Rising Edge AI Requirements Demand Higher Performance Solutions

A Lattice Semiconductor White Paper.

August 2019

Edge AI applications like presence detection and object counting are growing in popularity, but designers are increasingly pressed to implement Edge AI solutions with low power and a small form factor without compromising performance. The latest iteration of Lattice's sensAI technology stack, coupled with the ECP5 and iCE40 UltraPlus FPGAs, offers designers the hardware platforms, IP, software tools, reference designs and design services needed to deliver low power, high performance AI at the Edge.

Learn more:

www.latticesemi.com/sensAI

Contact us online:

www.latticesemi.com/contact
www.latticesemi.com/buy

Rev. 1.0 WP0019

TABLE OF CONTENTS

Executive Summary

The market for low cost, high performance Edge solutions is growing increasingly competitive. Leading market research firms are forecasting that over the next six years the market for Edge solutions will explode. IHS expects over 40 billion devices will be operating at the network Edge by 2025, while market intelligence firm Tractica predicts over 2.5 Billion Edge devices will ship annually by the same year.

As a new generation of Edge applications emerges, designers are increasingly pressed to develop solutions that combine low power and small form factor without compromising performance. Driving demand for these new AI solutions are a growing range of Edge applications such as presence detection for home control in applications like smart doorbells and security cameras, object counting for retail applications like inventory, and object and presence detection in industrial applications. On one hand, the market demands that designers develop solutions that operate at performance rates higher than ever before. On the other hand, latency, bandwidth, privacy, power and cost issues constrain them from relying on computational resources in the cloud to perform analytics.

At the same time, performance, power and cost limitations vary enormously from application to application. As data requirements for always-on Edge applications continue to drive demand for cloud-based services, designers must address traditional concerns about power, footprint and cost. How can developers address rising system constraints on power budgets that run in the mW range and form factors that can range from 5 mm2 to 100 mm2. The wide range of performance requirements alone is difficult to track.

Leveraging FPGA Advantages

Lattice's FPGAs are uniquely positioned to address the rapidly changing market for Edge devices. One way designers can quickly bring more computational resources to the Edge without relying on the cloud is to use the parallel processing capabilities inherent in FPGAs to accelerate neural network performance. Moreover, by using lower density FPGAs optimized for low power operation and available in compact packages, designers can meet the stringent power and footprint limitations associated with new consumer and industrial applications. As an example, Lattice's iCE40 UltraPlus™ and ECP5™ product families support development of Edge solutions that consume anywhere from 1 mW to 1 W on compact hardware platforms ranging from 5.5 mm2 to 100 mm2. By combining ultra-low power, high performance and accuracy with comprehensive legacy interface support, these FPGAs give Edge device developers the flexibility they need to address changing design requirements.

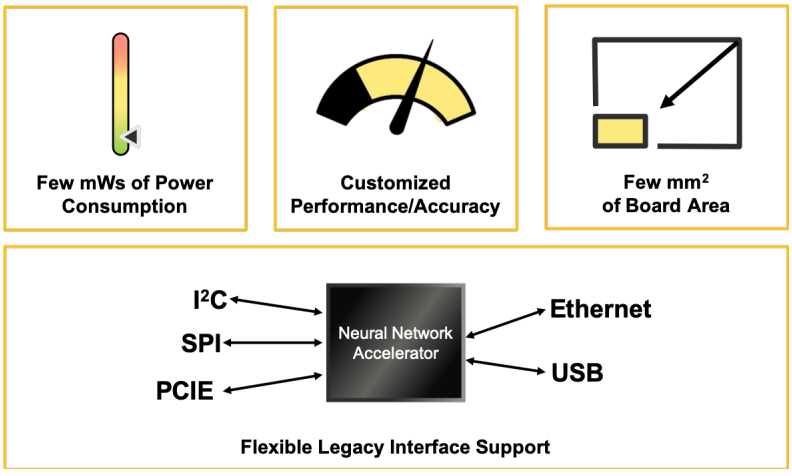

Figure 1: Lattice Semiconductor's low power, small form factor FPGAs provide the right blend of performance and features for supporting Edge AI applications.

To address this need and accelerate development, Lattice has brought together sensAI™, the industry's first technology stack that gives designers all the tools they need to develop low power, high performance Edge devices for smart homes, smart factories, smart cities, and smart cars. Designed to address the growing need for Edge devices with AI support, sensAI offers a comprehensive hardware and software solution for implementing low power, always-on AI functionality in smart devices operating at the network Edge. Introduced in 2018, it was designed to seamlessly create a new design or update an existing one with low power AI inferencing optimized for these new application requirements.

What's in this comprehensive design ecosystem? First, modular hardware platforms from Lattice such as the iCE40 UPduino 2.0 with the HM01B0 Shield and the ECP5-based Embedded Vision Development Kit (EVDK) provide a solid foundation for application development. The UPduino can be used to target AI designs that consume just a few mWs of power, while the EVDK supports applications requiring more power but generally operating under 1W.

Soft IP can be easily instantiated into an FPGA to speed up neural network development. Accordingly, the sensAI development package includes Compact Convolutional Neural Network (CNN) accelerator IP that allows designers to implement deep learning applications in the iCE40 UltraPlus FPGA. sensAI also offers a full CNN parameterizable accelerator IP core that can be implemented into Lattice's ECP5 FPGAs. These IPs provide support for variable quantization. That, in turn, empowers designers to make tradeoffs between data accuracy and power consumption.

Lattice's sensAI stack allows designers to explore design options and tradeoffs with an easy-to-use tool flow. Designers can perform network training using industry standard frameworks such as Caffe, TensorFlow and Keras. The development environment also provides a neural network compiler that maps the trained network model into a fixed point representation with support for variable quantization of weights and activations. Designers can use the compiler to help analyze, simulate, and compile different types of networks for implementation on Lattice's accelerator IP cores without prior RTL experience. Designers can then use traditional FPGA design tools like Lattice's Radiant and Diamond to implement the overall FPGA design.

To speed up design implementation, the sensAI stack offers a growing array of reference designs and demos. This list includes designs and demos for facial recognition, hand gesture detection, key phrase detection, human presence detection, face tracking, object counting and speed sign detection. Finally, design teams often need unique expertise to complete a design. To meet that need Lattice has built relationships with a number of design service partners in various geographical locations to support customers locally if they don't have the required AI/ML expertise.

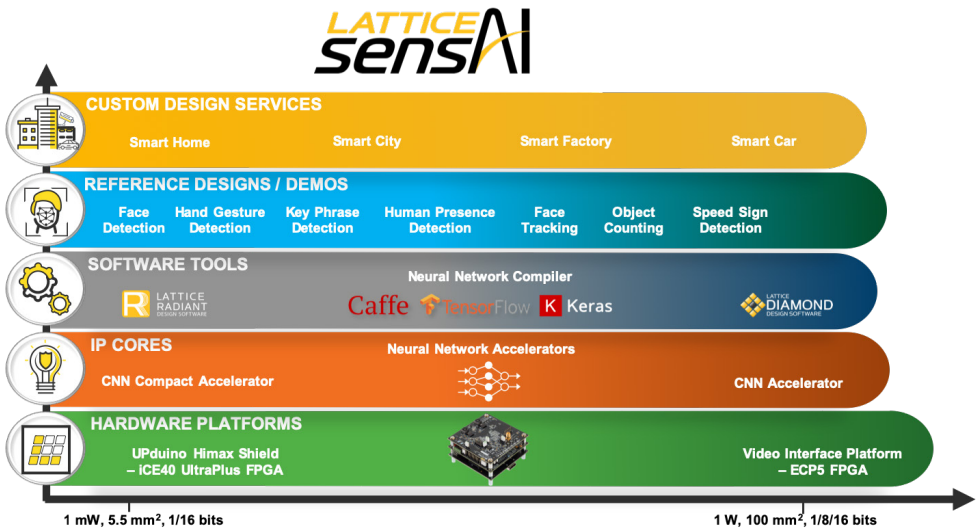

Figure 2: Lattice sensAI is a complete hardware and software solutions stack for development of Edge AI applications.

Major Enhancements

To meet these rapidly rising performance requirements for Edge AI, Lattice announced major performance and design flow enhancements to sensAI in 2019. The revised technology stack now delivers a 10X improvement in performance over the initial sensAI stack. This performance boost was driven by multiple factors, including optimizing memory access using an updated CNN IP and neural network compiler with features like 8-bit quantization, smart layer merging and a dual DSP engine.

In the latest version, the memory access sequence was significantly improved by updating the Neural Network Compiler to support 8-bit input data. Not only does this reduce the amount of external memory access by half, it also enables the use of higher resolution images as data inputs. By using higher resolution images, solutions are more accurate.

To further accelerate performance, Lattice optimized the convolution layer in the sensAI Neural Network to reduce the total time it takes to compute convolutions. Lattice has doubled the amount of convolution engines allowed in their devices, which can reduce convolution time by an estimated 50 percent.

Given that Lattice has improved performance without increasing power consumption, designers now have the opportunity to move to a lower gate count device in Lattice's ECP5 FPGA product line without increasing power. Optimized demos help convey this step-up in performance. A human presence detection demo, optimized for low power operation and featuring a CMOS image sensor, offers a resolution of 64 x 64 x 3 over a VGG8 network. The system runs at five frames per second whole dissipating a meager 7 mW using an iCE40 UltraPlus FPGA. A second, performance-optimized demo, targeted at human counting applications and also featuring a CMOS image sensor, provides a resolution of 128 x 128 x 3 over a VGG8 network. This demo performs at 30 frames per second while dissipating 850 mW using an ECP5-85K FPGA.

Human Presence Detection
Power Optimized

- Sensor: CMOS image sensor
- Resolution: 64x64x3
- Network: VGG8
- Speed: 5 frames per second
- Power: 7 mW on iCE40 UltraPlus

Human Counting
Performance Optimized

- Sensor: CMOS image sensor
- Resolution: 128x128x3
- Network: VGG8
- Speed: 30 frames per second
- Power: 850 mW on ECP5-85K

Figure 3: These reference designs illustrate the power vs. performance options sensAI provides.

At the same time, sensAI's seamless user experience supports the use of new neural network models, machine learning frameworks, and faster design cycles. New customizable reference designs simplify the development of popular Edge solutions like object counting and presence detection while a growing family of design partners provides important design services. With these designs Lattice now offers developers all the crucial components they need to easily replicate or modify their design. As an example, the block diagram below illustrates the comprehensive group of components Lattice now offers, including training models, training data sets, training scripts, updated neural network IP and an updated neural network compiler.

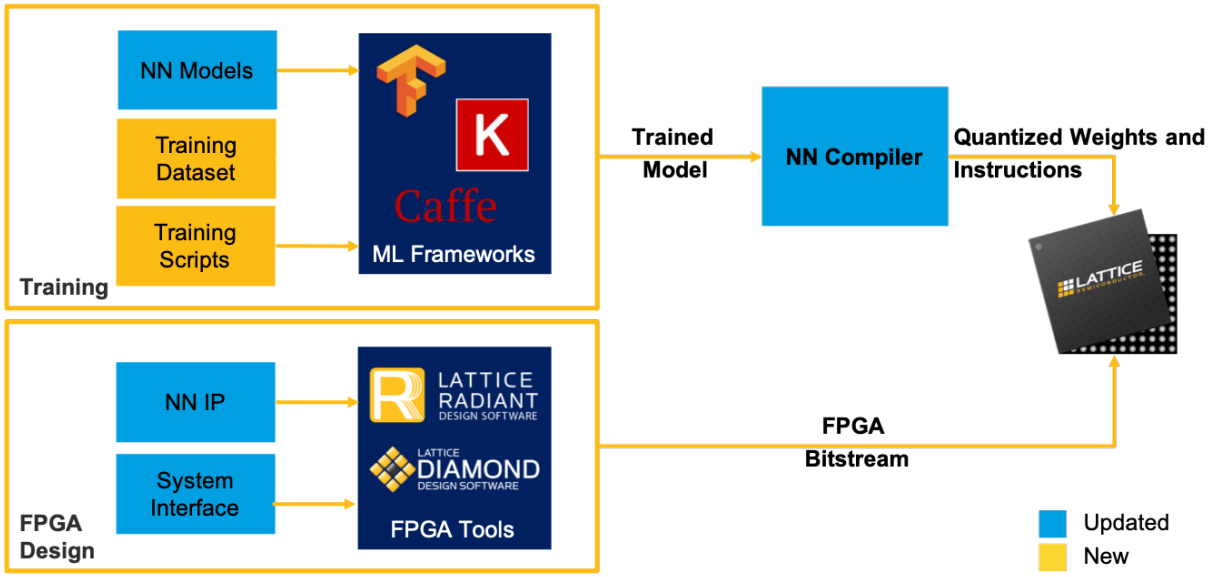

Figure 4: The sensAI design flow includes industry-leading machine learning frameworks, training data and scripts, and neural network IP required for designing and training Edge AI devices.

As part of its focus on delivering a seamless user experience, Lattice has extended its support for machine learning frameworks. While the initial sensAI stack supported Caffe and TensorFlow, its successor adds support for Keras, an open source neural network written in Python and designed to run on top of TensorFlow, Microsoft Cognition Toolkit or Theano. Intended to help engineers quickly experiment with deep neural networks, Keras facilitates fast prototyping by offering an environment that is highly user-friendly, modular and extensible. It was originally conceived as an interface rather than a standalone machine-learning framework and offers developers a high level of abstraction that accelerates the development of deep learning models.

To further enable ease of use, Lattice updated the sensAI Neural Network Compiler tool to automatically select the most accurate fraction bits when converting a machine learning model to the firmware file. The updates sensAI stack also comes with a hardware debugging tool that users to read and write to every layer in the network. After software simulation, engineers will want to know how their network will perform on live hardware as well. With this tool, engineers can see results in live hardware within minutes.

Furthermore, the latest version of sensAI is supported by a growing number of companies that supply design services and product development skills optimized for lower power, always-on Edge devices. These companies can help customers deliver Edge AI devices, either by seamlessly updating existing designs or develop entirely new solutions for specific applications.

sensAI Use Cases

Lattice sees its new higher performance stack being used in any of four different accelerator use cases. In the first case (see figure 5 below) designers will use the stack to build solutions designed to operate in standalone mode. This system architecture offers designers the opportunity to develop always-on, integrated solutions on Lattice's iCE40 UltraPlus or ECP5 FPGAs in low latency, secure implementations where FPGA resources can be used for system control. A typical application for this use case could be a stand-alone sensor for human presence detection or counting.

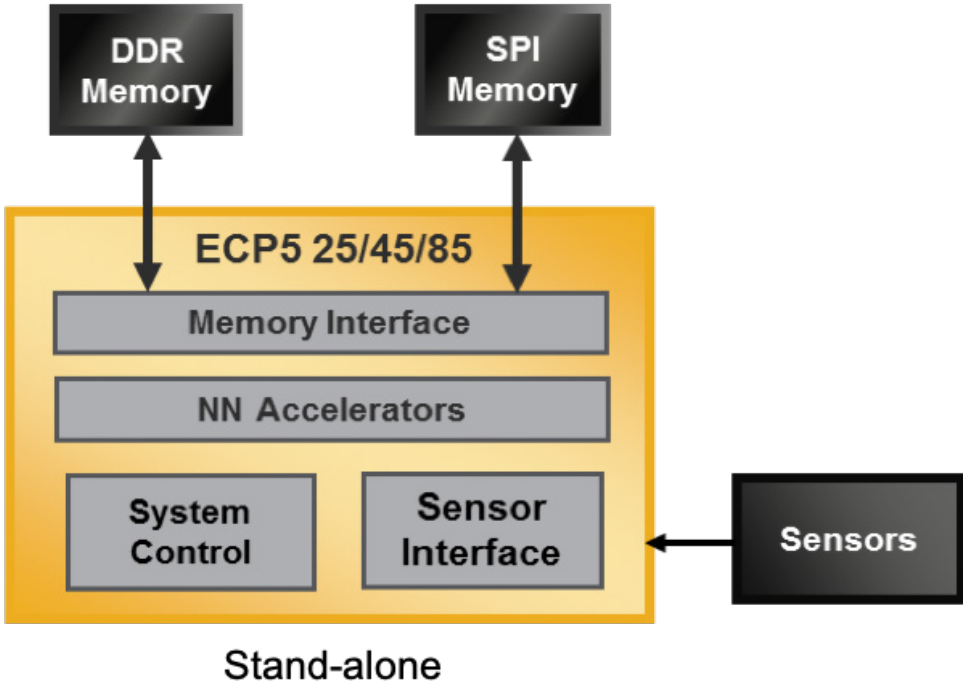

Figure 5: The diagram depicts the sensAI stack as a standalone Edge AI processing solution.

Designers are also using the sensAI stack to develop two different types of pre-processing solutions. In the first (see figure 6 below) designers are minimizing the cost of sending data to a SoC or the cloud to perform analysis using Lattice's sensAI stack and a low power iCE40 UltraPlus FPGA to pre-process sensor data. If used in a smart doorbell, for example, the sensAI solution would perform the initial read on incoming image sensor data. If the initial read is defined as non-human, such as a cat, the system would not wake the SoC or connect to the cloud for further analysis and, thereby, minimize data costs and power usage. If the pre-processing system identifies the subject at the door as human, it would wake up the SoC for further analysis. This approach dramatically reduces the amount of data the system must analyze and the power requirements to perform that function, particularly in always-on Edge applications.

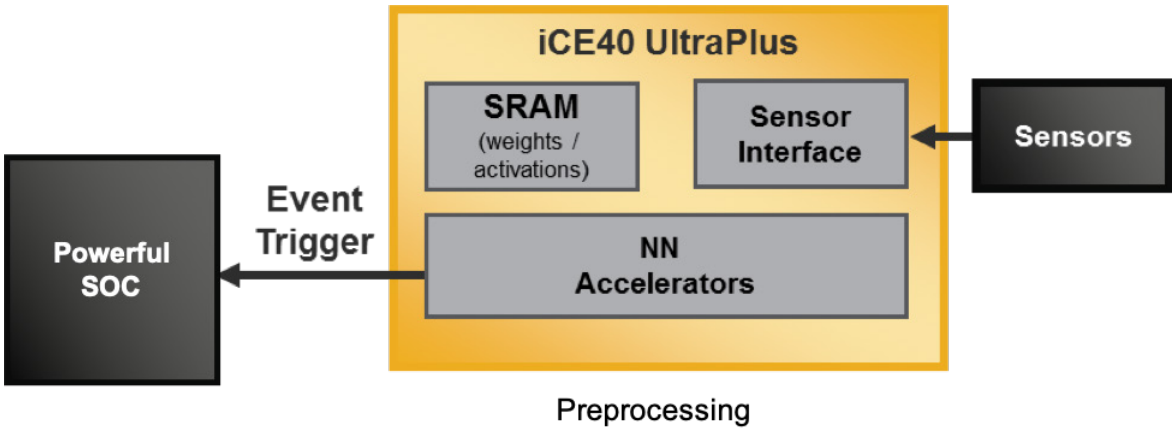

Figure 6: In this example, the sensAI stack would pre-process sensor data to determine if it should be sent to the SoC for further processing.

In a second pre-processing application designers could use an ECP5 FPGA to perform neural network acceleration (see figure 7). In a growing number of cases companies who have developed a proven MCU-based legacy solution are finding they want to add some form of AI capability without replacing components or rebuilding their design. In some instances their MCU is relatively underpowered. A typical example might be a smart industrial or home application that needs additional filtering on an image before performing analytics. Here designers can either add another MCU and go through the time-consuming process of proofing their design, or they can add an accelerator between the MCU and the data center to perform post processing and minimize the amount of data sent to the cloud. This approach is particularly attractive to developers of IoT devices who want to add AI capability.

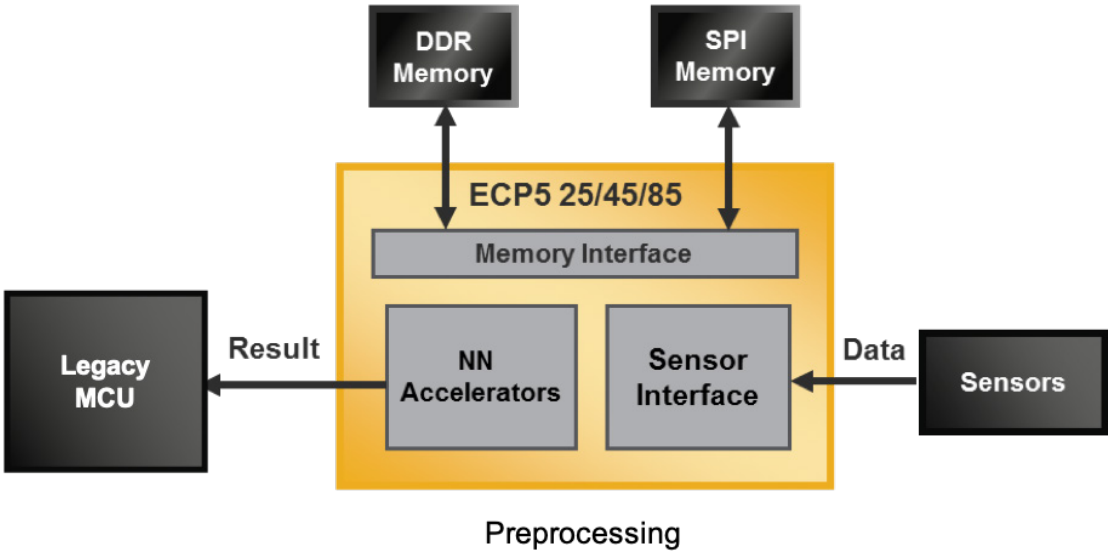

Figure 7: In a second system architecture using pre-processing, designers can use the ECP5 and the sensAI stack to pre-process sensor data and accelerate overall neural network performance.

Designers can also use sensAI accelerators in a post-processing role (see figure 8). Instead of sending data to cloud for processing, ECP5 FPGA can be used as an accelerator at the Edge. For example in security camera and smart door bells, ECP can sit next to the main system SoC and run pattern detection on data from image sensor after the SoC cleans up image data using internal ISP. In this use case there is no need to run pattern detection on the cloud saving bandwidth and reducing security and privacy concerns.

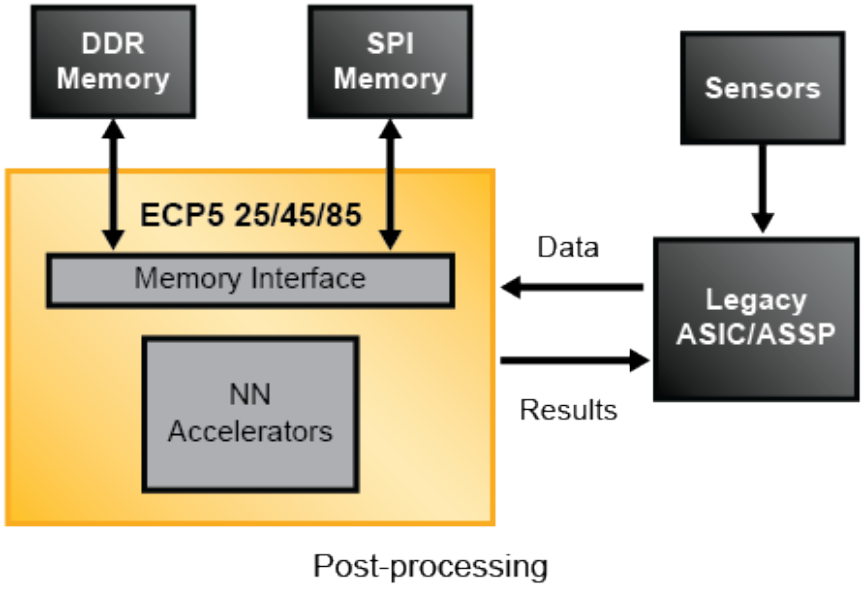

Figure 8: Pattern detection at the Edge using ECP5 eliminating the need to send data to the cloud for processing.

Conclusion

Clearly, the next few years will prove crucial to the development of the always-on, smart Edge device market. As applications grow increasingly complex, designers will need tools capable of supporting higher levels of performance at low power. The latest iteration of Lattice's sensAI technology stack, coupled with the ECP5 and iCE40 UltraPlus FPGAs, offers designers the hardware platforms, IP, software tools, reference designs and design services they'll need to beat their competitors and rapidly develop successful solutions.

Learn more:

www.latticesemi.com/sensAI

Contact us online:

www.latticesemi.com/contact
www.latticesemi.com/buy