Accelerate Deep Learning Inference Using Intel Technologies: Introduction to Smart Video

May 2018

Core and Visual Computing Group

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.

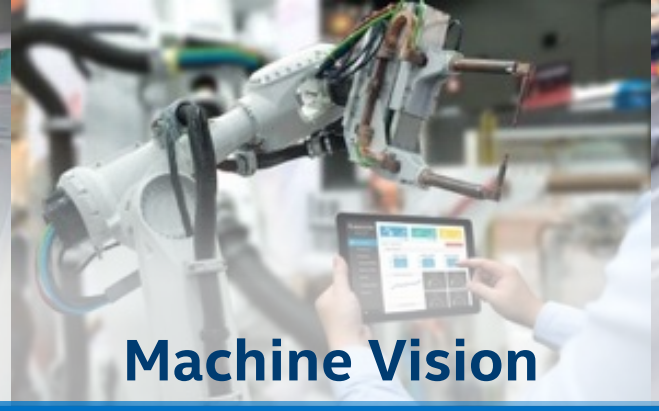
Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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.



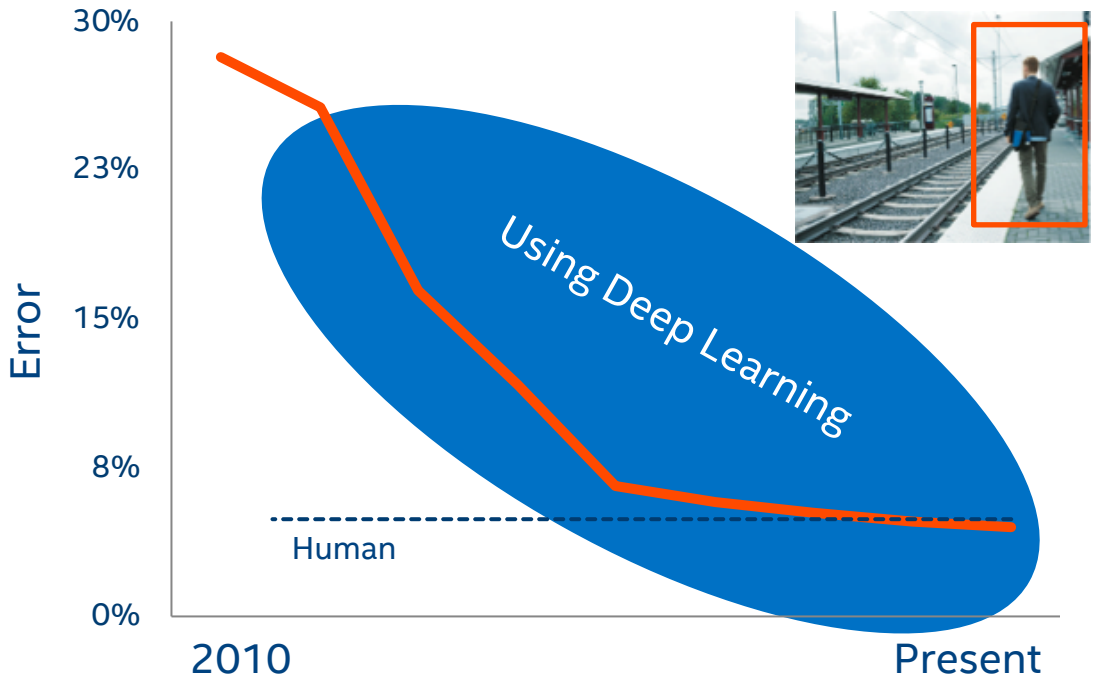
Video: The "Eye of IOT"

Use of video, computer vision, and deep learning is growing rapidly

Deep Learning Usage Is Increasing

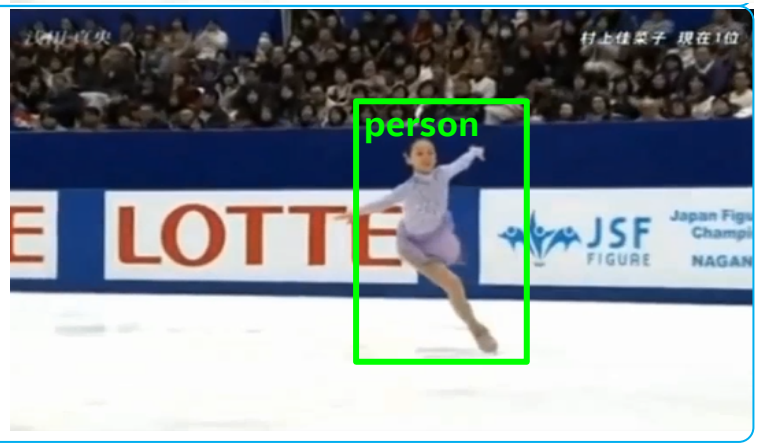
Deep learning revenue is estimated to grow from $655M in 2016 to $35B by 2025[1].

Image Recognition



Traditional Computer Vision
Object Detection

Deep Learning Computer Vision
Person Recognition

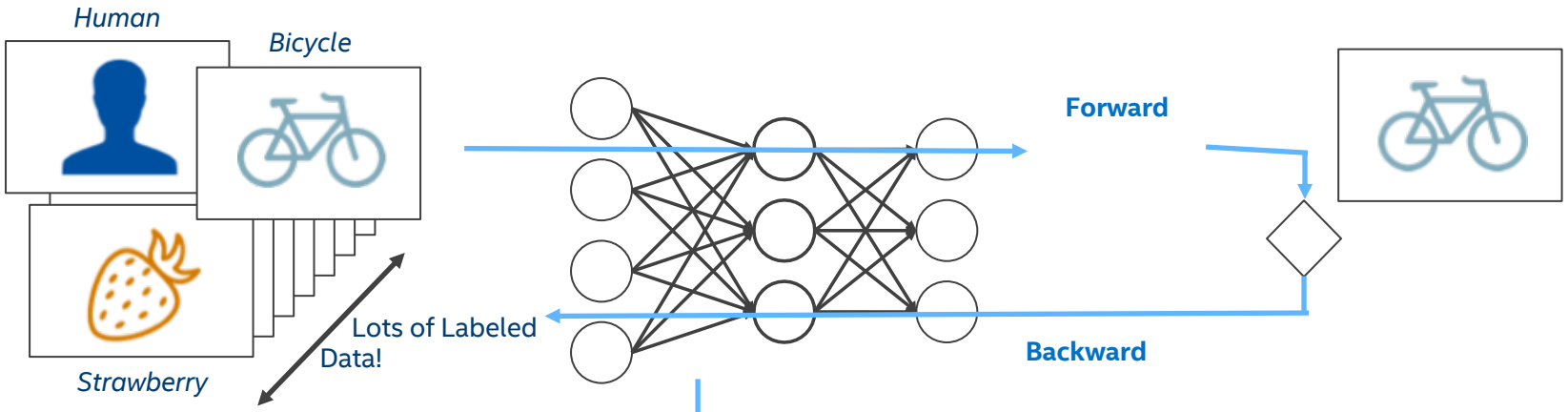


Market Opportunities + Advanced Technologies Have Accelerated Deep Learning Adoption

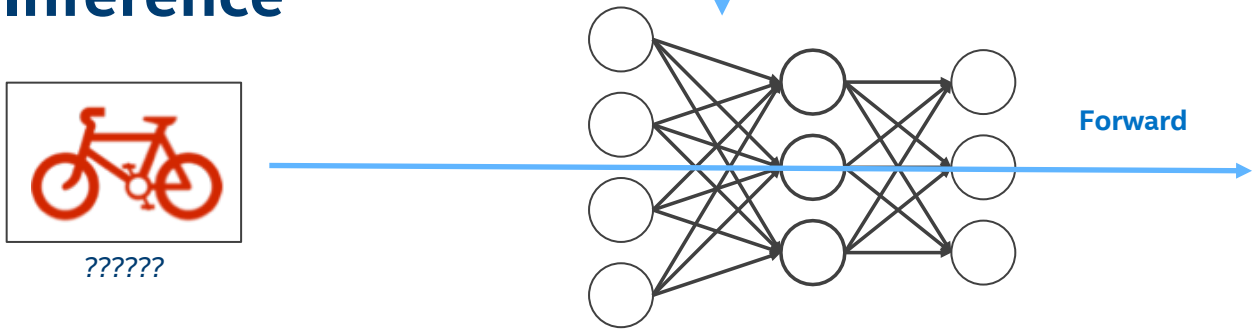
[1]Tractica* 2Q 2017

Deep Learning: Training vs. Inference

Training

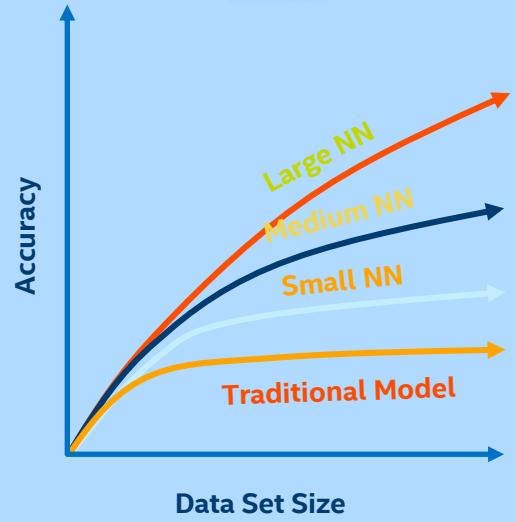


Inference



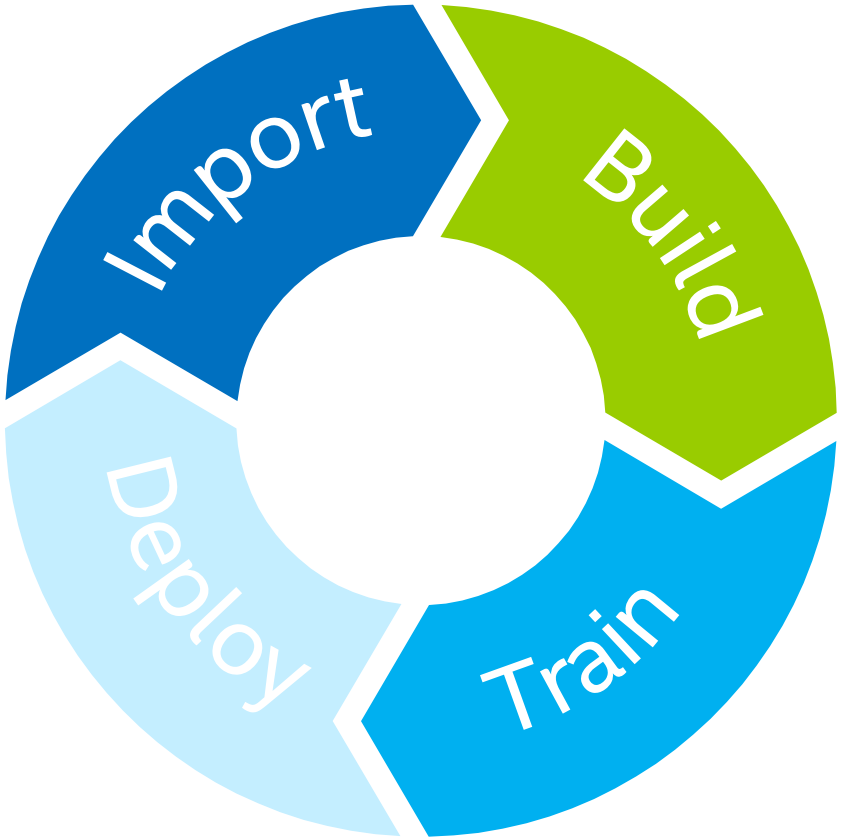
Did You Know?

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



Artificial Intelligence Development Cycle

Data aquisition and organization



Create models

Integrate trained models with application code

Adjust models to meet performance and accuracy objectives

Intel® Deep Learning Deployment Toolkit Provides Deployment from Intel® Edge to Cloud

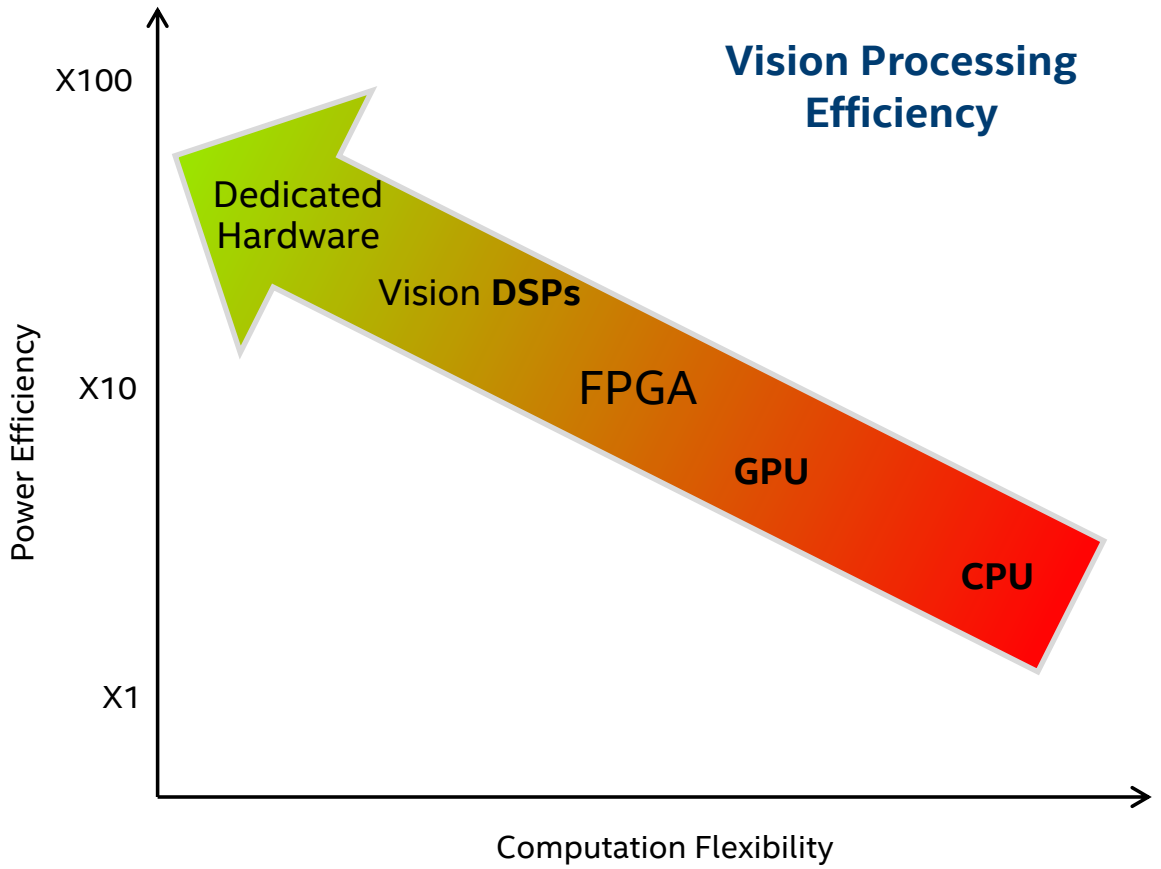
Choosing the “Right” Hardware

Power/Performance Efficiency Varies

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

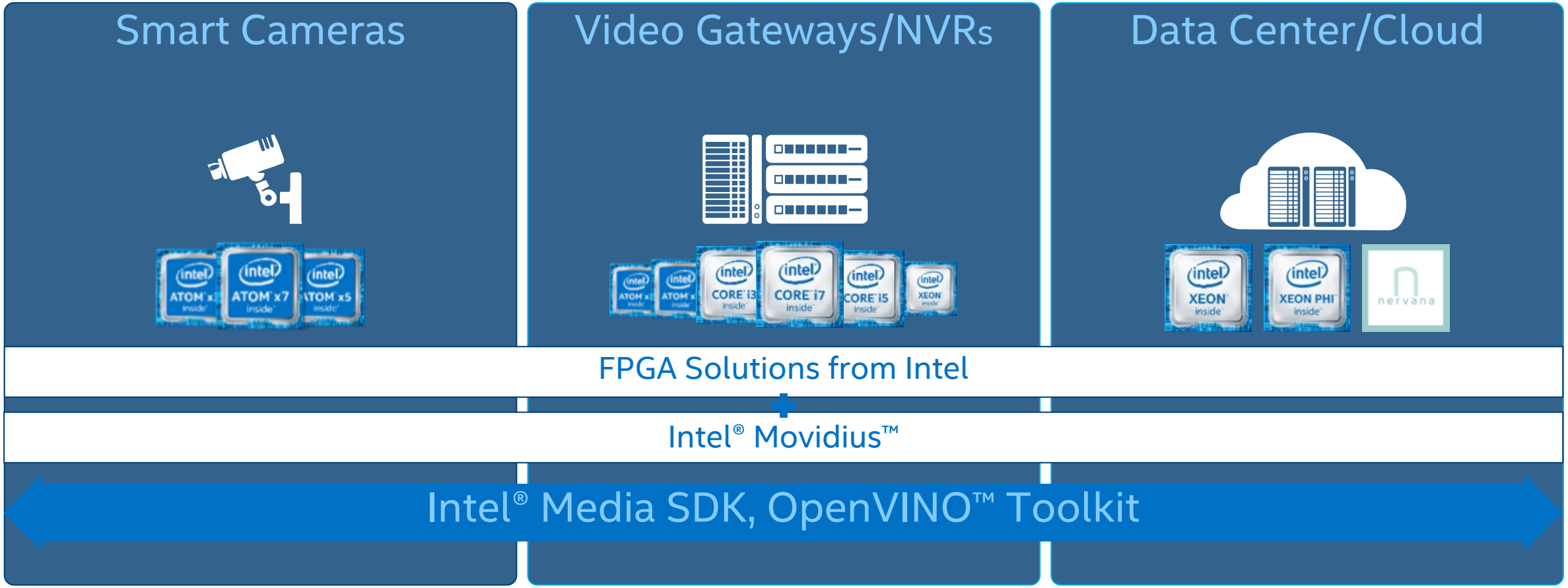
Tradeoffs

- Power/performance
- Price
- Software flexibility, portability



Intel Internet of Things (IoT) Video Portfolio

Intel Invests in AI, Computer Vision, and Deep Learning for IoT



Industry's Broadest Media and Computer Vision and Deep Learning Portfolio

Open Visual Inference and Neural Network Optimization (OpenVINO™) Toolkit and Components

Inside the OpenVINO™ Toolkit

TensorFlow*
Caffe*
mxnet*
Trained Models

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert and Optimize

IR

Inference Engine
Optimized Inference

CPU | GPU | FPGA | VPU

Component tools

Traditional Computer Vision—All SDK Versions
Optimized Computer Vision Libraries

OpenCV* **OpenVX***

Increase Processor Graphics Performance—Linux* Only

Intel® Media SDK
(Open Source Version)

OpenCL™
Intel Integrated Graphics Drivers and Runtimes

Linux for FPGA Only

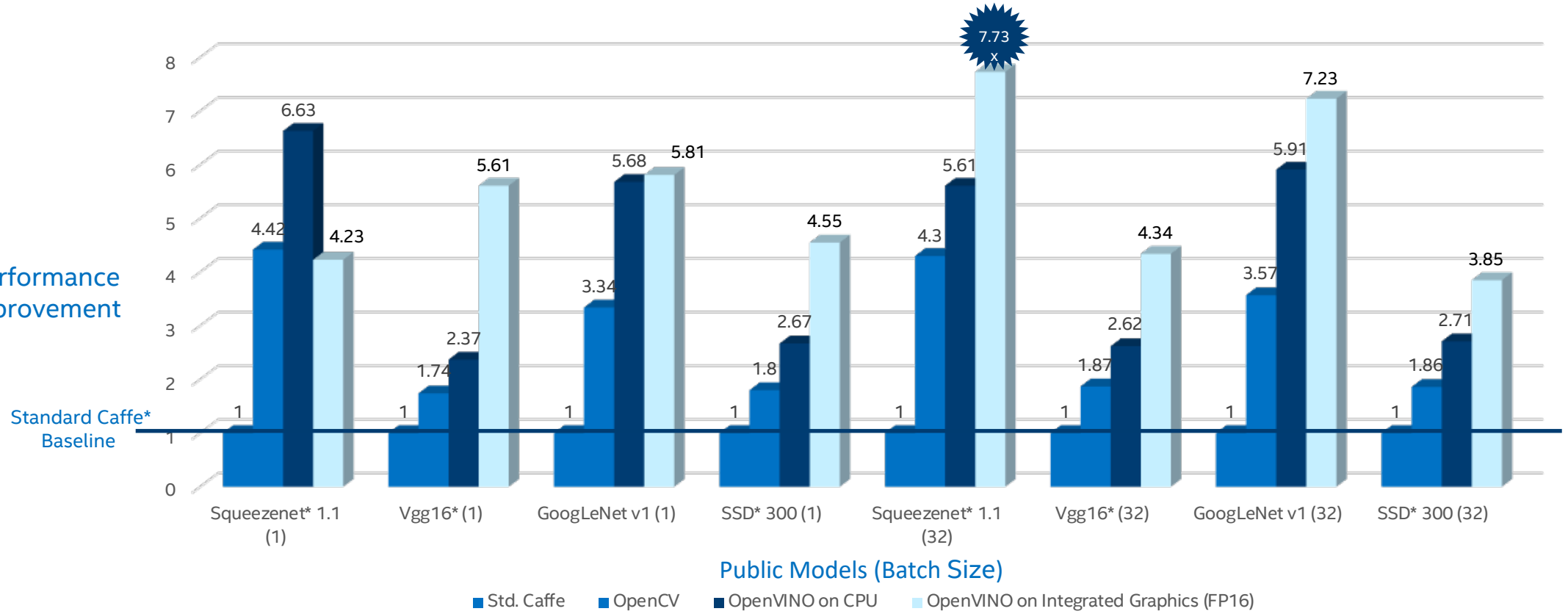
FPGA Runtime Environment (RTE)
(from Intel® FPGA SDK for OpenCL™)

Bit Streams

GPU = CPU with Intel Integrated Graphics Processing Unit
VPU = Intel® Movidius™ Vision Processing Unit

Performance Improvement Using the OpenVINO™ Toolkit

Comparison of Frames per Second (FPS)



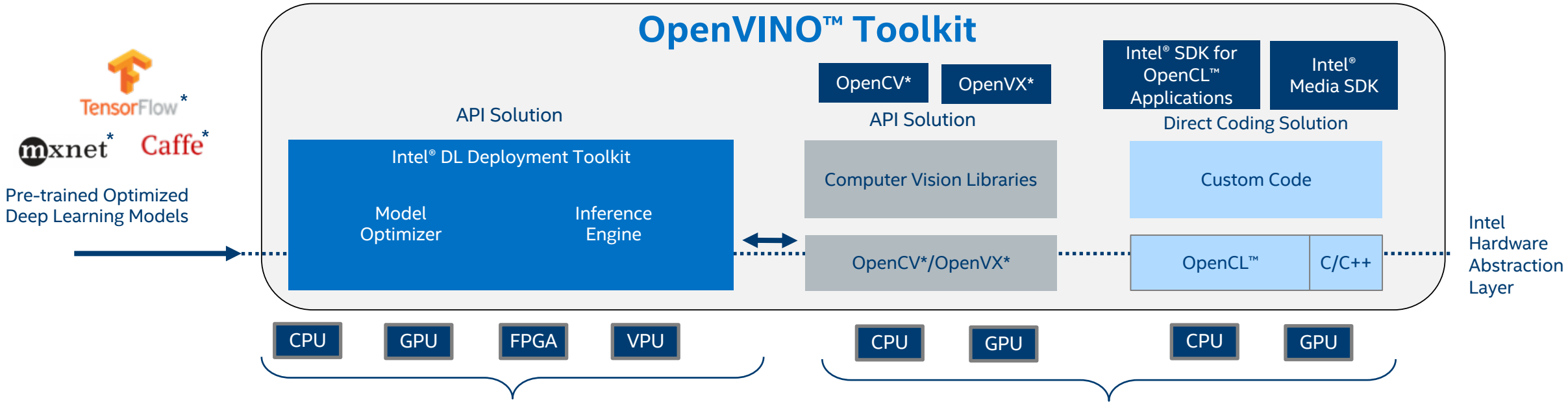
Faster Results on Intel Hardware

[1]Accuracy changes can occur w/FP16

The benchmark results reported in this deck may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. **Configuration:** Intel® Core™ i7-6700K CPU @ 2.90 GHz fixed, GPU GT2 @ 1.00 GHz fixed Internal ONLY testing, performed 4/10/2018 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. Benchmark Source: Intel Corporation.

Deep Learning vs. Traditional Computer Vision

OpenVINO™ Toolkit End-to-End Vision Pipeline



DEEP LEARNING Computer Vision

- Based on the application of a large number of filters to an image to extract features.
- Features in the object(s) are analyzed with the goal of associating each input image with an output node for each type of object.
- Values are assigned to output node representing the probability that the image is the object associated with the output node.

TRADITIONAL Computer Vision

- Based on selection and connections of computational filters to abstract key features and correlating them to an object.
- Works well with well-defined objects and controlled scene.
- Difficult to predict critical features in larger number of objects or varying scenes.

Computer Vision Application development with OpenVINO™ Toolkit

Train

Train a DL model.
Currently supports:

- Caffe*
- Mxnet*
- TensorFlow*

TensorFlow
Caffe
mxnet

Prepare Optimize

Model optimizer:

- Converting
- Optimizing
- Preparing to inference

(device agnostic, generic optimization)

.prototxt
.caffemodel

Run Model Optimizer

IR
.xml
.bin

Inference

Inference engine lightweight API to use in applications for inference.

User Application

Inference Engine

Optimize/ Heterogeneous

Inference engine supports multiple devices for heterogeneous flows.

(device-level optimization)

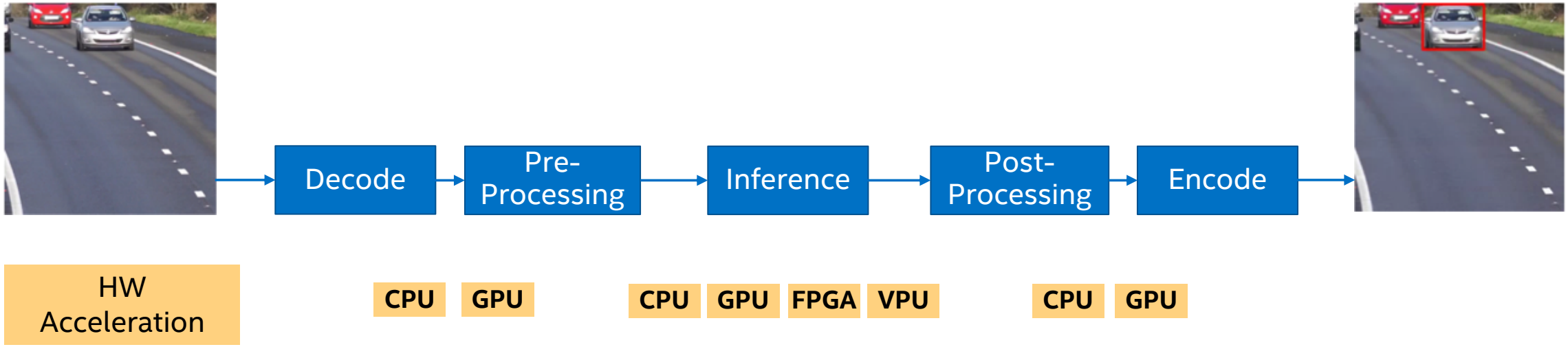
- MKL-DNN → CPU: Intel® Xeon®/Intel® Core™/Intel Atom®
- cl-DNN → GPU
- DLA → FPGA
- Intel® Movidius™ API → Movidius Myriad™ 2/X

Extend

Inference engine supports extensibility and allows custom kernels for various devices.

- CPU: Extensibility C++
- GPU: Extensibility OpenCL™
- FPGA: Extensibility OpenCL™/TBD
- Movidius Myriad™ 2/X: Extensibility TBD

Full Pipeline Optimization



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*

yocto PROJECT

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

Theory of Operation: Intel® Media SDK/Intel® Media Server Studio

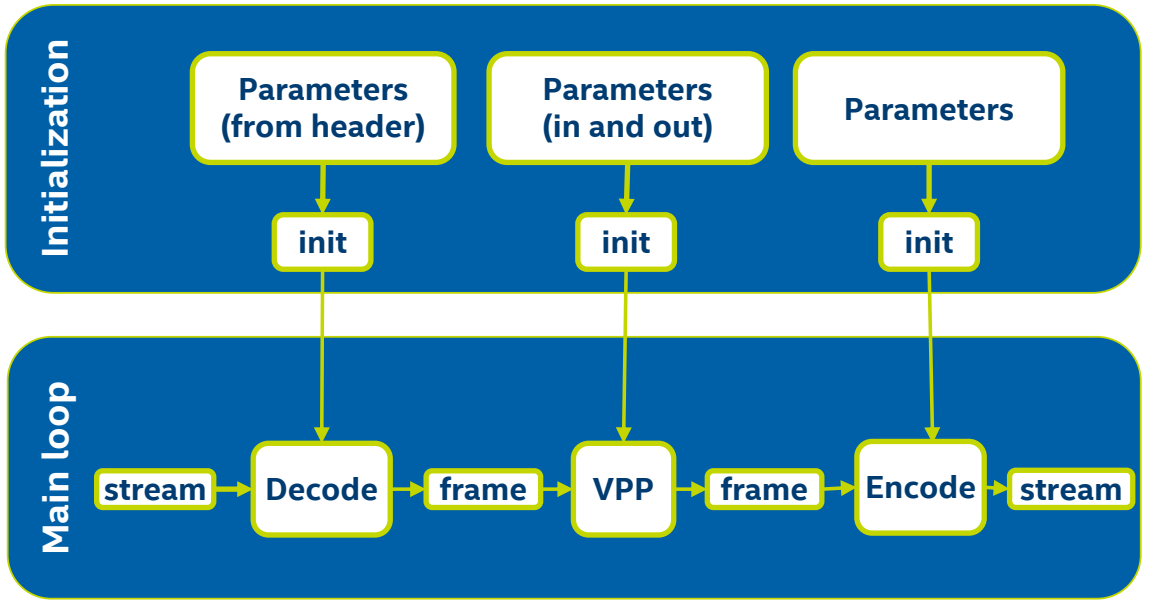
Media accelerator framework

- Codec based
- High level/parameter interface
- Three operations

Good option for:

- Accelerated video encode, decode
- (and short list of frame processing)

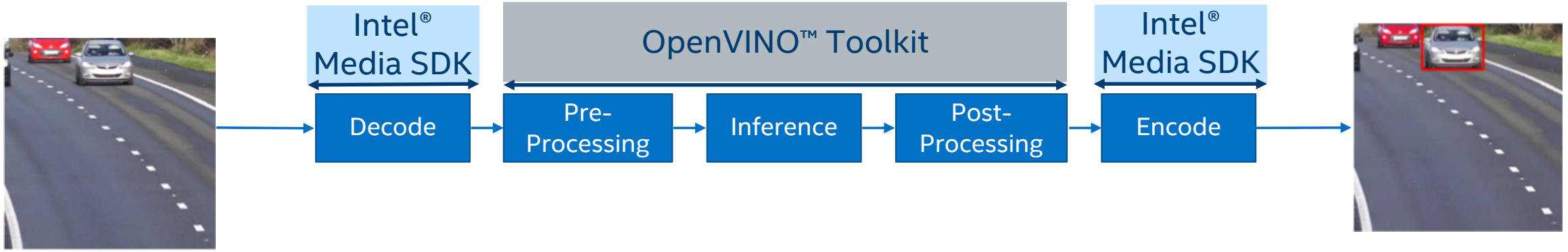
Out of scope: audio, containers, networking…



- Intel® Media Server Studio
- Intel® Media SDK
- Intel® Media Codec Samples

Accelerate Streaming Performance, Integrate Video Analytics Computer Vision Needs Intel® Media SDK

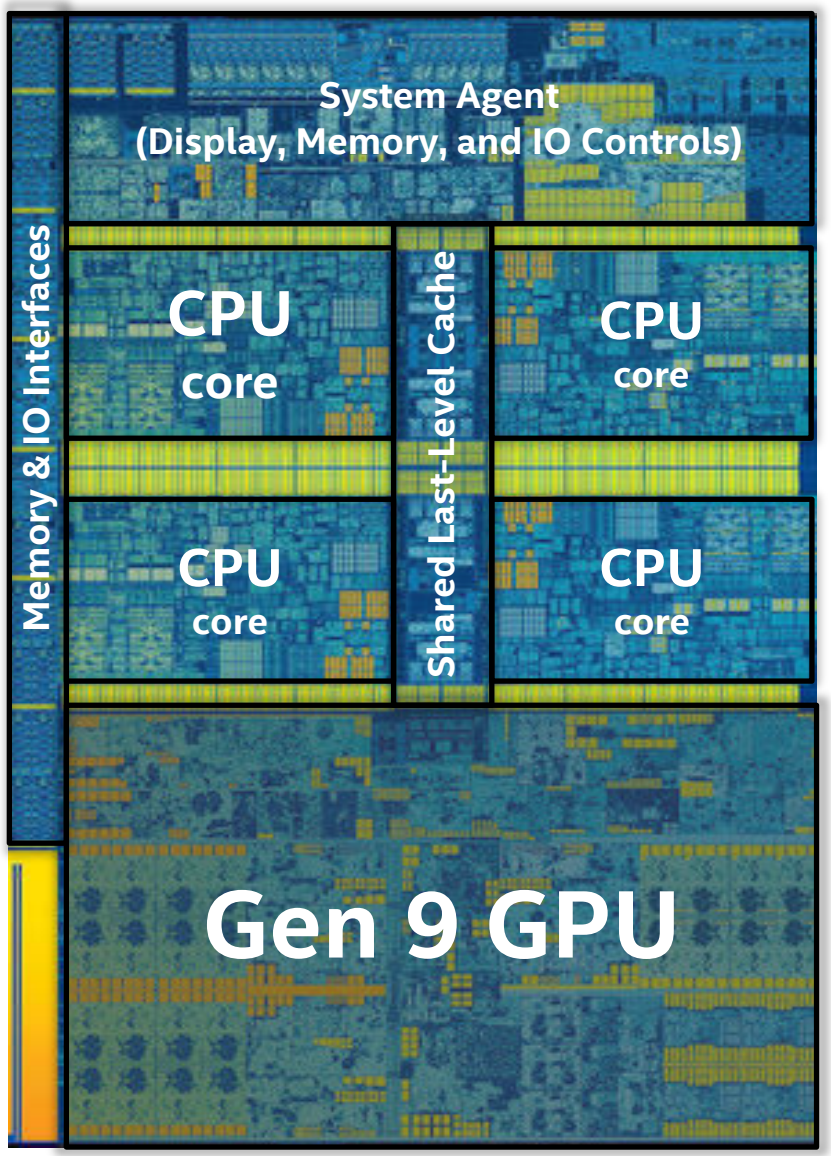
Using Intel® Media SDK and the OpenVINO™ toolkit together enables customers to build high performance, intelligent vision solutions.



Intel Integrated Graphics

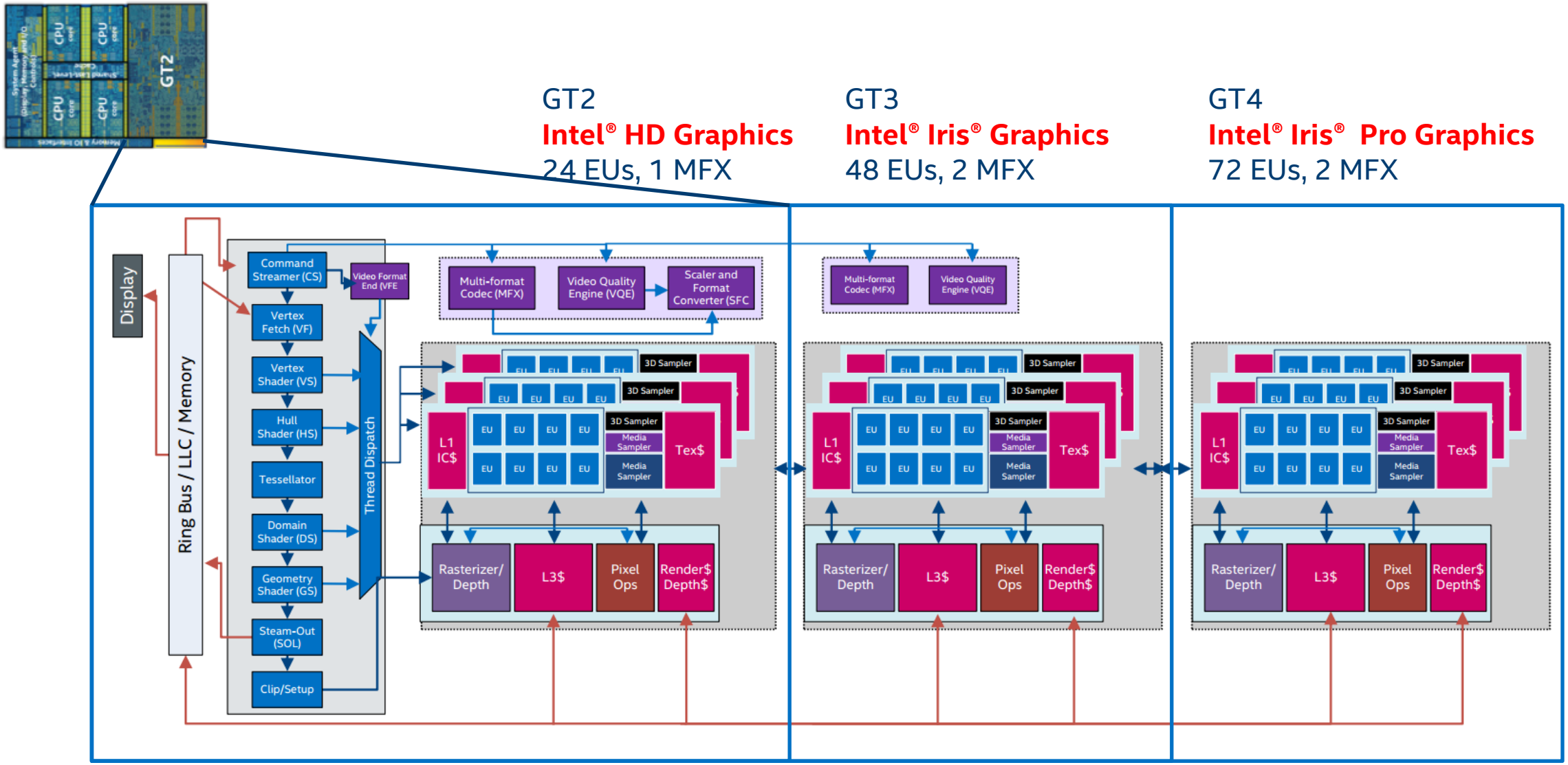
Gen is the internal name for Intel's on-die GPU solution. It's a hardware ingredient with various configurations.

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



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

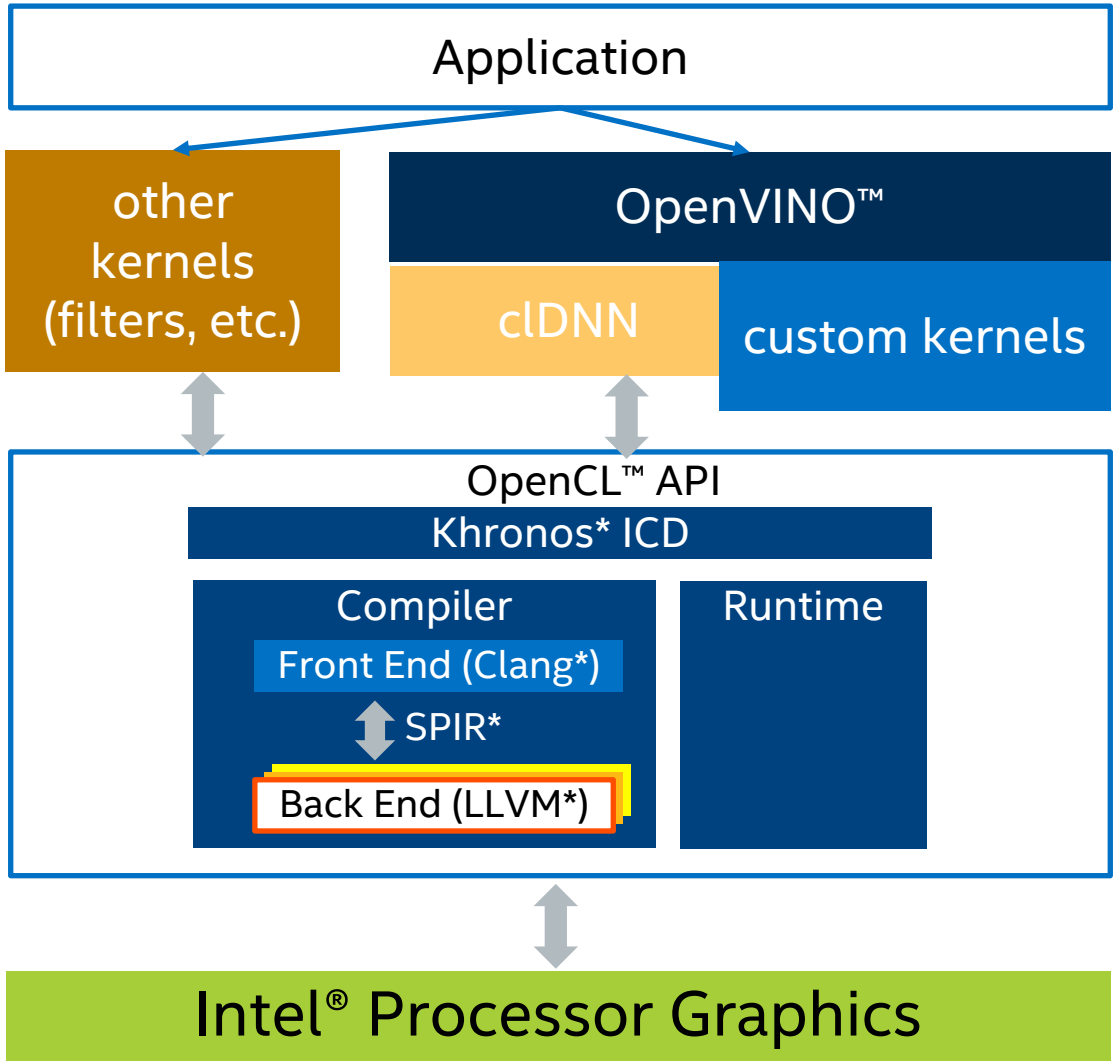
Intel GPU Configurations



OpenCL™

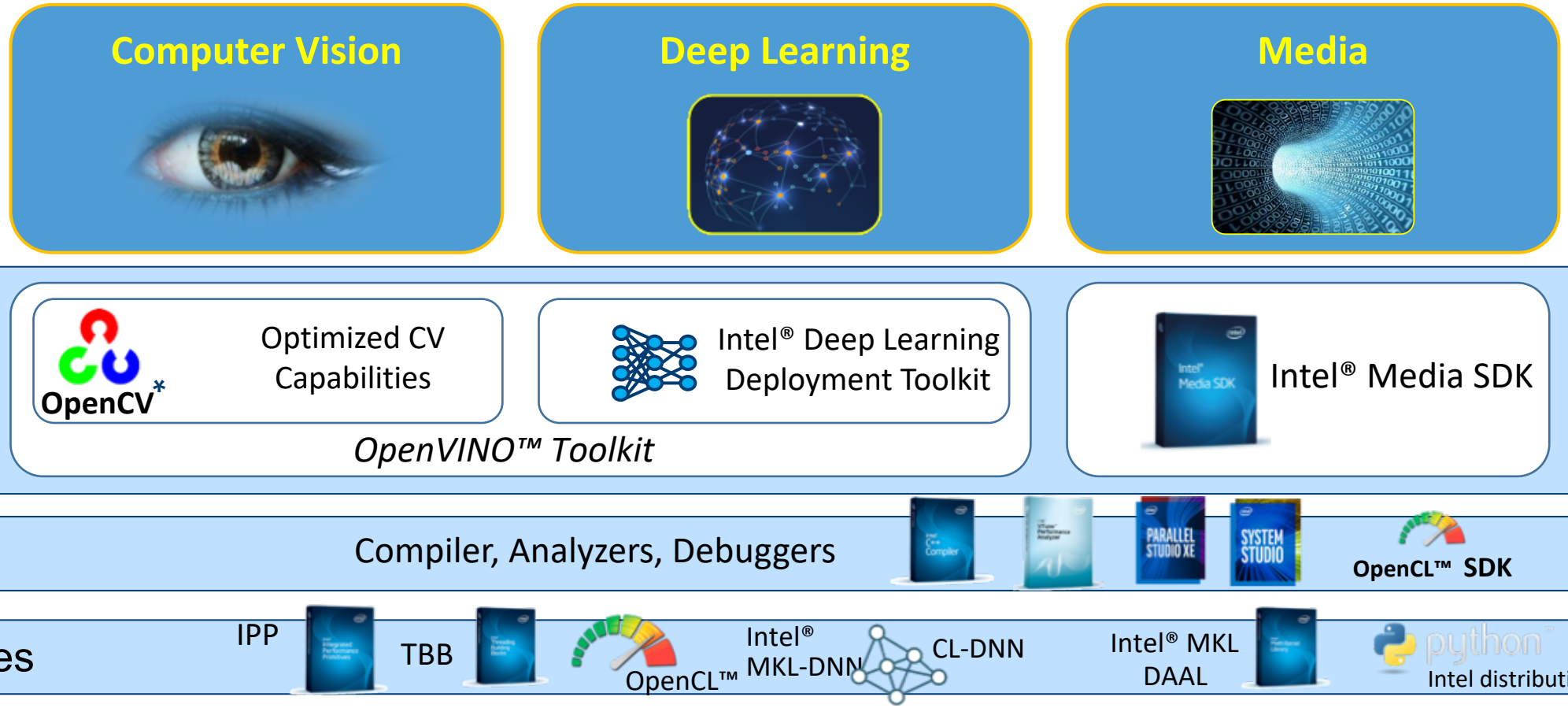
OpenCL™:

- 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



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.

OpenVINO™ 101

2. Basic End-to-End Object Detection Example

Hardware Acceleration

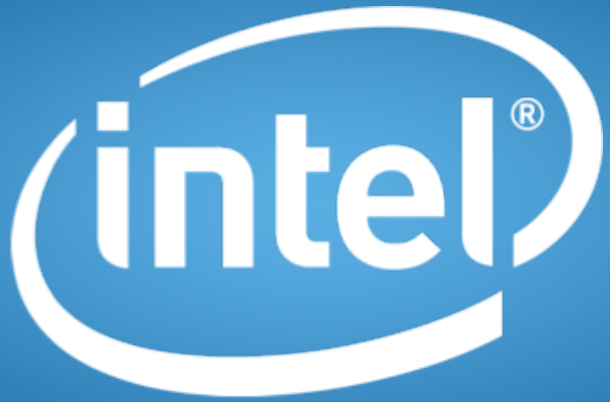
3./4./5. Hardware Acceleration with CPU, Integrated GPU, Intel® Movidius™ NCS, FPGA

Optimization

6. Optimization Tools and Techniques

Application

7. Advanced Video Analytics


intel