WHITE PAPER

Artificial Intelligence
Intel® FPGAs
Intel® AI Builders

Machine Learning on Intel® FPGAs

Table of Contents

Introduction

Artificial intelligence (AI) originated in classical philosophy and has been loitering in computing circles for decades. Twenty years ago, AI surged in popularity, but interest waned as technology lagged. Today, technology is catching up, and AI's resurgence far exceeds its past glimpses of popularity. This time, the compute, data sets, and technology can deliver, and Intel leads the AI pack in innovation.

Among Intel's many technologies contributing to AI's advancements, field-programmable gate arrays (FPGAs) provide unique and significant value propositions across the spectrum. Understanding the current and future capabilities of Intel® FPGAs requires a solid grasp on how AI is transforming industries in general.

AI Is Transforming Industries

Industries in all sectors benefit from AI. Three key factors contribute to today's successful resurgence of AI applications:

- Large data sets
- Recent AI research
- Hardware performance and capabilities

The combination of massive data collections, improved algorithms, and powerful processors enables today's ongoing, rapid advancements in machine learning, deep learning, and artificial intelligence overall. AI applications now touch the entire data spectrum from data centers to edge devices (including cars, phones, cameras, home and work electronics, and more), and infiltrate every segment of technology, including:

- Consumer devices
- Enterprise efficiency systems
- Healthcare, energy, retail, transportation, and others

Some of AI's largest impacts are found in self-driving vehicles, financial analytics, surveillance, smart cities, and cyber security. Figure 1 illustrates AI's sizable impact on just a few areas.

Figure 1. Examples of how AI is transforming industries.

To support AI's growth today and well into the future, Intel provides a range of AI products in its AI ecosystem. Intel® FPGAs are a key component in this ecosystem.

Intel's AI Ecosystem and Portfolio

As a technology leader, Intel offers a complete AI ecosystem that concentrates far beyond today's AI—Intel is committed to fueling the AI revolution deep into the future. It's a top priority for Intel, as demonstrated through in-house research, development, and key acquisitions. FPGAs play an important role in this commitment.

Intel's comprehensive, flexible, and performance-optimized AI portfolio of products for machine and deep learning covers the entire spectrum from hardware platforms to end user applications, as shown in Figure 2, including:

- Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN)
- Compute Library for Deep Neural Networks
- Deep Learning Accelerator Library for FPGAs
- Frameworks such as Caffe* and TensorFlow*
- Tools like the Deep Learning Deployment Toolkit from Intel

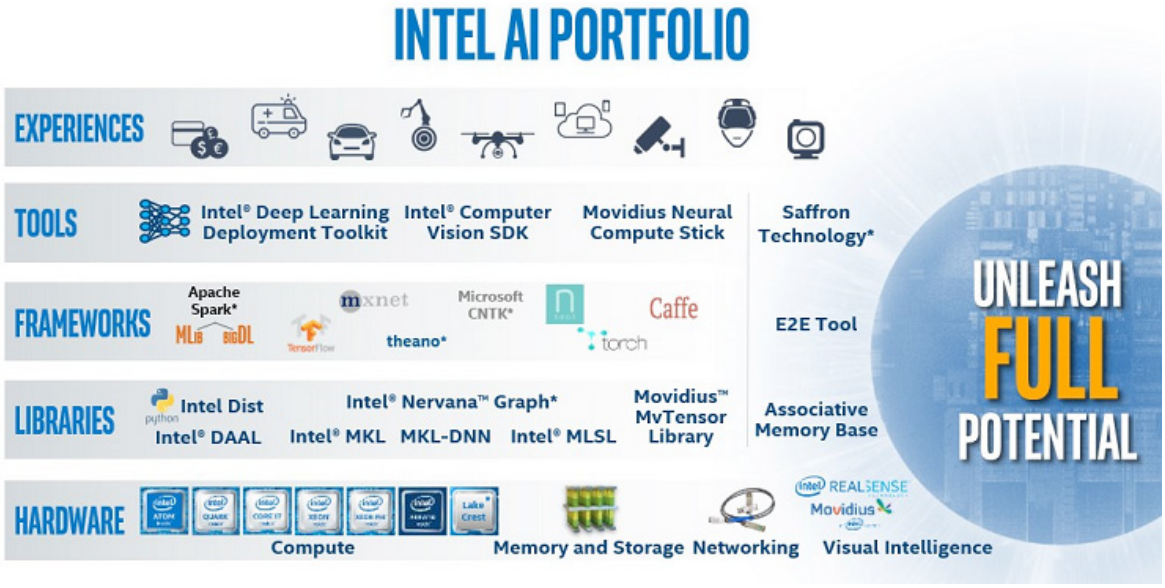

Figure 2. Intel's AI Portfolio of products for machine and deep learning.

Overall, Intel provides a unified front end for the broad variety of backend hardware platforms, enabling users to develop a system with one device today and seamlessly switch to a newer, different hardware platform tomorrow. This comprehensive

nature of the Intel's AI Ecosystem and portfolio means Intel is uniquely situated to help developers at all levels access the full capacity of Intel hardware platforms, both current and future. This approach empowers hardware and software developers to take advantage of the FPGAs' capabilities with machine learning, leading to increased productivity and shorter design cycles.

The Intel® FPGA Effect

Intel® FPGAs offer unique value propositions, and they are now enabled for Intel's AI ecosystem. Intel® FPGAs provide excellent system acceleration with deterministic low latency, power efficiency, and future proofing, as illustrated in Figure 3.

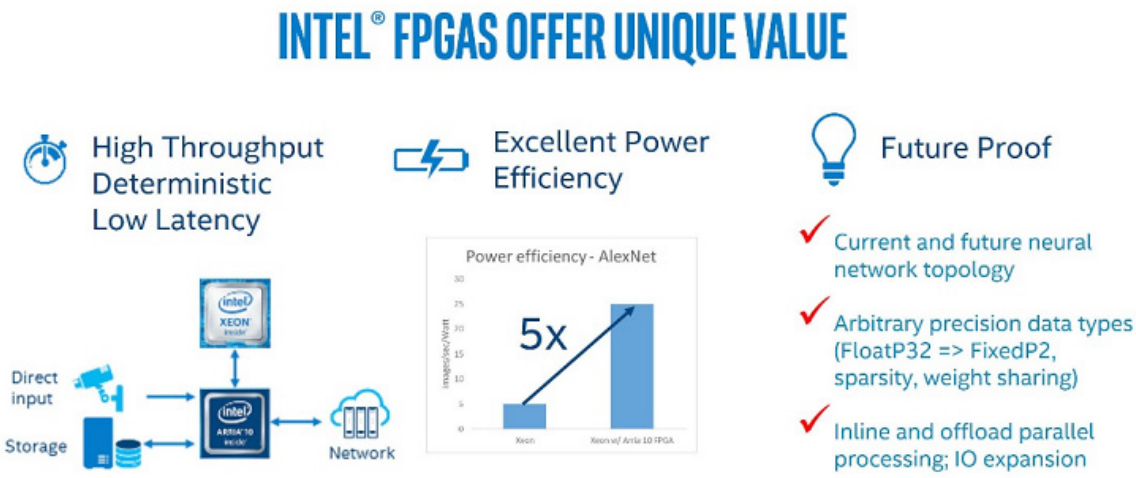

Figure 3. Intel® FPGAs offer unique value propositions for AI.

System Acceleration

Today, people are looking for ways to leverage CPU and GPU architectures to get more total operations processing out of them, which helps with compute performance. FPGAs are concerned with system performance. Intel® FPGAs accelerate and aid the compute and connectivity required to collect and process the massive quantities of information around us by controlling the data path. In addition to FPGAs being used as compute offload, they can also directly receive data and process it inline without going through the host system. This frees the processor to manage other system events and provide higher real time system performance.

Real time is key. AI often relies on real-time processing to draw instantaneous conclusions and respond accurately. Imagine a self-driving car waiting for feedback after another car breaks hard or a deer leaps from the bushes. Immediacy has been a challenge given the amount of data involved, and lag can mean the difference between responding to an event and missing it entirely.

FPGAs' flexibility enables them to deliver deterministic low latency (the guaranteed upper limit on the amount of time between a message sent and received under all system conditions) and high bandwidth. This flexibility supports the creation of custom hardware for individual solutions in an optimal way. Regardless of the custom or standard data interface, topology, or precision requirement, an FPGA can implement the exact architecture defined, which allows for unique solutions and fixed data paths. This also equates to excellent power efficiency and future proofing.

Power Efficiency

FPGAs' ability to create custom solutions means they can create power-efficient solutions. They enable the creation of solutions that address specific problems, in the way each problem needs to be solved, by removing individual bottlenecks in the computation, not by pushing solutions through fixed architectures.

Intel® FPGAs have over 8 TB/s of on-die memory bandwidth. Therefore, solutions tend to keep the data on the device tightly coupled with the next compute. This minimizes the need to access external memory, which results in running at significantly lower frequencies. These lower frequencies and efficient compute implementations result in very powerful and efficient solutions. For example, FPGAs show up to an 80% power reduction when using AlexNet* (a convolutional neural network) compared to CPUs.

Future Proofing

In addition to system acceleration and power efficiency, Intel® FPGAs help future proof systems. With such a dynamic technology as machine learning, which is evolving and changing constantly, Intel® FPGAs provide the flexibility unavailable in fixed devices. As precisions drop from 32-bit to 8-bit and even binary/ternary networks, an FPGA has the flexibility to support those changes instantly. As next generation architectures and methodologies are developed, FPGAs will be there to implement them. By reprogramming an FPGA's image, its functionality can be changed completely. Dedicated ASICs can provide a higher

total cost of ownership (TCO) in the long run, and with such a dynamic technology, there is a higher and higher threshold to warrant building them, especially if FPGAs can meet a system's needs.

Some markets demand longevity and high reliability from hardware with systems being deployed for 5, 10, 15, or more years in harsh environments. For example, imagine putting smart cameras on the street or compute systems in automobiles and requiring the same 18 month refresh cycle that CPUs and GPUs expect. The FPGAs flexibility enables users to update the hardware capabilities for the system without requiring a hardware refresh. This results in longer lifespans of deployed products. FPGAs have a history of long production cycles with devices being built for well over 15 to 20 years. They have been used in space, military, and extremely high reliability environments for decades.

For these reasons and more, developers at all levels need to understand how the Intel's AI Ecosystem and portfolio employs Intel® FPGAs. This knowledge will enable developers to use Intel® FPGAs to accelerate and extend the life and efficiency of AI applications.

Increased Productivity and Shortened Design Cycles

Most developers know FPGAs are flexible and robust devices providing a wide variety of uses:

- FPGAs can become any digital circuit as long as the unit has enough logic blocks to implement that circuit.
- Their flexible platform enables custom system architectures that other devices simply cannot efficiently support.
- FPGAs can perform inline data processing, such as machine learning, from a video camera or Ethernet stream, for example, and then pass the results to a storage device or to the process for further processing. FPGAs can do this while simultaneously performing, in parallel, compute offload.

But not all developers know how to access Intel® FPGAs' potential or that they can do so with shorter-than-ever design cycles (illustrated in Figure 4).

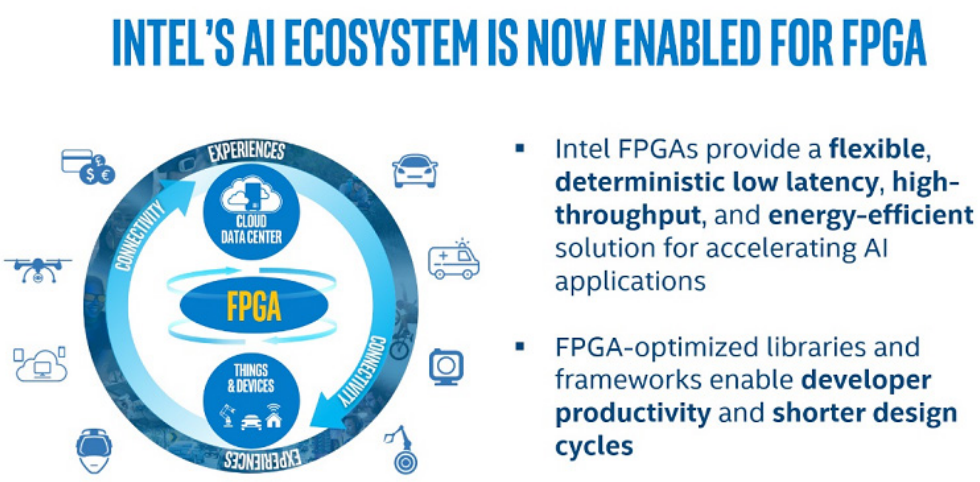

Figure 4. Intel's AI ecosystem is now enabled for FPGA.

To help developers bring FPGAs to market running machine learning workloads, Intel has shortened the design time for developers by creating a set of API layers. Developers can interface with the API layers based on their level of expertise, as outlined in Figure 5.

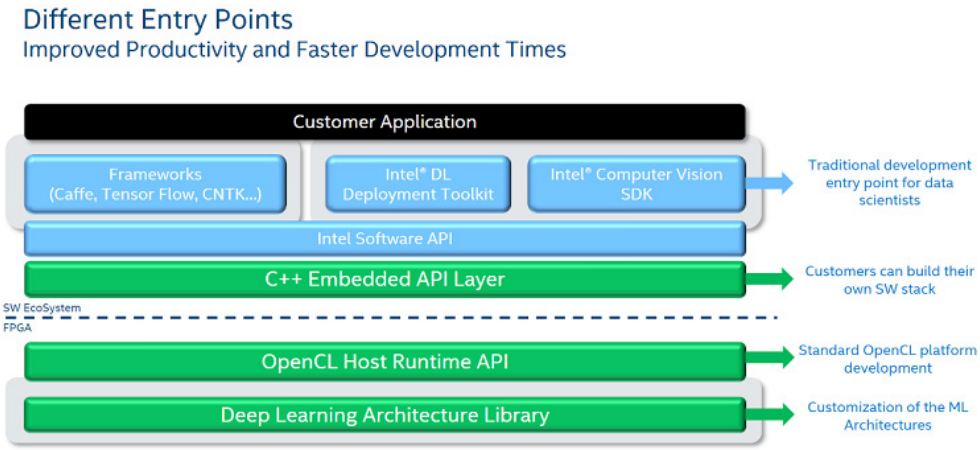

Figure 5. Four Entry Points for Developers

Typical users can start at the SDK or framework level. More advanced users, who want to build their own software stack, can enter at the Software API layer. The Software API layer abstracts away the lower-level OpenCL™ runtime and is the same API the libraries use. Customers who want to build their own software stack can enter at the C++ Embedded API Layer. Advanced platform developers who want to add more than machine learning to their FPGA—such as support for asynchronous parallel compute offload functions or modified source code—can enter in at the OpenCL™ Host Runtime API level or the Intel Deep Learning Architecture Library level, if they want to customize the machine learning library.

Several design entry methods are available for power users looking to modify source code and customize the topology by adding custom primitives. Developers can customize their solutions by using traditional RTL (Verilog or VHDL), which is common for FPGA developers, or the higher level compute languages, such as C/C++ or OpenCL™. By offering these various entry points for developers, Intel makes implementing FPGAs accessible for various skillsets in a timely manner.

Conclusion

Intel is uniquely positioned for AI development—the Intel's AI Ecosystem offers solutions for all aspects of AI by providing a unified front end for a variety of backend technologies, from hardware to edge devices. In addition, Intel's ecosystem is now fully enabled for FPGA. Intel® FPGAs provide numerous benefits, including system acceleration opportunities, power efficiency, and future proofing, due to FPGAs' long lifespans, flexibility, and re-configurability. Finally, to help propel AI today and into the future, Intel AI solutions allow a variety of language-agnostic entry points for developers at all skillset levels.