
ACCELERATE SMART VISION DEPLOYMENTS
WITH SCALABLE INDUSTRIAL READY KITS &
EASY-TO-USE INFERENCE ENGINE SOFTWARE
Intel® Programmable Solutions Group
Jean-Michel Vuillamy – May 22, 2019

Title or Abstract

Productize Your AI with the OpenVINO™ Toolkit on FPGA

Deep learning boom is now a few years old and while it remains an important research topic, the technology is now mature enough to arrive to production.

We present a toolkit, named OpenVINO™, which optimizes deep learning models for inference in order to execute them on a wide spectrum of hardware in terms of power consumption, cost, and inference performance. This leaves full flexibility for the end user to integrate AI into their products given their given hardware constraints. In this presentation, we will provide a general overview about the tool. We will then describe the general graph optimizations given by our tool and designed to create a smaller and leaner graph read to use for inference. Those optimizations include batch normalization fusion, patterns, and subgraphs replacement.

After that, we will dive deep into the deep learning accelerator (DLA), included in the OpenVINO™ toolkit, and understand how it powers neural network acceleration in FPGAs. We explain how DLA is used to execute neural network graphs on programmable logic even without the need to have detailed FPGA knowledge in the first place. But we will also give insights about how the functionality is physically implemented in the device and how it can be customized by the developer as needed given the fluid nature of neural network development these days.

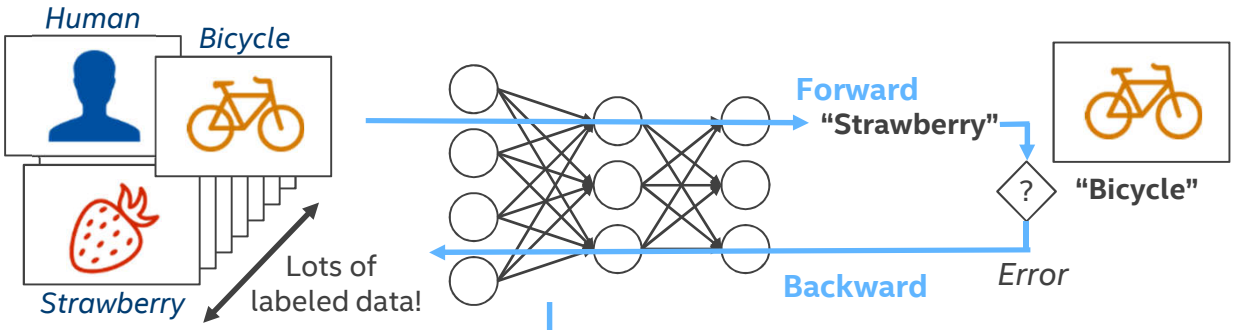
Finally, we will perform a face detection demo based on a SqueezeNet SSD topology running on multi-streaming videos where we demonstrate that FPGA technology enables the scaling of number of channels whilst maintaining frame rate, providing significant acceleration compared to CPU.

Agenda

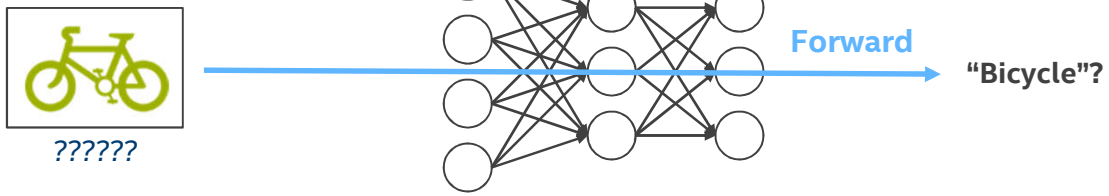
- OpenVINO™ toolkit
  - Overview
  - Graph Optimizations
  - Inference Engine
- Deep learning acceleration
  - Machine Learning on FPGA
  - Intel® FPGA Deep Learning Acceleration Suite
  - Execution on the FPGA
- Demo

Deep Learning: Training vs. Inference

TRAINING

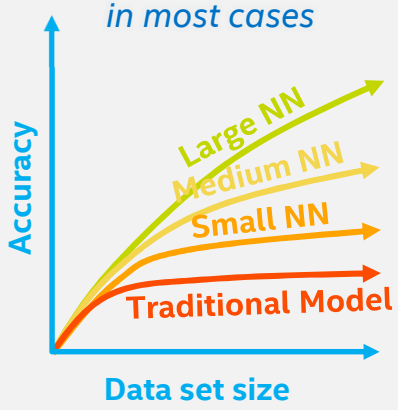


INFERENCE

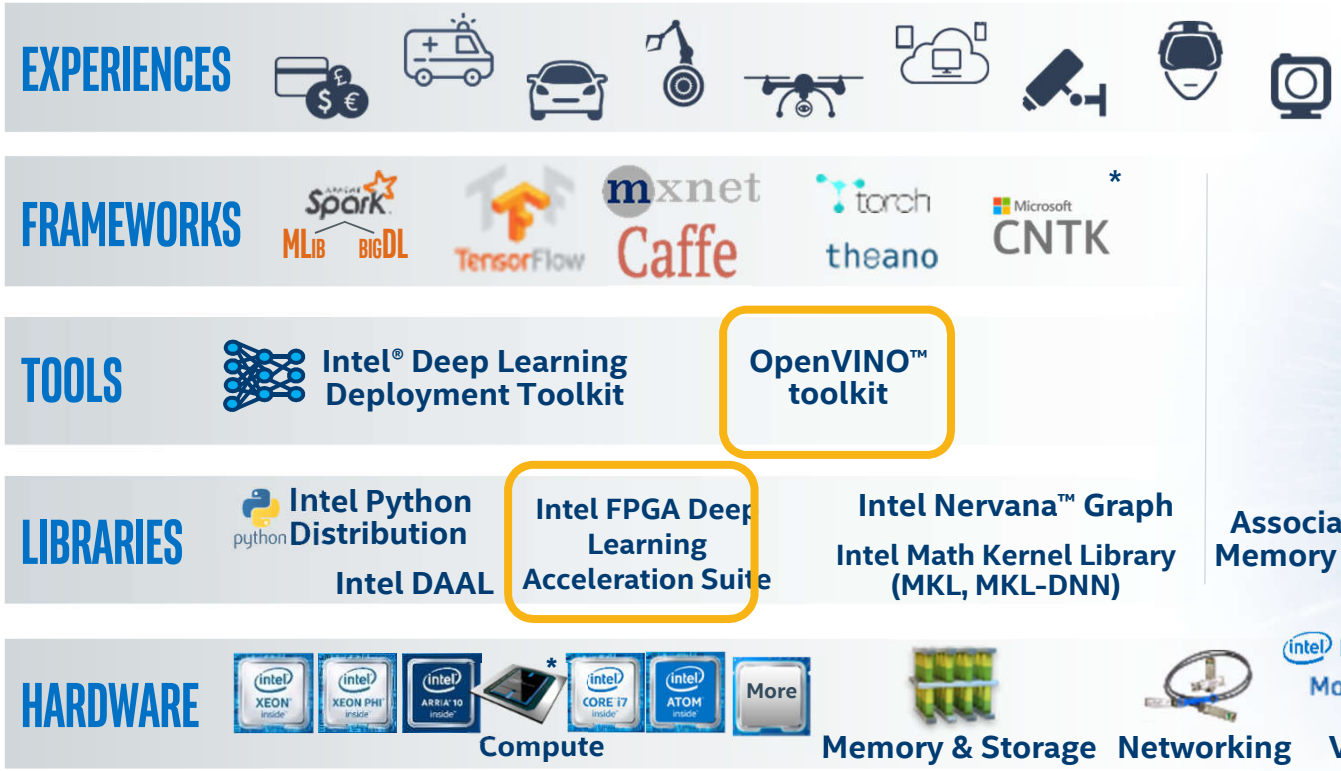


DID YOU KNOW?

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



Intel® AI Portfolio



UNLEASH FULL POTENTIAL

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

OpenVINO™ Toolkit

- A toolkit to accelerate the development of **high-performance computer vision and deep learning into vision applications**
- Increase application performance through Intel® accelerators and flexible heterogeneous architectures (CPU, CPU w/ integrated GPU, FPGA, and Intel Movidius™ Vision Processing Unit)
- Easy deployment across multiple Intel platforms using **one single API**
- Accelerate workloads for a wide range of solutions and vertical use cases
- Drive power, cost, and development efficiencies to designs and applications

What's Inside Intel® Distribution of OpenVINO™ Toolkit

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

30+ Pre-trained Models

Computer Vision Algorithms

Samples

IR = Intermediate Representation file

Traditional Computer Vision

Optimized Libraries & Code Samples

OpenCV*

OpenVX*

Samples

For Intel® CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

Intel Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For GPU/Intel Processor Graphics

Optimize Intel FPGA (Linux* only)

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

Bitstreams

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)

Intel Architecture-Based Platforms Support

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

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

Intel® Deep Learning Deployment Toolkit

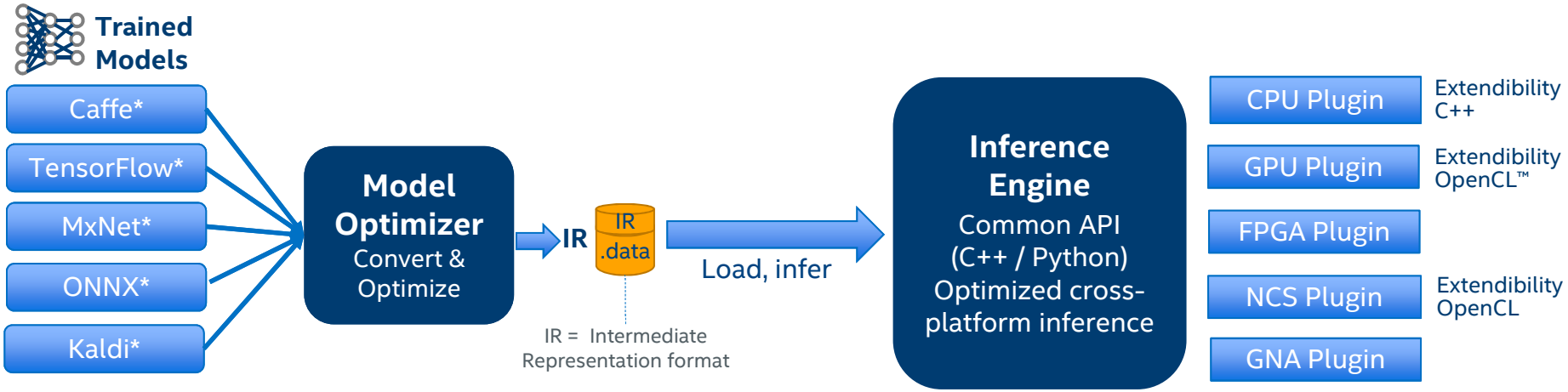
For Deep Learning Inference

Model Optimizer

- **What it is**: A python-based tool to import trained models and convert them to Intermediate representation.
- **Why important**: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.

Inference Engine

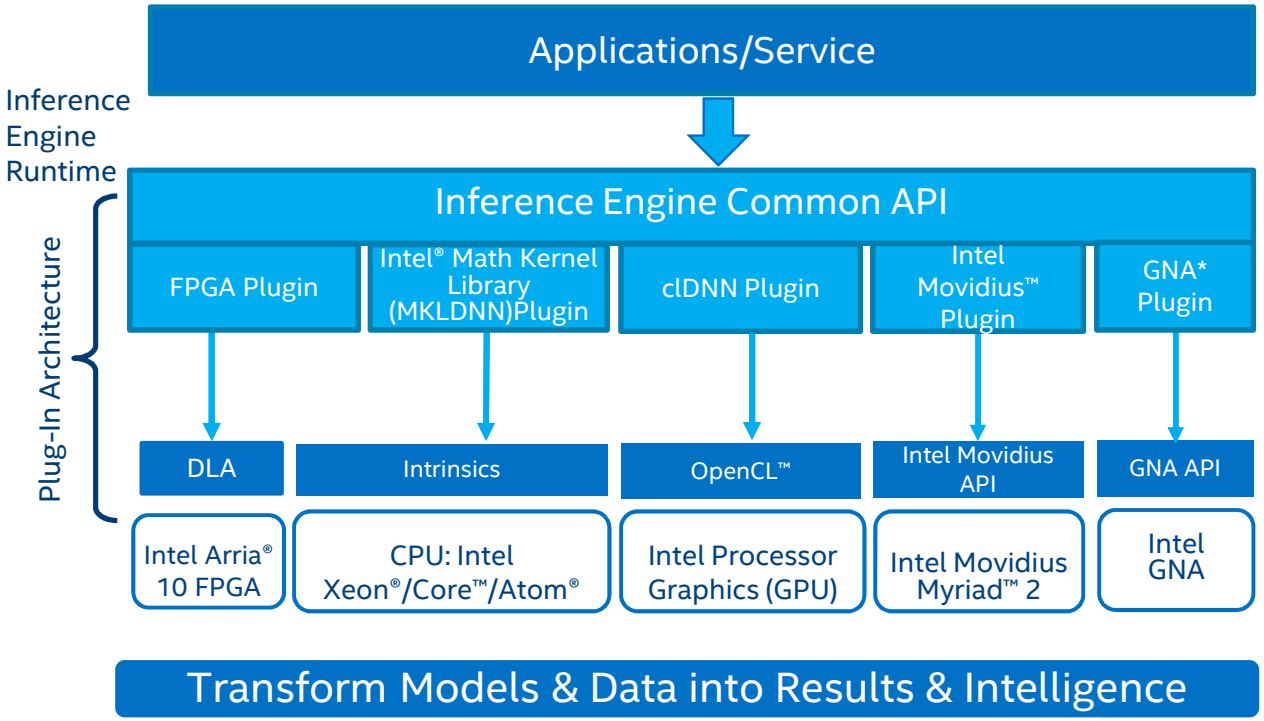
- **What it is**: High-level inference API
- **Why important**: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.



GPU = Intel CPU with integrated graphics processing unit/Intel Processor Graphics

Optimal Model Performance Using the Inference Engine

- Simple and unified API for inference across all Intel® architecture
- Optimized inference on large IA hardware targets (CPU/GEN/FPGA)
- Heterogeneity support allows execution of layers across hardware types
- Asynchronous execution improves performance
- Future-proof or scale your development for future Intel processors



GPU = Intel CPU with integrated graphics/Intel® Processor Graphics/GEN
GNA = Gaussian mixture model and Neural Network Accelerator

Speed Deployment with Intel Optimized Pre-trained Models

The OpenVINO™ toolkit includes optimized pre-trained models to expedite development and improve deep learning inference on Intel® processors. Use these models for development and production deployment without the need to search for or to train your own models.

Pre-Trained Models

- 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
- Text Detection
- Vehicle Detection
- Retail Environment
- Pedestrian Detection
- Pedestrian & Vehicle Detection
- Person Attributes Recognition Crossroad
- Emotion Recognition
- Identify Someone from Different Videos – standard & enhanced
- Facial Landmarks
- 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)

Save Time with Deep Learning Samples and Computer Vision Algorithms

Samples

Use Model Optimizer and Inference Engine for public models and Intel® pretrained models.

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

Computer Vision Algorithms

Start quickly with highly-optimized, ready-to-deploy, custom-built algorithms using Intel pretrained models.

- Face Detector
- Age & Gender Recognizer
- Camera Tampering Detector
- Emotions Recognizer
- Person Re-identification
- Crossroad Object Detector
- License Plate Recognition
- Vehicle Attributes Classification
- Pedestrian Attributes Classification

Agenda

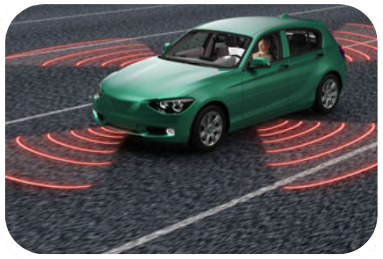
- OpenVINO™ toolkit
  - Overview
  - Graph Optimizations
  - Inference Engine
- **Deep learning acceleration**
  - **Machine Learning on FPGA**
  - **Intel® FPGA Deep Learning Acceleration Suite**
  - **Execution on the FPGA**
- Demo

Solving Machine Learning Challenges with Intel® FPGA

EASE OF USE

SOFTWARE ABSTRACTION, PLATFORMS & LIBRARIES

Intel® FPGA solutions enable software-defined programming of customized machine learning accelerator libraries.



REAL TIME

DETERMINISTIC LOW LATENCY

Intel FPGA hardware implements a deterministic low-latency datapath unlike any other competing compute device.

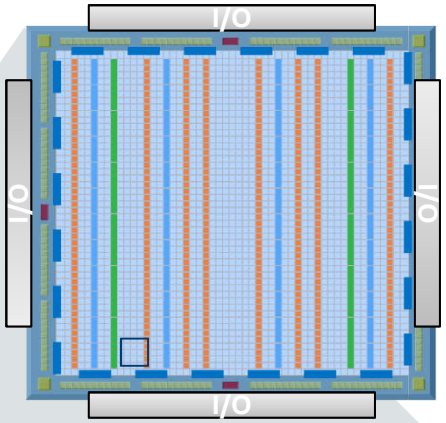
FLEXIBILITY

CUSTOMIZABLE HARDWARE FOR NEXT GEN DNN ARCHITECTURES

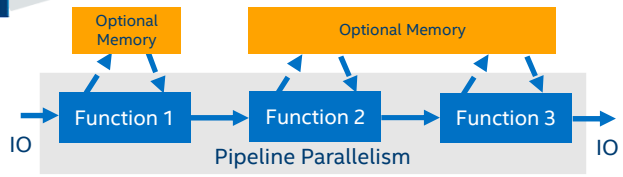
Intel FPGAs can be customized to enable advances in machine learning algorithms.

Why Intel® FPGAs for Machine Learning?

Convolutional Neural Networks are Compute Intensive



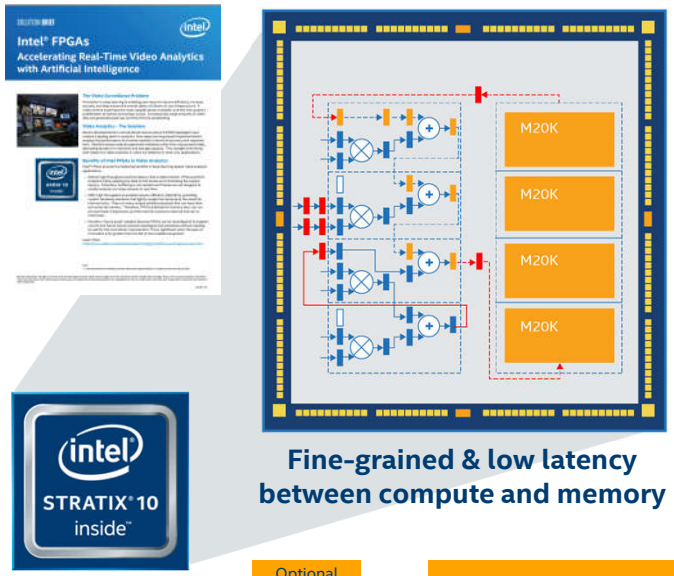
Fine-grained & low latency between compute and memory



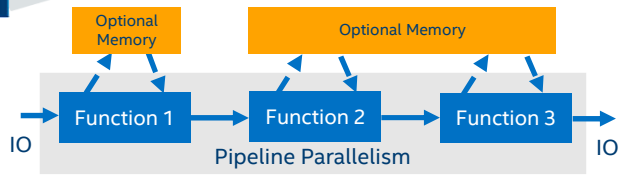
Feature	Benefit
Highly parallel architecture	Facilitates efficient low-batch video stream processing and reduces latency
Configurable Distributed Floating-Point DSP Blocks	FP32 10+ TFLOPS & FP16, FP11 support Accelerates computation by tuning compute performance
Tightly coupled high-bandwidth memory	>50 TBps on-chip SRAM bandwidth, random access, reduces latency, minimizes external memory access
Programmable Datapath	Reduces unnecessary data movement, improving latency and efficiency
Configurability	Support for variable precision (trade-off throughput and accuracy). Future proof designs and system connectivity

Why Intel® FPGAs for Machine Learning?

Convolutional Neural Networks are Compute Intensive



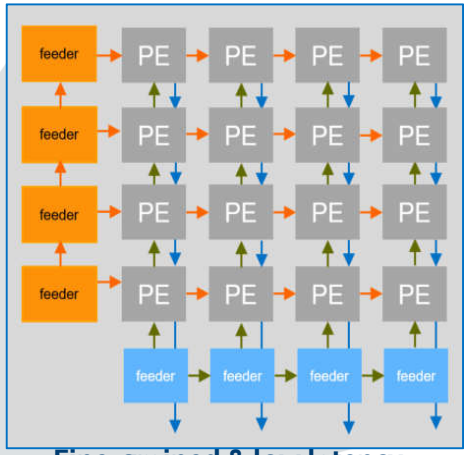
Fine-grained & low latency between compute and memory



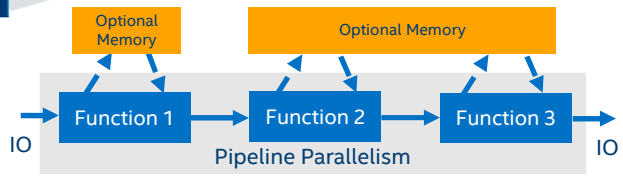
Feature	Benefit
Highly parallel architecture	Facilitates efficient low-batch video stream processing and reduces latency
Configurable Distributed Floating-Point DSP Blocks	FP32 10+ TFLOPS & FP16, FP11 support Accelerates computation by tuning compute performance
Tightly coupled high-bandwidth memory	>50 TBps on-chip SRAM bandwidth, random access, reduces latency, minimizes external memory access
Programmable Datapath	Reduces unnecessary data movement, improving latency and efficiency
Configurability	Support for variable precision (trade-off throughput and accuracy). Future proof designs, and system connectivity

Why Intel® FPGAs for Machine Learning?

Convolutional Neural Networks are Compute Intensive

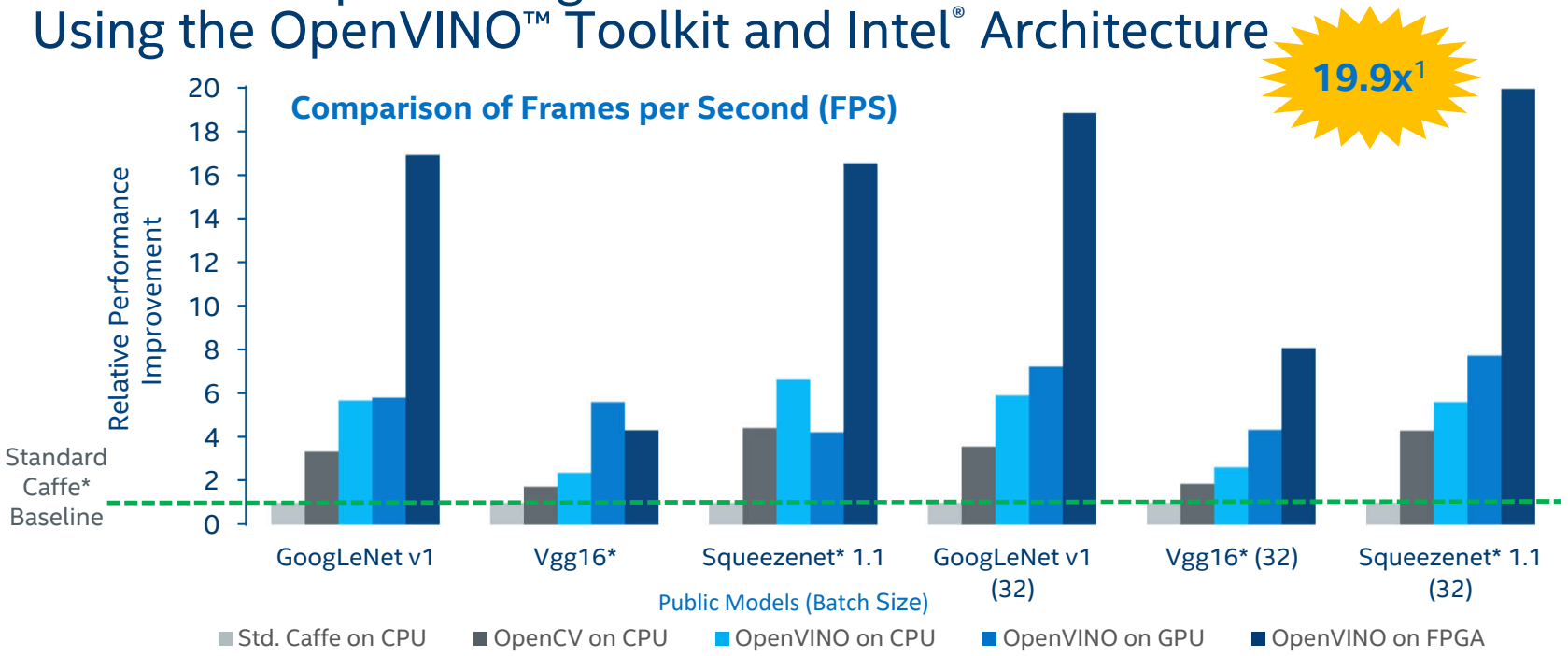


Fine-grained & low latency between compute and memory



Feature	Benefit
Highly parallel architecture	Facilitates efficient low-batch video stream processing and reduces latency
Configurable Distributed Floating-Point DSP Blocks	FP32 10+ TFLOPS & FP16, FP11 support Accelerates computation by tuning compute performance
Tightly coupled high-bandwidth memory	>50 TBps on-chip SRAM bandwidth, random access, reduces latency, minimizes external memory access
Programmable Datapath	Reduces unnecessary data movement, improving latency and efficiency
Configurability	Support for variable precision (trade-off throughput and accuracy). Future proof designs, and system connectivity

Increase Deep Learning Workload Performance on Public Models Using the OpenVINO™ Toolkit and Intel® Architecture

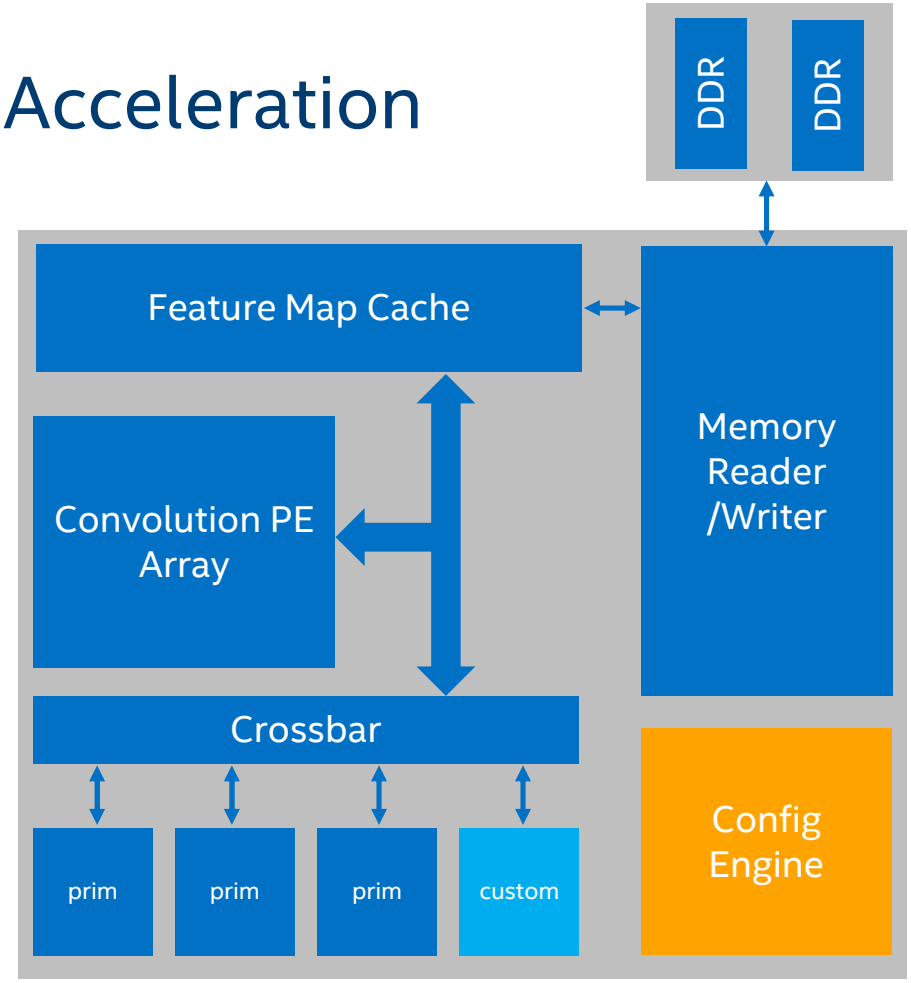


Get an Even Bigger Performance Boost with Intel® FPGA

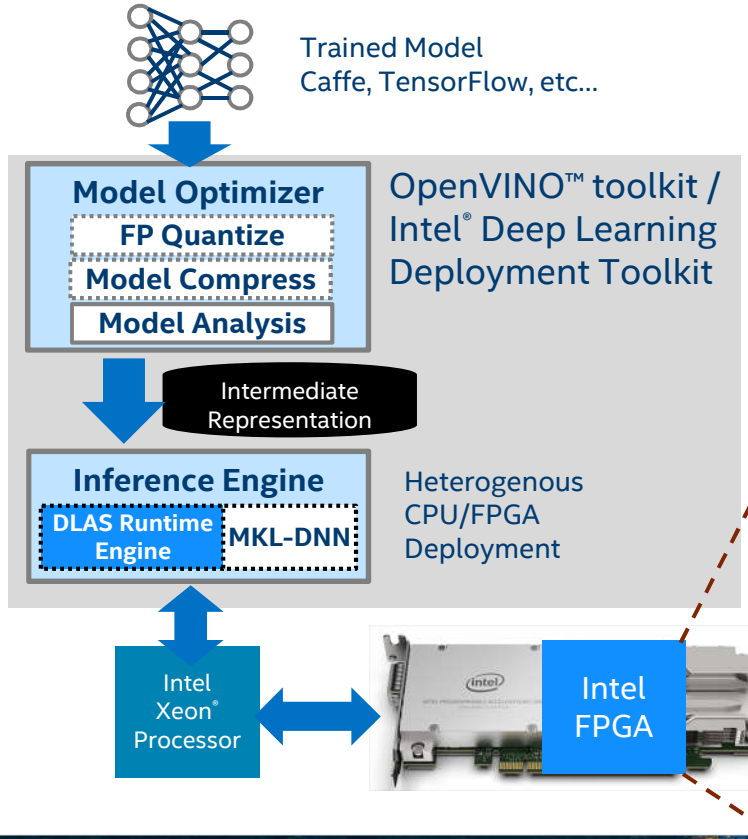
[1]Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. The benchmark results reported in this deck may need to be revised as additional testing is conducted. 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 @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, performed 4/10/2018 Test v312.30 – Ubuntu* 16.04, OpenVINO™ 2018 RC4. Intel® Arria 10-1150GX FPGA. 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.

Intel® FPGA Deep Learning Acceleration Suite Features

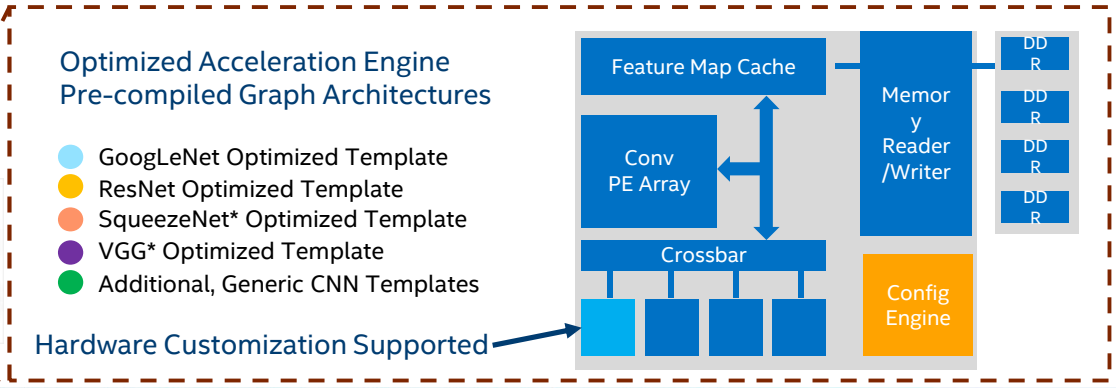
- CNN acceleration engine for common topologies executed in a graph loop architecture
  - AlexNet, GoogleNet, LeNet, SqueezeNet*, VGG16*, ResNet, Yolo, SSD…
- Software Deployment
  - No FPGA compile required
  - Run-time reconfigurable
- Customized Hardware Development
  - Custom architecture creation with parameters
  - Custom primitives using OpenCL™ flow



FPGA Usage with OpenVINO™ Toolkit



- Supports common software frameworks (Caffe, TensorFlow)
- Model Optimizer enhances model for improved execution, storage, and transmission
- Inference Engine optimizes inference execution across Intel® hardware solutions using unified deployment application programming interface (API)
- Intel® FPGA Deep Learning Acceleration Suite provides turn-key or customized CNN acceleration for common topologies



Application Acceleration with Intel® FPGA Powered Platforms

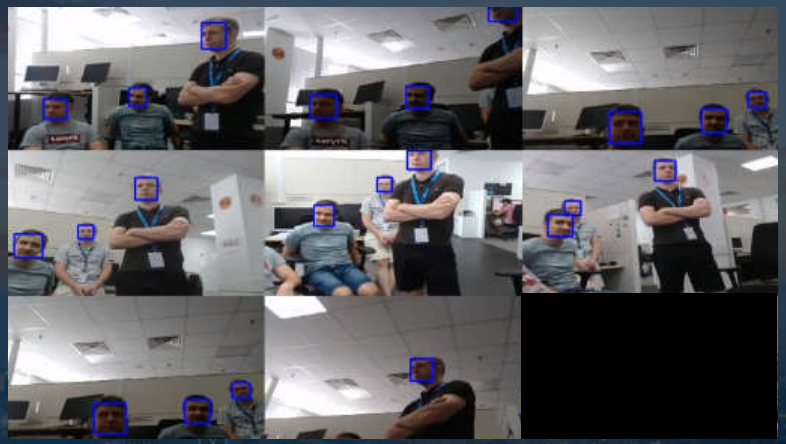
SUPPORTED PLATFORMS FOR FPGA	 Intel® Arria® 10 FPGA Development Kit	 Intel Programmable Acceleration Card with Intel Arria 10 GX FPGA	 Mustang F-100
INTERFACE	PCI Express* (PCIe x8)*	PCIe x8	PCIe x8
CURRENTLY MANUFACTURED BY[1]	INTEL®	INTEL®	iEi
SOFTWARE TOOLS	**OPENVINO™ TOOLKIT** Develop NN Model; Deploy across Intel® CPU, GPU, VPU, FPGA; Leverage common algorithms		

*Please contact Intel representative for complete list of ODM manufacturers. Other names and brands may be claimed as the property of others.

Agenda

- OpenVINO™ toolkit
  - Overview
  - Graph Optimizations
  - Inference Engine
- Deep learning acceleration
  - Machine Learning on FPGA
  - Intel® FPGA Deep Learning Acceleration Suite
  - Execution on the FPGA
- **Demo**

MultiChannel Demo



FPGA acceleration enables scaling of number of channels whilst maintaining frame rate

Summary

- FPGAs provide a flexible, deterministic low-latency, high-throughput, and energy-efficient solution for accelerating AI applications
- Intel® FPGA Deep Learning Acceleration Suite supports CNN inference on FPGAs
- Accessed through the OpenVINO™ toolkit
- DLA architecture can be customized for best performance
- Available for Intel® Arria® 10 FPGAs today

Call to Action, Resources

Download ▸

[Free Intel® Distribution of OpenVINO™ toolkit]

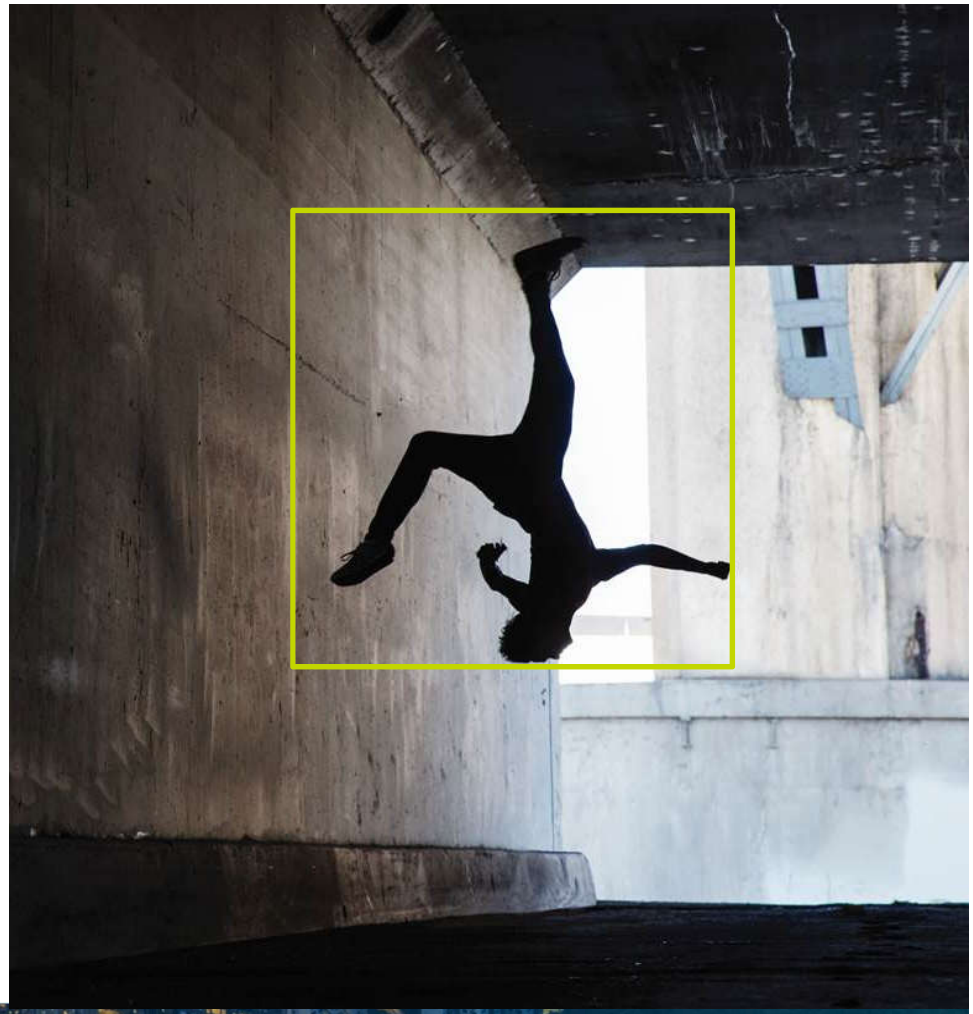
Get started quickly with

- OpenVINO toolkit developer resources
- Intel technology, decoded online webinars, tools, how-tos, and quick tips
- Hands-on developer workshops
- Intel® FPGA resources
- AI powered by Intel FPGAs

Support

- Connect with Intel engineers and computer vision experts at the public Community Forum

Select Intel customers may contact their Intel representative for issues beyond forum support.



Legal Disclaimers


intel®