

Boost Performance of Video Analytics with the Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

Build high-performance computer vision applications with integrated deep learning inference

Overview

Artificial intelligence (AI) is driving the next big wave of computing, transforming the way businesses operate and changing how people engage in every aspect of their lives. Machine learning and deep learning are components of AI that use data to train models and build inference engines. These engines apply trained models for data classification, identification, and detection. Low-latency solutions allow the inference engine to process data faster, increasing overall system response times for real-time processing.

Computer vision is fast becoming an integral source of key AI data and this, in turn, is enabling new Internet of Things (IoT) applications and use cases from smart cities to retail. In order to meet the needs of real-time analytics and cost-efficient processing, much of the resulting data processing must be accomplished at the “edge”—at the edge appliance or on-premise server—as well as at the cloud or data center. Intel® Vision Accelerator Designs bring a new family of “blueprints” for specific inference accelerator cards primarily aimed at AI computer vision solutions and deep neural network inference from the edge to the cloud.

Intel is driving an open software standard which provides a common framework for developing AI computer vision applications and identifying the appropriate silicon technology.

The Intel Vision Accelerator Design with Intel® Arria® 10 FPGA offers developers and solution providers acceleration and flexibility to speed time to results and to market. It is ideal for creating applications supporting edge video analytics, as well as many IoT use cases.

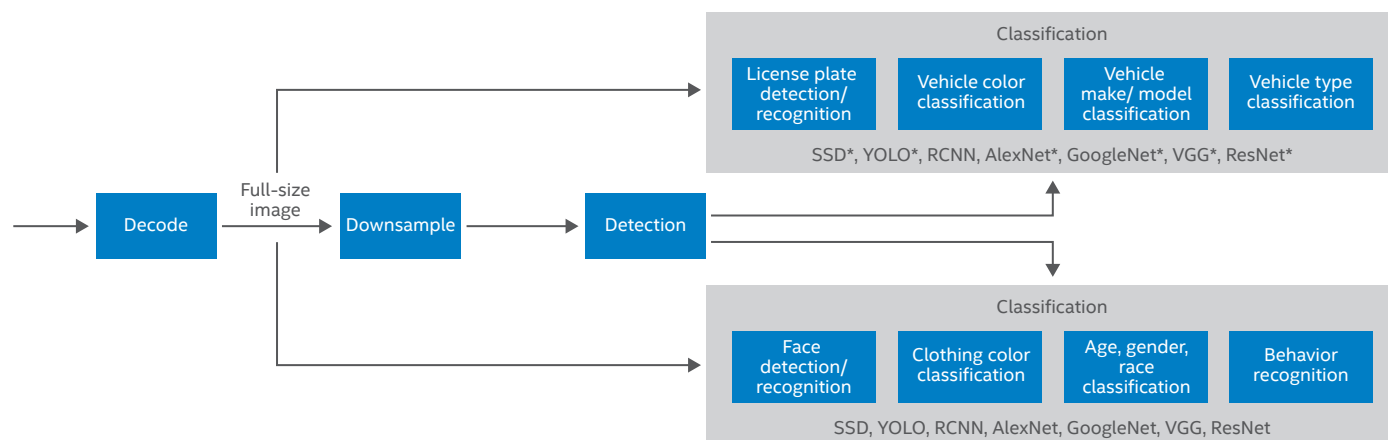


Figure 1. The Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA supports deep neural network inference for AI use cases



AI computer vision at the edge (edge appliance or on-premise server)

- The power of FPGA acceleration in a high-volume-ready PCIe* card form factor, providing an easy, modular approach for application development.
- System-level optimization—customizable datapath and inference precision create energy-efficient dataflow.
- Optimized for low-latency systems.
- Fine-grained parallelism enabling high throughput on low-batch workloads.
- Increase productivity and enable fast code reuse across Intel Atom® processors and Intel® Core™ processors.



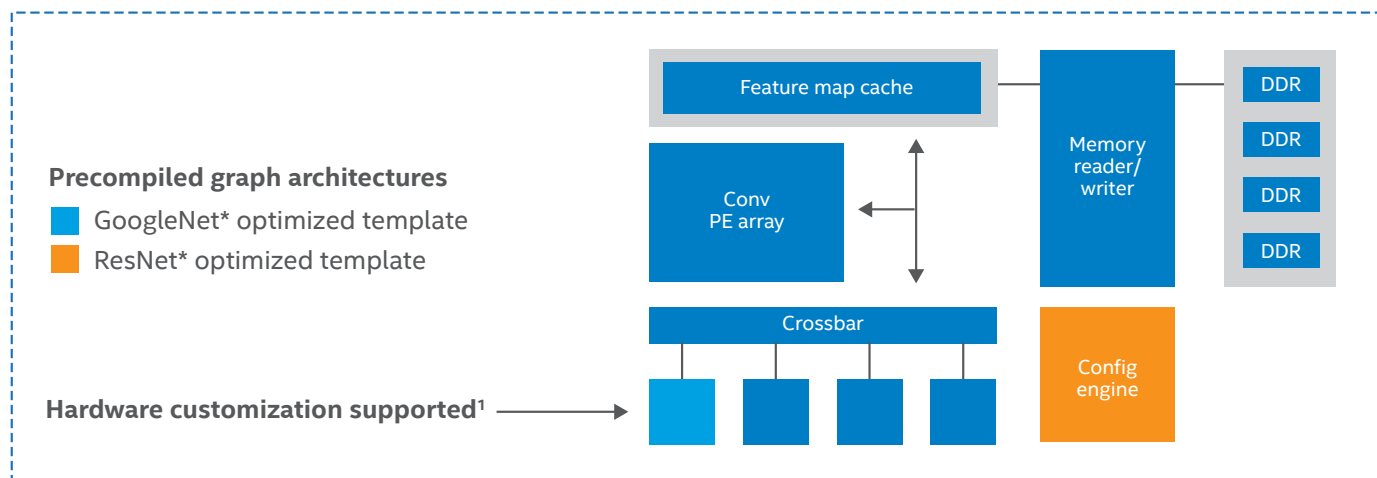
AI computer vision at the cloud

- API and drivers allow software developers to program at a higher level of abstraction for Intel® Xeon® Scalable processors with FPGA platforms.
- Intel® FPGAs combine with Intel Xeon Scalable processors to provide low-latency implementation with a small footprint.
- Increase productivity and enable fast code reuse across Intel Xeon Scalable processors.

Intel Vision Accelerator Design with Intel Arria 10 FPGA benefits include:

- **Optimize vision analytics at the edge:** Assemble vision analytics systems for a wide variety of end uses on-premise and locally.
- **Efficient video acceleration:** Analyze multiple video streams at the edge cost-effectively (high video frames/W/\$), where temperature-controlled environments are not always possible.
- **Reduce TCO:** Develop solutions that consume less power, while operating reliably over extended periods of time in many thermal environments.
- **Deliver multifunction solutions:** FPGAs are designed to handle multiple functions. With the Intel Vision Accelerator Design, developers have simplified path to market, with immediate access to functionality, as well as the option to code and customize functions.
- **Speed video decode and encryption:** Accelerate real-time decode and encode with fast, simultaneous processing across multiple camera streams. With Intel® FPGAs, soft IP provides real-time performance and the number of dedicated decoders can scale up or down for greater efficiency.

Optimized acceleration engine



Encrypted DLA source code license required; sold separately

Figure 2. Support for a growing range of topologies and algorithms gives developers more choice

Intel Vision Accelerator Design with Intel Arria 10 FPGA

Programmable, software-defined Intel FPGAs offer exceptional performance and flexibility, along with low power consumption for deep learning and computer vision solutions, whether at the edge appliance, on-premise server, or cloud. Using Intel FPGAs can help ensure continual performance optimization—taking advantage of bitstream updates—without necessitating hardware upgrades. They are designed for long product lifetimes and can handle harsh and/or outdoor environments.

The solution works seamlessly with the OpenVINO™ toolkit, offering a simplified path for developers to run customized topologies that can be optimized for heterogeneous hardware platforms. Intel FPGAs are designed to handle data without the hassle of operating systems and applications. Fine-grained parallelism and high-throughput capability are designed into the hardware architecture, allowing for extremely low-batch latency applicable to a multitude of workloads. The extremely high, fine-grained, on-chip memory bandwidth can more efficiently solve memory challenges.

The FPGA's soft IP provides performance for decoding and encryption in real time—something that is not possible with software running on a GPU or CPU—to achieve high-performance images per second at reduced power, and provide dynamic flexibility, consistent power consumption, future-proofing for custom or new workloads, and low latency.

Unlike fixed function devices, functionality in Intel FPGAs can always be changed or modified to increase or deepen intelligence. This allows them to be architected to solve very specific problems. And when a problem is consuming a lot

of bandwidth and performance, this accelerates the entire system. The FPGAs support a wide spectrum of vision use cases and applications. Intel has already customized the FPGA into a performant accelerator for convolutional neural network (CNN) workloads, so developers do not have to.

Intel FPGA cards can support more than 40 channels of video, along with vision use cases such as facial and vehicle detection, license plate recognition (LPR), estimation of head position for automotive, facial video analytics, and person and vehicle video analytics. Supported network topologies include GoogleNet*, ResNet*, SqueezeNet*, VGG-16*, and MobileNet*. Customizable datapaths and inference precision create energy-efficient data flow for system-level optimization.

The Intel Vision Accelerator-Design with Intel Arria 10 FPGA is ideal for complex or large workloads and customized applications; use cases where you may want to add your own primitives or sub-layer configurations; and workloads not supported by Intel® Movidius™ Myriad™ X VPU.

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 the Intel Vision Accelerator Design with Intel Arria 10 FPGA invaluable for complex, massive deep learning analysis and visual intelligence.

Intel Vision Accelerator Design with Intel Arria 10 FPGA supports a rapidly growing set of common topologies and algorithms for AI computer vision solutions. Developers can experiment and easily add algorithms and topologies.

TOPOLOGIES
GoogleNet*
ResNet-18*
ResNet-50*
ResNet-101*
SqueezeNet*
SqueezeNext*
VGG-16*
DenseNet*
MobileNet*
Tiny YOLO*
SSD300*
SSD512*

ALGORITHMS	
Face detection	Face selection
Face recognition	Hand tracking
Facial attribute classification	Stereo matching
People detection	Camera pose
People tracking	3D reconstruction
People attributes	Person re-identification
Age recognition	Visual SLAM
Gender recognition	Change detection
Object detection	Multi-camera tracking
Object tracking	Sensor fusion
Object recognition	Optical character/word recognition
Multi-target tracking	Behavior detection
Body detection	Gesture recognition
Body recognition	Activity recognition
	Abandoned object recognition

Sample use cases

Retail



Face and object recognition are used to identify demographics, such as age and gender, time spent in the store, and repeat and VIP customers. Customer behavior data is used to adjust merchandise mix and refine heat maps; detect merchandise for stocking and inventory; help prevent shrinkage; and improve flow of customers through store.

Traffic management



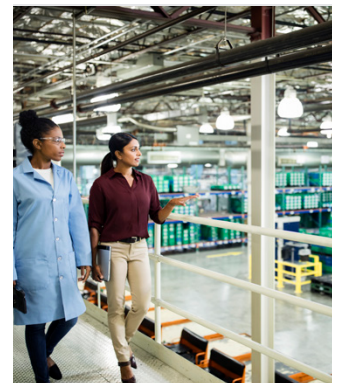
License plate recognition (LPR), identification of persons, gender, and age, and detection of behavior are utilized for a wide range of purposes from traffic pattern analysis to law enforcement.

Smart cities



Urban environments are using video analytics from multiple camera streams to detect and track people in real time for increased public safety. Anomaly detection indicates if citizens are in trouble, and require medical or emergency responder assistance.

Industry 4.0



Manufacturing facilities are using multiple cameras for visual inspection of factory operation and/or resources.

Maximize performance. Minimize development time.

Intel Vision Accelerator Design with Intel Arria 10 FPGA combines with the OpenVINO toolkit to fast-track the development of high-performance computer vision and deep learning applications. The OpenVINO toolkit is designed to increase performance and reduce development

time for computer vision solutions. It simplifies access to the benefits from the rich set of hardware options available from Intel, which can increase performance, reduce power, and maximize hardware utilization—letting you do more with less and opening new design possibilities.

OpenVINO toolkit

The OpenVINO toolkit 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.

Increase deep learning workload performance up to 19.9X on public models using the OpenVINO toolkit and Intel® architecture.¹

- **Accelerate performance:** Access Intel computer vision accelerators and speed code performance. Supports heterogeneous processing and asynchronous execution.
- **Integrate deep learning:** Unleash convolutional neural network (CNN)–based deep learning inference using a common API and more than 40 pretrained models.
- **Speed development:** Reduce development time with a library of optimized OpenCV and OpenVX functions and premade samples, and leverage common algorithms.
- **Write once:** Develop once and deploy for current and future Intel architecture-based devices.
- **Innovate and customize:** Use the increasing repository of points in OpenCV to add your unique code.

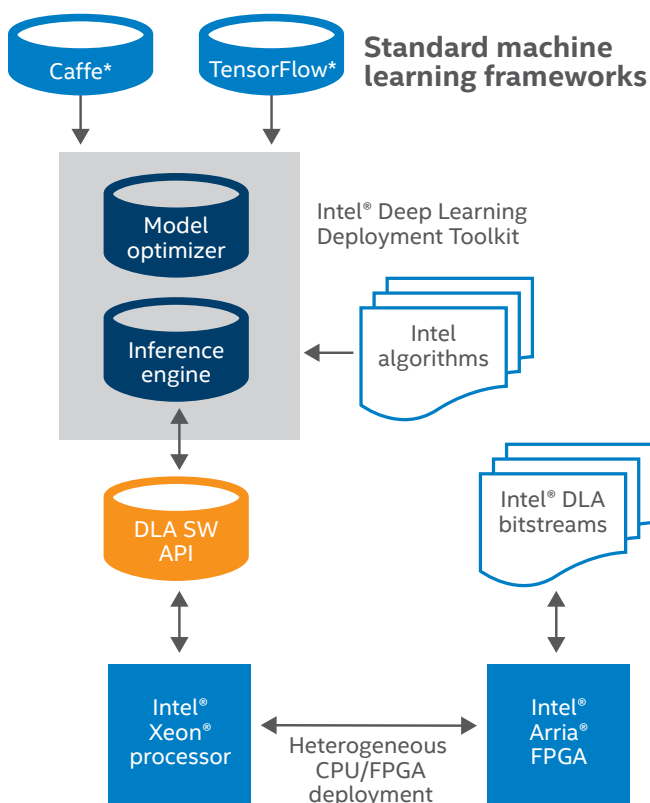


Figure 3. The OpenVINO™ toolkit allows developers to further optimize performance of deep learning inference

Intel® Deep Learning Deployment Toolkit

The OpenVINO toolkit includes the Intel Deep Learning Deployment Toolkit. Inputting trained models from standard frameworks, such as Caffe*, TensorFlow*, and MXNet*, the Model Optimizer converts these models into a unified intermediate representation (IR) file that is then run on the inference engine. Loading Caffe or other frameworks is not required if using standard layers or user-provided custom layers. The API for the inference engine has abstracted the hardware and is common across all hardware types. This allows for testing across different accelerators without necessitating recoding. The inference engine also allows heterogeneity by providing fallback from custom layers on an FPGA to the CPU.

In essence, the Model Optimizer optimizes for performance and space with conservative topology transformations. The inference engine interface is implemented as dynamically loaded plugins for each hardware type, thereby delivering optimal performance for each type, without requiring you to implement and maintain multiple code pathways.

Accelerate deep learning inference performance

- Model Optimizer converts Caffe, TensorFlow, and MXNet to IR files
- Inference engine with plugins for CPU, GPU, FPGA, and VPU

OpenCV

- Precompiled OpenCV 3.3 with Intel® CPU optimizations
- Intel® Photography Vision Library, with face detection/recognition, blink detection, and smile detection

OpenVX

- Graph-based implementation of a short list of traditional computer vision operations and CNN primitives
- Khronos OpenVX* Neural Networks Extension 1.2
- Visual Algorithm Designer (VAD)
- Eclipse* plugin enabling integrated OpenVX application development

OpenCL™

- Drivers and runtimes, and media drivers are included to simplify working with Intel® Media SDK and Intel® FPGA SDK for OpenCL™ applications for computer vision.

Models

- More than 40 pretrained models and samples are included in the package, in addition to a model downloader that will also download public models.

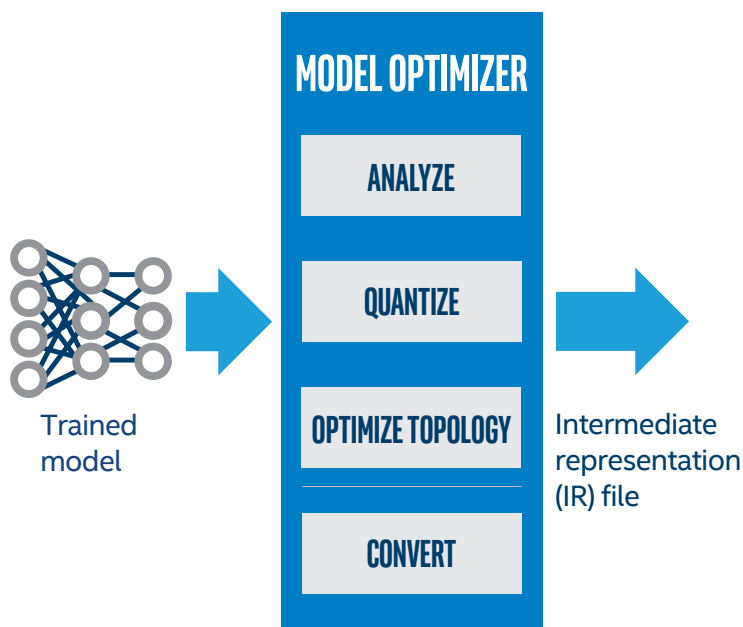


Figure 4. The Model Optimizer converts deep learning frameworks such as Caffe*, TensorFlow*, and MXNet* to intermediate representation (IR) files for efficient processing

Key specifications

Intel Arria 10 FPGAs for Deep Learning

- Intel Arria 10 1150GX PCIe add-in card with OpenVINO™ toolkit support and bitstream update tool
- Provides various inference precision options to support different performance targets using the same hardware
- Bitstream updates on a regular cadence for performance enhancements over time

Hard floating point, arbitrary precision DSP blocks

- Up to 1.5 TFLOPS on Intel Arria 10 FPGAs and 10 TFLOPS on Intel® Stratix® 10 FPGAs
- Arbitrary precision data types (FP16 => FP9) offer 2 TOPS to 20 TOPS on Intel Arria 10 (26 TOPS at maximum)

Distributed, fine-grain DSP, memory, and logic

- Orders of magnitude higher on-chip bandwidth compared to GPUs
- Reduced data movement enables power efficiency
- Deterministic low latency

Core and I/O programmability enables flexibility

- Arbitrary deep learning architectures
- Versatility in I/O configurations and use models
- Multifunction acceleration

Ecosystem solutions

A growing list of ecosystem partners offer application-tailored solutions based on the Intel® Programmable Acceleration Card (Intel® PAC) and Intel Vision Accelerator Design with Intel Arria 10 FPGA.

IEI Video Analytics Accelerator Card*

- Designed for high-performance, low-latency applications
- Supports multiple network topologies (GoogleNet, ResNet, etc.)
- OpenCL BSP for accelerator customization
- Intel Arria 10 GX 1150KLE FPGA
- Two memory banks (DDR4 at 4 GB each)
- Passive or active cooling
- ½ L, ½ H PCIe (Gen3 x8, x8 mechanical)

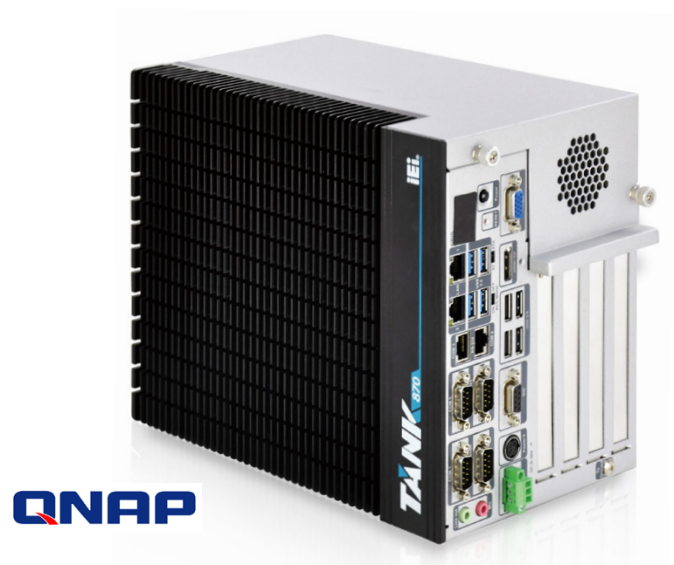


Figure 5. IEI and QNAP offer solutions based on the Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA

The Intel Vision Accelerator Design portfolio

Intel delivers more choice of hardware acceleration for deep learning inference on the edge with the new family of hardware acceleration boards. The Intel Vision Accelerator Design is also available with the Intel Movidius Myriad X VPU.

You can use either the Movidius or FPGA solution or both depending on your application and innovation needs. With the Intel Vision Accelerator Design products, Intel is helping to remove barriers to entry from silicon to software stack.

	INTEL® VISION ACCELERATOR DESIGN WITH INTEL® MOVIDIUS VPU	INTEL® VISION ACCELERATOR DESIGN WITH INTEL® ARRIA® 10 FPGA
	<i>Specialized processors designed to deliver high-performance machine vision and awareness at ultra-low power.</i>	<i>Flexible, customizable processors designed to adapt to advanced display, video, and image processing workloads.</i>
Select when you need	High compute and efficiency, low power consumption, and excellent computer vision and deep learning performance/W/\$.	- Dynamic flexibility and high performance on a single chip, consistent power consumption, future-proofing for custom or new workloads, and low latency. - Look for additional deep learning optimization features like compute-intensive networks (VGG*, ResNet-101*), and new network topologies.
Use cases	Excels in camera and video appliance use cases that have power, size, and/or cost constraints, and mainstream topologies that can be optimized into an ASIC.	Excels in video appliance and server use cases that benefit from soft IP for nearly instantaneous functional modification combined with high performance.
Configuration	For cameras and other power/size-constrained systems.	For on-premise video appliances and video analytics servers.
Supported streams	Typically supports 2 video streams per VPU.	Aggregation of 3 to 32 video streams per device
Batch size	Batch sizes of 1+	Batch sizes of 1+
Power consumption	Per VPU power consumption typically 2W to 3W. (Designs ranging from ~4W to 25W).	More consistent power consumption (typically, ~42W for Intel® Arria® 10 1150 device).
Efficiency	High efficiency (inferences/sec/W).	Good efficiency (inferences/sec/W) vs. GPGPU.
Network memory	<250M parameters	>250M parameters
Precision	Supports fp16 precision networks.	Supports lower precision networks (e.g., fp16/11/9).
Customization	Hardware-optimized for generic cases.	Custom hardware architecture for specific topologies.
Partitioning	Requires partition across multiple devices for large networks (e.g., YOLO* v1).	No need to partition across multiple hardware devices.

Conclusion

The Intel Vision Accelerator Design with Intel Arria 10 FPGA offers a cost-effective and efficient platform on which to develop software applications for AI computer vision solutions. It brings the performance and flexibility of Intel FPGAs, and works with the OpenVINO toolkit for optimized CNN-based deep learning inference. The Intel Vision Accelerator Design with Intel Arria 10 FPGA is one of the Intel solutions designed to speed and improve a breadth of edge analytics applications. Experience the power of a single open software environment that supports key silicon technologies, and easily adopt the technology that best suits your requirements.

With Intel, developers and solution providers can achieve better value through lower TCO while seizing the AI-based video analytics opportunity.

Learn more

Explore Intel® Vision products at intel.com/visionproducts.

Discover Intel FPGAs at intel.com/fpga.

Find out more about Intel innovation for AI at intel.com/ai.

Download the free [OpenVINO toolkit](#).

Contact IEI to place orders and get support at sales@usa.ieiworld.com.

Development tip

For distributed inference at the gateway, edge, or endpoint, start with an Intel® CPU and add targeted FPGA acceleration for higher throughput and/or throughput per watt. Integrated Intel® processor graphics are a general way to further boost throughput. The FPGA allows for a lot of platform customization, including I/O, multistream aggregation, in-line processing, and a mix of deep learning traditional sensor processing acceleration.



1. Performance increase comparing certain standard framework models vs. Intel-optimized models in the Intel® Deep Learning Deployment Toolkit.

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.

Estimated results reported above 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.

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 information about benchmarks and performance test results, go to www.intel.com/benchmarks.

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 intel.com/visionproducts.

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.

Intel, the Intel logo, Intel Atom, Intel Core, Intel Movidius, Intel Movidius Myriad X, Intel Optane, Arria, OpenVINO, Stratix, and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

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

© Intel Corporation

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

1018/BH/CMD/PDF

338179-001US