

ACCELERATE DEEP LEARNING INFERENCE USING INTEL TECHNOLOGIES

INTRODUCTION: SMART VIDEO

August 2019

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

Hardware Acceleration on laptop and devcloud

Optimization

Application

Custom layers

2. Basic End-to-End Object Detection Example

3./4./5. Hardware Acceleration with CPU, Integrated GPU, Intel® Movidius™ NCS, FPGA

6. Optimization Tools and Techniques

7. Advanced Video Analytics

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 of 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.

Copyright © 2018 Intel Corporation. All rights reserved.



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.





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

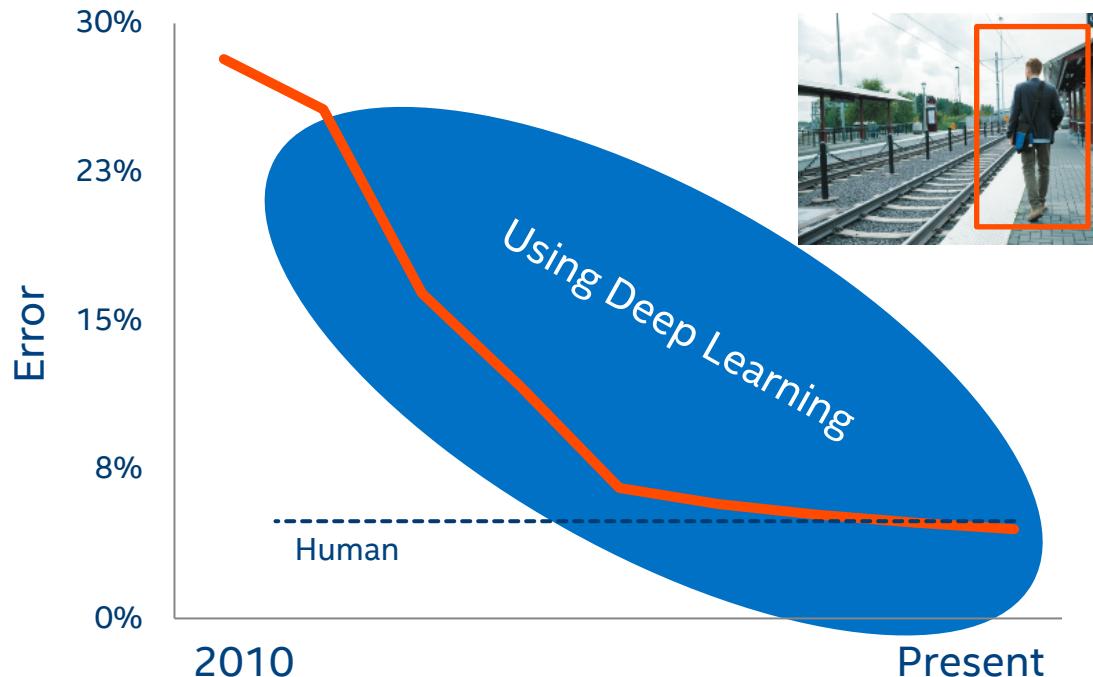


PUBLIC SECTOR

Deep Learning Usage Is Increasing

Deep learning revenue is estimated to grow from \$655M in 2016 to **\$35B** by 2025¹.

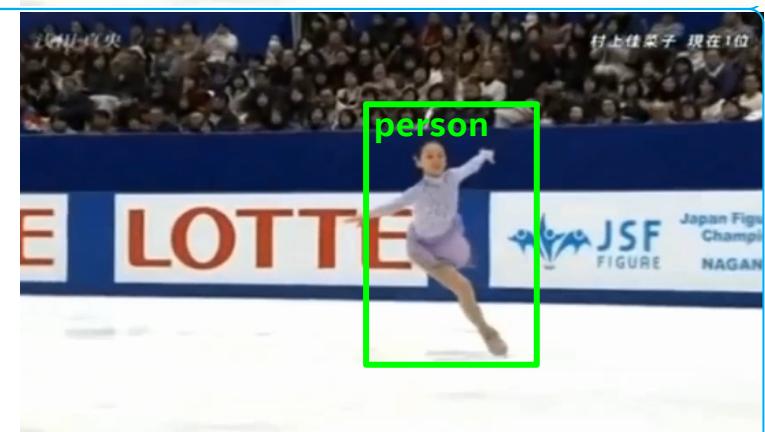
Image Recognition



Traditional Computer Vision Object Detection



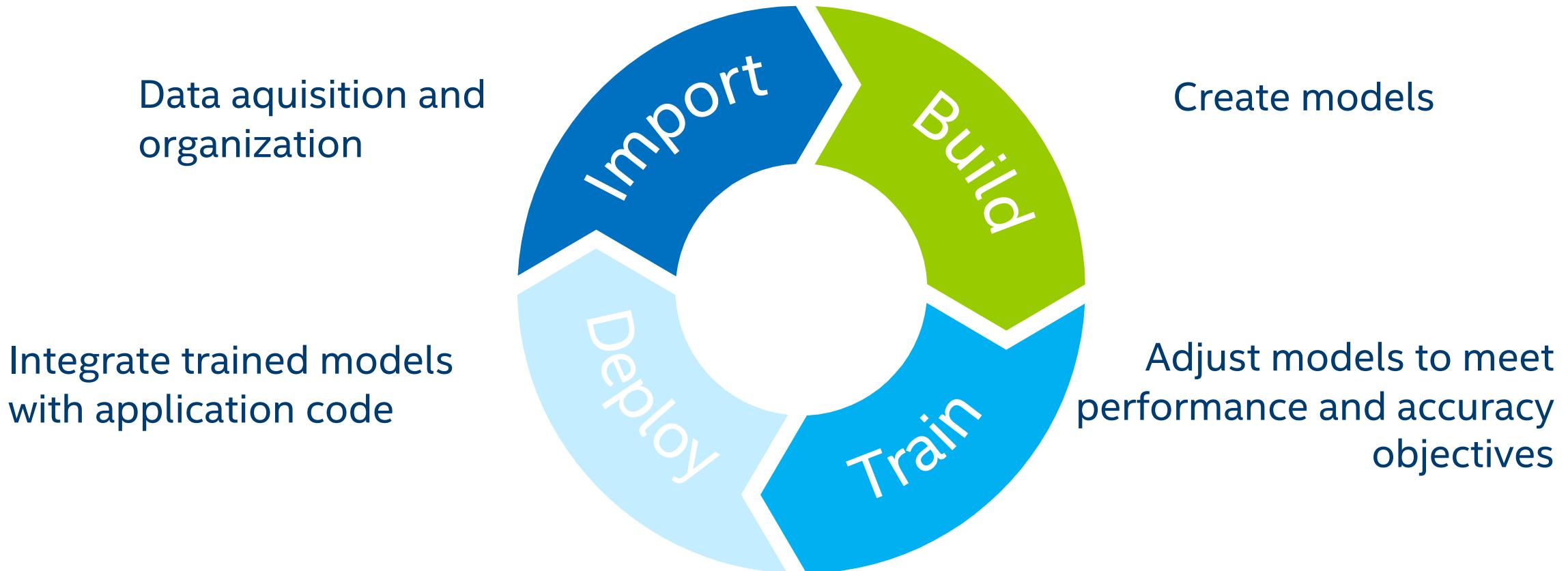
Deep Learning Computer Vision Person Recognition



Market Opportunities + Advanced Technologies Have Accelerated Deep Learning Adoption

¹Tractica* 2Q 2017

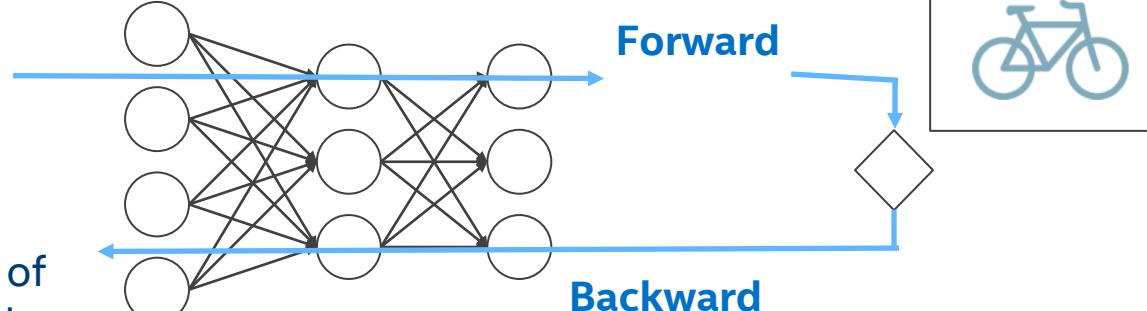
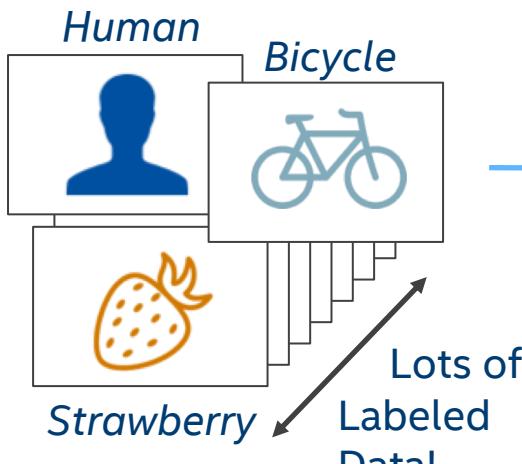
Deep Learning Development Cycle



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

Deep Learning: Training vs. Inference

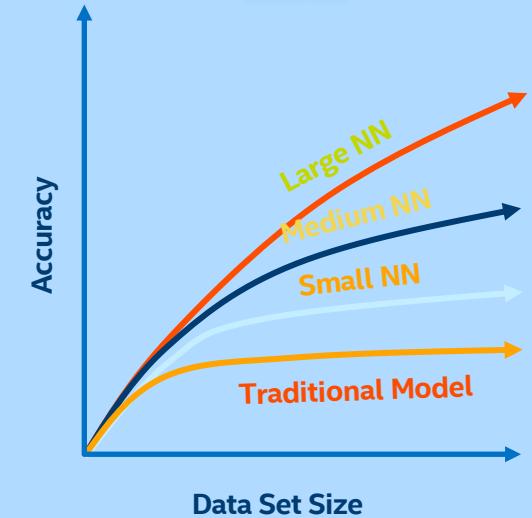
Training



Inference



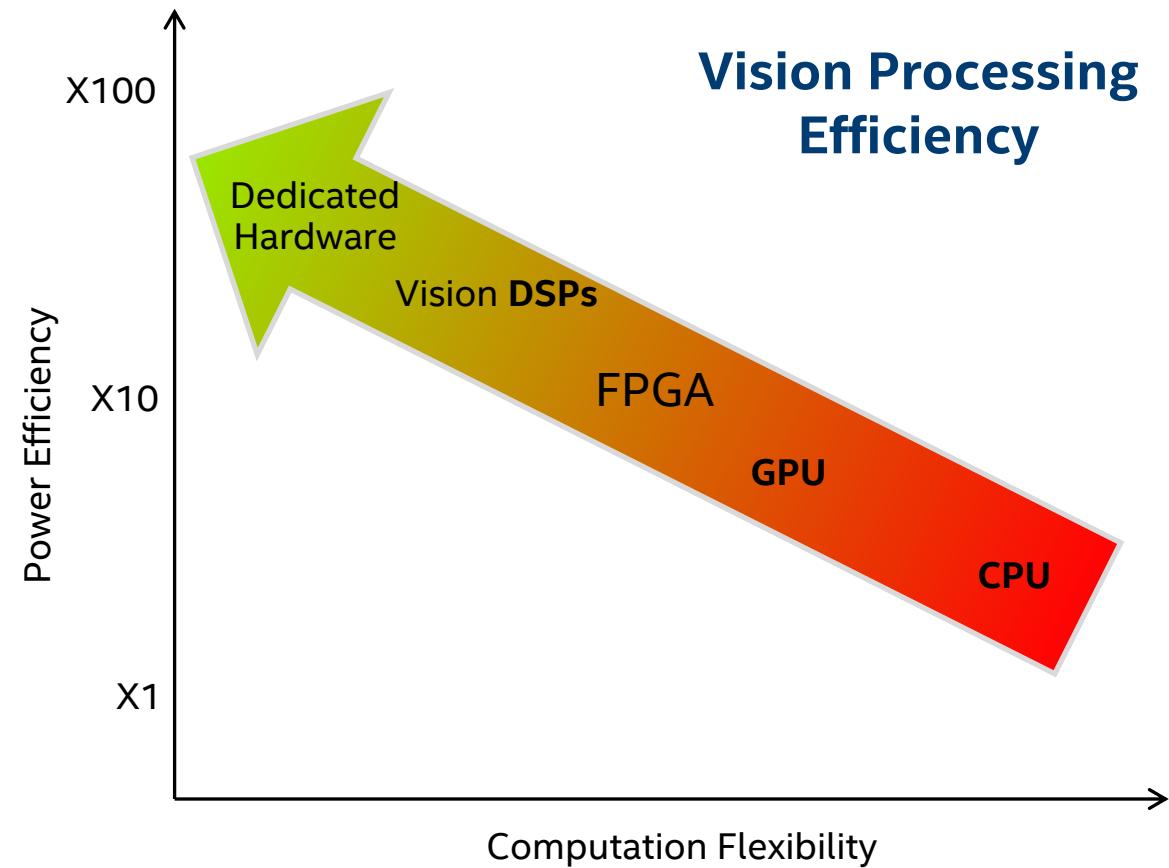
Training requires a very large data set and deep neural network (many layers) to achieve the highest accuracy in most cases



Choosing the “Right” Hardware

Power/Performance Efficiency Varies

- Running the right workload on the right piece of hardware → higher efficiency
- Hardware acceleration is a must
- Heterogeneous computing?



Tradeoffs

- Power/performance
- Price
- Software flexibility, portability

Intel Computer Vision Portfolio

EXPERIENCES



TOOLS

Intel® Parallel Studio XE
Intel® System Studio
Intel® SDK for OpenCL™ Applications

Intel® Media SDK / Media Server Studio
Intel® Distribution of OpenVINO™ toolkit

FRAMEWORKS



theano



Caffe



ONNX

LIBRARIES

Intel® Data
Analytics
Acceleration
Library

Intel®
Distribution for python

Intel® Math Kernel Library

Intel® Nervana™ Graph

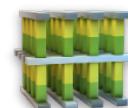


Movidius Stack

HARDWARE



Compute



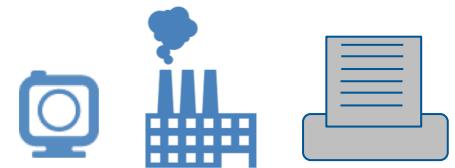
Memory & Storage



Networking



Visual Intelligence



UNLEASH
FULL
POTENTIAL

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

Take your computer vision solutions to a new level
with deep learning inference intelligence.

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.

Who needs this product?

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

Usages

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



HIGH PERFORMANCE, PERFORM AI AT THE EDGE



STREAMLINED & OPTIMIZED DEEP LEARNING INFERENCE

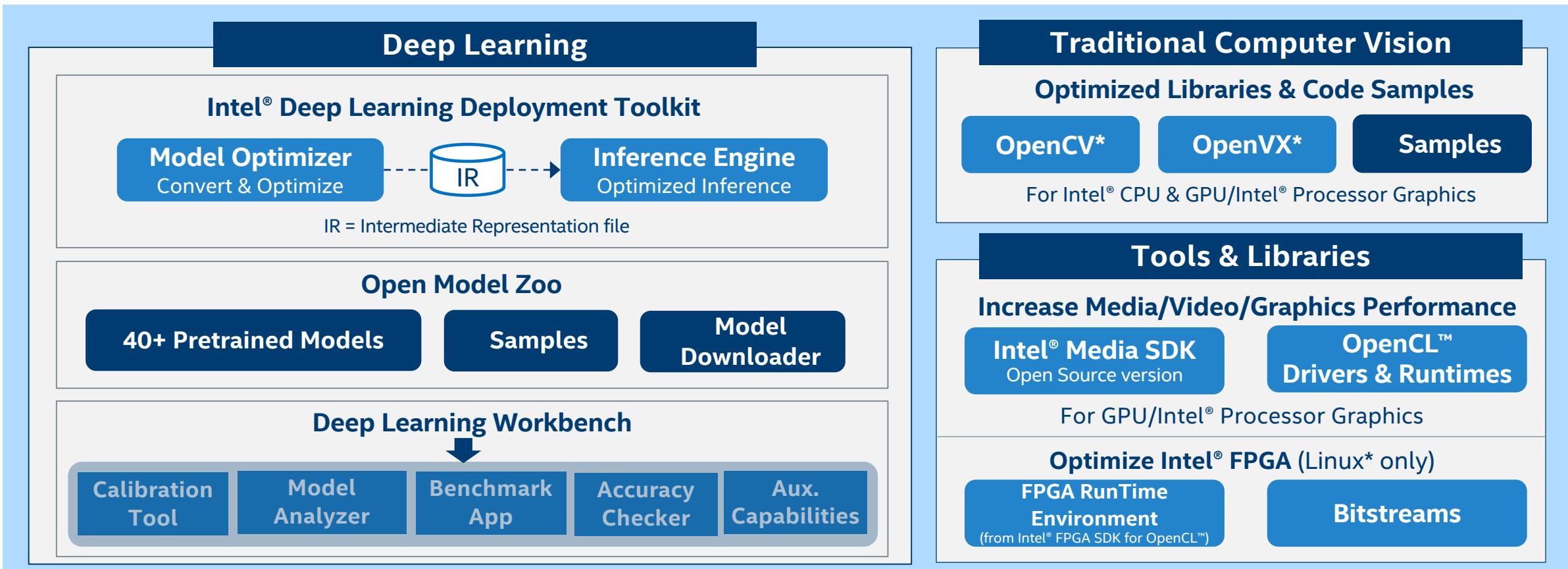


HETEROGENEOUS, CROSS-PLATFORM FLEXIBILITY

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

Open Source version ▶ 01.org/openvinotoolkit

What's Inside Intel® Distribution of OpenVINO™ toolkit



OS Support: CentOS* 7.4 (64 bit), Ubuntu* 16.04.3 LTS (64 bit), Microsoft Windows* 10 (64 bit), Yocto Project* version Poky Jethro v2.0.3 (64 bit), macOS* 10.13 & 10.14 (64 bit)

Intel® Architecture-Based Platforms Support



Intel® Vision Accelerator Design Products & AI in Production/Developer Kits

An open source version is available at 01.org/openvino/toolkit (some deep learning functions support Intel CPU/GPU only).

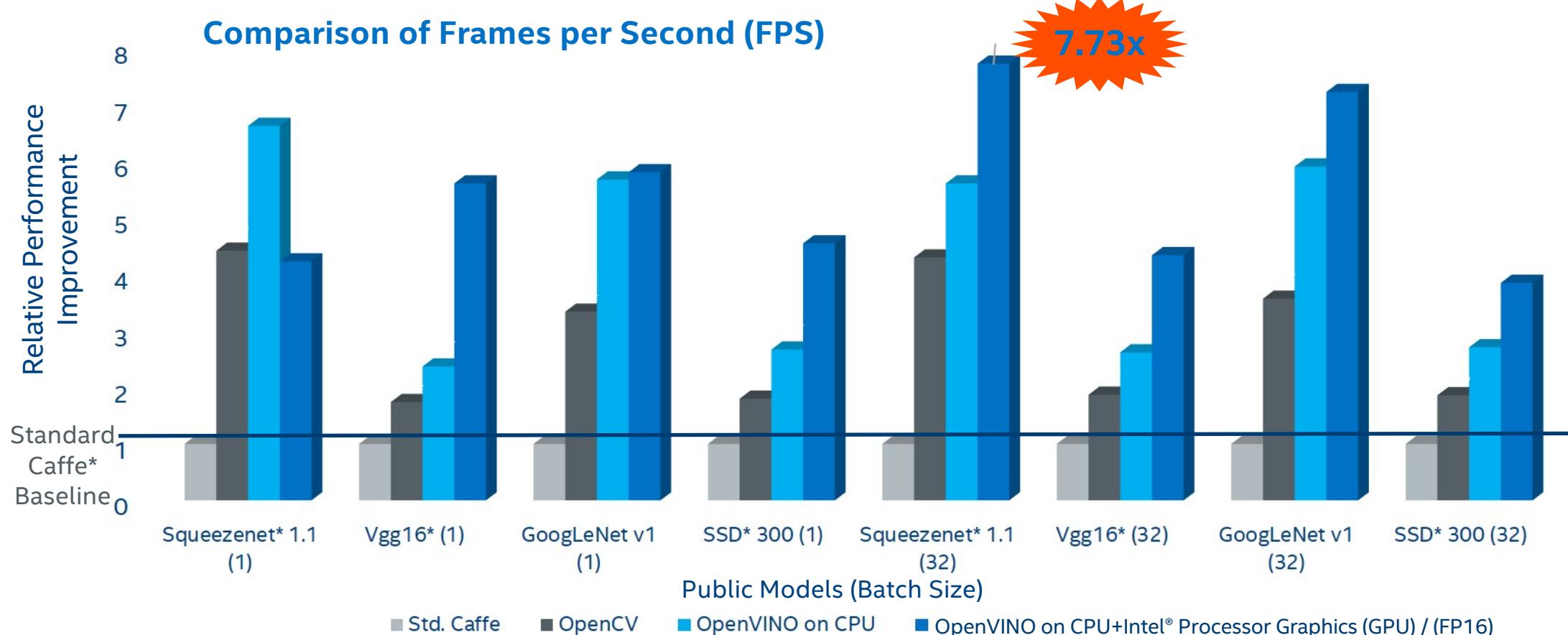
Quick Guide: What's Inside the Intel Distribution vs Open Source version

Tool/Component	Intel® Distribution of OpenVINO™ toolkit	OpenVINO™ toolkit (open source)	Open Source Directory
Installer (including necessary drivers)	✓		
Intel® Deep Learning Deployment toolkit	✓	✓	
Model Optimizer	✓	✓	/opencv/dldt/tree/2018/model-optimizer
Inference Engine	✓	✓	/opencv/dldt/tree/2018/inference-engine
Intel CPU plug-in	✓ Intel® Math Kernel Library (Intel® MKL) only ¹	✓ BLAS, Intel® MKL ¹ , jit (Intel MKL)	/opencv/dldt/tree/2019/inference-engine
Intel GPU (Intel® Processor Graphics) plug-in	✓	✓	/opencv/dldt/tree/2019/inference-engine
Heterogeneous plug-in	✓	✓	/opencv/dldt/tree/2019/inference-engine
Intel GNA plug-in	✓	✓	/opencv/dldt/tree/2019/inference-engine
Intel® FPGA plug-in	✓		
Intel® Neural Compute Stick (1 & 2) VPU plug-in	✓	✓	/opencv/dldt/tree/2019/inference-engine
Intel® Vision Accelerator based on Movidius plug-in	✓		
40+ Pretrained Models - incl. Model Zoo (IR models that run in IE + open sources models)	✓	✓	https://github.com/opencv/open_model_zoo
Samples (APIs)	✓	✓	/opencv/dldt/tree/2018/inference-engine
Demos	✓	✓	https://github.com/opencv/open_model_zoo
Traditional Computer Vision			
OpenCV*	✓	✓	https://github.com/opencv/opencv
OpenVX (with samples)	✓		
Intel® Media SDK	✓	✓ ²	https://github.com/Intel-Media-SDK/MediaSDK
OpenCL™ Drivers & Runtimes	✓	✓ ²	https://github.com/intel/compute-runtime
FPGA RunTime Environment, Deep Learning Acceleration & Bitstreams (Linux* only)	✓		

¹Intel MKL is not open source but does provide the best performance

²Refer to readme file for validated versions

Increase Deep Learning Workload Performance on Public Models using Intel® Distribution of OpenVINO™ toolkit & Intel® Architecture



Fast Results on Intel Hardware, even before using Accelerators

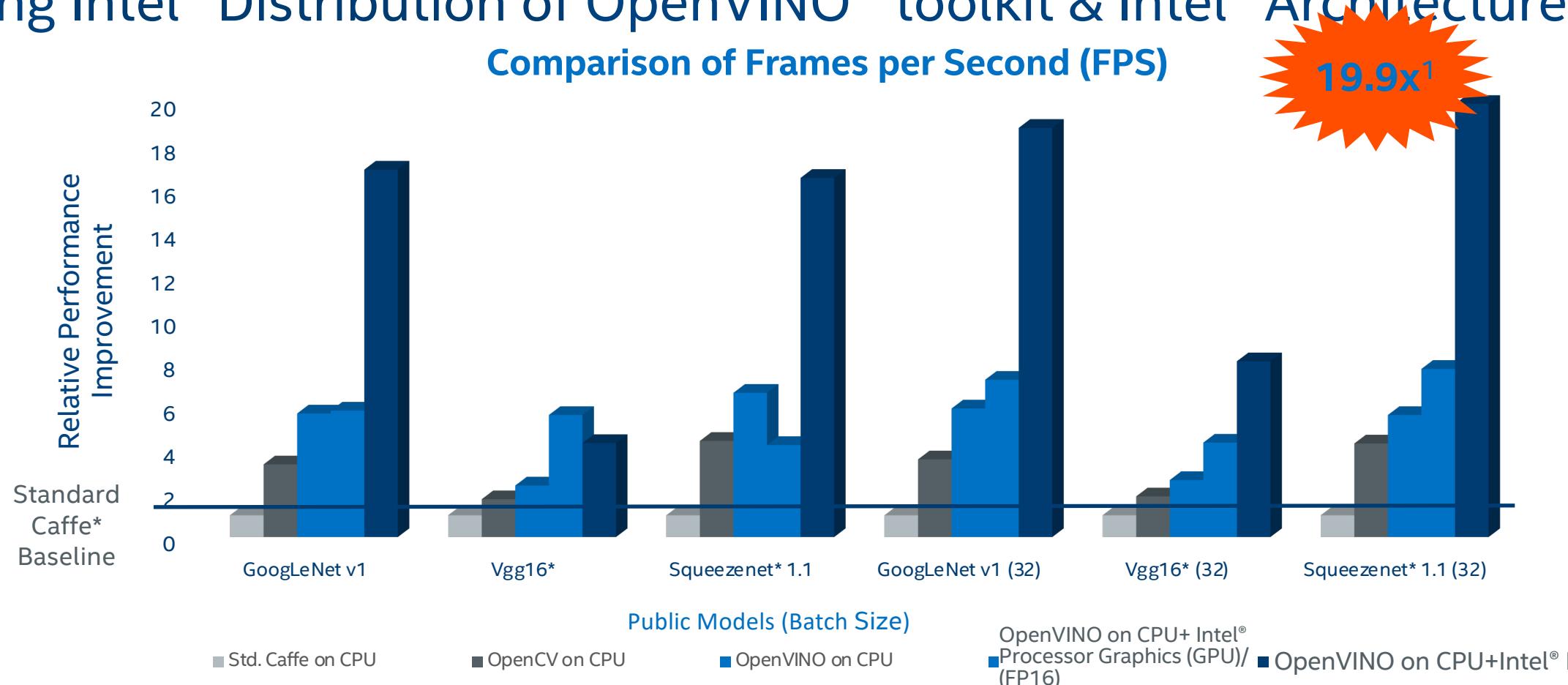
1Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. The benchmark results reported in this deck may need to be revised as additional testing is conducted. Performance results are based on testing as of April 10, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Configuration: Testing by Intel as of April 10, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v312.30 – Ubuntu* 16.04, OpenVINO™ 2018 RC4. Tests were based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

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 of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this fixed product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804



Increase Deep Learning Workload Performance on Public Models using Intel® Distribution of OpenVINO™ toolkit & Intel® Architecture



Get an even Bigger Performance Boost with Intel® FPGA

¹Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. Performance results are based on testing as of June 13, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Configuration: Testing by Intel as of June 13, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v3.15.21 – Ubuntu* 16.04, OpenVINO 2018 RC4, Intel® Arria® 10 FPGA 1150GX. Tests were based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

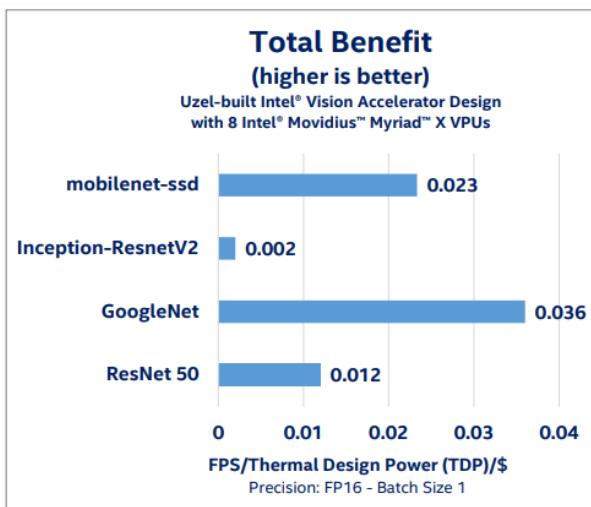
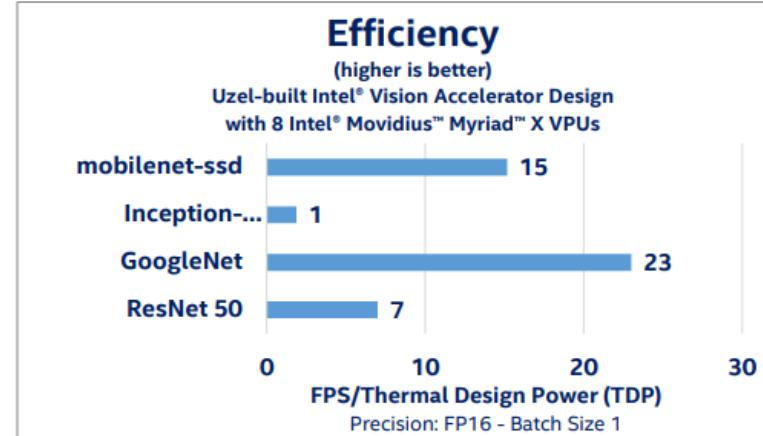
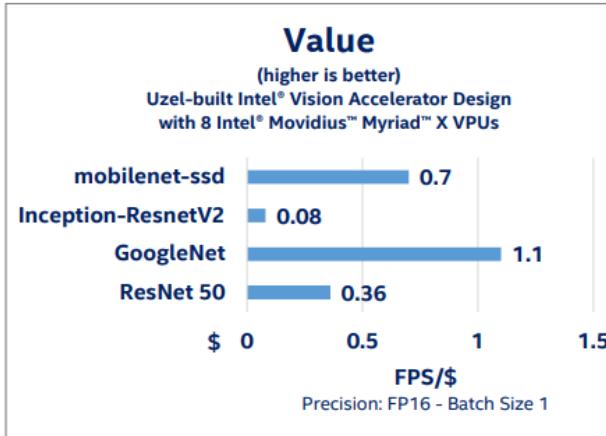
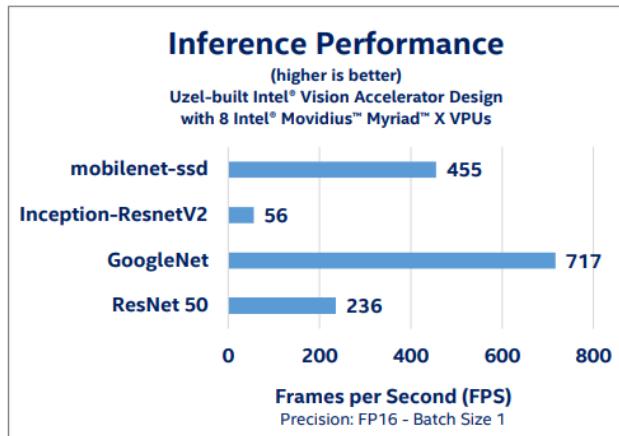
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of 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. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804



New Benchmarks for HDDL Accelerators – June 2019

Increase Deep Learning Model Performance with Intel® Movidius™ VPU & Intel® Distribution of OpenVINO™ toolkit

Location: [Intel® Distribution of OpenVINO™ toolkit product site](#) > Hardware > [Performance Benchmarks](#)



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. Performance results are based on testing as of March 29, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information, see [Performance Benchmark Test Disclosure](#).

CONFIGURATIONS: Testing by Intel as of March 29, 2019 [See slide 59 for configuration details](#).

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 of 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. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Speed Deployment with Pretrained Models & Samples

Expedite development, accelerate deep learning inference performance, speed production deployment

Pretrained Models in Intel® Distribution of OpenVINO™ toolkit		
<ul style="list-style-type: none">▪ Age & Gender▪ Face Detection—standard & enhanced▪ Head Position▪ Human Detection—eye-level & high-angle detection▪ Detect People, Vehicles & Bikes▪ License Plate Detection: small & front facing▪ Vehicle Metadata▪ Human Pose Estimation▪ Action recognition – encoder & decoder	<ul style="list-style-type: none">▪ Text Detection & Recognition▪ Vehicle Detection▪ Retail Environment▪ Pedestrian Detection▪ Pedestrian & Vehicle Detection▪ Person Attributes Recognition Crossroad▪ Emotion Recognition▪ Identify Someone from Different Videos—standard & enhanced▪ Facial Landmarks▪ Gaze estimation	<ul style="list-style-type: none">▪ Identify Roadside objects▪ Advanced Roadside Identification▪ Person Detection & Action Recognition▪ Person Re-identification—ultra small/ultra fast▪ Face Re-identification▪ Landmarks Regression▪ Smart Classroom Use Cases▪ Single image Super Resolution (3 models)▪ Instance segmentation▪ and more...
Binary Models		
<ul style="list-style-type: none">▪ Face Detection Binary▪ Pedestrian Detection Binary	<ul style="list-style-type: none">▪ Vehicle Detection Binary	<ul style="list-style-type: none">▪ ResNet50 Binary



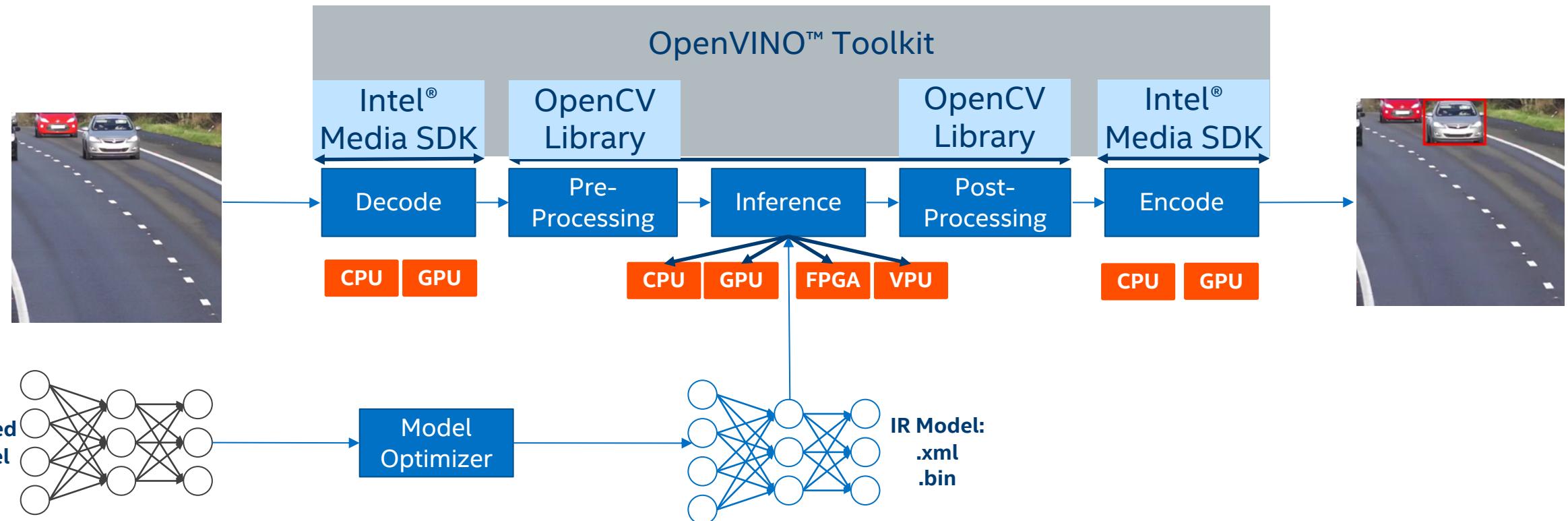
Save Time with Deep Learning Samples

Use Model Optimizer & Inference Engine for public models & Intel pretrained models

- Object Detection
- Standard & Pipelined Image Classification
- Security Barrier
- Object Detection SSD
- Neural Style Transfer
- Object Detection for Single Shot Multibox Detector using Asynch API+
- Hello Infer Classification
- Interactive Face Detection
- Image Segmentation
- Validation Application
- Multi-channel Face Detection

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.



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

Embedded Linux*



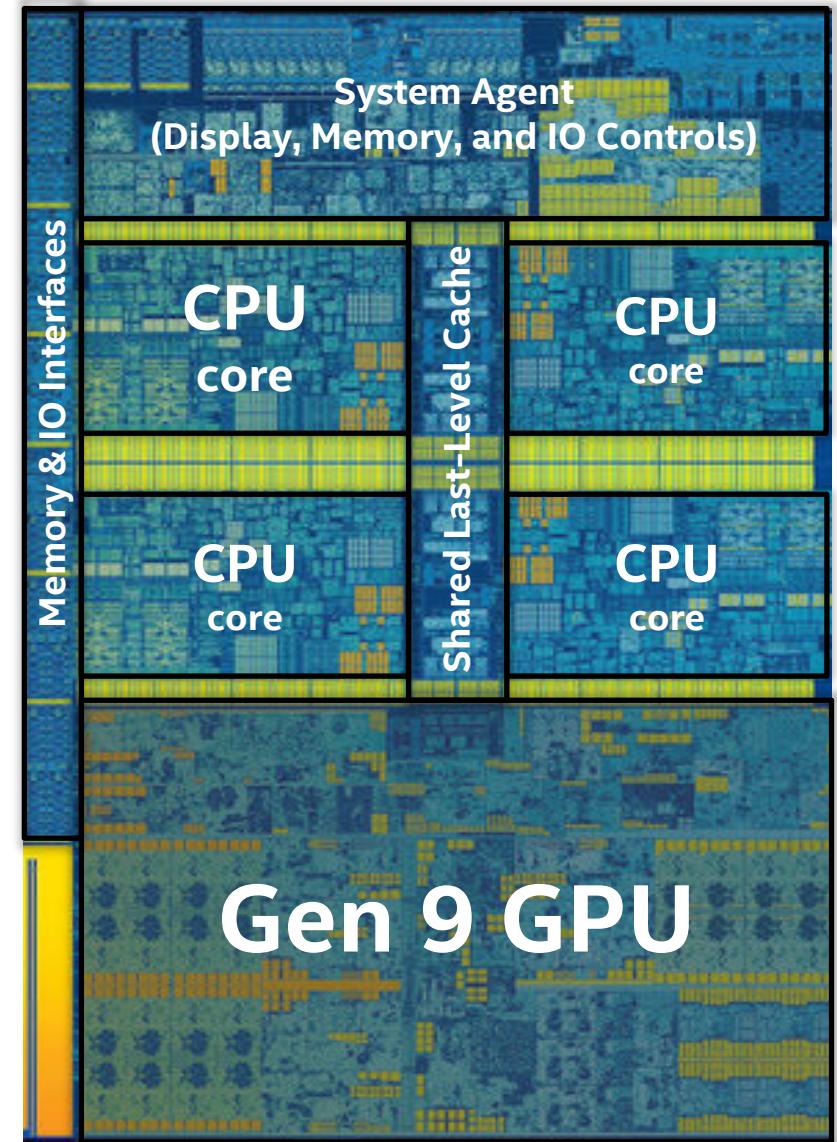
¹ Intel® Celeron® Processor N3350, Intel® Pentium® Processor N4200, Intel Atom® E3930, E3940, E3950 processors

Intel Integrated Graphics

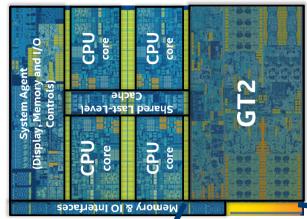
Gen is the internal name for Intel's on-die GPU solution. It's a hardware ingredient with various configurations.

- 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™.

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



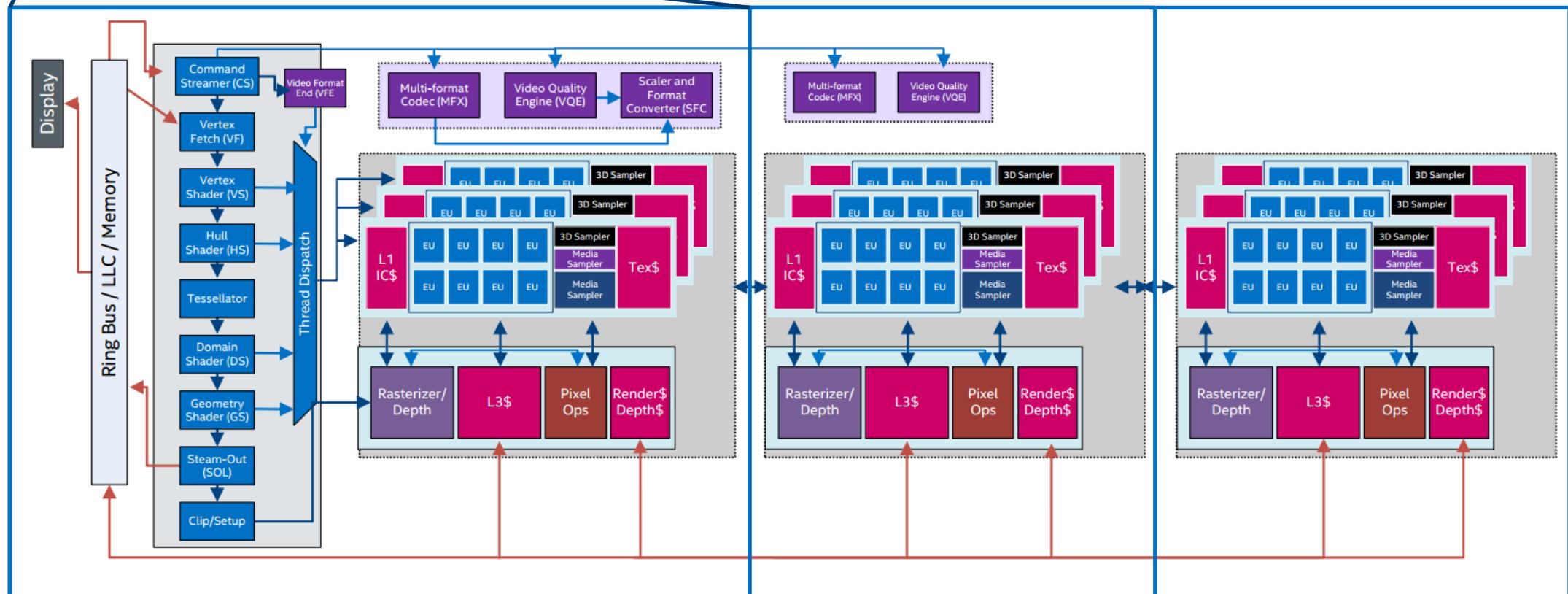
Intel GPU Configurations



GT2
Intel® HD Graphics
24 EUs, 1 MFX

GT3
Intel® Iris® Graphics
48 EUs, 2 MFX

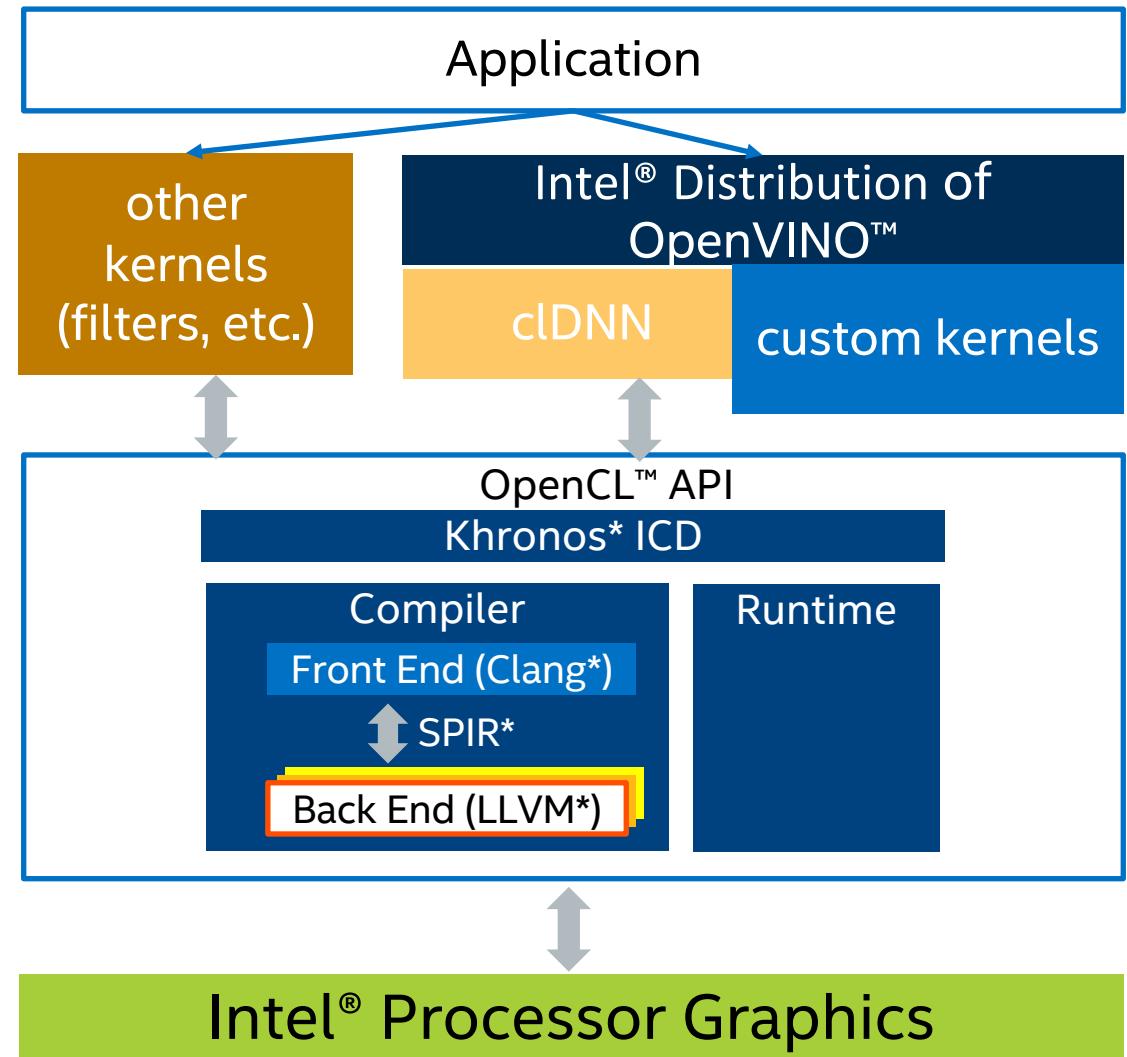
GT4
Intel® Iris® Pro Graphics
72 EUs, 2 MFX



OpenCL™

OpenCL™:

- Required to run with a GPU target (clDNN) using Intel® Processor Graphics
- Custom kernels
- Other kernels can be used for other non-inference pipeline stages, such as color conversions



Putting It All Together

- A major challenge is to get all these tool and libraries to work together in the best possible way to minimize development time and optimize system power/performance.
- A good way to abstract that workload is using an end-to-end pipeline

Computer Vision



Deep Learning



Media



SDKs



Optimized CV
Capabilities

Intel® Distribution of OpenVINO™ Toolkit



Intel® Deep Learning
Deployment Toolkit



Intel® Media SDK

Tools

Compiler, Analyzers, Debuggers



OpenCL™ SDK

Libraries

IPP



TBB



OpenCL™

Intel® MKL-DNN



Intel® MKL
DAAL



