YoseUe: “trimming„ Random Forest’s training towards resource-constrained inference on FPGA

A novel approach for inference on large decision tree models, tailored for resource-constrained FPGAs

Alessandro Verosimile
alessandro.verosimile@polimi.it

Andrea Damiani
andrea.damiani@polimi.it

Alessandro Tierno
alessandro.tierno@mail.polimi.it

Marco Domenico Santambrogio
marco.santambrogio@polimi.it

POLITECNICO MILANO 1863

ASP-DAC Conference 2024
Incheon, 22-Jan-2024

ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

POLITECNICO MILANO 1863
NECST laboratory

Problem definition

Decision Tree
Ensemble models

Inference
optimization

Embedded
systems

Resource-constrained
scenario

alessandro.verosimile@polimi.it

YoseUe contributions

- A domain-specific architecture for the inference of Random Forests on Embedded systems
- A new Decision tree based ML model, designed to optimize its inference on the developed architecture

Results

YoseUe is able to support DT Ensemble models with 2 orders of magnitude more trees

alessandro.verosimile@polimi.it

Agenda

- Context definition
- Design methodologies and SoA works
- YoseUe architecture
- Training co-optimization: the MDRFC
- Experimental results

Agenda

Context definition

Design methodologies and SoA works

YoseUe architecture

Training co-optimization: the MDRFC

Experimental results

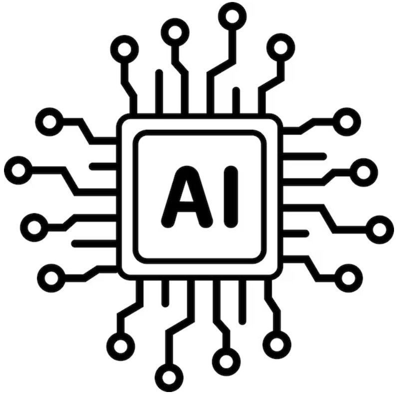
alessandro.verosimile@polimi.it

Artificial Intelligence of Things

Combinations of two worlds:

IoT

Artificial Intelligence

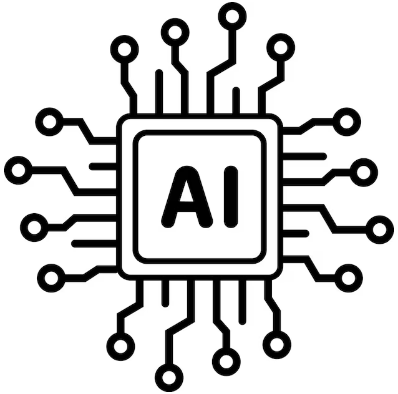


Artificial Intelligence of Things

Combinations of two worlds:

IoT

Artificial Intelligence



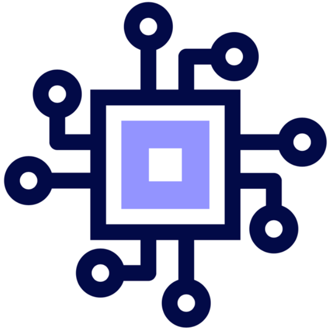
It finds application in multiple fields, such as:

- Video surveillance
- Automotive

Weaknesses & Risks of AIot

- Instability of latency/throughput
- Privacy and security risks

IoT device



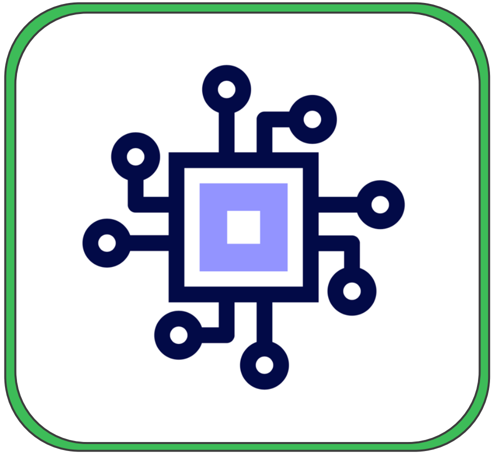
Internet connection

Cloud Infrastructure

alessandro.verosimile@polimi.it

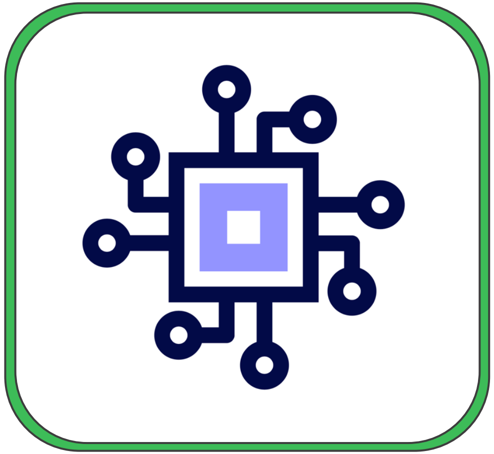
Ideal solution

Perform all the computations (the inference) on the device



Ideal solution

Perform all the computations (the inference) on the device

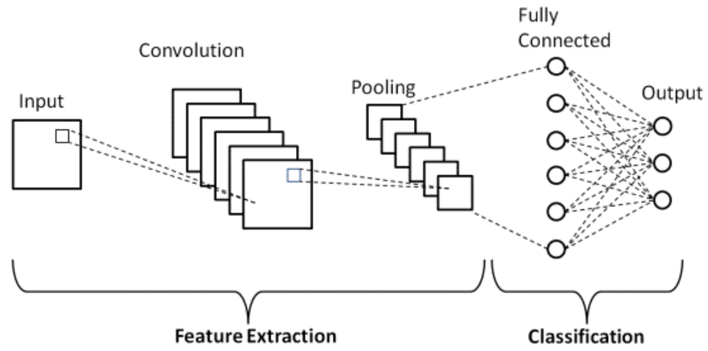


Resource-Constrained

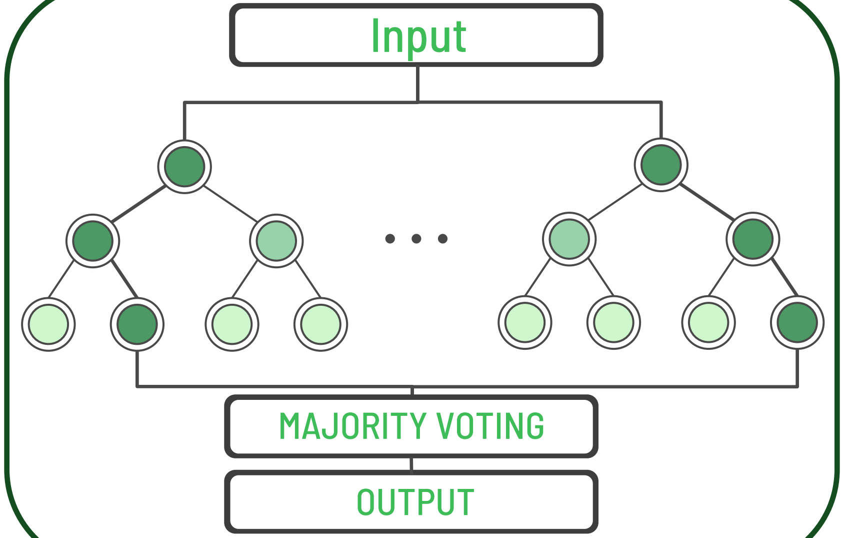
Low computational power

Machine Learning Models

Artificial Neural Networks

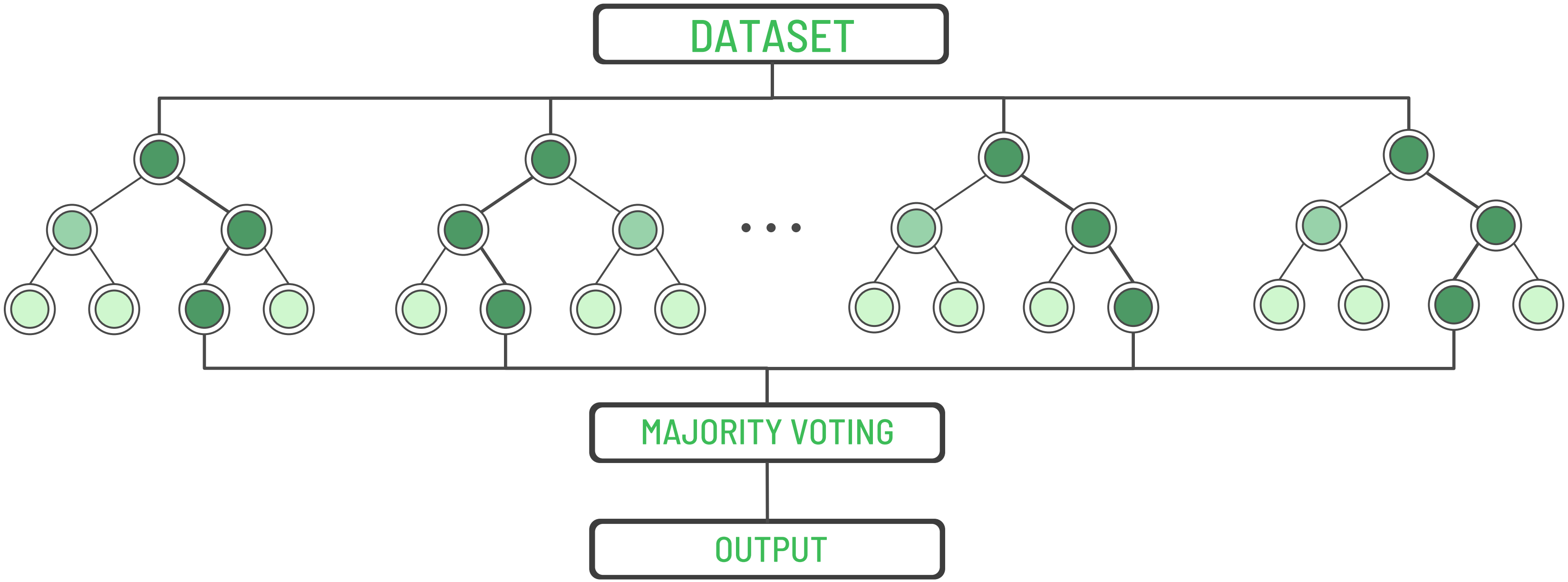


Decision Tree – based models



Random Forest

"Tree ensembles are arguably among the most accurate ML models in use nowadays" [1]

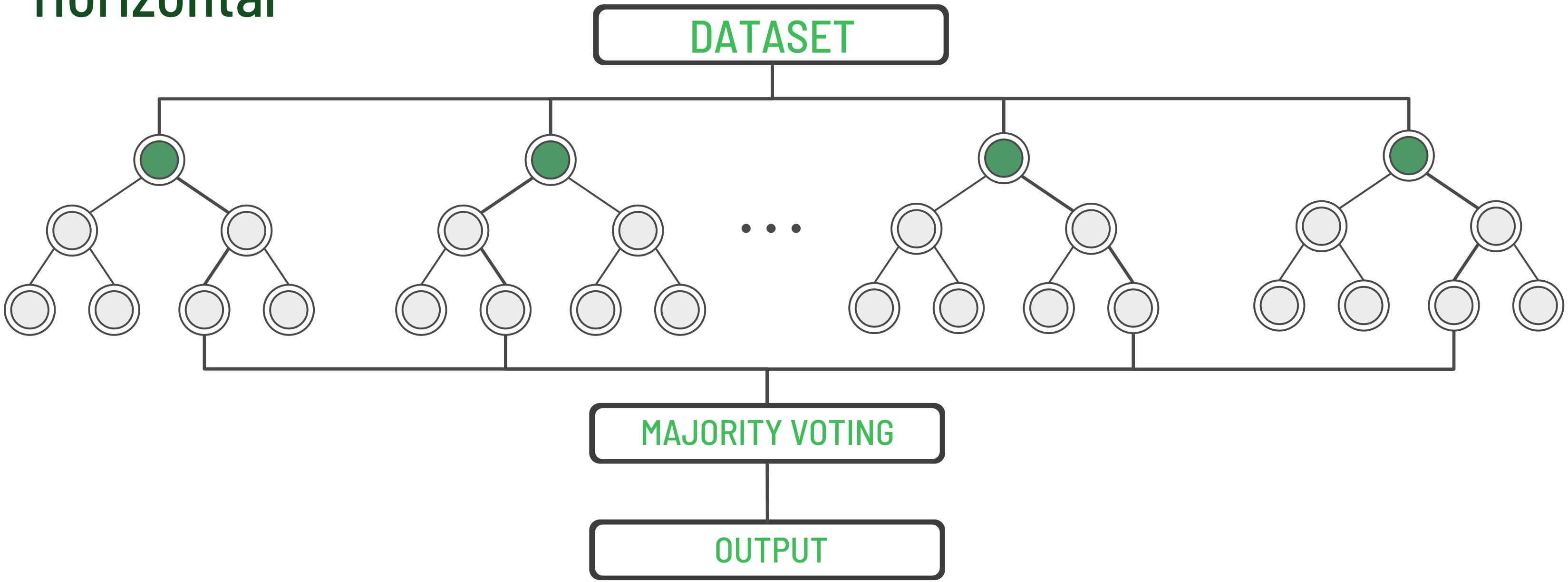


[1] A. B. Arrieta, N. D́ ıaz-Rodŕ ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garć ıa, S. Gil-Ló pez, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

alessandro.verosimile@polimi.it

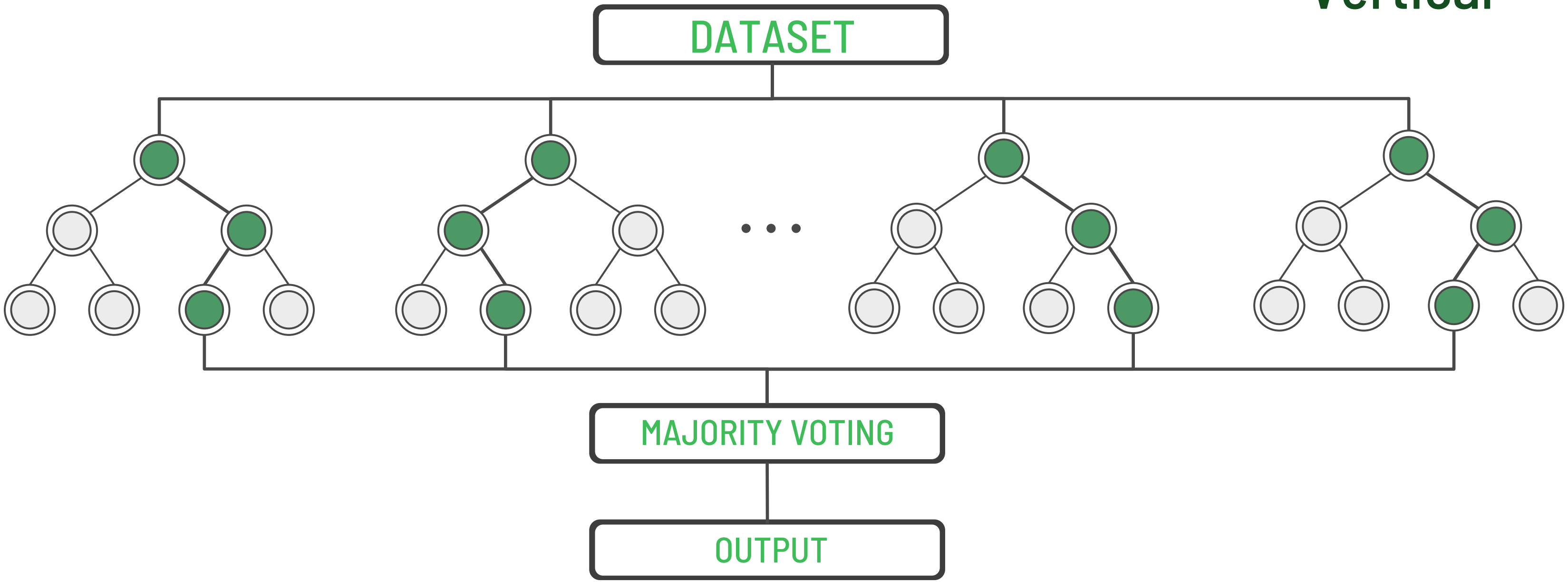
Random Forest

Horizontal



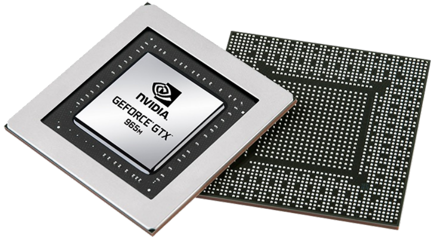
Random Forest

Vertical



Alternative architectures

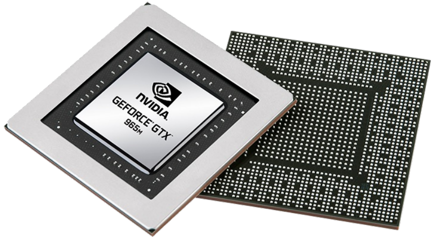
CPU & GPU



FPGA & ASIC

Alternative architectures

CPU & GPU

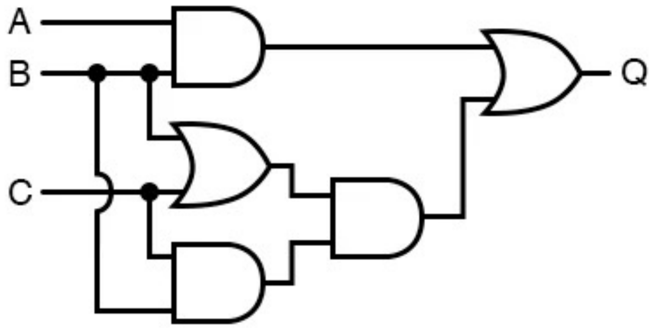


FPGA & ASIC

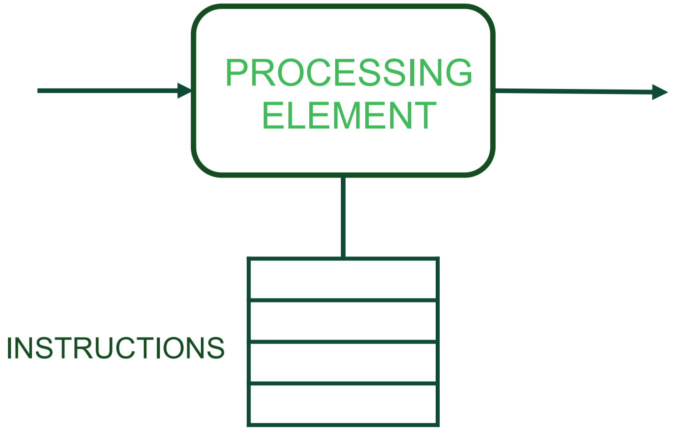
alessandro.verosimile@polimi.it

Design methodologies and SoA works

Logic-Centric



Memory-Centric



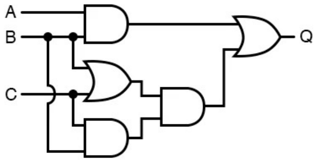
Agenda

- Context definition
- Design methodologies and SoA works
- YoseUe architecture
- Training co-optimization: the MDRFC
- Experimental results

alessandro.verosimile@polimi.it

Design methodologies and SoA works

State of the art about Logic-Centric approach

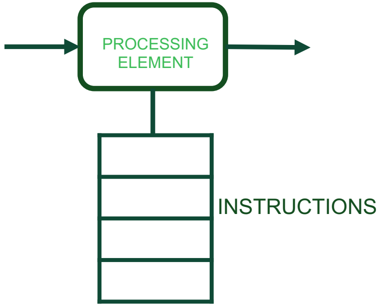


- Conifer [2]: the entire model is hardcoded in the hardware leading to high throughput but high resource consumption

- Entrée [3]: employs partial dynamic reconfiguration to fit larger models

[2] S. Summers, G. Di Guglielmo, J. Duarte, P. Harris, D. Hoang, S. Jin- dariani, E. Kreinar, V. Loncar, J. Ngadiuba, M. Pierini et al., “Fast inference of boosted decision trees in fpgas for particle physics,” Journal of Instrumentation, vol. 15, no. 05, p. P05026, 2020.
[3] A. Damiani, E. Del Sozzo, and M. D. Santambrogio, “Large forests and where to “partially” fit them,” in 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC). IEEE, 2022, pp. 550–555.

Design methodologies and SoA works



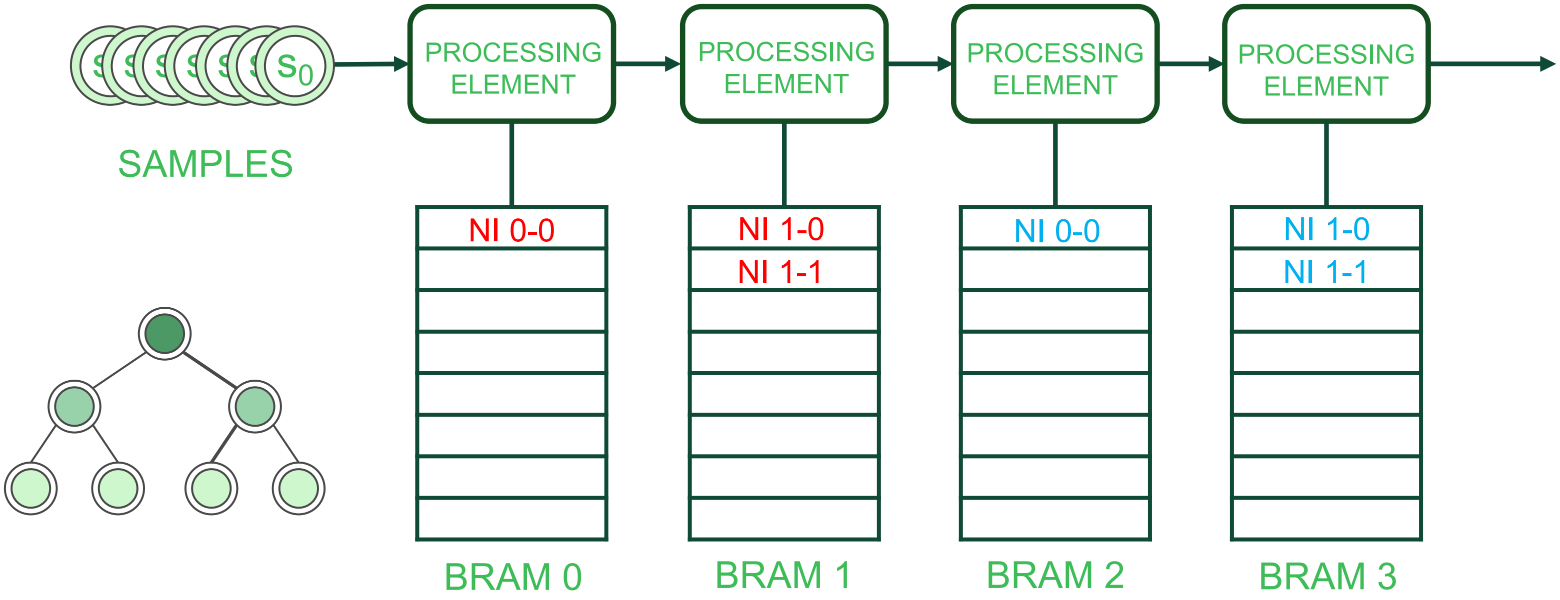
State of the art about Memory-Centric approach

- RF-RISA [4]: implements a specific co-processor and instruction set for RF acceleration. However, it still as significant room for improvement regarding resource-consumption

[4] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, “Rf-risa: A novel flexible random forest accelerator based on fpga,” Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

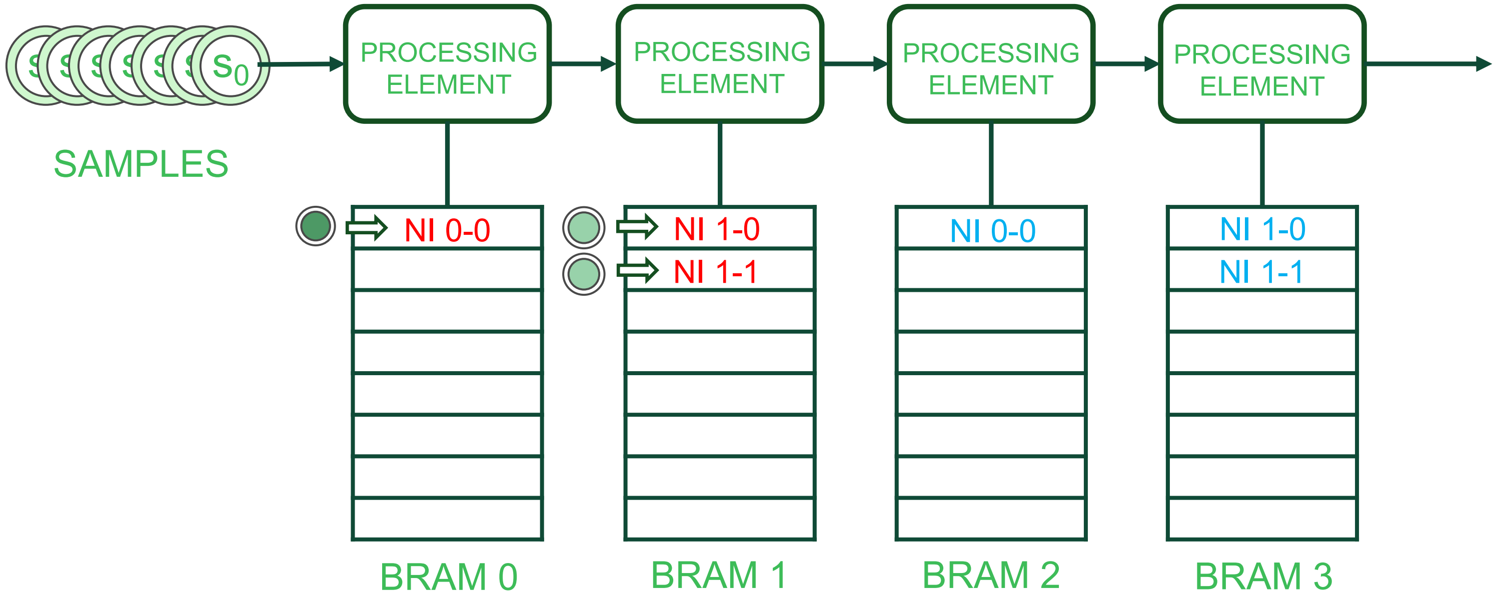
Design methodologies and SoA works

Example: RF-RISA architecture inferring a random forest with two trees of depth 2



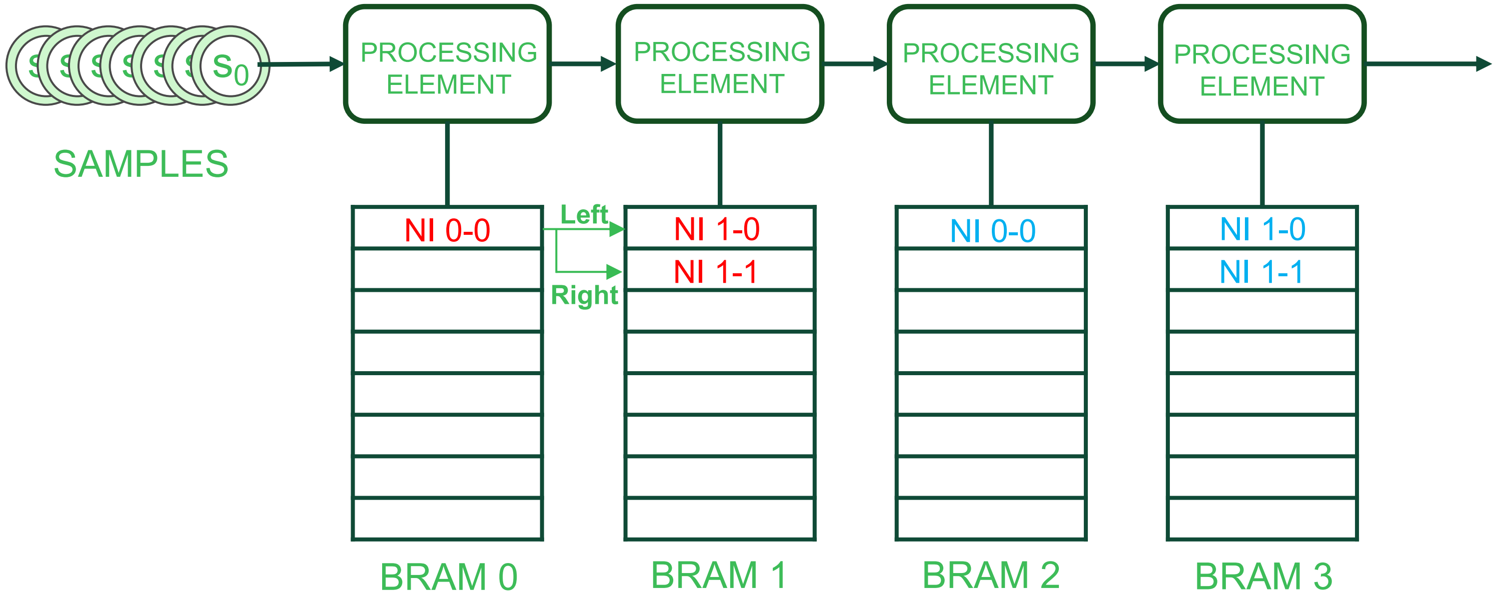
Design methodologies and SoA works

Example: RF-RISA architecture inferring a random forest with two trees of depth 2



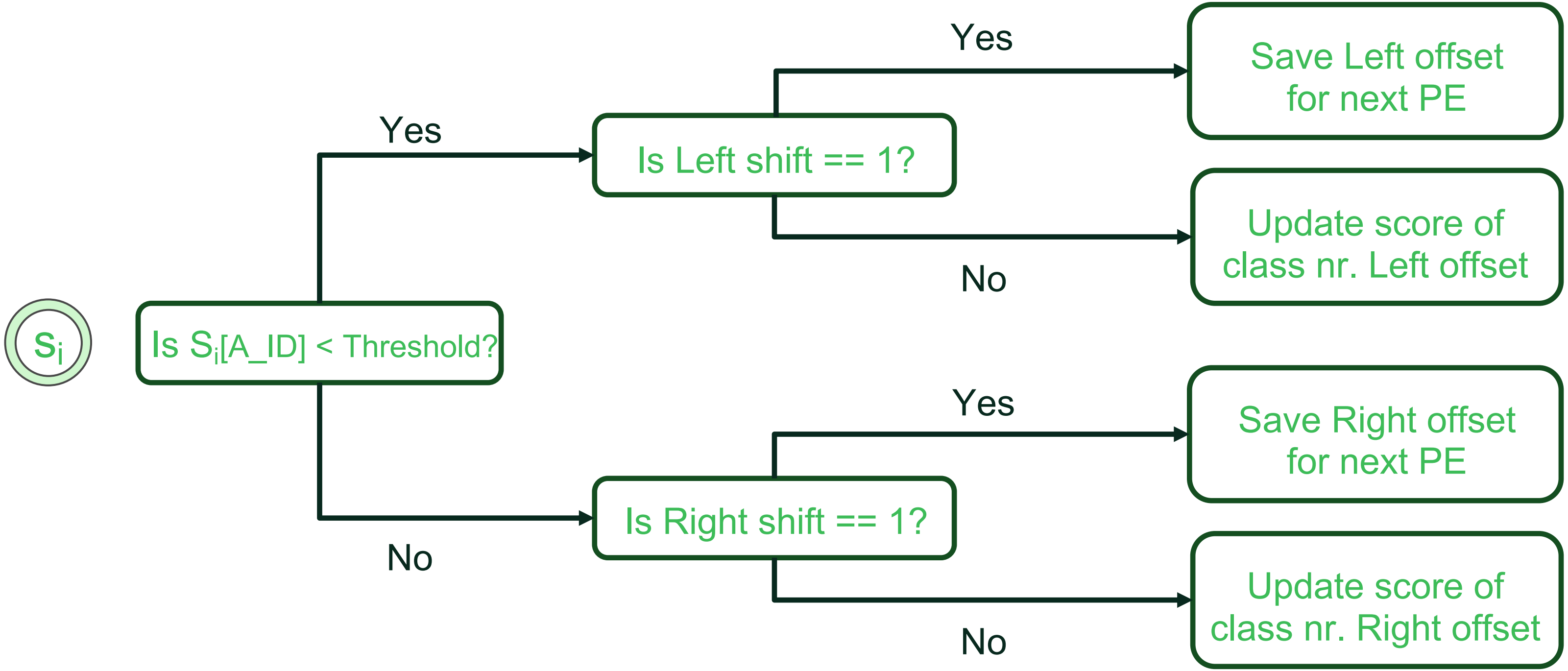
Design methodologies and SoA works

Example: RF-RISA architecture inferring a random forest with two trees of depth 2



alessandro.verosimile@polimi.it

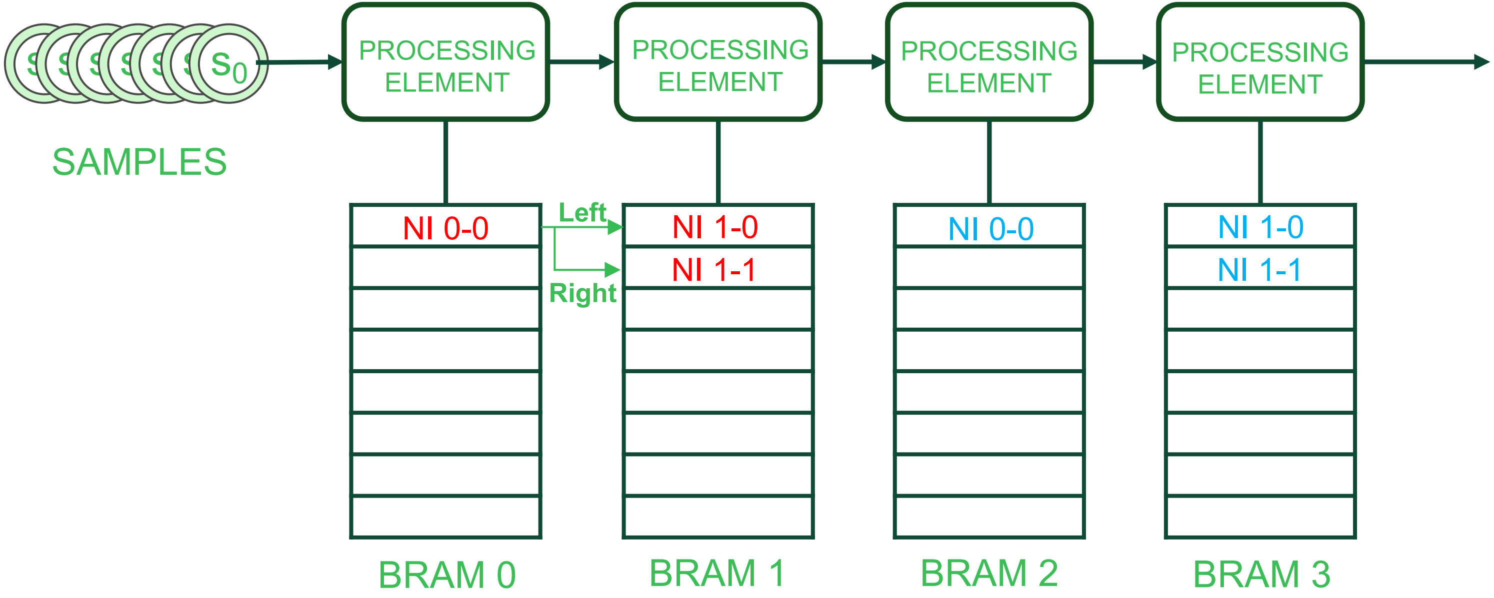
Algorithm



Node RA	A_ID	Threshold	Left shift	Left offset	Right shift	Right offset

Design methodologies and SoA works

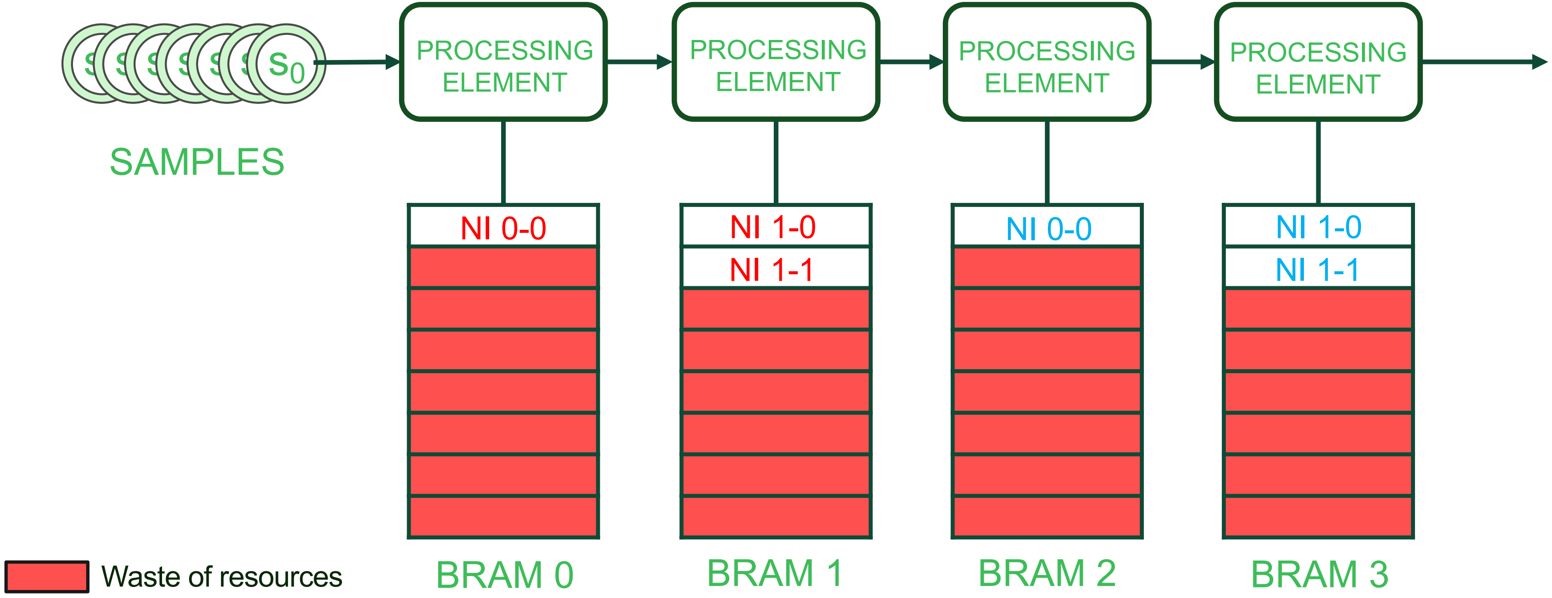
Example: RF-RISA architecture inferring a random forest with two trees of depth 2



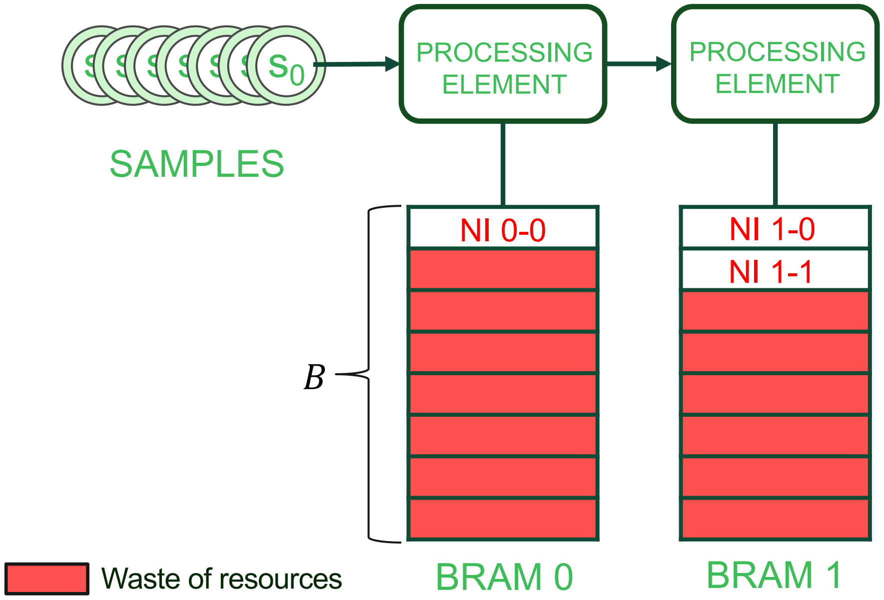
alessandro.verosimile@polimi.it

Design methodologies and SoA works

Example: RF-RISA architecture inferring a random forest with two trees of depth 2



Design methodologies and SoA works



Formally, the ratio of unused memory of RF-RISA is:

$$1 - \frac{|NoInst| \cdot (2^{\hat{d}} - 1)}{|B| \cdot \hat{d}}$$

Legend

- $|B|$ is the size of a memory block
- $\hat{d}$ is the maximum depth of the DTs
- $|NoInst|$ is the size of a Node instruction

Agenda

Context definition

Design methodologies and SoA works

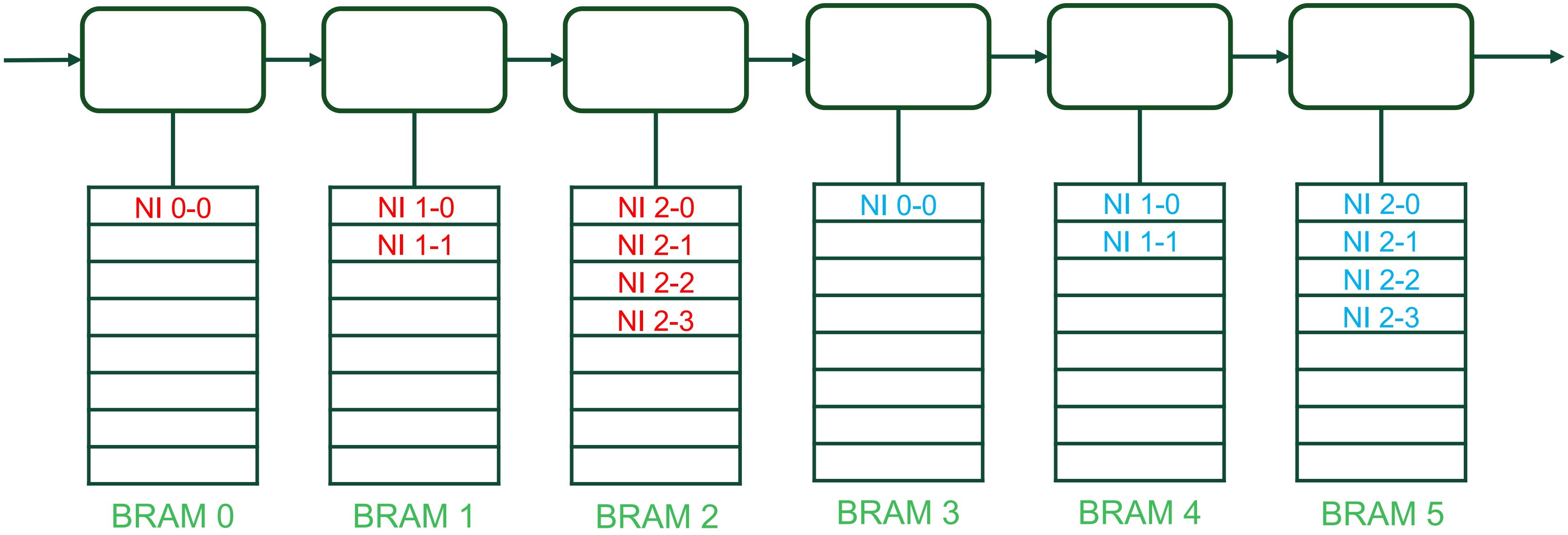
YoseUe architecture

Training co-optimization: the MDRFC

Experimental results

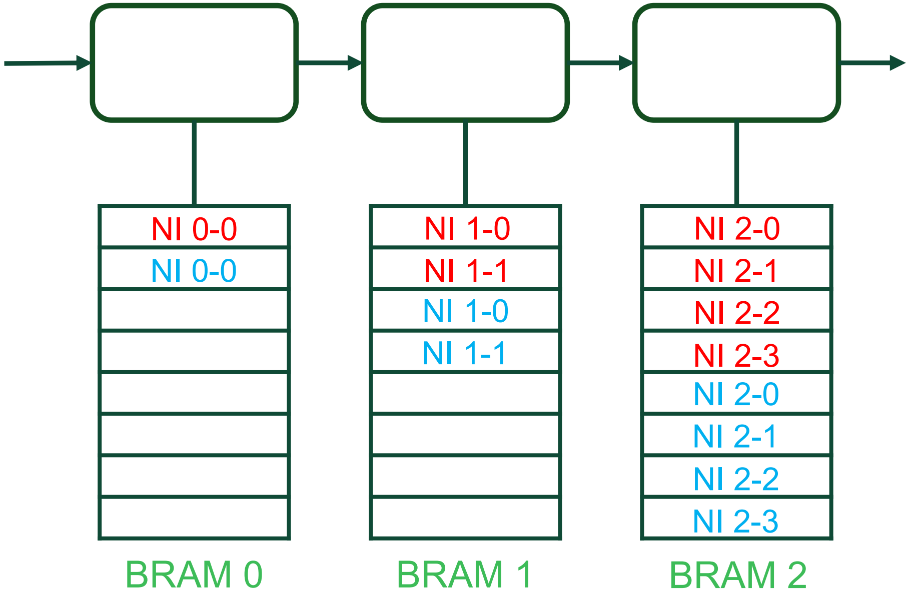
alessandro.verosimile@polimi.it

YoseUe architecture

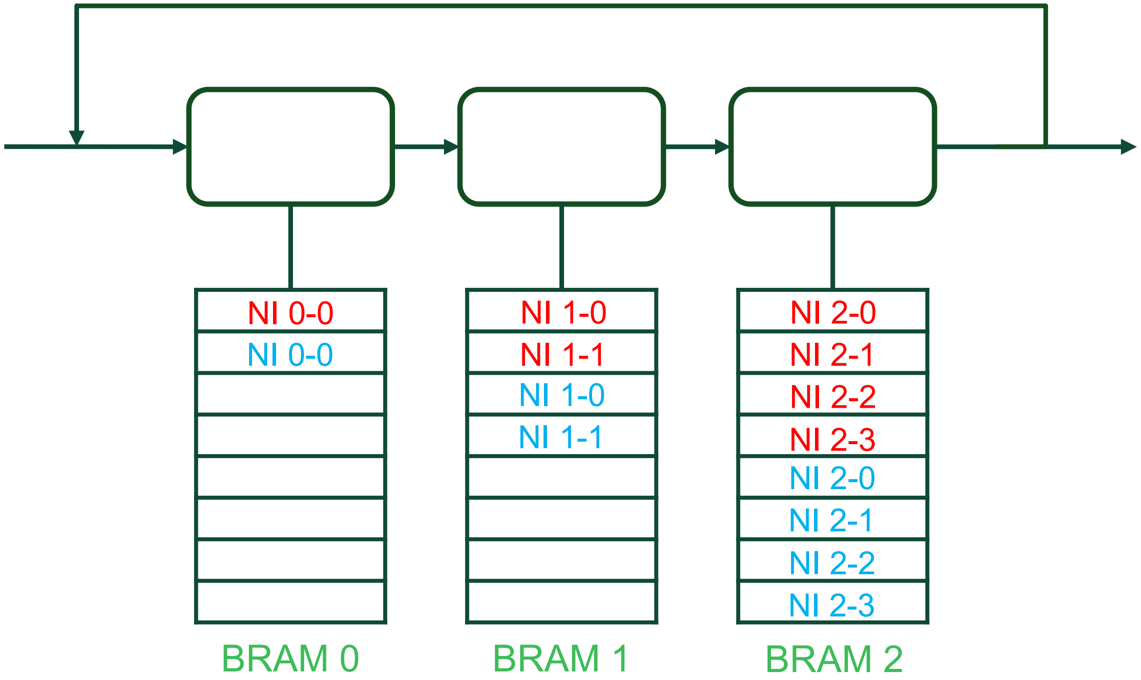


alessandro.verosimile@polimi.it

YoseUe architecture

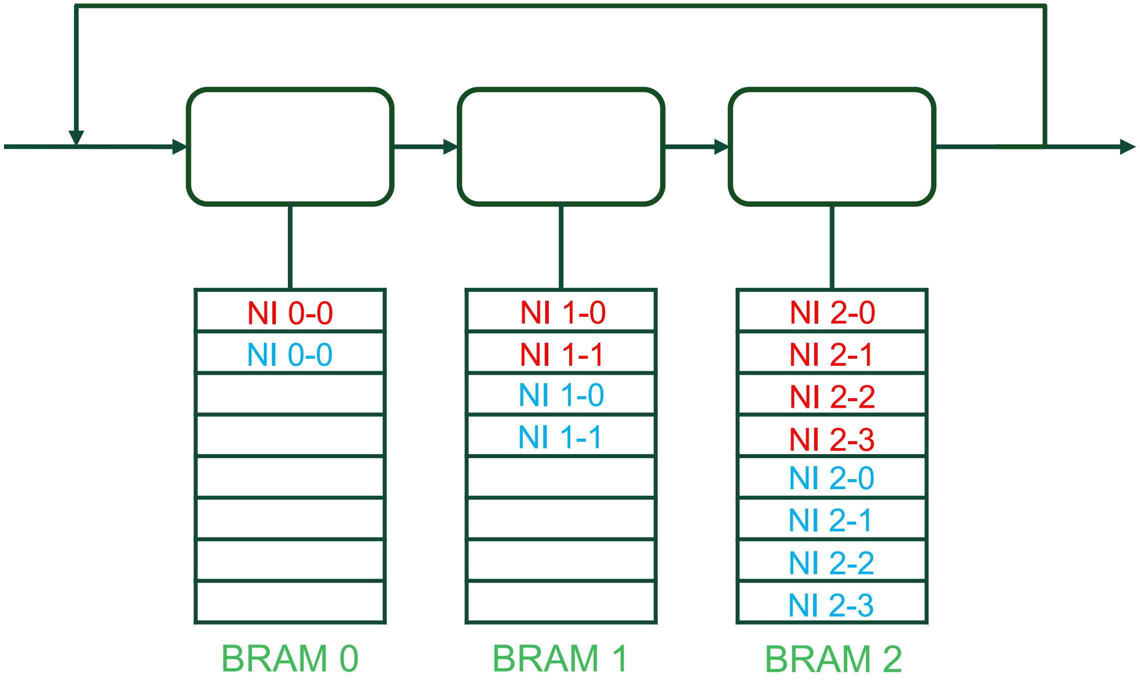


YoseUe architecture



alessandro.verosimile@polimi.it

YoseUe architecture



With our improvement, the ratio of unused memory becomes:

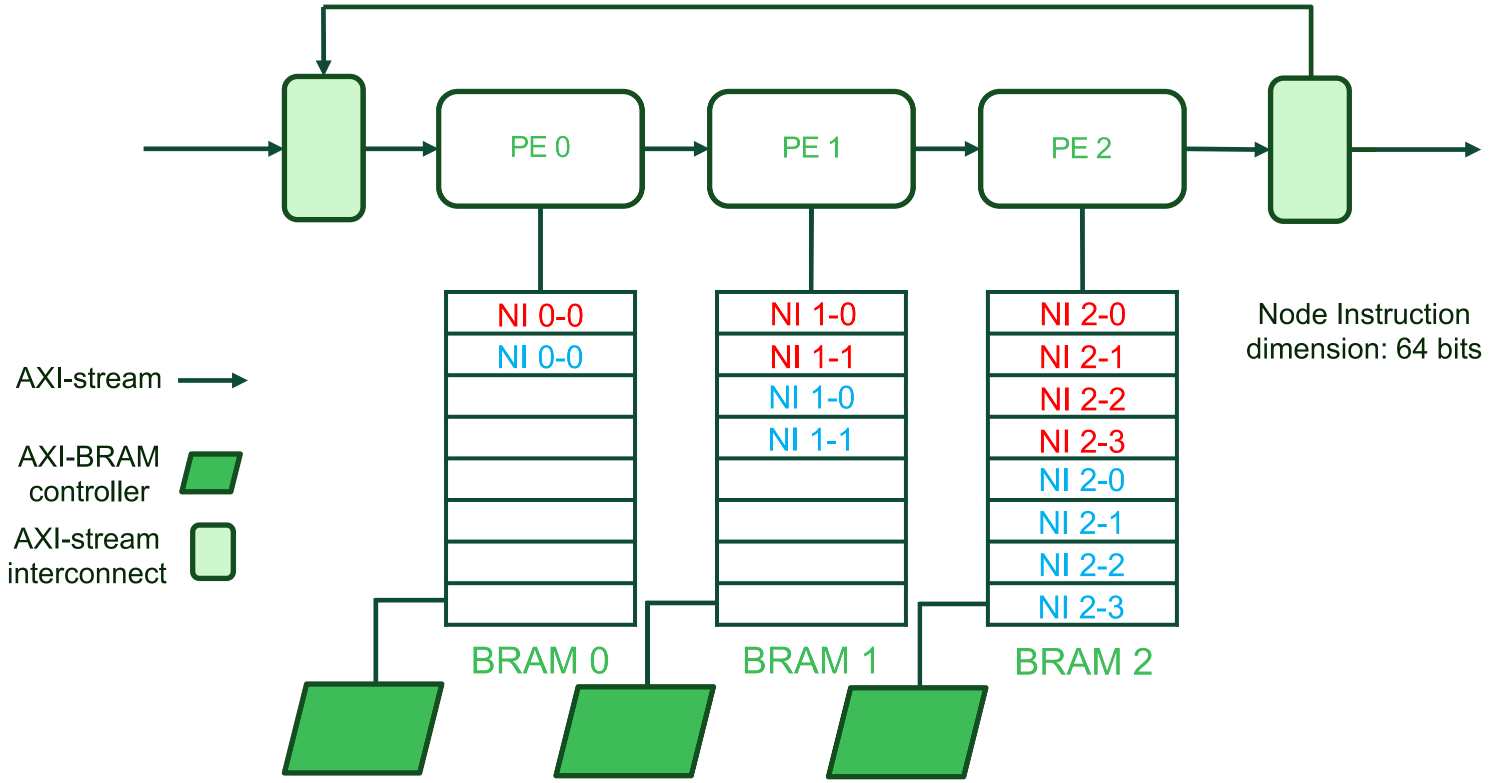
$$1 - \boldsymbol{T}\frac{|NoInst| \cdot (2^{\hat{d}} - 1)}{|B| \cdot \hat{d}}$$

Where $\boldsymbol{T}$ is the number of trees that can fit into a set of PEs, that can be calculated as:

$$\frac{|B|}{|NoInst| \cdot 2^{\hat{d}-1}}$$

alessandro.verosimile@polimi.it

YoseUe architecture



alessandro.verosimile@polimi.it

Agenda

Context definition

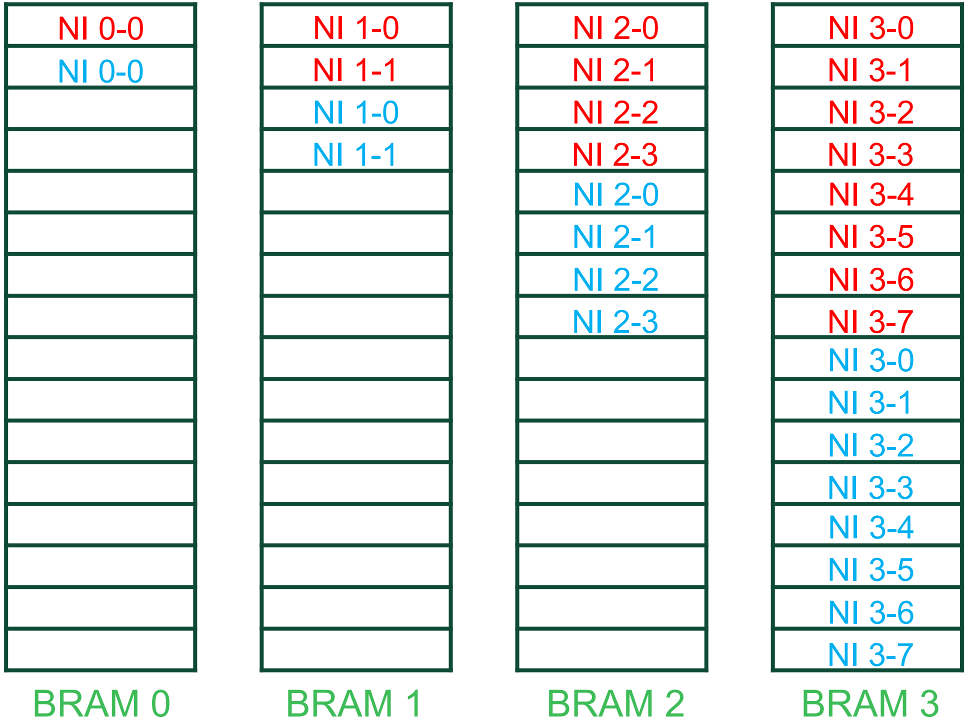
Design methodologies and SoA works

YoseUe architecture

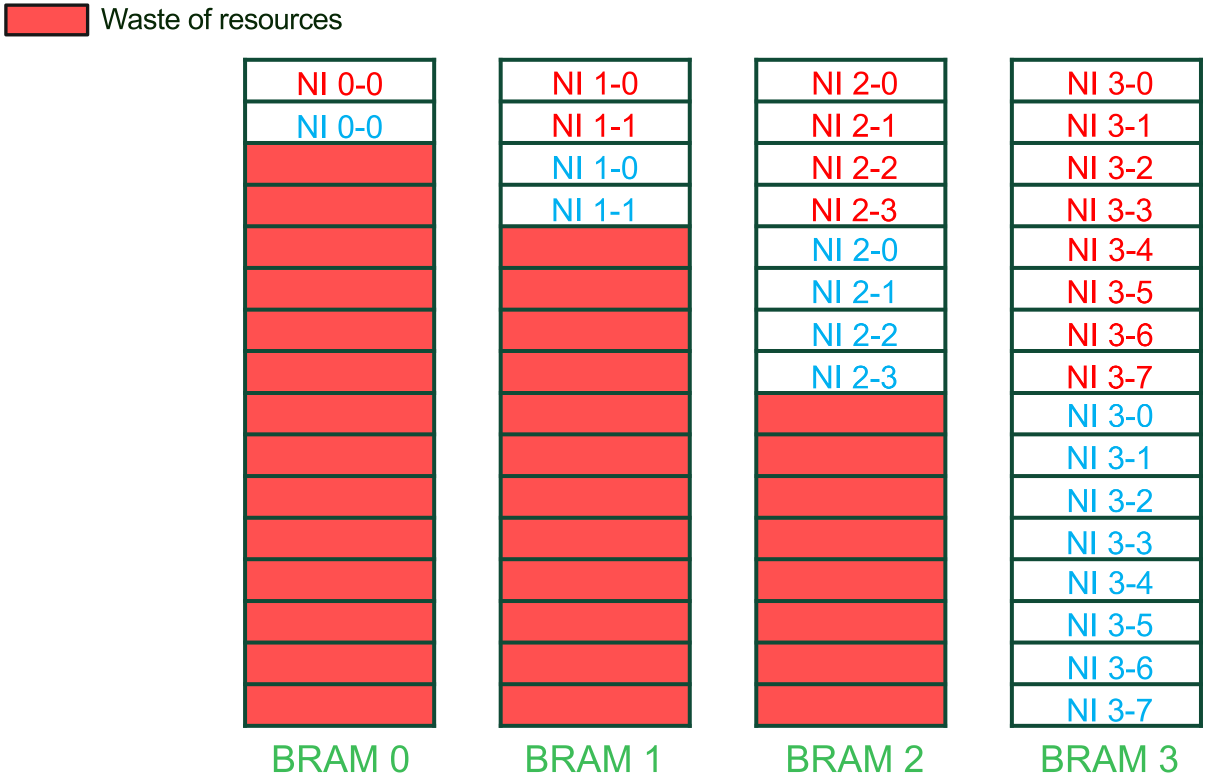
Training co-optimization: the MDRFC

Experimental results

Training co-optimization: the MDRFC



Training co-optimization: the MDRFC



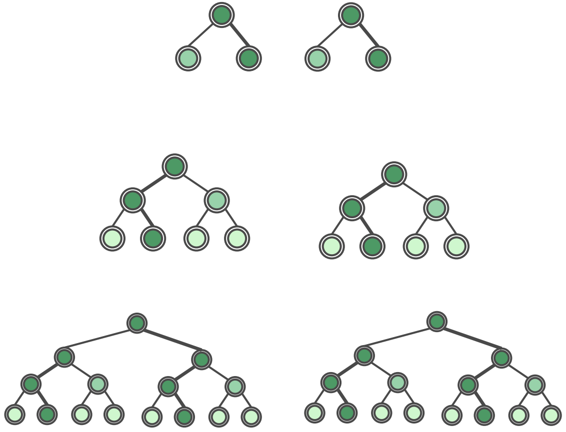
alessandro.verosimile@polimi.it

Training co-optimization: the MDRFC

NI 0-0	NI 1-0	NI 2-0	NI 3-0
NI 0-0	NI 1-1	NI 2-1	NI 3-1
NI 0-0	NI 1-0	NI 2-2	NI 3-2
NI 0-0	NI 1-1	NI 2-3	NI 3-3
NI 0-0	NI 1-0	NI 2-0	NI 3-4
NI 0-0	NI 1-1	NI 2-1	NI 3-5
NI 0-0	NI 1-0	NI 2-2	NI 3-6
NI 0-0	NI 1-1	NI 2-3	NI 3-7
	NI 1-0	NI 2-0	NI 3-0
	NI 1-1	NI 2-1	NI 3-1
	NI 1-0	NI 2-2	NI 3-2
	NI 1-1	NI 2-3	NI 3-3
	NI 1-0	NI 2-0	NI 3-4
	NI 1-1	NI 2-1	NI 3-5
	NI 1-0	NI 2-2	NI 3-6
	NI 1-1	NI 2-3	NI 3-7
BRAM 0	BRAM 1	BRAM 2	BRAM 3

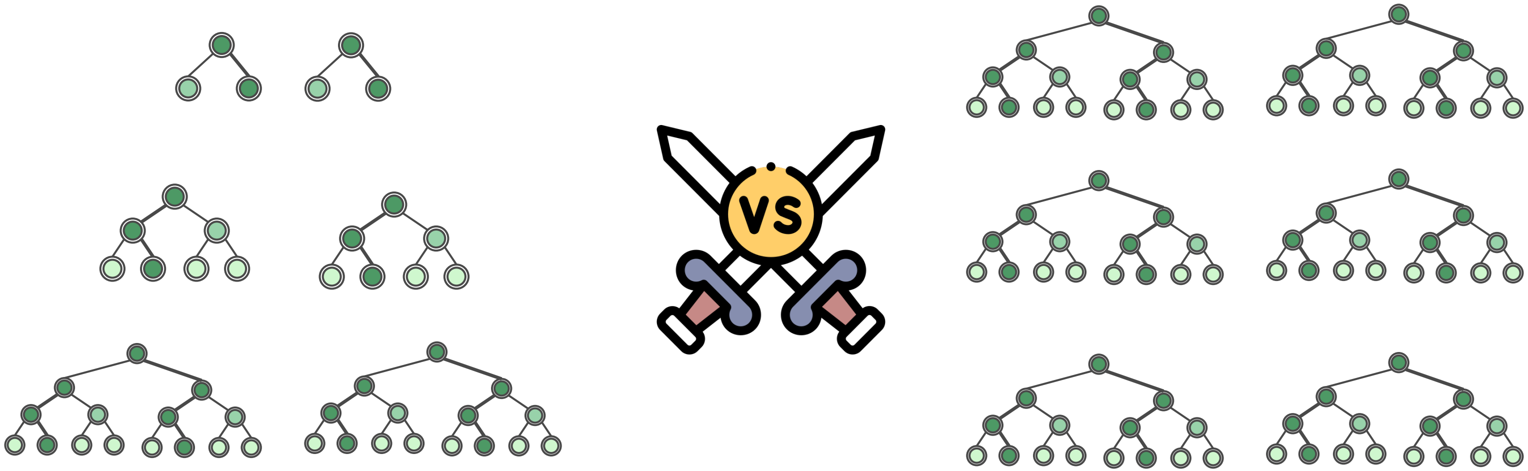
alessandro.verosimile@polimi.it

Training co-optimization: the MDRFC

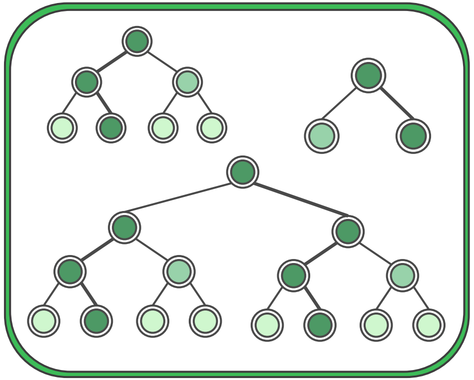


BRAM 0	BRAM 1	BRAM 2	BRAM 3
NI 0-0	NI 1-0	NI 2-0	NI 3-0
NI 0-0	NI 1-1	NI 2-1	NI 3-1
NI 0-0	NI 1-0	NI 2-2	NI 3-2
NI 0-0	NI 1-1	NI 2-3	NI 3-3
NI 0-0	NI 1-0	NI 2-0	NI 3-4
NI 0-0	NI 1-1	NI 2-1	NI 3-5
NI 0-0	NI 1-0	NI 2-2	NI 3-6
NI 0-0	NI 1-1	NI 2-3	NI 3-7
	NI 1-0	NI 2-0	NI 3-0
	NI 1-1	NI 2-1	NI 3-1
	NI 1-0	NI 2-2	NI 3-2
	NI 1-1	NI 2-3	NI 3-3
	NI 1-0	NI 2-0	NI 3-4
	NI 1-1	NI 2-1	NI 3-5
	NI 1-0	NI 2-2	NI 3-6
	NI 1-1	NI 2-3	NI 3-7

Training co-optimization: the MDRFC



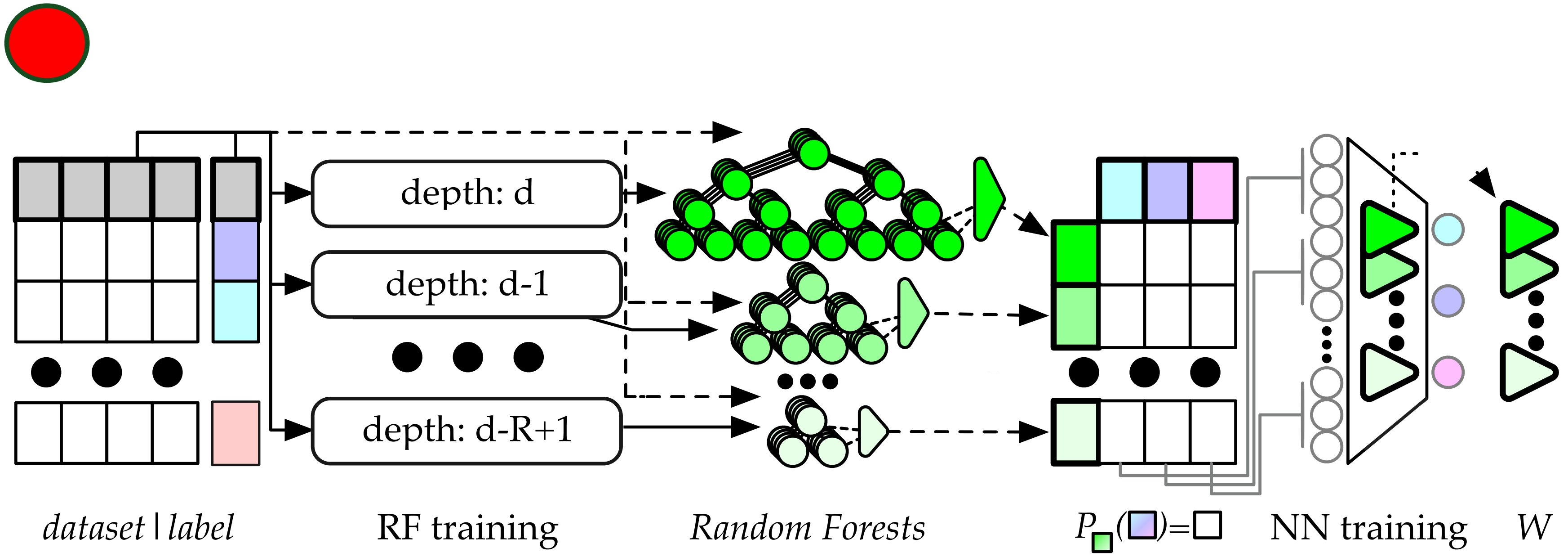
Training co-optimization: the MDRFC



Multi-depth RF

Model performance drop

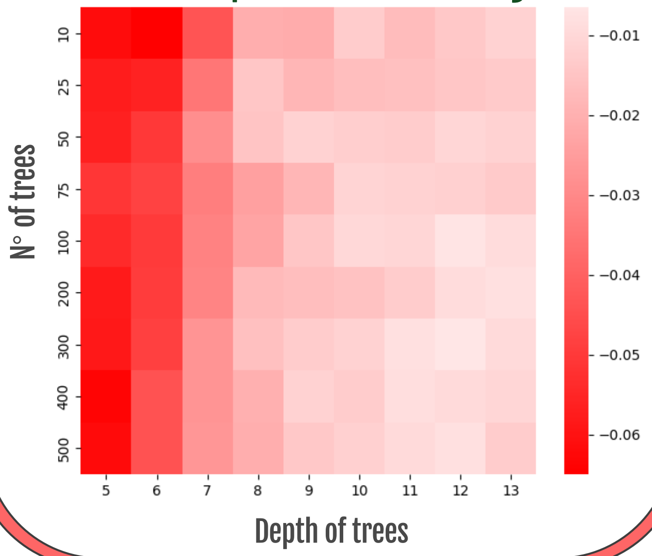
Training co-optimization: the MDRFC



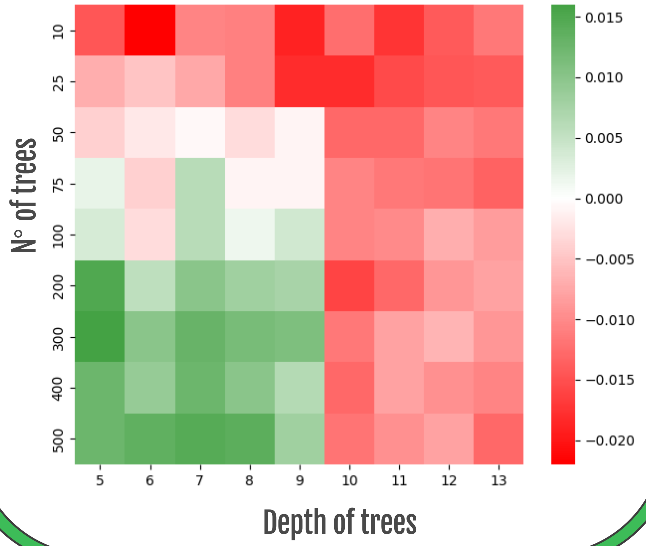
Model results

The tests are performed over 5 datasets, **Satellite**, Shuttle, Accelerometer, Vehicles, Vowel from the UCI ML Repository [5]

Delta of accuracy with multi-depth without weights



Delta of accuracy with multi-depth with weights



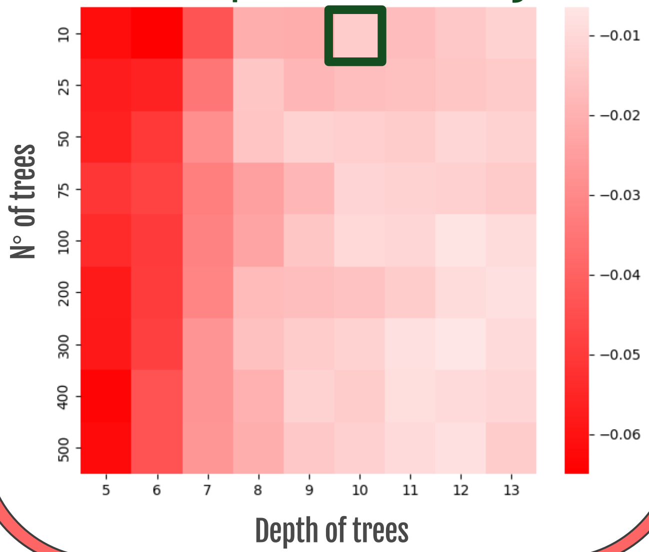
[5] K. N. Markelle Kelly, Rachel Longjohn, "The uci machine learning repository." [Online]. Available: https://archive.ics.uci.edu

Model results

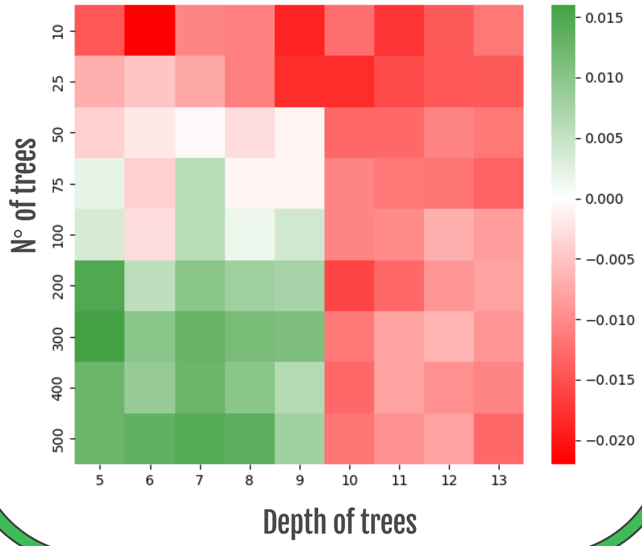
Better than reference

Worse than reference

Delta of accuracy with multi-depth without weights



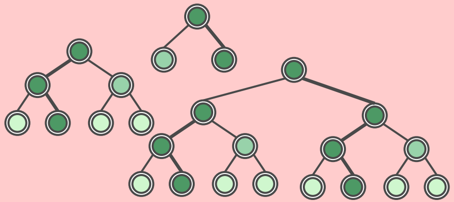
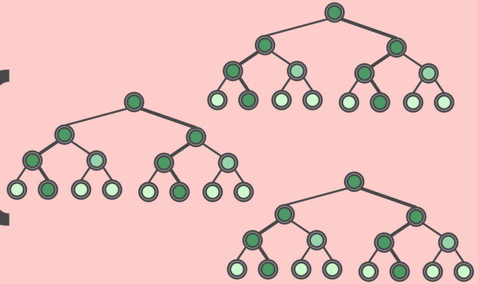
Delta of accuracy with multi-depth with weights



alessandro.verosimile@polimi.it

Model results

Delta =

Acc () – Acc ()

Multi-Depth Random Forest

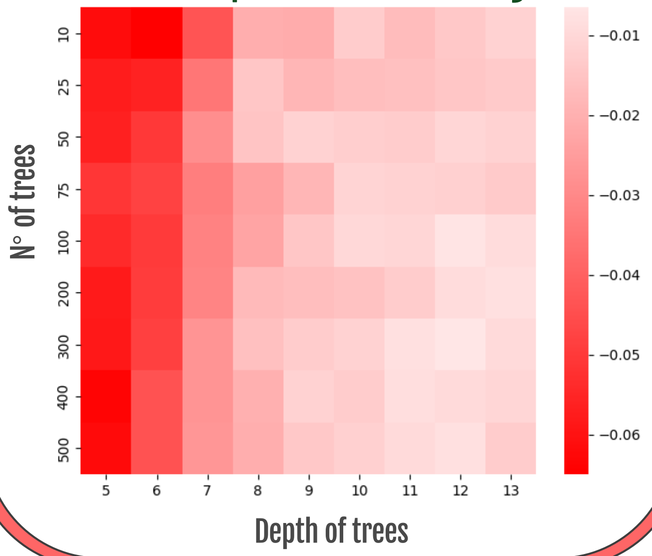
Standard Random Forest

Model results

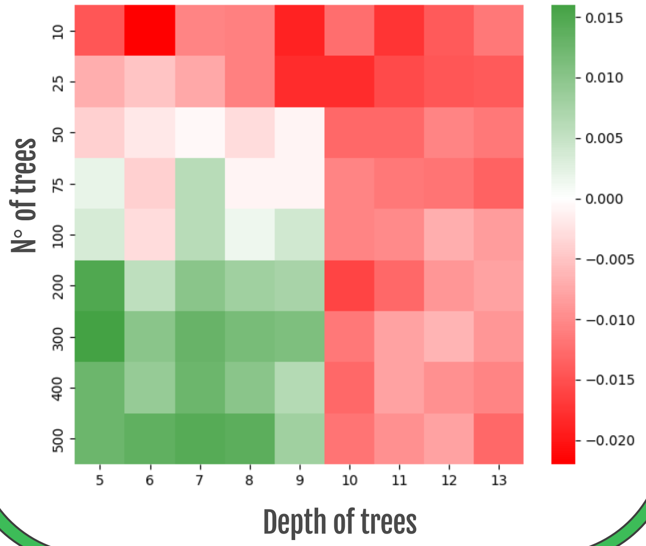
Better than reference

Worse than reference

Delta of accuracy with multi-depth without weights



Delta of accuracy with multi-depth with weigths



alessandro.verosimile@polimi.it

Model results

Better than reference

Worse than reference

Delta of accuracy with multi-depth with weigths

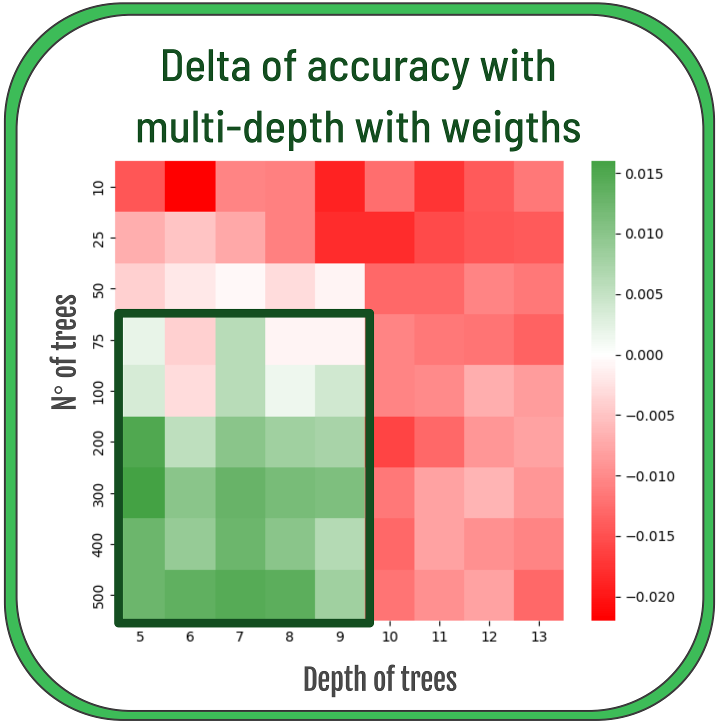
alessandro.verosimile@polimi.it

Model results

Better than reference

Worse than reference

Widely used

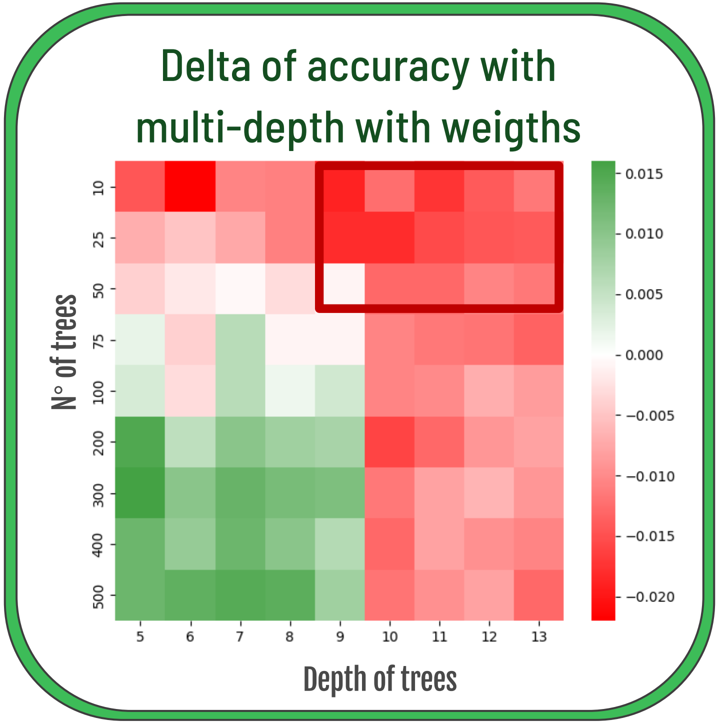


alessandro.verosimile@polimi.it

Model results

Better than reference

Worse than reference



Strongly discouraged

alessandro.verosimile@polimi.it

Agenda

Context definition

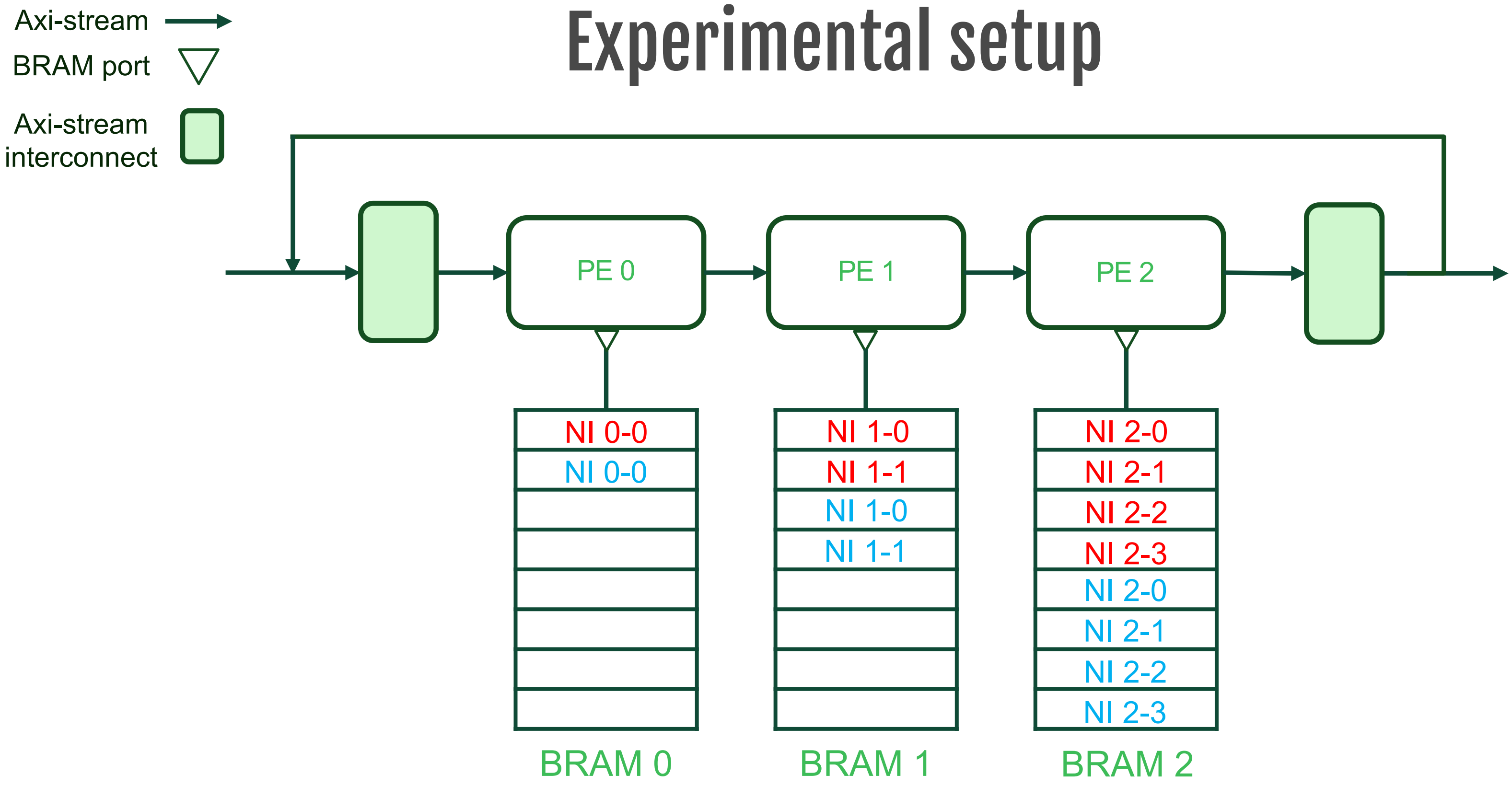
Design methodologies and SoA works

YoseUe architecture

Training co-optimization: the MDRFC

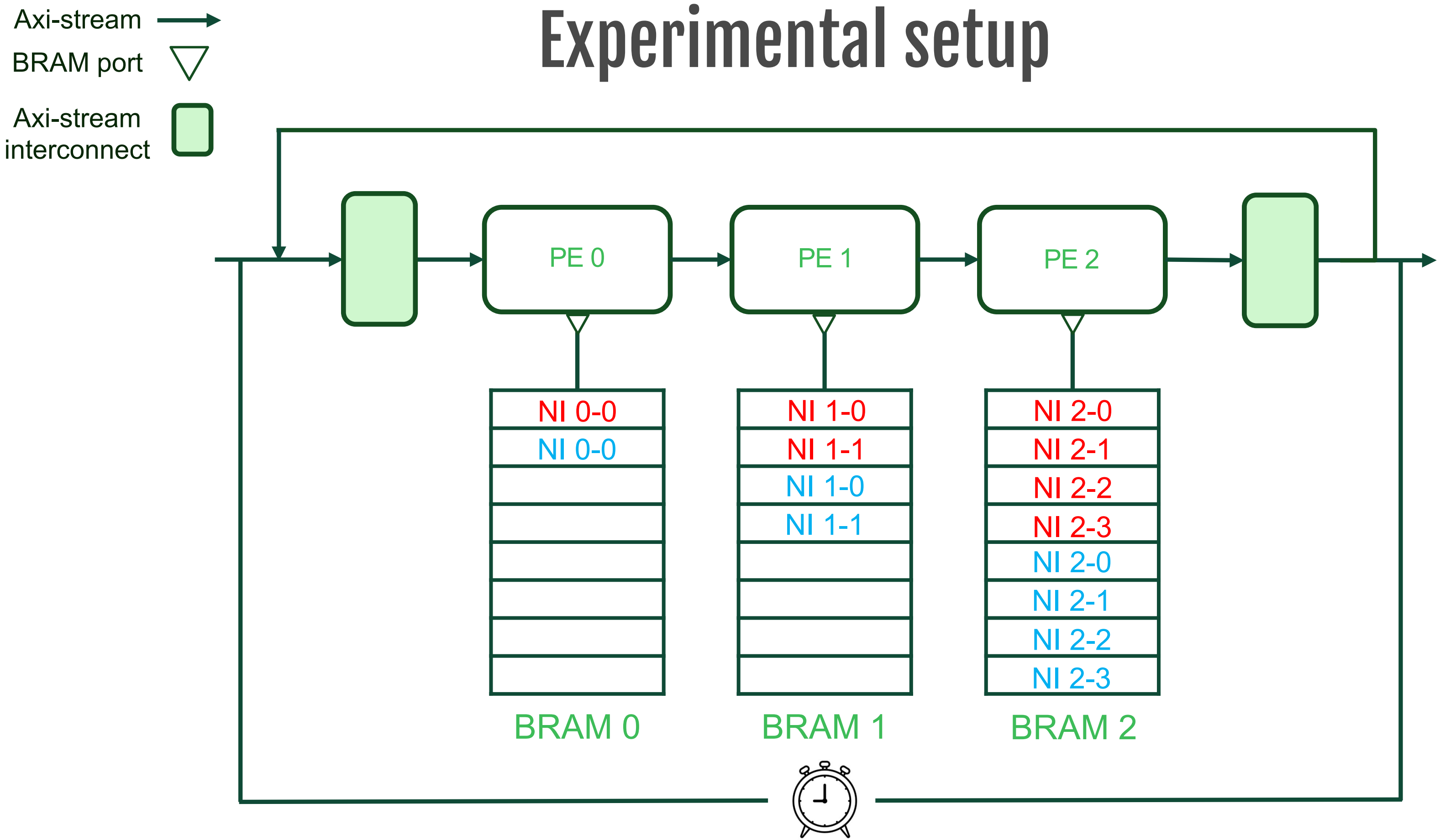
Experimental results

Experimental setup



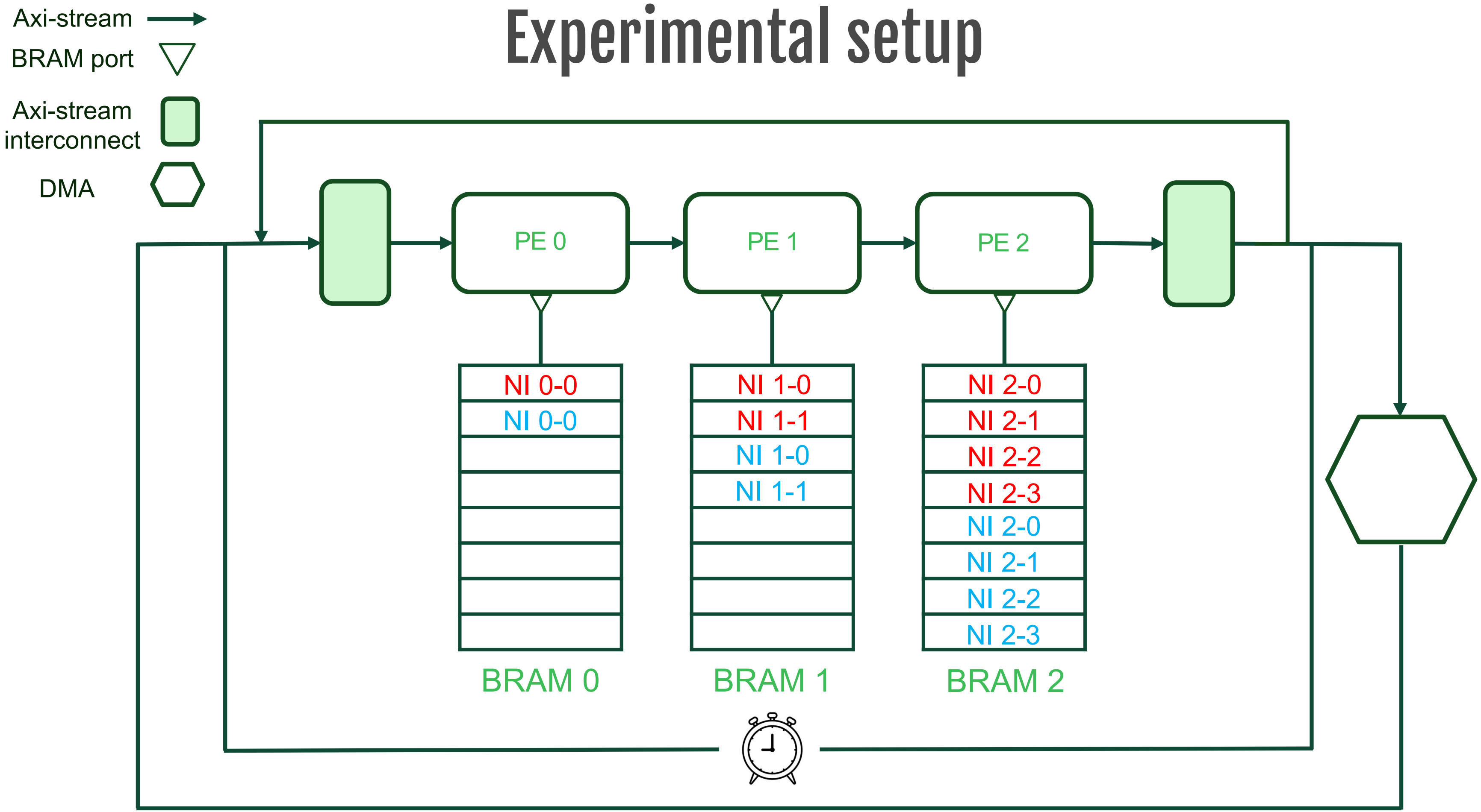
alessandro.verosimile@polimi.it

Experimental setup



alessandro.verosimile@polimi.it

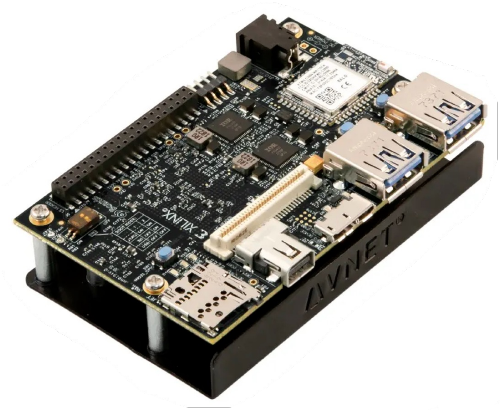
Experimental setup



alessandro.verosimile@polimi.it

Experimental setup

AVNET Ultra96-v2



141,160

FFs

70,560

LUTs

36K BRAMs

216

Zynq Ultrascale+ MPSoC
XCZU3EG

alessandro.verosimile@polimi.it

Experimental results

Throughput

Depth of trees \ N° of trees	3	4	5	6	7	8
5	34M	34M	⊗	⊗	⊗	⊗
7	34M	⊗	⊗	⊗	⊗	⊗
9	⊗	⊗	⊗	⊗	⊗	⊗

RF-RISA [4] re-implementation on Ultra96-v2

Depth of trees \ N° of trees	5	10	25	50	75	100	200	300
5	9.2M	7.6M	7.6M	7.6M	7.6M	9.3M	9.3M	9.3M
7	9.7M	8.7M	8.7M	11.4M	16.1M	⊗	⊗	⊗
9	10.3M	13.4M	⊗	⊗	⊗	⊗	⊗	⊗

YoseUe on Ultra96-v2

$M = x10^9$

Tables refer to an implementation with samples composed of 5 attributes (ap_fixed<32,16>)

[4] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, “Rf-risa: A novel flexible random forest accelerator based on fpga,” Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

alessandro.verosimile@polimi.it

Experimental results

Synthesizable configurations

RF-RISA [4] re-implementation on Ultra96-v2

Depth of trees \ N° of trees	3	4	5	6	7	8
5	✓	✓	✗	✗	✗	✗
7	✓	✗	✗	✗	✗	✗
9	✗	✗	✗	✗	✗	✗

YoseUe on Ultra96-v2

Depth of trees \ N° of trees	5	10	25	50	75	100	200	300
5	✓	✓	✓	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✗	✗	✗
9	✓	✓	✗	✗	✗	✗	✗	✗

Tables refer to an implementation with samples composed of 5 attributes (ap_fixed<32,16>)

[4] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, “Rf-risa: A novel flexible random forest accelerator based on fpga,” Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

alessandro.verosimile@polimi.it

Conclusions

- **MDRFC**: new Ensemble DT-based model for **efficient inference**

- A DSA to infer the MDRFC, allowing the inference of **two orders of magnitudes** more trees w.r.t. SoA

Future Work

1.

Allowing the execution of an unbounded number of trees by adding pre-fetching modules for dynamically loading new instructions into memories.

2.

Study different topologies of the architecture to carry out inference in a more efficient way.

alessandro.verosimile@polimi.it

Thanks for your attention!

Alessandro Verosimile
alessandro.verosimile@polimi.it

Andrea Damiani
andrea.damiani@polimi.it

Alessandro Tierno
alessandro.tierno@mail.polimi.it

Marco Domenico Santambrogio
marco.santambrogio@polimi.it

Thanks for your attention!

Alessandro Verosimile
alessandro.verosimile@polimi.it

Alessandro Tierno
alessandro.tierno@mail.polimi.it

Andrea Damiani
andrea.damiani@polimi.it

Marco Domenico Santambrogio
marco.santambrogio@polimi.it

FURTHER DETAILS

LUTs utilization

RF-RISA[1] re-implementation on Ultra96-v2

N° of trees

Depth of trees	3	4	5	6	7	8
5	56.35%	76.55%	⊗	⊗	⊗	⊗
7	80.19%	⊗	⊗	⊗	⊗	⊗
9	⊗	⊗	⊗	⊗	⊗	⊗

The total number of LUTs on the Ultra96-v2 is 70560

YoseUe on Ultra96-v2

N° of trees

Depth of trees	5	10	25	50	75	100	200	300
5	42.59%	42.60%	42.60%	42.63%	42.70%	42.33%	65.14%	65.14%
7	49.85%	49.88%	49.89%	82.10%	82.10%	⊗	⊗	⊗
9	62.05%	61.92%	⊗	⊗	⊗	⊗	⊗	⊗

In each configuration, an additional 10% of LUTs is consumed due to the experimental setup

Tables refer to an implementation with samples composed of 5 attributes (ap_fixed<32,16>)

[1] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, "Rf-risa: A novel flexible random forest accelerator based on fpga," Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

alessandro.verosimile@polimi.it

FFs utilization

RF-RISA[1] re-implementation on Ultra96-v2

N° of trees

Depth of trees	3	4	5	6	7	8
5	40.14%	53.92%	⊗	⊗	⊗	⊗
7	57.24%	⊗	⊗	⊗	⊗	⊗
9	⊗	⊗	⊗	⊗	⊗	⊗

The total number of FFs on the Ultra96-v2 is 141120

YoseUe on Ultra96-v2

N° of trees

Depth of trees	5	10	25	50	75	100	200	300
5	32.44%	32.47%	32.51%	32.55%	32.58%	32.59%	48.73%	48.73%
7	38.11%	38.15%	38.20%	61.01%	61.01%	⊗	⊗	⊗
9	46.15%	46.19%	⊗	⊗	⊗	⊗	⊗	⊗

In each configuration, an additional 7.7% of FFs is consumed due to the experimental setup

Tables refer to an implementation with samples composed of 5 attributes (ap_fixed<32,16>)

[1] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, “Rf-risa: A novel flexible random forest accelerator based on fpga,” Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

alessandro.verosimile@polimi.it

BRAMs utilization

RF-RISA[1] re-implementation on Ultra96-v2

Depth of trees \ N° of trees	3	4	5	6	7	8
5	18.06%	38.79%	⊗	⊗	⊗	⊗
7	41.2%	⊗	⊗	⊗	⊗	⊗
9	⊗	⊗	⊗	⊗	⊗	⊗

YoseUe on Ultra96-v2

Depth of trees \ N° of trees	5	10	25	50	75	100	200	300
5	14.81%	14.81%	14.81%	14.81%	14.81%	14.81%	19.44%	19.44%
7	16.67%	16.67%	16.67%	23.15%	23.15%	⊗	⊗	⊗
9	18.52%	18.52%	⊗	⊗	⊗	⊗	⊗	⊗

The total number of BRAMs on the Ultra96-v2 is 216

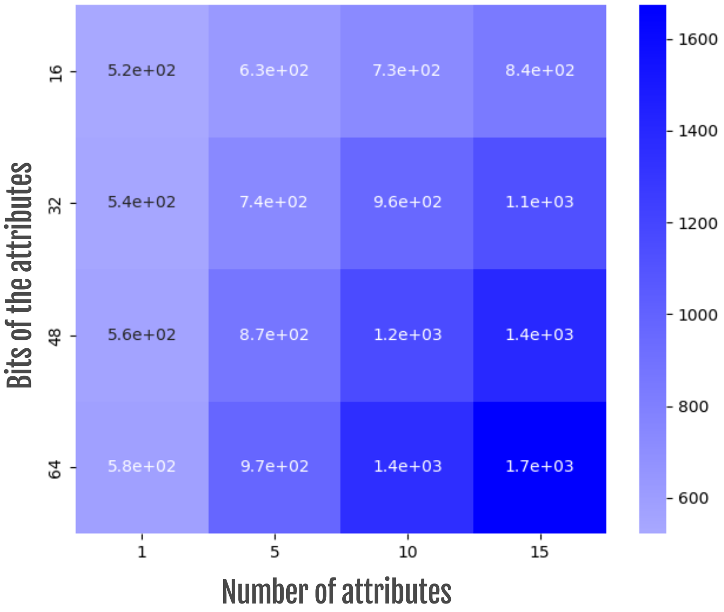
Tables refer to an implementation with samples composed of 5 attributes (ap_fixed<32,16>)

[1] S. Zhao, S. Chen, H. Yang, F. Wang, and Z. Wei, “Rf-risa: A novel flexible random forest accelerator based on fpga,” Journal of Parallel and Distributed Computing, vol. 157, pp. 220–232, 2021

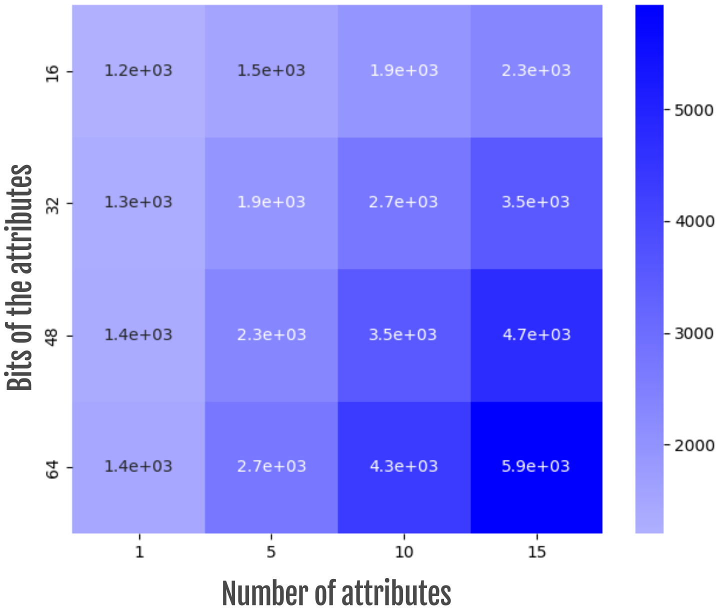
PE’s resource utilization

How does the number of attributes of the samples and their dimension influences LUTs and FFs utilization?

LUTs



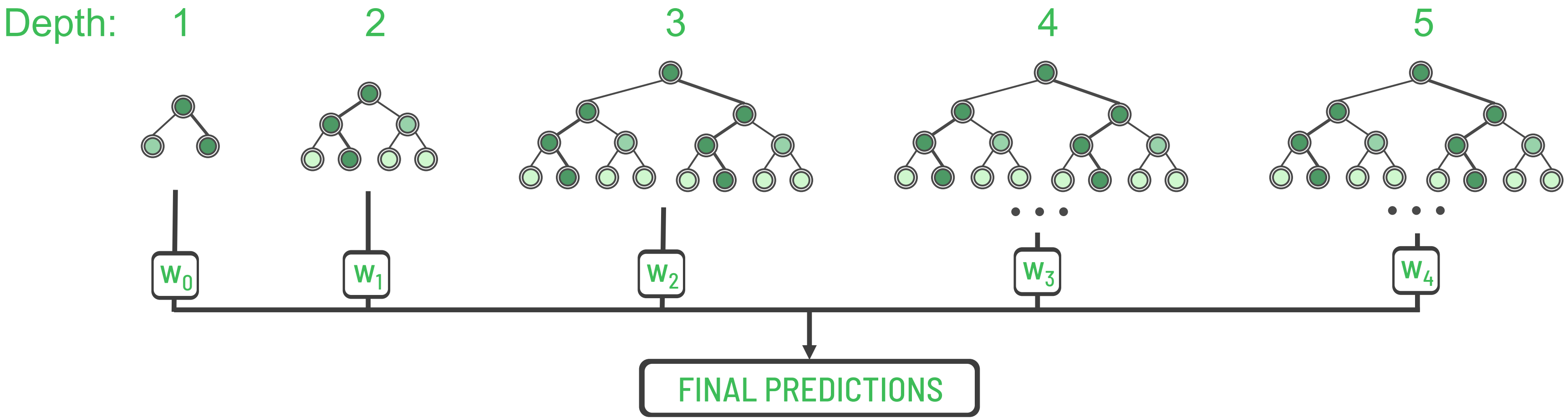
FFs



alessandro.verosimile@polimi.it

Training co-optimization

What happens if we weight the predictions of the single sub-models with different depths?

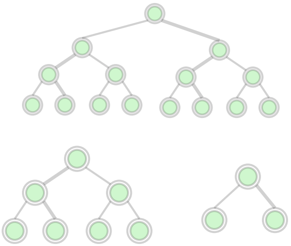


The weights are obtained through the training of a small neural network

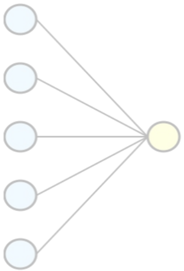
alessandro.verosimile@polimi.it

Model steps

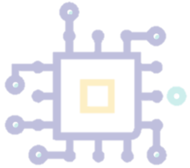
Dataset



Multi depth Random Forest training



Training a small NN to weigh the different depths

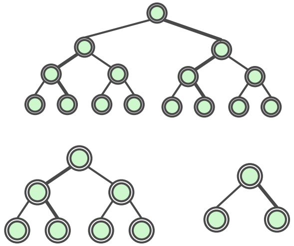


Inference on FPGA

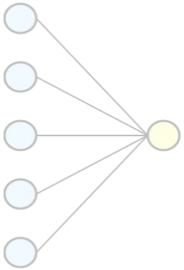
alessandro.verosimile@polimi.it

Model steps

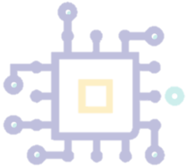
Dataset



Multi depth Random Forest training



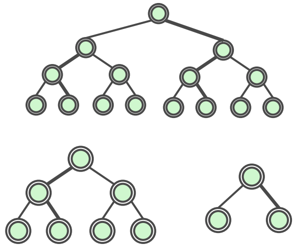
Training a small NN to weigh the different depths



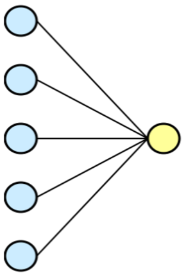
Inference on FPGA

Model steps

Dataset



Multi depth Random Forest training

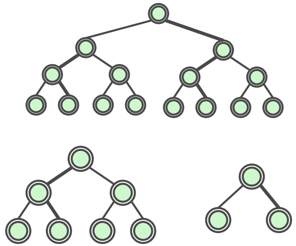


Training a small NN to weigh the different depths

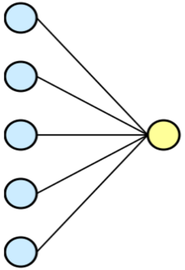
Inference on FPGA

Model steps

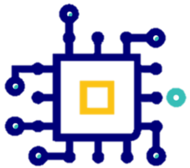
Dataset



Multi depth Random Forest training



Training a small NN to weigh the different depths



Inference on FPGA

alessandro.verosimile@polimi.it