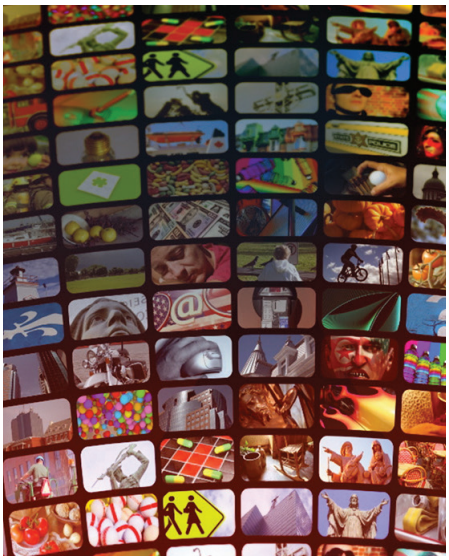
SOLUTION BRIEF

Intel® OpenVINO™ with FPGA Support Through the Intel FPGA Deep Learning Acceleration Suite

Intel® FPGA Deep Learning Acceleration Suite enables Intel FPGAs for accelerated AI optimized for performance, power, and cost.



Introduction

Artificial intelligence (AI) is driving the next big wave of computing, transforming both the way businesses operate and how people engage in every aspect of their lives. Intel® FPGAs offer a hardware solution that is capable of handling extremely challenging deep learning models at unprecedented levels of performance and flexibility. The Intel OpenVINO™ toolkit is created for the development of applications and solutions that emulate human vision, and it provides support for FPGAs through the Intel FPGA Deep Learning Acceleration Suite. The Intel FPGA Deep Learning Acceleration Suite is designed to simplify the adoption of Intel FPGAs for inference workloads by optimizing the widely used Caffe* and TensorFlow* frameworks to be applied for various applications, including image classification, computer vision, autonomous vehicles, military, and medical diagnostics.

Intel FPGAs offer customizable performance, customizable power, deterministic low latency, and flexibility for today's most widely adopted topologies as well as programmability to handle emerging topologies. Unique flexibility, for today and the future, stems the ability of Intel FPGAs to support emerging algorithms by enabling new numeric formats quickly. What makes FPGAs unique is its ability to achieve high performance through parallelism coupled with the flexibility of hardware customization – which is not available on CPU, GPU, or ASIC architectures.

Turnkey Solution Offers Faster Time to Market

The Intel Programmable Acceleration Card with Intel Arria® 10 GX FPGA (Intel PAC with Intel Arria 10 GX FPGA) combined with the Intel FPGA Deep Learning Acceleration Suite offer an acceleration solution for real-time AI inference with low latency as a key performance indicator. By featuring industry-leading FPGAs that meet Intel's stringent production quality standards while offering familiar support for common AI frameworks, this FPGA-based solution simplifies the process of deploying new hardware into an existing infrastructure. This is done without requiring specialized accelerator knowledge, FPGAs or others, to utilize. The solution provides a major leap for AI developers by offering a unique combination of energy-efficient inference, scalable throughput, variable-precision support including non-power-of-two floating-point sizes, and low latency to enable fast responses.

FPGAs for Deep Learning

Intel FPGAs are the "Accelerators of Choice" for AI because they allow deep learning algorithms to be implemented in custom hardware that is future proof and tailored to the specific algorithm versus an "off-the-shelf" multi-core approach. Custom hardware allows FPGAs to support the massive data bandwidth required by AI applications with data flow efficiency. These devices offer high throughput with deterministic, low latency and excellent power efficiency. The reconfigurable logic of FPGAs also provides a future-proof platform that can be updated to support AI architecture advancements, including support for arbitrary precision data types, new sparsity and weight sharing schemes, and reconfigurable I/O for inline and offload processing.

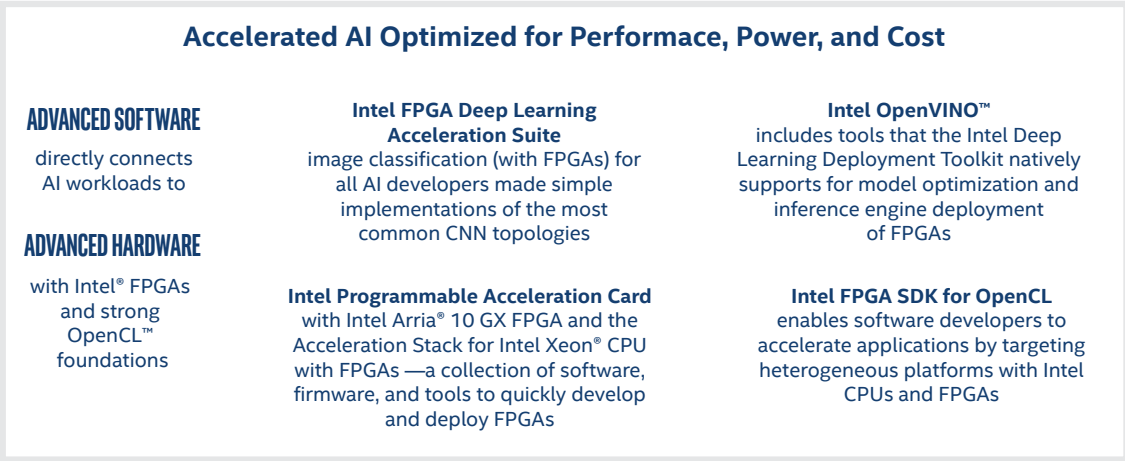


Figure 1. Comprehensive Solution for Low-Latency Real-Time AI Inference

Design Challenge

FPGA design has historically required an RTL design flow performed by experienced hardware designers, putting it out of reach of a vast community of software and application developers. The challenge is to bring heterogeneous programming through a high-level design environment, such as OpenCL™, to be used with application-specific frameworks including Caffe and TensorFlow. This challenge has been met by Intel's comprehensive Acceleration Stack for Intel Xeon® CPU with FPGAs, in particular the Intel FPGA Deep Learning Acceleration Suite, and the Intel OpenVINO toolkit.

Intel Deep Learning Deployment Toolkit

The Intel Deep Learning Deployment Toolkit is built for data center applications for real-time or offline video/image recognition and classification. It accelerates computationally intensive convolutional neural network (CNN) primitives optimized for FPGA hardware. The development and deployment of AI applications are supported through the Intel Deep Learning Deployment Toolkit that supports industry-standard machine learning frameworks including Caffe and TensorFlow.

The Intel Deep Learning Deployment Toolkit is a fully validated, AI framework that makes it easy for developers to target FPGAs into their real-time AI applications with a software-centric flow. This framework includes a runtime deployable deep learning acceleration engine that is optimized for real-time performance and low latency for the Intel PAC with Intel Arria 10 GX FPGA. The Intel FPGA Deep Learning Acceleration Suite supports today's most popular deep learning architectures including AlexNet, GoogleNet, SqueezeNet, VGG16, and ResNet for a variety of numeric precisions.

The Intel FPGA Deep Learning Acceleration Suite includes pre-programmed intellectual property (IP) cores for FPGAs which accelerate the following six CNN primitives, critical for high throughput and low-latency inference:

- conv – convolution
- fc – fully connected
- relu – rectified linear unit
- pool – pooling (maximum and average)
- norm – local response normalization (LRN)
- concat – concatenation

These primitives are enabled through Caffe and the Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) which provide unified deep learning application programming interfaces (APIs). The whole software stack is optimized for performance. Reconfiguration to the primitives are made at the upper layers of the stack, through Caffe and the Intel MKL-DNN, abstracting the low-level FPGA programming complexity. Network topologies that use primitives beyond these six are configured on the host and computed with a sophisticated usage of both the CPU and FPGA. Validated network topologies include AlexNet, GoogleNet, SqueezeNet, VGG-16, CaffeNet, and LeNet.

All software layers of the stack are packed into an installer, which installs all components with a few clicks to greatly facilitate installation and system setup from the user end. Once installed, OpenCL (BSP and runtime) enables the communication with the hardware from the host side. The six primitives in the FPGA are enabled through the Intel MKL-DNN, which is designed to provide a unified deep learning API for Intel devices with optimized performance. With the integration of the Intel MKL-DNN to Caffe, users can build deep learning applications through the Intel FPGA Deep Learning Acceleration Suite using the Caffe framework or directly using the Intel MKL-DNN primitive API.

Intel OpenVINO

Intel OpenVINO is a comprehensive toolkit for developing and deploying vision-oriented solutions on platforms from Intel. It provides an environment that spans CPUs, GPUs, and FPGAs. It contains OpenCV for vision programming to run on a CPU or a CPU-GPU chip, while using the Intel Deep Learning Deployment Toolkit to provide access to FPGA capabilities for deep neural network (DNN) programming.

Intel OpenVINO includes optimized deep learning tools for high-performance inferencing, the popular OpenCV library for computer vision applications and acceleration of machine perception, and Intel's implementation of the OpenVX* API. Included are the Deep Learning Inference Engine that features a unified (i.e. OpenVX-agnostic) API to integrate the inference with application logic, and a Deep Learning Model Optimizer tool to help with the deployment of CNNs.

Intel OpenVINO provides for segmentation, overlays, vision, and much more to be handled by the host, with programming via the Intel Deep Learning Deployment Toolkit for image classification connecting the low-latency, low-power, deterministic, and flexible capabilities of the FPGA into the workflow. The combination is powerful and easy to manage thanks to Intel OpenVINO and the Intel Deep Learning Deployment Toolkit components of the comprehensive Acceleration Stack for Intel Xeon CPU with FPGAs.

Acceleration Platforms for Intel Xeon CPU with FPGAs

The Acceleration Stack for Intel Xeon CPU with FPGAs brings together all this software into a single comprehensive support system that allows software developers to leverage the power of FPGAs much easier than before. A core component is the FPGA Interface Manager, which provides performance-optimized connectivity between an Intel FPGA and an Intel Xeon processor. The FPGA can be directly transacted on with the Intel Acceleration Engine with Open Programmable Acceleration Engine (OPAE) Technology, which provides thread-safe APIs that can be called from within virtual machines and containers. This relieves developers of crafting customized drivers and debugging interfaces, enabling them to focus on their core expertise – algorithm development – and develop their solutions faster and with greater confidence.

Together, the Acceleration Stack for Intel Xeon CPU with FPGAs and the Intel Programmable Acceleration Card (Intel PAC) enables developers to harness the power of FPGA acceleration with a comprehensive productivity suite to match the growing demands on data centers.

Intel PAC

The Intel PAC offers both inline and lookaside acceleration to simplify the use of FPGAs in servers. This card can be deployed in various servers with its low-profile form factor, low-power dissipation, and passive heat sink. OpenCL support is baked into this turnkey well-integrated solution – including the necessary bitstream for the FPGA. Its plug-and-play architecture allows for a simple insertion and configuration in only minutes. The card also contains a networking interface for accelerating workloads, streaming analytics and video transcoding, and dedicated banks of DDR4 memory with error correction. This card includes the FPGA Interface Manager, and seamlessly pairs with an Intel Xeon processor over a PCI Express* bus.

The Intel PAC provides the performance and versatility of FPGA acceleration and is one of several platforms supported by the Acceleration Stack for Intel Xeon CPU with FPGAs. This acceleration stack provides a productive developer environment as well as eases application deployment. Together with acceleration libraries and development tools, the acceleration stack saves developers time and enables code reuse across multiple Intel FPGA platforms.

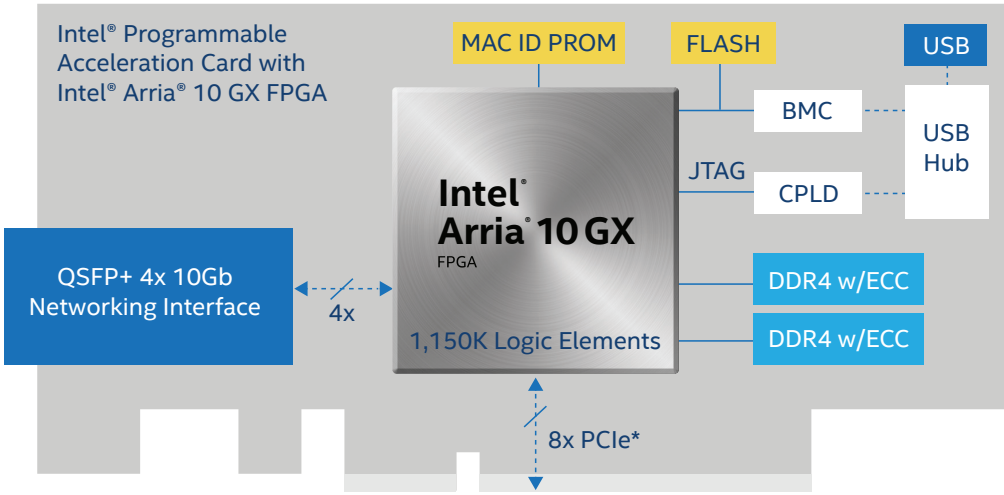


Figure 2. Intel Programmable Acceleration Card with Intel Arria 10 GX FPGA

Figure 3. Intel Programmable Acceleration Card with the Intel Arria 10 GX FPGA

Summary

Intel's technology leadership stands out in today's increasingly complex and heterogeneous computing world that demands both general-purpose and specialized solutions that can connect, interoperate, and offer longevity and high reliability. We are entering a smart and connected world, where all things are expected to be captured as a piece of data, measured in real time, and accessible from anywhere. The breadth of Intel's product portfolio allows a better fit for specific needs, and FPGAs are a critical component in engineering the accelerated solutions.

Software developers can enjoy high-quality software development tools and libraries that deliver great performance in familiar software environments, including most popular AI development platforms. Intel products are made even more exciting by Intel's proven abilities to deliver production quality and the ability to build hybrid devices including CPU-FPGA products to complement discrete devices for PCI Express acceleration cards.

Related Links

Learn More about Intel's FPGA solutions for real-time AI.

- Learn more about the Intel FPGA Deep Learning Acceleration Suite
- Learn more about the Intel OpenVINO Toolkit
- Installation Guide: Install the Intel OpenVINO on Linux, with FPGA
- Configuration Guide: Preproduction Reconfigurable Reference Design Setup Guide
- OpenVINO details: Inference Engine Developer Guide
- Intel FPGAs Power Micrsoft Project Brainwave AI
- Microsoft* Turbocharges AI with Intel FPGAs. You Can, Too
- Intel FPGAs Bring Power to Artificial Intelligence in Microsoft Azure
- Read the Intel NewsByte: Intel Eases Use of FPGA Acceleration: Combines Platforms, Software Stack and Ecosystem Solutions to Maximize Performance and Lower Data Center Costs
- Read the blog: Supercharging Data Center Performance while Lowering TCO: Versatile Application Acceleration with FPGAs
- Start designing today with the Intel Acceleration Stack for Intel Xeon CPU with FPGAs
- Learn more about the Intel Programmable Acceleration Card with Intel Arria 10 GX FPGA
- See more of the excitement in the Intel FPGA Acceleration Hub
- Read about the Intel Xeon Processor family
- New to FPGAs? Read an Overview of FPGAs
- Learn about Intel FPGAs

ss-1092-1.0