Flexible and Scalable Acceleration Techniques for Low-Power Edge Computing (and More)

Francesco Conti *† and Luca Benini *†
* Energy-Efficient Embedded Systems Laboratory – University of Bologna (Italy)
† Integrated Systems Laboratory – ETH Zurich (Switzerland)

Next-generation sensors will extract from the environment an unprecedented amount of sensory data, due to the availability of more and more novel sensors capable to extract more information within an ever-decreasing energy budget. The sheer size of the compound amount of data makes it impractical to transfer, collect and analyse all of it using well-known data mining analytic pipelines - especially for battery-limited or energetically autonomous Internet-of-Things sensor nodes, and for higher-end embedded devices with strong real-time requirements, such as those found in autonomous vehicles.

A proposed solution to this is the paradigm of edge computing, where at least part of the computation necessary to extract semantically relevant information out of raw data streams is performed directly on the sensor nodes. For IoT nodes, low-power microcontrollers currently on the market lack both the flexibility and the computing power to perform much more than very naïve data analytics schemes, whereas complex but successful algorithms such as those based on machine learning are entirely out of reach. We propose a platform and methodology to to perform significant data analytics directly at the sensor's edge within a power budget of few tens of milliwatts in active mode, compatible with the constraints of IoT devices.

Our technique is based on the PULP (Parallel Ultra-Low Power) platform, a small-scale cluster of simple in-order RISC cores coupled with a shared L1 scratchpad and extended with specialized computing engines able to further boost the energy efficiency of particularly critical workloads. The technique is based on cluster-level coupliong of accelerators As a case study, we bring the PULP-based *Fulmine* chip [1], fabricated in 65nm technology, which couples four cores with two engines dedicated respectively to Convolutional Neural Networks and AES security. *Fulmine* is able to perform complex CNN-based workloads within a 15mW power envelope.

This same methodology for hardware acceleration can be scaled or tailored to a diverse set of targets, ranging from low-power ASICs to FPGAs for high-performance embedded systems. In particular, we show-case how the same methodology was used in the FPGA scenario to develop *Neuraghe* [2], a Zynq-based accelerator for CNNs. Finally, we show how the same architectural paradigm can be upscaled to bigger systems targeted at multiple clusters and more challenging applications such as CNN training, as in the *Neurostream* platform [3].

[1] *F. Conti et al.,* An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics, *IEEE Transactions of Circuits and Systems I,* https://arxiv.org/abs/1612.05974

[2] *P. Meloni et al.,* A high-efficiency runtime reconfigurable IP for CNN acceleration on a mid-range all-programmable SoC. *Proceedings of ReConFig 2016,* http://ieeexplore.ieee.org/abstract/document/7857144/

[3] *E. Azarkhish et al.,* Neurostream: Scalable and Energy Efficient Deep Learning with Smart Memory Cubes, https://arxiv.org/abs/1701.06420