Lattice SensAI 4.1: Tools and IP Transform Low-Power FPGAs into Intelligent AI/ML Edge Computing Engines

Abstract

Explosive growth of edge devices is driving the development of new applications that transform massive amounts of raw data into useful, actionable information for real-time decision making. Lattice's sensAI 4.1 solution stack provides ready-to-use AI/ML tools, IP cores, hardware platforms, reference designs and demos, and custom design services to bring these edge devices and applications to market quickly.

SPONSORED BY LATTICE

Introduction

No doubt, you've read or heard about the growing tsunami of data streams being generated by the exploding number of edge devices, including autonomous vehicles, IoT devices, consumer electronics, and even laptops and PCs. By many estimates, there will be many tens of billions of IoT devices in operation by 2025. These edge devices send myriad forms of data to the cloud in the form of continuous data streams sent at widely varied data rates. In aggregate, these billions of devices are generating massive amounts of raw data, which will only increase over time.

Video imagers in security cameras, autonomous vehicles, and PCs generate high-rate, high-resolution video streams. IoT devices generate moderate-rate data, which aggregate into large data streams. Many other types of IoT sensors (measuring temperature, pressure, position, light levels, etc.) generate low-rate data streams, but there soon will be billions of these sensors. Therefore, even these individual low-rate streams can aggregate into large, high-rate data streams before entering the cloud.

Emerging 5G wireless and other high-speed networking technologies, including picocells, long-distance IoT-specific networks, such as LoRaWAN, and global networked satellite constellations, such as SpaceX's expanding StarLink broadband network and Swarm Technologies' nascent satellite-based IoT network, provide ever more pervasive and ever faster cloud access. (Note: StarLink acquired Swarm Technologies in August 2021.) These varied communication and networking technologies accelerate the development of emerging, new, edge-computing devices and applications.

Emerging new edge devices and applications include autonomous vehicles, robotics, automated manufacturing, remote monitoring, supply-chain and logistics systems, and video monitoring for public and private security. The demand for these edge systems is exploding because they boost efficiency, cut operating costs, and improve user experiences, but no matter how much wireless and wired communications infrastructure we add, the data tsunami threatens to overwhelm and clog all of these shiny new data pipes leading to the clouds.

Processing on the Edge Unclogs the Data Pipes

These trends highlight the need to perform far more processing at the network's edge, as close to where the data is being generated as possible, so as to reduce the amount of data that must be sent into the cloud. The explosive growth of IoT and other network-connected devices is a major force driving the design of new edge devices, which further spurs the development of new applications to transform raw data into useful, actionable information to support the speedy decision making that permits real-time responses to changing situations.

Early in the development of edge computing, companies focused on the costs of delivering data to data centers over long distances. Initially, the need to access data stored in the cloud and to other computers connected to the cloud defined edge applications. These early edge applications were usually not real-time applications; response times in hundreds of milliseconds or even seconds were tolerable. However, the development of IoT devices and the growing need for real-time processing, analysis, and response at the edge are now driving edge technology to advance ever more forcefully, accompanied by much greater design challenges.

Edge processing brings computation and data storage closer and closer to the devices where the data is being gathered, instead of performing the analysis and decision making at a data center that can be thousands of miles away. Real-time applications at the edge generally cannot tolerate latency delays, so the processing, analysis, and decision making must move into the edge device itself. There are many edge devices, including autonomous vehicles, IoT sensors, security cameras, smartphones, laptop computers and PCs. So the edge computing opportunity is very large.

The Cloud Can't Do Everything. There's Too Much to Do.

Edge computing's development has been spurred by the exponential growth of smartphones and IoT devices, which are omnipresent and must connect to the Internet to send or receive information to and from the cloud. Some IoT devices, such as video cameras, generate enormous amounts of data during the course of their operations.

Other IoT devices, like temperature sensors, individually create small amounts of data but there are billions of these sensors, creating a lot of data for the cloud to handle. Given this situation, edge-based processing has simply become essential to reduce the costs of the network communications to the cloud and cloud storage costs, and to prevent the eventual overloading of the pipes leading to the cloud.

Developers of these edge products and applications increasingly employ artificial intelligence and machine learning (AI/ML) algorithms for complex pattern matching and recognition to help analyze data and make decisions based on that analysis. In fact, the rise in the use of AI/ML techniques has been meteoric.

AI/ML algorithms are now viewed as essential to the efficient processing of raw data because they identify and recognize complex, multidimensional data patterns that are increasingly difficult to isolate and identify using conventional algorithmic programming. Specific AI/ML applications include the detection, recognition, identification, and counting of people and objects; asset and inventory tracking, environmental sensing, sound and voice detection and identification, system health monitoring, and the scheduling of system maintenance.

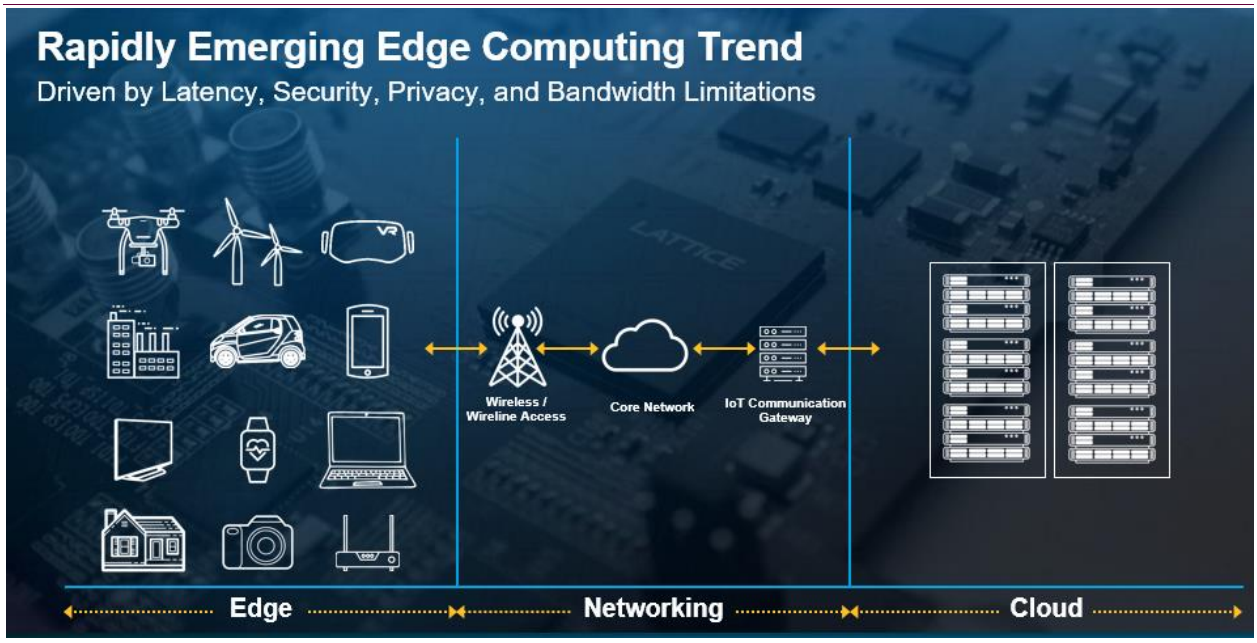

Figure 1 Edge Computing Trends. (Source: Lattice)

Many edge applications that can benefit from AI/ML deployment must function under extremely challenging energy constraints. Often, these widely distributed devices must run on battery power. Such applications abound in many edge environments including factories, farms, office buildings, retail stores, hospitals, warehouses, streets, and residences. Increasingly, as their numbers grow, devices used in these environments must operate for long periods, perhaps even months or years, on one battery charge or on harvested and stored energy.

As a result, many of these devices need to spend a significant amount of the time in a sleep or hibernation state, where most of their circuitry is placed in a low-power idle mode while the device is not actively operating. An activation event then starts up the edge device when it's needed. In such applications, an essential circuitry that runs on very little power must stay alert, awaiting for an activation event, which then powers up the rest of the device as needed.

FPGAs Offer Low-Power Methods for Implementing AI/ML

It might seem as though the need for low operating power and AI/ML algorithmic execution present opposing requirements for low-power edge device designs. However, these two complex sets of design requirements need not be at odds. Lattice's newest FPGAs – the low-power, small-footprint, high-performance Lattice CertusPro-NX family of devices – are specifically tailored to meet the many design requirements of low-power edge devices. These FPGAs can support multiple sensors, displays, high-resolution video, networking, and edge AI/ML processing.

Meanwhile, the latest release of the company's sensAI solution stack, version 4.1 provides ready-to-use AI/ML tools, IP cores, hardware platforms, reference designs and demos, and custom design services that design teams need to develop and bring new edge devices to market quickly. This latest version of the sensAI solution stack supports CertusPro-NX FPGAs.

The Lattice sensAI solution stack facilitates end-to-end AI/ML model training, validation, and compilation. Lattice added the sensAI Studio design environment, a GUI-based tool that helps developers build accelerated machine learning applications quickly, to its sensAI solution stack in version 4.0, which was announced earlier in 2021. Configured with the tools in the Lattice sensAI 4.1 solution stack, edge-computing designs can deliver real-time AI/ML performance while consuming very little power – as low as 1mW to 1W – when based on Lattice iCE40 UltraPlus, CrossLink-NX, ECP5, and CertusPro-NX FPGAs.

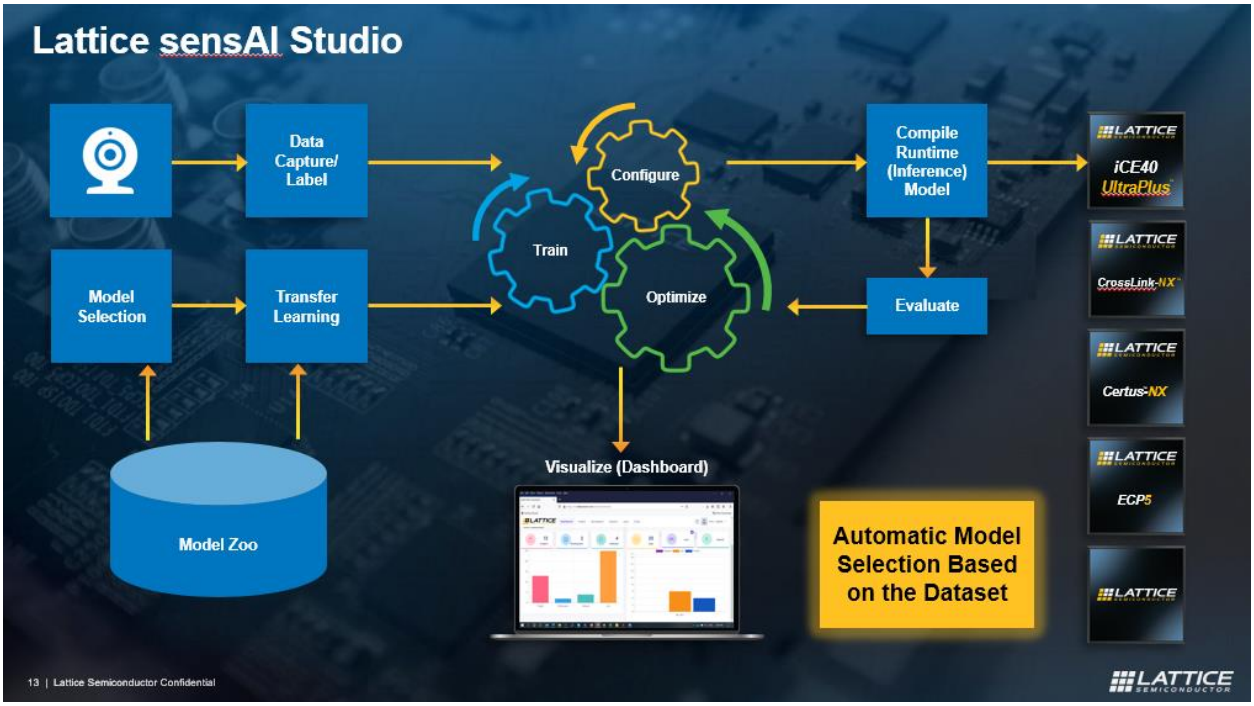

Figure 2 The Lattice sensAI Studio design environment facilitates end-to-end AI/ML model training, validation, and compilation. (Source: Lattice)

Support for the Lattice CertusPro-NX FPGA family in the sensAI 4.1 solution stack has allowed Lattice to add performance enhancements such as the ability to classify multiple objects simultaneously in real time, in addition to existing object detection and tracking capabilities. The sensAI 4.1 solution stack includes an updated neural network (NN) compiler and is compatible with other widely-used ML platforms, including the latest versions of Caffe, Keras, TensorFlow, and TensorFlow Lite.

IP cores in the Lattice sensAI 4.1 solution stack include three types of convolutional neural network (CNN) accelerators- CNN, CNN Plus, and CNN Compact – and a CNN coprocessor engine. The CNN IP core allows developers to use many of the widely used CNNs published by others, such as Mobilenet v1/v2, Resent, SSD, and VGG, or to implement custom CNN models where needed. The sensAI 4.1 CNN accelerators simplify implementation of ultra-low power AI designs by leveraging the parallel processing capabilities, distributed memory, and DSP resources of Lattice FPGAs. The accelerator cores take advantage of the FPGAs' programmable logic to implement low-power NNs including extremely efficient binary neural networks (BNNs) to implement CNNs with extremely low power consumption, in the mW range.

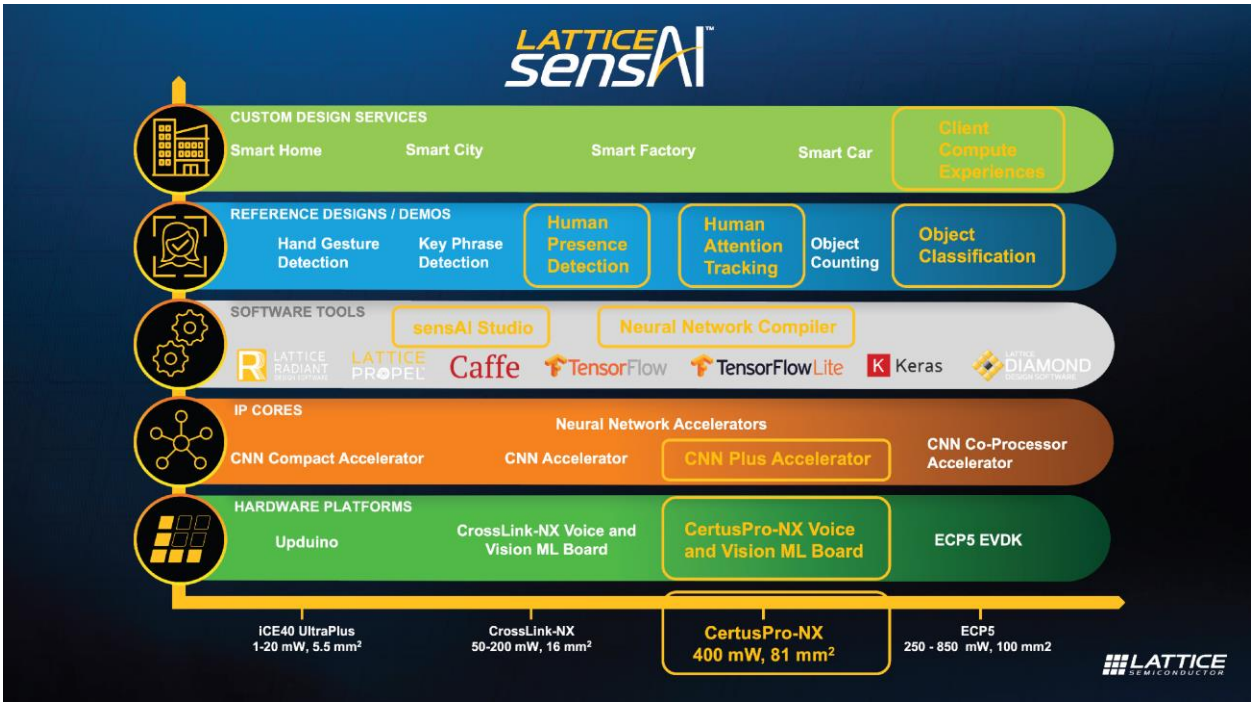

Figure 3 The Lattice sensAI solution stack permits the development of AI/ML devices based on Lattice FPGAs. (Source: Lattice)

Lattice sensAI 4.1 Reference Designs

Lattice FPGAs provide programmable I/O that can be configured to support many different electrical interface standards commonly used for sensor interfaces. The company also offers many hard and soft IP blocks to support different sensor communications protocols. Because FPGAs have long excelled at sensor fusion, the Lattice sensAI solution stack 4.1 is specifically designed to ease the development of AI/ML inference features based on multiple sensors for edge devices, enabling the development of intelligent sensor fusion. The sensAI 4.1 solution stack includes many example reference designs that demonstrate intelligent sensor fusion use cases that can run concurrently to enable deeper context awareness. These reference designs include:

- *Hand Gesture Detection*

This reference design implements a low power AI-based system to detect hand gestures using an IR image sensor. A training dataset, scripts for training using common NN training tools, and an NN model are provided to enable modification.

- *Key Phrase Detection*

This reference design continuously searches for a key phrase utterance using a digital MEMS microphone. Designers can add wake-word capability to systems by updating

the supplied training dataset using deep learning frameworks, such as Caffe, Tensorflow, or Keras. This reference design includes a training dataset, training scripts for common NN training tools, and an NN model are provided to enable modification.

- *Human Face Detection*

This reference design implements CNN-based human face identification using an image sensor and can be retasked to identify other object types by modifying the supplied training database.

- *Human Presence Detection*

This reference design continuously searches for the presence of a human using a CMOS image sensor. AI systems based on this design can detect and locate any object of interest by updating the supplied training model using deep learning frameworks, such as Caffe or Tensorflow. This reference design includes an NN model, a training dataset, and scripts for training using common training tools.

- *Object Detection, Classification, Tracking, and Counting*

This reference design provides examples for object detection, classification, tracking, and counting and includes complete designs with FPGA RTL for Lattice development boards, NN models, a sample training dataset, and scripts to recreate and update the designs.

Obvious and Non-Obvious Edge Applications That Can Employ AI

While the advantages of using AI/ML algorithms to improve the performance of many edge devices, such as autonomous robots, environment controls, and video security cameras, is fairly apparent, other types of edge equipment can also benefit – PCs and laptops, for example. Lattice is working with partners and customers to harness multimodal, intelligent sensor fusion and AI/ML techniques in the never-ending quest to enhance the PC/laptop user's experience and to significantly reduce operating power for laptop PCs and increase battery life by as much as 28% in some applications.

What device characteristics make these non-obvious benefits possible?

PC and laptop usage varies widely over a 24-hour period, with concentrated usage during daytime work hours. However, even during work hours, there are times when PCs or laptops aren't in use. People take breaks and lunch hours, and they usually leave their computers running during those times to keep their place in the various applications they have open on the screen.

Making the PC or laptop aware of its surroundings through intelligent sensor fusion by coupling AI/ML analysis and decision making with the computer's existing sensors –

cameras and microphones – allows the computer to decide when it can power down its display and CPU, and when it should power them back up.

The simplest use of presence detection is to power down the computer when no one is around. Attention tracking allows a computer to dim its screen and activate a lower-power mode when the user looks away from the screen for an extended period of time. A low-power, small-footprint FPGA acting as an intelligent sensor hub can accept input from the computer's sensors and then decide which components to power up given the situation.

Addressing Privacy and Security Concerns

These same capabilities can enhance a computer's privacy and security as well. The computer's built-in conferencing camera can be used to monitor the background behind a user and to detect when someone is looking over the user's shoulder. If the computer is configured for privacy, it might alert the user with a pop-up warning or even dim the screen when someone behind the authorized user appears to be looking at the computer's screen. Note that all inferencing data is kept locally within the FPGA with these solutions. Only metadata is passed to the SoC, which further enhances privacy and reinforces security.

Enhancing the User Experience

AI/ML features can also enhance the computer user's general experience. For example, an AI/ML-based face-framing feature can take advantage of the extremely high resolution of today's built-in, video-conferencing cameras to crop and center the user's image so that it's properly framed for video conferencing sessions. Session participants can even move during a conference while their image remains centered. Similarly, gesture recognition can add contactless operating features to the laptop or PC, or to any other video-enabled IoT device.

Health and Wellness Benefits

Many corporations are now explicitly responsible for their employees' health and wellness, and AI/ML-powered awareness features can help to avoid repetitive stress injuries through pop-up reminders and other measures, making use of the computer's video sensors to ensure that employees actually take the suggested work breaks.

AI/ML applications can also be used to detect a user's posture, which can be another factor influencing repetitive stress injuries. These features, which make use of active sensor feedback, enable the creation of wellness applications that are demonstrably superior to the simple timed alerts currently used in corporate environments to combat stress-related workplace injuries.

All of these AI/ML-enabled features can help vendors create PCs and laptops that are more attractive to corporate buyers, and all of these features can be implemented by

using the sensAI 4.1 solution stack to harness the capabilities of low-power Lattice FPGAs.

Using FPGAs in this manner transcends the sensor interfacing and fusion functions that have long been a hallmark of FPGA development and adds sensor signal analysis and decision making based on proven AI/ML algorithms. Adding AI/ML capabilities allows the FPGA to become a low-power system controller that manages system functions, enhances user experience, and substantially boosts battery life by reducing overall system operating power.

Conclusion: Billions to be served at the Edge

With its multiple low-power FPGA families and the family-spanning sensAI 4.1 solution stack, Lattice is firmly focused on the use of AI/ML technologies in billions of edge devices. Consequently, edge applications are an excellent market to target.

By many estimates, the planet's widespread geographies require tens of billions of edge devices to serve the needs of these myriad edge markets, and those are very attractive unit volumes in the FPGA business – they're great unit volumes in any business. Lattice has squarely targeted these edge applications and markets with the release of its sensAI 4.1 solution stack and its low-power, small-footprint FPGA families. Lattice's sensAI 4.1 solution stack is an innovative development tool for edge applications that allows system developers to create flexible, application-specific, FPGA-based, AI/ML inferencing solutions for a wide variety of markets.