Deep-Learning Oriented Smart Sensing for the Next Generation of Embedded Applications

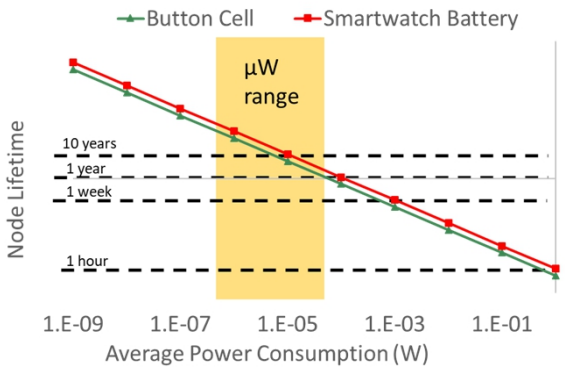
Manuele Rusci, Francesco Conti,
Alessandro Capotondi, Luca Benini

Energy-Efficient Embedded Systems Laboratory
Dipartimento di Ingegneria dell'Energia Elettrica e dell'Informazione "Guglielmo Marconi"

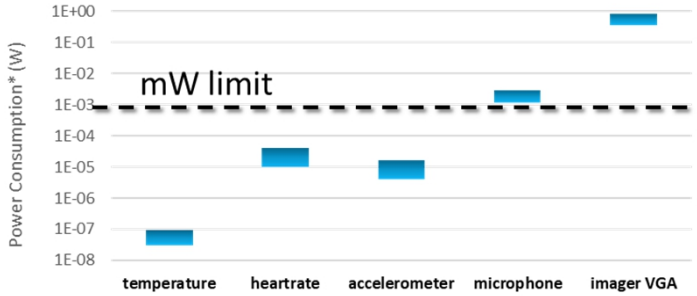
IWES18 – Siena, 14 Settembre 2018

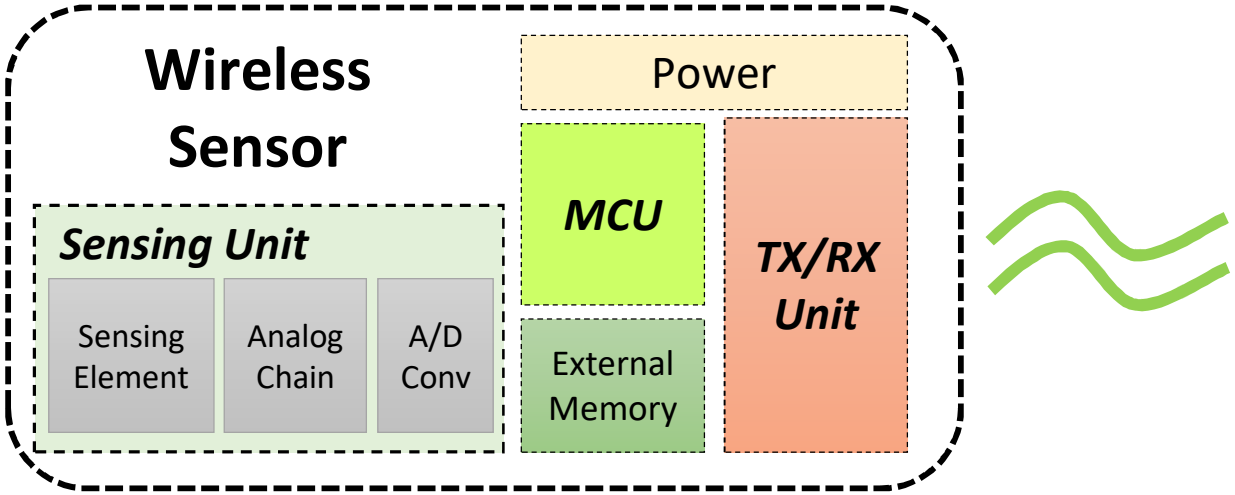
From data collectors…

Node average power budget



Wireless Sensing

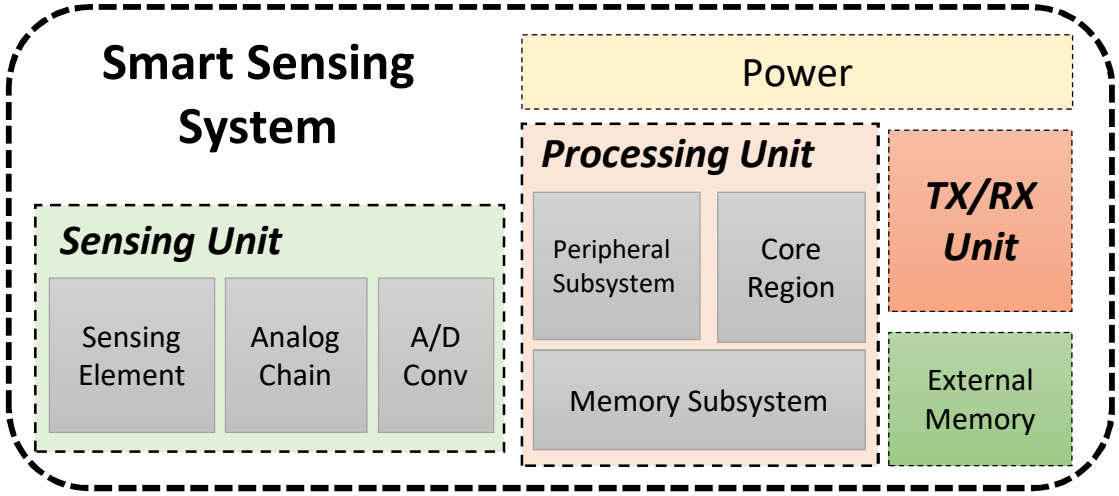




[Alioto, Massimo. "IoT: Bird's Eye View, Megatrends and Perspectives." *Enabling the Internet of Things*. Springer International Publishing, 2017. 1-45.]

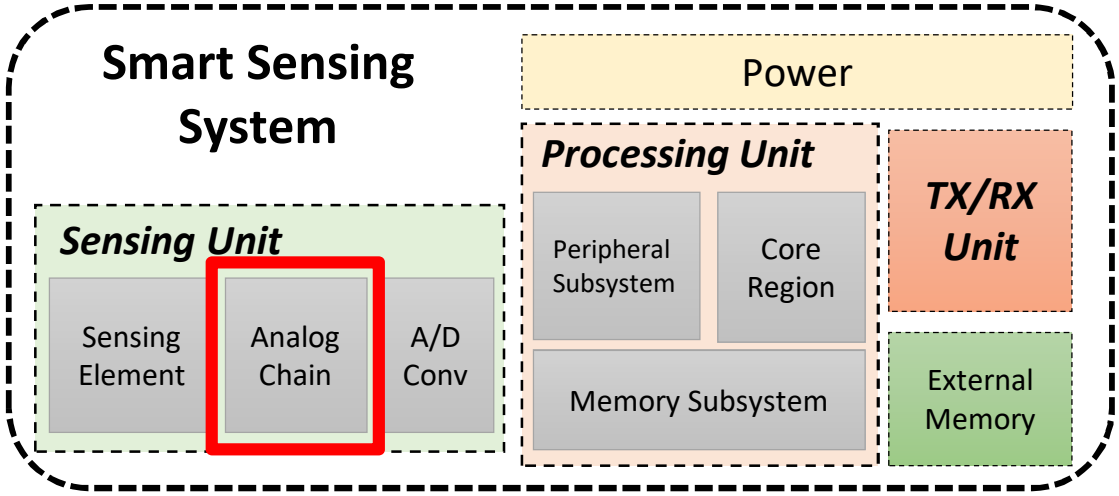
...to always-ON smart sensors

Challenge: bringing **intelligence** in-the-node at **mW** cost



...to always-ON smart sensors

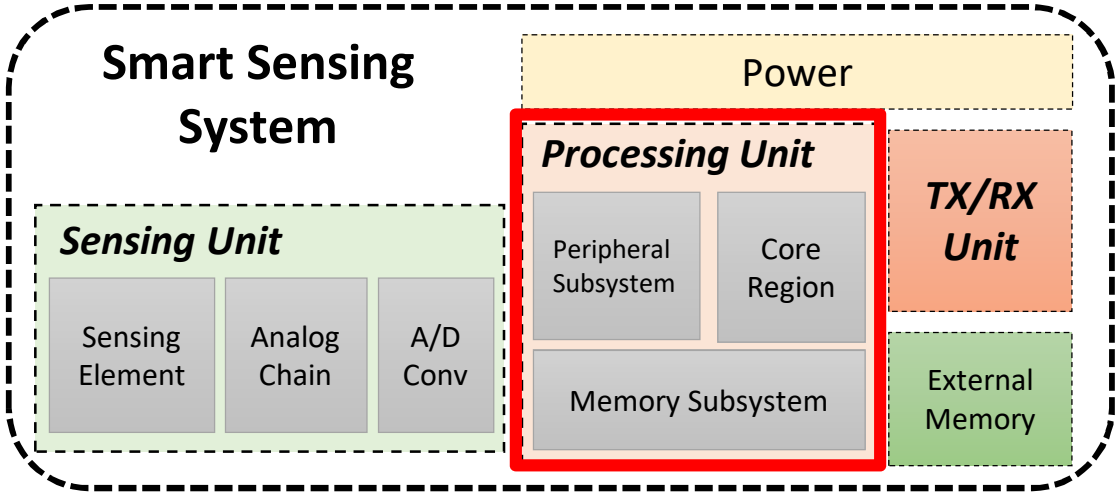
Challenge: bringing **intelligence** in-the-node at **mW** cost



1. **low-power "feature" / event extraction on sensor**

...to always-ON smart sensors

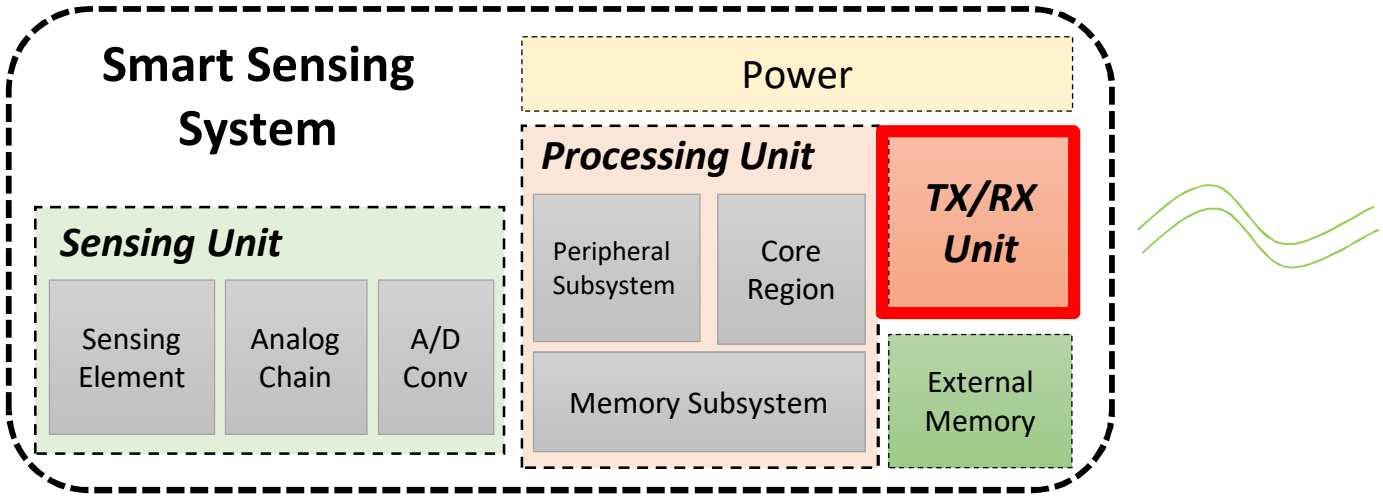
Challenge: bringing **intelligence** in-the-node at **mW** cost



1. low-power “feature” / event extraction on sensor
2. **event-based near-sensor processing**

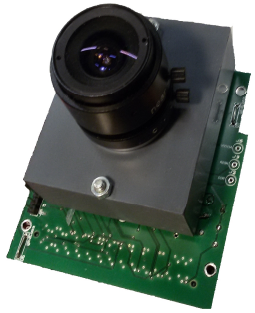
...to always-ON smart sensors

Challenge: bringing **intelligence** in-the-node at **mW** cost



1. low-power “feature” / event extraction on sensor
2. event-based near-sensor processing
3. **“slim” and uncommon transmission of high-level features**

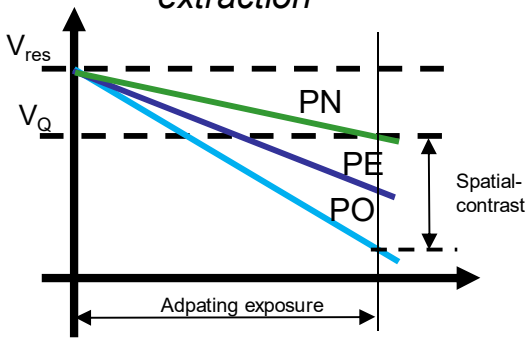
Ultra-Low Power Imaging (GrainCam)



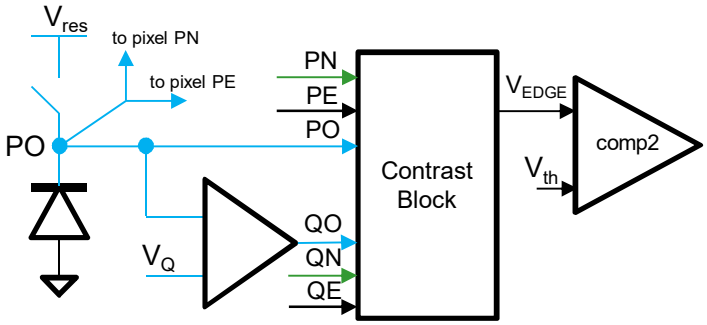
Focal Plane Processing. Moving an early computation stage into the sensor die to reduce the power costs of the imaging task.

Imager performing **spatial filtering** and **binarization** on the sensor die through **mixed-signal sensing**!

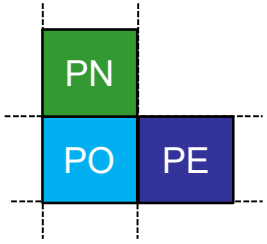
Gradient extraction



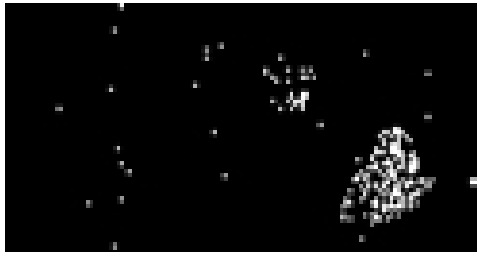
Per-pixel circitut for filtering and binarization



'Moving' pixel window



Traditional Camera

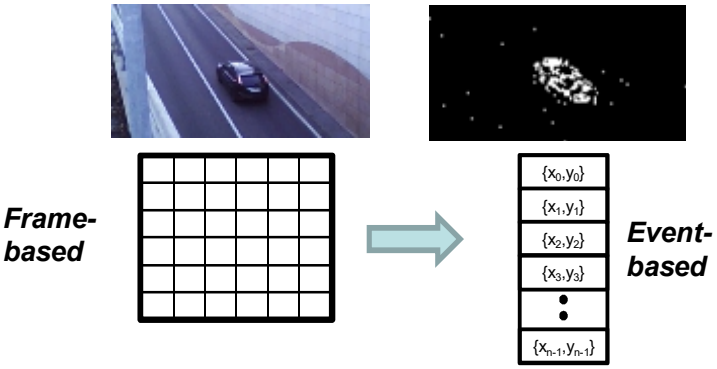


Graincam w/ motion detection

FONDAZIONE BRUNO KESSLER

Event-Based Paradigm

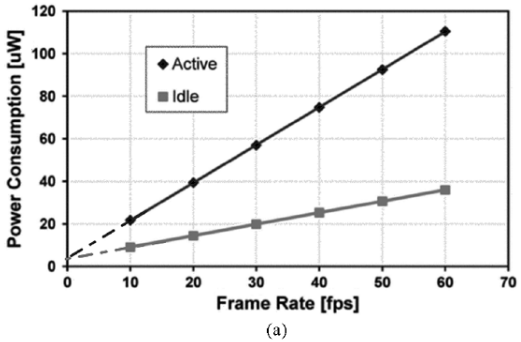
Event-based sensing: output frame data bandwidth depends on the external context-activity



Readout modes:

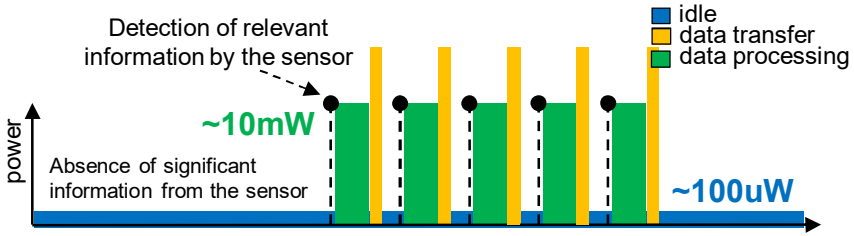
- **IDLE**: readout the *counter of asserted pixels*
- **ACTIVE**: sending the addresses of asserted pixels (Address-Coded Representation, AER)

Ultra-Low Power Consumption <100uW



<10x wrt SoA imagers

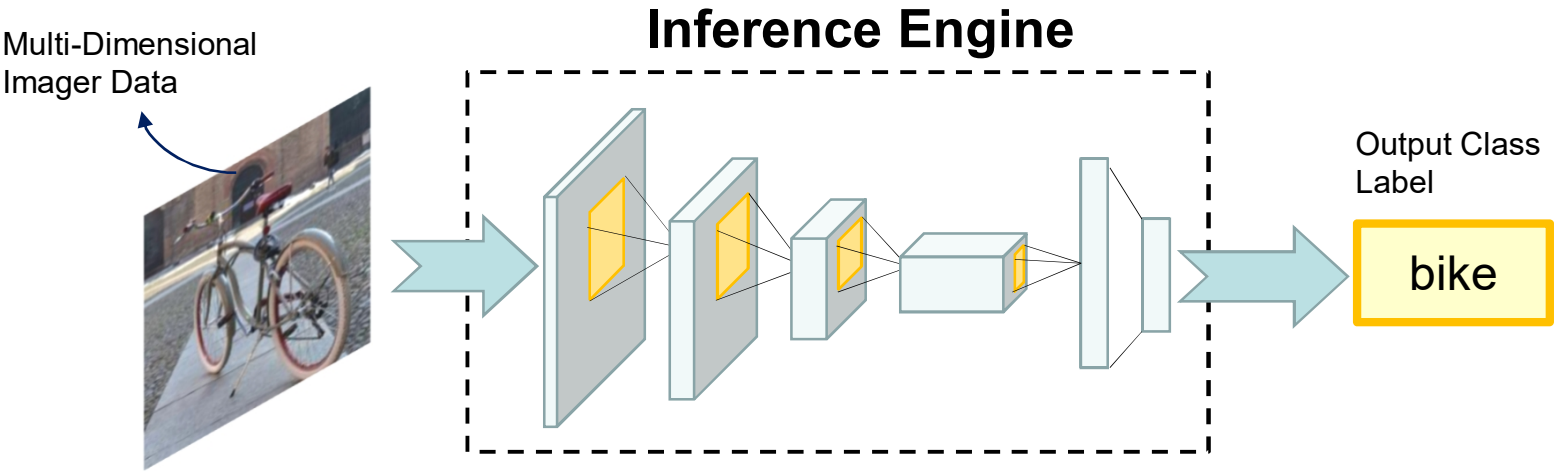
Event-Based Data Processing



M. **Rusci** et al. "A sub-mW IoT-endnode for always-on visual monitoring and smart triggering," in *IEEE Internet of Things Journal, 2017*

Deep Learning at the Edge

Convolutional Neural Networks are state-of-the art for visual recognition, detection and classification tasks

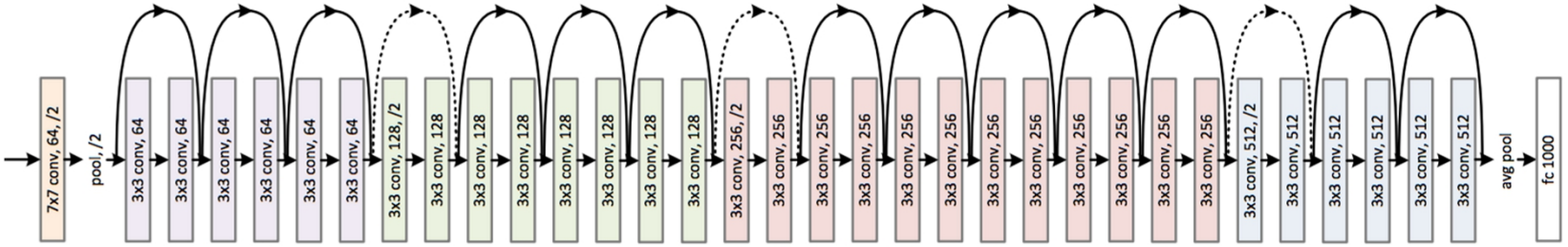


How to exploit CNNs on always-on devices with a power envelope of few mWs or sub-mW ?

Issues:

- Large memory footprint to store weights (the 'program') and intermediate results (up to hundreds of MBs), greater than memory footprint available on ultra-low power engines (100's kBs)
- High-complexity CNN implementation, demanding floating-point precision
- Imager Power costs of tens to hundreds of mWs

Deep Learning at the Edge



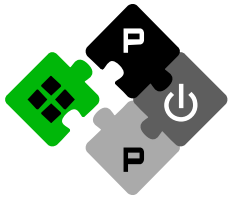
"Extreme" example: **ResNet-34**

- classifies 224x224 images into 1000 classes
- ~ trained human-level performance
- ~ **21M** parameters
- ~ **3.6G MAC**

Performance for 1 fps: **~3.6 GMAC/s**

Energy efficiency for 1 fps @ 20 mW: **~180 GMAC/s/W = ~5pJ/MAC**

Specialized HW



parallelism and ***HW acceleration*** are key paradigms to achieve low energy

Quantization

Precision	Accuracy loss
full precision / 8bit	0
6bit	-1.3%
4bit	-3.3%

VGG-16 @ CIFAR-10

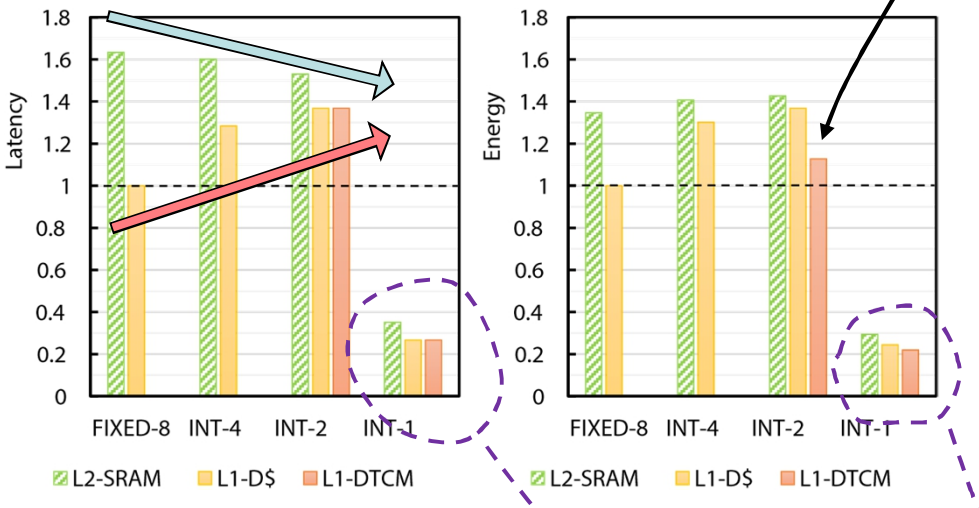
Quantization: *no free lunch*

Running INT-Q convolution on a ARM Cortex-M7 core

-> huge opportunity for HW/SW codesign

lower bandwidth from L2-SRAM impacts on low-bitwidth precision

Lower power consumption when fitting into L1 thanks to compression

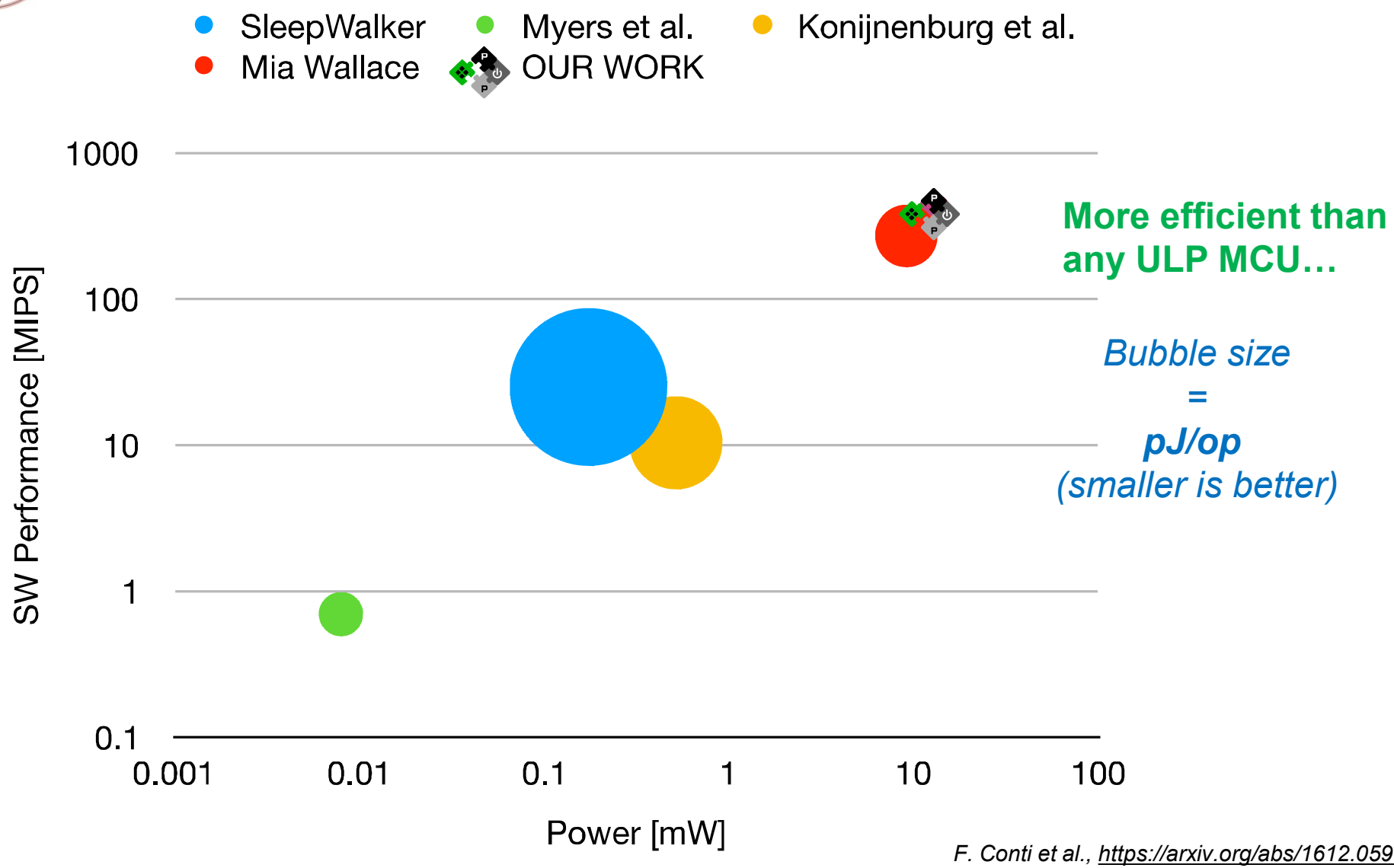


overhead for casting INT-4/2 to INT-16 for 2x16bit vectorized MAC instructions

INT-1 kernel exploits bitwise operations and does not pay casting overhead because XNOR convolutions are supported by the ISA

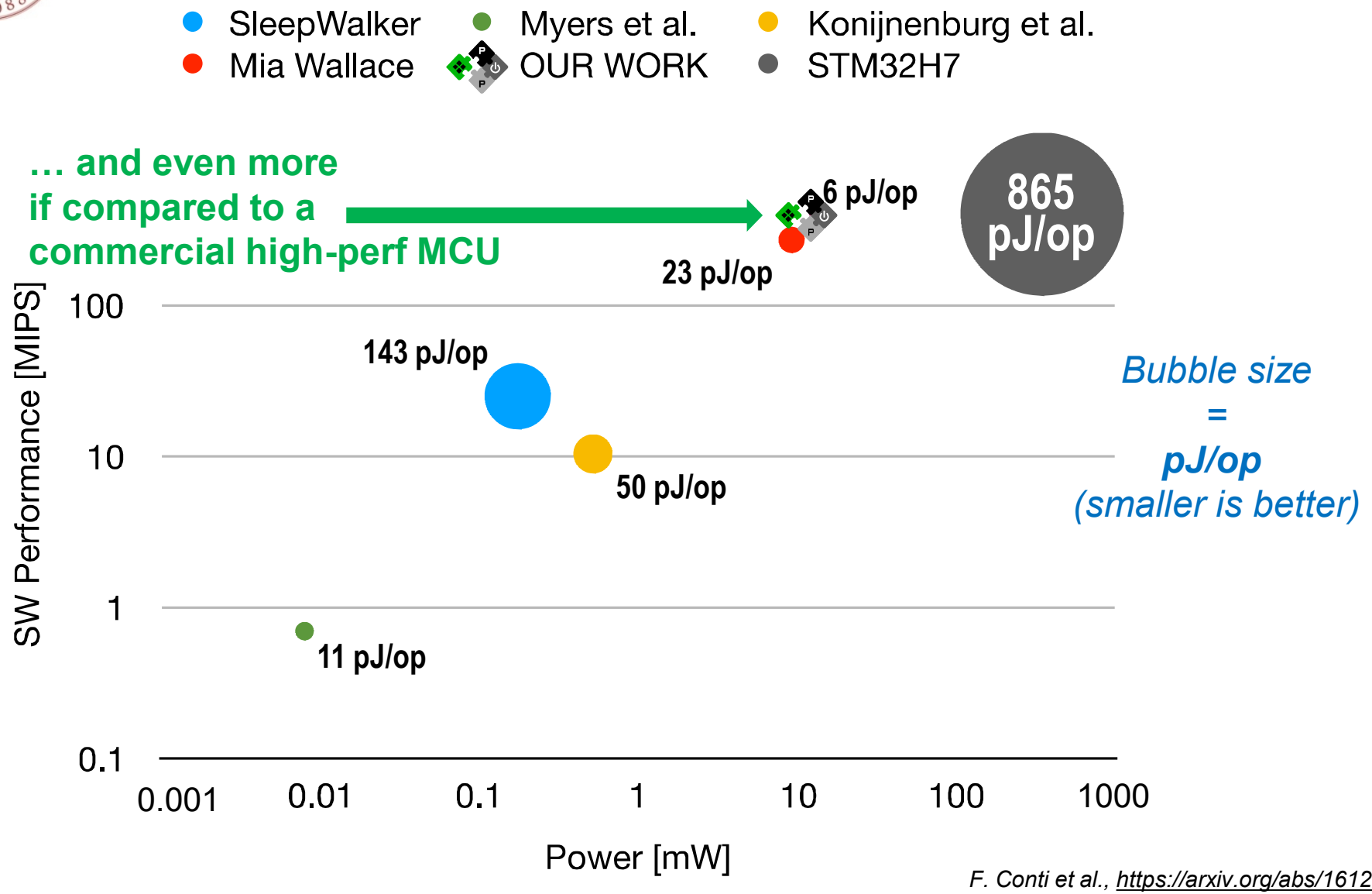
Open Source:
https://github.com/EEESlab/CMSIS_NN-INTQ

Quantization + Acceleration = ♥



F. Conti et al., https://arxiv.org/abs/1612.05974

Quantization + Acceleration = ♥



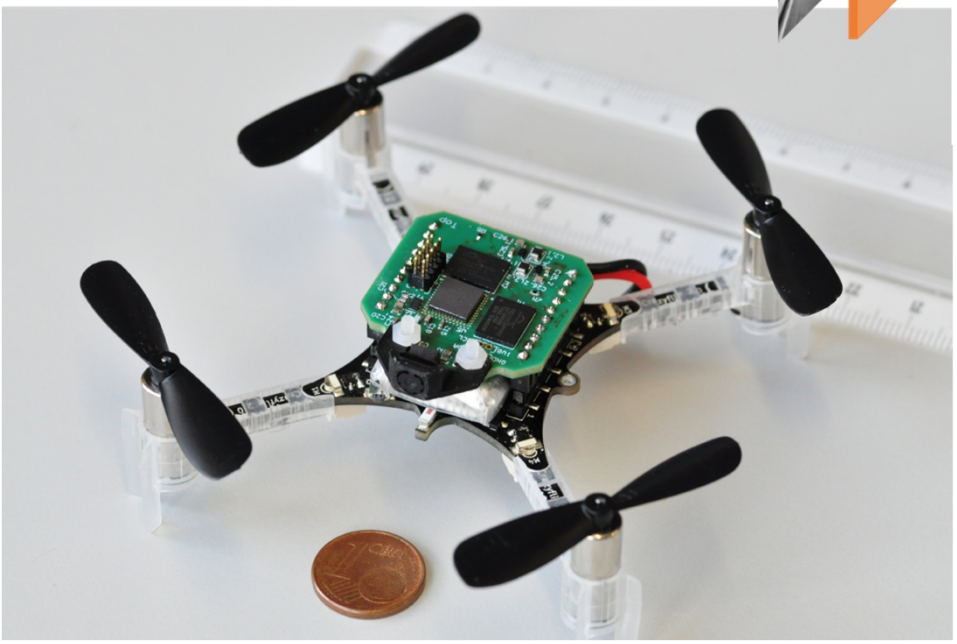
F. Conti et al., https://arxiv.org/abs/1612.05974

Flying a Drone with DL (**in <10mW**)

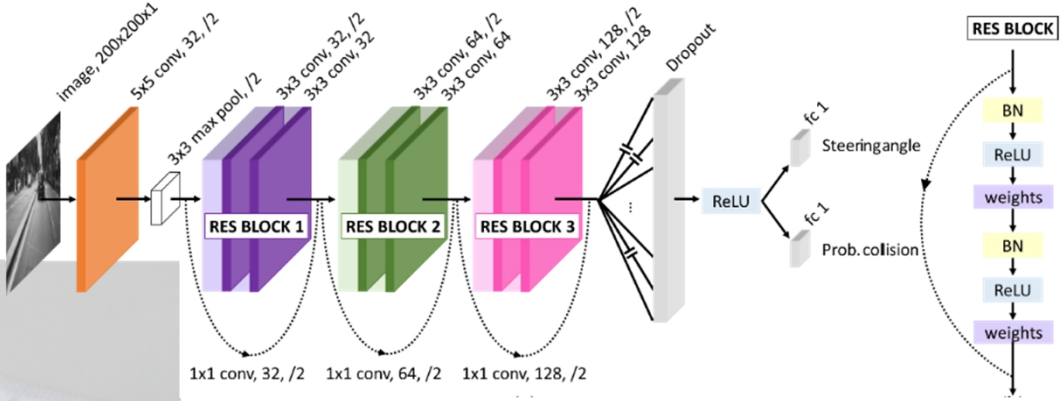
DroNet: a ResNet-based CNN to drive a drone in the environment

- original implementation: **20fps** on external CPU, requires a big drone (e.g. DJI, Parrot)

	GAP8 – 8 Cores (200MHz)	GAP8 – HWCE (200MHz)
FPS	**32 fps**	**51 fps**



Example nano-drone from D. Palossi et al., https://arxiv.org/abs/1805.01831



DroNet on GAP8/PULP:

- Fixed-Point 16bit (Q3.13)
- Removed Batch Normalization
- Max Pooling layer 2x2
- Striding support in HW
- Support for HWCE
- Comparable accuracy w.r.t. baseline

GREENWAVES TECHNOLOGIES

ETH zürich

Thanks for your attention.

Questions?

Special acks to: Davide Rossi (UNIBO), Daniele Palossi (ETHZ), Eric Flamand (GreenWaves Technologies), all the PULP team

https://**github.com/pulp-platform**

Twitter **@pulp_platform**