Design ASIC architectures for generic, self-learning and reliable neuromorphic AI accelerators.

Martin Andraud

Assistant Professor, UC Louvain (Belgium)

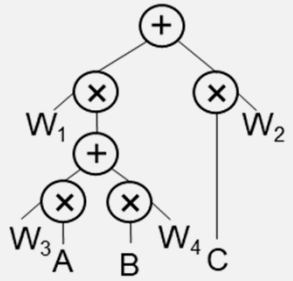
Visiting Professor, Aalto University (Finland)

My background (I)

- Originally from Clermont-Ferrand, France
- Assistant professor in ***UCLouvain*** since 01/2024
  - Visiting professor at Aalto University

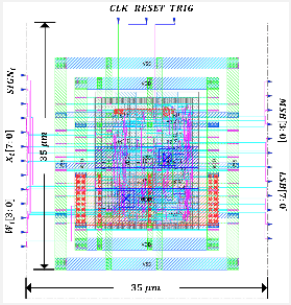
Main research interests:

HW-SW integration and probabilistic AI



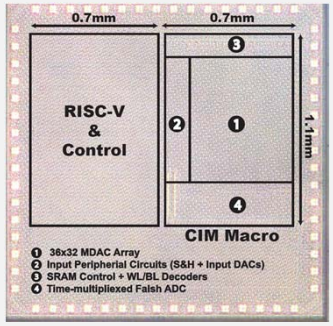
[UAI'23]

Digital and time-based AI processors



[NORCAS'24]

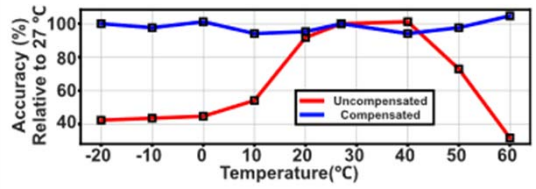
Mixed-signal In-Memory Computing



[Acore, AICAS'24]

Reliability of IMC hardware

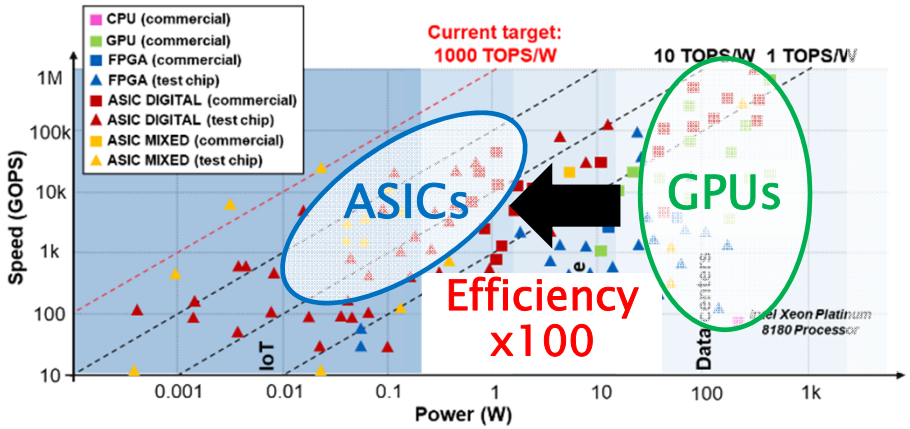
Self-calibration methods against process or temperature



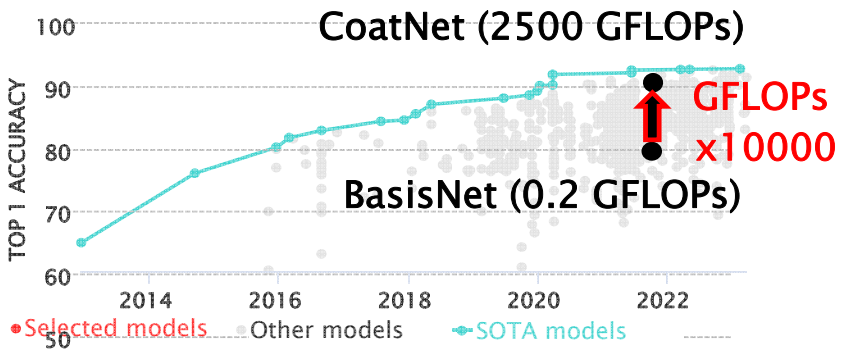
[ISCAS'23, ETS'24]

Edge AI: a hardware/software issue

[1] https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator
[2] https://paperswithcode.com/sota/image-classification-on-imagenet
[3] Thompson et al., Deep Learning's Diminishing Returns, 2021



Hardware "accelerators" [1]



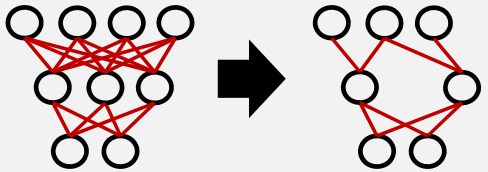
Top 1 accuracy on ImageNet [2]

- More efficient accelerators but even bigger NN models ➔ We need to explore **alternatives**
- Deep NNs have known limitations (reliability, explainability, hardware efficiency) and could be well **complemented or replaced** (in specific tasks) by **other models**…
- This is my **research focus**

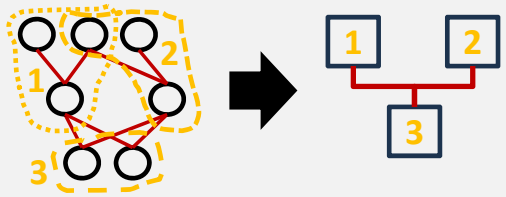
Recent models are sparse…

1. More and more Deep Learning models shift towards **sparse** matrix/tensor computation:

Pruned DNNs



Compressed DNNs

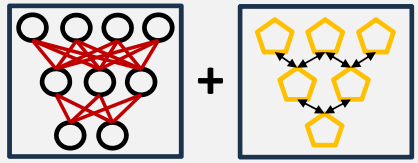


[4,5]

Irregular workloads are not optimal for classical accelerators
➔ Group of tensors

2. **Combining** "emerging" models on the same architecture is gaining a lot of interest as well:

Neurosymbolic AI



DNN **"Logic" NN**

Bidirectional
Expresses uncertainty
…

Probabilistic reasoning

Encodes uncertainty in a principled way

Explainable AI
Trustworthy AI

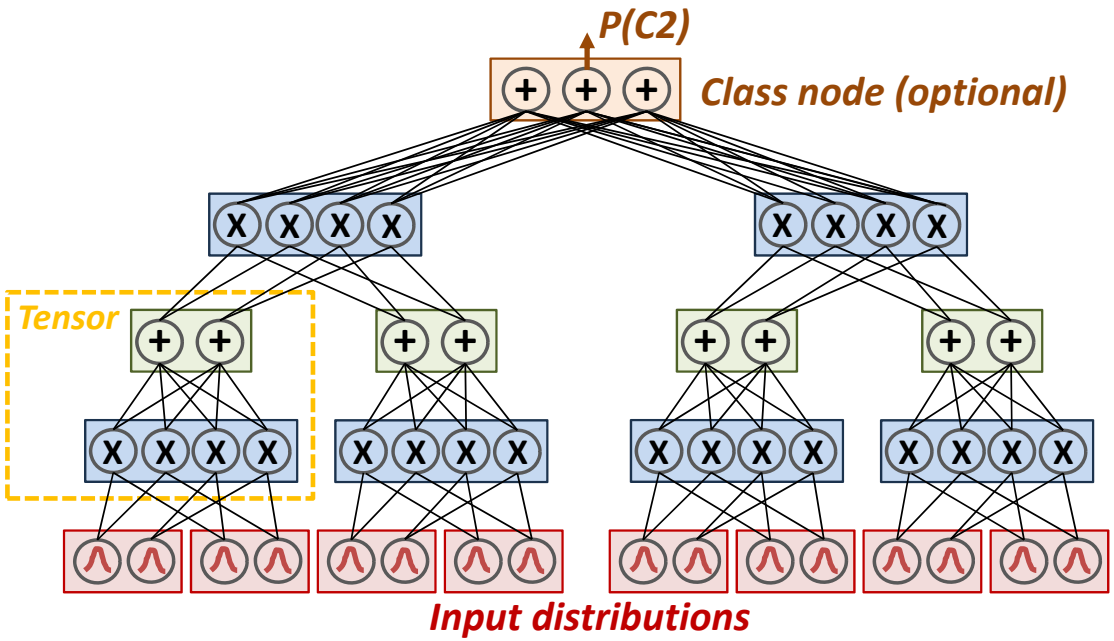
Efficiently compute these emerging models requires hardware solutions

Emerging models on-chip

3. These "Emerging" models can be now implemented with **efficient and tensorized** forms:

Probabilistic circuits (PCs)

- PCs are computational graphs (similar to DNNs) but they are **encoding a probability distribution (for inference).**
- Composed of basic computations (**Weighted sums**, **products**, ***max***, etc.)
- Can be **learned from data** (as DNNs)
- Can be used for **logic/neurosymbolic AI**
- They have **tensorized** versions available
  - Einsum Networks, RAT-SPNs…



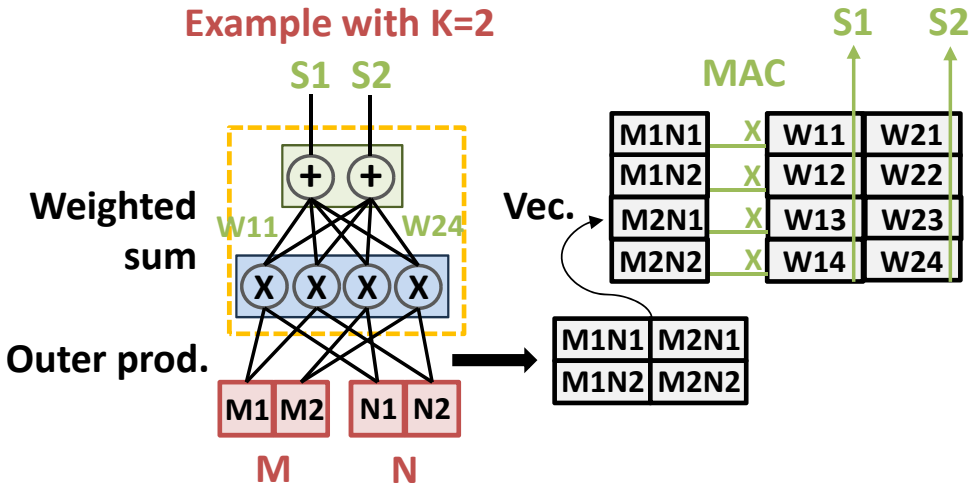
1+2+3: find a **common computation primitive** and a **single hardware for all these applications**

A tensor computing primitive

Computing this tensor would fit various AI models:

Compute this tensor (vector-matrix multiplication) of dimension K*K:

- Outer product of inputs (**K^2 PRODs**)
- Vector **MAC** operation
  - Connect results to K sum nodes with weights (**K^3 PRODs**)
  - Accumulate (**K SUMs or K^3-1 accumulations**)

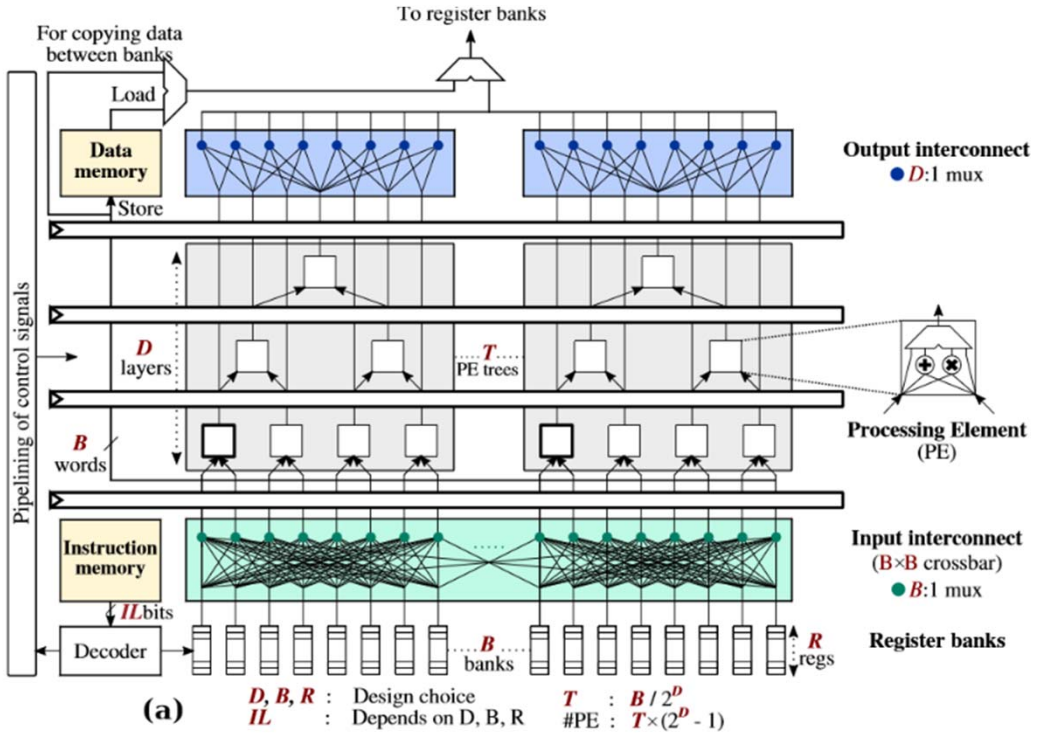


Trees of tensors are explicictly used in **Sparse DNN** computation **[4], Compressed NNs [5], probabilistic circuits [6,7] ➔ Accelerator dedicated to this?**

Baseline architecture/ DPU

[8] N.Shah et al. “DPU-v2: Energy-efficient execution of irregular directed acyclic graphs” MICRO 2022

- DPU is a state-of-the-art HW accelerator for irregular graph processing [8]
  - Targets the execution of irregular Direct Acyclic Graphs (PCs, sparse solvers)
  - Computation in 8b/16b/32b Floating point
- DPU decomposes graphs in **multiple blocks executed on trees of PEs**
  - Each PE can do an add, mult. or be bypassed
- Performance
  - Irregular graphs **10 GOPS, 3-6 GOPS/W**
  - “regular” graphs: **17 GOPS, 42 GOPS/W**



Main ideas to explore

- Modify DPU to efficiently execute these tensors (➔ **Generic architecture**)
  - One tree = 1 tensor unit computation (**ongoing**)
- Include the computation of DNNs in floating point (➔ **self-learning**)
  - Self-learning (self-training) possibilities (**ongoing**)
- Evaluate the possibility of (analog) In-memory computing for the computation with emerging memories (➔ **neuromorphic**)
  - Compatibility of analog IMC with higher resolution computation? (planned)
  - "Analog probabilistic reasoning" possible? (planned)
- Integrate self-calibration mechanisms for efficiency (➔ **reliable**)
  - Self-calibration seems necessary for these architectures (ongoing)

TRL 4 to 5 (fully digital)

TRL 3 or 4 (emerging memories)

Collaboration perspectives

- Collaborate on this type of generic architecture
  - Find other computation primitives for more models
  - Include the possibly of online training
- Collaborate on analog In-Memory Computing architectures
  - How to make AIMC really reliable?
  - Can we use AIMC for high(er) resolution computing?
- Find applications of combined AI model execution
  - Explainable AI, uncertainty estimation, etc.

ICOS International Cooperation On Semiconductors

European Commission

EU – SOUTH KOREA – Joint Researchers Forum on Semiconductors

This project has received funding from the European Union's Horizon Europe research and innovation programme under GA Nº 101092562

www.icos-semiconductors.eu

References

[1] https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator

[2] https://paperswithcode.com/sota/image-classification-on-imagenet

[3] Thompson et al., “Deep Learning's Diminishing Returns”, 2021

[4] Nandeeka Nayak, et al. “TeAAL: A Declarative Framework for Modeling Sparse Tensor Accelerators”. In 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO ’23)

[5] J. Gu, B. Keller, J. Kossaifi, A. Anandkumar, B. Khailany, and D. Z. Pan, “HEAT: Hardware-Efficient Automatic Tensor Decomposition for Transformer Compression.” arXiv, Nov. 30, 2022. doi: 10.48550/arXiv.2211.16749.

[6] R. Peharz et al. “Random Sum-Product Networks: A Simple and Effective Approach to Probabilistic Deep Learning”, Conference on Uncertainty in Artificial Intelligence, 2019.

[7] Robert Peharz et al. “Einsum Networks: Fast and scalable learning of tractable probabilistic circuits”. arXiv preprint, 2020.

[8] N.Shah et al. “DPU-v2: Energy-efficient execution of irregular directed acyclic graphs” MICRO 2022