Accelerating Implementation of Low Power Artificial Intelligence at the Edge

A Lattice Semiconductor White Paper

November 2018

The emergence of smart factories, cities, homes and mobile are driving shifts in systems architectures and new applications that require increased intelligence at the edge. Artificial Intelligence/Machine Learning silicon solutions are essential to meeting the requirements for this new generation of AI-based edge computing applications.

Designers building computing solutions on the edge are challenged to meet new requirements for flexibility, low power, small form factor and low cost, without compromising performance and in hotly competitive market conditions. Systems that integrate low power inferencing close to the source of IoT data utilizing lower density FPGAs optimized for low power operation, can meet the stringent performance and power limitations imposed on the network edge with a time to market advantage.

The new Lattice sensAI™ stack, a comprehensive developmental ecosystem, simplifies the task of building flexible inferencing solutions optimized for the edge. With variety of IP, tools, reference designs, and design expertise developers can utilize the ecosystem to get to market quickly with innovative solutions.

Lattice Semiconductor
www.latticesemi.com

Table of Contents

Architectural Shifts & Demand for Increased Intelligence at the Edge

Since the invention of the first computer, identifying the ideal system architecture has been no simple task. Examination of the history of computing shows system architects continually shifting back and forth from centralized configurations where computational resources are located far from the user to distributed architectures where processing resources are located closer to the user. Server-based approaches popular in the 1970s and 80s, for instance, adopted a highly centralized approach to pool computational power and storage capabilities. But that philosophy quickly fell to the wayside in the 1980s and 1990s with the rise of low cost PCs and the emergence of the Internet. In this new architectural model computational tasks were increasingly delegated to the individual PC.

The highly distributed approach built around the PC looked to be on solid ground until the rise of mobility in the form of smart phones, tablets and laptops. Suddenly carrying around your computational hardware and storage resources became a liability. Slowly system architects began moving tasks to the cloud where they could take advantage of its virtually unlimited computational and storage resources, high reliability and low cost.

Organizations are using the cloud to reduce capital costs and more efficiently manage operational and maintenance costs associated with an IT infrastructure. As companies adopt machine learning techniques and introduce higher levels of artificial intelligence, the cloud will play a central role. The coming generations of smart factories, smart cities and smart homes need the cloud to efficiently manage machine vision systems, coordinate traffic patterns and minimize power consumption.

Not all applications will be run from the cloud. Industry experts note that the first indications of another architectural shift from centralized to distributed systems are already apparent. Whether the next transition is coming or not, one thing is clear. Lower latency requirements, escalating privacy concerns and communication bandwidth limitations are driving demand for increased intelligence at the edge. As designers add higher levels of intelligence to applications on the edge, they need systems that response faster to changing environmental conditions. When an autonomous car enters a smart city, for example, it can’t take the time to ask the cloud how to avoid a collision. It must act immediately and make a decision on its own. Similarly, when an AI-based security camera in a home identifies movement in the house, it must use its on-device resources to determine if a break-in is occurring and call the police.

These new applications will require AI/Machine learning-based computing solutions located closer to the source of IoT sensor data than the cloud. How large is this need? Some believe it's huge. Analysts at Gartner estimate that by 2022 up to 50% of enterprise-generated data will be processed outside a traditional centralized data center or cloud (see figure 1).

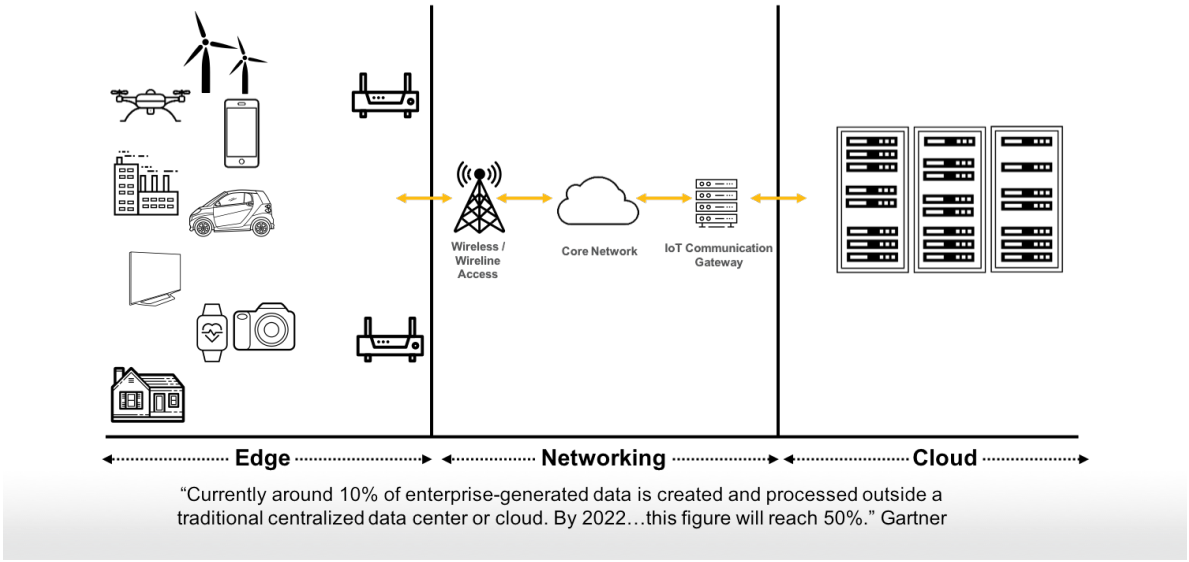

Fig. 1. Rapidly emerging edge computing trend, driven by latency, privacy, and network bandwidth limitations.

Edge Computing Requirements

One of the toughest challenges designers building computing solutions on the edge face is meeting a unique mix of flexibility, low power, small form factor, and low-cost requirements (see figure 2).

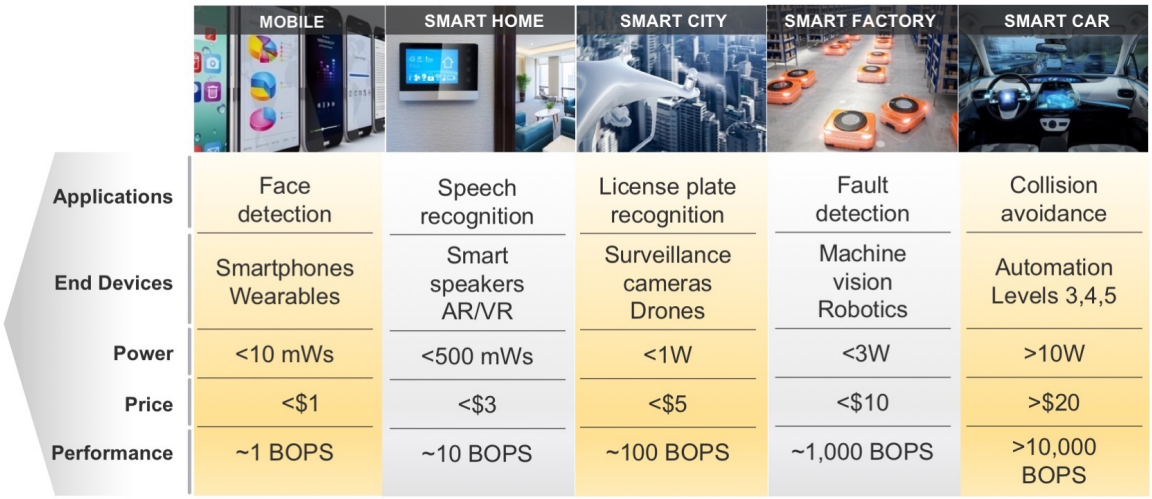

Fig. 2. A new generation of AI-based edge computing applications demand a unique mix of requirements.

How can developers build edge solutions that consume little power and occupy a minimal footprint at low cost without compromising performance? First and foremost, they need silicon that gives them maximum design flexibility. They need silicon solutions that help them take advantage of rapidly evolving neural network architectures and algorithms. They also need silicon that supports a wide range of I/O interfaces. Finally, they need solutions that, through custom quantization, allow them to trade off accuracy for power consumption.

Given the footprint constraints at the edge, designers also need silicon that allows them to build compact high-performance AI devices that deliver excellent performance without violating footprint or thermal management constraints. Cost is also a crucial factor. Any solution must compete against other high-volume edge solutions. Finally, even on the edge, time-to-market rules apply. Those who bring solutions to market first reap a tremendous edge. Accordingly, any perspective solution must have access to the resources designers need to customize solutions and shorten their development cycle – whether it be demos, reference designs, or design services.

FPGA based Machine Learning Inferencing at the Edge

What role does FPGAs play on the edge? Machine learning typically requires two types of computing workloads. Systems in training learn a new capability from existing data. A facial detection function, for example, learns to recognize a human face by collecting and analyzing tens of thousands of images. This early training phase is by nature highly compute intensive. Developers typically employ high performance hardware in the data center to process such large amounts of data.

The second phase of machine learning, inferencing, applies the system's capabilities to new data by identifying patterns and performing tasks. For example, the facial detection function previously discussed would continue to refine its ability to correctly identify a human face as it was put to work in the field. In this phase, the system learns as it works and increases its intelligence over time. Given the many constraints to performing on the edge, designers cannot afford to perform inferencing in the cloud. Instead, they must extend system's intelligence by performing those computational tasks close to the source of the data on the edge.

But how can designers replace the vast computational resources available in the cloud to perform inferencing on the edge? One way is to take the parallel processing capability inherent in FPGAs and use it to accelerate neural network performance. By using lower density FPGAs specifically optimized for low power operation, designers can meet the stringent performance and power limitations imposed on the network edge. Lattice's ECP5 and iCE40 UltraPlus FPGAs meet this need. By using ECP5 FPGAs to accelerate neural networks under 1W and iCE40 UltraPlus FPGAs to accelerate neural networks in the mW range, designers can build efficient, AI-based edge computing applications (see figure 3).

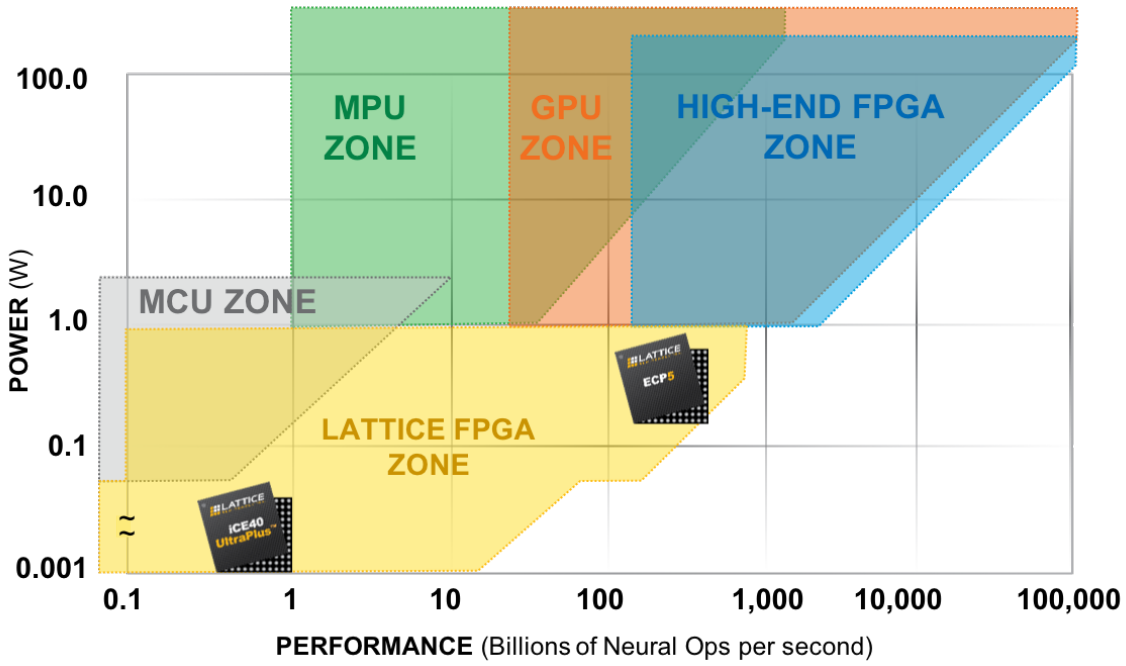

Fig. 3. Lattice FPGA based low power machine learning inferencing, from under 1mW-1W.

Introducing the Lattice sensAI™ Stack

In addition to access to computational hardware, designers need a variety of IP, tools, reference designs, and design expertise to build effective solutions and get them to market quickly.

To help developers address this growing challenge Lattice is now offering a new comprehensive developmental ecosystem based on both the iCE40 UltraPlus and ECP5 FPGA families. Designed to help developers quickly build AI edge solutions for smart home, smart city, smart factory, smart car and mobile applications, the Lattice sensAI stack offers flexible inferencing solutions optimized for the edge.

As Figure 4 below illustrates, by combining modular hardware platforms, neural network IP cores, software tools, reference designs and custom design services from eco-system partners, the Lattice sensAI stack simplifies the task of building flexible inferencing solutions optimized for low power consumption from 1 mW to 1 W, in package sizes starting as small as 5.5 mm2, and priced for high volume production.

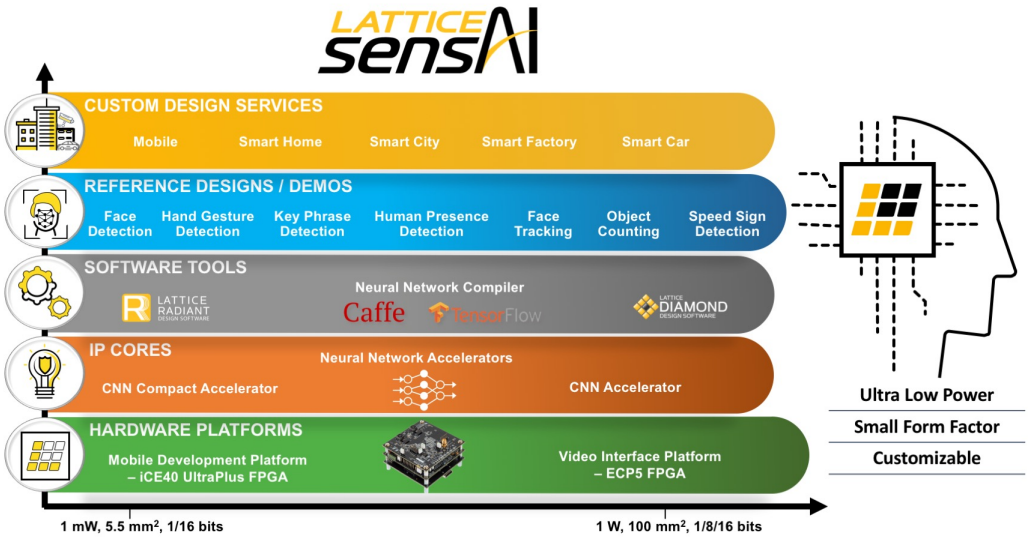

Fig. 4: Lattice's sensAI stack offers developers a solid foundation for developing edge computing solutions.

As depicted in figure 4 above, the Lattice sensAI stack begins with the Lattice hardware kits. To serve this function Lattice offers new modular hardware platforms to speed up prototyping of machine learning designs across a wide range of performance and power requirements. For AI designs consuming just a few mWs of power, Lattice is offering its Mobile Development Platform (MDP) based on its low power iCE40 UltraPlus FPGAs. The MDP features a variety of on-board sensors including image sensors, microphones, compass/pressure/gyros and others. For applications requiring more power but generally operating under 1W, Lattice offers its modular Video Interface Platform (VIP) based on the ECP5 FPGA family. The VIP provides connectivity over an array of interfaces including MIPI CSI-2, embedded DisplayPort (eDP), HDMI, GigE Vision and USB3. One of the first hardware platforms Lattice is offering is its award-winning Embedded Vision Development kit. This modular platform combines a CrossLink input board with an ECP5 processor board and an HDMI output board. With the recent introduction of new eDP and USB3 GigE I/O boards, designers can easily swap out the output boards to support other applications.

On top of the hardware layer Lattice offers new neural network accelerator IP cores that designers can easily instantiate on a FPGA. This lineup of soft IP includes Convolutional Neural Network (CNN) Accelerator IP core optimized for ECP5 FPGAs, and CNN Compact Accelerator IP core optimized for iCE40 UltraPlus FPGA. These fully parameterizable cores provide support for variable quantization, which allows designers to tradeoff accuracy for power consumption.

As depicted in figure 5 below, the Lattice sensAI stack enables users to perform rapid design space exploration and tradeoffs with easy-to-use tool flow. Network training can be done with industry standard frameworks such as Caffe and TensorFlow. The Neural Network Compiler tool then maps the trained network model into fixed point representation, with support for variable quantization of weights and activation. Additionally, the Neural Network Compiler helps analyze, simulate and compile different types of networks for implementation on Lattice's CNN/CNN Compact Accelerator IP cores, without requiring any prior RTL experience. Traditional FPGA design software tools such as Radiant and Diamond are then used to implement overall FPGA design, including the rest of the pre/post processing blocks.

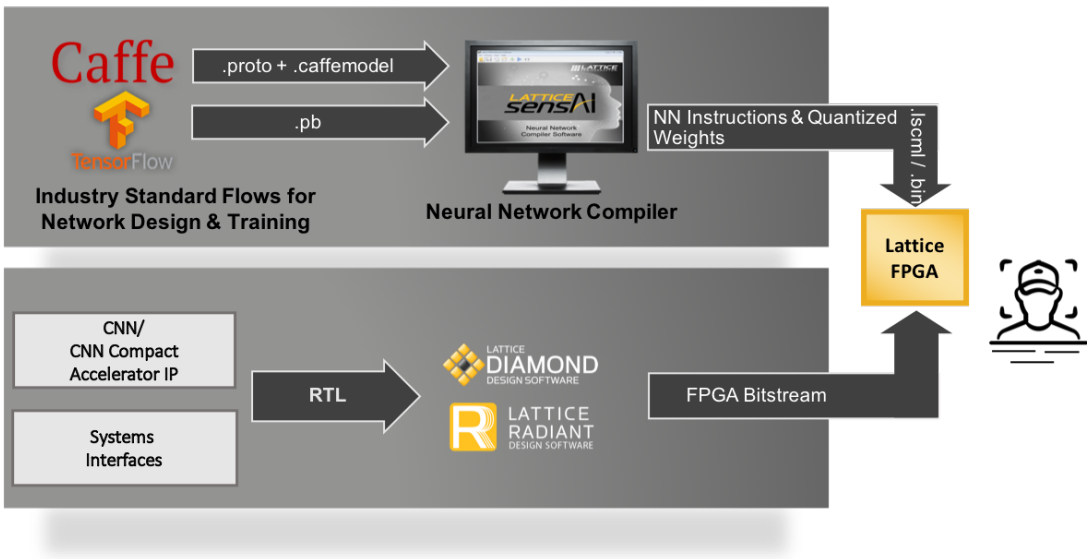

Fig. 5: Rapid design space exploration and tradeoffs with easy-to-use Lattice sensAI tool flow.

To simplify implementation of common AI functions, the Lattice sensAI stack includes a broad range of reference designs and demos that utilize the stack's hardware platforms, IP cores and software tools. Examples include:

Low Power Human Face Detection – This demo includes an accelerated, low power human face detection function designed to work at the network edge using a neural network model. Utilizing binary weights and activations, this iCE40 UltraPlus FPGA based demo helps designers implement face detection under 1mW of power consumption.

Automotive Aftermarket Camera – This demo is targeted at the emerging automotive aftermarket camera market. It illustrates how designers can implement a speed sign detection function using the inherent parallelization of FPGAs. In this demo a

convolutional neural network implemented in an ECP5 FPGA is trained to read passing traffic signs. Once training is complete the camera can detect and display speed limits when it passes a sign.

Converting Voice Commands into Systems Actions – This demo gives designers a blueprint for converting voice commands into system actions. This sub 5 mW key phrase detection function uses a Binarized Neural Network integrated into an iCE40 UltraPlus FPGA. The demo describes how to connect a digital microphone directly to the Lattice inference engine to enable always-on listening with key phrase detection.

Object Detection Solution for Face Tracking – This demo delves into the design of an AI-based object detection solution for face tracking applications. The demo describes the use of a Lattice ECP5-85 FPGA for Convolution Neural Network (CNN) acceleration with eight convolution layers implemented in 8 Neural Network engines. The solution operates standalone based on Lattice's Embedded Vision Development kit and runs at 14fps with 90 x 90 RGB input after power up. Total ECP5 power consumption is only 0.85 W.

Custom Design Services

Development teams often need the unique expertise of design services partners to develop custom solutions. The AI market is no exception. To meet that need Lattice has established relationships with a number of design service partners in areas that range from smart factories, smart cities and smart cars to smart homes and mobile applications. As an example, one of Lattice's certified partners is VectorBlox, a developer of neural network-based inference solutions. As part of their partnership, VectorBlox and Lattice recently implemented a neural network in less than 5000 LUTs of an iCE40 UltraPlus FPGA for a facial detection application. Using an open sourced RISC V soft processor with custom accelerators, the solution dramatically reduces power consumption while shortening response time.

To more rapidly implement inferencing solutions in Lattice FPGAs, developers may need design services experts who can bring knowledge of neural network design and training. Often times this knowledge must be paired with experience in frameworks such as Caffe and Tensor flow as well as traditional RTL design. To simplify the search for this expertise, Lattice offers the Lattice sensAI Design Service Program. This program provides access to design services companies need to accelerate designs that implement deep learning on Lattice FPGAs. These firms provide the expertise to develop and train networks along with the ability to develop RTL for an application.

The third-party firms participating in this program have worked with Lattice to demonstrate their skills in neural network development and training and implementations in hardware.

Conclusion

A revolution in edge computing awaits developers with expertise in AI-based systems. As users look for higher levels of intelligence, demand will grow for systems that integrate low power inferencing close to the source of the IoT dat. The Lattice sensAI stack provides flexible, ultra-low power, small form factor, and production priced solutions optimized for the edge. Lattice ultra-low power FPGAs, backed by an extensive set of hardware platforms, soft IP, design tools, reference designs and third-party experts, offer the surest, fastest path to success.

Learn more: visit www.latticesemi.com/sensAI