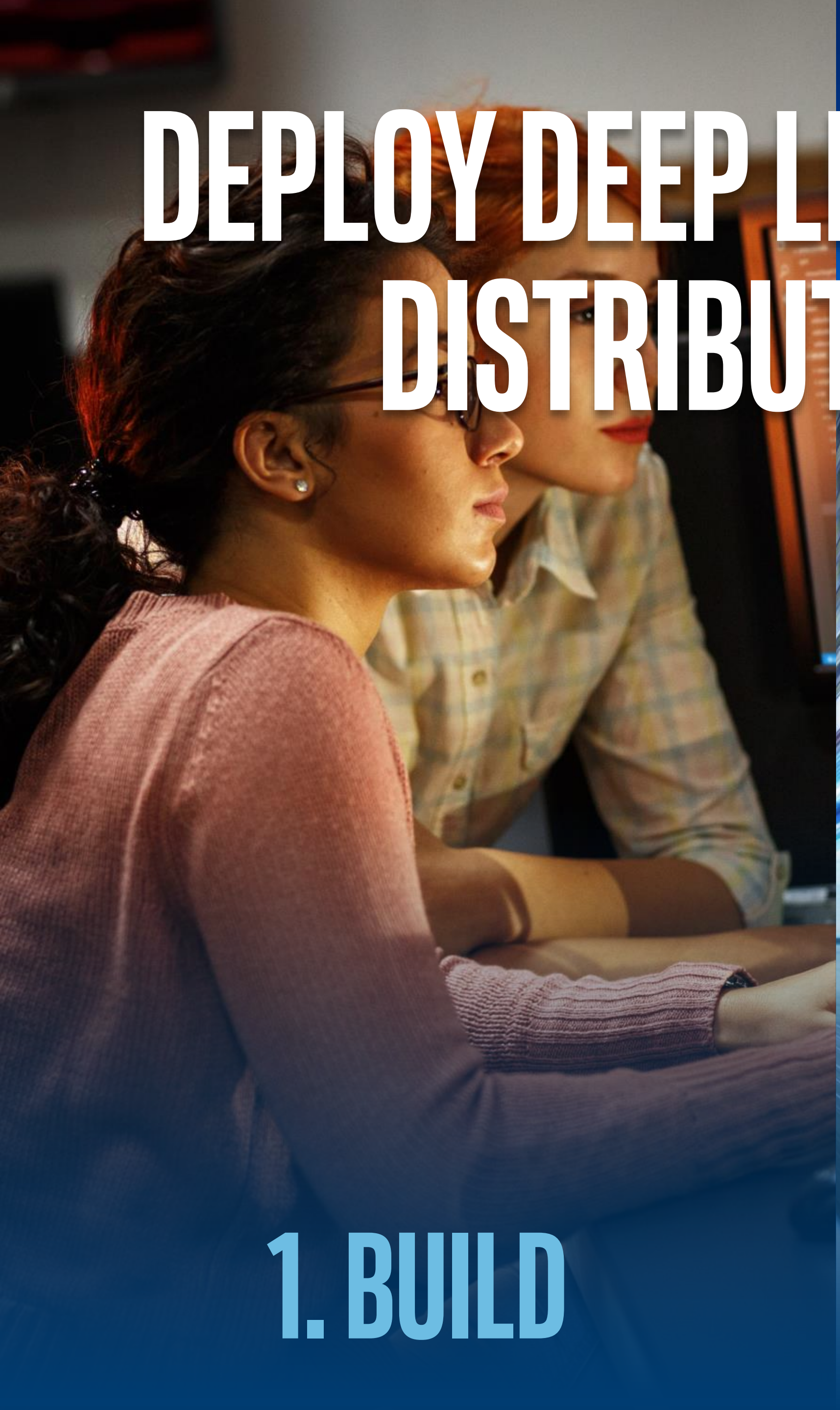
Faster, more accurate real-world results using high-performance, AI and computer vision inference deployed into production across Intel® architecture from edge to cloud

High-Performance,
Deep Learning Inference

Streamlined Development,
Ease of Use

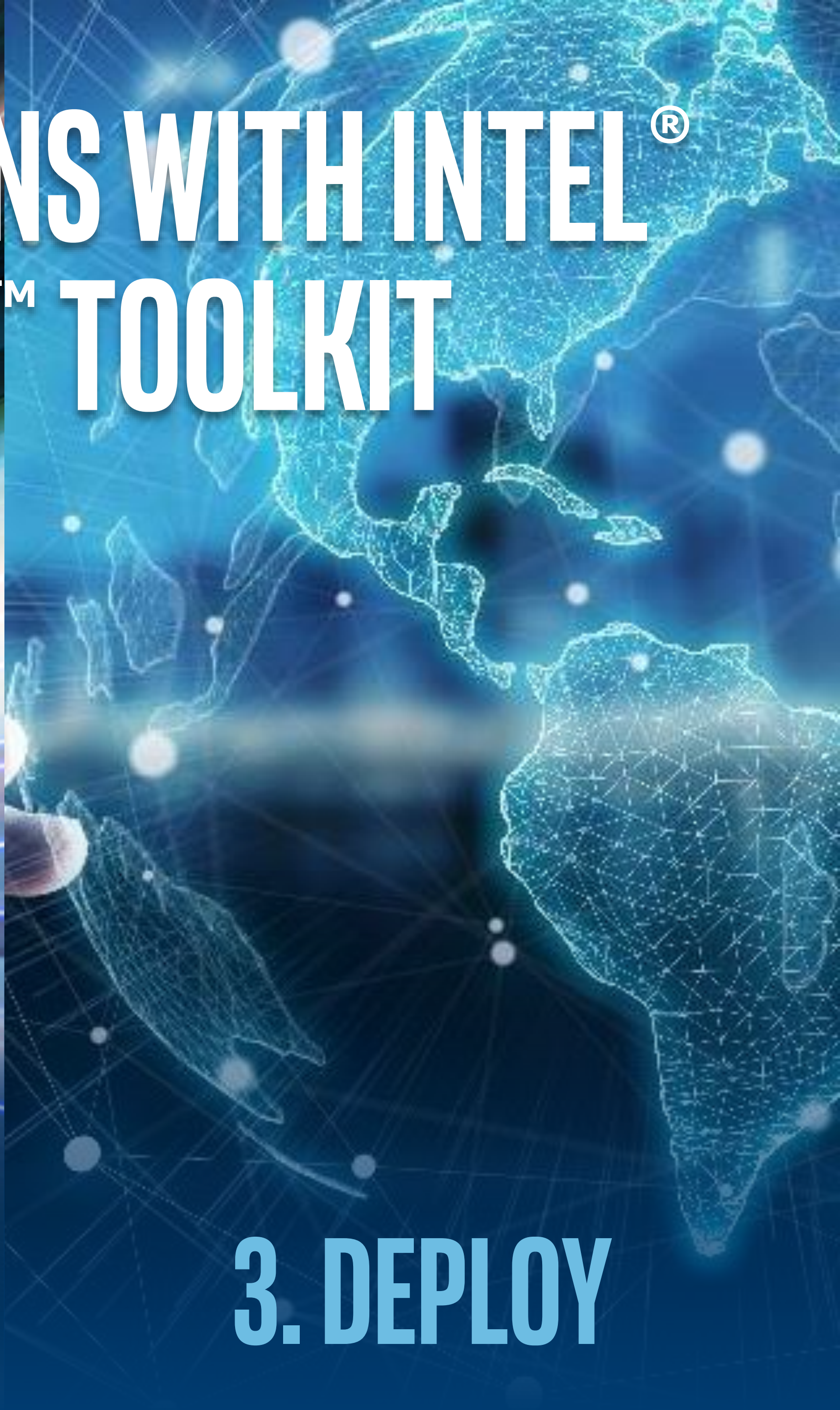
Write Once,
Deploy Anywhere

DEPLOY DEEP LEARNING SOLUTIONS WITH INTEL® DISTRIBUTION OF OpenVINO™ TOOLKIT

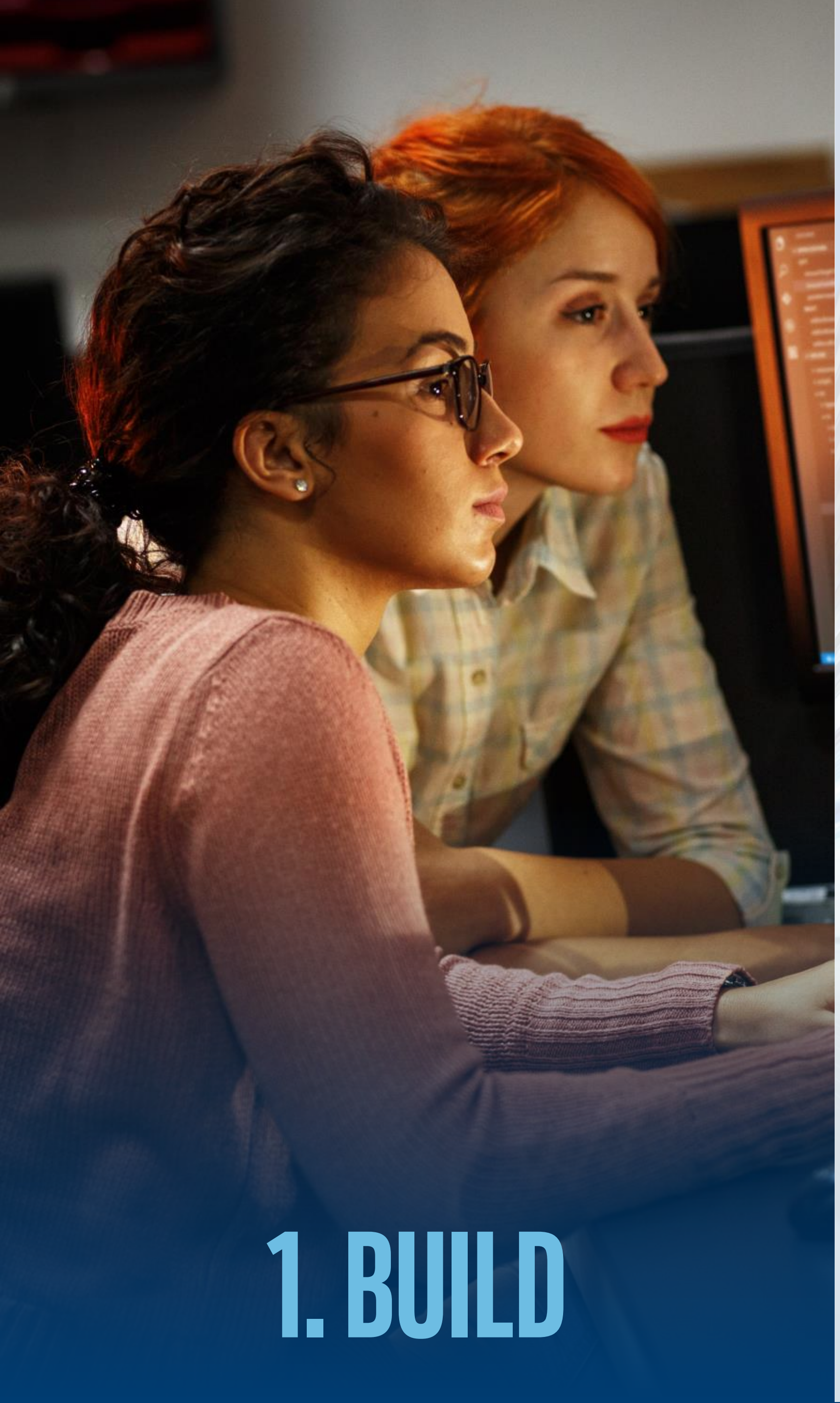


1. BUILD

2. OPTIMIZE



3. DEPLOY

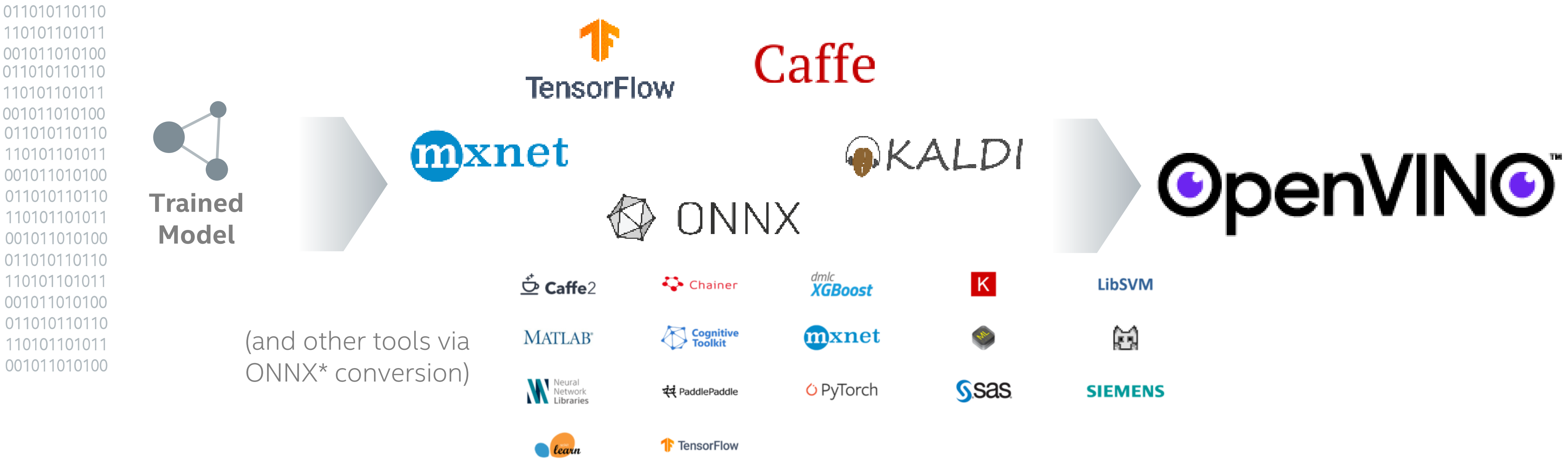


1. BUILD

2. OPTIMIZE

3. DEPLOY

BREADTH OF SUPPORTED FRAMEWORKS MAXIMIZES DEVELOPMENT



Trained Model

(and other tools via ONNX* conversion)

Supported Frameworks and Formats ▸ https://docs.openvinotoolkit.org/latest/_docs_IE_DG_Introduction.html#SupportedFW

Configure the Model Optimizer for your Framework ▸ https://docs.openvinotoolkit.org/latest/_docs_MO_DG_prepare_model_Config_Model_Optimizer.html

Optimization Notice

1. BUILD

2. OPTIMIZE

3. DEPLOY

FROM OPTIMIZATION TO DEPLOYMENT

OpenVINO

Model Optimizer

- A Python-based tool to import trained models and convert them to Intermediate Representation
- Optimizes for performance or space with conservative topology transformations
- Hardware-agnostic optimizations

Development Guide ▸
https://docs.openvinotoolkit.org/latest/_docs_MO_DG_Deep_Learning_Model_Optimizer_DevGuide.html

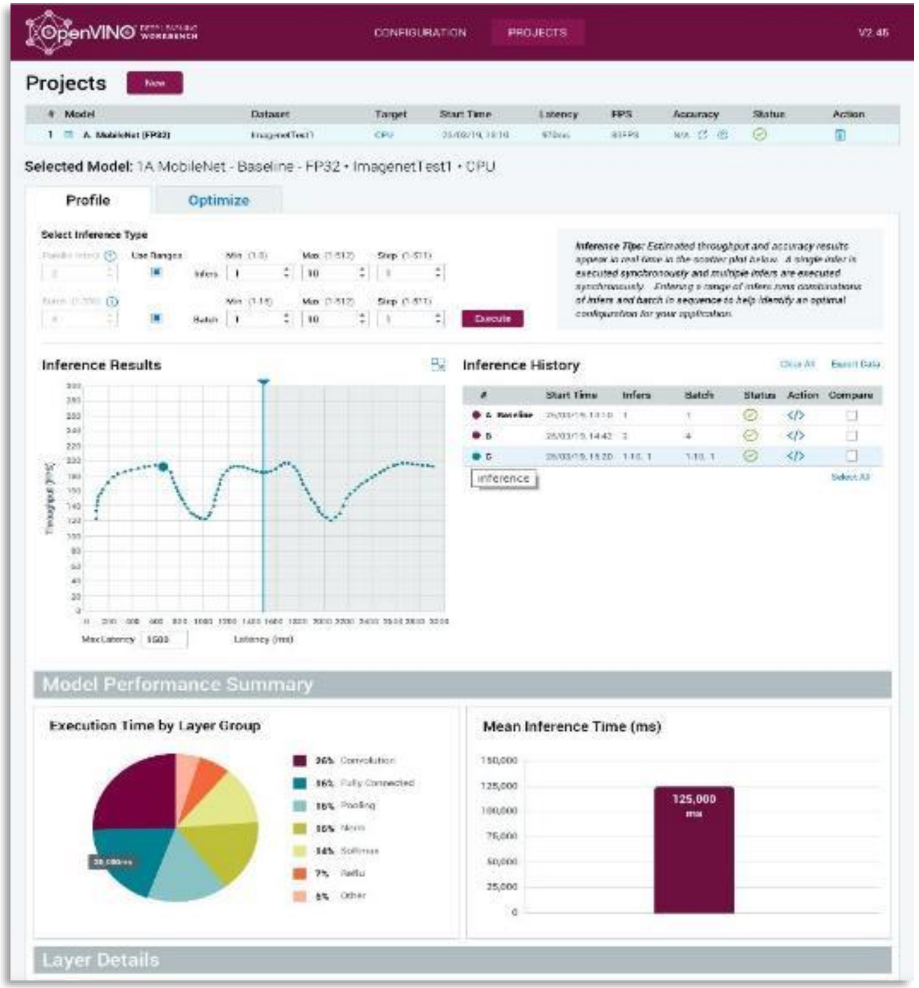
Inference Engine

- High-level, C/C++ and Python, inference API
- Interface is implemented as dynamically loaded plugins for each hardware type
- Delivers best performance for each type without requiring users to implement and maintain multiple code pathways

Development Guide ▸
https://docs.openvinotoolkit.org/latest/_docs_IE_DG_Deep_Learning_Inference_Engine_DevGuide.html

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



Projects

#	Model	Dataset	Target	Start Time	Latency	FPS	Accuracy	Status	Action
1	A. MobileNet (FP32)	ImagenetTest1	CPU	25/03/19, 13:10	970ms	83FPS	N/A	✓	

Selected Model: 1A MobileNet - Baseline - FP32 • ImagenetTest1 • CPU

Select Inference Type

Use Ranges ☑ Infers Min (1-1): 1 Max (1-512): 10 Step (1-511): 1

☑ Batch Min (1-15): 1 Max (1-512): 10 Step (1-511): 1

Inference Tips: Estimated throughput and accuracy results appear in real time in the scatter plot below. A single infer is executed synchronously and multiple infers are executed asynchronously. Entering a range of infers runs combinations of infers and batch in sequence to help identify an optimal configuration for your application.

Inference Results

Max Latency 1500

Inference History

#	Start Time	Infers	Batch	Status	Action	Compare
● A. Baseline	25/03/19, 13:10	1	1	✓	</>	□
● B	25/03/19, 14:42	3	4	✓	</>	□
● C	25/03/19, 15:20	1:10, 1	1:10, 1	✓	</>	□

Model Performance Summary

Execution Time by Layer Group

Mean Inference Time (ms)

Layer Details

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

1. BUILD

2. OPTIMIZE

3. DEPLOY

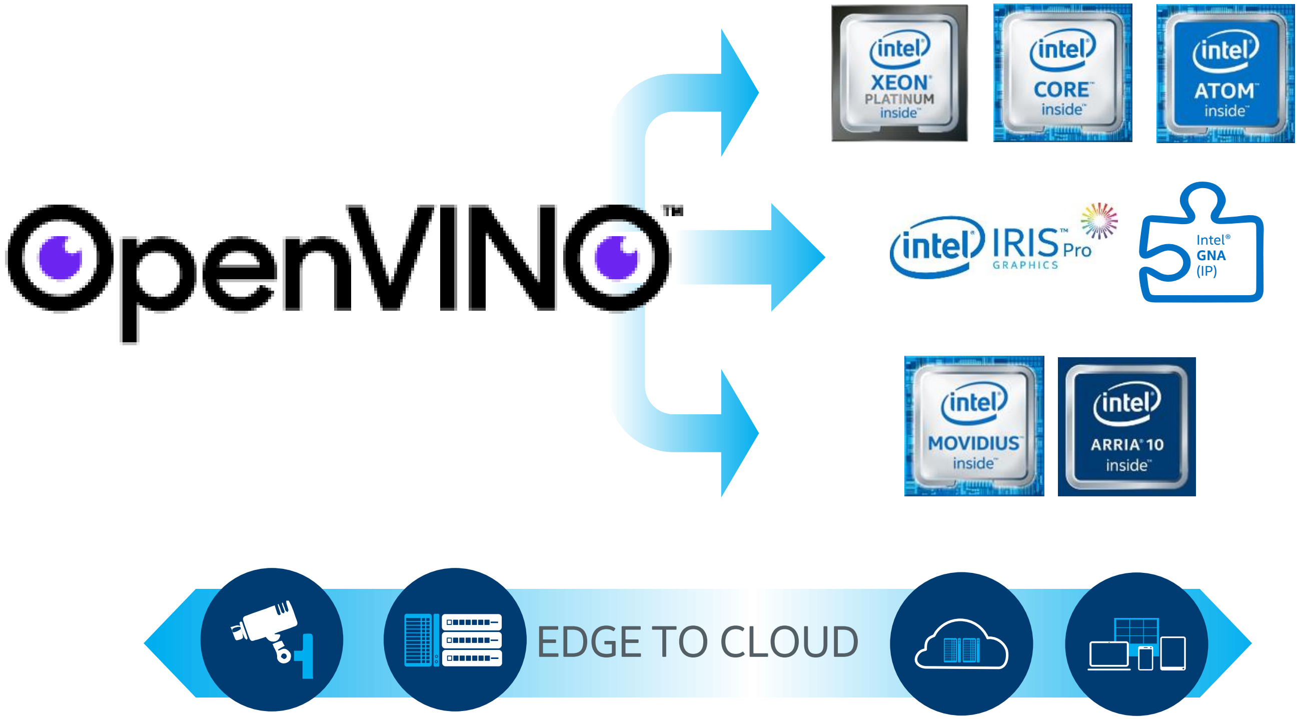
WRITE ONCE, DEPLOY ANYWHERE

Cross-Platform Flexibility on Intel® Distribution of OpenVINO™ toolkit

Write once, deploy across different platforms with the same API and framework-independent execution

Consistent accuracy, performance and functionality across all target devices with no re-training required

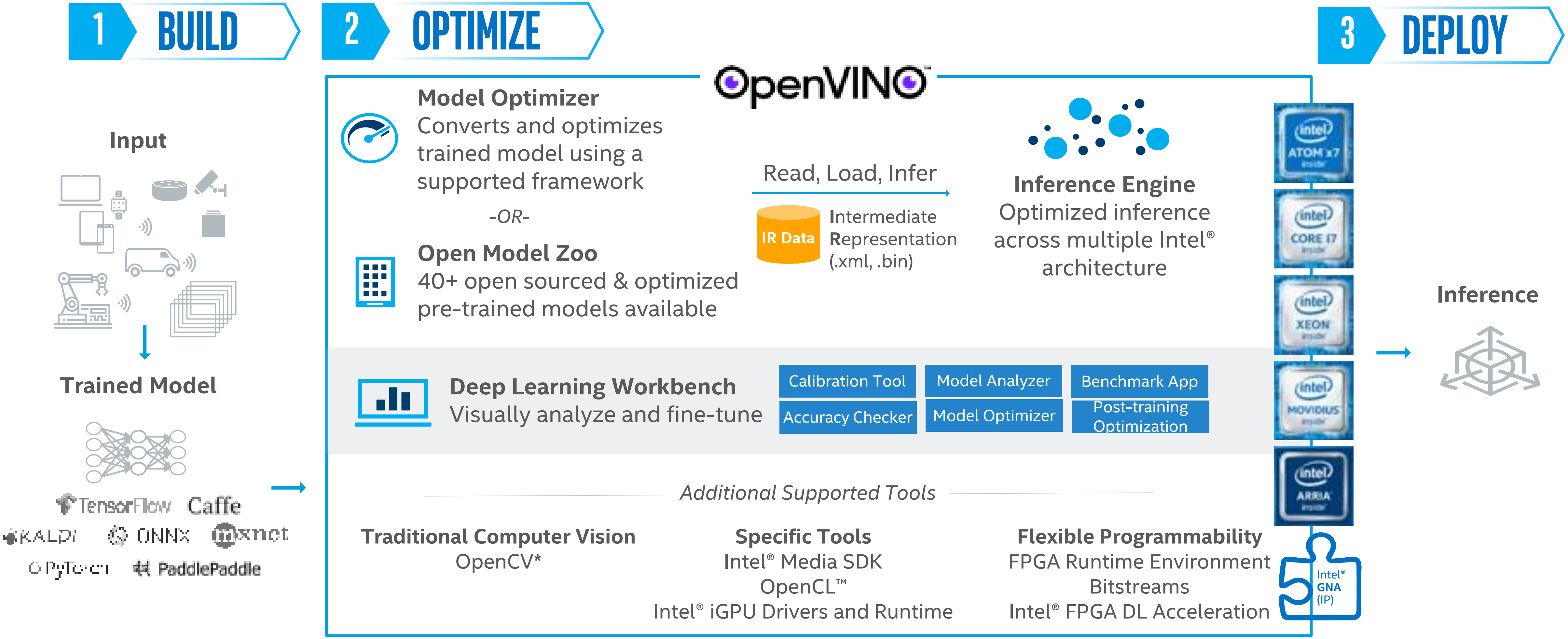
[NEW] Full environment utilization, or multi-device plugin, across available hardware for greater performance results



Introduction ▸ https://docs.openvinotoolkit.org/latest/_docs_IE_DG_supported_plugins_HETERO.html

Optimization Notice

STREAMLINED AND OPTIMIZED AI INFERENCING WORKFLOW



1 BUILD

2 OPTIMIZE

3 DEPLOY

Input

Trained Model

TensorFlow Caffe KALDI ONNX mxnet PyTorch PaddlePaddle

Model Optimizer
Converts and optimizes trained model using a supported framework

-OR-

Open Model Zoo
40+ open sourced & optimized pre-trained models available

Read, Load, Infer

IR Data Intermediate Representation (.xml, .bin)

Inference Engine
Optimized inference across multiple Intel® architecture

Deep Learning Workbench
Visually analyze and fine-tune

Calibration Tool	Model Analyzer	Benchmark App
Accuracy Checker	Model Optimizer	Post-training Optimization

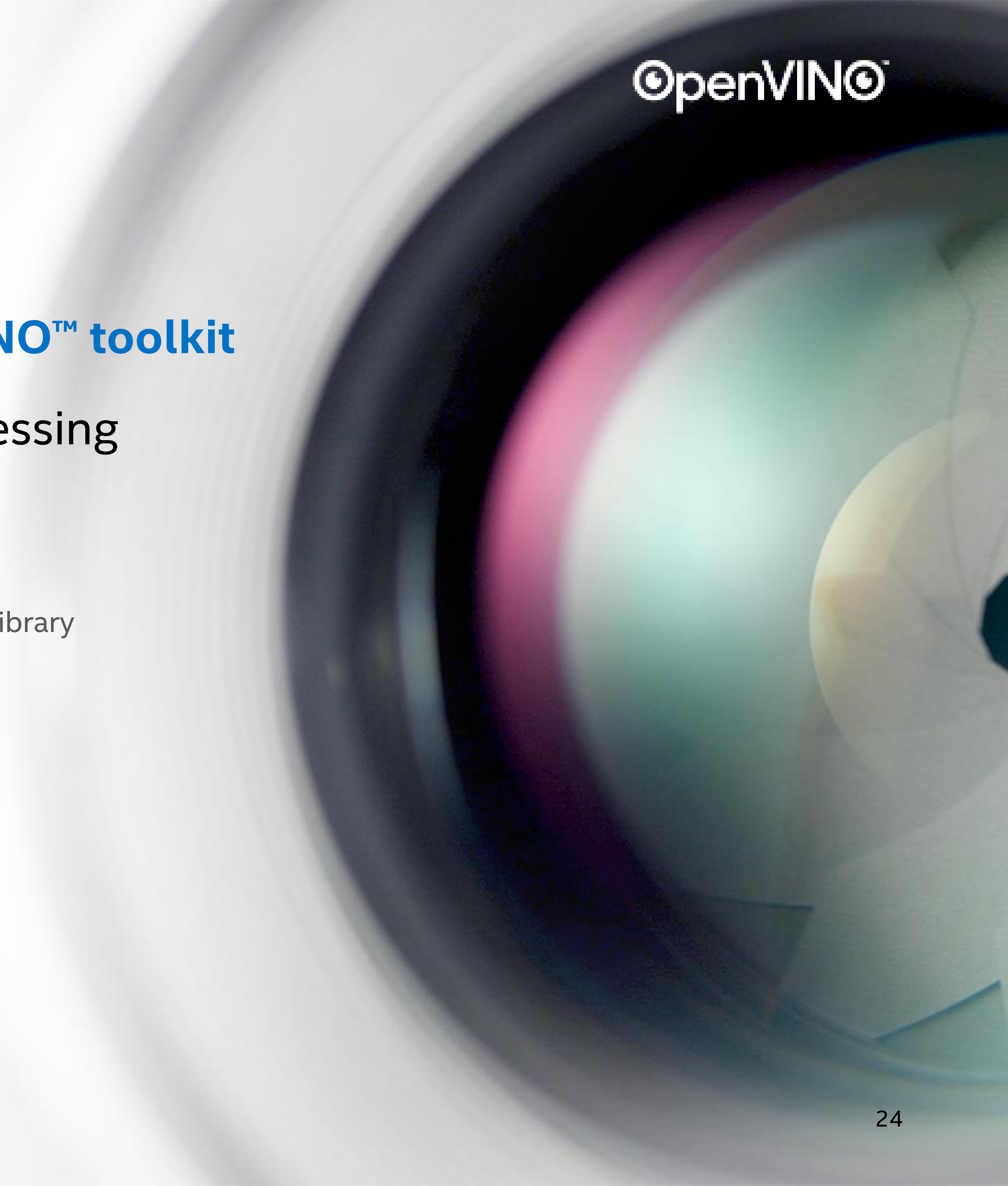
Additional Supported Tools

Traditional Computer Vision
OpenCV*

Specific Tools
Intel® Media SDK
OpenCL™
Intel® iGPU Drivers and Runtime

Flexible Programmability
FPGA Runtime Environment
Bitstreams
Intel® FPGA DL Acceleration

Inference



TRADITIONAL COMPUTER VISION

Powered by the Intel® Distribution of OpenVINO™ toolkit

Accelerate and optimize low-level, image-processing capabilities using OpenCV

https://opencv.org/

- Open sourced computer vision and machine learning library
- 2500+ algorithms for a common infrastructure and to accelerate time-to-market
- Large number of primitives for customizability

TOOLS TO SPEED UP TEST CYCLES AND DEVELOPMENT

[NEW] **Post-training Optimization**

- Reduce model size into low precision data types, such as INT8
- Reduces model size while also improving latency

Model Analyzer

- Provides theoretical data on models: computational complexity (flops), number of neurons, memory consumption

Benchmark App

- Measure performance (throughput, latency) of a model
- Get performance metrics per layer and overall basis

Deployment Manager

- Generate an optimal, minimized runtime package for deployment
- Deploy with smaller footprint compared to development package

Accuracy Checker

- Check for accuracy of the model (original and after conversion) to IR file using a known data set

Model Downloader

- Provides an easy way of accessing a number of public models as well as a set of pre-trained Intel models

Get Started ▸ https://docs.openvinotoolkit.org/latest/_docs_IE_DG_Tools_Overview.html –or- by using the Deep Learning Workbench

OpenVINO

SPEED UP DEVELOPMENT USING THE OPEN MODEL ZOO

Open source resources with pre-trained models, samples and demos

Computer Vision

Object detection

Object recognition

Reidentification

Semantic segmentation

Instance segmentation

Human pose estimation

Image processing

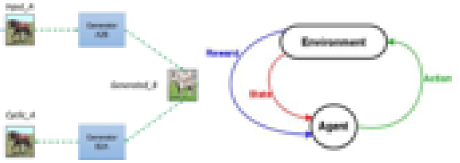
Audio, Speech, Language

Text detection

Text recognition

Recommender

Action recognition



Other

(Data Generation, Reinforcement Learning)

Compression models

Image retrieval

And more..

PRE-TRAINED MODELS

https://github.com/opencv/open_model_zoo

Optimization Notice

SPEED UP DEVELOPMENT USING THE OPEN MODEL ZOO

Open source resources with pre-trained models, demos, and tools

The Open Model Zoo demo applications are console applications that demonstrate how you can use your applications to solve specific use-cases.

Smart Classroom
Recognition and action detection demo for classroom settings

Multi-Camera, Multi-Person
Tracking multiple people on multiple cameras for public safety use cases

Gaze Estimation
Face detection followed by gaze estimation, head pose estimation and facial landmarks regression.

Super Resolution
Enhances the resolution of the input image

Action Recognition
Classifies actions that are being performed on input video

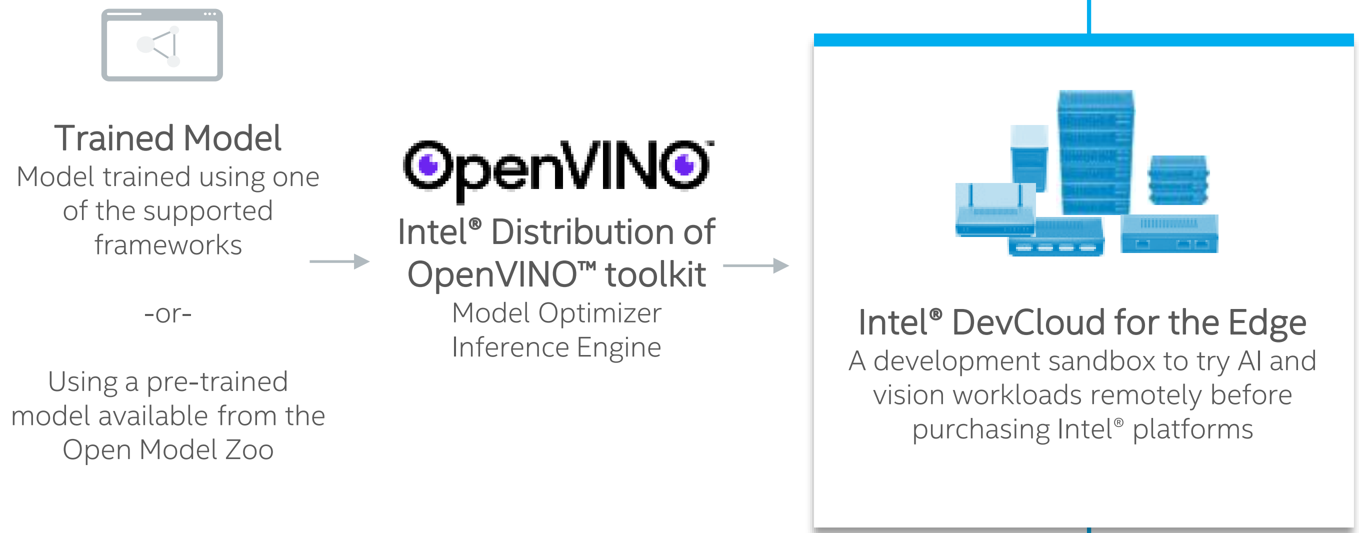
And more..

DEMO APPLICATIONS
https://github.com/opencv/open_model_zoo

Optimization Notice

TEST HARDWARE WITH THE INTEL® DEVCLOUD FOR THE EDGE

Powered by Intel® Distribution of OpenVINO™ toolkit



Trained Model
Model trained using one of the supported frameworks

-or-

Using a pre-trained model available from the Open Model Zoo

→

OpenVINO
Intel® Distribution of OpenVINO™ toolkit
Model Optimizer
Inference Engine

→

Intel® DevCloud for the Edge
A development sandbox to try AI and vision workloads remotely before purchasing Intel® platforms

- Prototype on the latest hardware and software to future proof your solution
- Benchmark your customized AI application
- Run AI applications from anywhere in the world
- Reduce development time and cost

https://devcloud.intel.com/edge/

Deploy and Scale

INTEL® MEDIA SDK

Speed Up Video Encoding, Decoding and Processing

AN API TO ACCESS INTEL® QUICK SYNC VIDEO HARDWARE-ACCELERATED ENCODE/DECODE AND PROCESSING SUPPORTING

40+ video quality and performance pre-processing

Features:

- Color Conversion
- Scaling
- De-interlacing
- De-noising
- Frame Rate Conversion
- Composition
- Rotate
- On-Screen-Display and Composition
- De-warp

Codecs:
H.265/HEVC
H.264/AVC
JPEG/MJPEG
MPEG2
VP8/VP9

Rate-Control:
(9+ BRC methods)
CQP (I/P/B and manual)
CBR (target bit-rate)
VBR (AVBR and CVBR)

Build High-Performance Media Pipelines at Low Cost

Use hardware acceleration of Intel® Xeon®, Intel® Core™ and Intel Atom® processors for premium performance

Embed Enterprise-Grade Codecs for Quick Time to Market

Accelerated HEVC, AVC, and MPEG-2 decode, encode, and transcode.

AAC, MP3 and MPEG-audio codecs

Stay Competitive
Transition to
4K and HEVC

Deliver real-time 4K HEVC on latest platforms

Use HEVC software and GPU-acceleration to tune and optimize for specific scenarios

Use Analyzer and Test Tools to Save Time and Reduce Engineering/Development Effort

Validate for compliance and robustness with Intel® Stress Bitstreams and Encoder

Inspect and debug with Intel® Video Pro Analyzer

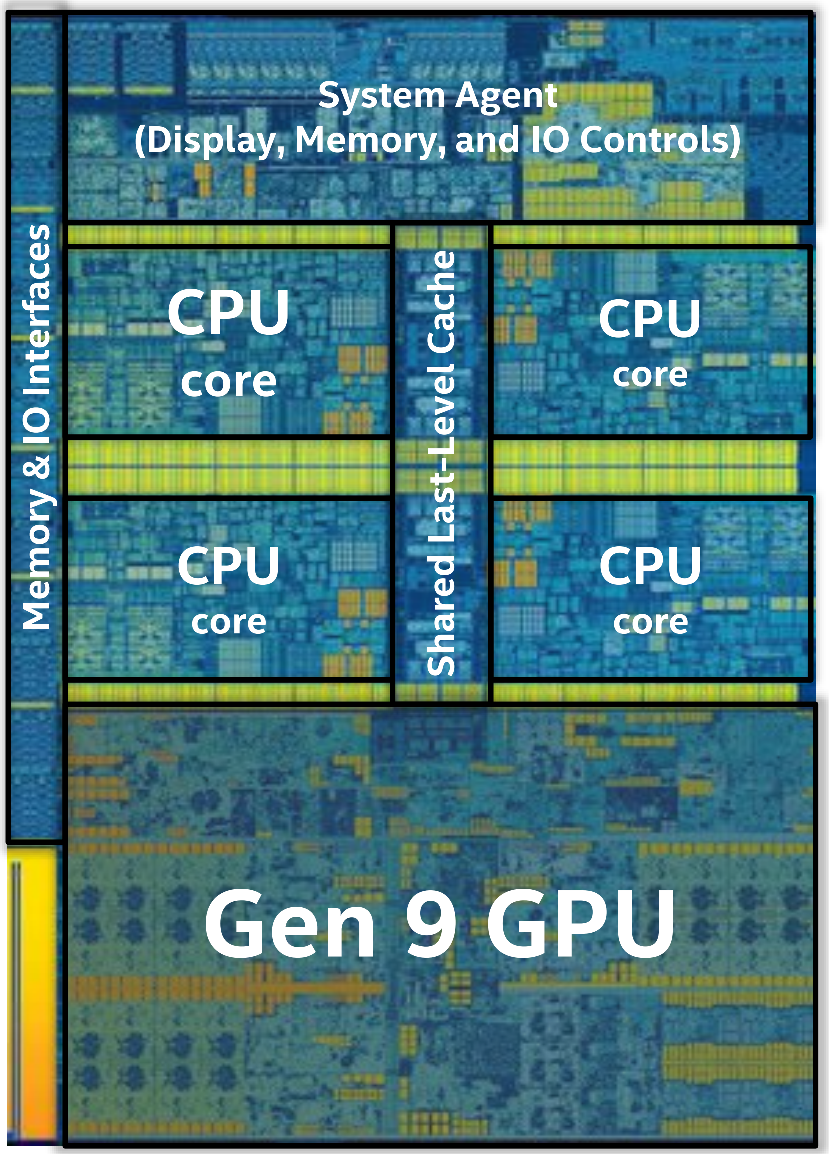
Documentation ▸ https://software.intel.com/en-us/articles/the-openvino-toolkit-and-the-intel-media-sdk-part-1

Optimization Notice

INTEL INTEGRATED GRAPHICS

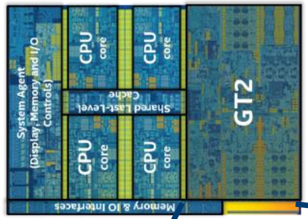
Gen is the internal name for Intel's on-die GPU solution. It's a hardware ingredient with various configurations.

- Intel® Core™ Processors include Gen hardware.
- Gen GPUs can be used for graphics and also as general compute resources.
- Libraries contained in the Intel® Distribution of OpenVINO™ toolkit (and many others) support Gen offload using OpenCL™.



6th Generation Intel® Core™ i7 (Skylake) Processor

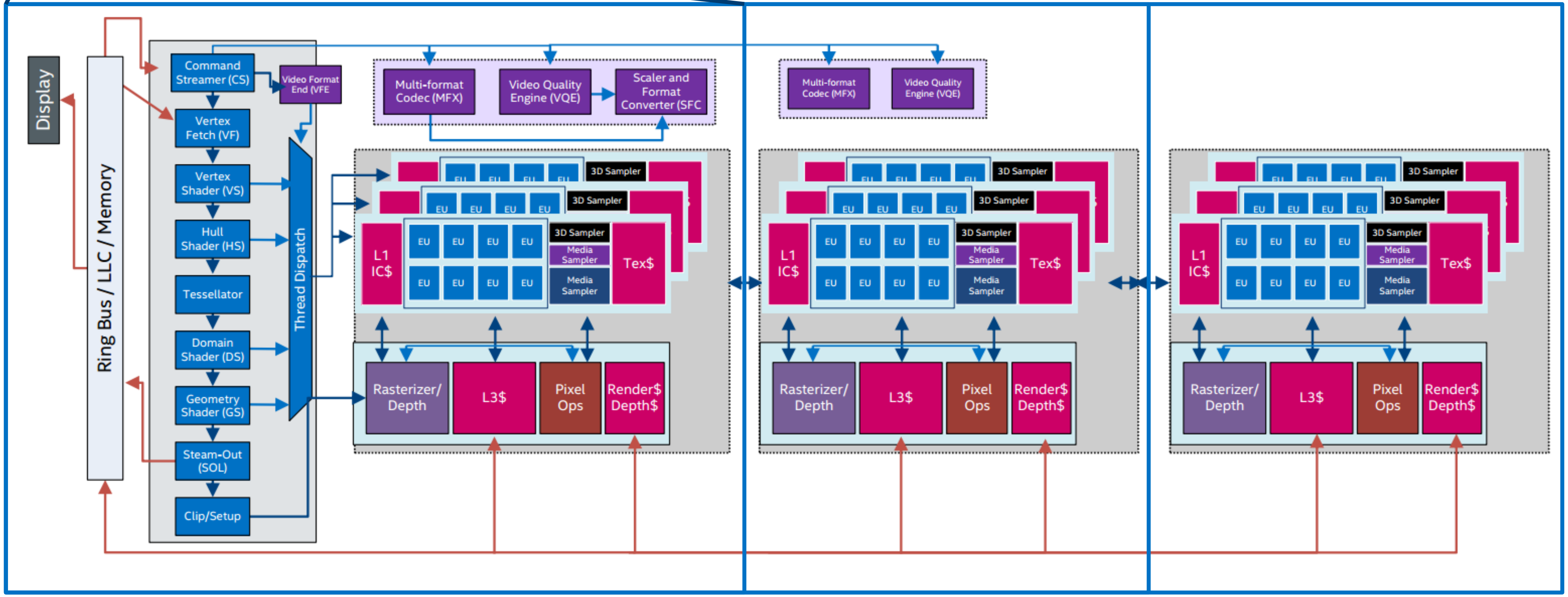
INTEL GPU CONFIGURATIONS



GT2
Intel® HD Graphics
24 EUs, 1 MFX

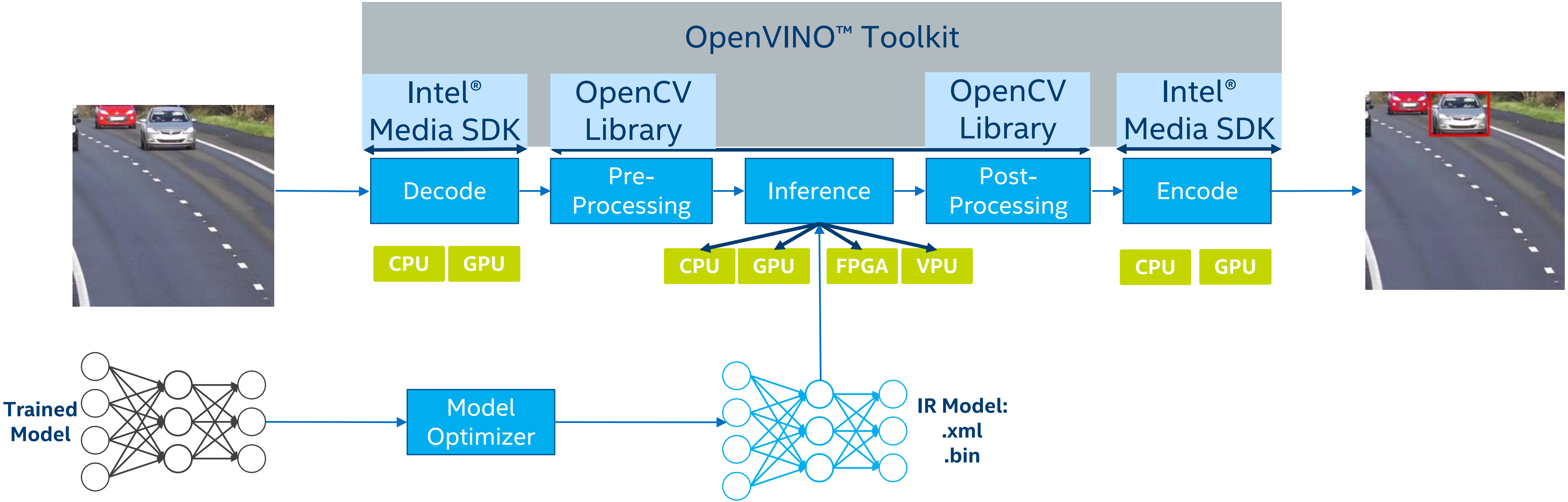
GT3
Intel® Iris® Graphics
48 EUs, 2 MFX

GT4
Intel® Iris® Pro Graphics
72 EUs, 2 MFX



Workflow of Applying OpenVINO™ in CV Applications, Accelerate Streaming Performance

Using Intel® Media SDK and the OpenVINO™ toolkit together enables customers to build high performance, intelligent vision solutions.



GETTING STARTED WITH INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

Recommendations to the customer or developer

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

intel®