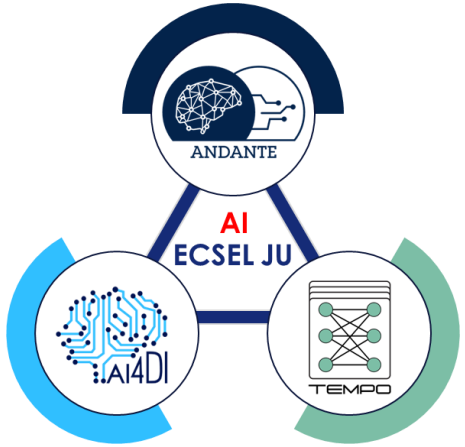
International Workshop on Embedded Artificial Intelligence Devices, Systems, and Industrial Applications (EAI)

MILAN
19-22 SEPTEMBER 2022
ESSDERC
ESSCIRC

Milan, Italy 19 September 2022

International Workshop on Embedded Artificial Intelligence Devices, Systems, and Industrial Applications (EAI)



Power Optimized Wafermap Classification for Semiconductor Process Monitoring

Ana Pinzari, Thomas Baumela, Liliana Andrade, Marcello Coppola and Frédéric Pétrot,
Grenoble INP, TIMA laboratory

MILAN 19-22 SEPTEMBER 2022 ESSDERC ESSCIRC

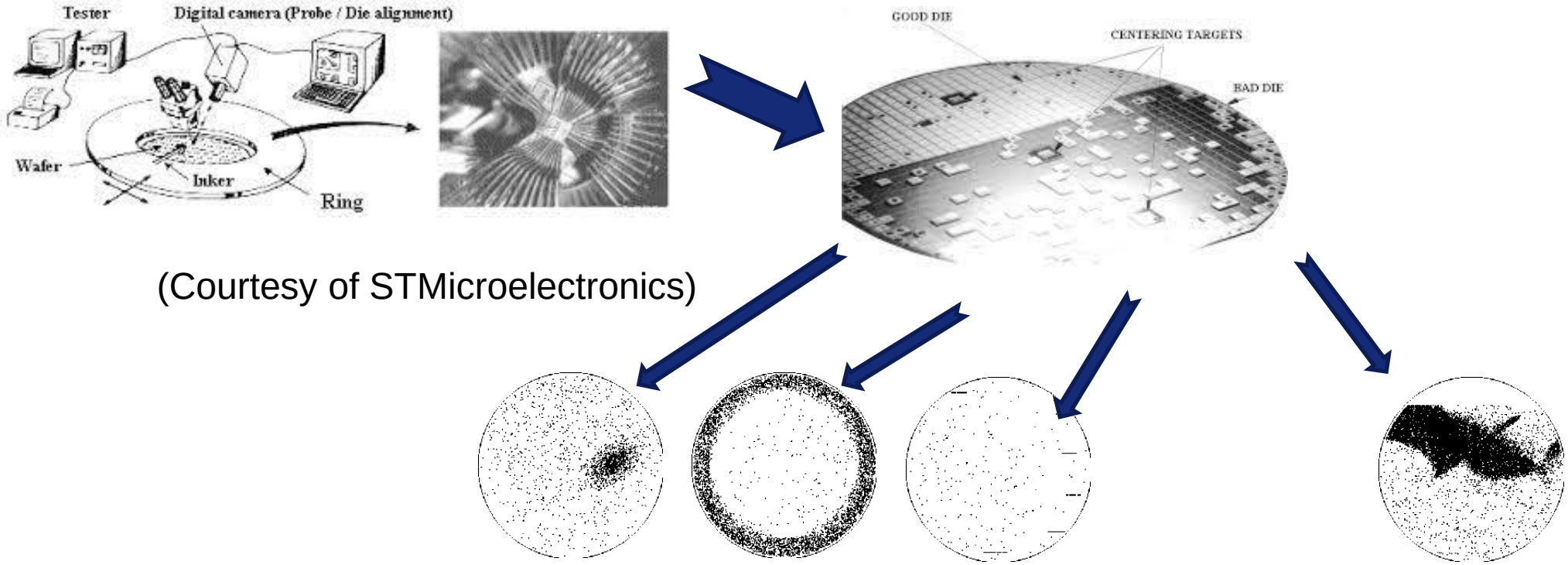
19 September 2022 Milan, Italy

Presentation Outline

- Introduction
- Background
- Research Goal
- Design and Implementation
- Experimental Results
- Discussions and Conclusions

Introduction - 1

- Wafer Manufacturing & Semiconductor Circuit Probe (CP) Testing
- Building Wafermaps : Binary map indicating erroneous chips

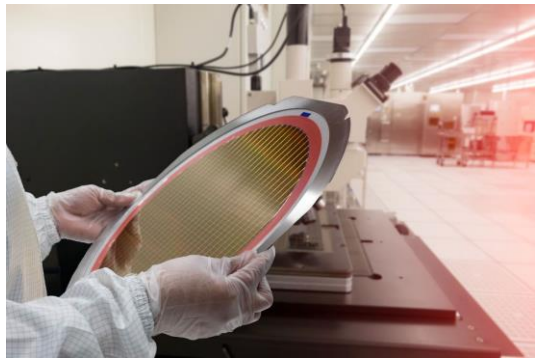


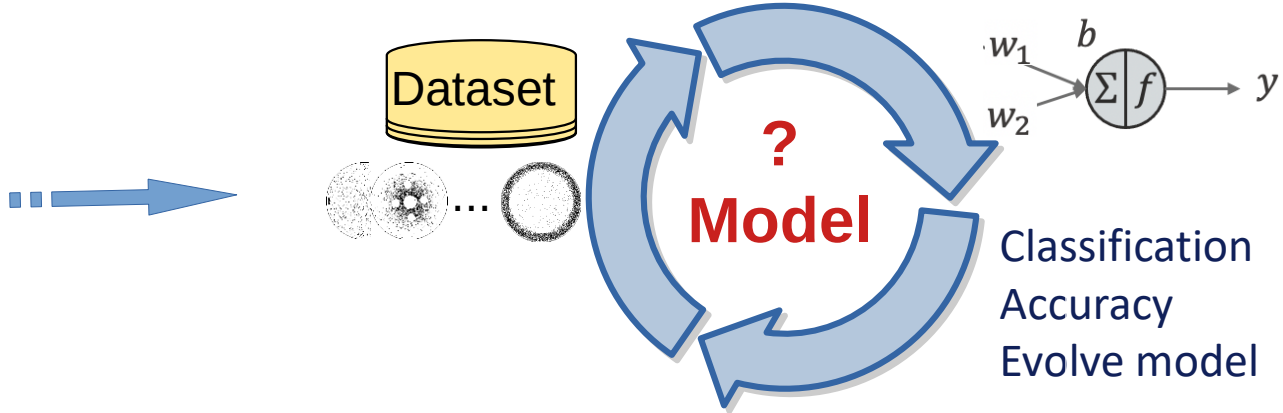
(Courtesy of STMicroelectronics)

- Identify root causes of process shifts

Introduction - 2

- Human Quality Control vs Automatic Classification





- Leveraging Machine Learning & Monitoring Performance
  - Analysis at tester throughput
  - Work that can and must be automated
- Simplified Neural Network for Low Power Consumption
  - Runs on tester at low energy/area footprint

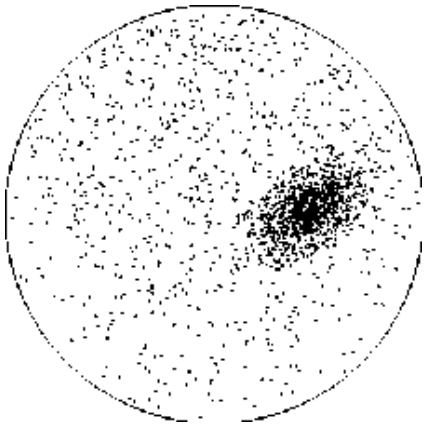
Background

Model Definition

- Data-set : 121,550 images,
  58 wafer fault classes
  (Multi-class Classification problem)

- Machine Learning & High-Level API platforms

TensorFlow Keras

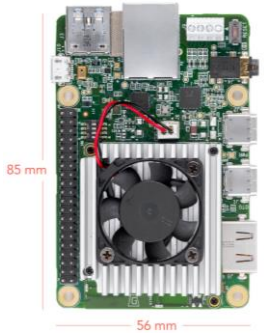


Experiments

- STMP1

- Google Coral TPU

Coral



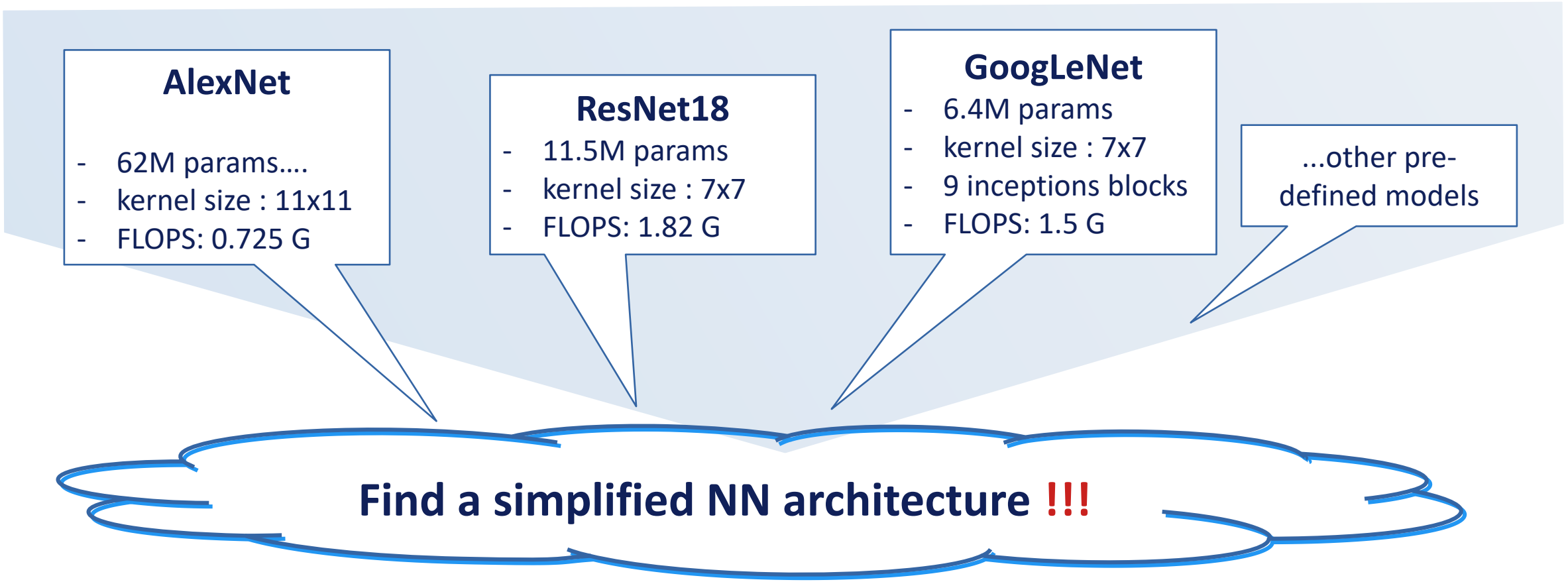
Validation & Results

- Model Predictive Accuracy
- Power Efficiency
- Real-time Performance

CLUSTER-BIG

CLUSTER-BIG

Research Goal



- ✓ Problem somehow relatively simple: binary wafermaps
- ✓ Goal to target **small edge devices**

Design & Implementation - 1

Data Pre-Processing

- Data-set already well-balanced due to pre-processing from ST (401 x 401 px)
- Size reduction (224 x 224 px)
- Binarization
- Rotation
- Notch removal

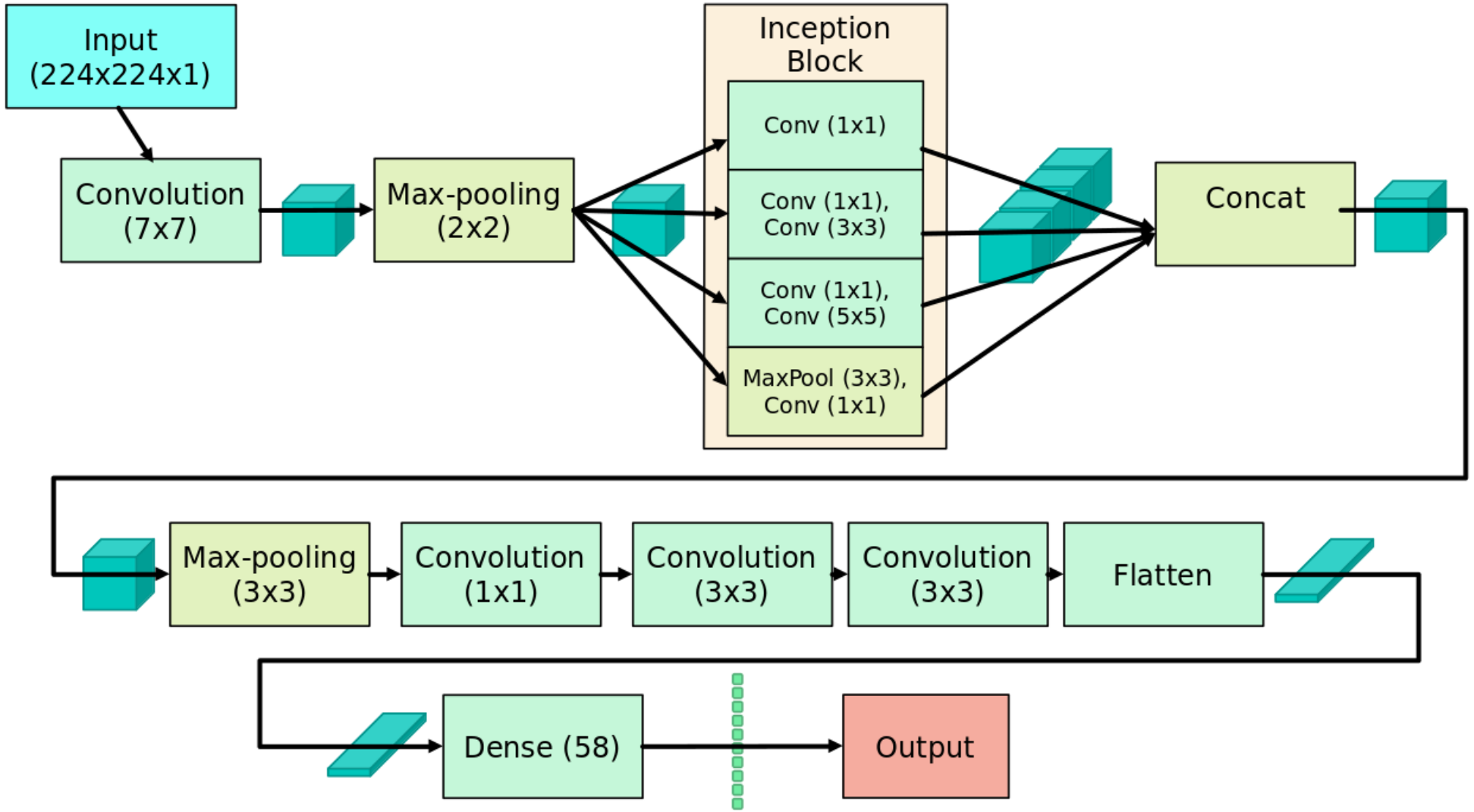
Model Definition

- 478,150 parameters
- 17 layers: 10 learnable layers
- Model size: 2 MB, Model FLOPS: 0,125

Quantization Techniques

- 8INT Post-Training Quantization
- Model size: 498 KB

Design & Implementation - 2

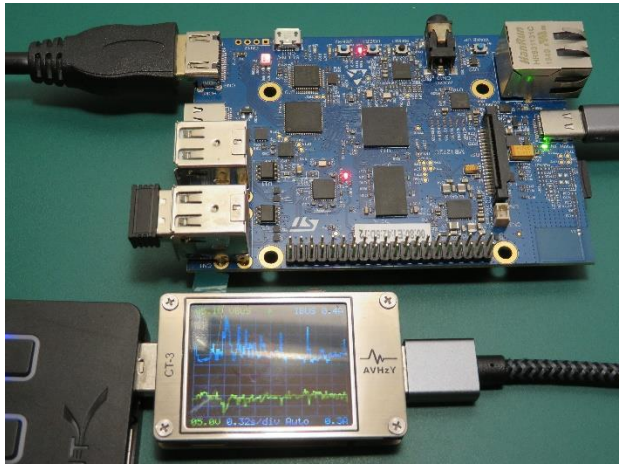


Experimental Results – 1 : Set-up

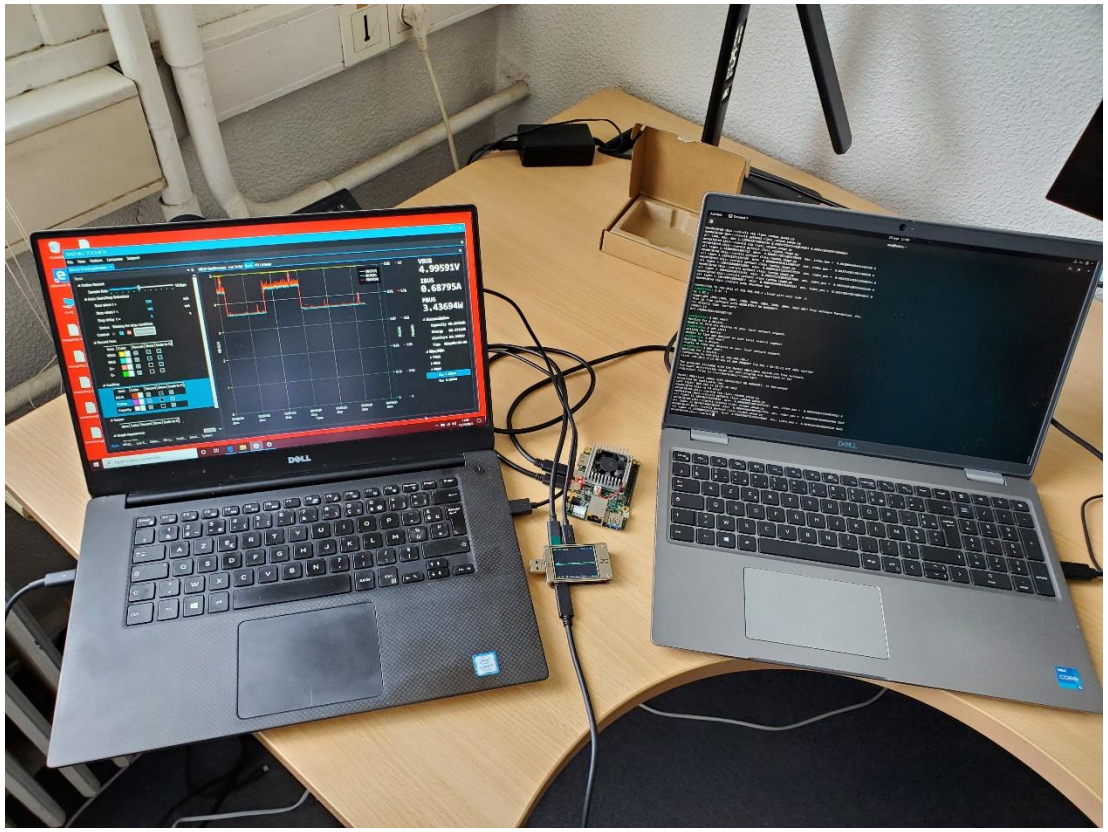
Power Measures

- AVHzY CT-3 USB 3.1 Power Meter
- Whole system: measure on power supply

STMicro MP1



Google Coral



Experimental Results – 2 : Throughput and accuracy

CPU
X86-64
7.5 inferences/s

MP1
Cortex A7
5.5 inferences/s

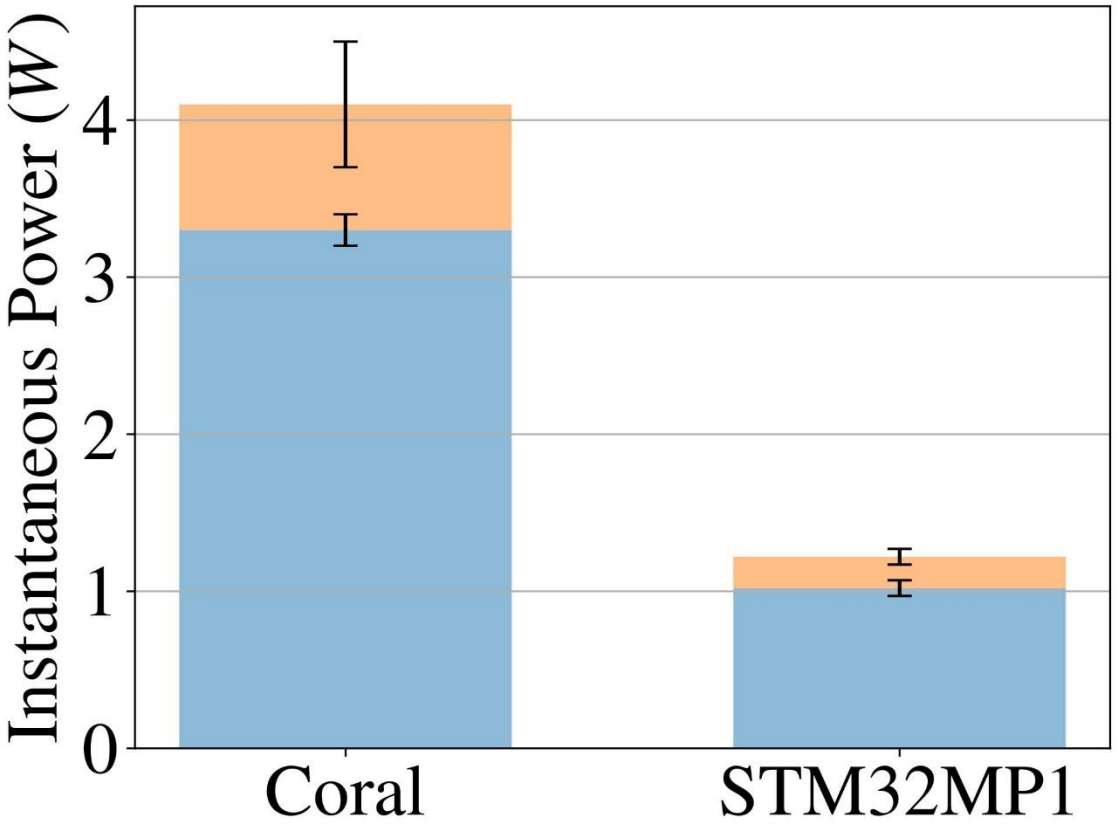
Coral
TPU
902 inferences/s

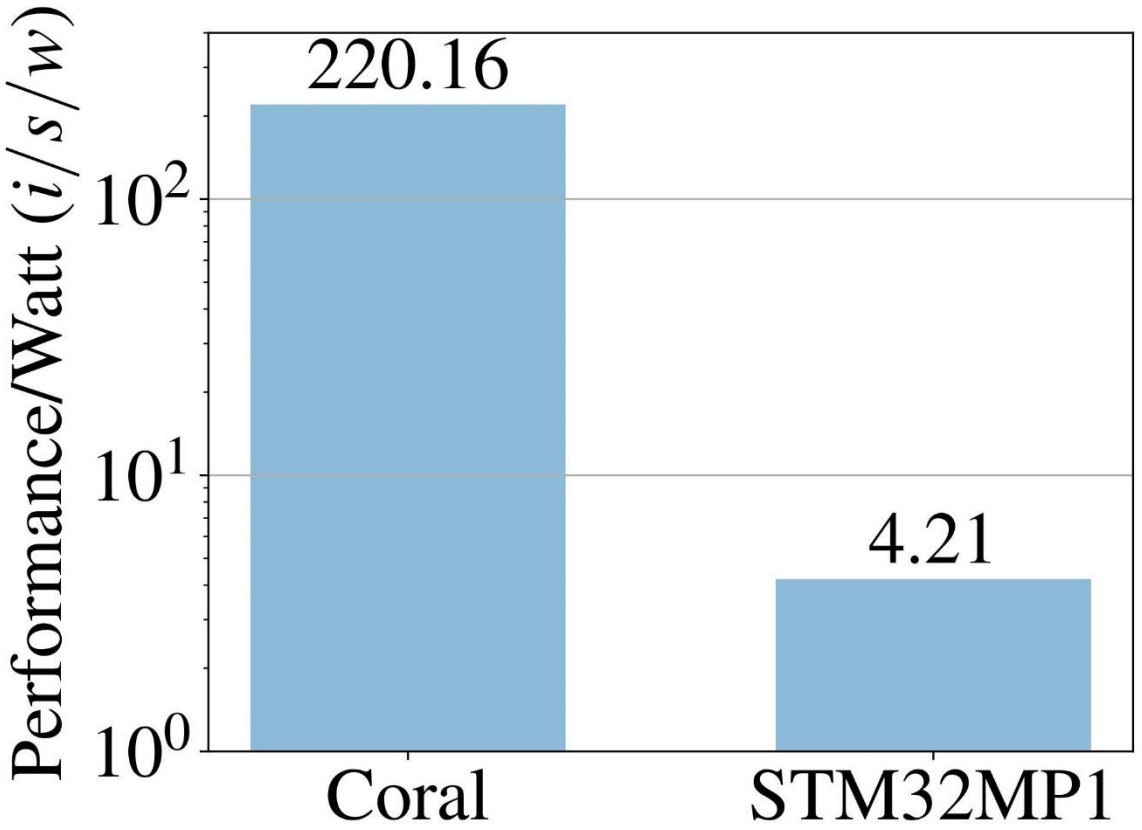
FP-32 : 99.84%

INT8 : 96.18% (PTQ)

Experimental Results – 3 : Power and power efficiency

Performance and power efficiency





Discussions & Conclusion

- Application specific neural network architectures :
  - Take benefit from binary maps to define ad-hoc NN architecture
  - Suitable for hardware devices limited in computation, memsize and power
- Leverage quantization techniques to use 8-bit weights with reasonable accuracy loss at no development cost
- Benefit from legacy HW acceleration: SIMD on processor, TPU when available
- Reach accurate real-time classification within a few watts
- ➔ Although not an embedded system per se, power efficiency matters!

Event Organisers

The Key Digital Technologies Joint Undertaking - the Public-Private Partnership for research, development and innovation – funds projects for assuring world-class expertise in these key enabling technologies, essential for Europe's competitive leadership in the era of the digital economy. KDT JU is the successor to the ECSEL JU programme. www.kdt-ju.europa.eu

The AI4DI project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the national authorities. www.ai4di.eu

The TEMPO project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Switzerland. www.tempo-ecsel.eu

The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.andante-ai.eu

Thank You

For your attention

@ frederic.petrot@univ-Grenoble-alpes.fr