White Paper

FPGAs and SoCs
AI Suite

altera
An Intel Company

Altera® FPGAs and SoCs with FPGA AI Suite and OpenVINO Toolkit Drive Embedded/Edge AI/Machine Learning Applications

Authors

Jahanzeb Ahmad
Sr. Solutions Architect
Intel Corp.

Mark Jervis
Sr. Solutions Architect
Intel Corp.

Rama Venkata
Sr AI Technical Marketing Manager
Intel Corp.

Table of Contents

As the speed of enterprise operations surge and expectations for quick responses rise, decision making increasingly migrates from data centers to the network's edge. Whether it's on the shop floor, where busy production lines must not stand idle, a doctor in a surgical suite waiting for answers, a fire crew needing orders to fight a frantically advancing wildfire, scientists looking for insights that will help a recovering coral reef, or a retail environment where impatient customers wait for help in making a purchase, enterprises must configure their systems to collect information, to develop actionable insights, and to provide decisions or answers in real time. It seems that more and more often, only fully automated decisions are timely enough to make a difference.

So much data is now being captured at the edge that Gartner predicts that by 2025, as much as 75 percent of all enterprise data will be generated outside traditional data centers[1]. According to Santhosh Rao, senior research director at Gartner, "Organizations that have embarked on a digital business journey have realized that a more decentralized approach is required to address digital business infrastructure requirements. As the volume and velocity of data increases, so does the inefficiency of streaming all this information to a cloud or data center for processing."[1]

Moving compute capabilities with artificial intelligence (AI) and machine learning (ML) algorithms closer and closer to this data, all the way to its origin point on the edge in many cases, enables new real-time use cases that potentially create fresh revenue streams, while preventing sensitive data from flowing through the network and into data centers at the same time. Achieving real-time responses to edge data relies on effectively combing at least four technologies:

- Edge computing
- AI
- High-speed networking
- The cloud

Enterprises must integrate these technologies throughout their infrastructure to realize the full benefits of edge-to-cloud intelligence. Putting more devices and compute power with AI capabilities at the edge creates the ability to process more data, while creating even more data at the same time, which enables more complex AI use cases and, in turn, enables the development of more actionable insight.

The edge encompasses all of the data gathering, processing, storage, and communications beyond the remote core, which may be located in an enterprise data center or the cloud. The edge consists of:

- Edge devices, which are assets that generate, collect, process, and/or consume data, including diverse devices such as smart cameras, industrial sensors, robots, autonomous vehicles, wearable devices, smartphones, smart speakers, and drones.
- Edge infrastructure, which includes devices that aggregate multiple data streams from different sources, such as local servers, gateways, and network video recorders.

[1] All of these edge devices can benefit from the incorporation of AI capabilities.

Defining the Edge

Edge devices are often small (for example a smart watch or a smart camera), with little or no space for hefty components. Also, they often need to run on limited power, which means that this edge hardware must be both space and power efficient. These devices must deliver high performance even while adding AI workloads to process locally collected data.

While edge devices can – and often do – operate independently while executing AI inference workloads, it can also be beneficial to connect multiple edge devices to enable federated learning for AI training. Federated learning allows edge devices to collaboratively learn and share prediction models while keeping all training data on the edge devices and removing the need to store all learning data in the cloud, which can improve data security.

Hardware that supports more holistic or complex edge computing in the form of edge clusters or network servers, for example, tend to offer higher performance than stand-alone edge devices. They may also make use of security or connectivity features as required to support their designated use case.

Here are just two real-world examples of the use of AI at the edge:

- There is huge potential for edge computing in manufacturing and industrial settings. For example, Audi's Neckarsulm factory assembles as many as 1,000 cars each day, with around 5,000 welds per car. That's five million welds per day that require inspection in just one factory. Manually inspecting millions of welds in a single day is costly, labor-intensive, and nearly impossible, yet Audi's goal is to inspect 100 percent of welds to an exceptional degree of accuracy.
- Smart cameras and video analytics provide a great opportunity to help monitor and protect endangered habitats, where the physical presence of conservation workers can be problematic. For example, coral reef recovery typically has been monitored by human divers who directly collect data underwater or manually capture video or images of the reef for later analysis. These data-collection methods run the risk of divers interfering with wildlife behavior and inadvertently affecting research results. Opportunity for data capture is limited as well because the divers are only able to safely spend around 30 minutes underwater at a time. Project: CoRaiL in the Philippines overcomes these issues by analyzing coral reef resiliency using smart cameras and AI-enhanced video analytics.

Use Cases for AI at the Edge

Examples of use cases for using AI at the edge include:

1. There are many potential uses for AI in healthcare, but one of the most popular is medical imaging. Thousands of medical images – such as CT scans, X-rays, and MRIs – are produced daily, each requiring careful analysis to identify anomalies and achieve accurate diagnoses.
2. At the retail level, machine vision can reliably read codes, text, and numbers to help manage, track, and analyze inventory levels and to ensure that critical materials reach the people who need them most. Connected, intelligent and responsive digital signs can suggest products and make offers to customers based on their behavior and interests, which in turn helps the retailer know when a message or offer is truly effective. Self-service kiosks and autonomous stores can provide customers with a range of services to help personalize their shopping experience. Meanwhile, machine learning can analyze multiple video streams from cameras set throughout a store to help identify potentially criminal behavior in real time.
3. Robots are being used in the US to help sterilize surfaces in hospitals using bursts of ultraviolet (UV) light, which is highly effective at killing off viruses but can also be harmful to humans. Robots can use AI to navigate around a hospital, and to check that a space is clear of people before it sterilizes the area with UV light. Here, the use of AI helps ensure that safety is maintained throughout the hospital while allowing busy areas to remain open as much as possible.
4. Smart cameras that incorporate AI can be invaluable in helping to automate repetitive, routine tasks to free up employees for more complex challenges. For example, AI-driven license plate recognition is at use in a variety of applications ranging from security, to prevent unauthorized entry, to car washes, where the cars owned by subscribers are automatically admitted into the car wash.

What Does It Take to Implement AI?

When implementing AI on a broader scale across a business, it is important to ensure that each of the three main infrastructure elements – edge devices, edge infrastructure, and cloud – are equipped with sufficient capabilities to handle AI workloads. The requirements are:

• High performance: AI workloads tend to be computationally intensive, so it is essential to have strong compute performance wherever AI training or inference takes place.

• Low latency: One of the strengths of AI is its ability to support real-time decision making. Moving at least some AI workloads to the edge can reduce decision latency.

• High capacity: AI depends on large volumes of data, so any infrastructure running AI helps to avoid bottlenecks by ensuring compute, storage, and memory capacity are up to the task.

• Robust security: AI workloads require large volumes of increasingly sensitive data (for example, in the healthcare or public safety industries). Whatever the AI workload, the devices and software that runs these workloads must be reliably secure.

Intel offers many technologies and solutions that enable organizations to support AI workloads from edge to cloud while meeting these requirements. Figure 1 shows a typical framework to begin an exploration of Intel AI solutions, but specific requirements such as low latency or custom board form factors can ultimately determine the final solution. Intel Edge technology solutions for AI enable high-performance inferencing on equipment ranging from on-premise servers to PCs, cameras, robots, and drones. Because no one size fits all when it comes to AI, the Intel portfolio of CPUs, GPUs, VPUs, and FPGAs are designed to deliver low-latency inference to help remove data bottlenecks. The Intel® AI Analytics Toolkit (AI Kit) for oneAPI and the Intel® Distribution of OpenVINO™ toolkit support a wide range of Intel computing devices with a unified suite of AI development tools.

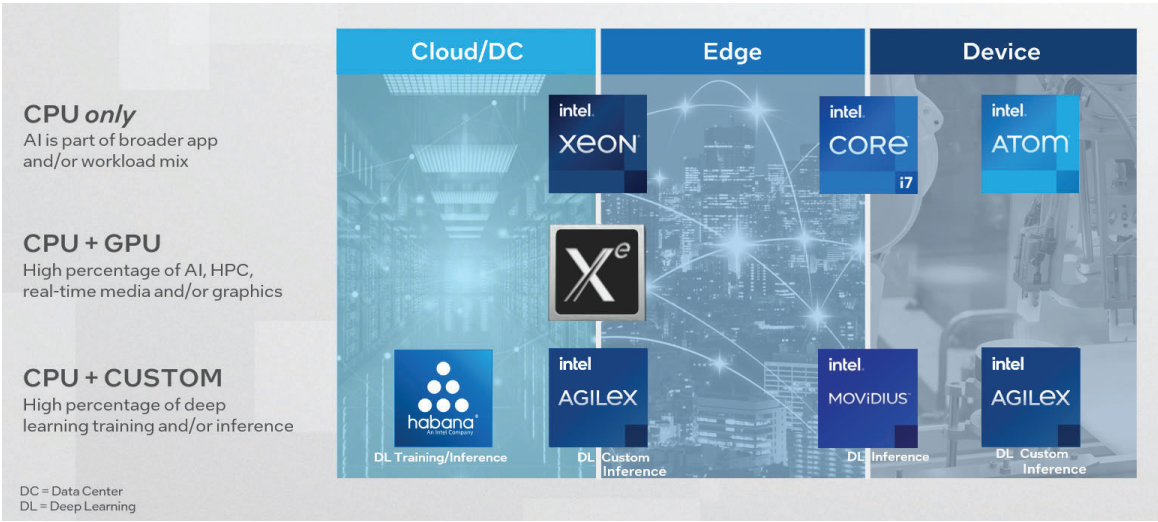


Figure 1. Framework for Intel's AI portfolio

Why FPGAs are especially good for implementing AI

If you examine the human brain, you'll see that it consists of nearly 100 billion neurons. Although that's truly a large number, the number of connections that organize these neurons into networks is on the order of 100 trillion and the number of connections significantly influences the brain's capabilities. The interconnectivity within an FPGA resembles the neural wiring in the human brain. The programmable-logic fabric within an FPGA is similarly connected, which is one reason why Intel® FPGAs are excellent implementation targets for neural networks and other AI workloads. The bit level dynamic programmability of FPGAs at the level of both logic and interconnecting wires is analogous to the flexibility of the brain to reorient attention to specific tasks at hand. In addition, FPGA external I/O have traditionally had enormous flexibility that is not available in other hardware architectures which can connect to sensors from variety of sources such as radar, audio, vibration, and vision. Add this up, and you can see how signals can enter and exit an FPGA in real-time with added intelligence, like the brain.

Intel FPGA families such as Intel® Cyclone® 10 GX FPGA, Intel® Arria® 10 GX FPGA, and Intel® Stratix® 10 GX FPGA offer traditional advantages in AI inference due to I/O flexibility, low power (or energy per inference), and latency. These advantages are further complemented by significantly higher AI inference performance in three newer Intel FPGA and SoC families. These three families are the Intel® Stratix® 10 NX FPGA and the newest members of the Intel Agilex® FPGA family: the Intel Agilex® 5 FPGAs D-Series and Intel Agilex® 5 FPGAs E-Series." These Intel FPGA and SoC families incorporate specialized DSP blocks that are optimized for tensor math, which is fundamental to performing AI calculations more quickly.

The first Intel FPGA to incorporate tensor blocks was the Intel Stratix 10 NX FPGA, introduced on June 18, 2020. The Intel Stratix 10 NX FPGA's tensor block architecture is optimized for performing the common matrix-matrix or vector-matrix multiplications and additions needed for AI calculations and is designed to work efficiently for both large and small matrix sizes. This tensor block supports INT8 and INT4 data computations, along with shared exponent support for FP16 and FP12 block floating point representations.

The Enhanced Digital Signal Processing with AI Tensor Block within the FPGA fabric of these new Intel Agilex FPGAs and SoC FPGAs inherit the design of the variable-precision DSP blocks in the earlier Intel Agilex device families, which already offer AI capabilities. In addition, it adds features derived from the tensor block used in the Intel® Stratix® 10 NX FPGAs. The Enhanced DSP with AI Tensor Block introduces two new important operations: the tensor processing capability for AI and complex number support for signal processing applications such as fast Fourier transform (FFT) and complex finite impulse response (FIR) filters.

The first mode enhances AI with the INT8 tensor mode, which provides twenty INT8 multiplications within one Enhanced DSP with AI Tensor Block, and increases INT8 compute density by 5X versus earlier Intel Agilex device families. The tensor mode uses a two-column tensor structure with both INT32 and FP32 cascade and accumulation capability, and also supports a block floating exponent for improved inference accuracy and low-precision training. In addition, the AI capability of the variable precision DSP functionality has also been enhanced. The vector mode has been upgraded from four INT9 multipliers to six INT9 multipliers. These modes are extremely useful for AI-centric tensor math and for various DSP applications.

Applications	Multiplier	Capabilities per DSP Block		Improvement*
		Earlier Intel Agilex Devices	Enhanced DSP with AI Tensor Block*	
AI, Signal Processing	INT8	4 OPS	20 OPS	5X
	INT9	4 Multipliers	6 Multipliers	50%
Signal Processing	16-bit Complex Multiplier	Needs 2 DSP Blocks	1 DSP Block	2X

Figure 2. Order of magnitude increase in AI and DSP compute density.

*Available in Intel Agilex® 5 FPGAs D-Series and Intel Agilex® 5 FPGAs E-Series.

The second new mode, the complex-number operation, doubles the performances of the tensor block when performing complex-number multiplication. Previously, two DSP blocks were needed for complex-number multiplication, but this new family of Intel Agilex FPGAs and SoC FPGAs can multiply 16-bit, fixed-point, complex numbers within one Enhanced DSP with AI Tensor Block.

FPGAs are a Perfect Fit for Many End Markets on the Network Edge and in the Core

There are many end markets outside of the data center where it makes sense to employ FPGAs to implement both the logic needed by the application and the AI compute capabilities needed to process the data locally. These end markets include:

- Health and life sciences, including medical monitors, 2D diagnostic equipment with image recognition and object detection including X-ray equipment and endoscopes, other types of pathology detection, genome sequencing, surgical robotics.
- Military and aerospace, including unmanned aerial vehicles (UAVs), target detection, radar detection and classification
- Industrial applications to add AI-based detection and real time control at the edge.
- ProAV, including videoconferencing cameras with face detection for automated panning/zooming and background removal, studio cameras with automatic face detection for precise focusing
- Broadcast video, including standard dynamic range to high dynamic range conversion, intelligent conversion among video resolutions, variable frame rate video capture and display
- Consumer applications, including 3D displays with eye detection and tracking for stereo imaging

Here are some in-depth examples of AI use in healthcare and industrial/manufacturing applications:

AI in Healthcare

Demographic shifts in both patient and healthcare provider populations and the desire to improve healthcare outcomes while reducing healthcare costs is driving the use of AI across a broad spectrum of healthcare applications. AI is helping to increase the accuracy of cancer diagnoses based on MRI and CT imaging, assisting surgeons in the surgical suite with both AI-enabled information systems and robotic surgical equipment, and improving the treatment of rare diseases through models based on global datasets.

One specific example of AI's use in healthcare is AI-enhanced endoscopic cameras, which are used in many healthcare specialties such as neurology, orthopedics, urology, and gynecology. These endoscopic camera systems increasingly support advanced imaging capabilities including edge enhancements and color correction that provide doctors with clearer, more easily interpreted images. Intel FPGAs provide the performance, small footprint, and low power consumption needed to add these real-time capabilities to endoscopic camera platforms.

Intel FPGAs also allow endoscopic camera manufacturers to support a variety of AI use cases, including the AI-enhanced detection of:

- Polyps during colorectal screenings
- Abnormal growths associated with Barrett's esophagus, observed during endoscopic esophageal screening

More than 16 million colonoscopies are performed in the US each year and there are more than 200,000 cases of Bartlett's esophagus diagnosed per year in the US alone. Consequently AI-enhanced imaging can have a huge positive impact on the efficiency and accuracy of these endoscopic procedures. Given these new AI-enabled capabilities and enhancements, it's quite likely that AI's use in endoscopy will continue to grow as a method of supporting physicians and helping to manage the rapid growth in demand for minimally invasive screenings and procedures.

AI in Industrial and Manufacturing Applications

Modern manufacturing is a complex system of systems, ranging from sensors, cameras, and actuators through hierarchies of connected and networked control. Intel FPGAs are used throughout this hierarchy to help meet hard real-time and safety requirements. In addition, manufacturing is undergoing a fourth industrial revolution that's merging operational technology (OT) systems with information technology (IT) systems to create smarter, more flexible factories that can provide more efficient and more autonomous production while requiring less human intervention.

Communication technologies used throughout industrial applications and manufacturing plants including 5G, industrial gateways, and smart network interface cards (NICs) all employ Intel FPGAs. They're used wherever I/O flexibility, ability to directly ingest data, deterministic computing capabilities, low operating power, and tolerance for harsh industrial conditions can benefit the workloads. AI technology runs throughout all these efforts and is increasingly employed in manufacturing applications for both vision and non-vision tasks.

Table 1 illustrates the breadth of industrial and manufacturing AI applications.

	Die Casting	Textiles	Electrical and Electronics	Manual Assembly
Task	• Package, part, or surface defect detection • Inline quality control/ assurance • Product defect detection • Raw material appearance inspection • Asset management	Predicting future outcomes based on historical data • Predict product quality Assess equipment health and predict maintenance • Predict yield fluctuation	• Identifying opportunities/ processes to improve upon • Safe worker collaboration • Pick and Place and Sorting • Palletization/Depalletization • Welding • Machine tending • Vision assisted robot • teaching/programming • Energy-optimized motion control • AMR perception for navigation and avoidance	• Control processes with soft real-time capabilities • Optimize process efficiency
Value	• Reduce production cost • Reduce customer returns	• Reduce factory downtime • Help prevent costly maintenance	• Improve factory production and efficiency	• Integrate multi-vendor solutions for interoperability • Manage cost of lifecycle management

Table 1. Example use cases for AI in the modern factory

Intel FPGAs used in edge equipment can implement power-efficient AI workloads either next to or in line with other sensing and control functions. For example, Intel FPGAs can directly ingest camera or sensor data to run workloads deterministically with low latency and high throughput. Intel FPGAs can directly ingest video from one or more cameras, pre-process that video for contrast and exposure, enhance edges, correct colors, grab frames of interest from the video stream, and detect features or defects – and all in real time. PLC (programmable logic controller) manufacturers are already incorporating small AI engines in their newest PLCs to add smart capabilities to their next-generation controllers. On the production line, AI-enhanced visual inspection is already finding defects and assessing quality faster and more accurately than human workers.

Tightly integrating AI enhancements with a video-processing pipeline or other functions is key to building an optimized, real-time industrial system where both low power consumption and low latency are key. Table 2 illustrates some of the many industrial and manufacturing uses for AI-enhanced image and video processing:

AI makes it easier to set up and teach a robot to sense its surroundings and react accordingly. Tasks in such a situation include learning by demonstration, picking and placing objects, controlling end-effector tools such as welding tools using direct feedback, and working safely in collaboration with other robots or humans. These same tasks apply to autonomous mobile robots (AMRs) because they must also use sensor data in real time to construct and constantly update a map of their surroundings and navigate accordingly. The AI features in these AMRs must consume very little power, so as not to overly tax the robot's batteries, yet they must also be reliable, exhibit low latency, and tightly integrate with other workloads.

The industrial edge includes many non-vision AI applications. Factory owners are keen to make these applications as efficient as possible to reduce total cost of ownership (TCO) and boost manufacturing yields. TCO factors include minimizing downtime through AI-based assessment of machine health and predictive maintenance. Using non-intrusive sensors such as voltage, temperature, vibration and sound, ML workloads can accurately detect emerging issues and predict the need for maintenance or repair before the problem becomes serious and affects or halts production.

Die Casting	Textiles	Electrical and Electronics	Manual Assembly
• Broken/Defective casting • Missing casting overflow • Hole blockage • Warpage deformation defect	• Warp and weft direction defect • Woven defect • Dirty point • Damaging defect	• Component existence mismatch • Component type mismatch • Component part number mismatch • Component placement mismatch • Circuit shortage	• Headcount detection • Production output detection • Human behavior detection

Table 2. Examples of industrial visual inspection and detection where AI enhancements are useful.

Intel FPGAs are very often used at the edge to process these sensor signals using AI and ML to reduce the factory's network bandwidth requirements and to quickly identify problems without the latency incurred from sending unprocessed data to the cloud and then waiting for a decision from the cloud. Leading original equipment manufacturers (OEMs) are using AI to dynamically calculate energy efficient movement paths – for example to control how a multi-axis robot moves an object from point A to point B within a prescribed amount of time – to both improve work efficiency and to minimize energy use. AI can be used to enhance closed-loop control algorithms, and low-latency AI algorithms and deterministic computing capabilities are critically important for such tasks. Intel FPGAs are especially well suited to implementing closed-loop algorithms with AI enhancements.

Around the world, industrial and manufacturing customers ask:

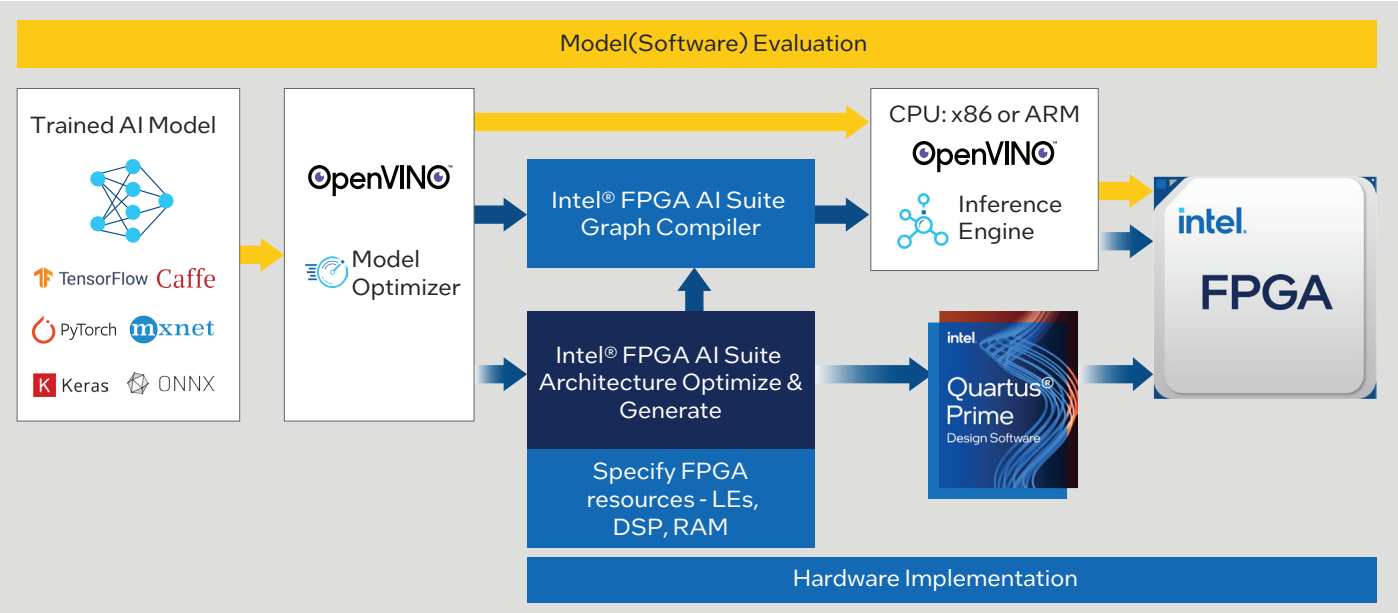
- How can we meet rising product quality requirements?
- How can we optimize factory operation for higher throughput and greater efficiency?
- How can we better predict and reduce equipment and system downtime?
- How can we respond and adapt faster to changes in market demands?

AI can help to provide answers to all these questions, which is why Intel FPGAs are quickly adding AI capabilities into their existing sensing and control capabilities. All of these features are required to create more effective industrial systems that can meet hard real-time requirements at the factory's edge.

Edge-ready AI toolkits for Intel FPGAs and SoCs

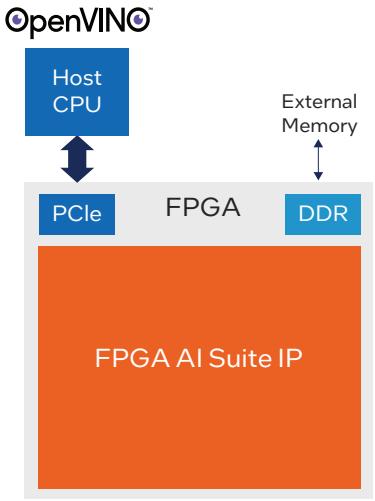
Distributed AI/ML edge solutions are among the most complex to develop. Intel development tools and software help developers streamline their workflows and speed the deployment of distributed edge solutions, with an emphasis on open standards and support for containerized and cloud-native development. Intel development tools for developing AI/ML applications using Intel FPGAs and SoCs include:

1. Intel® AI Analytics Toolkit (AI Kit) for oneAPI gives data scientists, AI developers, and researchers familiar Python tools and frameworks to accelerate end-to-end data science and analytics pipelines. Intel provides oneAPI libraries that support low-level compute optimizations with this toolkit to maximize performance for many workloads ranging from preprocessing through machine learning and provides interoperability for efficient model development.
2. The Intel® Distribution for Python helps create fast ML applications for multiple Intel computing platforms with support for common libraries and frameworks including TensorFlow, Keras, PyTorch, oneDNN and BigDL. These tools enable quick application development for a range of AI/ML workloads.
3. The Intel® Distribution of OpenVINO™ toolkit supports the development of deep learning applications essential for computer vision use cases at the edge.
4. The Intel® FPGA AI Suite enables FPGA designers, ML engineers, and software developers to efficiently optimize AI-enabled designs based on Intel FPGAs. Utilities in the Intel FPGA AI Suite speed up FPGA development for AI inference using familiar and popular industry frameworks such as TensorFlow or PyTorch and OpenVINO toolkit, while also leveraging robust and proven FPGA development flows with the Intel® Quartus® Prime Software.
5. The Intel FPGA AI Suite is flexible and can be configured for a variety of system-level use cases. Using a single push button flow, users can generate an optimized AI inference IP block for integration into the Intel Quartus Prime Software. Users can specify the device resources (DSP, memory, logic elements) and the throughput for the architecture optimizer in the Intel FPGA AI Suite. This unique, customization ability is critical to explore the design space and optimize the footprint of edge and embedded AI applications for size, weight, and power.

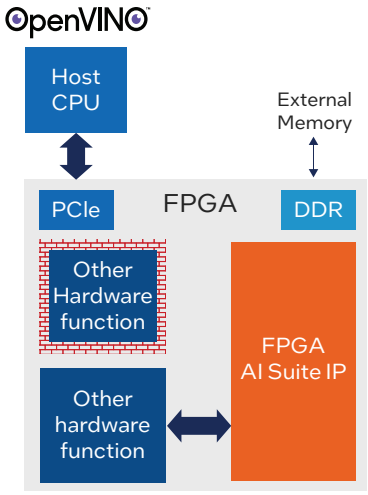


Four different ways to incorporate Intel FPGAs and AI/ML into a system using Intel development tools include:

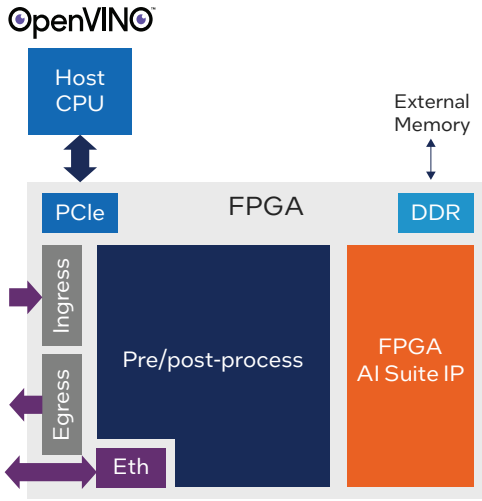
1: CPU offload with an FPGA-based AI/ML accelerator. The host CPU communicates with the AI/ML accelerator over a PCIe interface. Intel FPGAs directly support PCIe connections to Intel® CPU hosts.



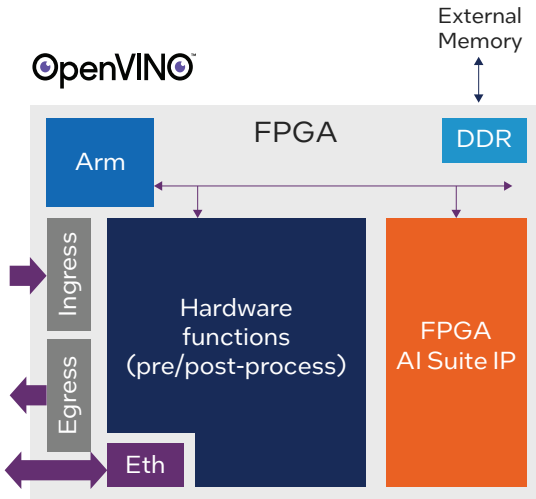
2: Multi-function CPU offload using an Intel FPGA to implement AI accelerators and additional logic. The Intel FPGA provides AI/ML acceleration to the host Intel CPU and implements any additional logic required by the application. As with example 1, the host CPU communicates with the AI/ML accelerator over a PCIe interface.



3: Ingest/Inline Processing + AI: The FPGA-based AI accelerator directly ingests data, processes it using AI and algorithmic workloads, and then streams processed data and inferences to a host Intel CPU over a PCIe connection.



4: An Intel® SoC FPGA acts as an AI/ML accelerator using its integrated CPU (an Arm or Nios® processor core) and directly ingests data, processes the data, implements AI/ML inferencing, and streams the processed data and inferences via Ethernet to the cloud through a network. The FPGA also implements any additional logic circuitry needed by the application.



Conclusion

Intel has accumulated decades of experience working across the entire edge value chain—from builder to integrator to cloud and network provider to developer. By working with customers on many types of use cases, Intel has developed solutions to common integration headaches, which has resulted in the creation of hundreds of preconfigured packages backed by a mature developer ecosystem that is constantly optimizing and innovating. Take advantage of this ecosystem to speed development time and accelerate your time to results by:

- **Using ready-to-deploy enterprise AI solutions**. Intel® AI Builders offers access to over 300 leading global AI software, hardware, and service providers with more than 150 solutions across diverse use cases and markets, enabling any business to quickly harness AI.
- **Ensuring optimal AI deployments**. Intel® Select Solutions for AI helps you simplify and accelerate infrastructure deployment with benchmark-tested and verified solutions optimized for Intel® Xeon® processors and other Intel platforms.
- **Reducing development and collaboration challenges**. Intel® AI: In Production helps accelerate the path to production with Intel technologies, software tools, development kits, code samples, and solutions from the broad Intel partner and developer ecosystem.
- **Intel FPGA AI Suite and OpenVINO toolkit** bridges the last mile to deployment on Intel FPGA and SoCs with a primary focus on ease of use of creation and integration of deep learning inference FPGA IP.

If you're developing edge or core equipment that could benefit from the addition of AI capabilities, please contact your local Intel field sales representative to find out how Intel can help your team.

References

- Rob van der Meulen, What Edge Computing Means for Infrastructure and Operations, October 3, 2018.
- Carl Zimmer, "100 Trillion Connections: New Efforts Probe and Map the Brain's Detailed Architecture," Scientific American, January 1, 2011.
- Intel® FPGA AI Suite melds with OpenVINO™ toolkit to generate heterogeneous inferencing systems

altera™
An Intel Company

WP- 01321-1.1