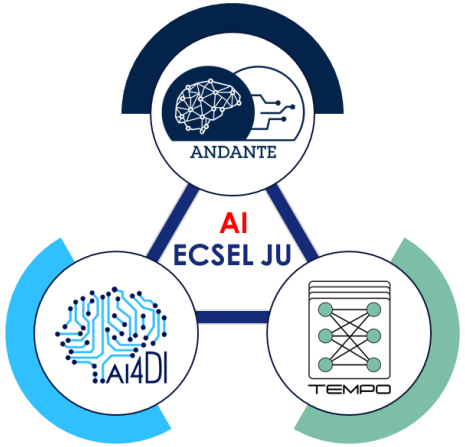
International Workshop on Embedded Artificial Intelligence Devices, Systems, and Industrial Applications (EAI)

MILAN
19-22 SEPTEMBER 2022
ESSDERC
ESSCIRC

Milan, Italy 19 September 2022

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



Tools and Methodologies for Edge-AI Mixed-Signal Inference Accelerators

Loreto Mateu, Johannes Leugering, Roland Müller, Yogesh Patil, Maen Mallah, Marco Breiling, Ferdinand Pscheidl

MILAN 19-22 SEPTEMBER 2022 ESSDERC ESSCIRC

19 September 2022 Milan, Italy

Copyright © 2022

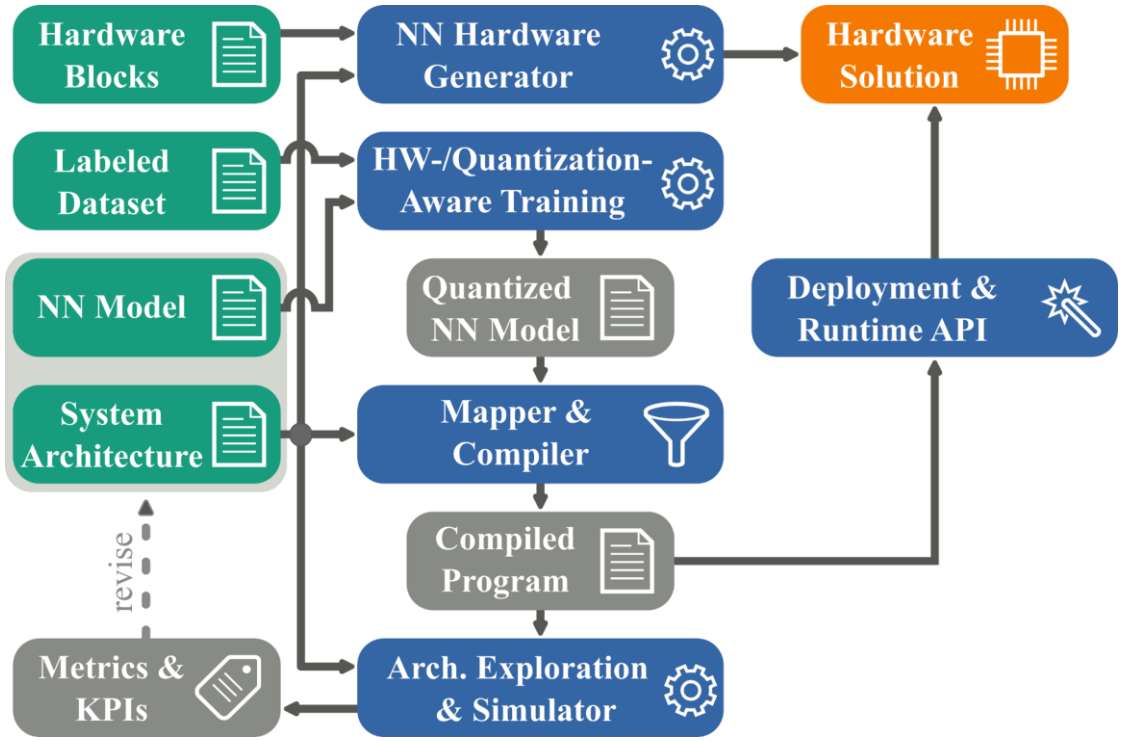
Outline

- Introduction
- Hardware-Aware Training
- Mapper & Compiler
- Architecture Exploration & Simulation
- NN Hardware Generator
- Conclusions and Future Work

Copyright © 2022

Introduction

- Custom inference accelerator ASICs provide better KPIs than off-the-shelf hardware for edge applications
- For achieving best performance with NN ASICs, hardware/software co-design is mandatory → Dedicated tools and workflows required
- Tools provide
  - Hardware-aware training
  - Automatic hardware generation
  - Instruction generation for the ASIC
  - Simulation and estimation of KPIs



Copyright © 2022

Hardware-Aware Training (HAT)

Quantization-Aware Training (QAT)

HAT extends on Xilinx Brevitas that injects quantization into the PyTorch graph.

Fault-Aware Training (FAT)

Any variation (e.g. noise, bit errors) can be injected to input, weights, bias and/or output of any layer.

Model Export

HAT exports the NN model into a custom ONNX format including quantization details.

Low Memory Footprint

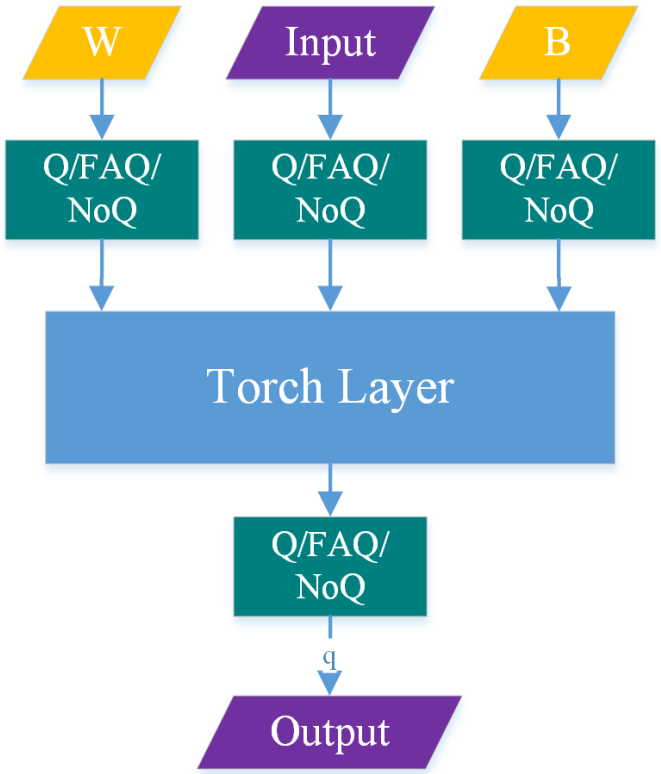
QAT enables training with very low bit resolution. Thus, NNs have lower memory footprint and energy consumption.

Robust NNs

Accuracy achieved by the inference accelerator matches accuracy from training.

Model Exchange

HAT produces a standard PyTorch graph with attached quantization parameters which can be trained, saved, loaded, and/or retrained. Thus, the exported NNs can be easily integrated in any workflow/tools.



Copyright © 2022

Hardware-Aware Training (HAT)

- Use case:
  - voice activity detection
  - Output: Voice/No Voice
  - Network: DNN: 5 conv and 2 FC layers
- QAT
  - 8-bit bias and 30-bit partial sum used
  - Quantization of weights and activation function with different bit-width
- HAT
  - Tolerance to hardware variations
  - DNN with 9-bit weights and 8-bit activation is trained/tested with:
    - Weight variation
    - ADC quantization error

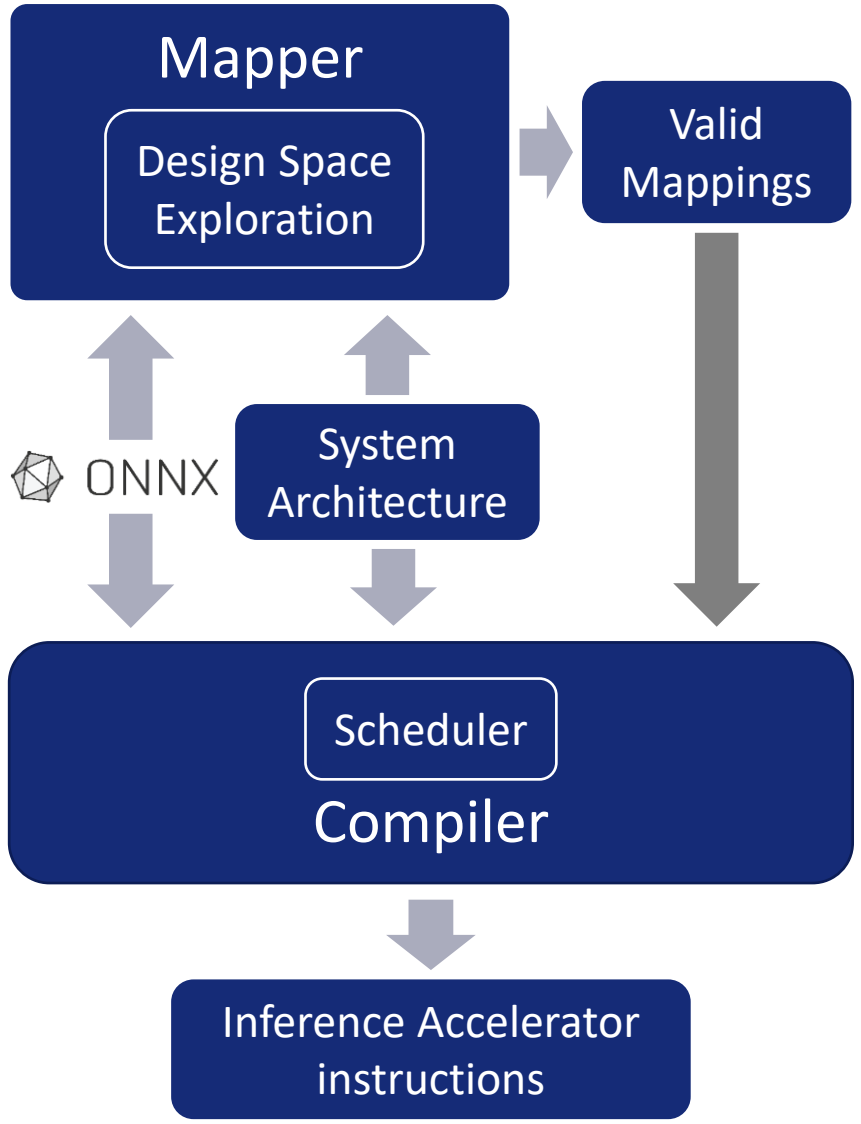
Activation \ Weight	Floating-point	9	5	3
Floating-point	91			
8		89.2	89.2	88
4		83.6	84.6	83.8

Training \ Testing	no variation	variation
no variation	88.7	54
variation	82.6	83

Copyright © 2022

Mapper & Compiler

- Bridges the gap between neural network quantized model and its deployment on the hardware
- Mapper -> Allocates the neural network model into the processing elements and cores
- Scheduler -> Assigns resources and controls data movement between the cores
- Compiler -> Converts scheduled mapping to instructions

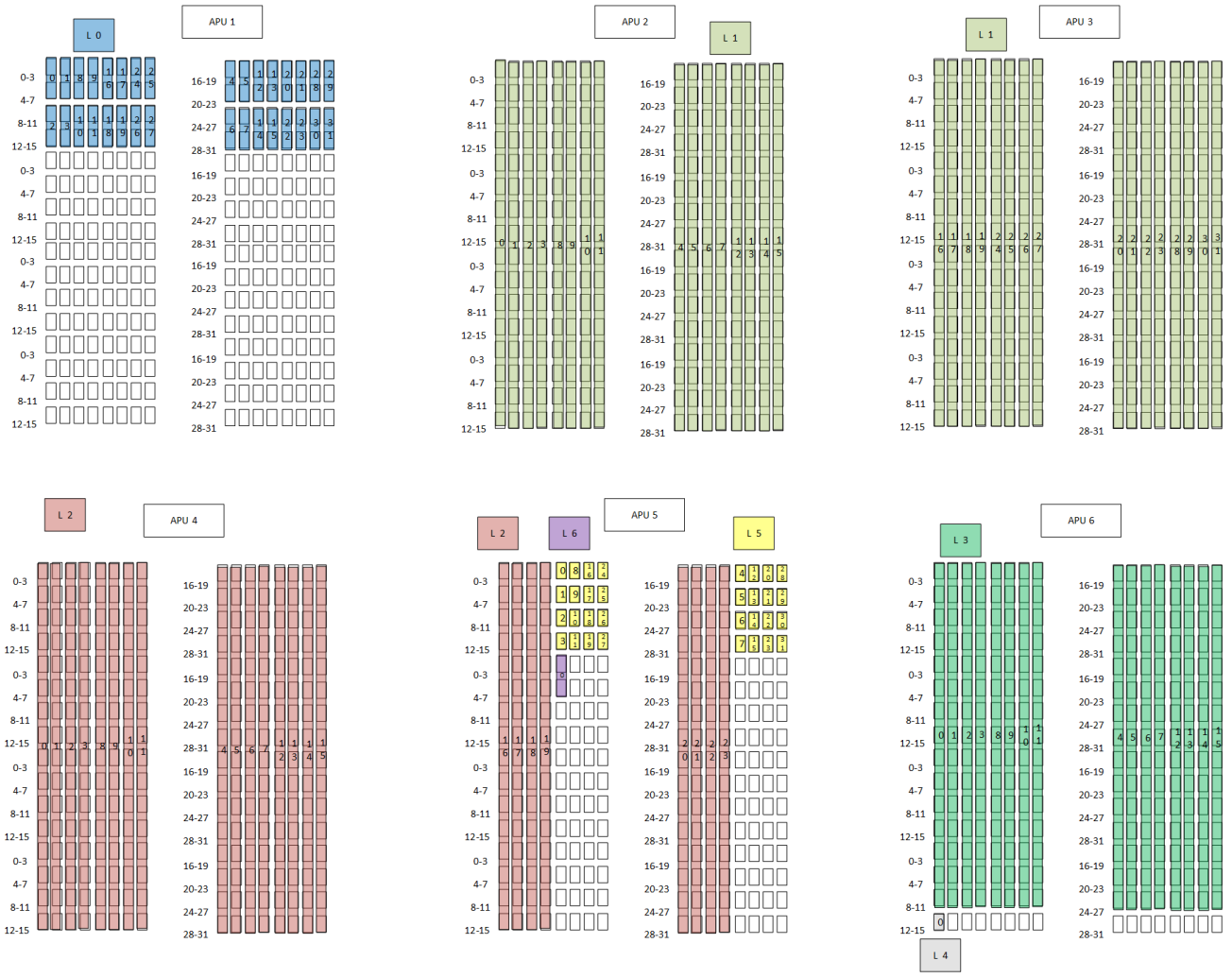


Copyright © 2022

Mapper & Compiler

- Mapping for In-Memory Computing can follow different strategies
  - Optimized data movement
  - Maximum utilization of hardware resources
- Mapper has a huge impact on KPIs like
  - Throughput
  - Energy consumption
- Layers are mapped into cores, where the filters containing the weights are distributed

VAD Network Mapping



Copyright © 2022

Architecture Exploration & Simulation Tool

Heterogeneous System C Model

- **Loosely timed** component models (not timing critical)
- **Cycle accurate interface** models (timing critical)

Simulation of KPIs

- **Latency, power, duty cycles** of different parts can be estimated per inference

Setup Files

- Hardware metrics
- Architecture configuration
- Compiled instructions
- Input data

High simulation speed

- Verification at individual module level
- Verification at system level
- Verification of Mapper & Compiler

KPI Optimization

- Impact on performance due to architecture changes can be evaluated quickly
- Architecture exploration, verification and optimization

Copyright © 2022

Architecture Exploration & Simulation Tool

Architecture Exploration finding best combination of ADC resolution and gain

ADC bit-resolution	max. ADC gain	Accuracy*	PASS/FAIL
8 bit	4x	63%	✗ FAIL
9 bit	1x	52%	✗ FAIL
9 bit	2x	54%	✗ FAIL
9 bit	4x	82%	✓ PASS
10 bit	1x	81%	✓ PASS

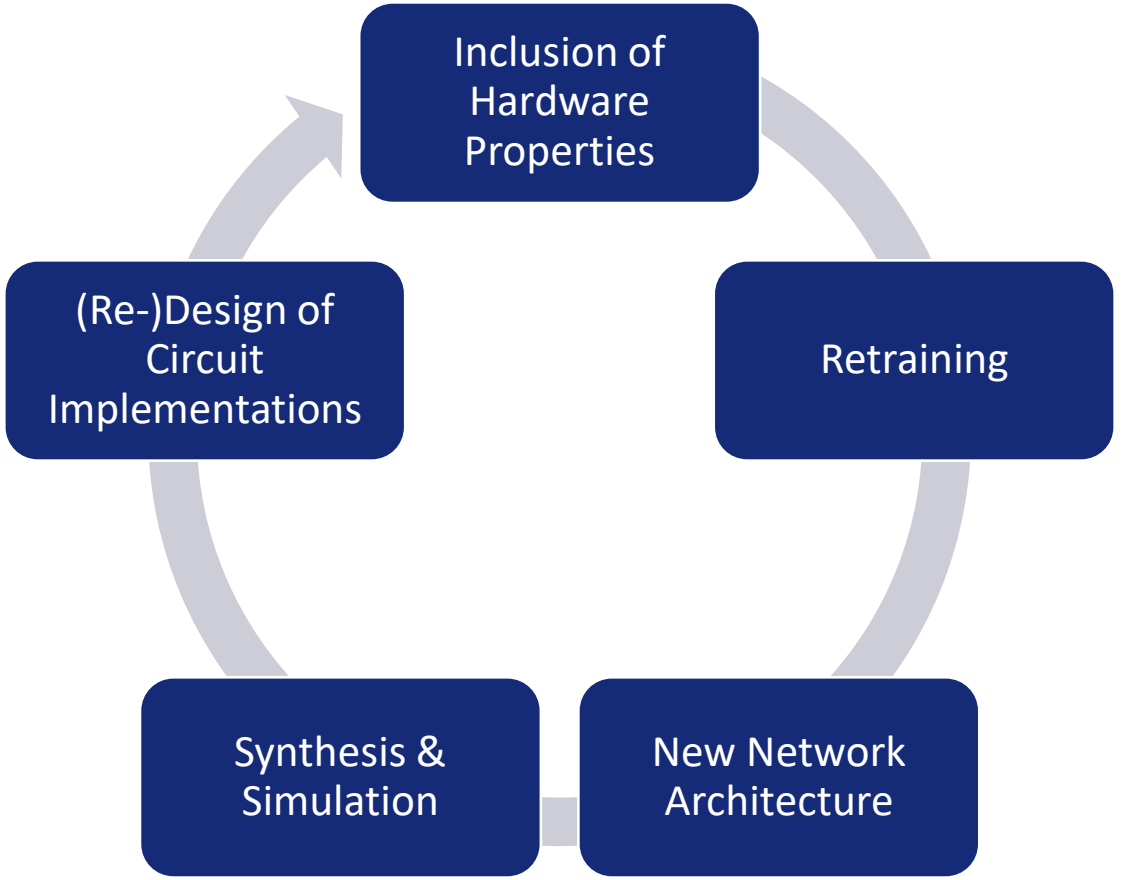
Best trade-off for this network

* Procedure: executed over 1000 inferences for VAD network in behavioural simulation

Copyright © 2022

Neural Network Hardware Generator Tool

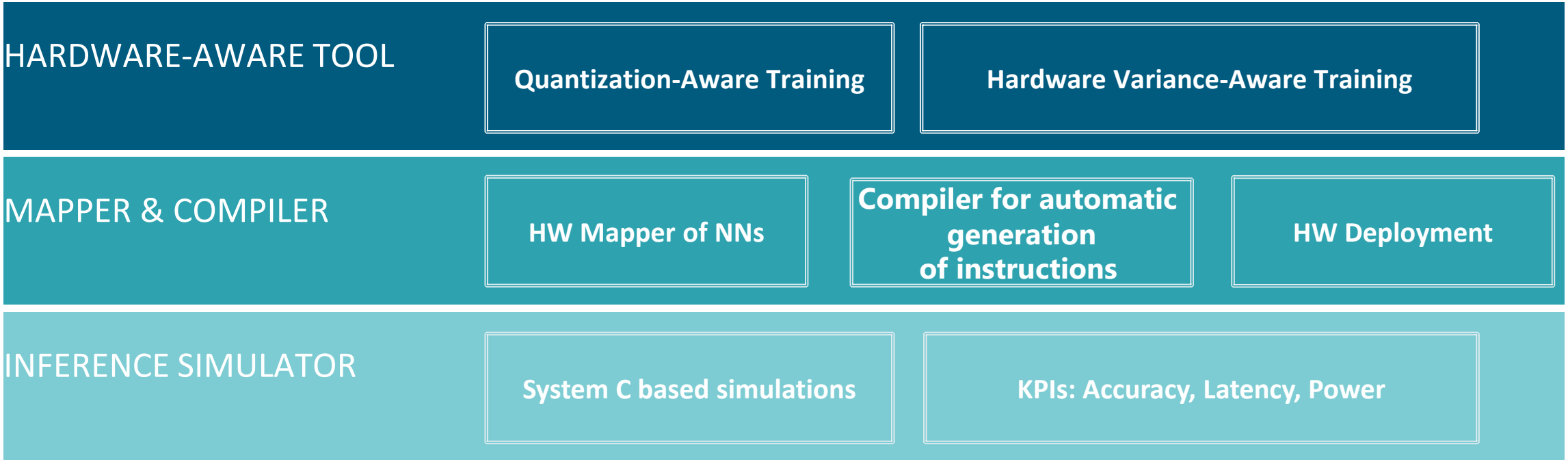
- Parallel development of network algorithm and circuit implementation
  - Regular changes
  - Reduction of design time
- Neural networks are regular structures
  - Ideal for automatic generation
- Stability of the design process
  - Multiple millions of circuit nodes
  - Avoid errors during circuit design
- Inclusion of improved circuit implementations
  - Synthesis with new circuit implementations
  - Improved performance



Copyright © 2022

Conclusions and Future Work

- For the design of a custom inference accelerator, a tool chain is required for allowing hardware/software co-design
- Users to adopt custom accelerator ASICs need a tool chain
- Integration of tools in a cohesive general framework



- Explore further mapping strategies, more architecture options and their impact in KPIs

Copyright © 2022

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. ANDANTE has also received funding from the German Federal Ministry of Education and Research(BMBF) under Grant No. 16MEE0116 and 16MEE0117.

Copyright © 2022

Thank You

For your attention

@ loreto.mateu@iis.fraunhofer.de

Copyright © 2022