WHITE PAPER

Intel® Vision Products
Deep Learning Inference Acceleration

Introducing the Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

Exceptional performance, flexibility, and scalability for deep-learning and computer-vision solutions

Authors

Richard Chuang, Ph.D.
Global Platform and Solutions Architect
Strategic Business Architect
Intel IOTG VMC Video

Omi Oliyide
Strategic Business Development Manager
Intel PSG VBG

Bob Garrett
Product Marketing Manager
Intel PSG PM

Table of Contents

Executive Summary

Rapid growth in the use of smart video cameras has also led to a massive increase in the volume of data that organizations must process and analyze to extract useful information and meaningful insights.

Yet, many organizations face significant challenges in developing scalable solutions with the deep-learning, computer-vision, and video-analytics capabilities that would enable them to manage their video data more efficiently. Whether those challenges arise from form-factor constraints, the cost of hardware, the size of their power envelope, or some combination of these and other factors, the result is the same. Organizations continue to struggle with the growing flood of data and their need to transform that data into insights they can use.

To help organizations overcome these challenges, Intel created the new Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA. These new cards offer exceptional performance, flexibility, and scalability for deep-learning and computer-vision solutions, but with significantly lower costs and power requirements and with a form factor half the size of Intel Arria 10 FPGA development cards. In addition, the Intel Vision Accelerator Design with Intel Arria 10 FPGA works seamlessly with the OpenVINO™ toolkit, offering developers a simpler path to deep learning and computer vision solutions and ensuring continual performance optimization through periodic FPGA bitstream updates.

1. Introducing the Intel Vision Accelerator Design with Intel Arria 10 FPGA

1.1 Challenges in Deep Learning

The proliferation of smart cameras and the explosion of video data, combined with longer retention periods and higher image resolution, poses a major challenge for many organizations as they struggle to collect, process, organize, and extract meaningful information and insights from these large datasets. Furthermore, this rapid increase in data is placing tremendous capacity and performance demands on compute, storage, and networking resources, leading to inefficiencies and higher costs, and stretching many existing infrastructures to their limits.

Video analytics solutions built on Intel Vision Accelerator Design products are making it possible for organizations to transform big data challenges into opportunities, by eliminating the need for human operators, consistently delivering fast and accurate results, and reducing storage and network requirements by orders of magnitude. The latest Intel Vision Accelerator Design products combine the silicon efficiency of Intel® FPGAs (field-programmable gate arrays) with a common deep-learning software toolkit, bringing compute efficiency to the network edge by enabling a new generation of deep-learning inference

applications in edge servers and making it possible to create solutions that can be easily scaled and updated over many years.

Intel Vision Accelerator Design products support a wide range of use cases—in smart cities, healthcare, transportation, industrial, and retail applications—distributing intelligence from camera to cloud, and accelerating deep learning inference on devices and edge appliances.

1.2 Intel Vision Accelerator Design with Intel Arria 10 FPGA Product Description

The Intel Vision Accelerator Design with Intel Arria 10 FPGA offers exceptional performance, flexibility, and scalability for deep-learning and computer-vision solutions—from NVRs (network video recorders) to edge deep-learning inference appliances to on-premises servers—at a fraction of the cost and with significantly lower power requirements than most of the existing FPGA PCIe cards. (See Figure 1.)

Programmable, software-defined Intel Arria 10 FPGAs ensure continual performance optimization—taking advantage of periodic FPGA bitstream updates provided by Intel—without necessitating hardware upgrades. The cards are designed for long product life (15 years of longevity for FPGA products) and can adapt to a wide range of work conditions, including harsh industrial and/or outdoor environments.

Unlike fixed-function devices, functionality in Intel Arria 10 FPGAs can always be changed or modified to increase or deepen intelligence, which allows them to be architected to solve very specific problems. And when used for deep learning inference, these FPGAs achieve high-performance images per second at reduced power while providing dynamic flexibility, consistent power consumption, and future-proofing for custom or new workloads as well as low latency across a wide spectrum of vision use cases and applications.

The Intel Vision Accelerator Design with Intel Arria 10 FPGA offers a simplified path for developers to run customized topologies in an optimal way and works seamlessly with the OpenVINO toolkit. Fine-grained parallelism enables high throughput on low-batch workloads. The extremely high, fine-grained, on-chip memory bandwidth can more efficiently solve memory challenges. With the OpenVINO toolkit, you don't need to be an FPGA expert to code applications integrating computer vision.

The Intel Vision Accelerator Design with Intel Arria 10 FPGA can support more than 20 channels of video inputs, along with vision use cases such as facial detection and recognition. Optimized network topologies include GoogLeNet v1*, ResNet-18/50/101*, SqeezeNet v1.1*, VGG-16*, and MobileNet v1*. More hardware accelerated topologies are coming with new OpenVINO toolkit releases. Customizable data paths and precision create energy-efficient dataflow for system-level optimization. The Intel Vision Accelerator Design with Intel Arria 10 FPGA is also ideal for complex or large workloads as well as customized applications and use cases where you may want to add your own primitives or sub-layer configurations.

Because of their innate adaptability, FPGAs are always on the cutting edge of new solutions and changing network topologies. This flexibility, and the ability to accelerate algorithm processing, make this Intel Vision Accelerator Design invaluable for complex, massive deep-learning analysis and visual intelligence.

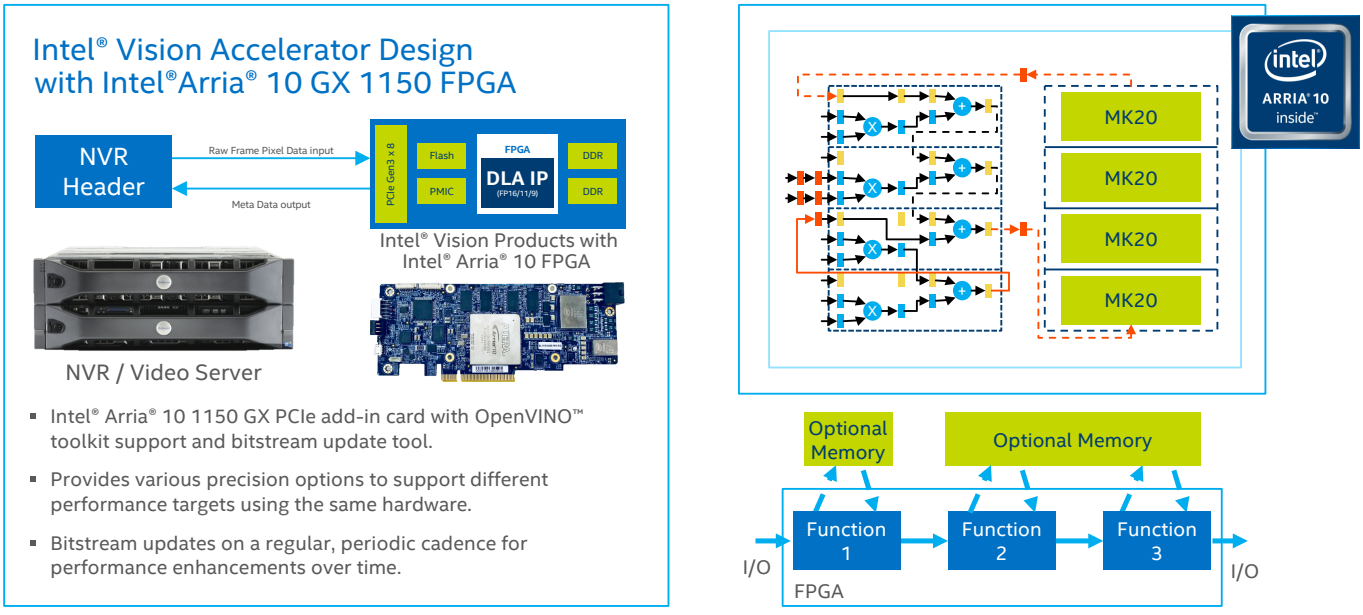


Figure 1. Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

Specifications

The Intel Vision Accelerator Design with Intel Arria 10 FPGA is half the height and half the length of a standard full-size FPGA add-in card, making it an ideal choice for anyone who needs a smaller form factor. Because the new cards are optimized for deep learning acceleration, Intel has been able to remove all components that do not contribute to that goal, reducing the size of the form factor and lowering the cost to an extremely competitive level. The Intel Vision Accelerator Design with Intel Arria 10 FPGA also offers a choice of active or passive thermal solutions, so you can select the design that best fits your needs.

Specifications of the Intel Vision Accelerator Design with Intel Arria 10 FPGA include:

- Deep learning network that requires larger memory footprint (> 2Mparams)
- Various precision options (FP16/11/9) to support different performance targets using the same hardware
- Intel Arria 10 1150 GX FPGA PCIe* add-in card with OpenVINO toolkit support and bitstream update tool
- Bitstream updates on a regular cadence for performance enhancements over time
- Batch size flexible (1 to N), especially strong in low-batch size setting for mission-critical applications
- Intel Arria 10 1150 GX FPGAs delivering up to 1.5 TFLOPs
- Interface: PCIe Gen3 x 8
- Form Factor: Standard Half-Height, Half-Length.
- Cooling: Active fan
- Operation Temperature: 0°C~65°C (ambient temperature)
- Operation Humidity: 5% to 90% relative humidity
- Power Consumption: < 60W, 38~42W typical
- Power Connector: 12V external power
- DIP Switch/LED Indicator: Indicating device number for multiple card support

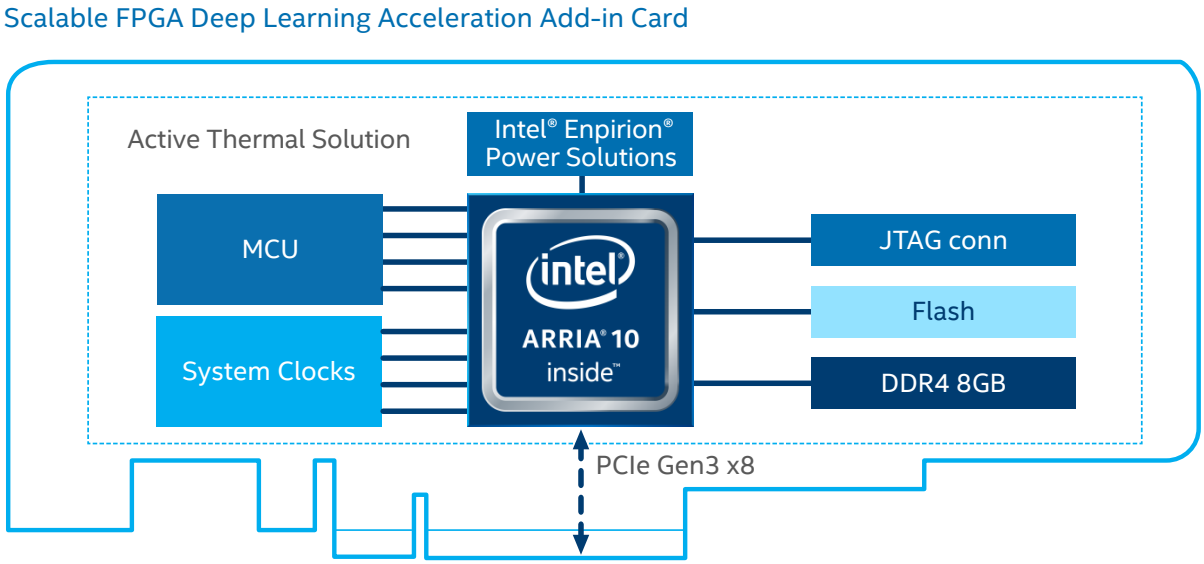


Figure 2. Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

1.3 Why FPGAs

Intel FPGAs feature a highly parallel architecture, tightly coupled high-bandwidth memory, and a programmable data path that reduces unnecessary data movement. The result is reduced latency, greater efficiency and flexibility, and exceptional low-batch DL inferencing throughput. The cards also feature configurable floating point DSP blocks to accelerate computation, particularly the computations involved in convolutions. Intel FPGA hardware provides deterministic low-latency and real-time inference, implementing a deterministic low-latency data path unlike any other competing compute device.

Deep learning is undergoing constant innovation, and efforts to improve throughput, performance, and efficiency are ongoing. This presents serious challenges when solutions are implemented on a fixed architecture such as a GPU. Intel FPGAs are future proof by design and can be customized to enable advances in machine learning algorithms.

Table 1.

FPGA FEATURE	BENEFIT
Highly Parallel Architecture	Reduces latency. Facilitates low batch processing.
Floating Point DSP Blocks	Accelerates computation (e.g., convolution).
Tightly Coupled High-bandwidth Memory	Reduces latency.
Programmable Data Path	Reduces latency, improves efficiency.
Configurability	- Optimizes latency v. accuracy and efficiency. - Future-proof designs. - Accelerate custom OpenCL kernels in FPGA

FPGAs can Leverage Parallelism Across the Entire Chip, Reducing the Compute Time to a Fraction

System Latency = I/O Latency + Compute Latency

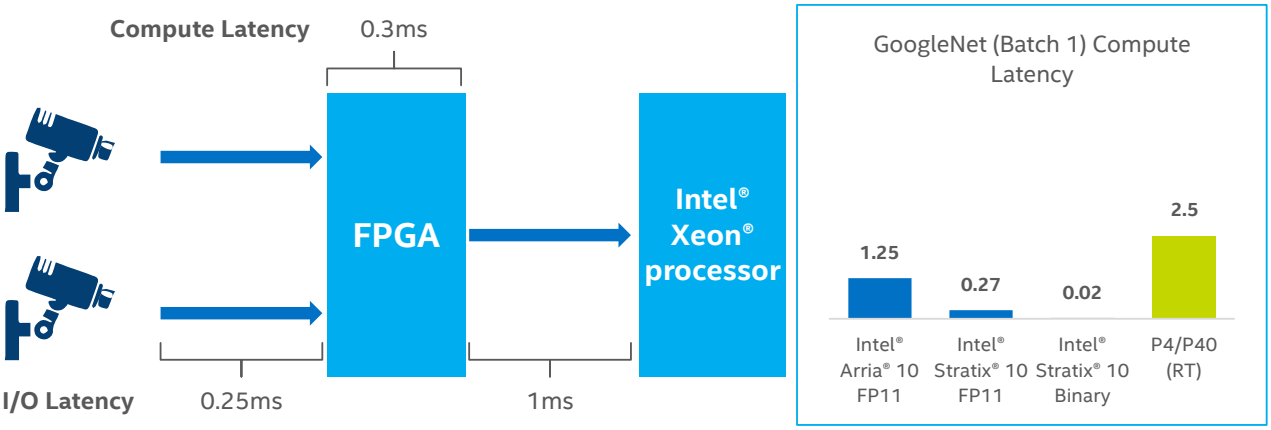


Figure 3. FPGAs Provide Deterministic System Latency

2. Introduction to Intel Arria 10 and Why it was Chosen

2.1 Introducing the Intel Arria 10 FPGAs

Intel Arria 10 FPGAs deliver more than a speed grade faster core performance, using publicly available OpenCore designs.[1] Intel Arria 10 FPGAs are up to 40 percent lower power than previous generation FPGAs and feature the hardened floating-point digital signal processing (DSP) blocks, with speeds up to 1.5 tera floating-point operations per second (TFLOPS).[1]

In addition to fast processing, high-performance video analytics, and deterministic low latency, Intel Arria 10 FPGAs also provide the highest performance 2,400 Mbps DDR4 SDRAM memory interface, and one of the industry's mid-range FPGAs with 25.78 Gbps transceivers. Ninety-six transceiver lanes deliver 3.3 Tbps of serial bandwidth.[1] These key features are everything needed to implement deep-learning acceleration (DLA) IPs in FPGA.

3. OpenVINO Toolkit for Computer Vision and Deep Learning

The OpenVINO toolkit is a free, downloadable toolkit that helps developers fast-track the development of high-performance computer vision and deep learning into flexible, cost-effective vision applications. It enables deep learning on hardware accelerators and streamlined heterogeneous execution across multiple types of Intel® platforms. It includes the Intel® Deep Learning Deployment Toolkit with a model optimizer and inference engine, along with optimized computer vision libraries and functions for OpenCV* and OpenVX*. This comprehensive toolkit supports the full range of vision solutions, speeding computer vision workloads, streamlining deep-learning deployments, and enabling easy, heterogeneous execution across Intel platforms from device to cloud.

The OpenVINO toolkit brings a new level of customization: you can match performance to the application requirements, and then deploy it across any product in the Intel Vision Accelerator Design family. With Intel, you can select the right accelerator for camera, edge appliance, or cloud while using consistent APIs and runtimes. Code developed on the OpenVINO toolkit can be deployed across any of Intel's architectures (CPU, CPU with integrated graphics, VPU or FPGA) to streamline time to market.

Download the OpenVINO toolkit for free today to optimize traditional computer vision functions and migrate your deep-learning algorithms to high-performing video applications across a variety of Intel hardware.

Key Benefits

- **Develop and deploy high-performance deep learning applications across Intel platforms:** With the OpenVINO toolkit and free deep learning on host platforms, developers can build high-performance computer vision applications and simplify deep-learning deployment on all Intel hardware.
- **Code portability:** Apps will run across all Intel® Vision products, while the Intel Vision Accelerator Design products add scalable acceleration.
- **Match inference performance to your product requirements:** Deploy trained networks across Intel's heterogeneous solutions with the OpenVINO toolkit.

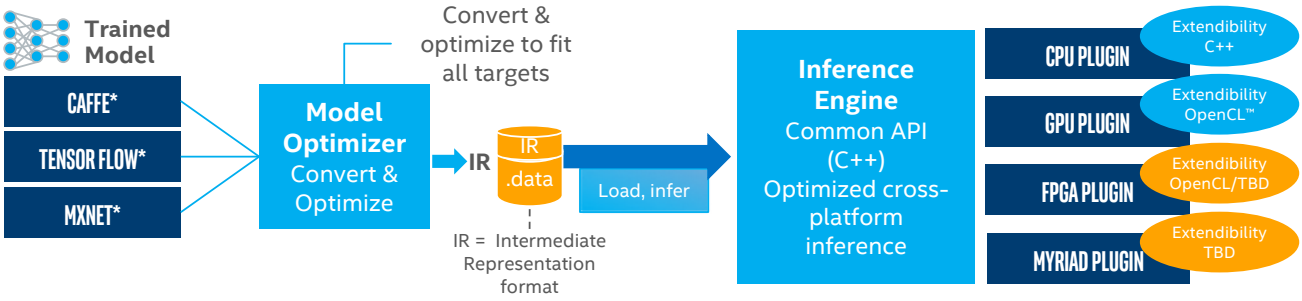
Take Full Advantage of the Power of Intel® Architecture with Intel® Deep Learning Deployment Toolkit

Model Optimizer

- What it is: Preparation step -> imports trained models
- 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
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

Figure 4. OpenVINO toolkit streamlines workflow that supports various popular DL frameworks, as well as abstracts the difference between hardware

4. Intel® FPGA Deep Learning Acceleration Suite

Intel® FPGA Deep Learning Acceleration (DLA) Suite includes a number of key components, including the Intel Deep Learning Deployment Toolkit, with its model optimizer and inference engine. Intel FPGA Deep Learning Acceleration Suite supports common software frameworks such as Caffe*, TensorFlow* and MXNet*, and provides turn-key or customized convolutional neural network (CNN) acceleration for common networks. In addition, the Intel® Deep Learning software stack provides network optimizations. (See Figure 7.)

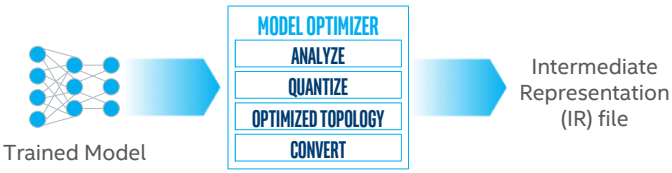
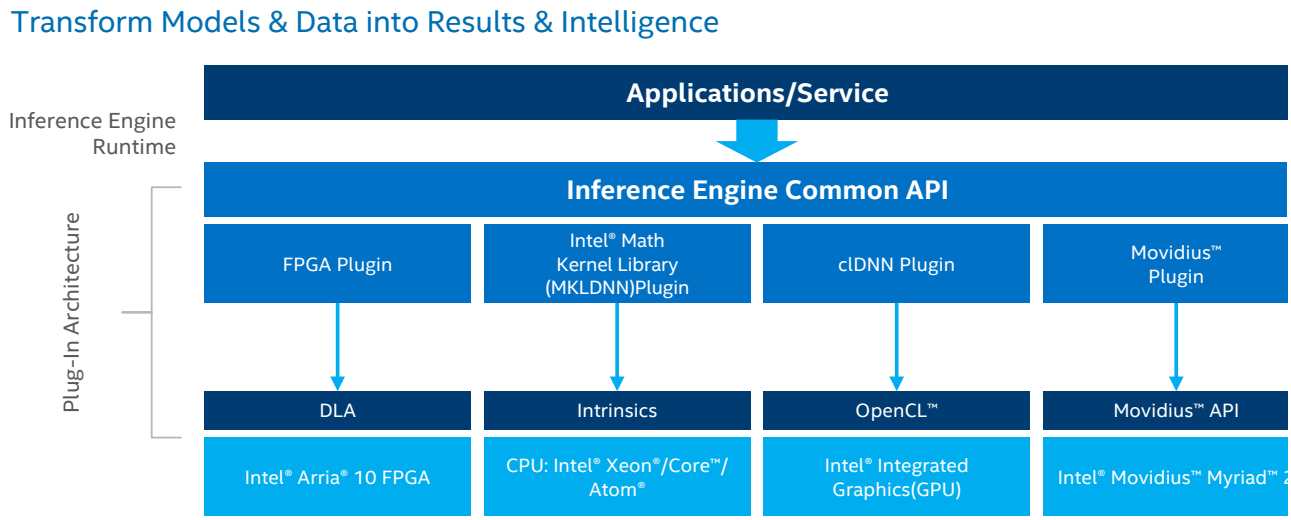


Figure 5. Workflow for optimizing trained model to Intel® architecture

The model optimizer imports a trained model—there are currently more than 100 validated models available—from a wide range of common software frameworks such as Caffe, TensorFlow and MXNet and converts it to an intermediate reciprocation (IR) file that describes the topology.

The inference engine functions as a high-level API that is optimized for cross-platform inference and includes dynamically loaded plug-ins for FPGAs, GPUs, CPUs and other types of hardware. When the IR file is submitted to the inference engine, the API translates the file for the correct plug-ins and automatically delivers the best performance for each type of hardware without requiring users to implement and maintain multiple code pathways. The inference engine decides where to deploy the file. That process is transparent to users, making the complexity of using the FPGA near zero.



GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics/GEN
OpenVX and the OpenVX logo are trademarks of the Khronos Group Inc.
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

Figure 6. Inference engine offers a unified API for developers to deploy workloads to various hardware accelerators

This architecture lets developers accelerate vision solution development with heterogeneous processing and asynchronous execution across multiple types of Intel® processors, using computer vision pre-optimized libraries and code samples validated on 100+ trained models. For example, developers can use the OpenVINO toolkit to streamline deep learning inference and deployment by designing a single algorithm and quickly porting it across many Intel-based hardware devices.

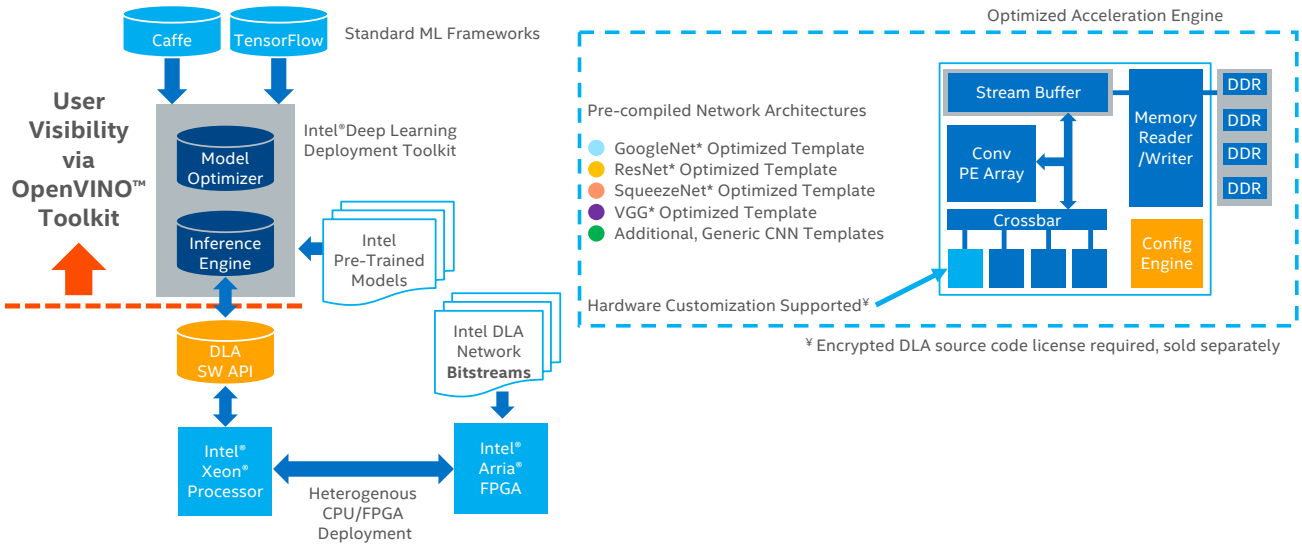


Figure 7. Intel® FPGA Deep Learning Acceleration Suite

The DLA optimized acceleration engine includes two circular paths—a smaller inner circle and a larger outer circle—that link several key components, including memory, a stream buffer, a crossbar (XBAR), a convolution PE array, and a computer engine array. In the smaller inner circle, data is continuously pulled from memory, moved to the stream buffer, and then sent through the convolution layer and other layers attached to the crossbar (such as activation or normalization layers). The data then circles back to the stream buffer to begin the next layer of computation, and the process continues to follow this circular course. When there is a new layer that needs to be supported, developers can use the crossbar and customized IP to add new features to the FPGA. The primary function of the convolution PE array and computer engine array is to accelerate the convolution compute.

The larger outer circle comes into play when there is new data coming into the DLA IP from the DDR (double data rate) channel, and the image is too large to fit into the stream buffer and follow the usual path. In that case, data is sliced into smaller pieces one at a time. Each slice is loaded into the stream buffer and processed until all crossbar operations are complete. The results are then written back to DDR memory and the next slice loaded to the stream buffer for processing. This feature means that you can input extremely large images into the Intel Vision Accelerator Design with Intel Arria 10 FPGA and never need to resize your images before sending the data to an accelerator.

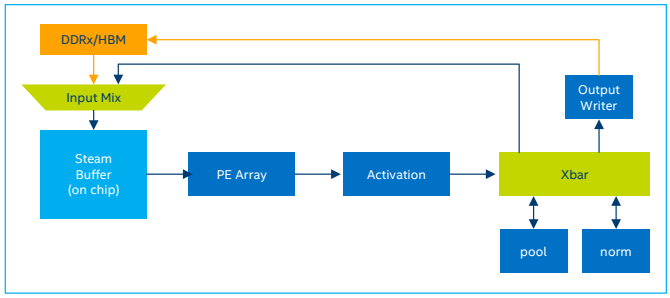


Figure 8. DLA Customization

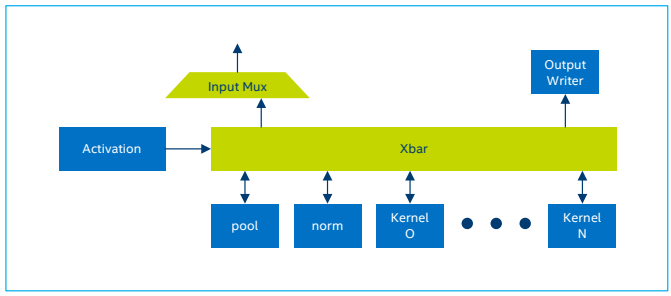


Figure 9. Auxiliary Kernels

The Intel Vision Accelerator Design with Intel Arria 10 FPGA is highly flexible and can be easily customized, allowing users to add new auxiliary kernels to the XBAR to fit their model's requirements. The XBAR is the computer bus you can use to connect, customize, and configure the layers to the DLA IP. This allows you to add more layers and achieve higher acceleration.

The flexibility of the Intel Vision Accelerator Design with Intel Arria 10 FPGA means that you are never locked down by the hardware. You can use the same board hardware for many years and add new acceleration features to the FPGA over time. Use the acceleration kernels needed to handle a specific workload, and adjust them as workload requirements change.

Table 2. Supported Primitives

HARDWARE ACCELERATED	Concat Conv Eltwise Add InnerProduct (Fully Connected Layer) Norm Pooling ReLU	Leaky ReLU PReLU Deconv (without slicing) Scale-shift Avg-Pool Max-Pool Depthwise Separable Convolutions
UPON REQUEST	Sigmoid	Tanh
FUTURE	ROI proporal	ROI pooling

Table 3. Supported Topology Types

– Out of box support for standard topologies

SUPPORTED	FUTURE
AlexNet GoogLeNet v1, v2¥, v3¥, v4¥ ResNet 18, 50, 101, 152¥ SqueezeNet v1.0¥, v1.1 MobileNet v1, v2 VGG-16, 19¥ TinyYolo v1 Yolo v2¥ SSD300, SSD512¥	LSTM/RNN FRCNN RemNet SqueezeNext and more to come in 2019

¥ Supported with good performance but not purposely optimized

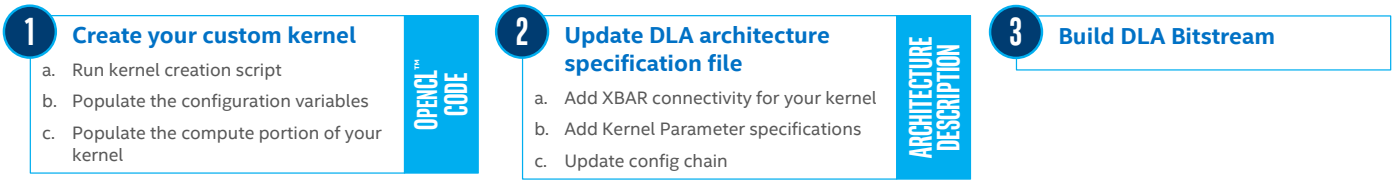


Figure 10. DLA Customization Steps

5. Where to Buy & How to Get Started

When you're ready to start creating flexible, high-performance, cost-effective deep-learning and computer-vision solutions, Intel makes it easy. You can purchase a branded version of the Intel Vision Accelerator Design with Intel Arria 10 FPGA from Intel partner IEI Integration Corp., and the OpenVINO toolkit is available as a free download from Intel. Check out the details below.

ieiworld.com

Get Started

Download Free OPENVINO™ toolkit

Get started quickly with:

- Developer resources
- Intel Tech.Decoded online webinars, tool how-tos and quick tips
- Hands-on in-person events

Support

Connect with Intel engineers & computer vision experts at the public Community Forum: **https://software.intel.com/en-us/blogs/2018/05/16/kits-to-accelerate-your-computer-vision-deployments**

or **https://software.intel.com/en-us/iot/hardware/iei-tank-dev-kit/get-started**

[1] Tests measure 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/benchmarks.

Printed in USA 1018/DW/MIM/PDF ♻ Please Recycle 338274-001

Appendix I

Choosing a Deep Learning Solution from Intel

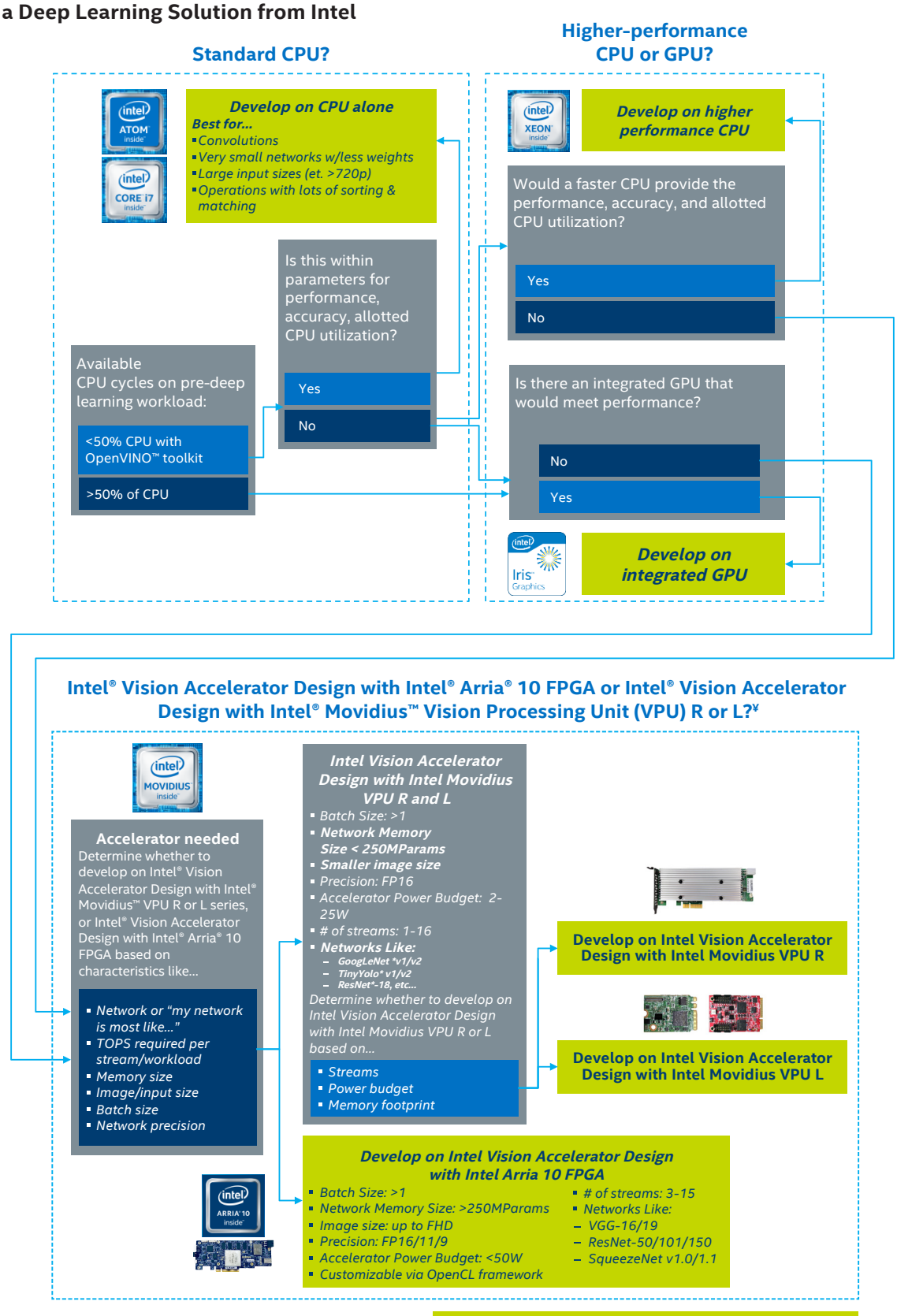


Figure 11. How to choose the right hardware based on the characteristics of your workloads

Appendix II

Figure 12. TANK-870 Industrial PC for use with Intel® Vision Accelerator Design Products

Specifications

CHASSIS	**Dimensions (WxHxD) (mm)**	121.5 x 255.2 x 205 mm (4.7” x 10” x 8”)
	Weight (Net/Gross)	4.2 kg (9.26 lbs)/ 6.3 kg (13.89 lbs)
MOTHERBOARD	**CPU**	– 6th Generation – Intel® Core™ i5-6500TE processor – (2.3 GHz quad core, 35 W thermal design power) – Intel® Core™ i7-6700TE processor – (2.4 GHz, quad core, 35 W thermal design power)
	Chipset	Intel® Q170
	System Memory	– 2 x 260-pin DDR4 SO-DIMM – 8 GB pre-installed (system max: 32GB)
I/O INTERFACES	**USB 3.0**	4
	USB 2.0	4
	Ethernet	– 2 x RJ-45 – LAN1: Intel® Ethernet Connection I219 – LAN2: (iRIS): Intel® Ethernet Controller I210
	COM Port	– 4 x RS-232 (2 x RJ-45, 2 x DB-9 w/2.5KV isolation protection) – 2 x RS-232/422/485 (DB-9)
	Digital I/O	8-bit digital I/O, 4-bit input / 4-bit output
	Display	– 1 x VGA – 1 x HDMI/DP – 1 x iDP (optional)
	Resolution	– VGA: Up to 1920 x 1200@60Hz – HDMI/DP: Up to 4096×2304@24Hz / 4096×2304@60Hz
	Audio	1 x Line-out, 1 x Mic-in
	TPM	1x Infineon TPM 2.0 Module
RELIABILITY	**Mounting**	Wall Mount
	Operating Temperature	– i7-6700TE -20°C ~ 45°C with air flow (SSD), – 10% ~ 95%, non-condensing – i5-6500TE -20°C ~ 60°C with air flow (SSD), – 10% ~ 95%, non-condensing
	Operating Vibration	MIL-STD-810G 514.6 C-1 (with SSD)
	Safety/EMC	CE/FCC/RoHS
OS	**Supported OS**	– Microsoft® Windows® 8 Embedded, Microsoft® Windows® Embedded Standard 7 E, – Microsoft® Windows® 10 IoT Enterprise – Linux Ubuntu 16.04