

# Effciera: Ultra-Low Power AI Inference Accelerator

## A practical AI implementation using FPGAs on the Edge with compact IP running on the device itself with no cloud connection

### Overview

LeapMind is a pioneer in “Extremely low bit quantization”, a low-power Deep Neural Network (DNN) technology for object detection and classification that minimizes ‘activation and weight’ bits, allowing:

- Reduction of complexity in convolution processing for DNN
- Reduction of power consumption, allowing energy savings and cost reduction
- Increase of area efficiency by flexible size IP core
- Reduction of memory requirements

In recent years, deep learning has become familiar and is used in many Internet of Things (IoT) devices at the Edge. Industry is challenged using artificial intelligence (AI) applications at the Edge because of high total cost of ownership (TCO), long training time, and the need to be dependent on a network connection.

Soichi Matsuda, CEO founded LeapMind to take on this challenge to create a platform of compact and simple deep-learning technologies that are easily accessible to anyone.

### LeapMind

In 2017 and 2018, LeapMind was selected as a Forbes 100 startup in Japan as recognition for their unique technology.

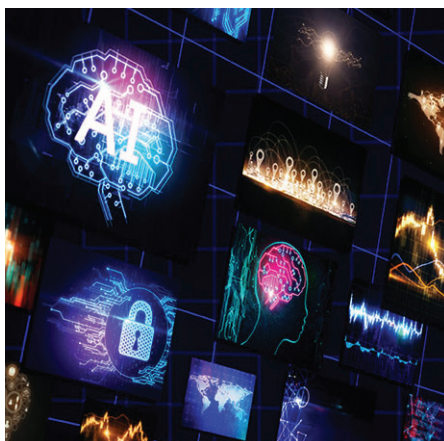
LeapMind’s solution to running deep learning in small edge devices has received attention from significant investors around the globe. LeapMind received investments from Intel Capital, Toyota, and Mitsui, to name a few, raising over \$46 million in total.

LeapMind is a pioneer in “extremely low bit quantization;” their technology uses only 1 bit to represent weight and 2 bits for activation. This level of resolution is the lowest limit that can be used for deep learning, allowing it to perform complex tasks using minimal resources.

### Challenge

The size of the IoT device market is steadily growing. The demand for edge devices that implement local deep learning, which is performing inference processing on the device itself instead of in the cloud, has expanded. LeapMind’s approach eliminates the need for a fast internet connection and latency reliability, while reinforcing security, because it doesn’t send data to the cloud.

This independent Edge appliance approach allows significant reduction of the TCO. LeapMind’s Edge appliance is powered by Intel® Edge-Centric FPGAs.



Energy efficient	Area efficient	Scalable
LeapMind's deep learning insights		



LeapMind's experience with customers and partners identified the following common set of challenges:

**Network Bandwidth**

Inferencing at the edge without sending data across a network is faster and sets the user free from unstable internet connections, a particular issue in remote locations.

**Real-Time Response/Latency**

In some cases, decisions need to be made immediately to prevent accidents, allowing little time for the data to be sent to the cloud and back again for a response.

**Security**

Sending personal and/or private information to the cloud increases the risk of leakage and hacking attacks.

**Resiliency**

Even if the network is down, local processing must work in case of emergency.

**Cost**

The cost of storing image and video data in the cloud, and the cost of sending and receiving the data, are significant.

**Quantization**

In the practical application of deep learning, there are several issues. One example is the problem of computing resources, such as the limited power available on the device restricting the amount of computation available during inference.

Convolutional Neural Networks (CNNs) iteratively repeat matrix operations using floating-point numbers. When these values include a large number of digits (high precision), it becomes significantly more complicated; the number of calculations needed increases, the processing speed decreases, and the power consumption rises.

LeapMind moves away from the standard 16-bit or 32-bit floating-point representations using "extremely low bit quantization" to reduce the weight of the model size.

Processing speed is a tradeoff relationship: the smaller the bit-width is, the lighter (less compute intensive) the model is, but the accuracy of inference deteriorates.

**Solution**

To resolve these challenges, LeapMind brings together technology and knowledge, based on extensive experience, to accelerate implementation of machine learning (ML) in society.

LeapMind engineers developed their solution based on the knowledge gained from the co-creation of ML solutions with more than 150 companies.

The commercialized form of LeapMind's "extremely low bit quantization," called "Effciera" is an ultra-low-power AI inference accelerator IP.

Effciera is used in the construction of an edge AI system that can add analysis functions to vision systems.

The 1-bit matrix multiplication can be implemented with a single XNOR gate, dramatically improving calculation efficiency with greatly reduced need for memory and multipliers.

Intel® Edge-Centric FPGAs with low-power consumption and parallel logic-rich architecture are an ideal fabric to implement Effciera.

The so called "Quantize Aware Training" performs the quantization to a low-bit implementation in the training stage, achieving a significant size reduction while hitting the desired accuracy of the inference system.

Effciera v1.0 was commercially introduced in October 2020. It is described as an RTL for use with both ASICs and FPGAs.

Effciera offers the following four features:

**Energy Efficiency** – By minimizing the volume of data transmitted and the number of bits, the power required for convolution operations is significantly reduced.

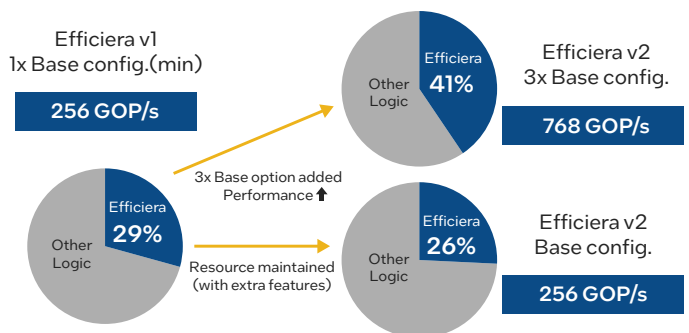
**High Performance** – By reducing the arithmetic logic complexity, the number of operation cycles is reduced and the arithmetic capacity per area/clock rate is improved.

**Smaller Footprint** – By minimizing the number of operated bits, the circuit-area, and SRAM-size per arithmetic logic unit are minimized.

**Scalability** – Since the computing performance can be fine-tuned by adjusting the circuit configuration, it is possible to optimize the configuration and maximize the performance of Effciera specifically to the task being performed.

Edge-Centric FPGAs are the ideal target devices for the small footprint of Effciera.

On a Cyclone® V SoC FPGA (110K logic elements (LE)), the implementation of the smallest 1X "Base Configuration" (single module instantiated) of Effciera v2 occupies approximately 26% of the logic, while the 3X "Base Configuration" consumes nearly 41%.



Using Cyclone V SoC FPGA to implement an "image signal processing (ISP) pipeline" with machine vision algorithms, could be a cost-effective option for large production deployments.

Cyclone V SoC FPGA embeds an Arm Cortex-A9 hard processor system to build a complete machine vision system on a chip with low power consumption.

Another benefit of Intel's Edge-Centric FPGAs is a long-term supply. Since most industrial customers require long life-time support, Intel® FPGAs are an ideal solution for mass production with their established track-record and reliability.

## AI Models

LeapMind provides pre-trained AI models and fine-tuning tools.

Reusing part of existing models to help train a new model, LeapMind helps customers get to market faster by addressing the specific needs of new use cases.

With the fine-tuning tool, it is possible to build new models starting from a smaller dataset and avoid the time-consuming task of creating the model from scratch.

By using these models in combination, the performance of Effciera can be maximized, leading to a faster development phase.

**Object Detection (OD) model** – Effciera's OD Model can capture the location of the detected object from the image, making it useful for a wide range of applications such as traffic flow or intruder detection.

**Anomaly Detection (AD) model** – Effciera's AD model can learn the inspection "ideal" target from dozens of normal images using the embedded camera system and detect "non-ideal" anomalies during the inspection. It can be used for a variety of targets such as product shape or surface inspection. The AD model can also automate visual inspection even when the specifications of the target change frequently or unknown abnormalities are detected.

**Noise Reduction (NR) model** - LeapMind's NR model can convert noisy images taken in ultra-low light into clear images without increasing the size of the image sensor or lens.

These models allow customers to quickly get the most out of Effciera.

## Target Applications

Using Effciera is ideal if the application needs: real-time response, handling of personal information, or communication isn't stable. Example applications include:

- Industrial and agricultural equipment such as construction vehicle, drones, and inspection machines.
- Surveillance cameras and public sector equipment.
- Small manufacturing machines and robots with restrictions on power, cost, and/or heat dissipation.
- Video production systems.

Machine vision is one of the ideal use-cases where AI technology opens further potential in industrial equipment as an intelligent edge device.

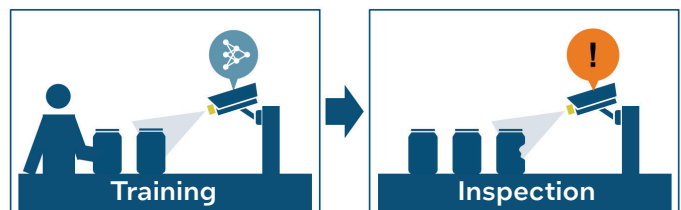
Among machine vision features, LeapMind focuses on the production line in a factory and develops an AD model to detect anomalies for visual inspection.

As a proof of this concept, Effciera has been implemented on the Kondo-Electronics KEIm-CVSoC Development Kit using an Intel® Cyclone® V SoC FPGA. Using an AD model, it demonstrates anomaly detection capability without requiring off-line AI training or even changes to the hardware.

Completely different inspection tasks can be trained on-site within 20 seconds and immediately perform inspection workloads. Inference results are visualized on the screen as a heatmap so that anomalies are easily recognizable. This helps in agile manufacturing (short batch productions) where long training times would be needed using other methods.

## LeapMind Edge AI Anomaly Detection

- Training complete with FPGA equipped edge device
- Learning with only dozens of normal product images
- Easy re-training on mixed production lines



## Summary

Effciera is leading the introduction of an innovative AI solution for edge applications requiring a real-time response but limited in network bandwidth and power. Combining Intel FPGAs and the features of Effciera it is possible to:

- Embed energy efficient and high-performance AI capability in edge applications
- Enable resilient operation of an edge application's intelligent function when the network isn't available
- Reduce the operational turn-around time of switching the inspection target in a factory line
- Maintain long term operation with an FPGA's upgradability, long product lifecycle, and reliability

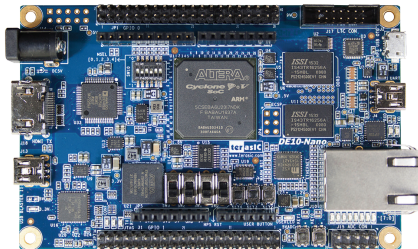
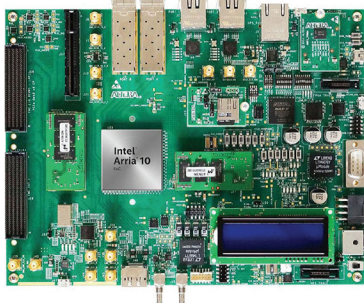
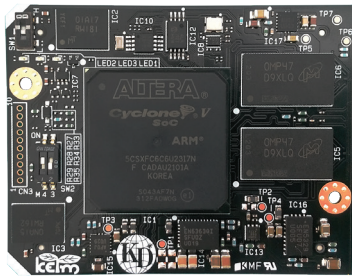
## Where to Get More Information

- [Watch webinar](#)
- [Intel® Cyclone® V SoC FPGAs](#)
- [Contact LeapMind](#)
- [Learn more about Effciera](#)
- [Amazon Web Services \(AWS\) blog](#)
- [Kondo Electronics KEIm-CVSoC](#)

## Solution Ingredients

LeapMind promotes the Effciera FPGA Partner's Program. As part of this program, LeapMind is working to make Effciera compatible with even more mass-produced boards. A wider compatibility shortens development time for deep learning capabilities and enables the mass production of "on-device" AI solutions.

The supported hardware with Effciera and its models are as follows:

<p><b>Intel® Cyclone® V SoC FPGA</b></p> <ul style="list-style-type: none"><li>• <a href="#">Terasic DE10-Nano Kit</a><ul style="list-style-type: none"><li>- Arm <a href="#">Cortex-A9</a> Dual, 110K LEs, 1 GB DDR3</li><li>- Effciera: 1X, 3X, and 4X base configuration</li></ul></li></ul>	 A blue printed circuit board (PCB) populated with various components. The central component is a large black integrated circuit (IC) labeled 'ALTERA Cyclone V SoC'. Other visible components include several smaller ICs, capacitors, and connectors along the edges of the board.
<p><b>Intel® Arria® 10 SoC</b></p> <ul style="list-style-type: none"><li>• <a href="#">Intel® Arria® 10 SoC Development Kit</a><ul style="list-style-type: none"><li>- Arm <a href="#">Cortex-A9</a> Dual, 660K LEs, 1 GB DDR4 + FPGA: 2GB DDR4</li><li>- Effciera: 1X, 3X, 4X, 12X, and 24X base configuration</li></ul></li></ul>	 A green PCB with a prominent silver square component labeled 'Intel Arria 10'. The board is densely packed with various electronic components, including capacitors, resistors, and connectors. A small blue LCD screen is visible on the right side of the board.
<p><b>Intel® Cyclone® V SoC FPGA</b></p> <ul style="list-style-type: none"><li>• <a href="#">Kondo-Electronics KEIm CVSoC SoM</a><ul style="list-style-type: none"><li>- Arm <a href="#">Cortex-A9</a> Dual, 110K LEs, 1 GB DDR3</li><li>- Effciera: 1X, 3X, and 4X base configuration</li></ul></li></ul>	 A black PCB with a large black IC labeled 'ALTERA Cyclone V SoC' in the center. The board features numerous smaller components, including capacitors and connectors, and is populated with various surface-mount components.



Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.