tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

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

Maen Mallah – Senior engineer, Fraunhofer Institute for Integrated Circuits IIS
Roland Müller – Senior engineer, Fraunhofer Institute for Integrated Circuits IIS

December 12, 2023

www.tinyML.org

Thank you, **tinyML Strategic Partners**,

for committing to take tinyML to the next Level, together

TINY ML TALKS webcast

Executive Strategic Partners

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

TINY ML TALKS webcast

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners


Build the
Future of tinyML
on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

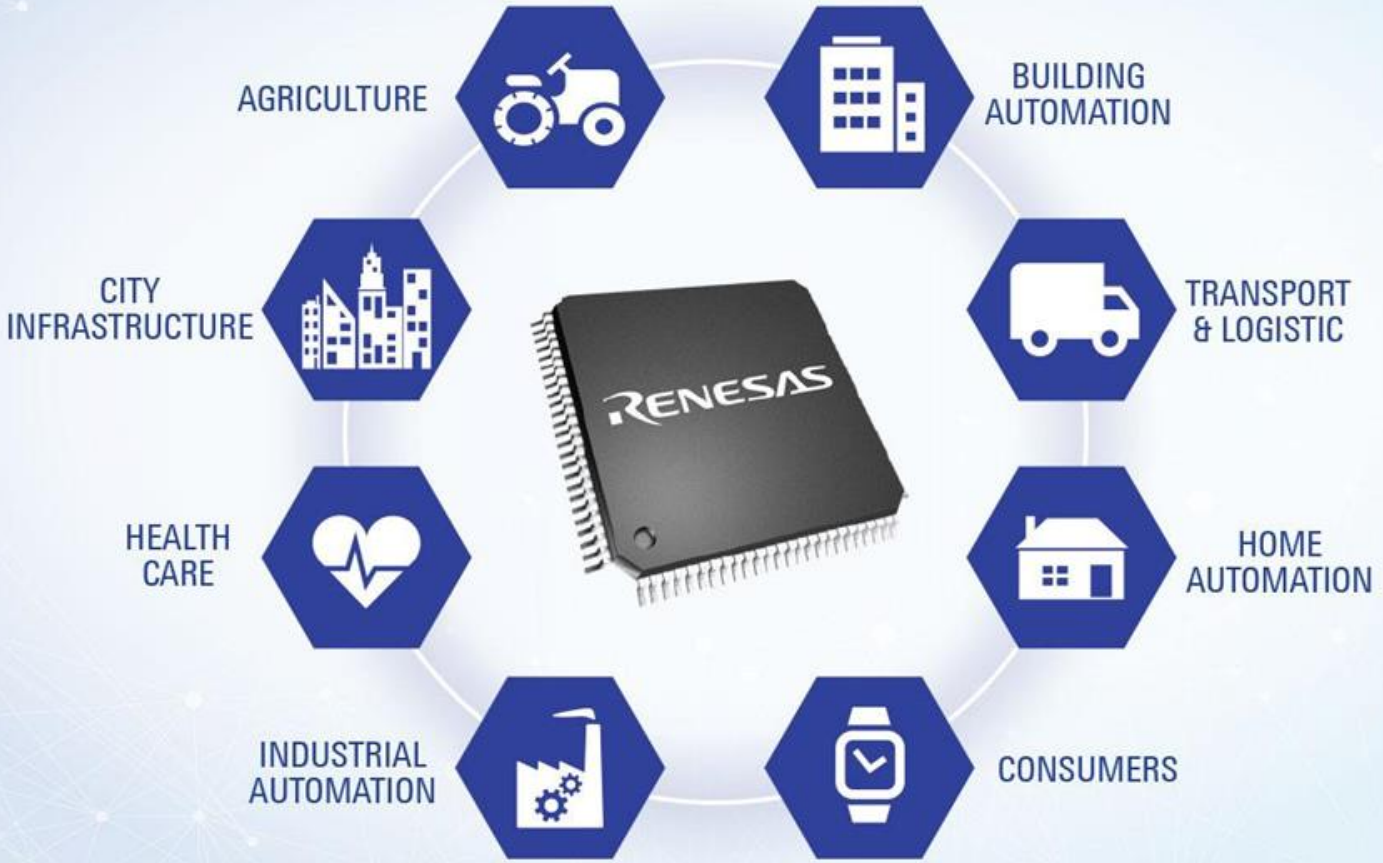
www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



renesas.com

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners

Join Growing tinyML Communities:

17.7k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

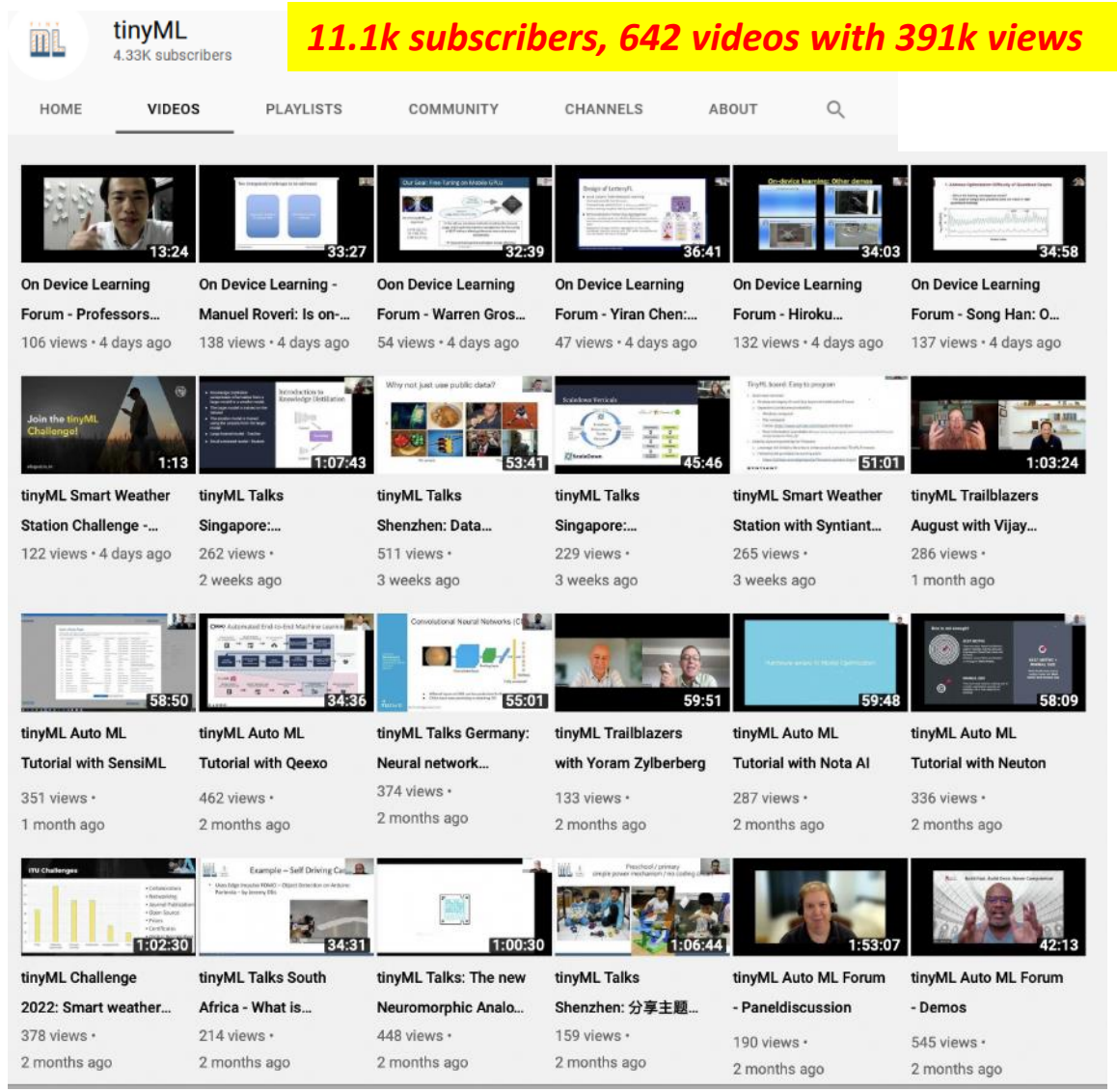
4k members
&
13k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

tinyML Research Symposium
April 22, 2024
Call for Papers

WEVOLVER

https://www.wevolver.com/article/2023-edge-ai-technology-report

TINY ML FOUNDATION

2023 Edge AI Technology Report

The guide to understanding the state of the art in hardware & software in Edge AI.

TINY ML TALKS webcast

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions

TINY ML TALKS webcast

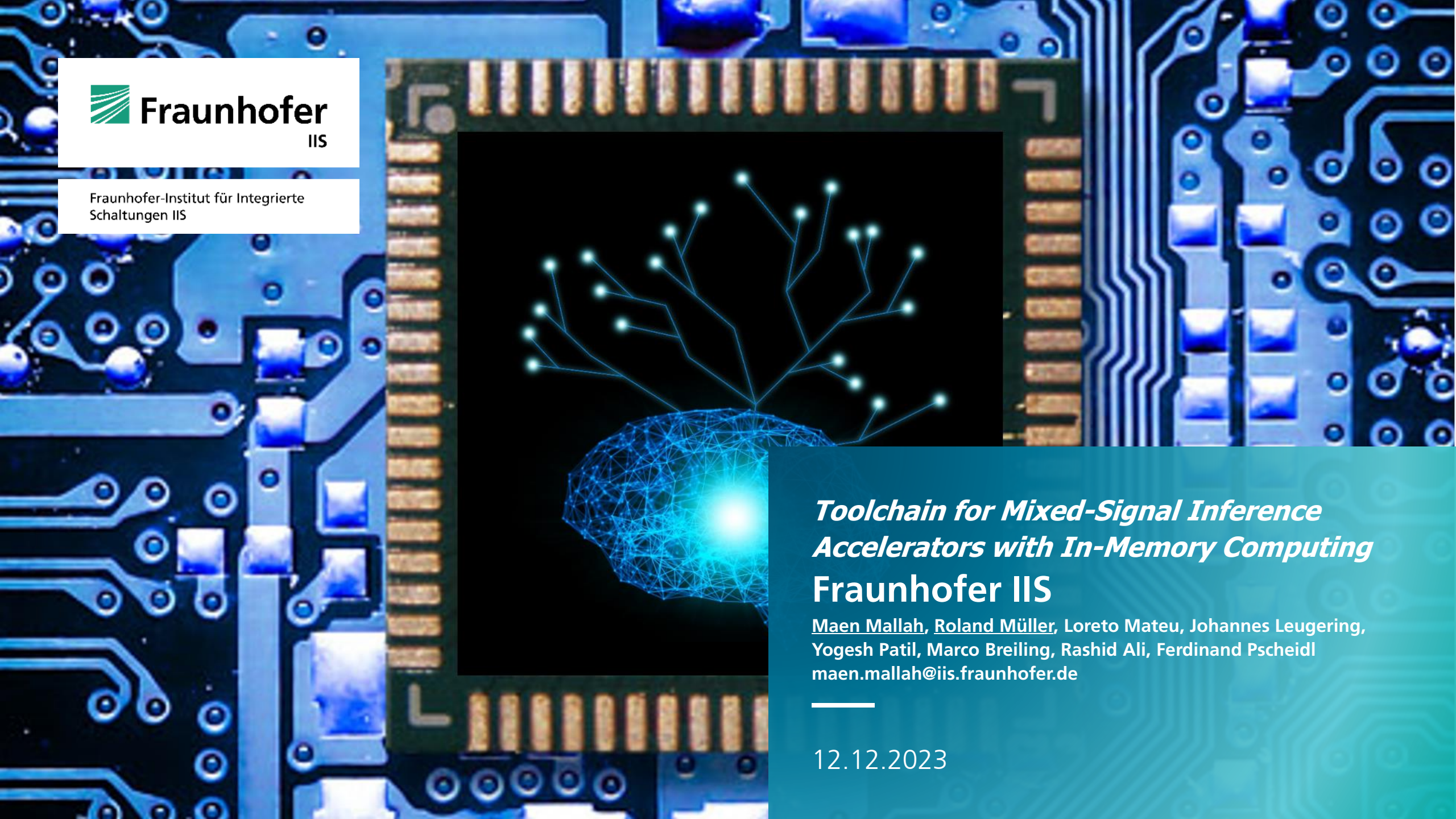
Maen Mallah

Mallah is a researcher in the area of embedded AI at the Fraunhofer Institute for Integrated Circuits (IIS). He obtained his B.Sc in Telecommunication Engineering in 2014 from An-Najah National University, Palestine and his M.Sc in Electrical Engineering in 2018 from Bilkent University, Turkey with a thesis titled “Multiplication Free Neural Networks”. In March 2018, He joined Fraunhofer IIS as an expert for eAI and focusing mainly on energy efficient NNs. His main work and interest focuses on implementing and optimizing NNs for Edge applications and designing the special SW tools required for such a task with a special focus on Quantization- and Fault-aware training.

TINY ML TALKS webcast

Roland Müller

Müller obtained his bis B.Eng. at the OTH Regensburg in 2017 and his M.Sc. in 2019 at the FAU Erlangen, both in electrical engineering. In May 2019, he joined the department of Integrated Circuits and Systems at Fraunhofer IIS, Erlangen (Germany), where he is working in the field of analog-mixed signal design of neural network accelerators and design automation for such circuits. Currently, he is pursuing his PhD. His main research interests include low power analog-mixed signal circuits, neuromorphic computing and electronic design automation.



Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Toolchain for Mixed-Signal Inference Accelerators with In-Memory Computing

Fraunhofer IIS

Maen Mallah, Roland Müller, Loreto Mateu, Johannes Leugering, Yogesh Patil, Marco Breiling, Rashid Ali, Ferdinand Pscheidl
maen.mallah@iis.fraunhofer.de

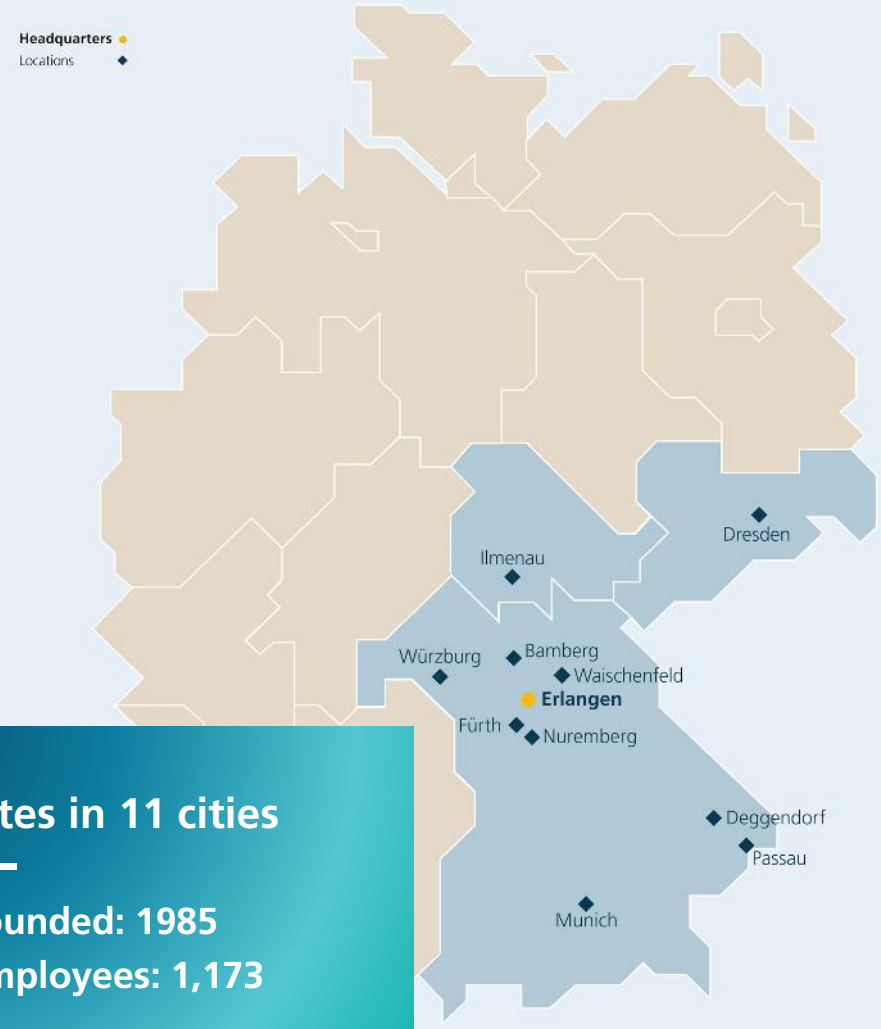
12.12.2023

Outline

- Introduction
- Mixed-signal Neuromorphic HW
- Software Tools - Workflow and Toolchain
  - Training
  - KPI Estimation
  - Mapper & Compiler
  - HW Generator
- ASIC Characterization and Demo
- Q&A

© Fraunhofer IIS

FhG IIS - Neuromorphic Computing



Sites in 11 cities

Founded: 1985

Employees: 1,173

https://www.iis.fraunhofer.de/en/ff/kom/ai/neuromorphic.html

Mixed-signal Neuromorphic ASICs for DNN/SNN

Ultra low energy consumption per Inference for longer battery life means analog MAC computation

Ultra-low processing time per inference for fast computation or high data processing

Hardware and Software co-design and Automated tool chain for achieving the targeted accuracy of the NN

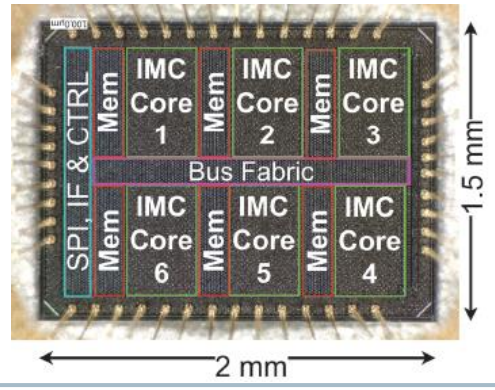
© Fraunhofer IIS

ADELIA Gen 2

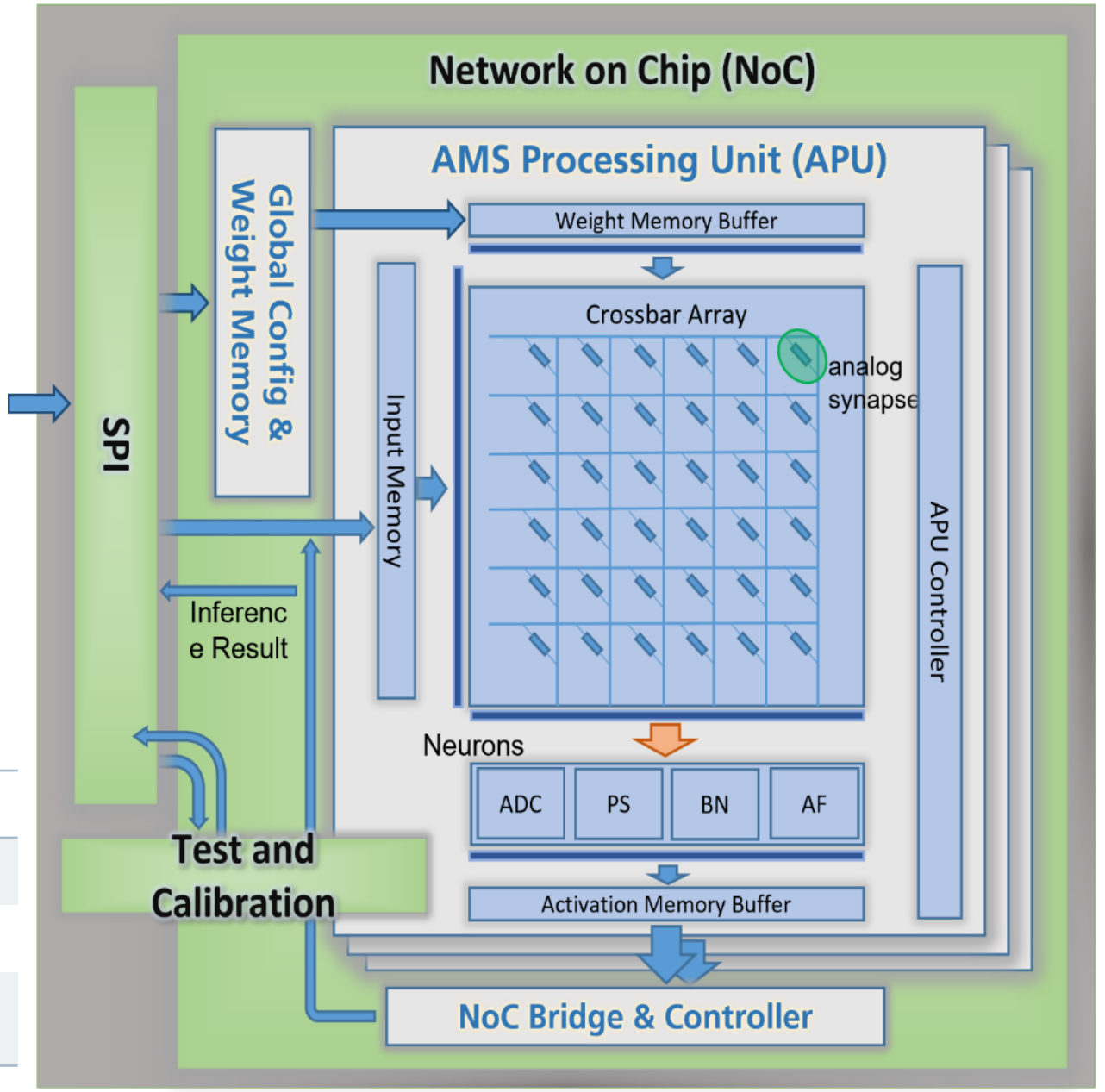
Key Features

Flexible and scalable mixed-signal architecture

- ADELIA Gen 2 supports scalable analog/mixed-signal processing cores (APUs) that can communicate using a cyclic bus fabric.
- Employs field programmable analog in-memory compute cores for fast and energy efficient MAC computation
- Digital interface using low power ADCs between APUs offers additional flexibility and efficiency
- Supports different weight and input precision up to 9 bits for each



Technology: 22nm FDSOI
Voltage: 0.8 V (core), 1.8 V (IO)
MAC Precision: 3 – 9 bit
337 TOPS/W IMC Energy efficiency (4 bit) 91 TOPS/W ASIC Energy efficiency (4 bit)



© Fraunhofer IIS

ADELIA Gen 2

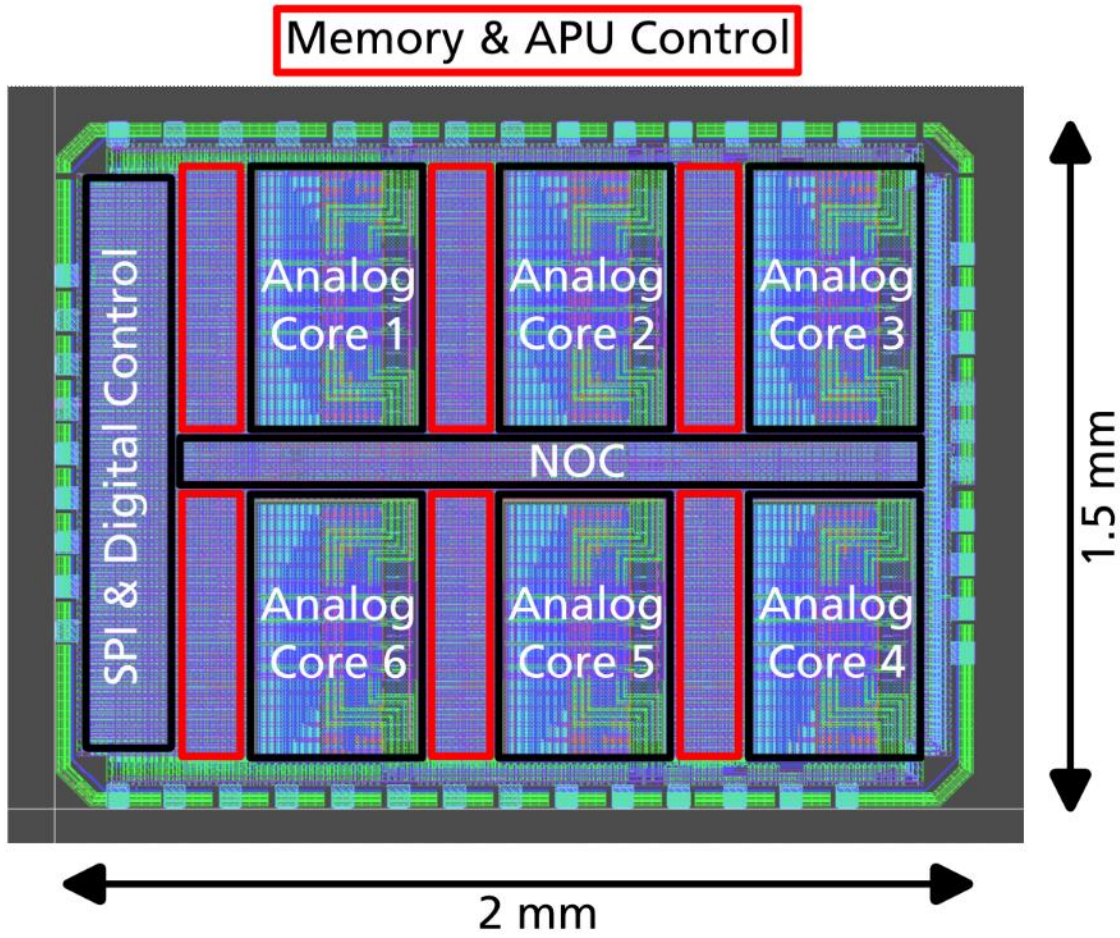
Key Hardware Parameters

ANDANTE ASIC / Network on Chip (NoC) has:

- 6 APUs, 1 KB Instruction Memory, NoC controller, digital and analog test circuits

Each APUs have:

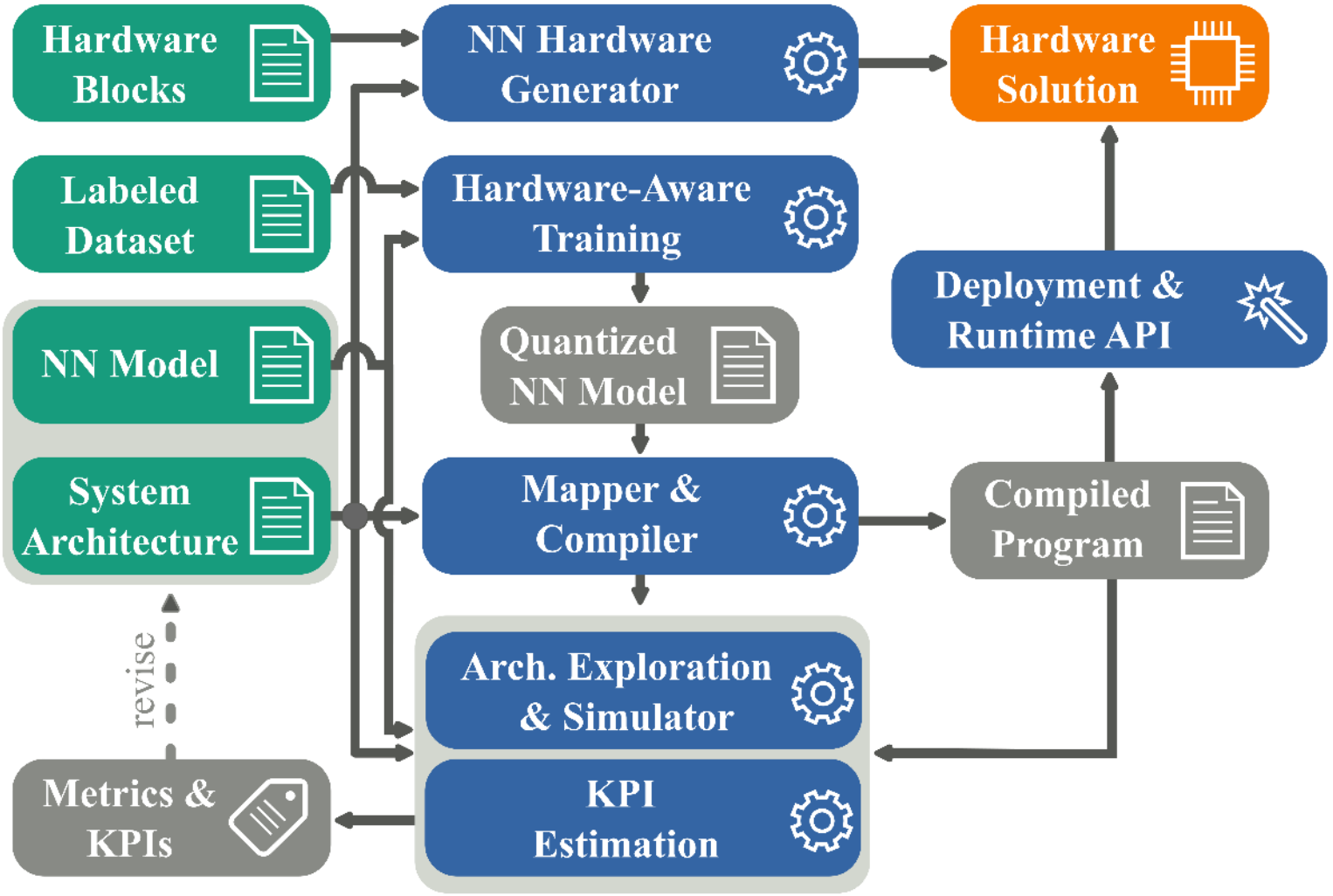
- An analog in-memory crossbar array having 256 x 64 synapses, each with 3 bit signed precision,
- Crossbar array with 8 KB distributed SRAM
- 8 KB Input memory, 8 KB weight memory, 1 KB Instruction Memory
- 32 low power ADCs, Shift and ADD blocks including Batch Norm capability



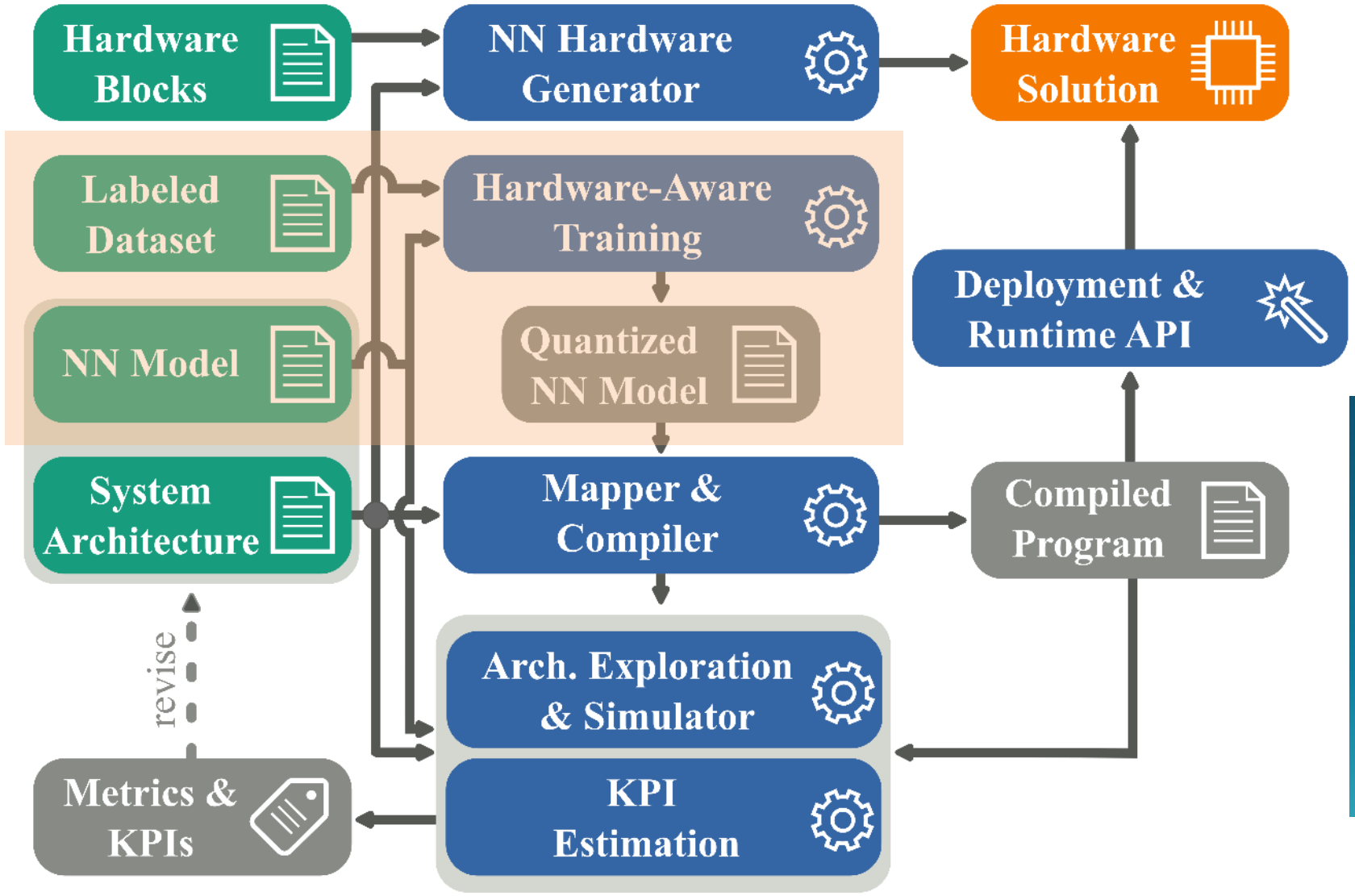
Package: JLCC68

© Fraunhofer IIS

Neuromorphic Computing Tool Chain



© Fraunhofer IIS



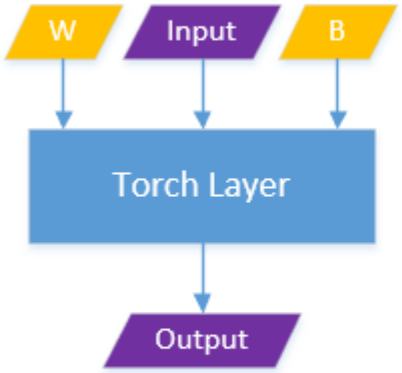
01

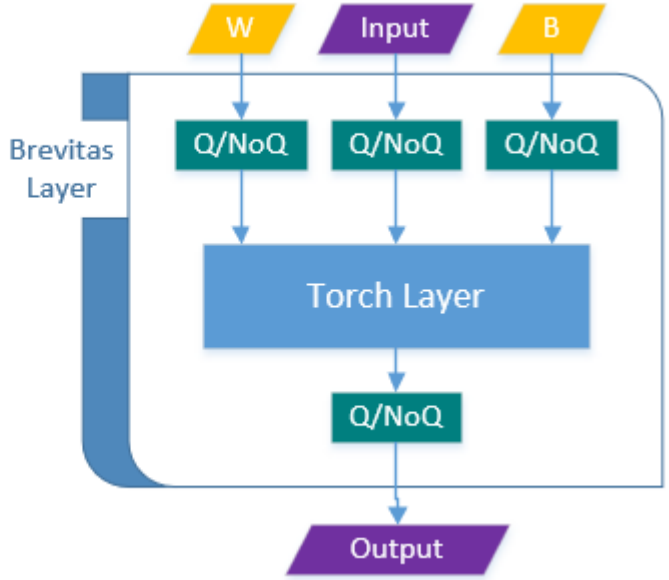
Training

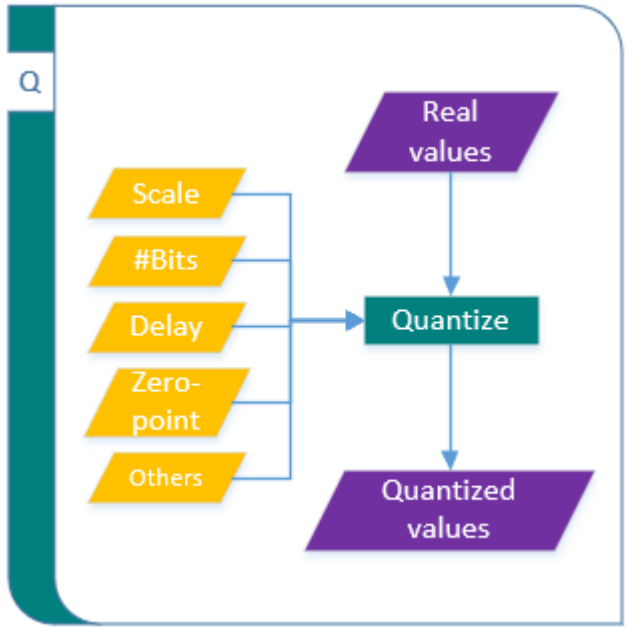
© Fraunhofer IIS

Hardware-Aware Training (HAT) Tool

Quantization-Aware Training







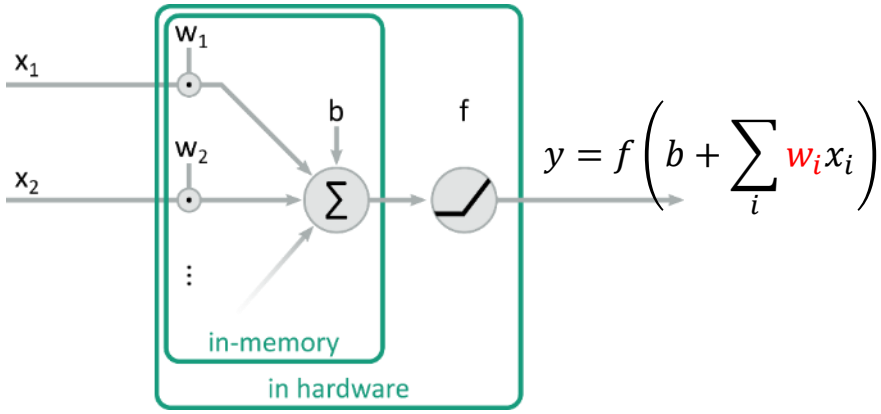
- Q: Quantizer
- NoQ: No Quantizer

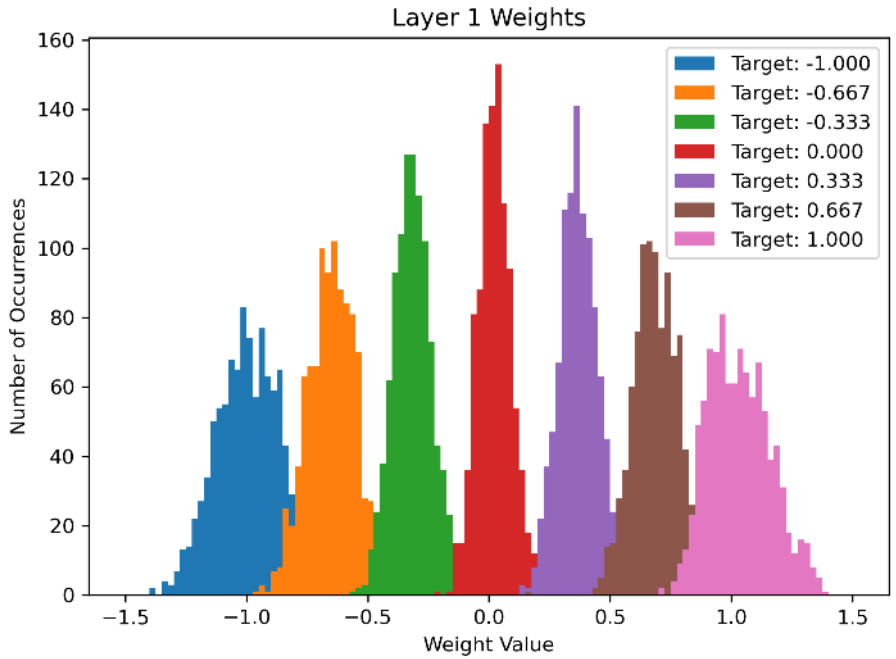
© Fraunhofer IIS

Hardware-Aware Training (HAT) Tool

Fault-Aware Training

- Accurate NN computation is necessary even with PVT variations and mismatch of weights

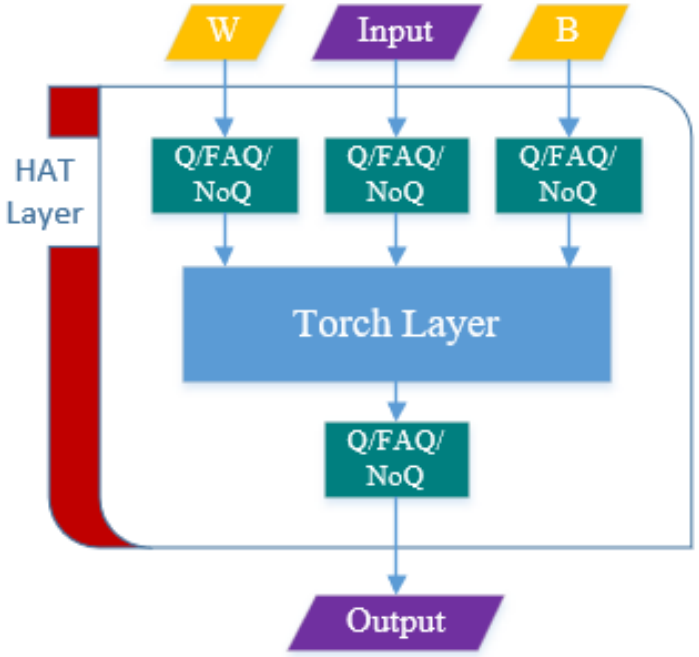


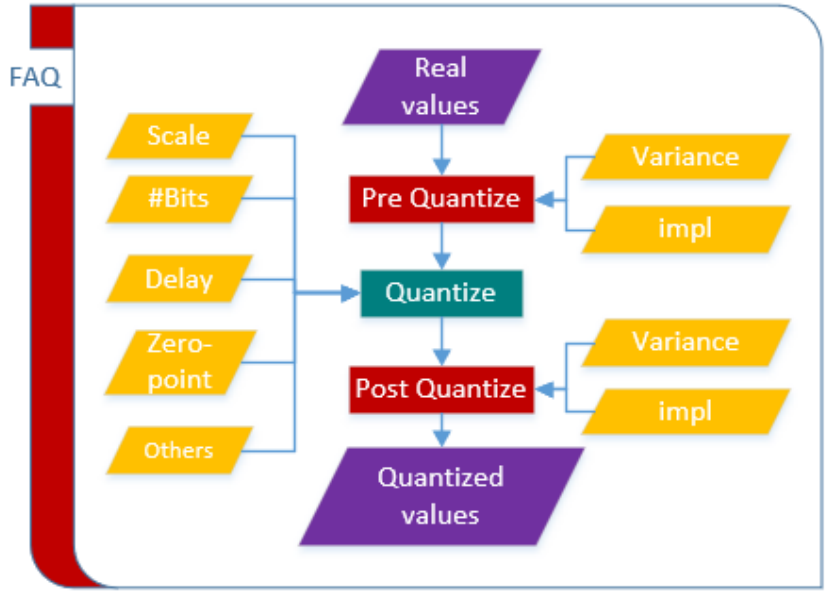


© Fraunhofer IIS

Hardware-Aware Training (HAT) Tool

Fault-Aware Training





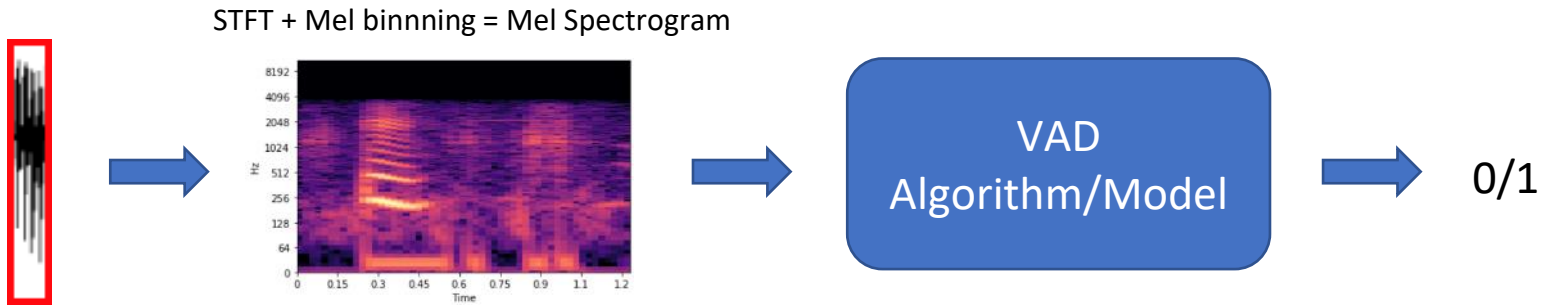
- Q: Quantizer
- NoQ: No Quantizer
- FAQ: Fault Aware Quantizer

© Fraunhofer IIS

Hardware-Aware Training (HAT) Tool

VAD use case results

- Voice Activity Detection UC
- Input: Log spectrograms of size: 13X64
- Output: Binary
- Network:
  - 5X conv2D layers and 2X FC layers
  - Params: 25.794K, MAC: 1.238M
  - Floating-point test accuracy: 89%
  - 2 types of Hardware variations:
    - Mismatch of analog synapses -> Relative variations
    - Bit error of the ADCs in the accumulation -> Absolute variations



Training \ Testing	w/o HW variations (quant only)	with HW variations
w/o HW variations (quant only)	88.0%	50.0%
with HW variations	83.1%	83.3%

© Fraunhofer IIS Public

Hardware-Aware Training (HAT) Tool

Summary

Quantization-Aware Training (QAT)

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

Low Memory Footprint

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

Fault-Aware Training (FAT)

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

Robust NNs

Accuracy achieved by the inference **accelerator matches** accuracy from **training**.

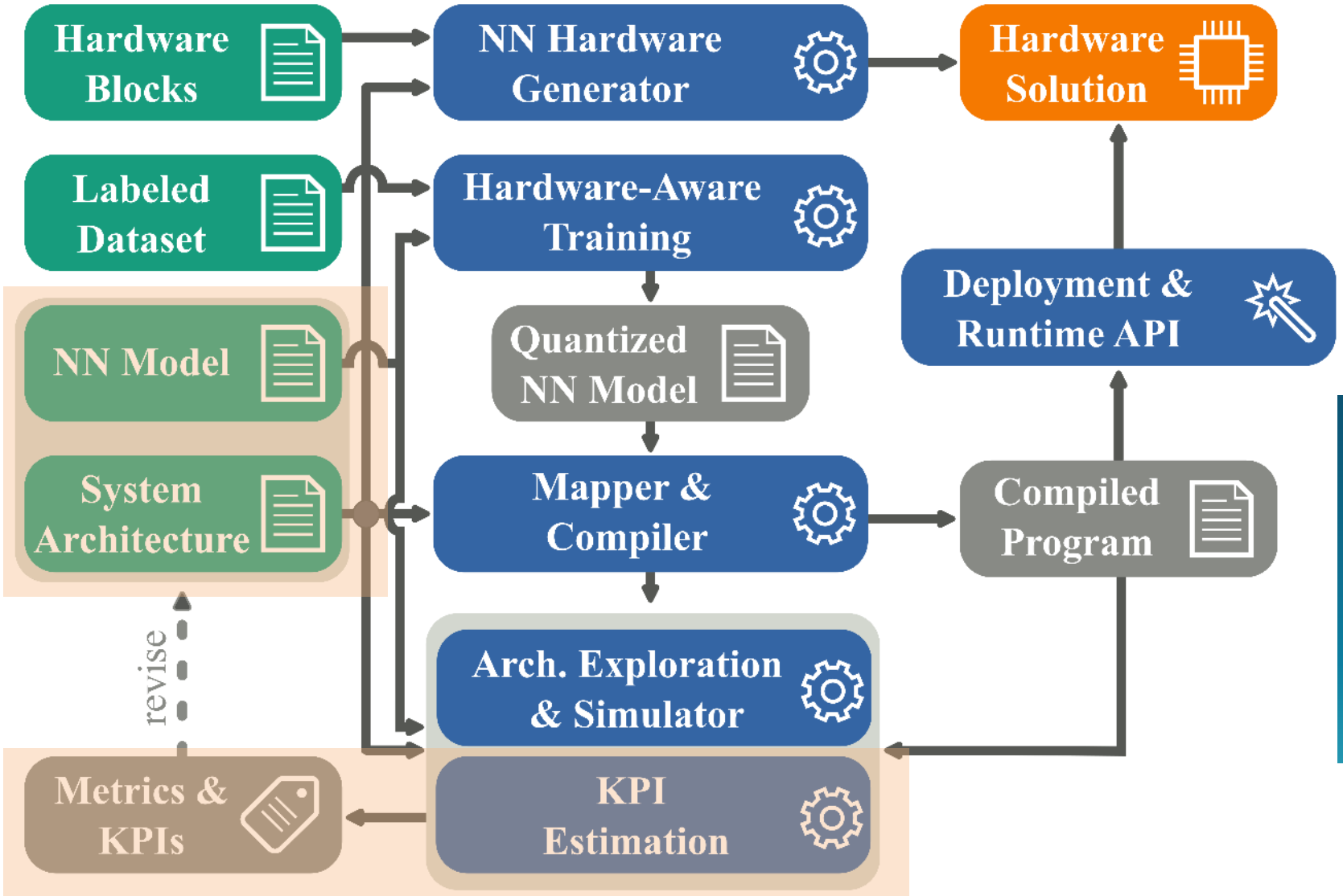
Model Export (ONNX)

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

Model Exchange

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

© Fraunhofer IIS



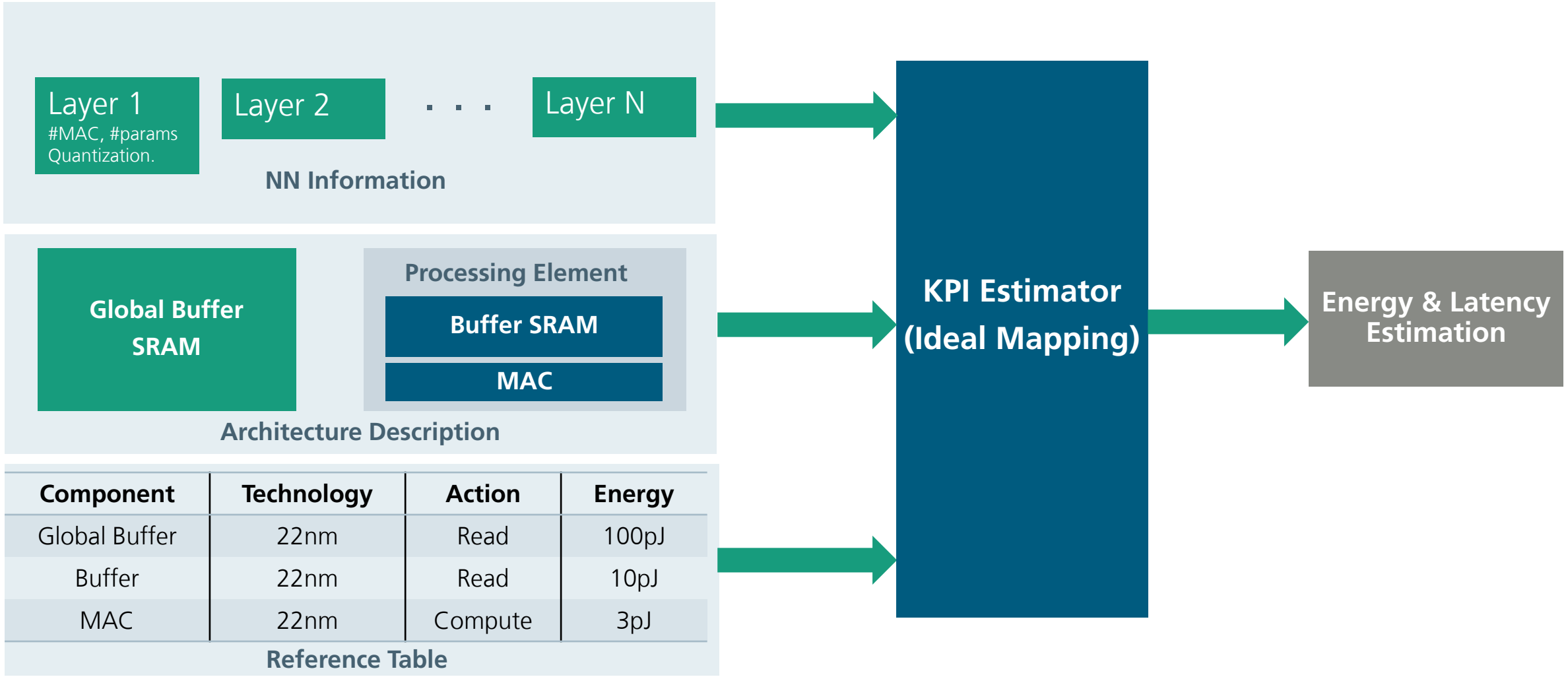
02

KPI Estimation

© Fraunhofer IIS

KPI Estimator

Mapping-aware Estimator



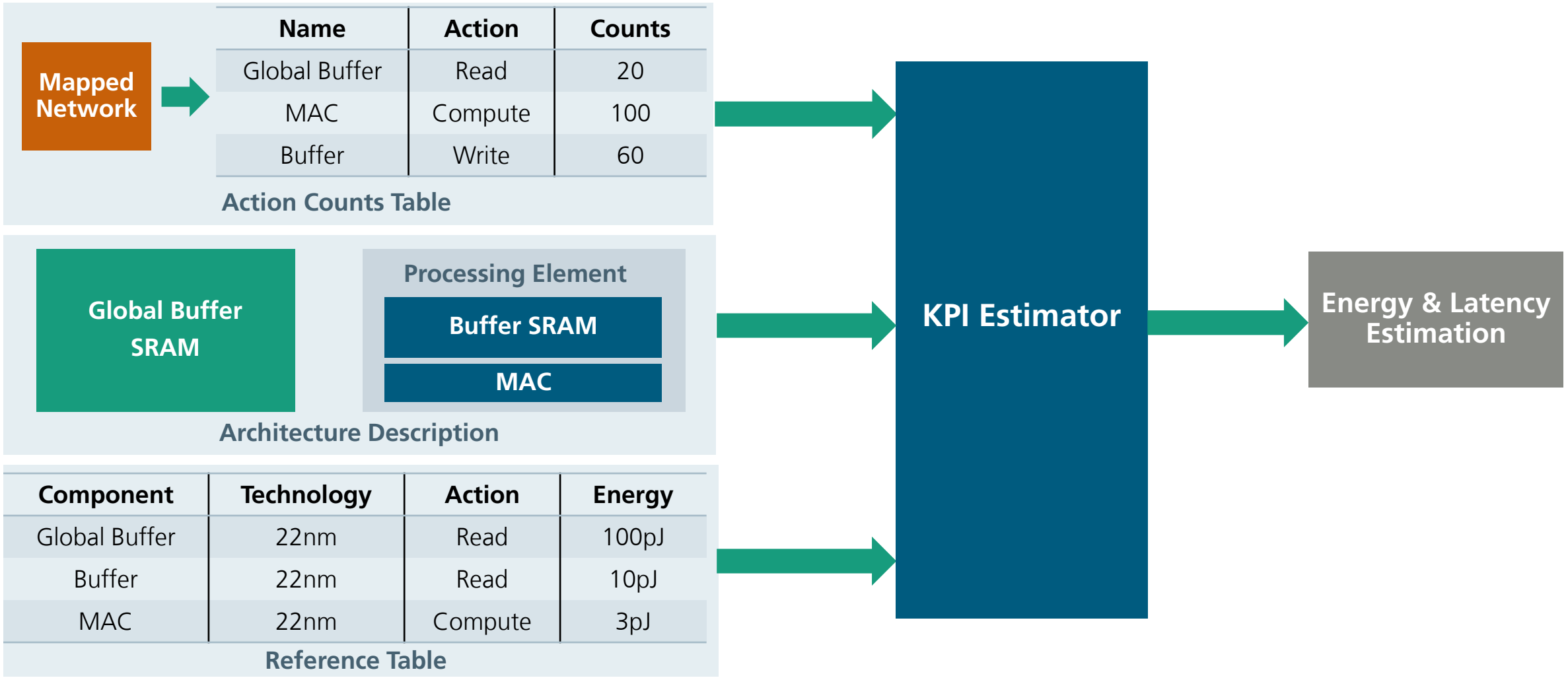
Component	Technology	Action	Energy
Global Buffer	22nm	Read	100pJ
Buffer	22nm	Read	10pJ
MAC	22nm	Compute	3pJ

Reference Table

© Fraunhofer IIS

KPI Estimator

Mapping-aware Estimator

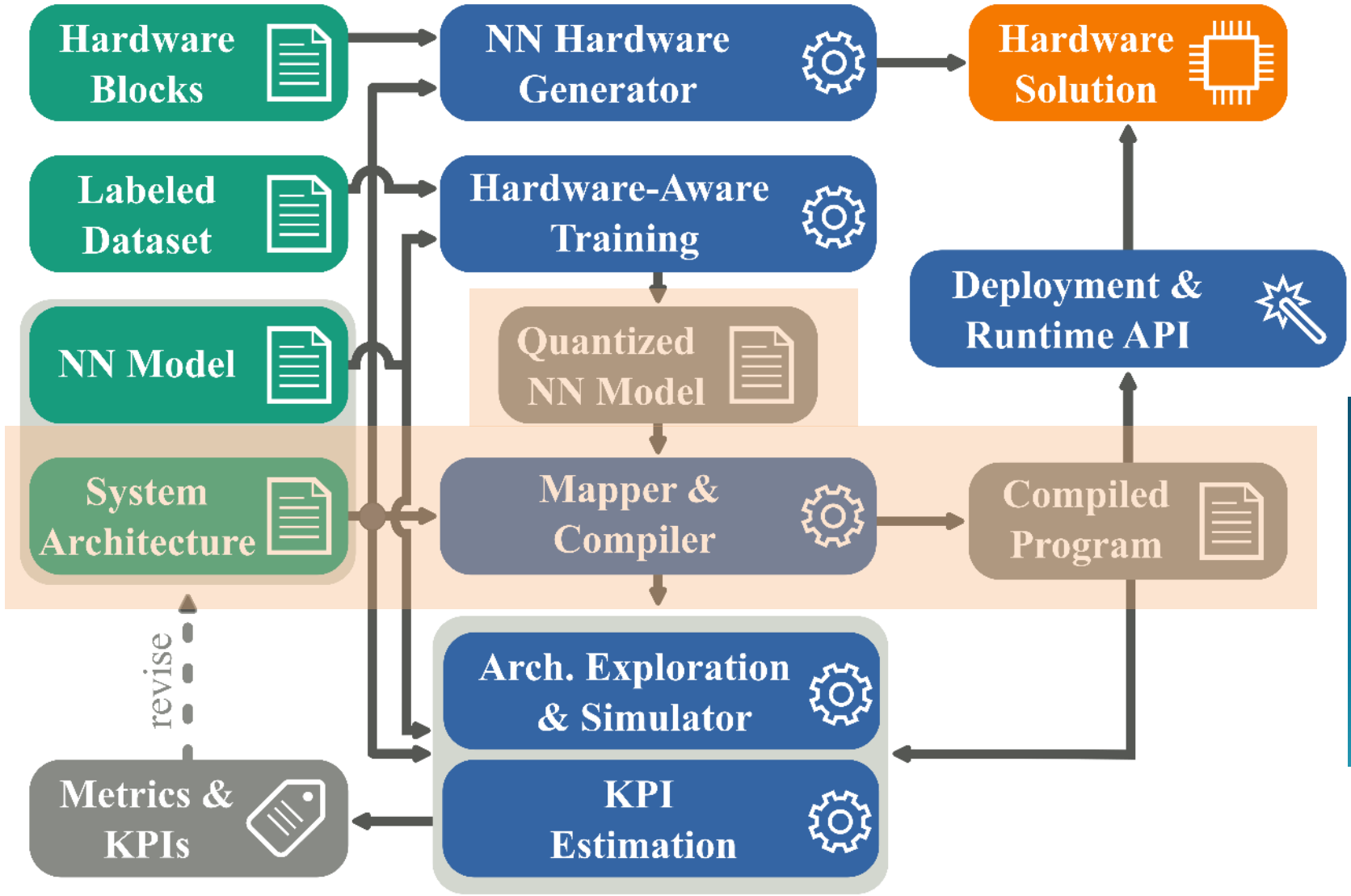


© Fraunhofer IIS

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Q&A

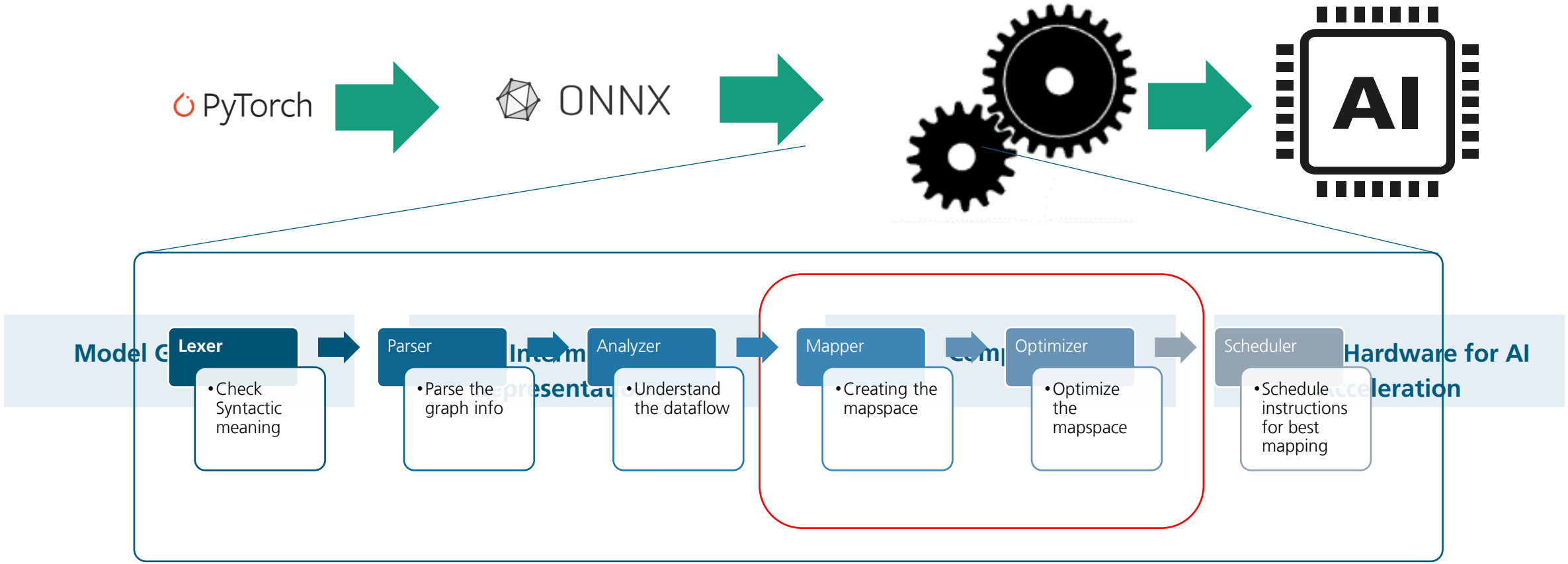


03

Mapper & Compiler

© Fraunhofer IIS

Mapper & Compiler Tool Chain



© Fraunhofer IIS

Mapper for DNN Accelerator

Multi-core accelerator

Neural Network layers can be mapped into different cores available on-chip.

Impact on KPIs

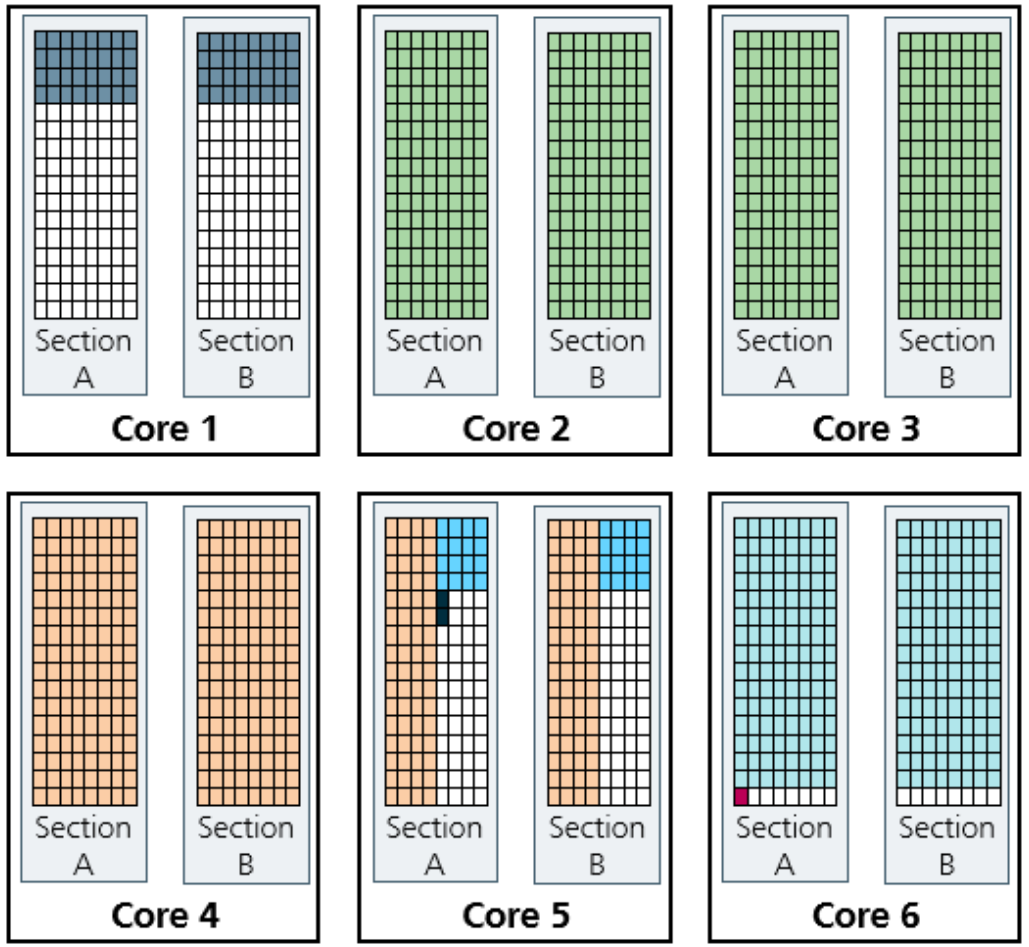
Very huge impact on KPIs like energy consumption, latency.

Mapping strategy

Reduced data movement on chip.
Maximum utilization of hardware resources like ADCs.

Animation GIF exported

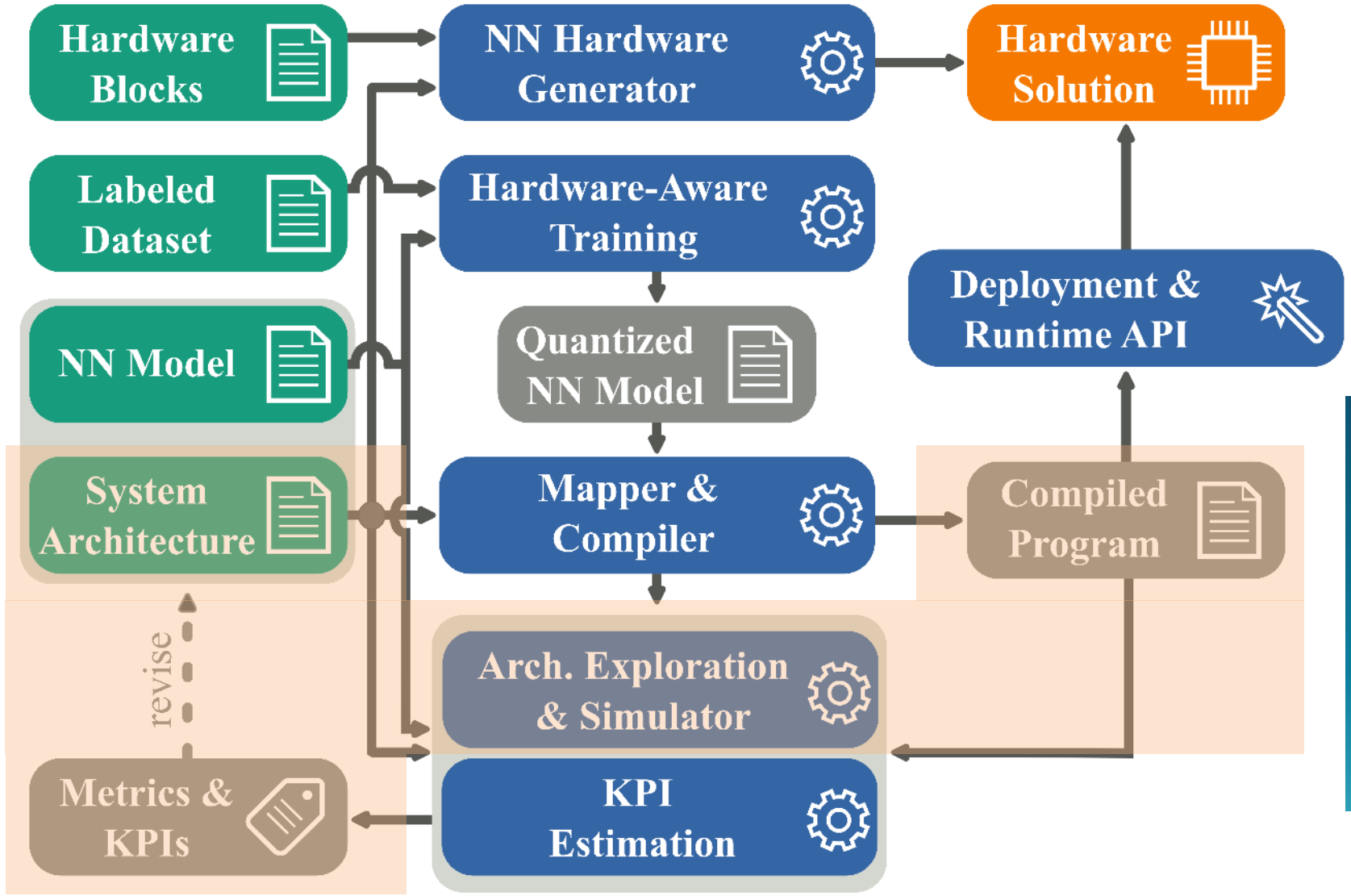
Better visualization for user.



Layer	Colour
1	
2	
3	
4	
5	
6	
7	

© Fraunhofer IIS

Fraunhofer IIS



04

Architecture Exploration & Simulation

© Fraunhofer IIS

Architecture Exploration & Simulator

Objectives

- Architecture *structure* exploration = Test different accelerator options at an early stage
  - Keep functionality abstract (behavioral) --> low implementation effort
  - Simulate timing only, where this is critical --> fast simulation
  - Use abstract interfaces (function calls) instead of actual RTL HW interfaces wherever possible --> modules can be changed and exchanged quickly
  - Explore number of compute cores, global vs. local weight caches, different on-chip bus options...
- Architecture *parameter* exploration = Make (behavioral) modules highly parametrizable
  - Explore RAM sizes, (bus) word-widths, crossbar dimensions...
- Calculate KPIs during simulation
  - Each operation has attributes about required energy and latency
  - Each module has attributes about static power consumption

© Fraunhofer IIS

Fraunhofer IIS

Architecture Exploration & Simulation Tool

Our solution

Heterogeneous System C Model

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

High simulation speed

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

Simulation of KPIs

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

KPI Optimization

Impact on performance due to architecture changes can be evaluated quickly

Architecture exploration, verification and optimization

Setup Files

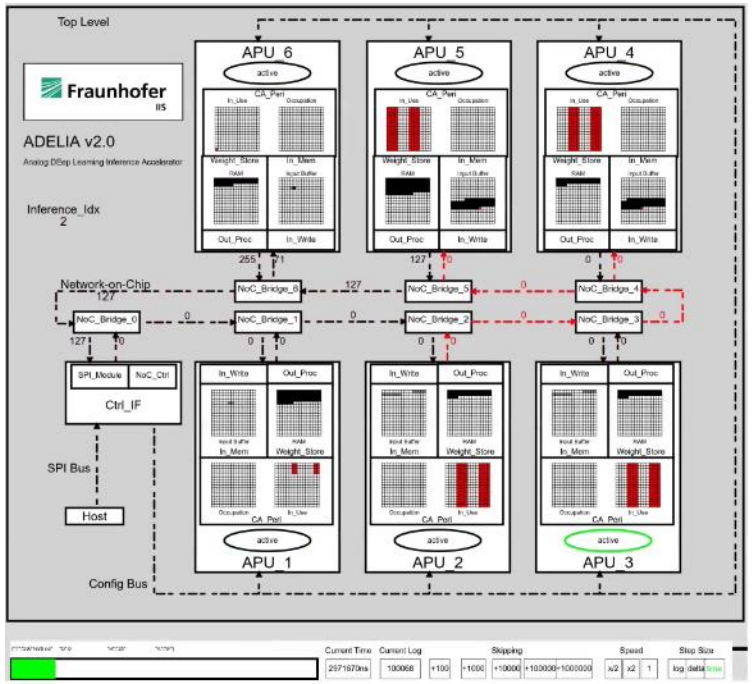
Hardware metrics
Architecture configuration
Compiled instructions
Input data

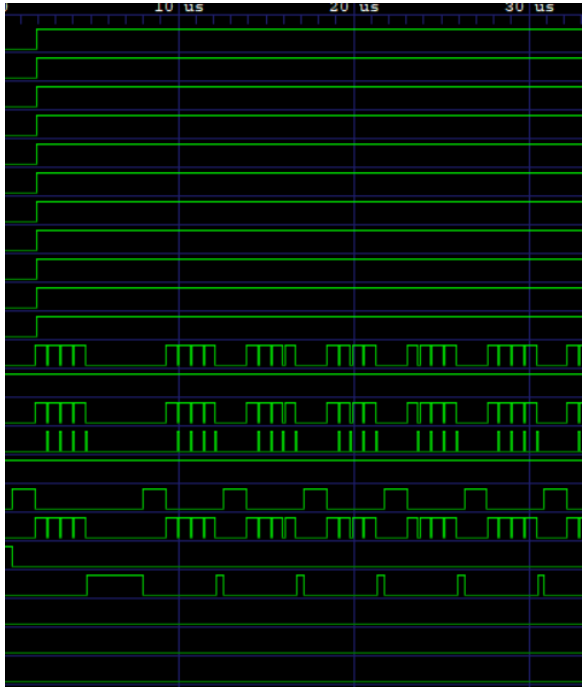
© Fraunhofer IIS

Architecture Exploration & Simulator

Outputs + Visualization

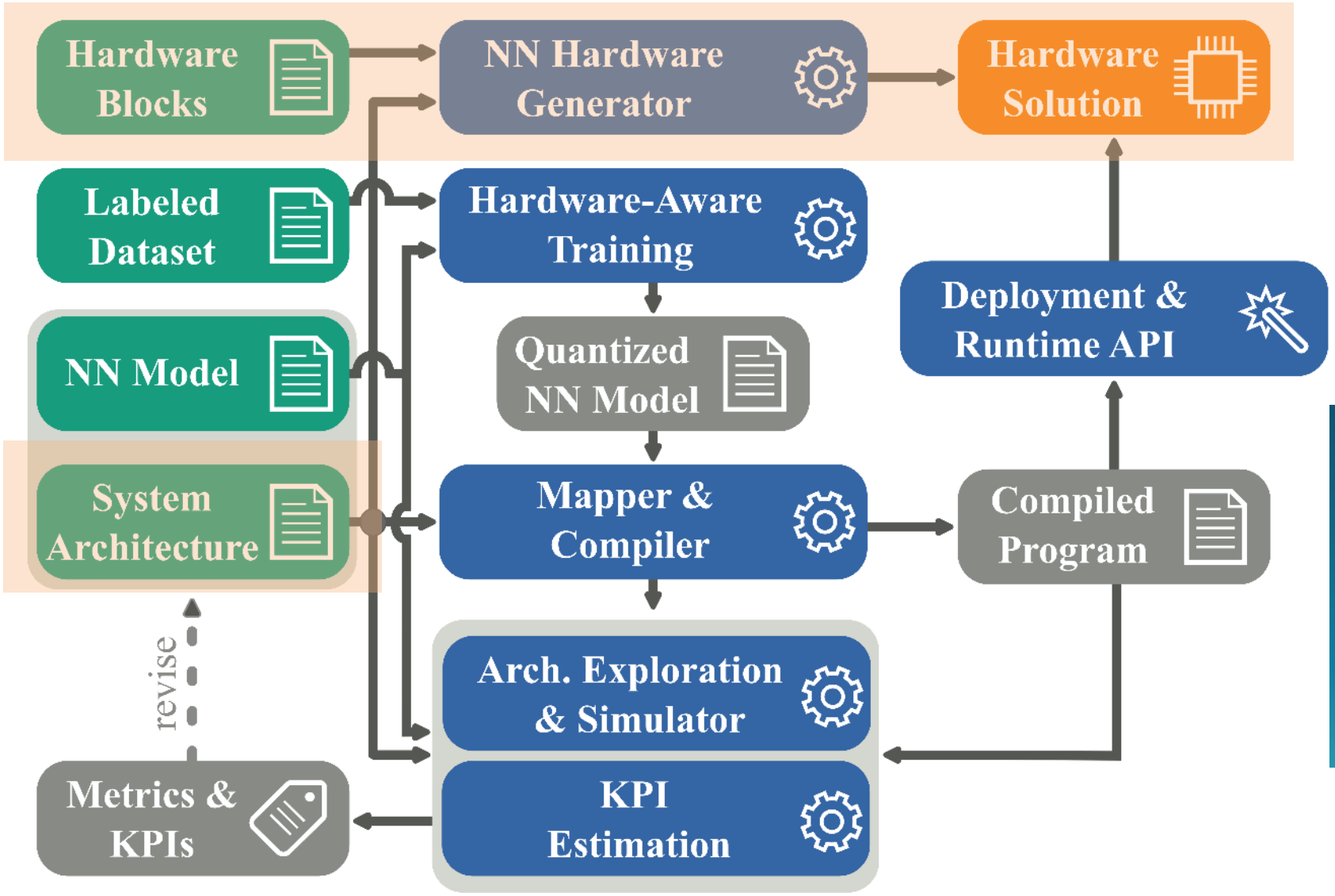
- KPI/Metric output
- Output of Neural Network's inferences + accuracy metric
- Waveform (but some signals are just losely-timed!)
- Visualization






```
{
  "APU_1": {
    "0_Module_KPIs": {
      "Accum_Static_Energy_Ex_vec_Submodules_J": 7.104728240410908e-09,
      "Duty_Cycle_Percent": 100.0
    },
    "ADC_Conversion": {
      "Accum_Energy_J": 3.743541717529297,
      "Comment": "Tracks each individual ADC conversion.",
      "Duty_Cycle_Percent": 22.0,
      "Operation_Count": 59520
    },
```


 © Fraunhofer IIS Public



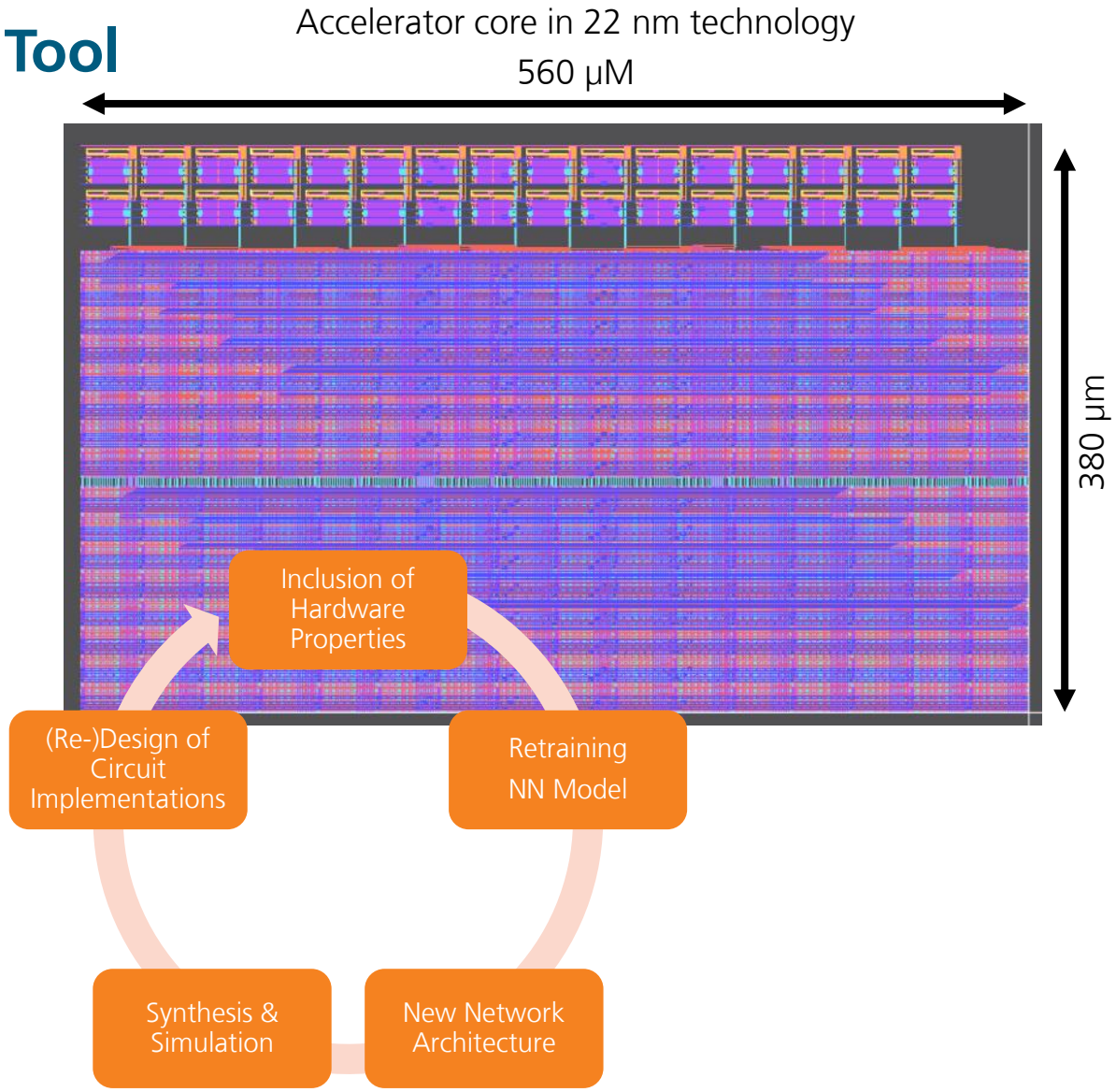
05

HW Generator

© Fraunhofer IIS

Neural Network-Hardware Generator Tool

- Analog/Mixed-Signal AI-Accelerators
  - 10 thousands of circuit instances
  - Long time-to-market
- Stability of the design process
  - Multiple millions of circuit nodes
  - Thousands of pins
  - Possibility of errors
- Design Cycle
  - Hardware/Software Co-Design
  - Parallel development of network algorithm and circuit implementation
  - Inclusion of improved circuit implementations

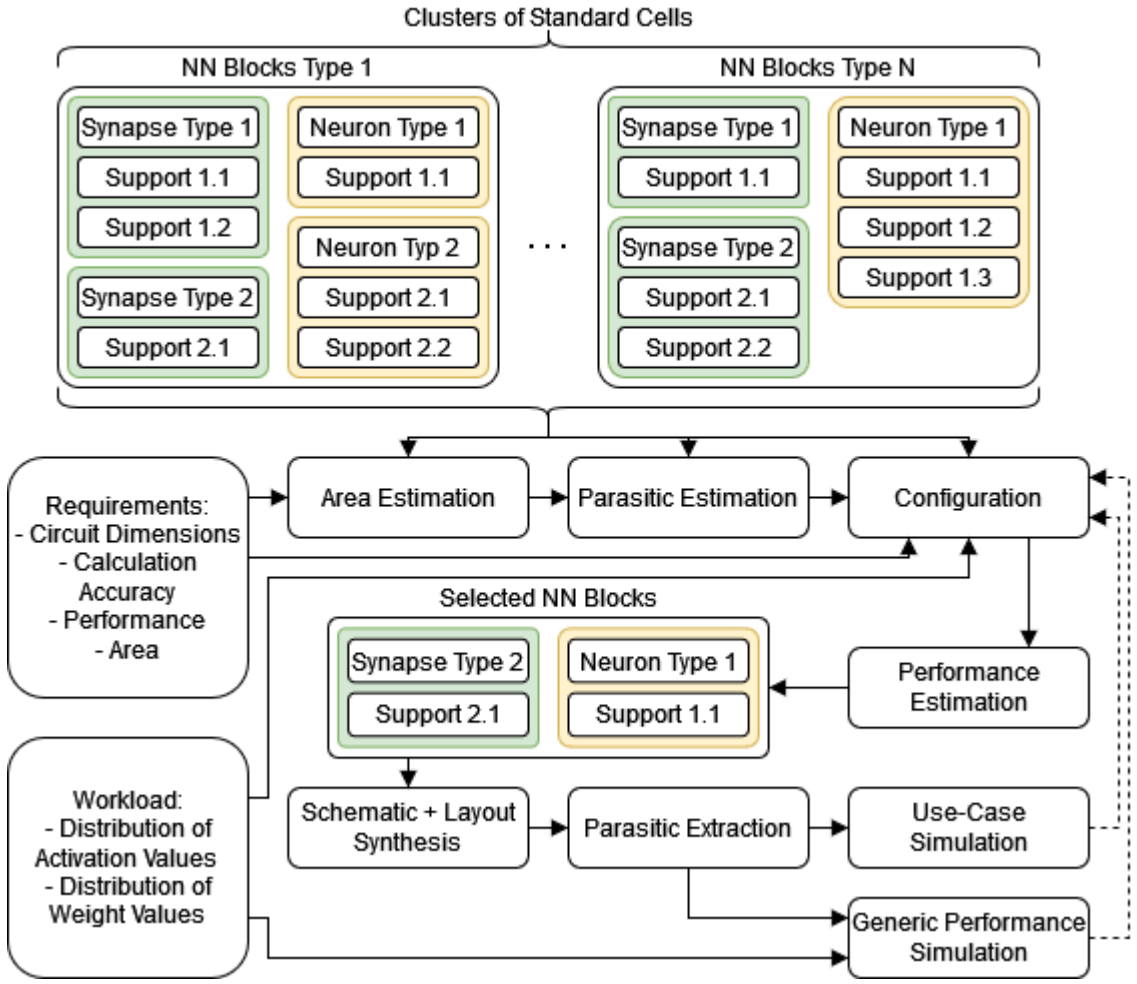


© Fraunhofer IIS

Neural Network-Hardware Generator Tool

Design Flow

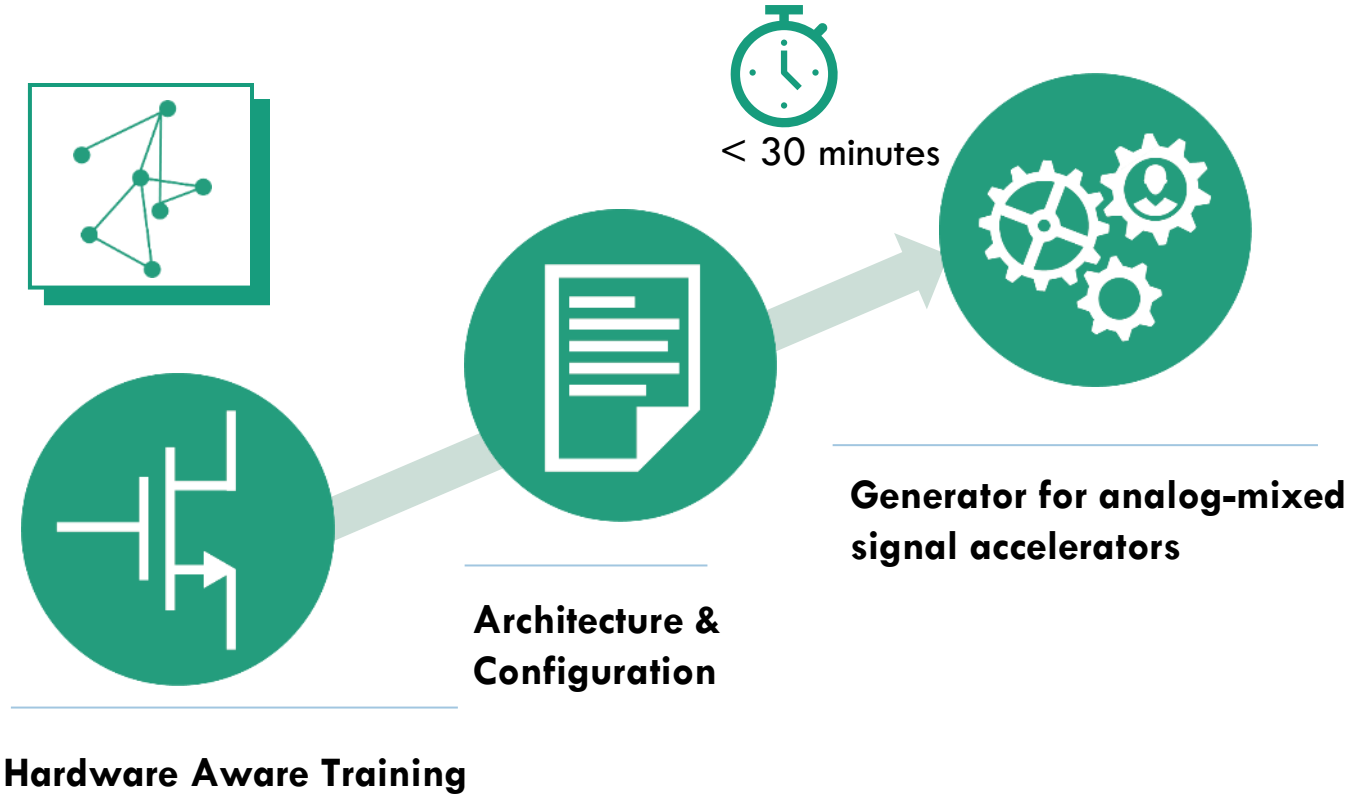
- Synthesis flow for analog/mixed-signal Neural Network (NN) accelerators
  - Multiple sets of NN standard cells (building blocks)
  - Cell selection and routing options based on:
    - Requirements
    - Parasitic estimation derived from area estimation
    - Estimated performance
  - Layout and schematic synthesis
  - Verification based on workload and use-case

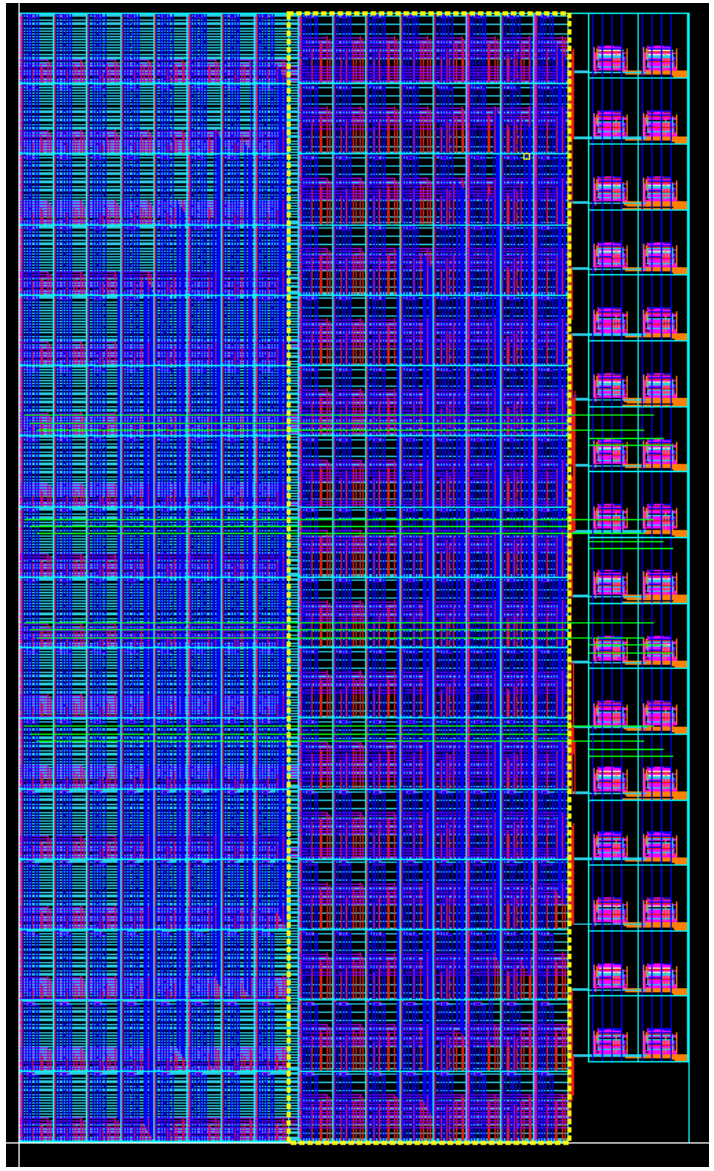


R. Müller, L. Mateu and R. Brederlow, "Analog/Mixed-Signal Standard Cell Based Approach for Automated Circuit Generation of Neural Network Accelerators," *2023 38th Conference on Design of Circuits and Integrated Systems (DCIS)*, Málaga, Spain, 2023, pp. 1-6, doi: 10.1109/DCIS58620.2023.10335979.

© Fraunhofer IIS

Neural Network-Hardware Generator Tool





© Fraunhofer IIS

Neural Network-Hardware Generator Tool

Results

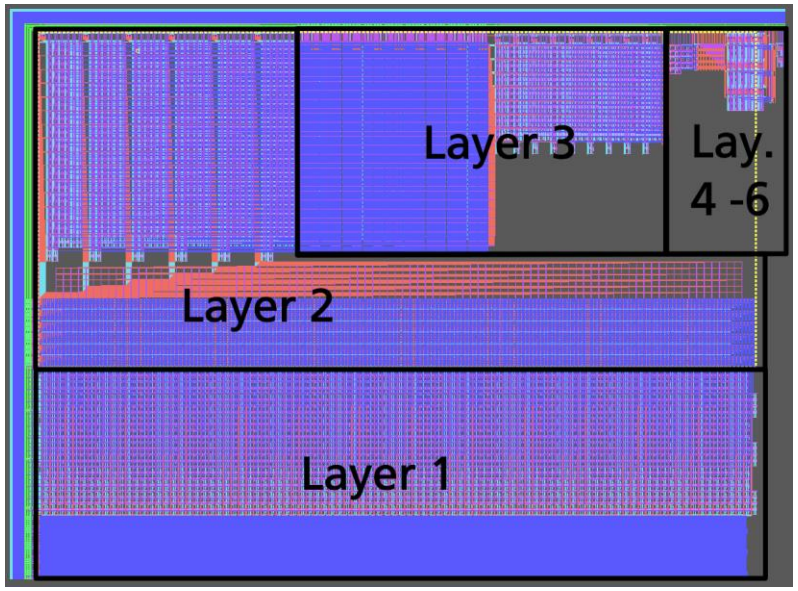
- Tested with different accelerator architectures
- Tested with different semiconductor technologies

- Great reduction in design time
- Limited area overhead

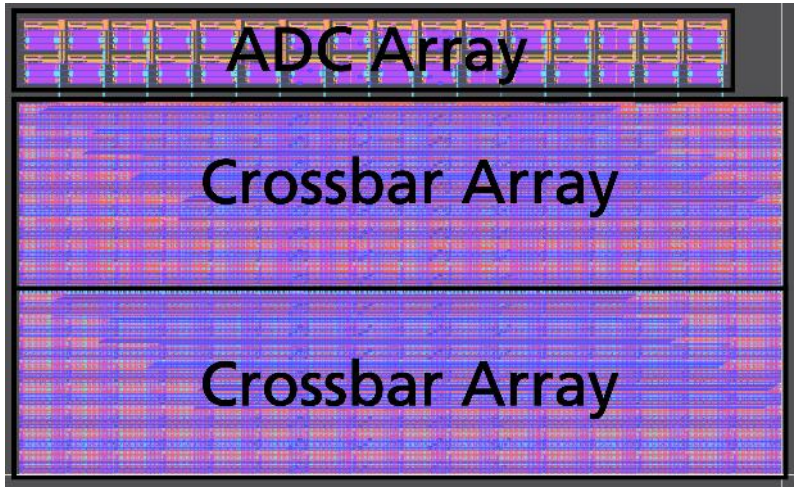
	ADELIA Gen1 22nm	ADELIA Gen1 28nm	ADELIA Gen2 22nm
No. of Synapses	72846	2480	16384
Runtime: Autom. Place & Route	~30min	~5min	~5min
Effort: Manual Place & Route	~3mths	~1mth	~2mths
Area overhead	~15%	~5%	~2%

R. Müller, L. Mateu and R. Brederlow, "Analog/Mixed-Signal Standard Cell Based Approach for Automated Circuit Generation of Neural Network Accelerators," *2023 38th Conference on Design of Circuits and Integrated Systems (DCIS)*, Málaga, Spain, 2023, pp. 1-6, doi: 10.1109/DCIS58620.2023.10335979.

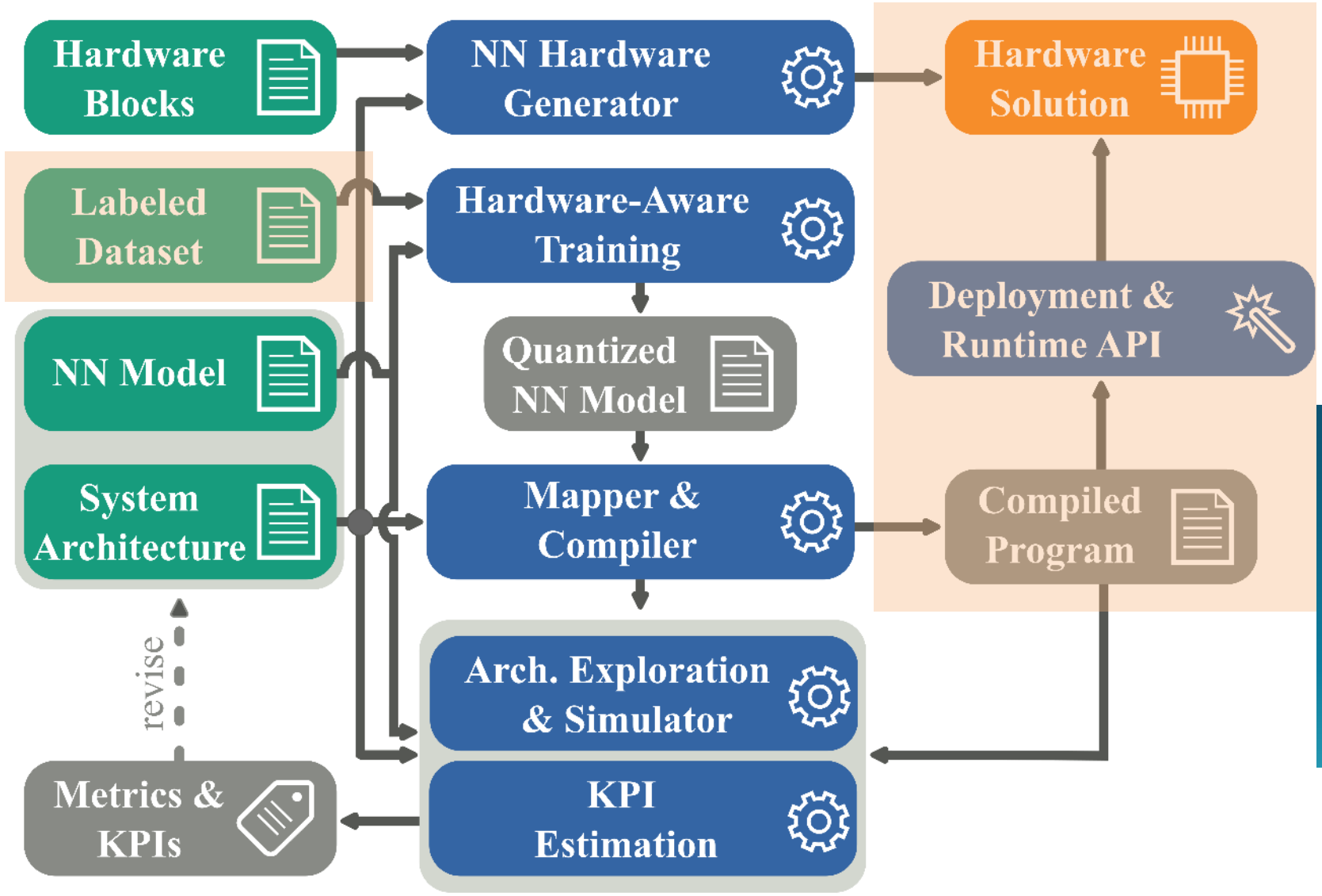
ADELIA Gen1 22nm



ADELIA Gen2 22nm



© Fraunhofer IIS



06

ASIC Characterization and Demo

© Fraunhofer IIS

Runtime API

ASIC Characterization

- Test setup to
  - Power up the chip, including logging of the power consumption per pin
  - Communicate via SPI
  - Log digital output
  - Log analog output
  - Apply analog input voltage (for calibration)
- Chip in a socket for easy changing
- Automated measurement flow
  - Run automated calibrations
  - Easily sweep through many different configurations
  - Record result of many different input data sets

Measurement equipment integrated in a PC

Evaluation Board with the chip

© Fraunhofer IIS

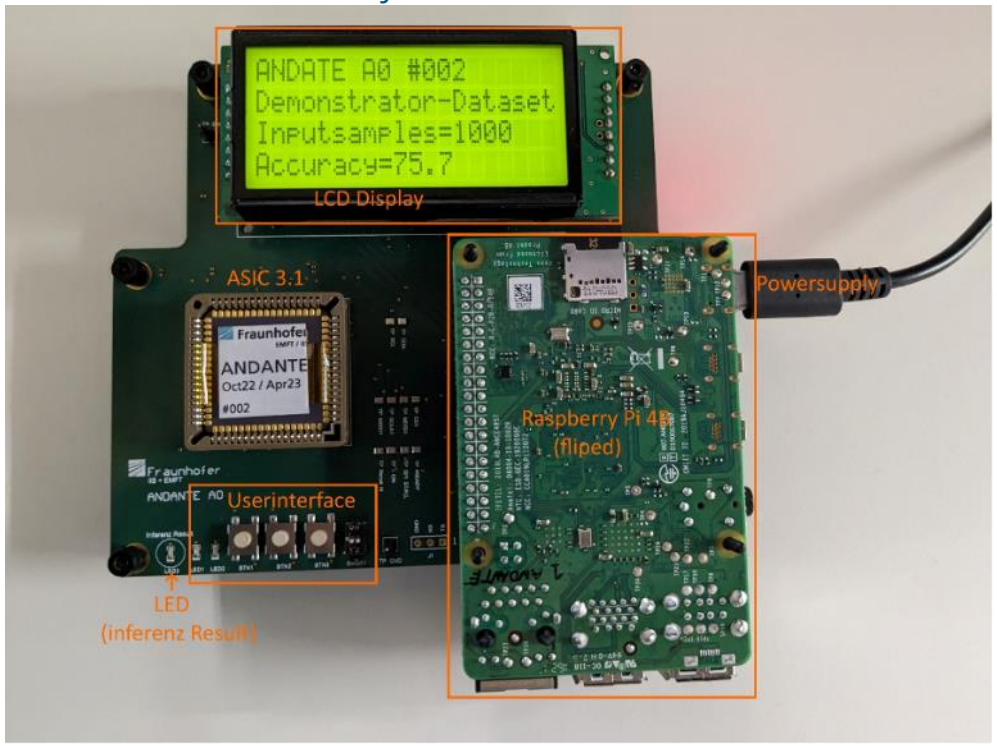
Runtime API

Demonstrator

Real World Demonstrator

- ADELIA Gen2 ASIC with Raspberry Pi as host-system
  - Live voice activity detection with microphone input
  - Power consumption measurement
  - Accuracy calculation in dataset mode

Voice Activity Detection Demonstrator



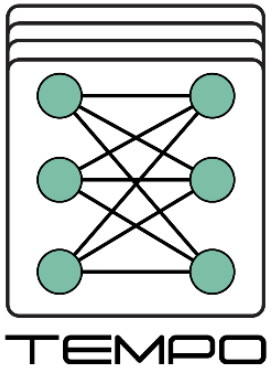
 © Fraunhofer IIS

Acknowledgement

ECSEL Joint Undertaking

Electronic Components and Systems for European Leadership

This 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 France, Belgium, Germany, Netherlands, Portugal, Spain, Switzerland. ANDANTE has also received funding from the German Federal Ministry of Education and Research (BMBF) under Grant No. 16MEE0116.



This 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.
TEMPO has also received funding from the German Federal Ministry of Education and Research (BMBF) under Grant No. 16ESE0405.

 © Fraunhofer IIS

Contact

Dr. Loreto Mateu
Head of Integrated Circuits and Systems Department
Division Smart Sensing and Electronics
Phone +49 9131 776-4456
loreto.mateu@iis.fraunhofer.de

Roland Müller
Senior Engineer – Advanced Analog Circuits
Division Smart Sensing and Electronics
Phone +49 9131 776-9214
roland.mueller@iis.fraunhofer.de

Maen Mallah
Senior Engineer – Embedded AI
Division Communication Systems
Phone +49 9131 776-6339
maen.mallah@iis.fraunhofer.de

Fraunhofer IIS
Fraunhofer-Institut für Integrierte Schaltungen IIS

Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Thanks for your attention

Q&A

Copyright Notice

This multimedia file is copyright © 2023 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org

Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org