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.

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.

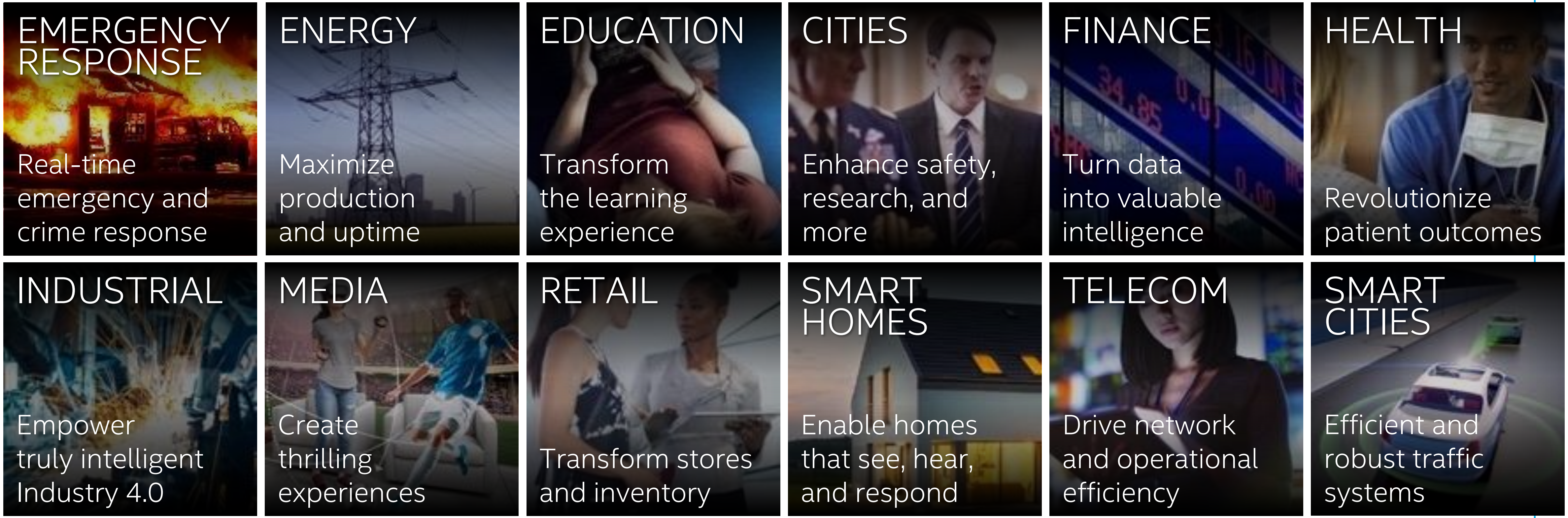
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

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.

Copyright © 2018, Intel Corporation. All rights reserved.

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

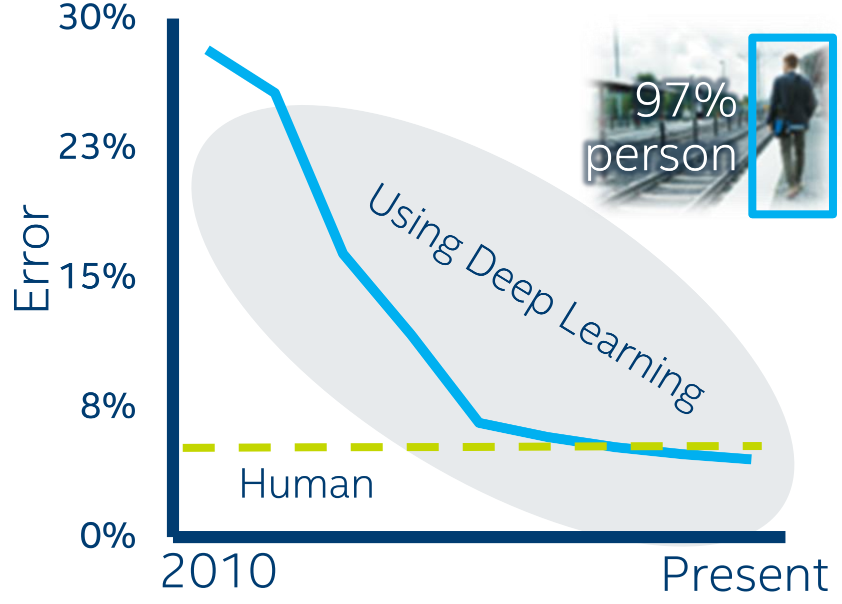
MANUFACTURING

PUBLIC SECTOR

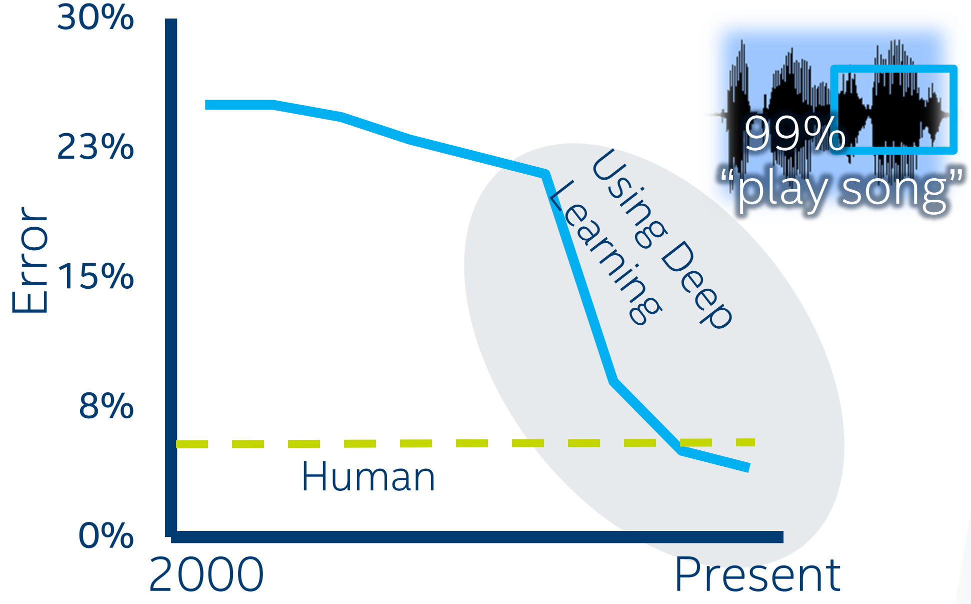
DEEP LEARNING BREAKTHROUGHS AND OPPORTUNITIES

Machines able to meet or exceed human image and speech recognition

Image Recognition



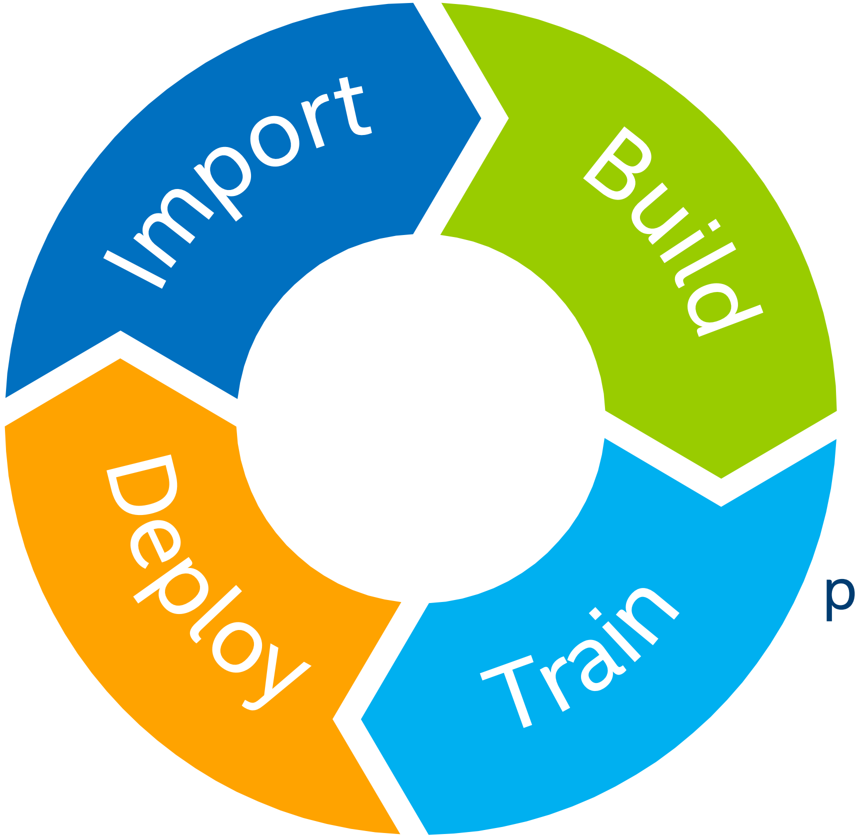
Speech Recognition



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

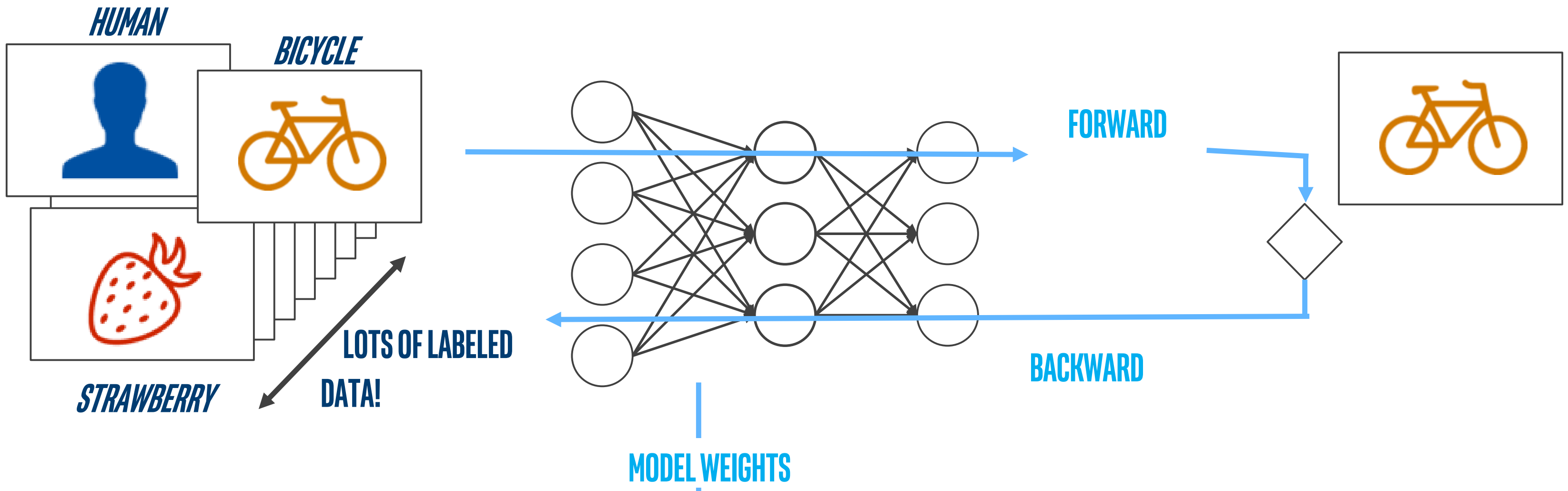
Adjust models to meet performance and accuracy objectives

Integrate trained models with application code

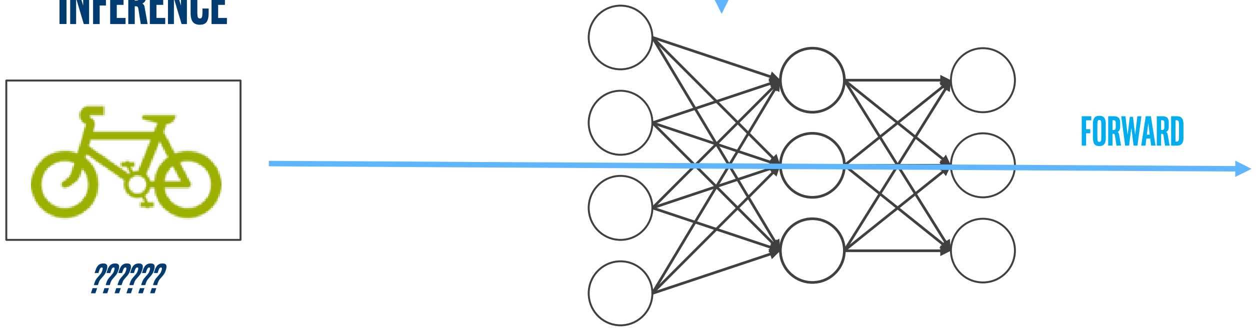
Intel® Distribution OpenVINO™ Toolkit Provides Deployment from Intel® Edge to Cloud

DEEP LEARNING: TRAINING VS. INFERENCE

TRAINING

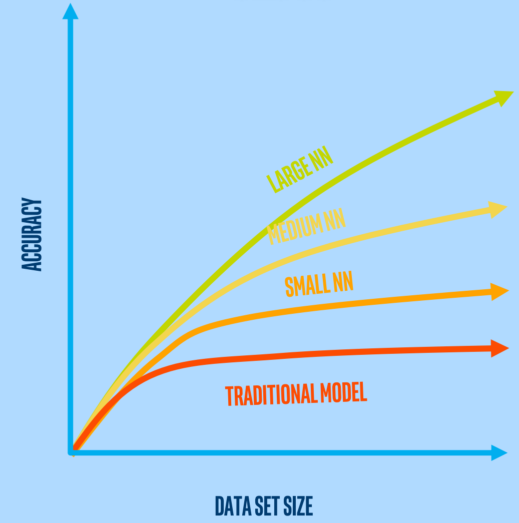

HUMAN
BICYCLE
STRAWBERRY
LOTS OF LABELED DATA!
FORWARD
BACKWARD
MODEL WEIGHTS

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

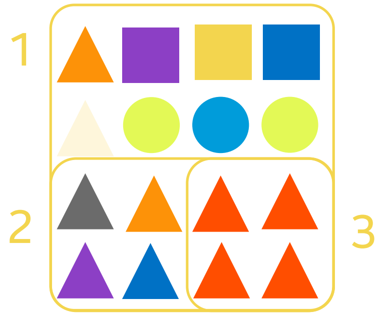
Diverse requirements for myriad use cases require unique approaches

Inability to meet use-case specific requirements

AI COMPUTE CONSIDERATIONS

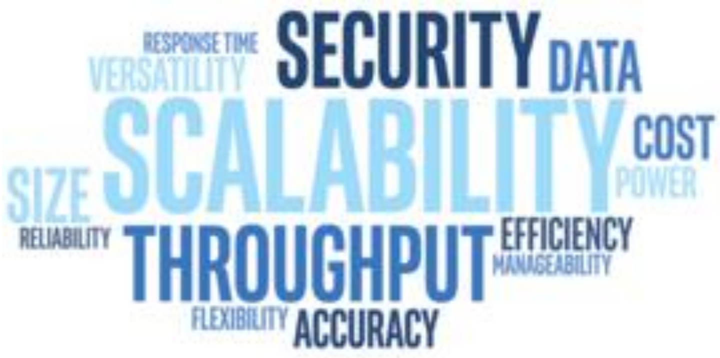
How do you determine the right computing for your AI needs?

WORKLOADS



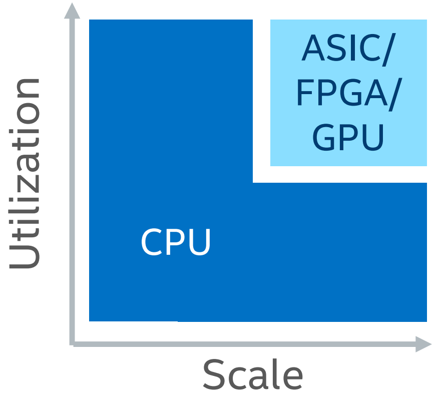
What is my workload profile?

REQUIREMENTS


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

What are my use case requirements?

DEMAND

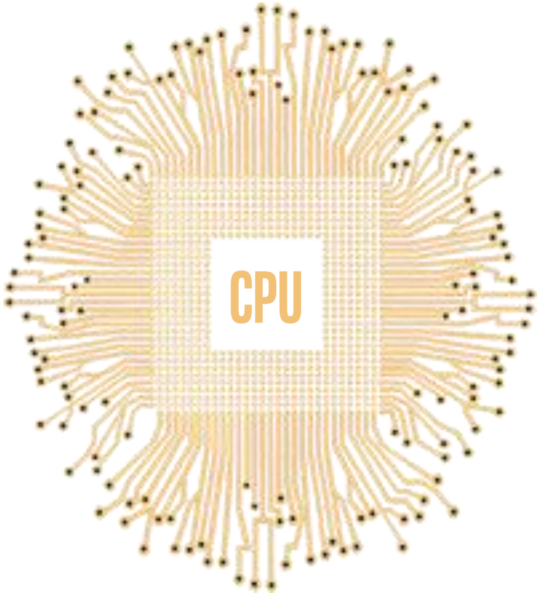

ASIC/
FPGA/
GPU
CPU
Utilization
Scale

How prevalent is AI in my environment?

Optimization Notice

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

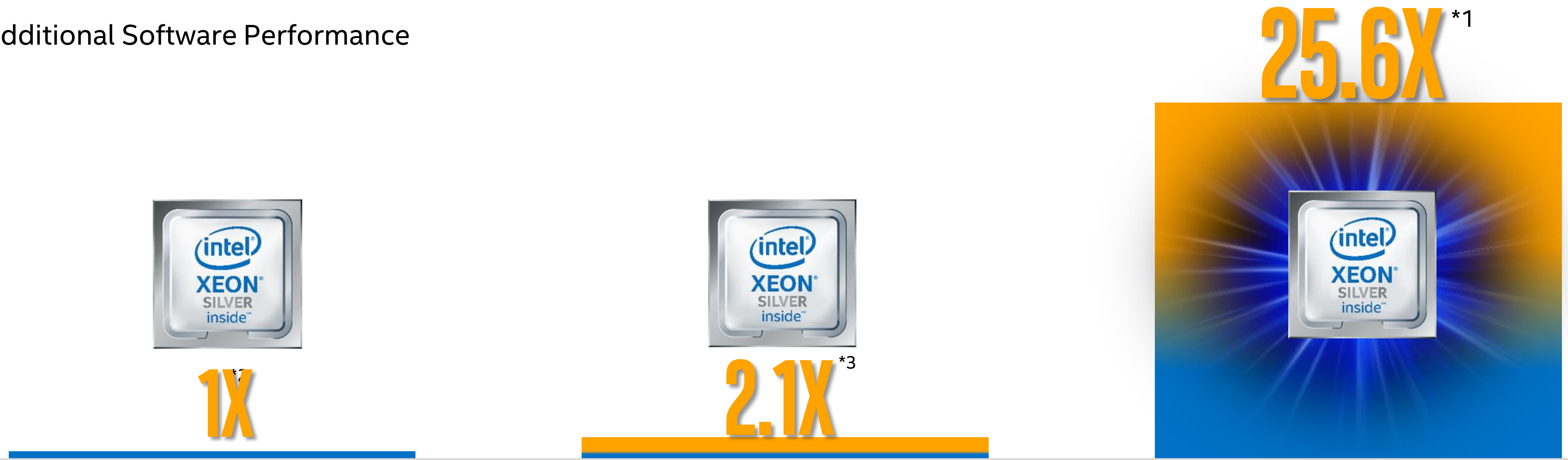
Write Once,
Deploy Anywhere

THE COMPOUNDING EFFECT OF BOTH HARDWARE AND SOFTWARE

Improvements Means Exponential Performance

🟦 Baseline Performance

🟧 Additional Software Performance



OpenVINO™ Release 2018 R1

OpenVINO™ Release 2019 R1

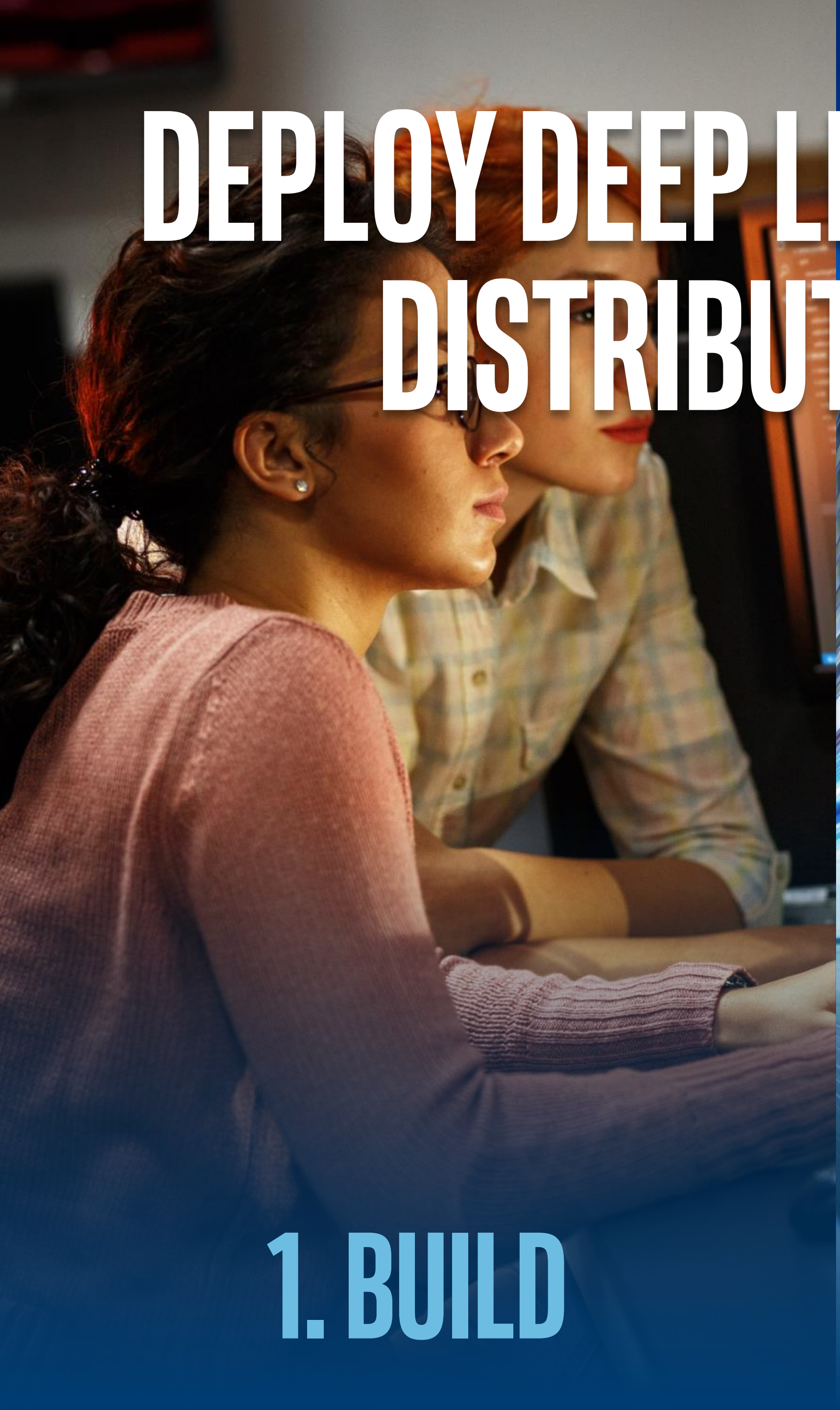
1st Generation Intel® Xeon Scalable Processor

OpenVINO™ Release 2019 R3

2nd Generation Intel® Xeon Scalable Processor

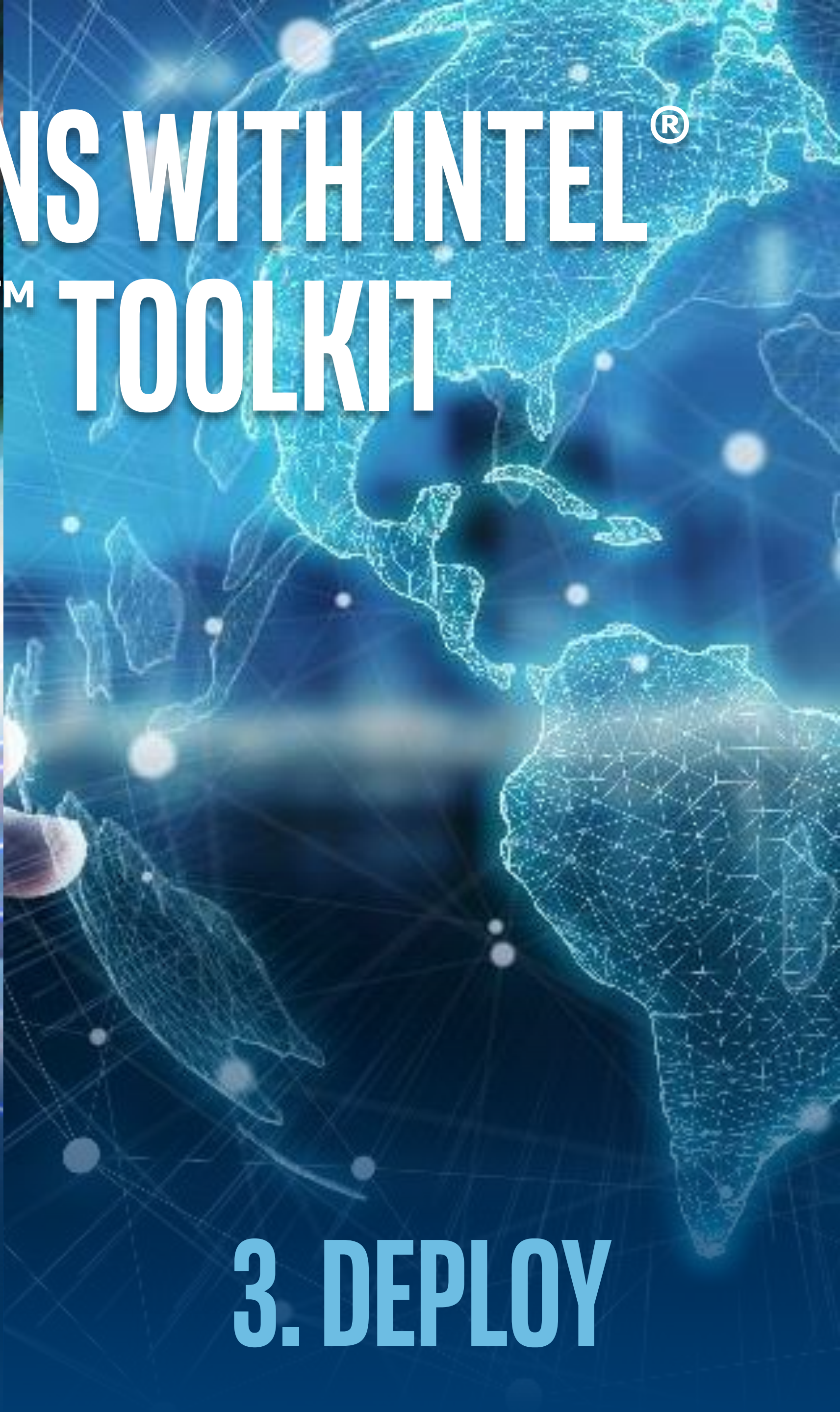
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See backup for configuration details.
Comparison of Frames Per Second utilizing Mobilenet SSD, Batch 1.

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

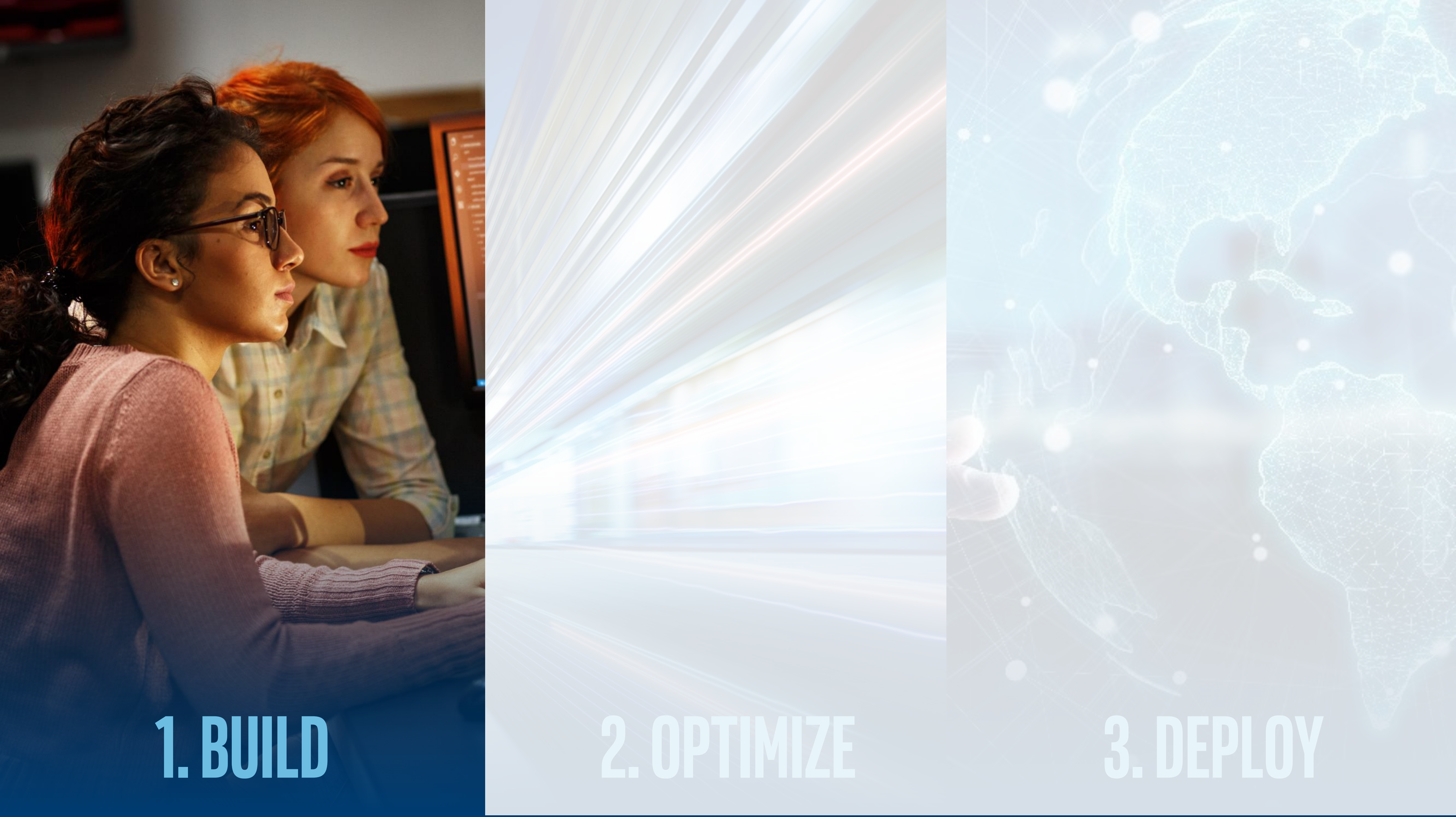


1. BUILD

2. OPTIMIZE



3. DEPLOY

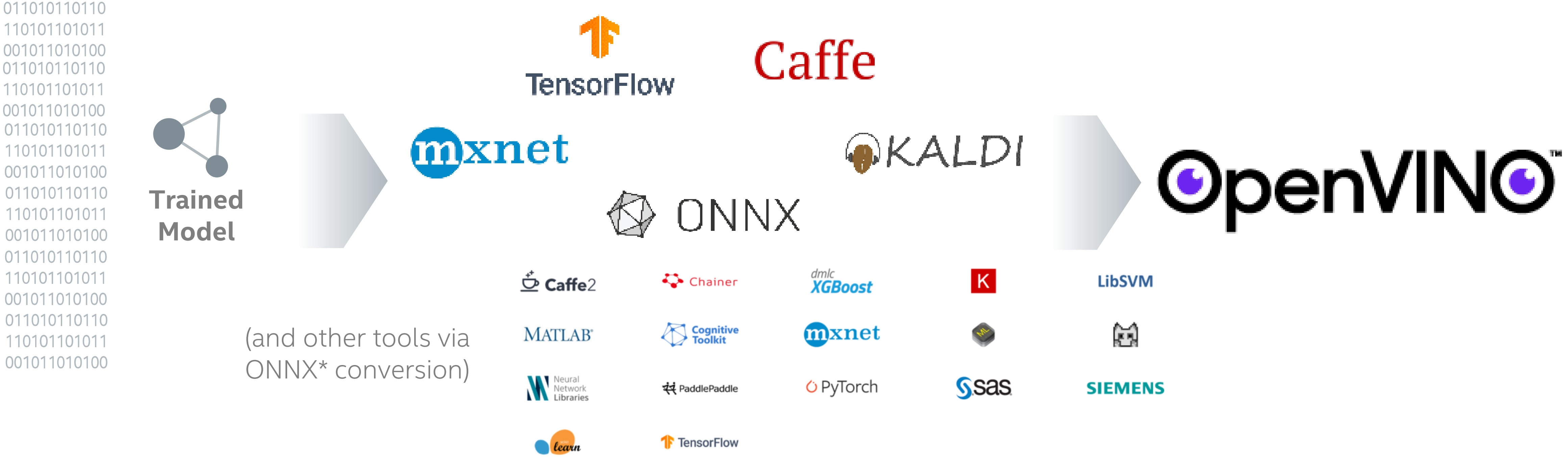


1. BUILD

2. OPTIMIZE

3. DEPLOY

BREADTH OF SUPPORTED FRAMEWORKS MAXIMIZES DEVELOPMENT


011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
Trained Model
TensorFlow
Caffe
mxnet
KALDI
ONNX
OpenVINO™
Caffe2
Chainer
dmlc XGBoost
K
LibSVM
MATLAB
Cognitive Toolkit
mxnet
PyTorch
sas
SIEMENS
Neural Network Libraries
PaddlePaddle
learn
TensorFlow

(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

1. BUILD

2. OPTIMIZE

3. DEPLOY

FROM OPTIMIZATION TO DEPLOYMENT

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

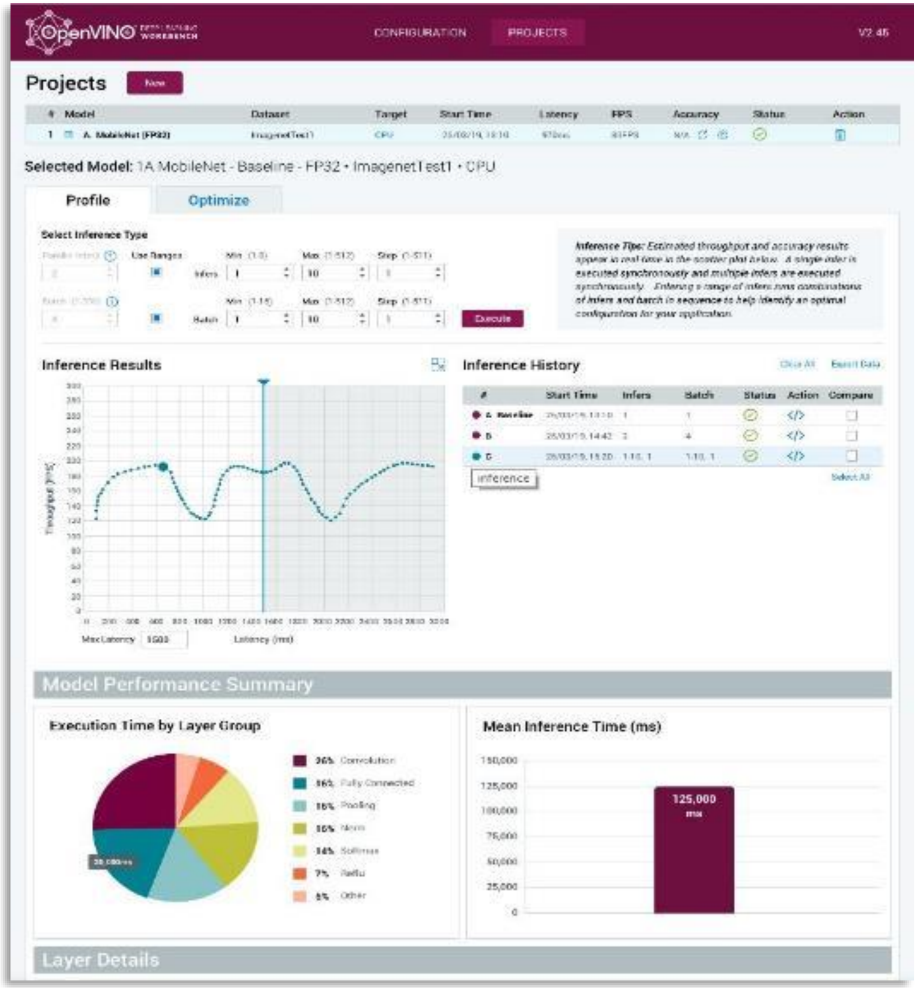
Inference Engine

- 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


OpenVINO DEEP LEARNING WORKBENCH
CONFIGURATION
PROJECTS
V2.45
Projects
New
Model
Dataset
Target
Start Time
Latency
FPS
Accuracy
Status
Action
Selected Model: 1A MobileNet - Baseline - FP32 • ImagenetTest1 • CPU
Profile
Optimize
Select Inference Type
Execute
Inference Results
Inference History
Latency (ms)
Model Performance Summary
Execution Time by Layer Group
26% Convolution
16% Fully Connected
16% Pooling
16% Norm
14% Softmax
7% ReLu
5% Other
Mean Inference Time (ms)
125,000 ms
Layer Details

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

For public use – OK for non-NDA disclosure

1. BUILD

2. OPTIMIZE

3. DEPLOY

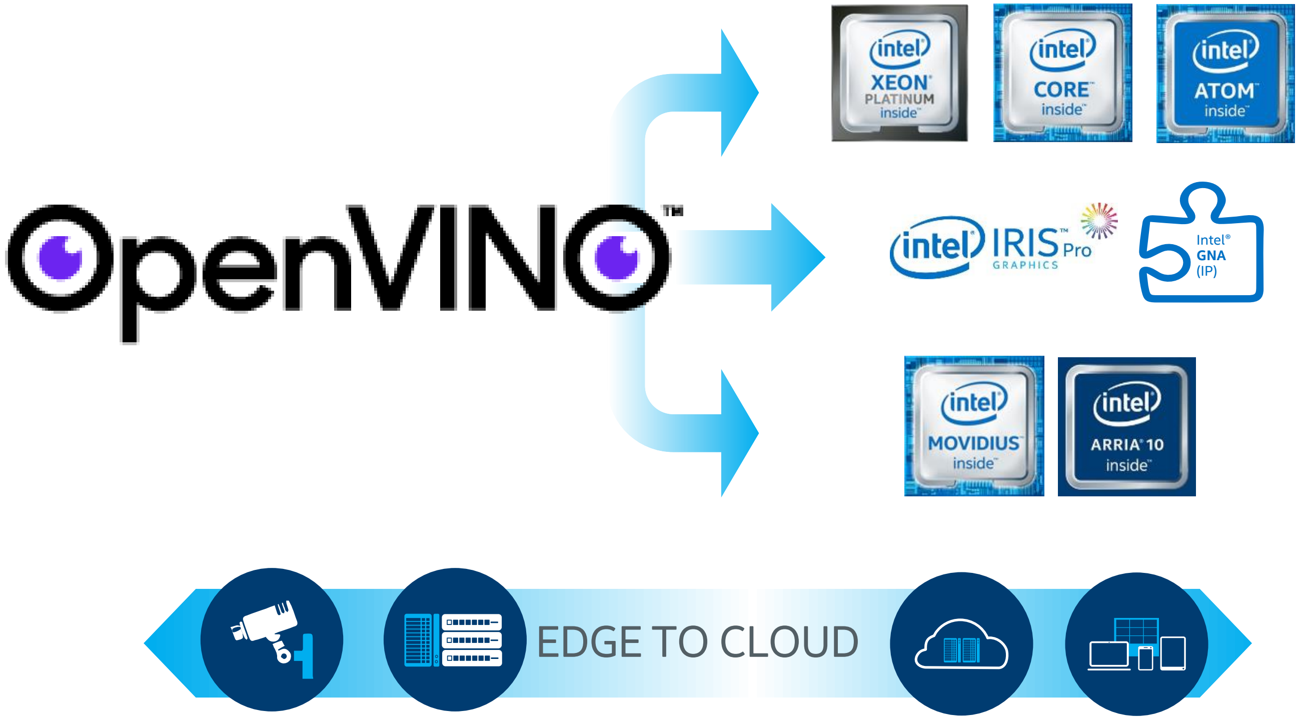
WRITE ONCE, DEPLOY ANYWHERE

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


OpenVINO™
intel XEON PLATINUM inside
intel CORE inside
intel ATOM inside
intel IRIS Pro GRAPHICS
Intel® GNA (IP)
intel MOVIDIUS inside
intel ARRIA 10 inside
EDGE TO CLOUD

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

Deployment Manager

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

Model Analyzer

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

Accuracy Checker

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

Benchmark App

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

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

Optimization Notice

SPEED UP DEVELOPMENT USING THE OPEN MODEL ZOO

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

Computer Vision

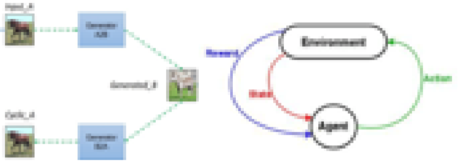
- Object detection
- Object recognition
- Reidentification
- Semantic segmentation
- Instance segmentation
- Human pose estimation
- Image processing

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

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

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

→

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

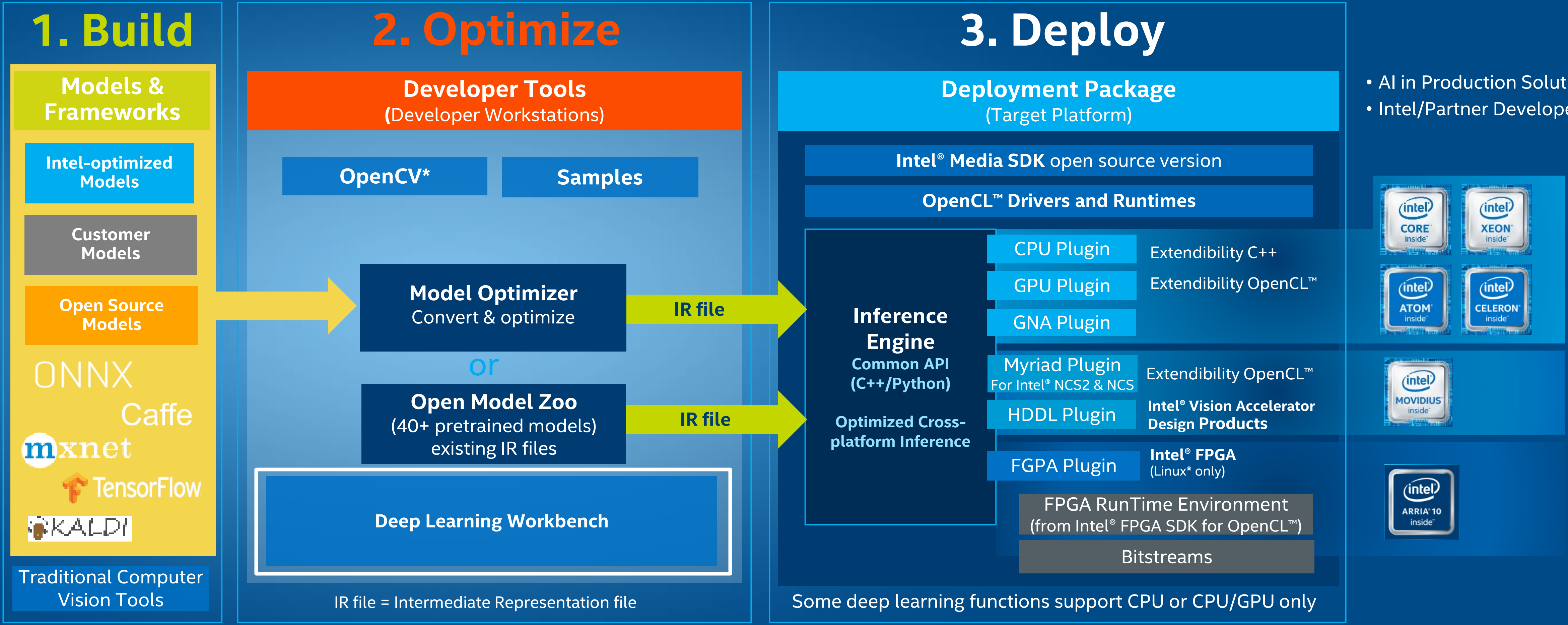
- Prototype on the latest hardware and software to future proof your solution
- Benchmark your customized AI application
- Run AI applications from anywhere in the world
- Reduce development time and cost

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

Deploy and Scale

USING THE INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

ADVANCED CAPABILITIES TO STREAMLINE DEEP LEARNING DEPLOYMENT



1. Build

Models & Frameworks

- Intel-optimized Models
- Customer Models
- Open Source Models

ONNX
Caffe
mxnet
TensorFlow
KALDI

Traditional Computer Vision Tools

2. Optimize

Developer Tools (Developer Workstations)

OpenCV*
Samples

Model Optimizer
Convert & optimize

IR file

or

Open Model Zoo
(40+ pretrained models)
existing IR files

IR file

Deep Learning Workbench

IR file = Intermediate Representation file

3. Deploy

Deployment Package (Target Platform)

Intel® Media SDK open source version

OpenCL™ Drivers and Runtimes

Inference Engine
Common API (C++/Python)
Optimized Cross-platform Inference

Plugin	Notes	Hardware
CPU Plugin	Extendibility C++	intel CORE inside intel XEON inside intel ATOM inside intel CELERON inside
GPU Plugin	Extendibility OpenCL™	
GNA Plugin		
Myriad Plugin For Intel® NCS2 & NCS	Extendibility OpenCL™	intel MOVIDIUS inside
HDDL Plugin	**Intel® Vision Accelerator Design Products**	
FGPA Plugin	**Intel® FPGA** (Linux* only)	intel ARRIA 10 inside

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

Some deep learning functions support CPU or CPU/GPU only

- AI in Production Solutions
- Intel/Partner Developer Kits

Intel® NCS = Intel® Neural Compute Stick (VPU)

Optimization Notice

OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

THE COMPOUNDING EFFECT IN PRODUCTION DEPLOYMENTS

Powered by the Intel® Distribution of OpenVINO™ toolkit

Improvements made by pairing together Intel® architecture-based systems and deep learning acceleration powered by the Intel® Distribution of OpenVINO™ toolkit

Alibaba Cloud

Defect Detection

Aluminum alloy die-casting factories improved defect detection accuracy 5X from manual detection to automatic detection. Learn more

GE Healthcare

Cardiac Examination

Cardiac magnetic resonance imaging (MRI) exams to evaluate heart function, heart chamber volumes and myocardial tissue accelerated by 5.5X. Learn more

PHILIPS HEALTHCARE

Medical Imaging

Medical imaging accelerated bone age prediction model by 188X and lung segmentation model by 38X in inference performance. Learn more

Cities Traffic Management

Thailand's Ministry of Transportation saw reduction of avg queue length by 30.5% & reductions in delay by 8.46-24.52% in intersections in Bangkok. Learn more

ADLINK
TECHNOLOGY INC.

Operational Improvement

11X increase in performance on Intel® architecture and 19X with Intel® Vision Accelerator that lead to operational improvements in manufacturing. Learn more

Autonomous Sea Navigation

Autonomous and assisted sea navigation for autonomous ships delivered 4.8X image throughput compared to unoptimized baseline. Learn more

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

INTEL® MEDIA SDK

API to Access Intel® Quick Sync Video: Hardware Accelerated Encoding, Decoding, and Processing

- H.265 (HEVC)
- H.264 (AVC)
- MPEG-2 and more

- Resize, scale, deinterlace
- Color conversion, composition
- Denoise, sharpen, and more

Benefits

- Outstanding performance
- Rich API to tune encoding pipeline
- Future proofed: support new processor without code changes

Targeting Digital Security and Surveillance, Connected Car Applications, and More

Smart Camera

Car Infotainment and Cluster Display

using

Intel Atom®, Pentium®, and Celeron®[1]

and

Embedded Linux*

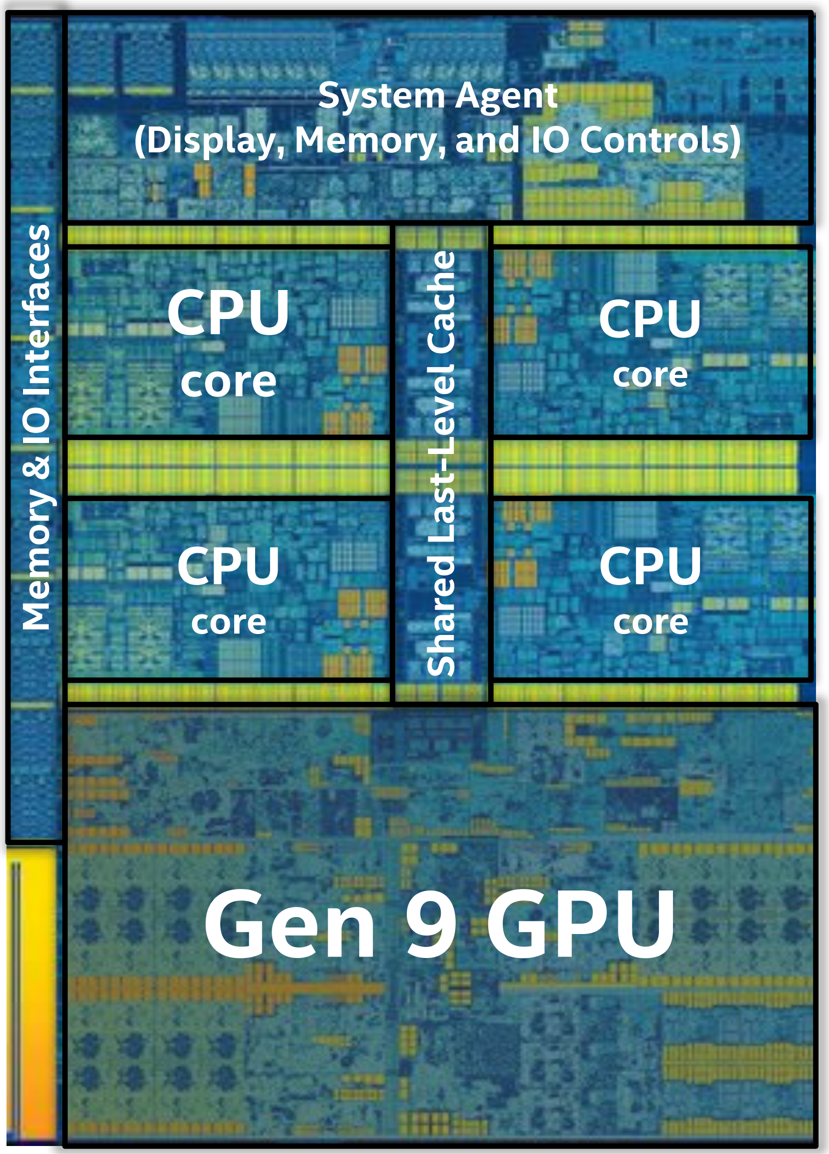
---

[1] Intel® Celeron® Processor N3350, Intel® Pentium® Processor N4200, Intel Atom® E3930, E3940, E3950 processors

INTEL INTEGRATED GRAPHICS

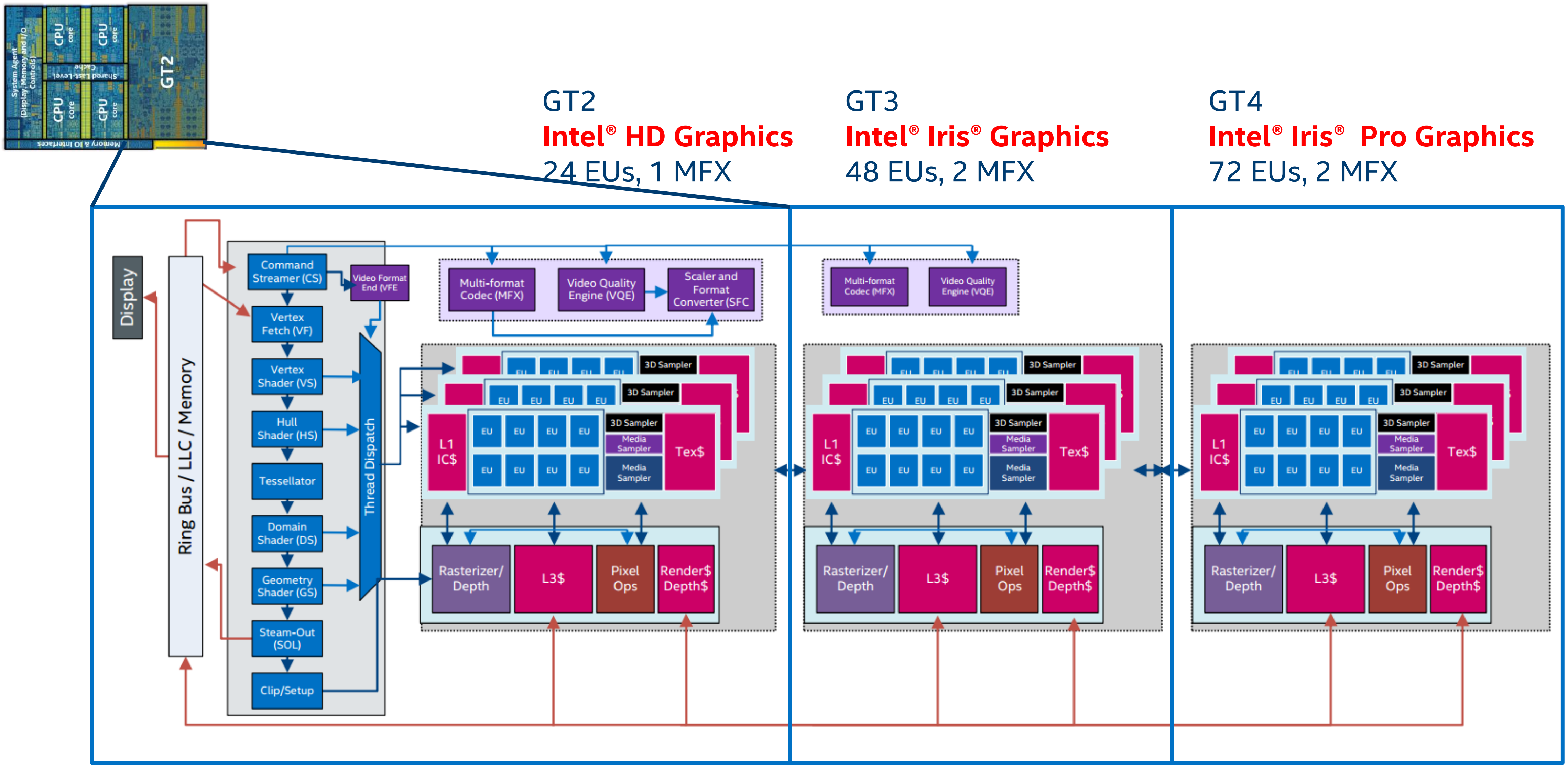
Gen is the internal name for Intel's on-die GPU solution. It's a hardware ingredient with various configurations.

- 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™.


System Agent
(Display, Memory, and IO Controls)
Memory & IO Interfaces
CPU core
Shared Last-Level Cache
CPU core
CPU core
CPU core
Gen 9 GPU

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

INTEL GPU CONFIGURATIONS



System Agent (Display, Memory and I/O Controls)
CPU core
CPU core
Shared Last-Level Cache
CPU core
CPU core
GT2
Memory & IO Interfaces
Display
Ring Bus / LLC / Memory
Command Streamer (CS)
Video Format End (VFE
Vertex Fetch (VF)
Vertex Shader (VS)
Hull Shader (HS)
Tessellator
Domain Shader (DS)
Geometry Shader (GS)
Steam-Out (SOL)
Clip/Setup
Thread Dispatch
Multi-format Codec (MFX)
Video Quality Engine (VQE)
Scaler and Format Converter (SFC
EU
3D Sampler
Media Sampler
L1 IC$
Tex$
Rasterizer/ Depth
L3$
Pixel Ops
Render$ Depth$

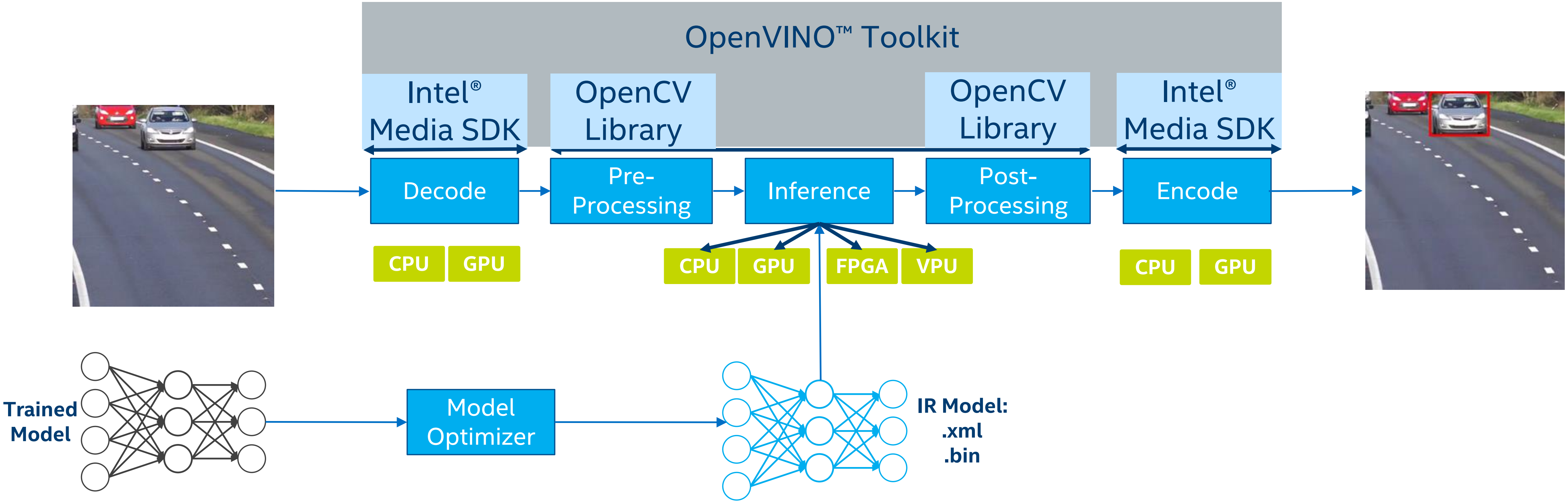
GT2
Intel® HD Graphics
24 EUs, 1 MFX

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.



OpenVINO™ Toolkit

Intel® Media SDK

OpenCV Library

OpenCV Library

Intel® Media SDK

Decode

Pre-Processing

Inference

Post-Processing

Encode

CPU GPU

CPU GPU FPGA VPU

CPU GPU

Trained Model

Model Optimizer

IR Model:
.xml
.bin

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

Optimization Notice

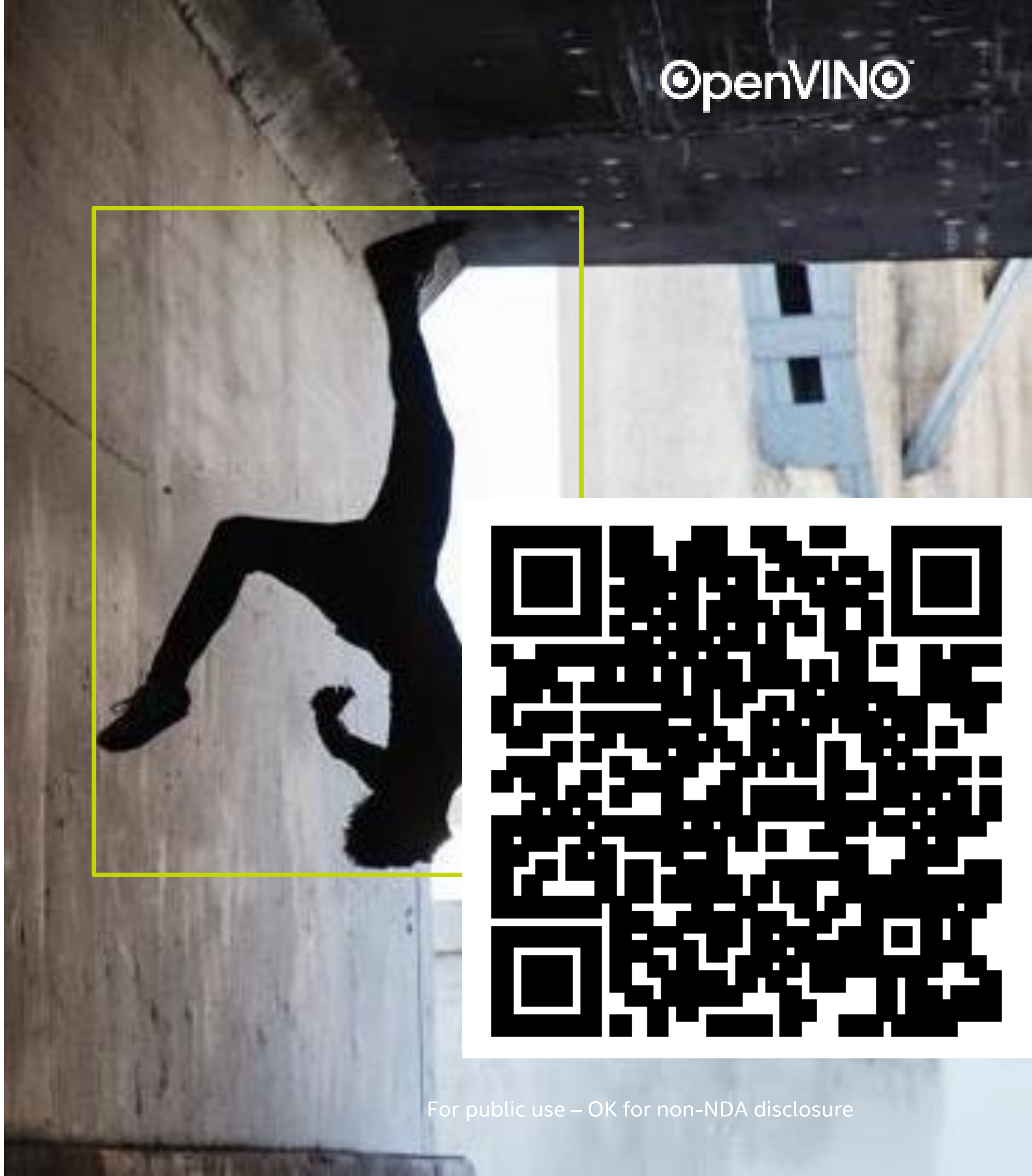
JUMPSTART DEEP LEARNING TODAY!

Download Free ▶

[Intel® Distribution of OpenVINO™ toolkit]

Also available from

[Docker] | [YUM] | [APT] | [NEW] [Anaconda Cloud]



For public use – OK for non-NDA disclosure