csem

FETA: a Flexible Low-Power AI/ML Accelerator for Time Series Signals

Stéphane Emery

CSEM White Paper • October 2024

ASICs for the Edge

Enabling time-series AI/ML functionalities in any device, at negligible power consumption

The large variety of data that is acquired by sensors on mobile, wearable, and IoT devices has enabled numerous new applications such as long-term medical monitoring, fitness tracking, and voice control. ML algorithms such as neural networks (NNs) are often used for processing time-dependent sensor data (time-series) from these sensors. However exploitation in edge devices is still limited by the inefficient processing of the vast amounts of sensor data.

Today, very few portable devices embed ML features and the rare ML tasks that are performed are often limited. Devices with tight power budgets are rarely enabled with ML functionalities at all. The numerous operations required by ML algorithms are in fact typically offloaded to the cloud, at the cost of power-hungry radio communication, long latency, and privacy risks. Thus, the design of ultra low-power NN accelerators is key to enable ML features in any battery powered device.

The development of optimized, yet flexible accelerators for NNs can unlock from 2x to 10x savings in power consumption. Thanks to the design of these circuits, the execution of computing-intensive algorithms can be made possible for any portable device and create unprecedented use-cases for edge devices.

Table of contents

FETA: a flexible RNN accelerator for any time-series signals

FETA is a digital accelerator that parallelizes the computation of recurrent neural networks (RNNs) which are commonly used in time-series applications as they are capable of identifying temporal correlation within the input signal. The parallelized computation is enabled by eight (8) processing elements (PEs) which compute multiply-and-accumulate (MAC) operations that are the most common calculations required by RNNs. Quantization of the computed neural networks is programmable and it can be equal to 4, 8, or 16 bits.

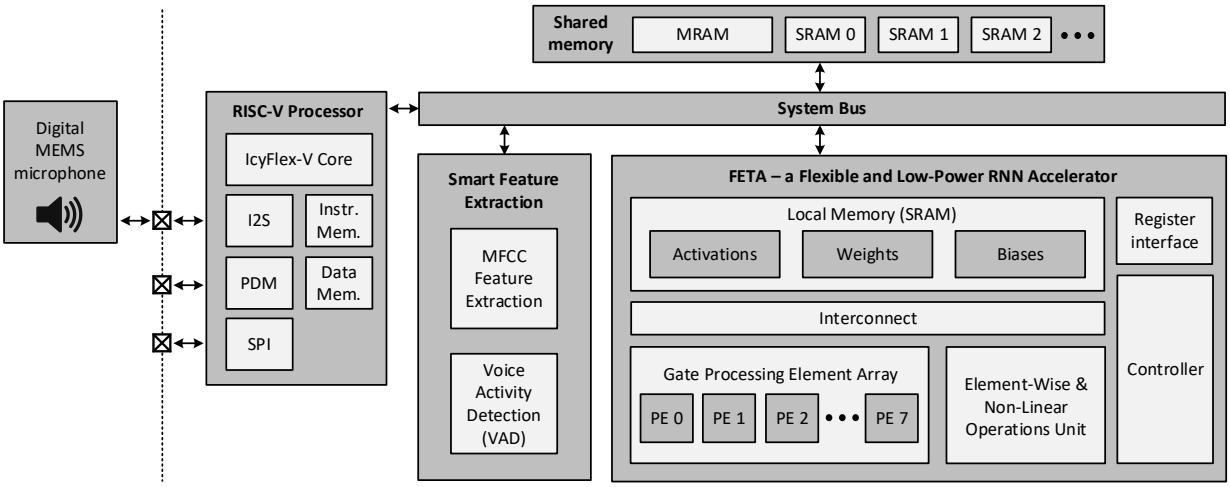


Figure 1: FETA accelerator and smart feature extraction block.

FETA is designed for flexibility to support a wide range of ML time-series applications, such as biomedical signals monitoring, predictive maintenance, and voice control. To achieve this, several parameters can be programmed, such as the number of RNN layers, the type of RNN cell (LSTM or GRU), and the non-linear activation function, such as hyperbolic tangent, sigmoid, and rectified linear unit (ReLU). Also, fully-connected deep neural networks based on fully-connected layers are supported. The FETA accelerator supports programmable quantization of the neural networks for either 4, 8, or 16 bits to save power consumption and unlock memory space for more weights and activations while ensuring a sufficient computational precision.

A smart feature extraction block is also available and designed for voice applications. This block extracts MFCC-based features from the raw voice data collected by the ADC and provides the essential information (the features) to the FETA accelerator. A voice activity detection (VAD) block can monitor either the raw voice data or the extracted features to enable the RNN acceleration by FETA only when voice is detected while keeping the accelerator in a low-power inactive state in the remaining time. The feature extraction block can also be bypassed to feed FETA with any type of data (e.g., raw audio data, output activations from another neural network, etc…).

Low-power operation

RNNs are inherently energy efficient as they operate in streaming mode. The result from an inference is obtained by recomputing the state of the network based on the new input data as well as on the previous state of the network. This operation mode is different from other types of neural networks (e.g., convolutional NNs) where the inference result is generated only if the whole data history is provided, therefore often with a higher computational cost.

Design optimizations have also been applied at circuit level. The accelerator relies on an energy-efficient data flow with spatial reuse of activations to minimize power-hungry memory accesses. According to this data flow, the same activation data is broadcasted to all the parallel processing elements (PEs) when

computing an inference. The choice on having eight parallel PEs maximize the energy benefits from data reuse and limits the leakage power consumption in logic.

Also, the data width of the SRAM units in the local memory system is matched to the equivalent data width of the parallel PEs. In the described accelerator, 128-bit wide SRAMs can provide data to eight PEs with 16-bit data width in a single access. This design choice reduces the read-access energy consumption per bit, especially for the memory units that store weights. Still, FETA can also use the system bus to access shared memory spaces where larger neural networks can be mapped.

A memory system is also available within the accelerator and it can be used to store both parameters and activations of the computed neural network for a low-power mode of operation. In this mode, the NN acceleration solely operates with the local memory (L1) to avoid power-hungry data transfer from/to shared memory systems (L2). The local memory system and the logic in the accelerator are implemented with low-leakage SRAMs and low-leakage standard cells, respectively. This design choice reduces the static power consumption with a latency penalty that can be accepted in time-series applications due to the slow nature of the signals (e.g., voice).

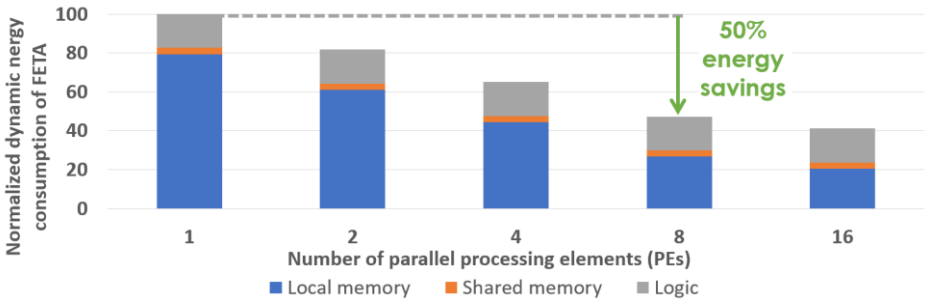


Figure 2: Energy benefits from data reuse with parallel PEs in FETA

Hierarchical processing

Battery-powered always-on devices are actually inactive most of the time, therefore power must be saved in sleep state and performance should be offered only when needed. In this context, a smaller neural network could be continuously computed for smart and low-power wakeup (e.g., single wakeword) that would trigger the on-demand execution of a second larger neural network (e.g., set of commands). Energy-constrained market products offer this flexibility only with costly reconfiguration (i.e., large power consumption and latency due to memory update). Instead, FETA can store an ultra-low-power neural network in the local memory and a more performant network in the shared memory to schedule their execution with hierarchical processing for large system energy savings.

In FETA different sleep modes were optimized by taking advantage of the embedded magnetoresistive memories (MRAMs). During extended sleep states, the weights of the on-demand NN can be stored in such non-volatile memory to power down the SRAM in the shared memory and reduce the overall leakage power consumption.

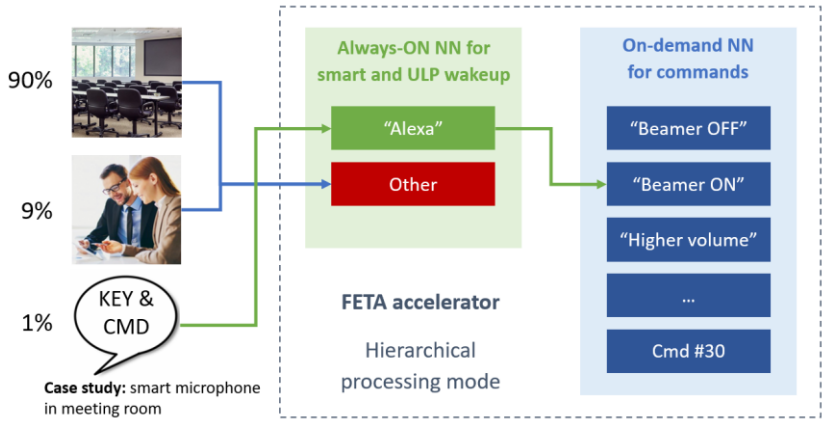


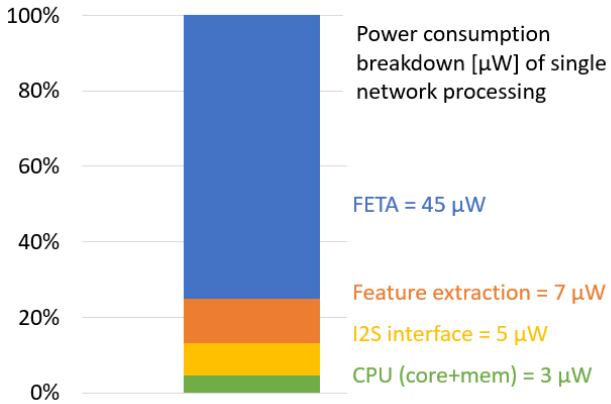
Figure 3: Hierarchical processing in FETA based on always-on and on-demand neural networks

Example of High-Performance System based on FETA

In recent years, several systems focused on audio have been developed. They perform several tasks, such as Voice Activity Detection (VAD), Keyword Spotting (KWS) and form of speech enhancement, such as noise cancelling and speaker identification. The performance of a CSEM system-on-chip (SoC) based on the FETA accelerator is depicted in Figure 4. The FETA SoC provides a low power consumption solution, a flexibility in terms of supported NNs and computing precision and a high accuracy required for a keyword spotting (KWS) application.

CSEM SoC with FETA	
Blocks	SPI, I2S, PDM, CLK gen, Hold Tank, FE, NN, Icyflex-V
NN topology	LSTM (118 cells) + FC (10 neurons)
KWS accuracy (Google 10-word dataset)	95%
Computing precision	4 / 8 / 16 bits
Core supply voltage	0.65 V
Process node	22 nm
Area	
FETA	0.2 mm^2 (logic) 0.1 mm2 (64 KB memory)
Shared memory (1 MB)	1.3 mm^2
Single network processing	
Task	100% of time @ 10-words
Power consumption	60 µW
Memory for parameters	64 KB (local L1 SRAM) 512 KB (shared L2 SRAM)
Hierarchical processing	
Task	99% of time @ 2-words KWS 1% of time @ 20-words KWS
Peak power consumption	20 uW @ 2-words KWS 343 uW @ 20-words KWS
Average power consumption	23 µW
Memory for parameters	2 KB (local L1 SRAM) 1070 KB (shared L2 MRAM)

(a)



(b)

Figure 4: (a) Features of a SoC based on FETA(b) breakdown of the energy per inference in the system

Ready for integration

The FETA accelerator is part of an IP library for the acceleration of edge AI/ML. This library offers a wide selection of hardware IPs for the design of modular and flexible SoCs that enable end-to-end inference on miniaturized systems. Available IP categories include ML accelerators, dedicated memory systems, the RISC-V based 32-bit processor core icyflex-V, and peripherals.

These ML accelerators enable parallel computing for dedicated ML tasks, from computer vision to time-series signals classification. The available memory systems are optimized for the accelerators and they are based on either SRAM, register files, or NVM. Thus, the best matching storage solution can be

chosen based on the tradeoff among power consumption, memory access, and storage density. A wide range of peripherals enable seamless integration with many external devices.

Tailoring the offered solutions to customers' needs is our priority at CSEM. These IPs often allow for design customization and flexible programmability (e.g., for size and precision). The modular nature of this IP library allows for fast and simple integration in any system. A software stack is also available, with firmware examples (e.g., face detection) as well as support for common ML flows and formats (TensorFlow, ONNX, PyTorch, Caffe).

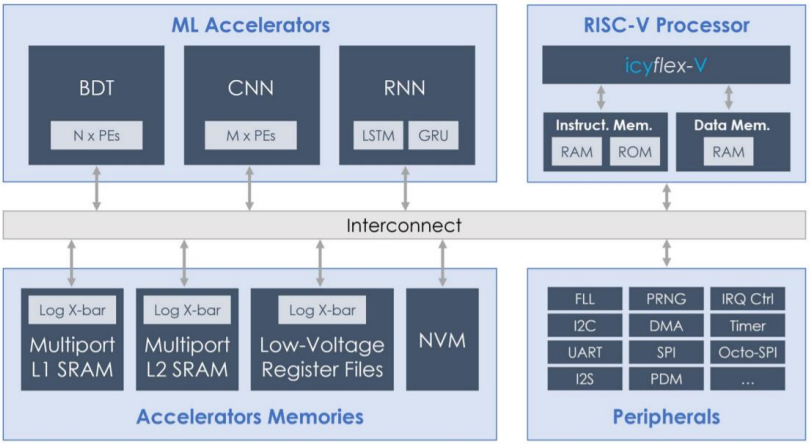


Figure 5: The CSEM IP library for the acceleration of edge AI/ML