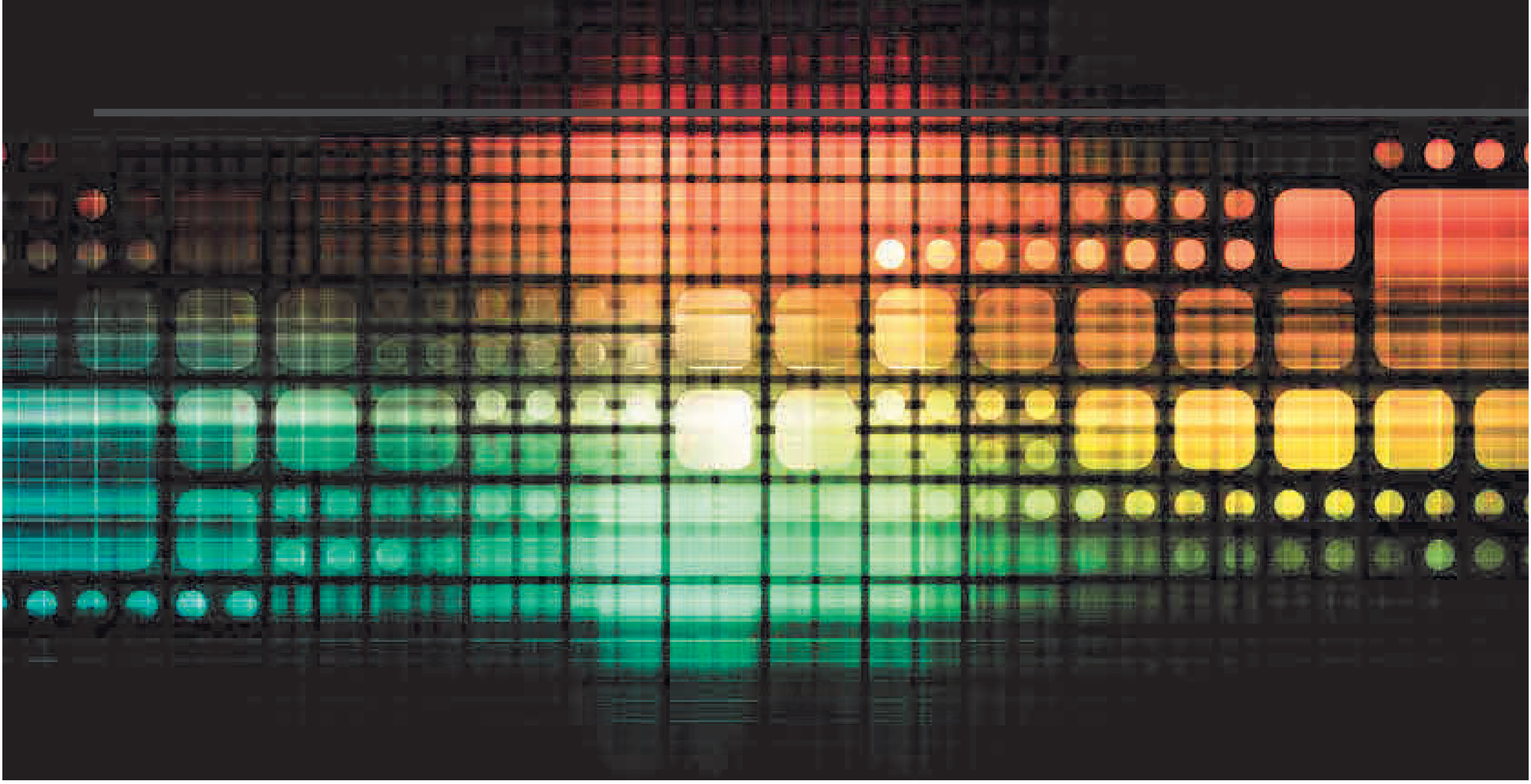
RTC

Real World Connected Systems Magazine

Second Quarter Summer 2020

Extension or no extension: what Class I manufacturers need to know about EU MDR

How Smart IoT Nodes at the Extreme Edge are Bringing Ultra- low Power Options to Embedded Vision Systems



How Smart IoT Nodes at the Extreme Edge are Bringing Ultra- low Power Options to Embedded Vision Systems

A new way of developing SoCs for smart IoT nodes could offer a more flexible way of harnessing neural networks for ultra-low-power embedded vision.

By Semir Haddad, Senior Director of Product Marketing, Eta Compute

Introduction

Historically, OEMs offering embedded vision have been obligated to rely mainly on cloud-based AI. Increasingly though, the cost-power-performance mix for these appli-cations can be balanced more favorably by using powerful application processors at the IoT endpoints.

Even better, solutions are now becoming available that can bring embedded vision to smarter IoT nodes that integrate micro-controllers (MCUs), enabling low-power vision applications like person detection, wake-on-approach, driver awareness and robotics.

This article will explain how it's now possible to perform machine vision in IoT endpoints in the milliwatt range as a result of an approach to System on Chip (SoC) design that moves away from conventional MCU architecture while retaining the cost and flexibility benefits of MCUs.

Why Do We Need Smart IoT Nodes?

Let's begin by considering a smart IoT node with a role in, for example, a person detec-tion solution where it would acquire data from a camera sensor, perform signal pro-cessing and feature extraction and run a machine learning algorithm.

A factor broadening the market for person detection has been the General Data Pro-tection Regulation, or GDPR, with its prohibition on capturing images of people without permission. For example, an image of goods on a shelf, needed for inventory manage-ment, cannot be part of inferencing if that image includes humans.

To compete successfully OEMs offering people detection and other machine learning solutions that use inferencing must meet customer demands to tamp down data stor-age and communication costs while assuring security and privacy.

Inferencing at the smart IoT node is advantageous for a number of reasons. Storage costs shrink because only actionable data is sent to the cloud. Also, the price paid to the network operator decreases corresponding to data being able to stay at the intel-ligent endpoint for inferencing rather than traveling to and from the cloud. In addition, data that need not travel to the cloud for inferencing does not risk security and priva-cy assaults or breaches during transfer. Latency is another bugaboo that would be avoided as would the negative impact on real-time capability caused by sending da-ta to the cloud unnecessarily.

A Power vs Performance Conundrum

However, these boons to cost reduction, security, and privacy have to this point not been as available as they could be. Smart IoT nodes are typically battery powered or rely on a limited power source, sometimes using energy harvesting. If the person detec-tion system in our example tried to rely just on traditional MCUs, as performance de-mands rose, power consumption would rise beyond the power capacity of the node. This would not be a suitable solution.

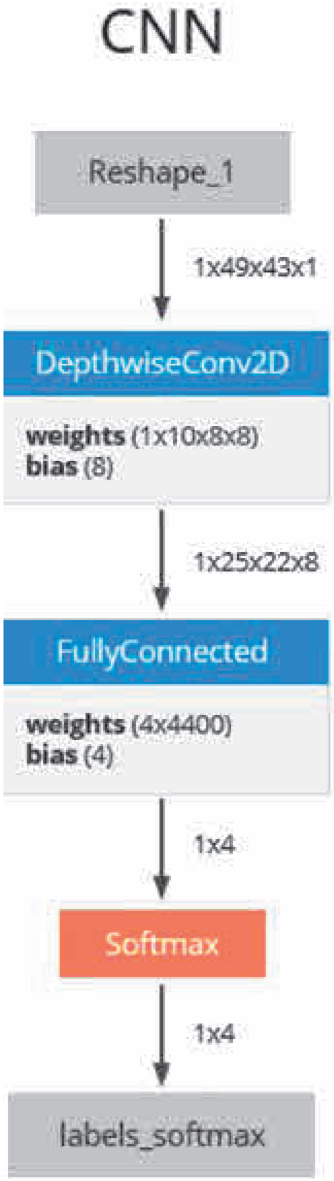


Figure 1: CNNs make heavy use of MMAC, making traditional microcontrollers inef-ficient for em-bedded vision systems.

For example, note the Convolutional Neural Network (CNN) on Figure 1. Widely used in machine vision for object classification, CNNs comprise several layers. Convolutions and fully connected are the most cycle-intensive operations, with heavy matrix multiply-accumulate (MMAC) use that MCUs are ill equipped to perform.

Attempts to get around the power-rise-with-higher-performance dilemma while cling-ing to the traditional microcontroller idea have led to no shortage of various neural networks for microcontrollers. But until now bringing out production-grade solutions that overcome performance and power constraints to create a smarter IoT node has prov-en elusive.

The Steps to Machine Vision in the 1mW Range

Taking on Workloads in Any Combination

So, how to significantly improve upon the efficiency possible using direct implementa-tion of neural networks on a standard microcontroller? One step is recognizing that smart IoT nodes face three workloads: a procedural one, another for digital signal pro-cessing, and a third for machine learning, making heavy use of MMAC operations.

To target each workload's unique demands, in the solution described here an Arm Cortex-M CPU handles the procedural load, while a dual MAC 16-bit DSP serves signal processing and machine learning needs. With this approach that takes full advantage of DSP benefits, doubling or even tripling neural network calculation performance is feasible (Figure 2).

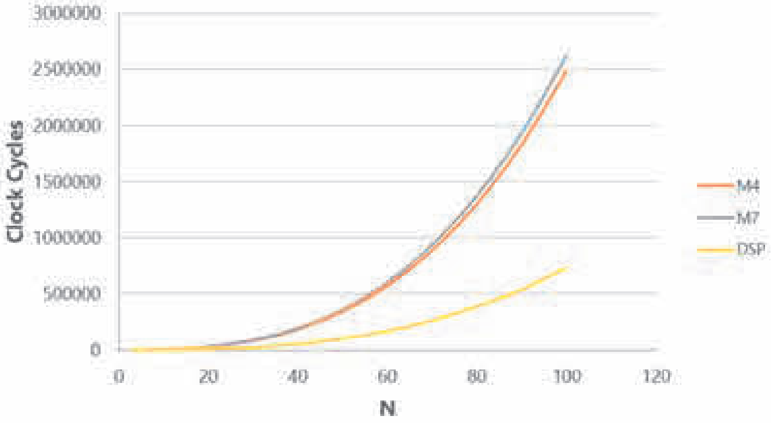


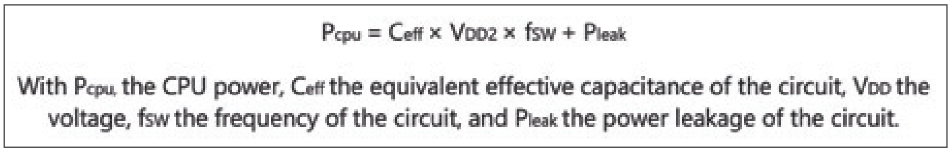
Figure 2: Matrix multiply (NxN) benchmark. DSP architecture strengths, including dual memory banks, zero loop overhead, and complex address generation, allow for over three times more efficiency at neural network calculation compared to MCUs with Cor-tex-M4 or Cortex-M7 cores.

The Hybrid Multicore architecture used with this approach can tackle workloads in any combination, including network stacks, RTOS, digital filters, time-frequency conversions, RNN, CNN, and traditional artificial intelligence like searches, decision trees, and linear regression.

Lowering Power Consumption to the mW Range

Developing a hybrid multicore architecture that enables neural networks to run more efficiently by accounting for the differences in various workloads is one component needed for more power-efficient and higher performing embedded vision systems. An-other is applying a new patented design technology to decrease power.

Developing a new way to lower power consumption so that OEMs could benefit from MCU strengths demanded a fresh look at the power and voltage relationship. The two relate to one another as expressed in Equation 1, such that lowering the voltage is an option for reducing power. Alas, the undesirable side effect – maximum frequency lowers when voltage is lowered – make it challenging to implement variable voltage schemes.



$$P_{cpu} = C_{eff} \times V_{DD2} \times f_{SW} + P_{leak}$$

With P_{cpu}, the CPU power, C_{eff} the equivalent effective capacitance of the circuit, V_{DD} the voltage, f_{SW} the frequency of the circuit, and P_{leak} the power leakage of the circuit.

Equation 1

Past attempts to keep power low while obtaining high performance include Dynamic Voltage Frequency Scaling (DVFS). However, DVFS works most effectively only within a voltage range of a few hundred mV and only for a handful of pre-defined discrete voltage levels. Another option, but one that is difficult to implement, is sub-threshold design.

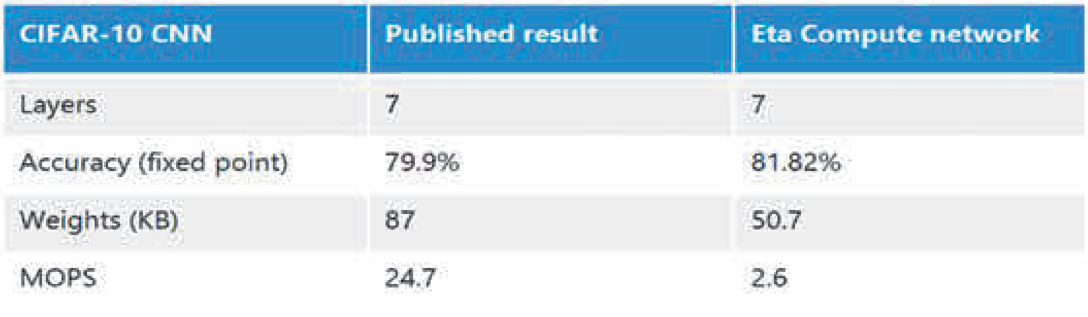
Now there is a new technology and approach empowering OEMs to take on the power-performance dilemma while preserving the benefits – simple product design and low cost - microcontrollers bring to embedded vision system designs.

A patented technology, Continuous Voltage and Frequency Scaling (CVFS) employs a scheme whereby the logic is self-timed, allowing each device to adjust voltage and frequency automatically, on a continuous scale. With CVFS, the SoC always operates at the most efficient voltage.

Now consider that the lower frequency and more efficient cores that a hybrid multicore architecture, discussed above, brings to the equation. That architecture magnifies CVFS advantages.

Beyond that, published neural networks created with no prioritization for the specific needs of the extreme edge can be optimized, capitalizing on innovative design tech-niques such as those Eta Compute and its partners are developing.

For example, Eta Compute optimized a CNN for the CIFAR10 dataset. Table 1 shows the results.



CIFAR-10 CNN	Published result	Eta Compute network
Layers	7	7
Accuracy (fixed point)	79.9%	81.82%
Weights (KB)	87	50.7
MOPS	24.7	2.6

Table 1: Note that while preserving similar accuracy, operations were divided by 10 and weights size by 2.

Results

Customizing for the Extreme Edge

Design techniques and optimization efforts that are deliberately tailored for the unique needs of the extreme edge are yielding results. For example, Eta Compute now offers a production-grade neural sensor processor, the ECM3532 (Figure 3).

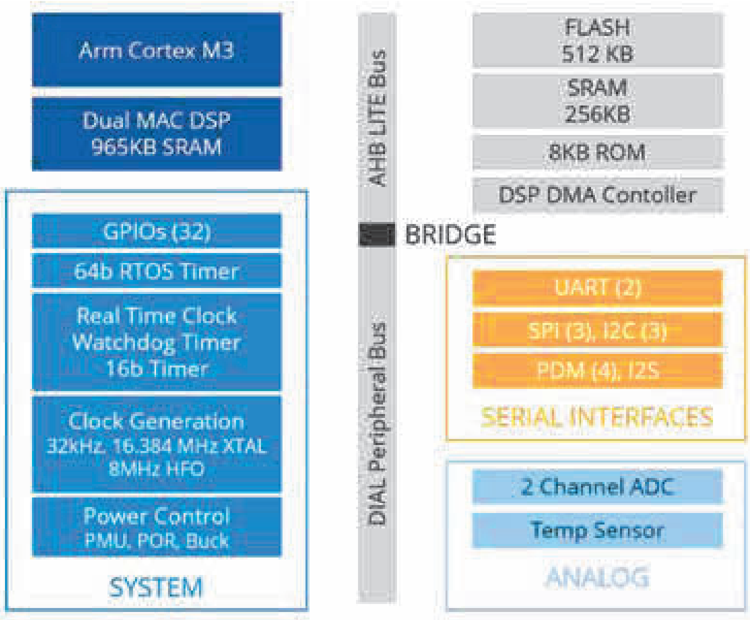


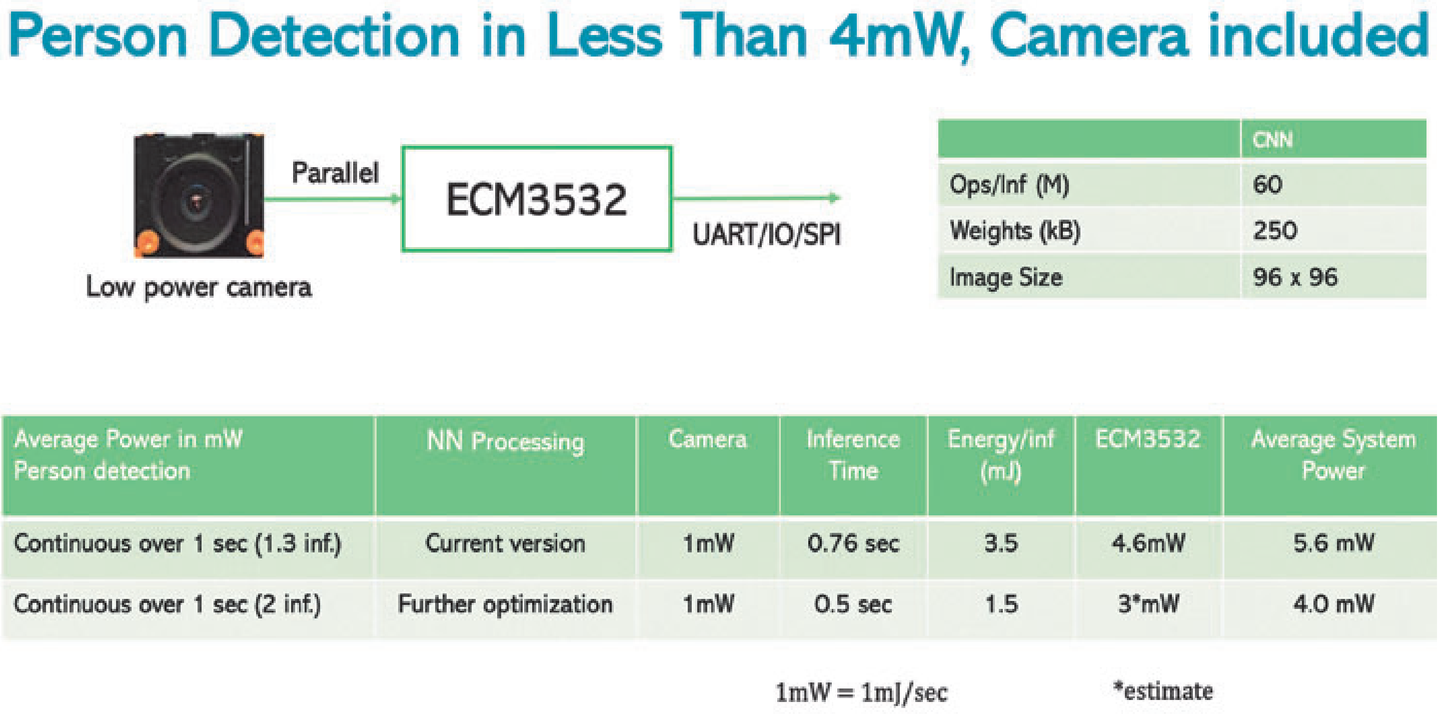
Figure 3: The Eta Compute ECM3532 neural sensor processor is a System on Chip (SoC) comprised of an Arm Cortex-M3 processor, an NXP CoolFlux DSP, 512KB of Flash, 352KB of SRAM, and supporting peripherals.

When the ECM3532 neural sensor processor powered a person detection model in a recent test (Figure 4), the test showed that the SoC could run the algorithm with an average power of 4.6mW, while the average system power was 5.6mW, including the camera – for an inference time of 0.7s (1.3 inference per second). We estimate that with further optimization, an average system power of 4mW can be reached with 2 infer-ences per second.

Just the Beginning

The capabilities that accrue from ultra-low power operation at the extreme edge can benefit applications that include people detection, as we have seen above, but also gaze detection for safer semi-autonomous driving, agricultural applications requiring accurate animal detection, people counting …the list goes on.

Free from the inefficiencies and latency associated with solutions where the heavy lift-ing must happen primarily in the cloud, OEMs will be able to offer customers in industrial, automotive, and consumer markets the leaps forward AI promises.



Person Detection in Less Than 4mW, Camera included

Parallel
ECM3532
UART/IO/SPI
Low power camera

	CNN
Ops/Inf (M)	60
Weights (kB)	250
Image Size	96 x 96

Average Power in mW Person detection	NN Processing	Camera	Inference Time	Energy/inf (mJ)	ECM3532	Average System Power
Continuous over 1 sec (1.3 inf.)	Current version	1mW	0.76 sec	3.5	4.6mW	5.6 mW
Continuous over 1 sec (2 inf.)	Further optimization	1mW	0.5 sec	1.5	3*mW	4.0 mW

1mW = 1mJ/sec

*estimate

Figure 4: The table shows results for a person detection system based on the Google Person Detection Model with 29 layers using an Eta Compute ECM3532 neural sensor processor and a Himax model HM01B0 low-power camera at 2.8V.