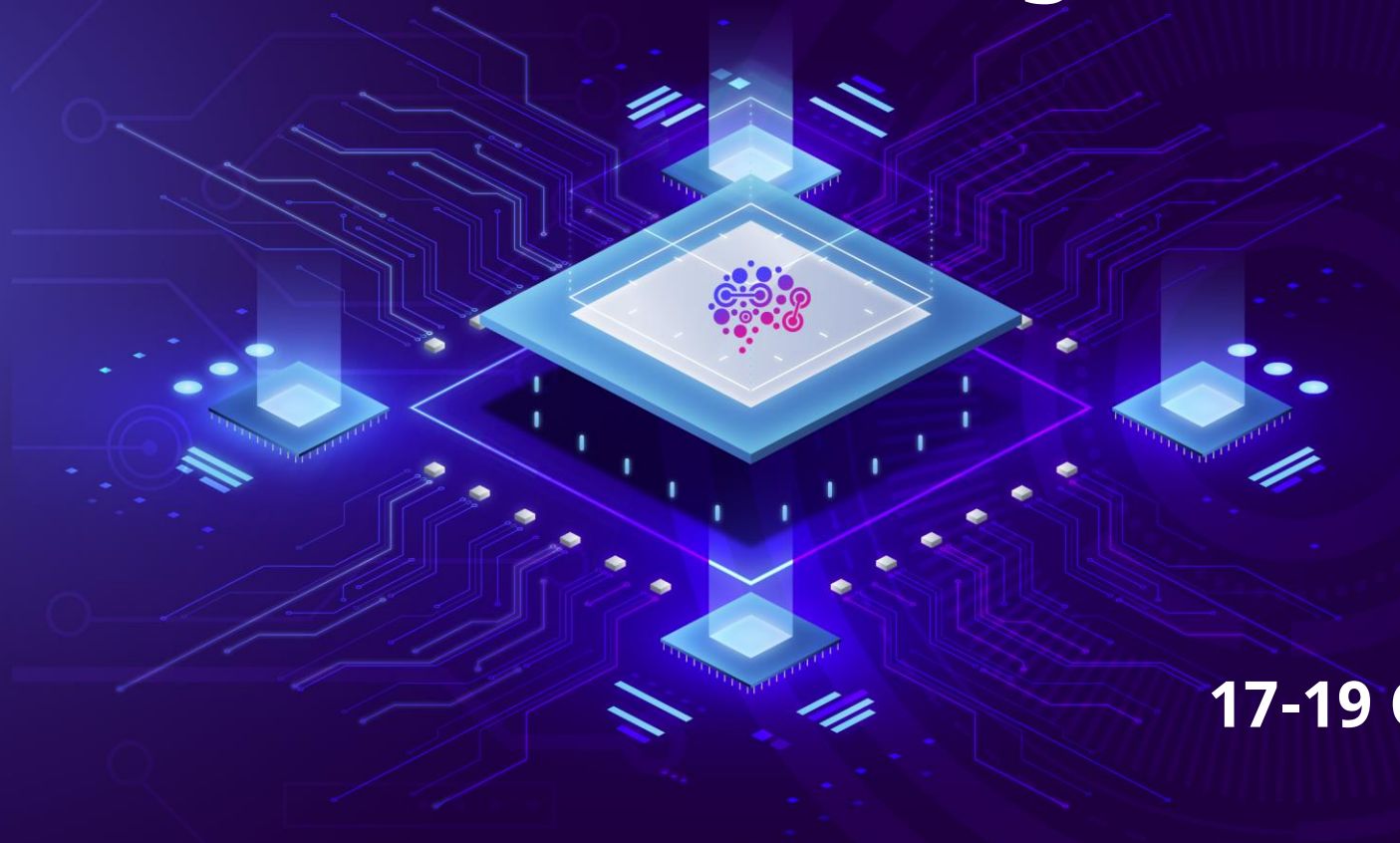
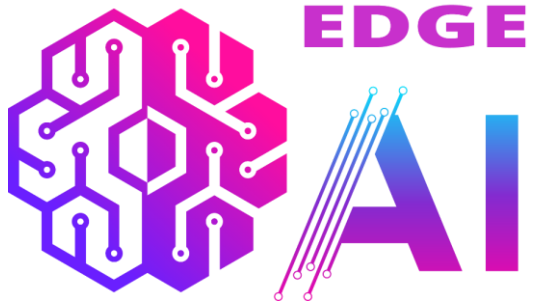
# European Conference on EDGE AI Technologies and Applications - EEAI



**Advancing emerging edge AI technologies and driving next generation intelligent applications.**

**17-19 October 2023, Athens, Greece**

# European Conference on EDGE AI Technologies and Applications - EEAI



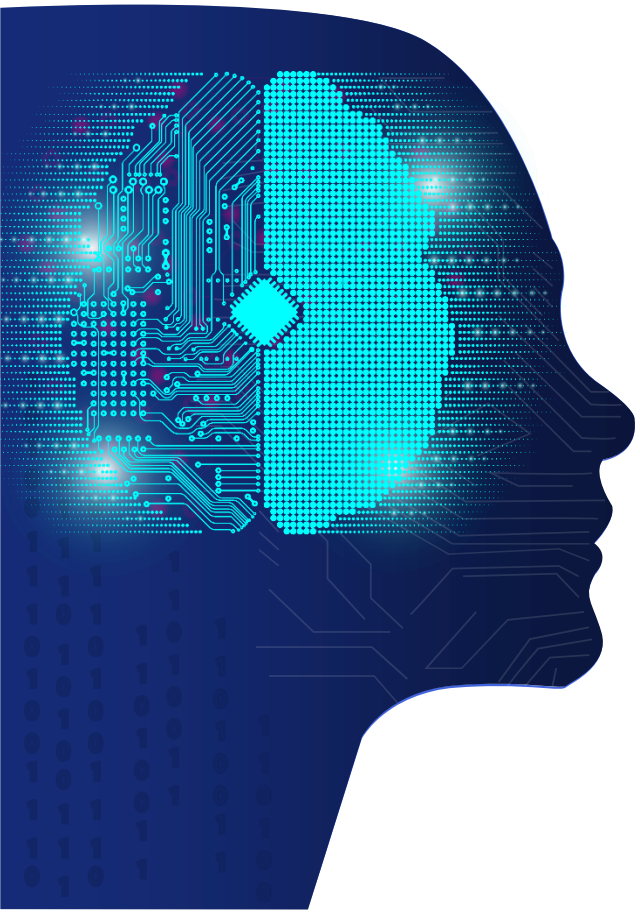

## Inside the AI Accelerators: From High Performance to Energy Efficiency

**Ana Pinzari, Adrien Prost-Boucle, Christelle Rabache, Frédéric Pétrot**
***Institute of Engineering University Grenoble Alpes, France***

**17-19 October 2023 Athens, Greece**

Copyright © 2023

# Presentation Outline

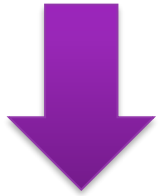


- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

# Introduction

- Immersion of AI-based solutions (recognition, classification etc)
- High capacity of predefined neural networks (VGG, Inception, ResNet)
- Tackle the huge amount of data
- Complex and diverse hardware architectures
- Exploration of HW accelerators and optimization techniques



**Achieve High Performance and Energy Efficiency ?**

Copyright © 2023

# State-of-the-art

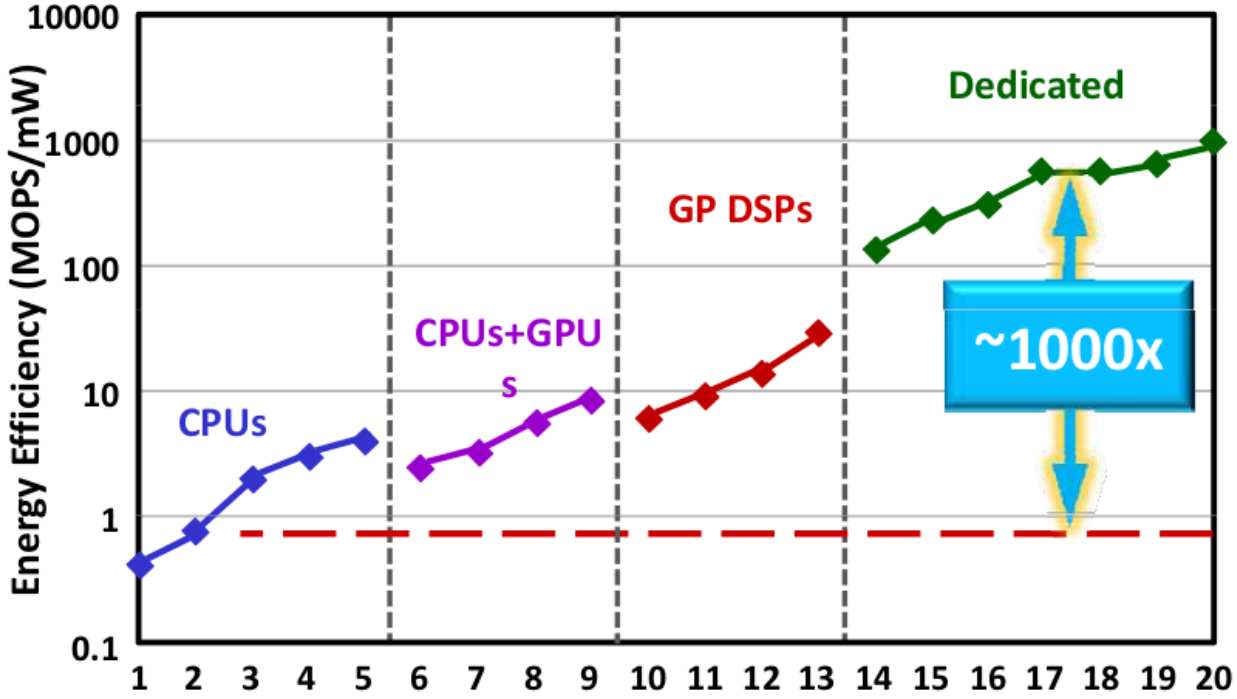
- Accelerated performance through **DL Compilers**:
  - CPU – Halide, XLA (TensorFlow), TVM (Apache), Glow (PyTorch), nGraph (Intel)
  - GPU – XLA, TVM, TC
  - FPGA - DNNVM, xfDNN, fpgaConvNet, our generator

- Custom ASIC (**TPU**)
- **Quantization** Techniques - QAT (hybrid quantization), PTQ
- **T**ensor **D**ecomposition
- Model **Optimizations,** knowledge distillation

Copyright © 2023

# Research Goal

- Facilitate/Optimize the NN portability on AI-edge platforms :
  - MCU, GPU, TPU and FPGA
- Evaluate neural network performance/accuracy and power/energy efficiency




["*Computing's Energy Problem (and what we can do about it)*", M.Horowitz, ISSCC 2014]

Copyright © 2023

# Methodology

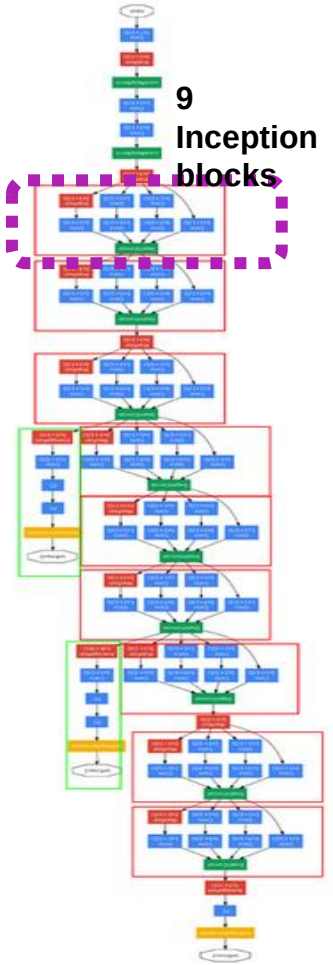
- Neural network design:
  - UC: supervised learning, multi-class (58) classification task, custom dataset
  - Deep CNN: find a lightweight architecture ?

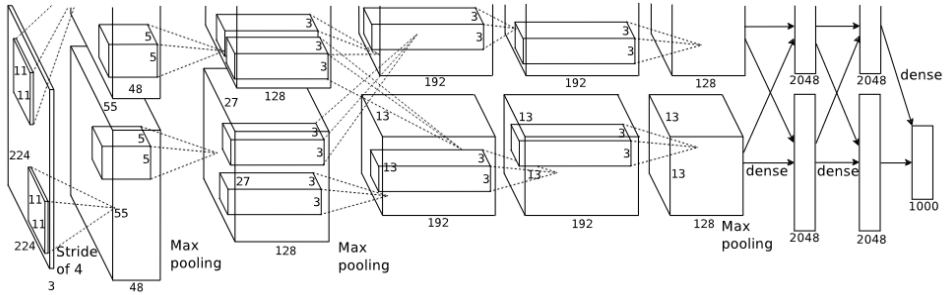
- Quantization:
  - 8 bits, 4 bits… 1 bit ?

Copyright © 2023

# Design and Implementation - 1

- From High Capacity to Lightweight CNN:



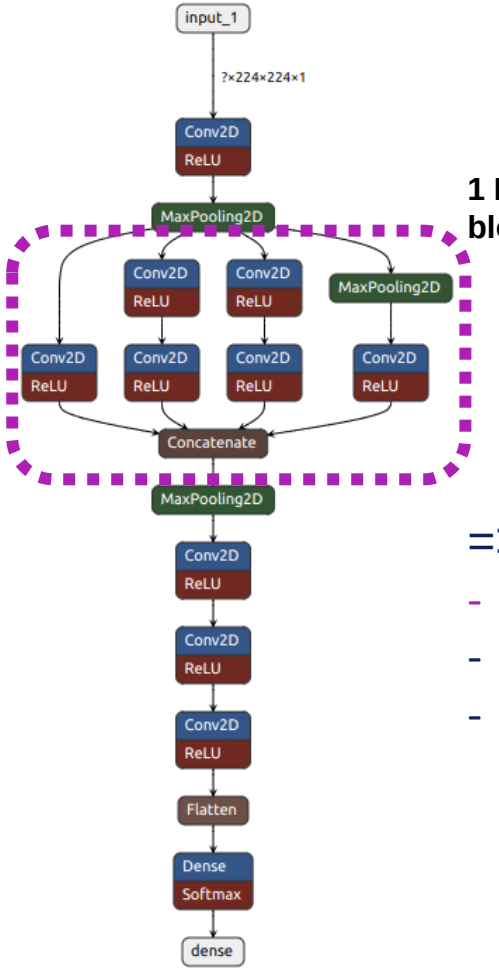




AlexNet – **61M** parameters (725M FLOPs)
(224 x 224 x 3) images

[Krizhevsky et al., NIPS 2012 ]

GoogLeNet – **6.9M** parameters (1.5B FLOPs)
(224 x 224 x 3) images

[Szegedy et al., CVPR 2015 ]




**1 Inception block**

=> Our CNN model:

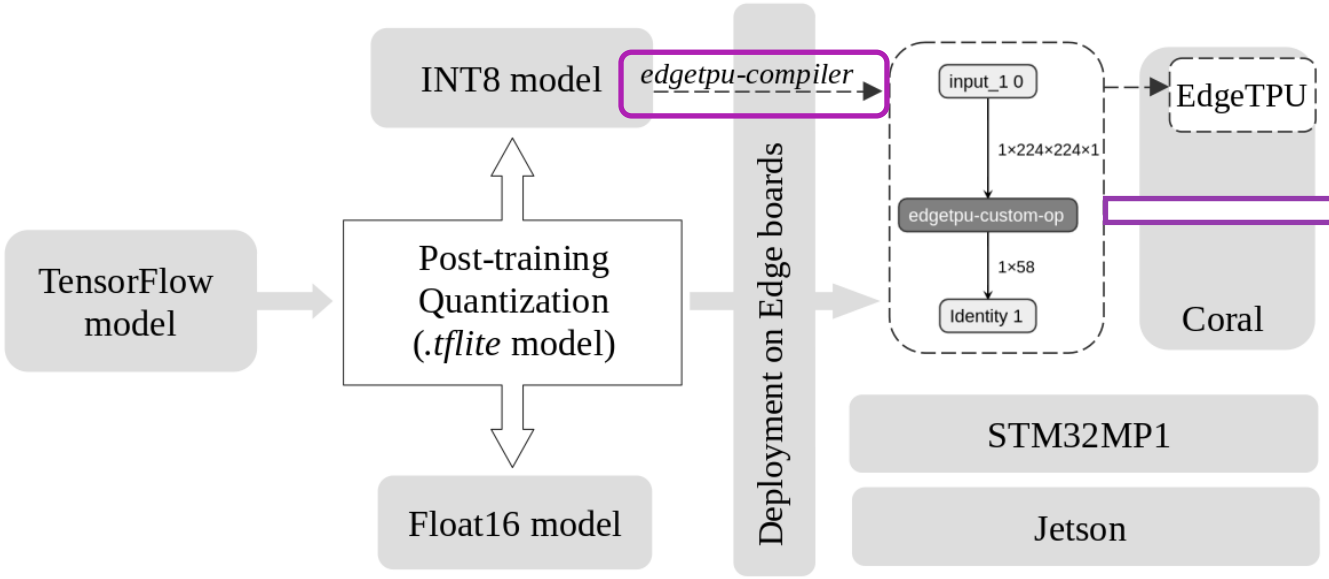
- **0.47M** parameters
- 10 learnable layers
- 125M FLOPs

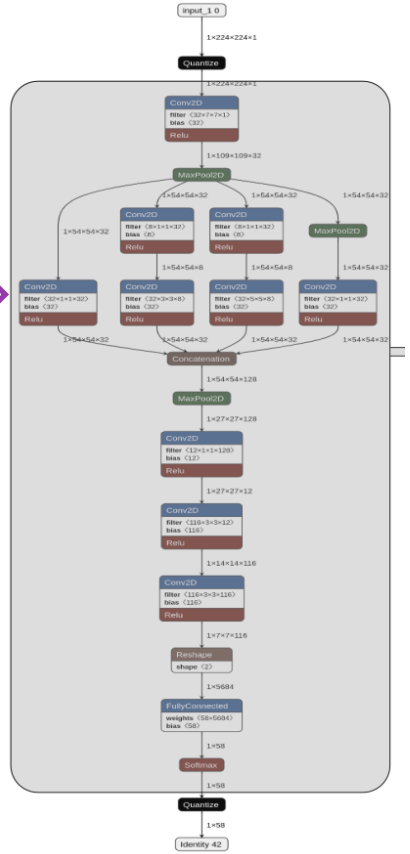
(224 x 224 x 1) images

Copyright © 2023

# Design and Implementation - 2

- Post-training integer quantization (8-bit, float16)
  - MP1 – GPU - TPU



- Conversion workflow to tflite model
- `Edgetpu-custom-op =>` all operations are supported and mapped to TPU

Copyright © 2023

# Design and Implementation - 3

- Quantization Aware Training:
  - 4 bits: 96 %
  - 2bits: 76%
  - **heterogeneous (4 bits, binary {-1, 1}): 94.6%**

QKeras – high-level API to define and train highly optimized NNs:

- *QConv2D, QActivation, QDense*

Example: QConv2D

Kernel_quantizer (Weights, biases)

+

QActivation

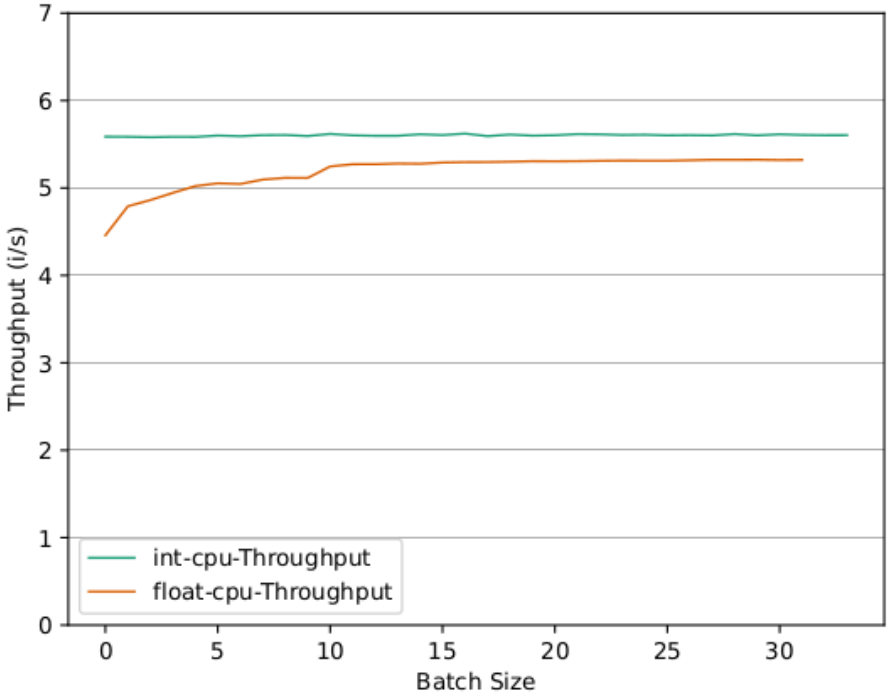
```
First Layer: kernel_quantizer =quantized_bits(4,0,1)
...
Middle layers: kernel_quantizer='binary()'
...
Last (Dense): kernel_quantizer =quantized_bits(4,0,1)
```

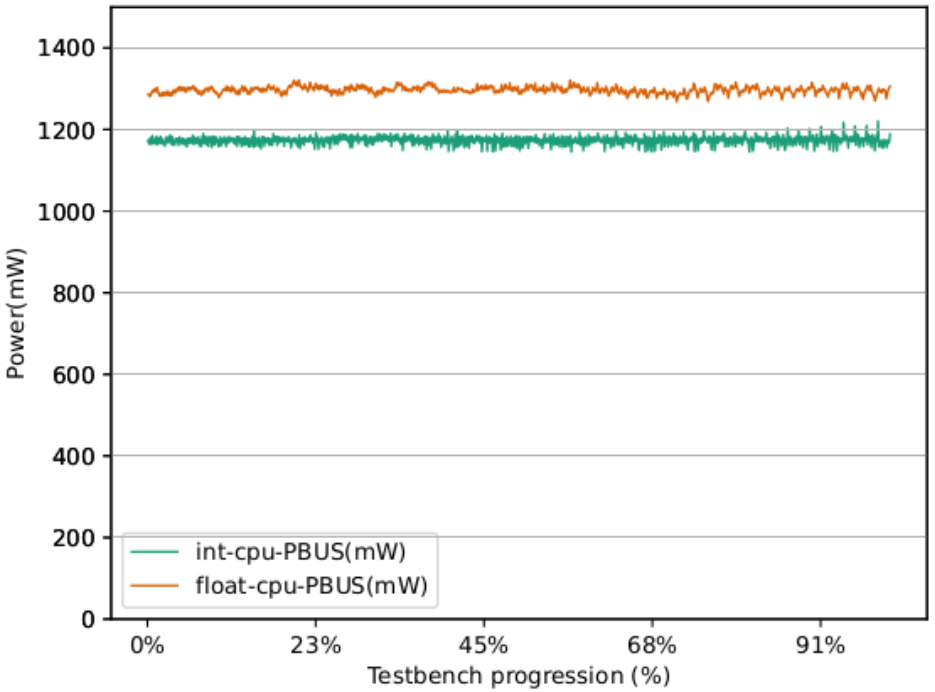
```
quantized_relu(4,0)
```

Copyright © 2023

# Experimental Results –1

- MCU - STM32MP1 CPU Cortex-A7 (int and float)
- PTQ (int, float ): 97% accuracy
  - **Batch=1: 4.7 images/s/W (int-cpu)**



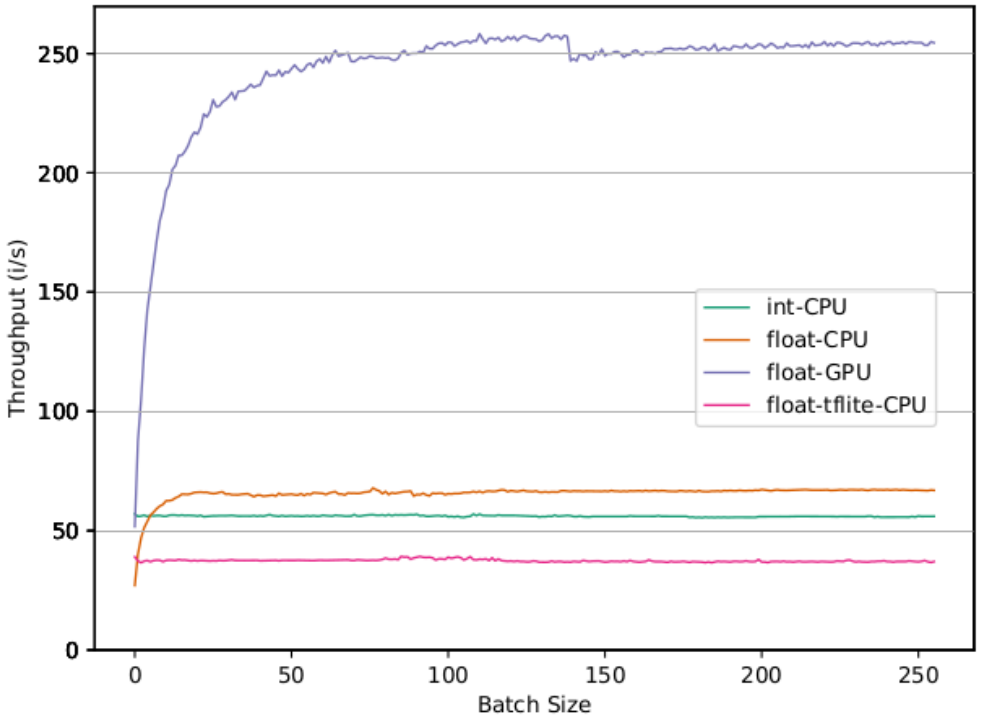




1 batch = 100 images

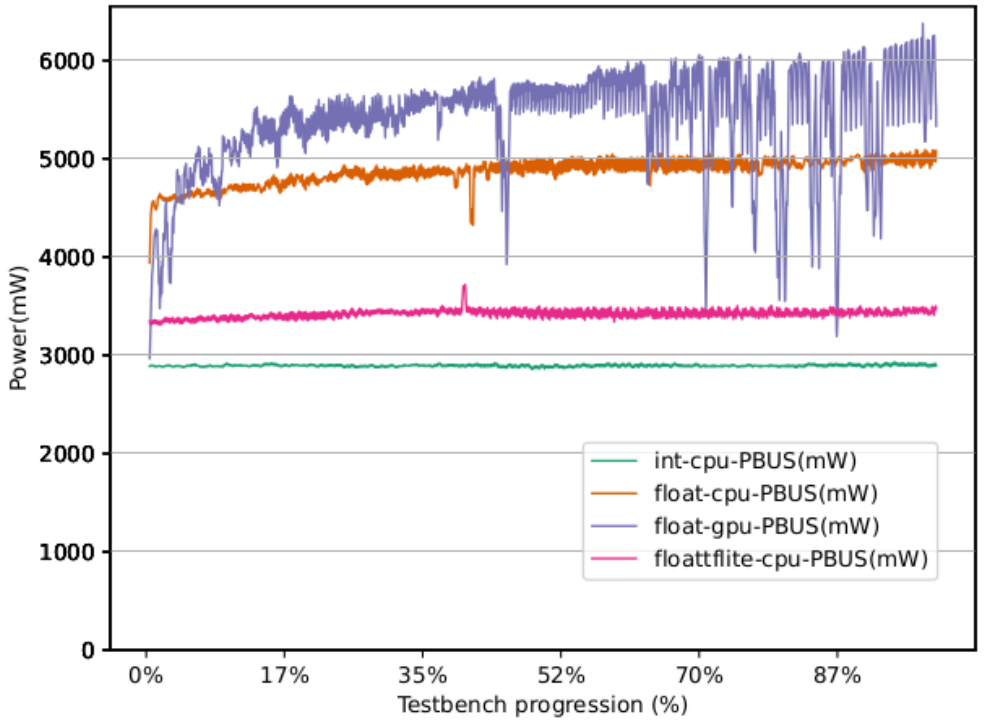
Copyright © 2023

# Experimental Results – 2

- GPU – Jetson Maxwell GPU (float), 128 CUDA cores and CPU (int and float) Quad Cortex-a53
- PTQ (int, float): 97% accuracy
  - **Batch=1: 44 images/s/W (float-GPU)**



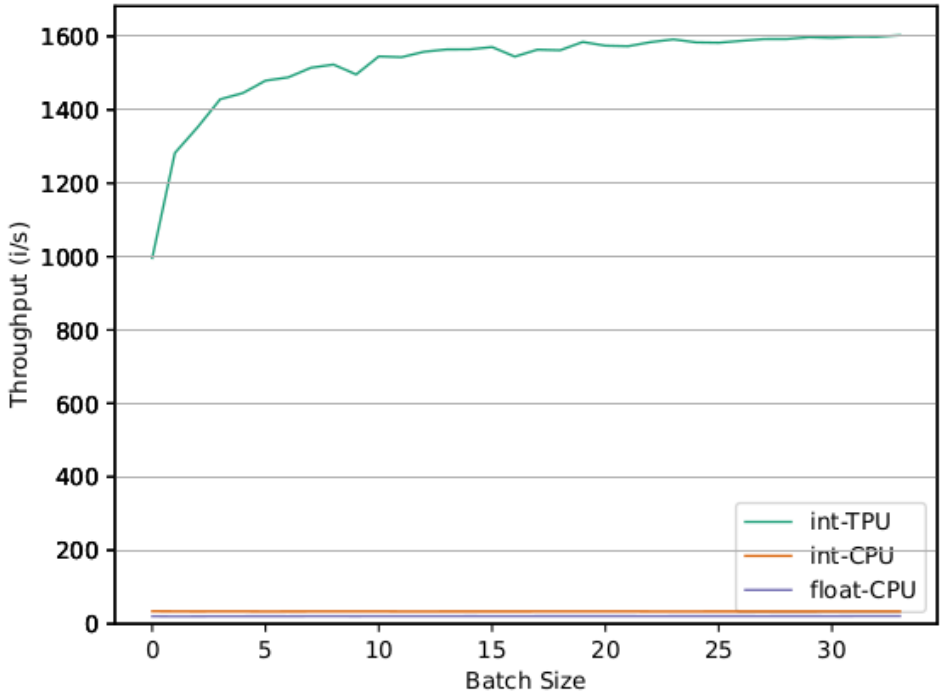

1 batch = 100 images



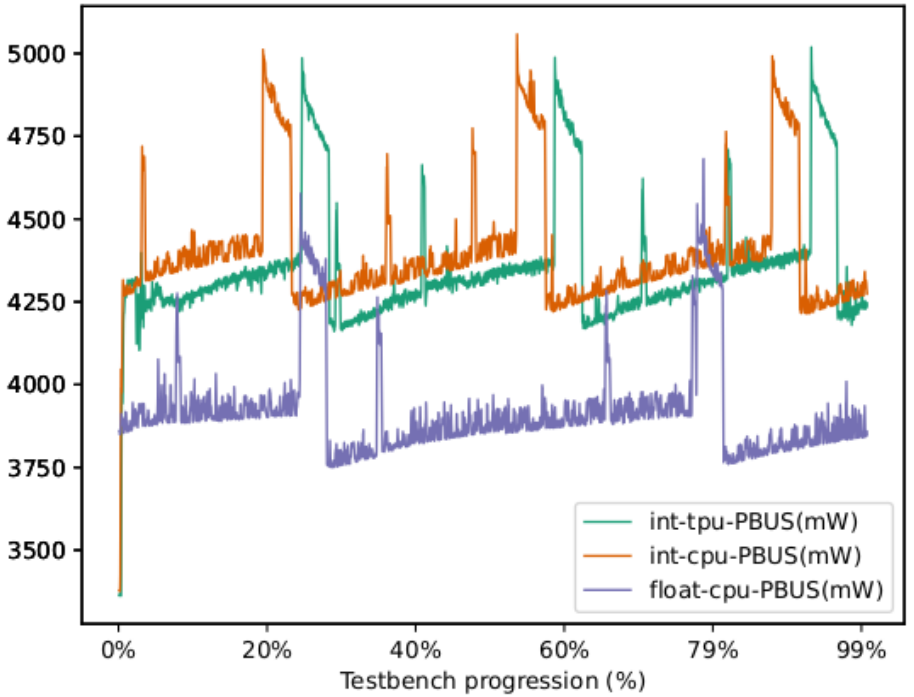

Copyright © 2023

# Experimental Results – 3

- TPU - Google Coral TPU coprocessor 4 TOPS (int) and CPU quad Cortex-A53 and a Cortex-M4F (int and float)
- PTQ (int, float): 97% accuracy
  - **Batch=1: 320 images/s/W (int-TPU)**




1 batch = 100 images




Copyright © 2023

# Experimental Results – 4

- FPGA – Zynq *Z010* (dual-core) 16nm, 2.6MB On-Chip memory
- Heterogeneous quantized model (4-bit, 1bit): 94.6% accuracy
  - **101.7 images/s/W (board)**
  - **741 images/s/W (FPGA only)**

| LUT (logic) | LUTRAM | Slice Registers | Block RAM | DSP cores |
|---|---|---|---|---|
| 9030 / 17600 (51.3%) | 4830 / 6000 (80.50%) | 11796 / 35200 (33.51%) | 60 / 60 (100%) | 37 / 80 (46.25%) |

| Performance | Latency | Power (FPGA only) | Power (Entire Chip) |
|---|---|---|---|
| 178 images/s | 26 ms | 0.24 W ~ 1.34 mJ /image | 1.75 W ~ 9.83 mJ / image |

Copyright © 2023

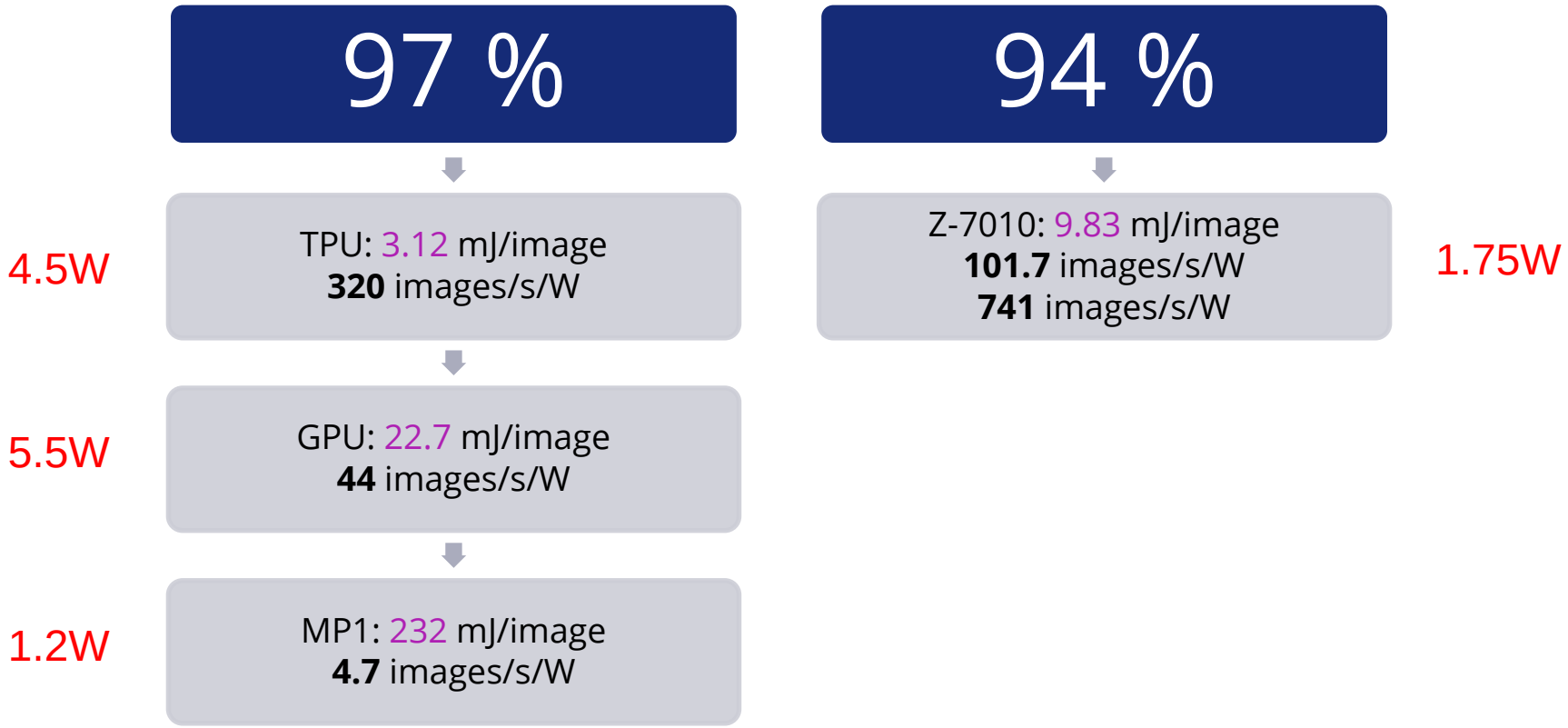
# Future Research

- Other data types (3-channel images, video)
- Other applications (e.g. detection)
- Benchmarking ad-hoc NN vs SoA (DL Compilers)

Copyright © 2023

# Conclusion

Model **Performance** vs Energy **Efficiency** vs Power **Consumption**

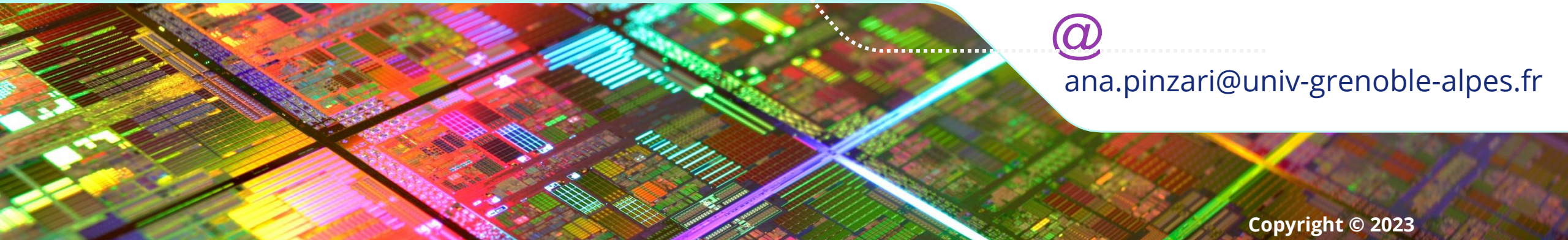



Copyright © 2023

EDGE AI

# Thank You

## For your attention



@ ana.pinzari@univ-grenoble-alpes.fr

Copyright © 2023