ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

XXXIV Cycle

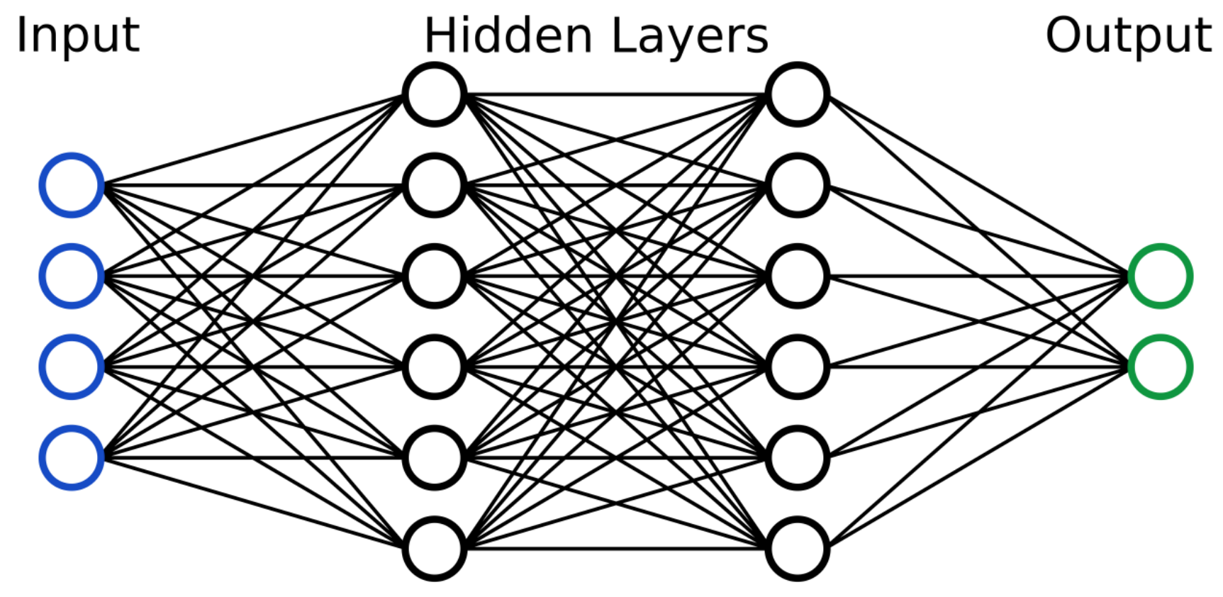
On-chip Machine Learning: Hardware Accelerator to support Deep Learning in Embedded Architecture

Maurizio Capra

Supervisor: Prof. Maurizio Martina

Research context and motivation

- **Machine Learning (ML)** is permeating many aspects of our society, drawing the attention and unleashing new powerful opportunities that go beyond all expectations from just a couple of years ago. In the specific, **Deep Learning (DL)** enables a large variety of **brain-inspired** applications such as image/speech recognition and object detection.
- However, constant data streams mixed with challenging algorithms make DL applications **computation-hungry**. This represents a barrier for the **Internet-of-Things (IoT)** development since many of its elements are based on limited resources.
- Moreover, being **Deep Neural Networks (DNNs) training** a time-consuming and power-hungry task, it is typically performed on servers, faraway from nodes who collect data for the training set.



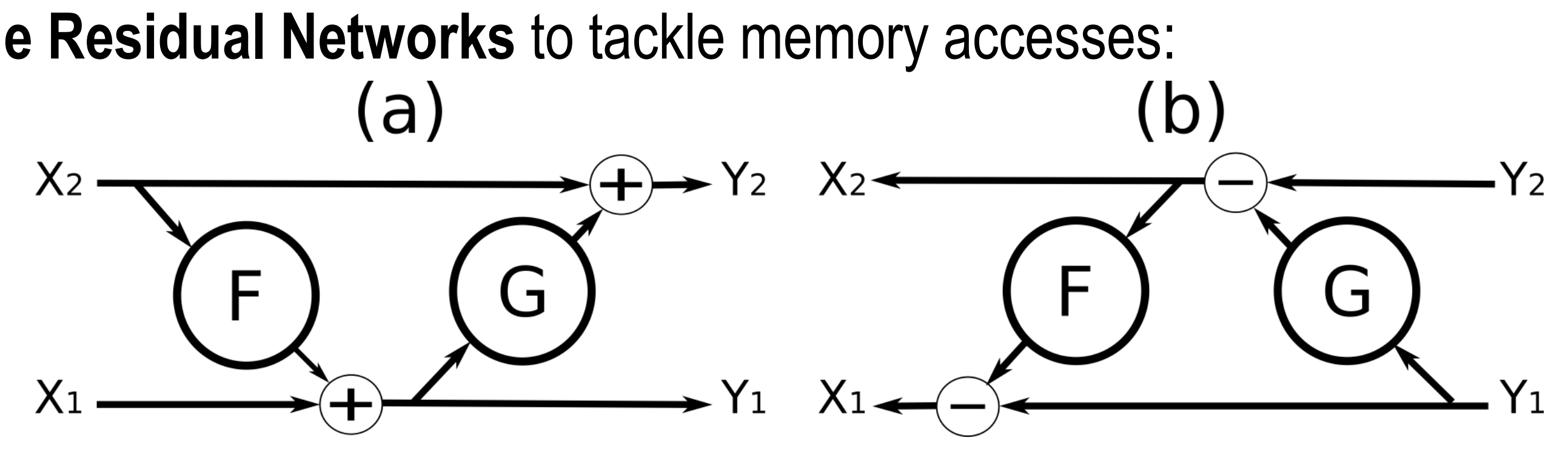
- In this scenario, arises the demand for suitable architectures able to handle both the computation and power demands. **Hardware accelerators** can play a fundamental role in enabling inference and training sessions on **Application Specific Integrated Circuit (ASIC)** -scale devices.

Addressed research questions/problems

- **Billions of multiply-and-accumulate (MACs)** occur inside a net (Inception V4, ResNet18), and each MAC needs to **access the memory** to extract data. Since fetching data is one of the most expensive tasks in terms of **energy**, memory accesses represent a **limitation for battery-powered devices**.
- Training requires to go back and forth through a net, fetching weights multiple times.
- In a DNN, according to the activation functions used (for example ReLU) **many of the weights or activations are null.** Is it possible to avoid negligible operations that involve zero values?

Adopted methodologies

- **Reversible Residual Networks** to tackle memory accesses:



(a) the forward, and (b) the reverse computations of a residual block.

(a) $Y_1 = X_1 + \mathbf{F}(X_2)$ (b) $X_2 = Y_2 - \mathbf{G}(Y_1)$

$Y_2 = X_2 + \mathbf{G}(Y_1)$ $X_1 = Y_1 - \mathbf{F}(X_2)$

- **Continual learning** and regularization techniques to part of the transfer training on the device;
- **PyTorch** Framework allows developing effective DNN training algorithm thanks to its **dynamic computation graph**.

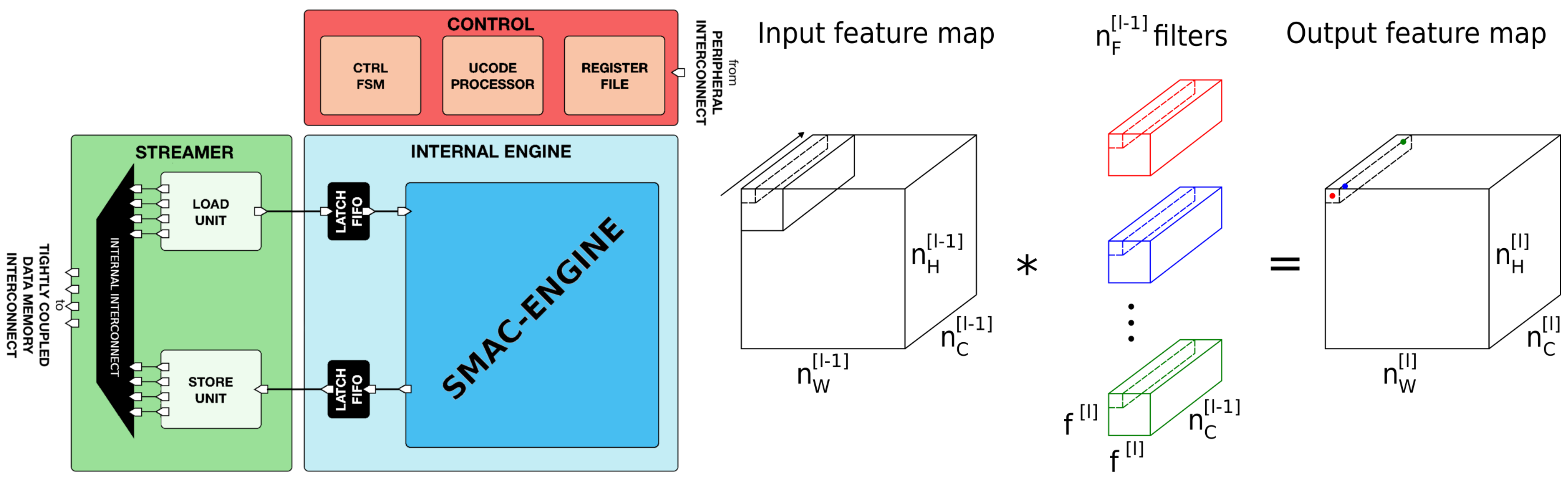
Submitted and published works

- Capra M., Peloso R., Masera G., Ruo Roch M., Martina M., "*Edge Computing: A Survey On the Hardware Requirements in the Internet of Things World*". Future Internet 2019, 11, 100
- Capra, Maurizio; Barone, Elena; De Matteo, Mario; Puppi, Martina; Garg, Ravin, "*Controlled microalgae culture as a reactive nitrogen filter: from ideation to prototyping*". CERN IdeaSquare Journal of Experimental Innovation, v. 3, n. 1, 2019, pp. 27-32
- Erik Anzalone, Maurizio Capra, Riccardo Peloso, Maurizio Martina, and Guido Masera, "*Low-power Hardware Accelerator for Sparse Matrix Convolution in Deep Neural Network*", WIRN, Vietri sul Mare, 2019
- Luigi Sole, Riccardo Peloso, Maurizio Capra, Massimo Ruo Roch, Guido Masera, and Maurizio Martina,"*VLSI architectures for the Steerable-Discrete-Cosine-Transform (SDCT)*", Applepies, Pisa, 2019

Novel contributions

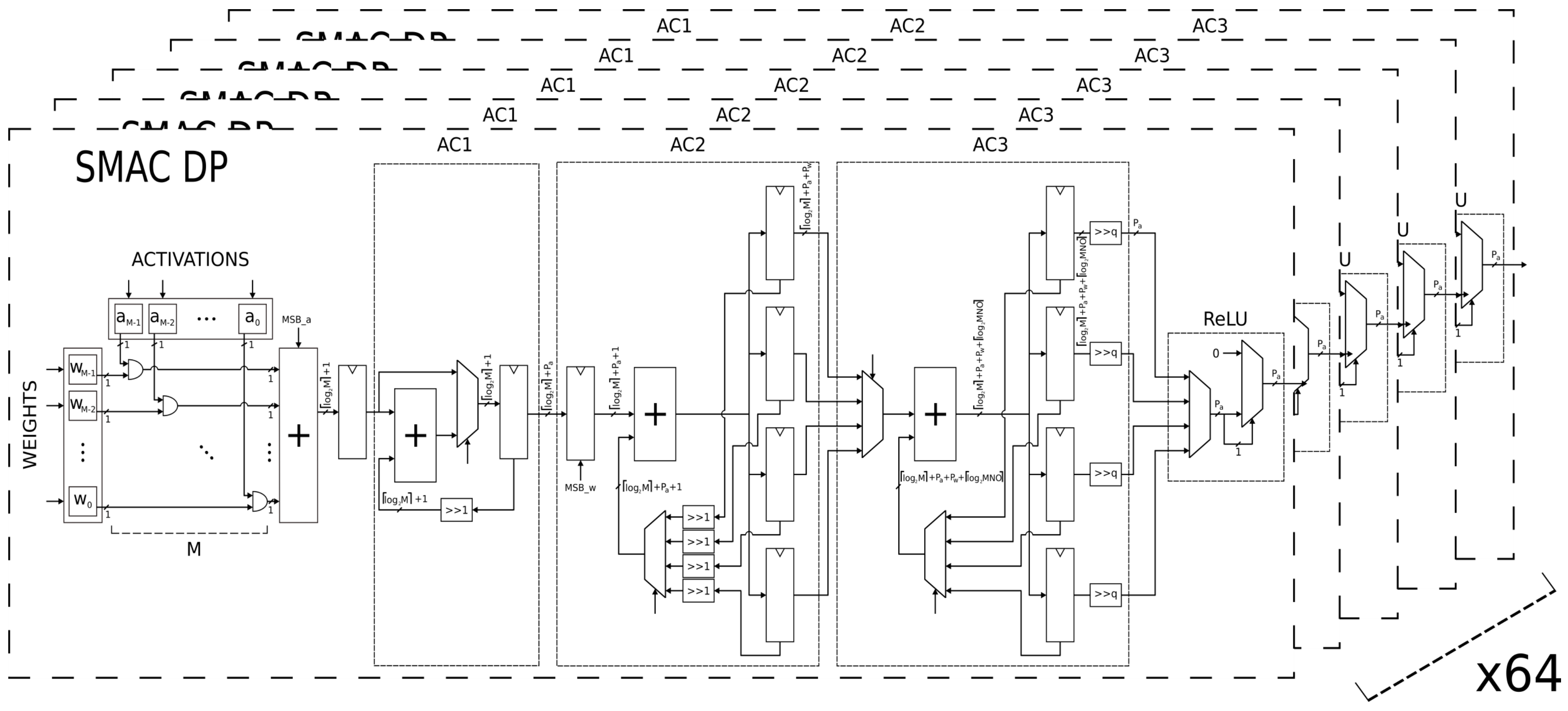
Serial-MAC-Engine (SMAC-E)

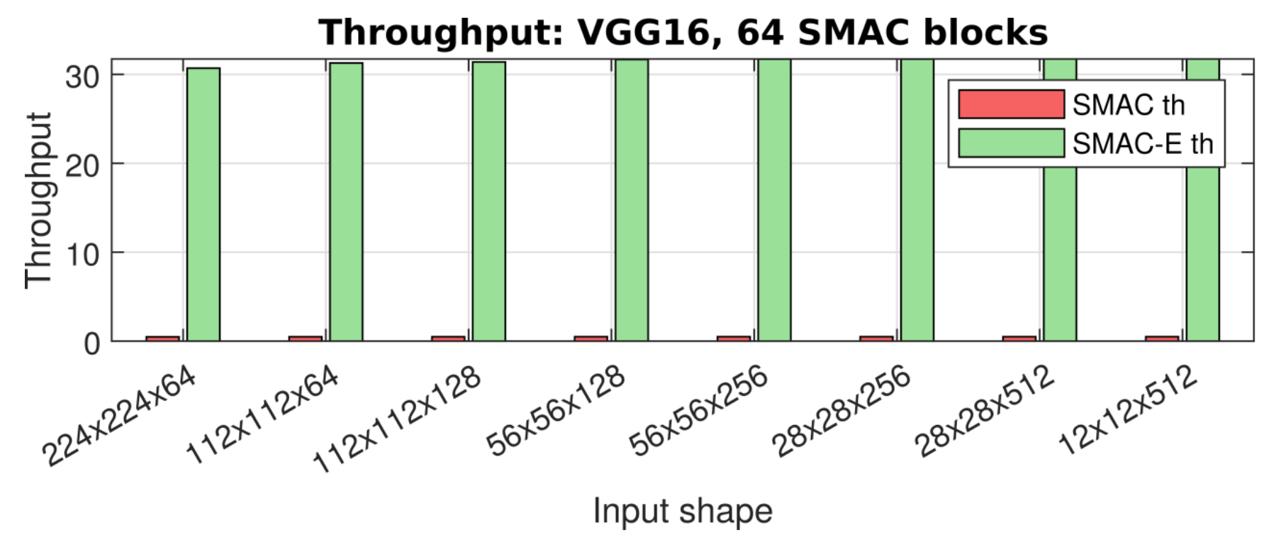
- The SMAC-Engine is a **multi-precision** Hardware Accelerator that is able to process convolutional and fully-connected layers in a **bit-serial** approach.
- The activation parallelism, called Pa, can be selected between 8 and 4 bits, while the weight one, Pw, among 8, 6 and 4 bits.
- Three accumulation stages allow the accelerator to process **256 filters** per convolution.

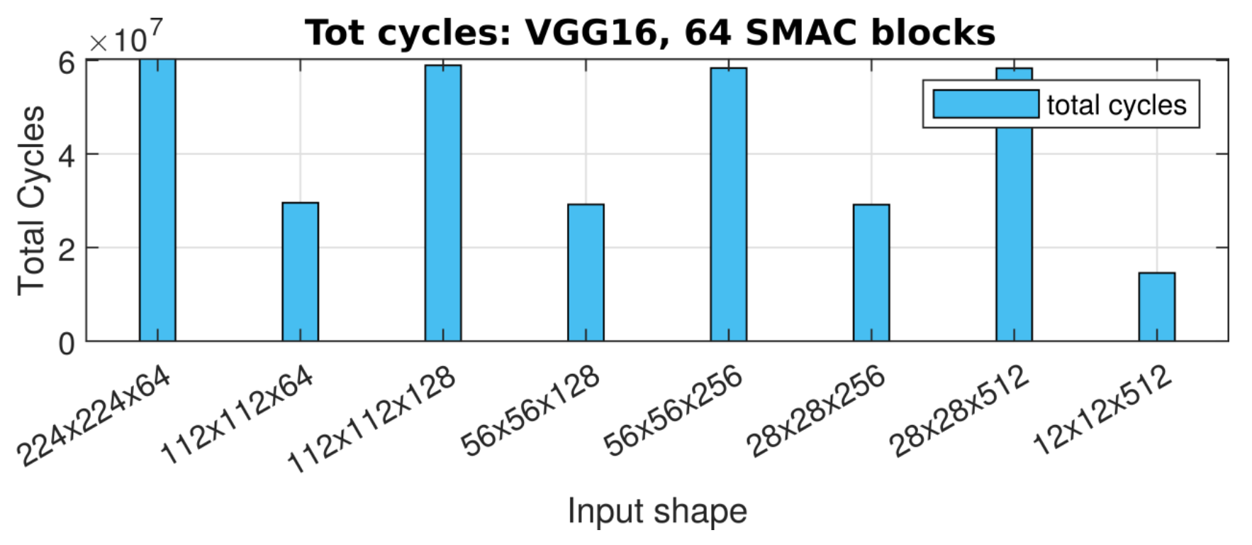


F. Conti, P. D. Schiavone and L. Benini, "XNOR Neural Engine: A Hardware Accelerator IP for 21.6-fJ/op Binary Neural Network Inference," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, no. 11, pp. 2940-2951, Nov. 2018

- **Output and input stationary** dataflow.
- Convolution performed by fetching the activations along the **direction of the channels**.
- The hardware accelerator has been integrated into **PULPissimo platform**.







HW solution	Frequency [MHz]	Area [mm²]	Throughput [GMAC/s]	Power [mW]	Energy Efficiency [pJ/MAC]
SMAC-E	416 (wc)	0.29	12.69	-	-
SMAC-E + HWPE @ 0.9V	285 (wc)	0.35	7.79	10.48	1.34
Fulmine @ 0.8V	108 (tc)	0.35	6.35	13	2.05
ShiDianNao	1000 (tc)	4.86	64	320	5
Eyeriss @ 1V	200 (tc)	12.25	23	278	12.09
XNE @ 1.2V	400 (tc)	0.092	35	5.92	0.15

SMAC-E has been **validated on VGG16** using Pa = 8 and Pw = 4.

After the synthesis on **UMC-65** the accelerator is able to reach **7.79 GMAC/s** consuming just **1.34 pJ/MAC @ 0.9 V.**

Future work

- **Develop Reversible Residual Network for continual learning** and test it on NVIDIA Jetson TX2.
- Adapt the accelerator to Reversible networks in order to perform **online training**.
- Transfer the application from Jatson to ASIC accelerator.
- Integrate a **scheduler** able to skip zero or near-zero value operations in the accelerator.

List of attended classes

- 01TEVRV – Deep learning (didattica di eccellenza)(04/06/19, 6 CFU)
- 03SGVRV – Entrepreneurship and start-up creation from University Research(04/07/19, 8)
- 03QTIIU – Mimetic learning(08/07/19, 4 CFU)
- 01QSCIU – Reconfigurable computing(11/02/19, 4 CFU)

POLITECNICO DI TORINO

PhD program in

Electrical, Electronics and Communications Engineering