The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

ANDANTE

AI
ECSEL JU

AI4DI

TEMPO

AI

Vienna, Austria

25-26 July 2022

The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

Benchmarking Neuromorphic Computing for Inference

Simon Narduzzi[1], Loreto Mateu[2], Petar Jokic1, Erfan Azarkhish[1], and Andrea Dunbar[1]

[1] CSEM, Switzerland

[2]Fraunhofer IIS, Germany

Vienna, Austria 25-26 July 2022

Presentation Outline

- Challenges of IoT/ Edge Computing
- Benchmarking: State-of-the-Art
- Unfair / Fair Benchmarking
- Use-Case dependent Benchmarking
- Conclusions and Outlook

Challenges of IoT / Edge Computing

- What is the best solution for my application?
  - Objective is **Hardware** comparison and KPIs estimation

- Applications for neuromorphic computing
  - Industrial plant → Condition Monitoring
  - Automotive → Autonomous Driving
  - People → Ambient Assisted Living
  - Ecosystems → Animals monitoring

- Conditions
  - Distributed deployment
  - Continuous Monitoring (low-latency)
  - Battery powered operation (low-power)
  - Wireless communication (without cloud connection)

Benchmarking: State-of-the-Art

KPIs

Task	Model	Hardware
Precision	Complexity	Power consumption
TPR	Parameters	TOPS
TNR	Precision	Memory size
FPR	Structure	Area
	SynOps	J/OP
		J/Spike
		Area efficiency
		Precision

Benchmarking: State-of-the-Art

Table 1.4 Typical display of performance comparison of neuromorphic hardware platforms, adapted from [34].

DLI Accelerator	Type	Target application	Performance
NVIDIA Jetson Nano	GPU	Embedded	472 GOPS @ 5 – 10 W
Nvidia Jetson TX2	GPU	Edge	1,3 TOPS @ 7,5 W
NVIDIA Jetson AGX Xavier	GPU	Edge	30 TOPS @ 30 W
NVIDIA Drive AGX Pegasus	GPU	Automotive	320 TOPS
Intel Movidius Myriad 2 bzw. Myriad X	Chip	Embedded/Edge DL/Vision	4 TOPS @ 1 W (Myriad X)
MobilEye EyeQ4	Chip	Automotive	2.5 TOPS @ 3 W
GreenWaves GAP8	Chip	Battery powered AI	200 MOPS bis 8 GOPS @ <100mW
Canaan Kendryte K210	Chip	Embedded Vision & Audio	250 GOPS @ 300mW
Google Coral Edge TPU	Chip	Edge	4 TOPS @ <2,5W
Lattice sensAI Stack	Soft IP-Core	Embedded	<1 mW – 1 W
Videantis v-MP6000UDXM	Soft IP-Core	Embedded DL/Vision	<6,6 TOPS @ 400 MHz

Table 1.5 Recent display of performance comparison of neuromorphic hardware platforms, adapted from [35].

		Eyeriss	ENVISION	Thinker	UNPU	This work	
Technology		65nm	28nm	65nm	65nm	65nm	
Area		1176k gates (NAND-2)	1950k gates (NAND-2)	2950k gates (NAND-2)	4.0mm×4.0mm (Die Area)	2695k gates (NAND-2)	
On-chip SRAM (kB)		181.5	144	348	256	246	
Max Core Frequency (MHz)		200	200	200	200	200	
Bit Precision		16b	4b/8b/16b	8b/16b	1b-16b	8b	
Num. of MACs		168 (16b)	512 (8b)	1024 (8b)	13824 (bit-serial)	384 (8b)	
DNN Model		AlexNet	AlexNet	AlexNet	AlexNet	sparse AlexNet	sparse MobileNet
Batch Size		4	N/A	15	N/A	1	1
Core Frequency (MHz)		200	200	200	200	200	200
Bit Precision		16b	N/A	adaptive	8b	8b	8b
Inference/sec	(CONV only)	34.7	47	-	346	342.4	-
	(Overall)	-	-	254.3	-	278.7	1470.6
Inference/J	(CONV only)	124.8	1068.2	-	1097.5	743.4	-
	(Overall)	-	-	876.6	-	664.6	2560.3

Benchmarking: Combined KPIs

KPIs

Task
- Precision
- TPR
- TNR
- FPR

Model
- Complexity
- Parameters
- Precision of parameters
- Structure
- SynOps

Hardware
- Power consumption
- TOPS
- Memory size
- Area
- J/OP
- J/Spike
- Area efficiency
- Precision

Combined
- Configuration time
- Latency
- Energy/Inference

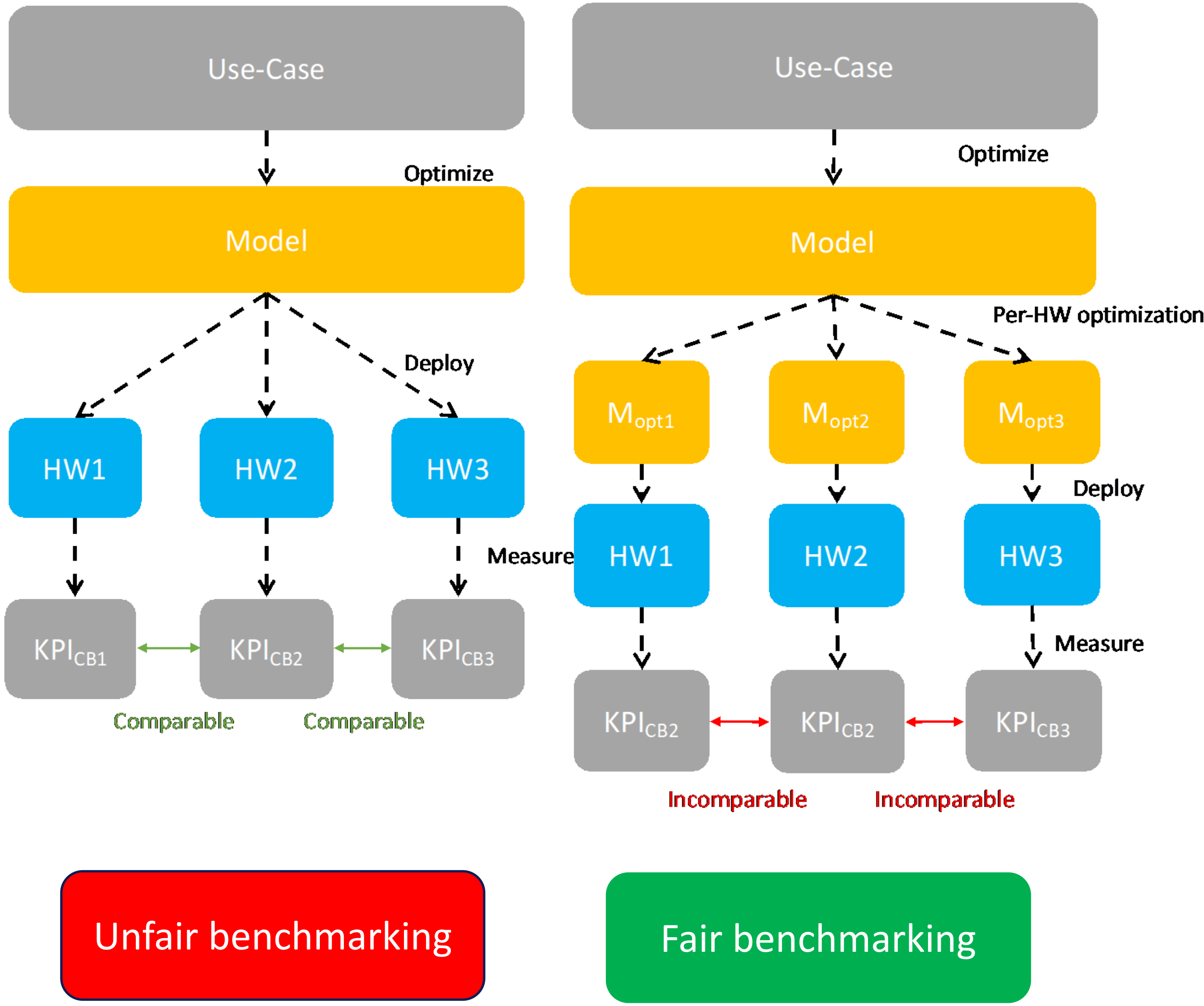
Unfair / Fair Benchmarking

Unfair Benchmarking

- Same NN is deployed in each hardware platform
- Hardware KPIs can be compared since Task and Model KPIs are same
- However, all capabilities of the hardware are not exploited and the models are not optimized and the comparison is misleading

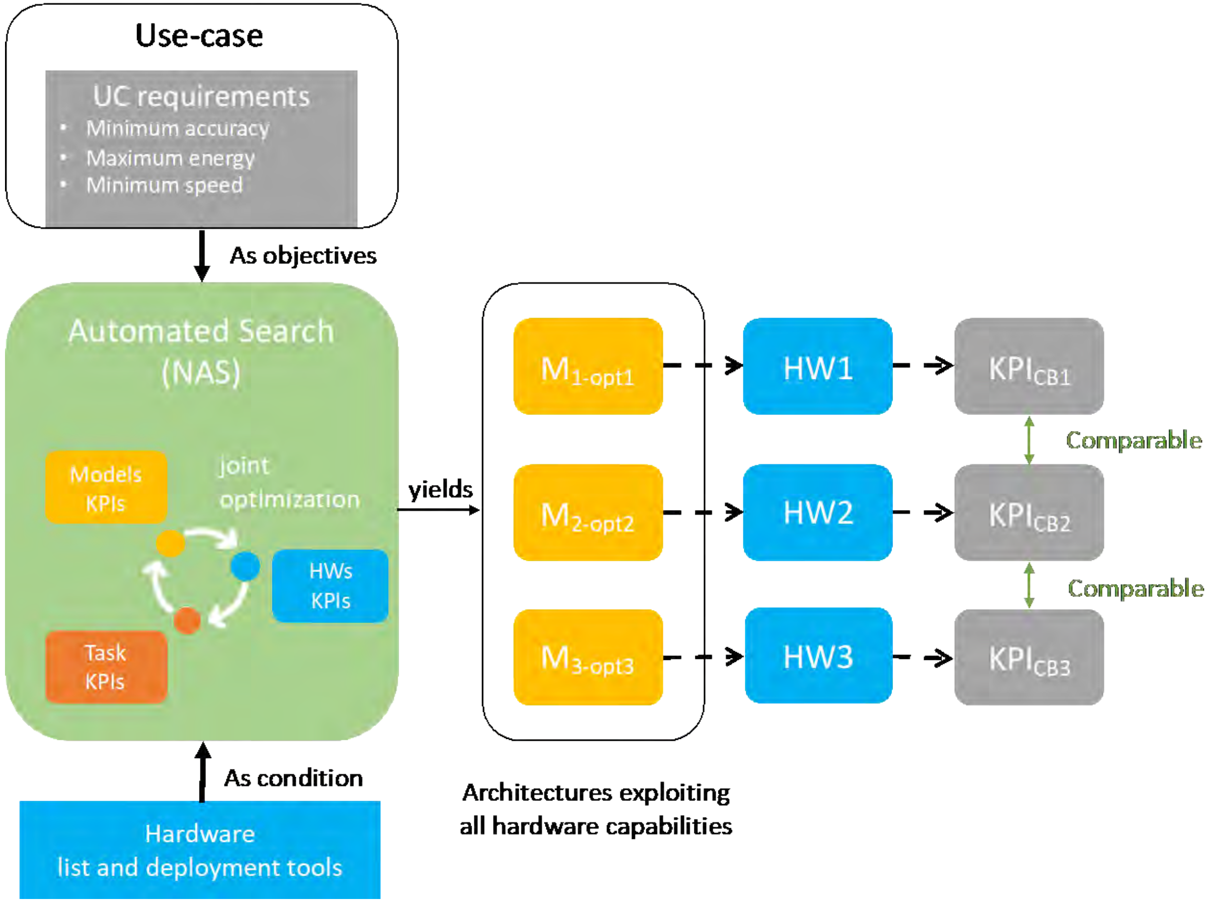
Fair benchmarking

- Exploits all capabilities of the hardware
- Different NN models are deployed on the hardware
- Thus Model and Hardware KPIs cannot be compared
- Over-optimization makes the comparison impossible.



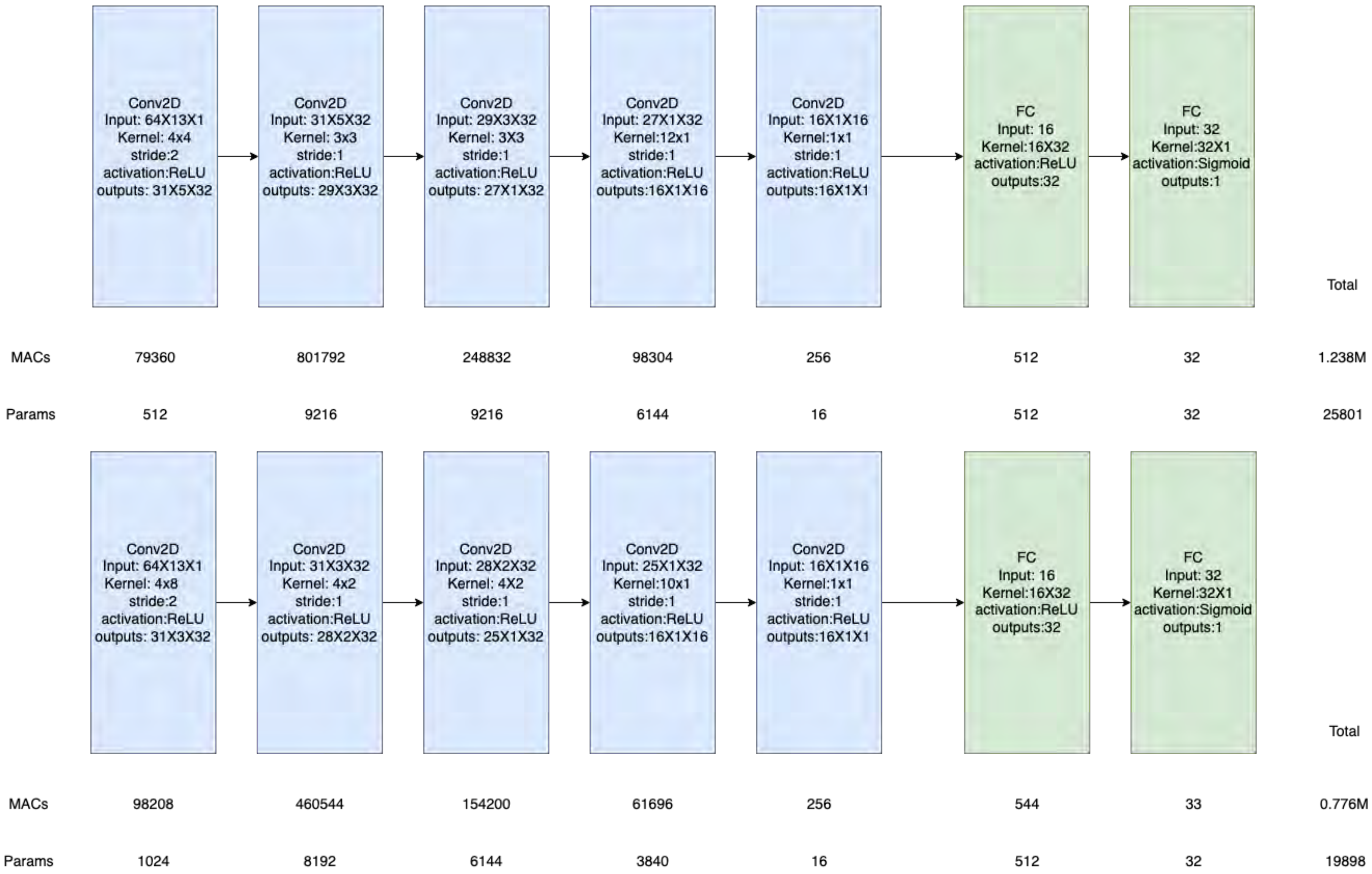
Use-Case dependent Benchmarking

- A benchmarking framework is necessary
- Automated Search is necessary to optmize KPIs
- Each hardware suports different network sizes and layer types
- Benchmarking should compare only architectures exploring all hardware capabilities for the requirements of a certain use case
  - Accuracy, latency, energy per inference
- Hardware device can be selected based on the resulting KPIs



Use-Case dependent Benchmarking

NN	Value
Quantization	8, 4, 2 bits
SRAM	200 kBytes
Num. Layers	≤ 10
Layer types	Conv1D, Conv2D, Flatten (1x1 Conv), FC, BN
Activation types	ReLU, Sigmoid
Max. input size FC layer	1152
Accuracy	≥92%
Mean Power consumption	< 1 mW
Latency	< 5 ms



Conclusion and Outlook

- Benchmarking neuromorphic hardware is a hard task
  - Variety of devices
  - Variety of software tools
  - (Un)comparable KPIs
- Not all KPIs are relevant, combined KPIs are more informative
- Use-case based benchmarking should be used
- Benchmarking framework and complete software tool chain are necessary
  - Hardware-aware training
  - Automated search
  - Mapper and compiler

Event Organisers

ECSEL Joint Undertaking

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

AI4DI

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

TEMPO

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

ANDANTE

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.

Thank You

For your attention

@ loreto.mateu@iis.fraunhofer.de