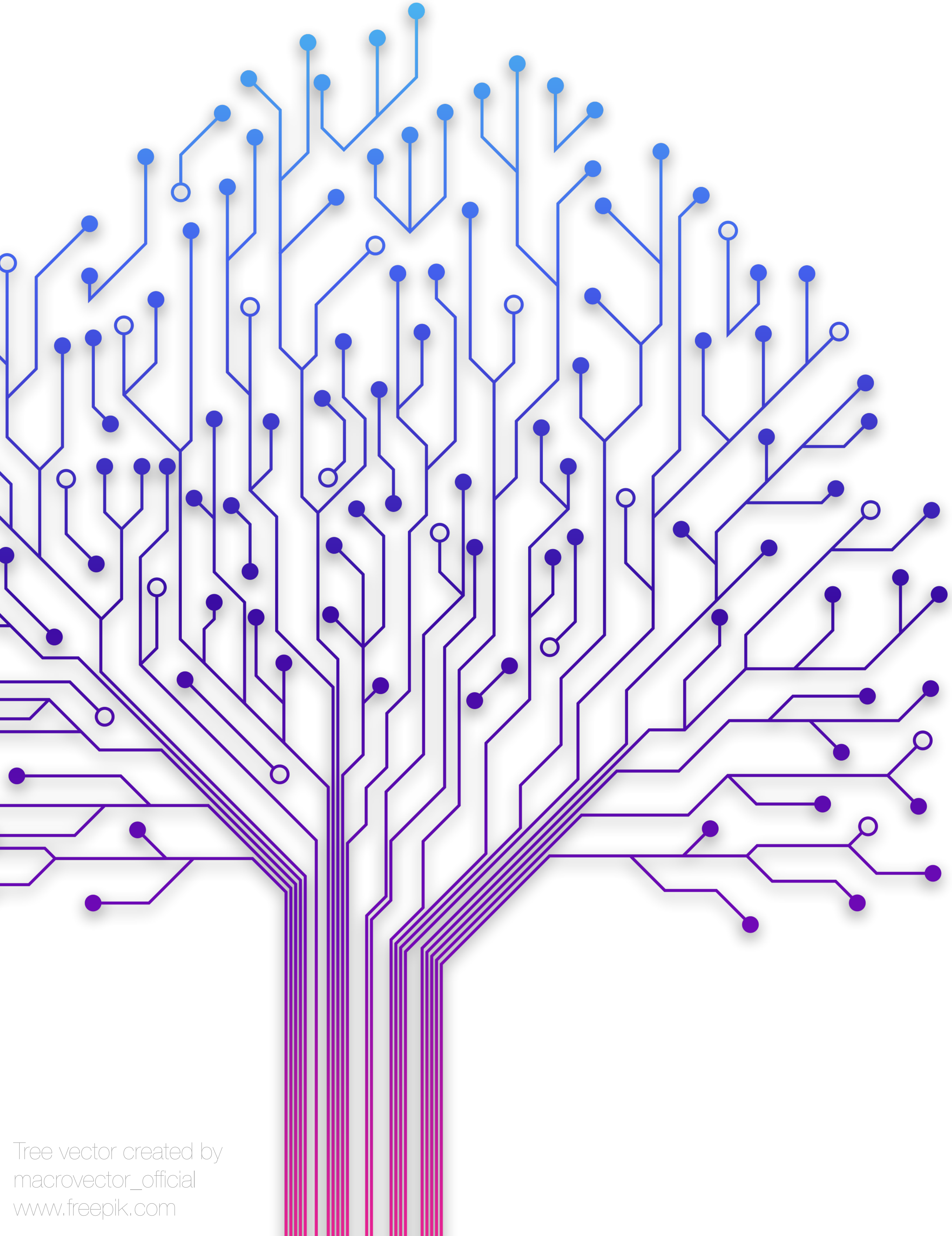


Tree vector created by
macrovector_official
www.freepik.com

Large Forests and Where to “Partially” Fit Them

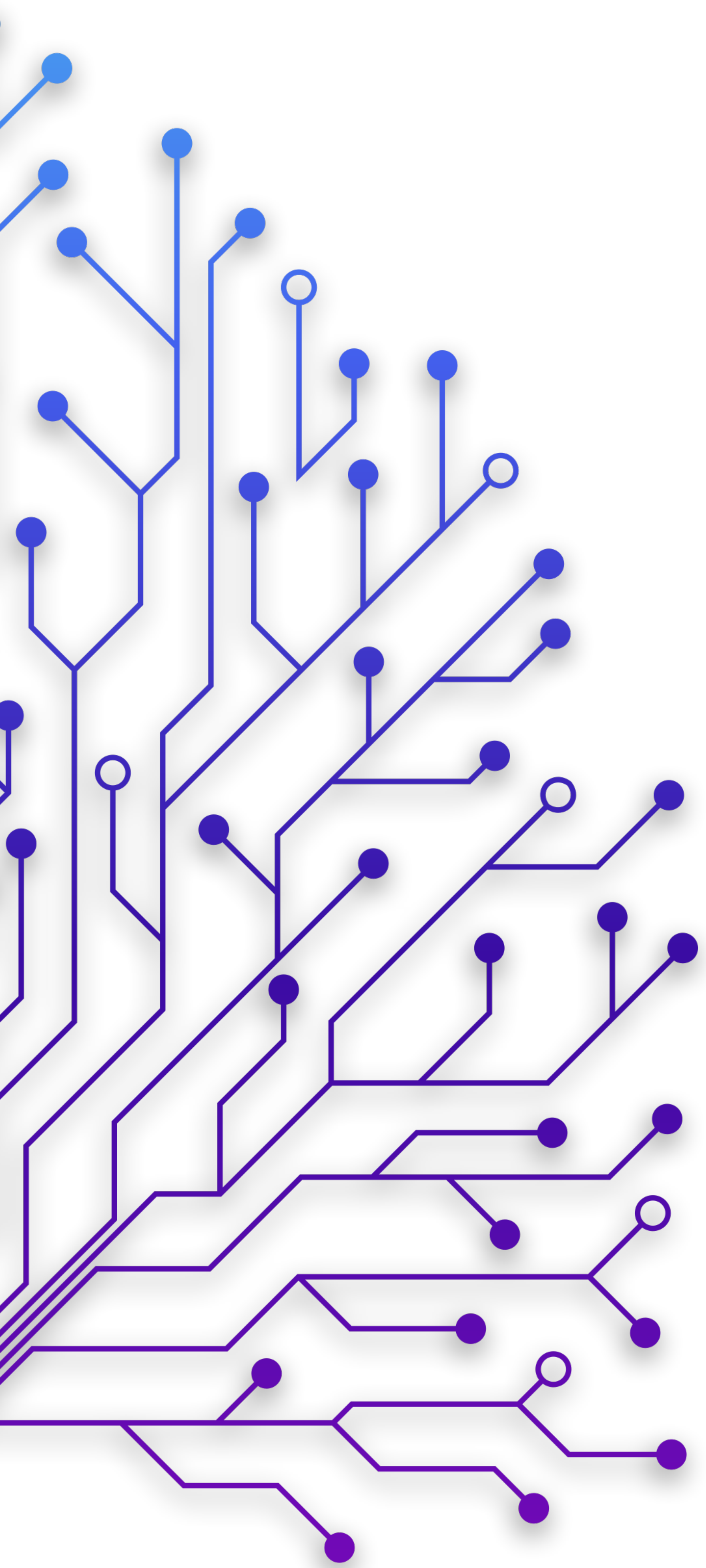
Exploiting partial dynamic reconfiguration towards explainable AIoT

Andrea Damiani - andrea.damiani@polimi.it
Emanuele Del Sozzo - emanuele.delsozzo@polimi.it
Marco D. Santambrogio - marco.santambrogio@polimi.it

January, 2022 - ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

POLITECNICO
MILANO 1863

POLITECNICO MILANO 1863
NECST laboratory



CAN ALL THE INTELLIGENT THINGS

RELY ON A BRAIN

THAT SITS THOUSANDS OF KILOMETERS AWAY

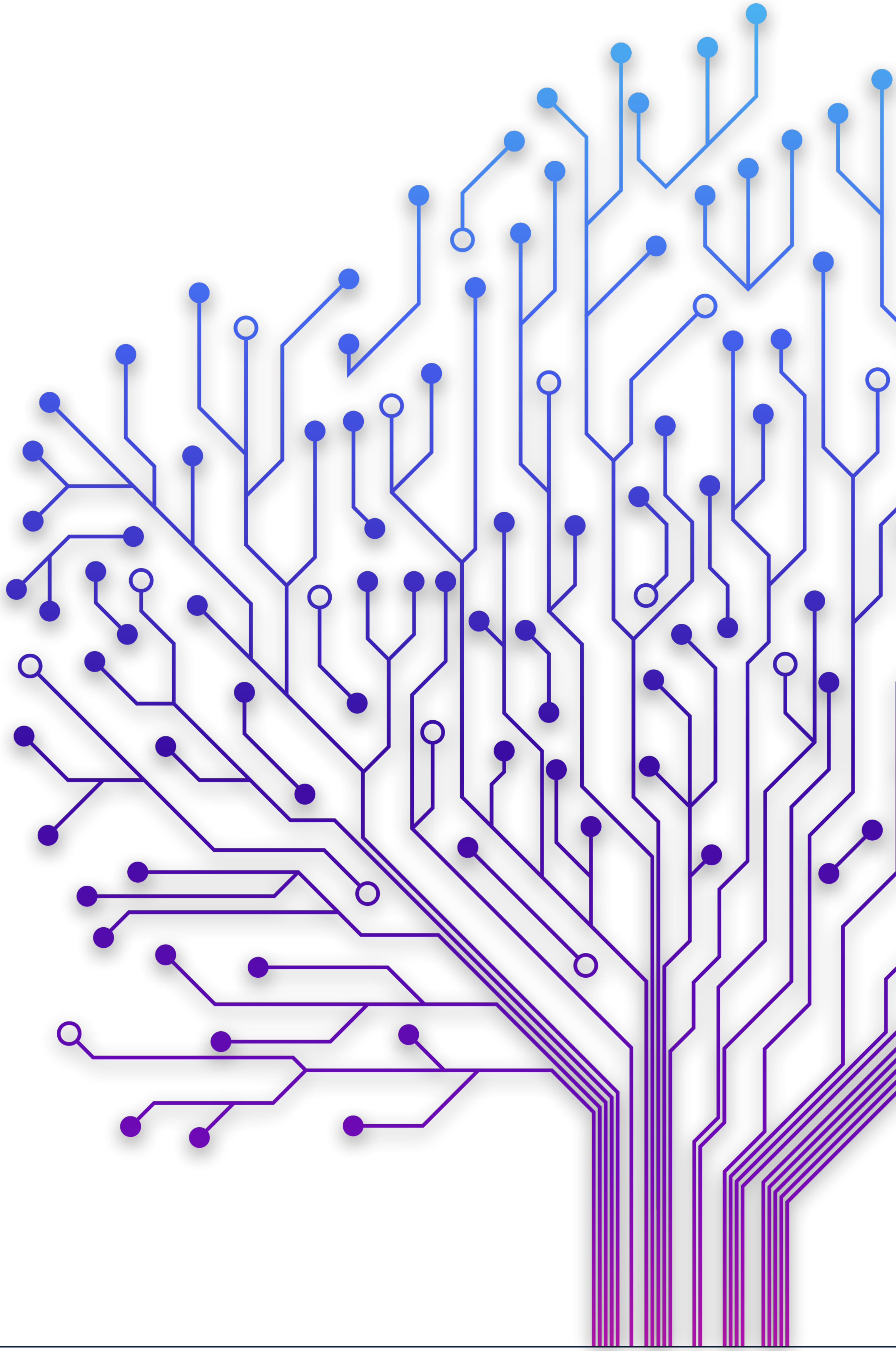
AND MAY AS WELL ANSWER

AFTER FEW MICROSECONDS OR MANY SECONDS?

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

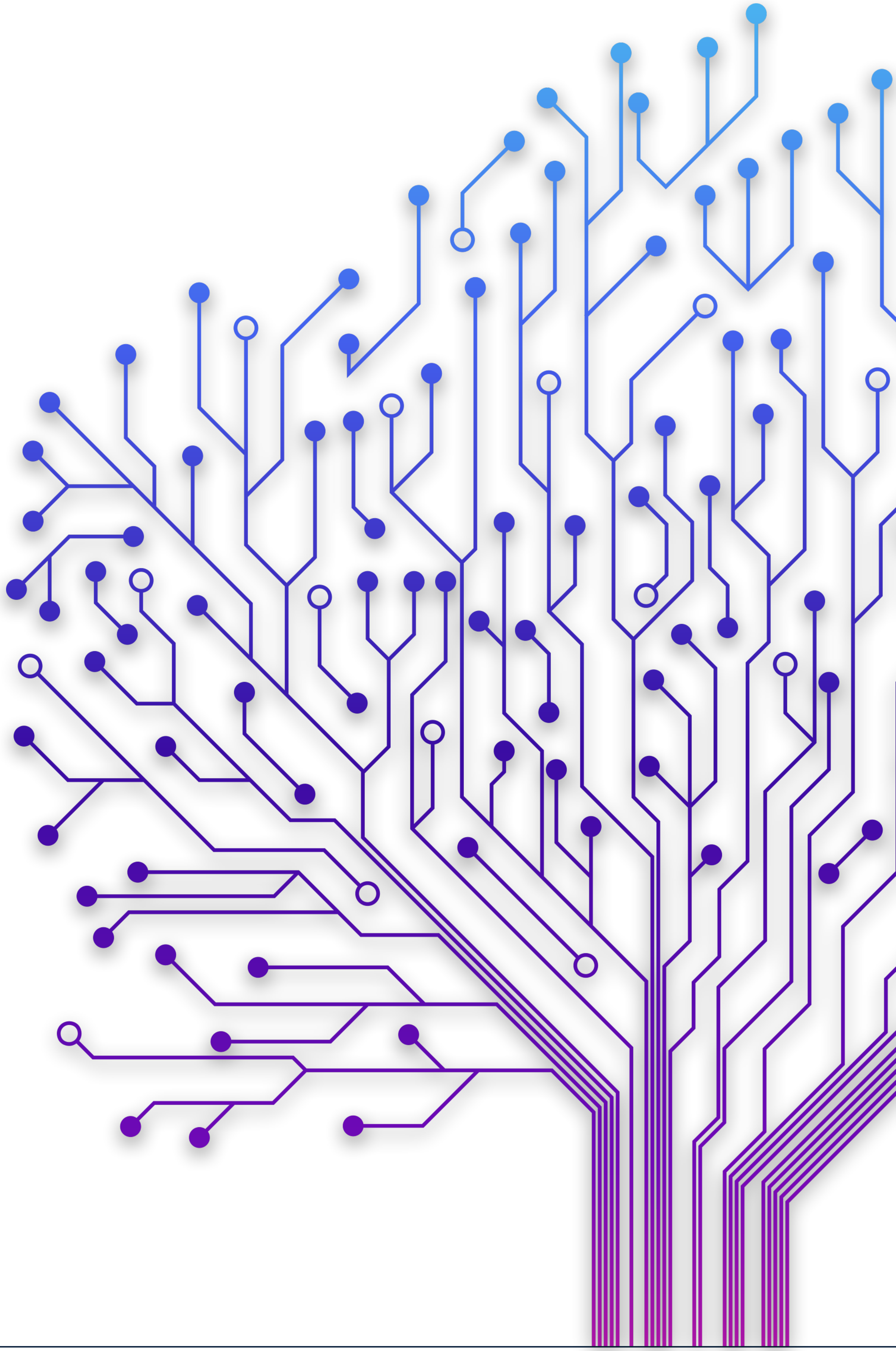
Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction

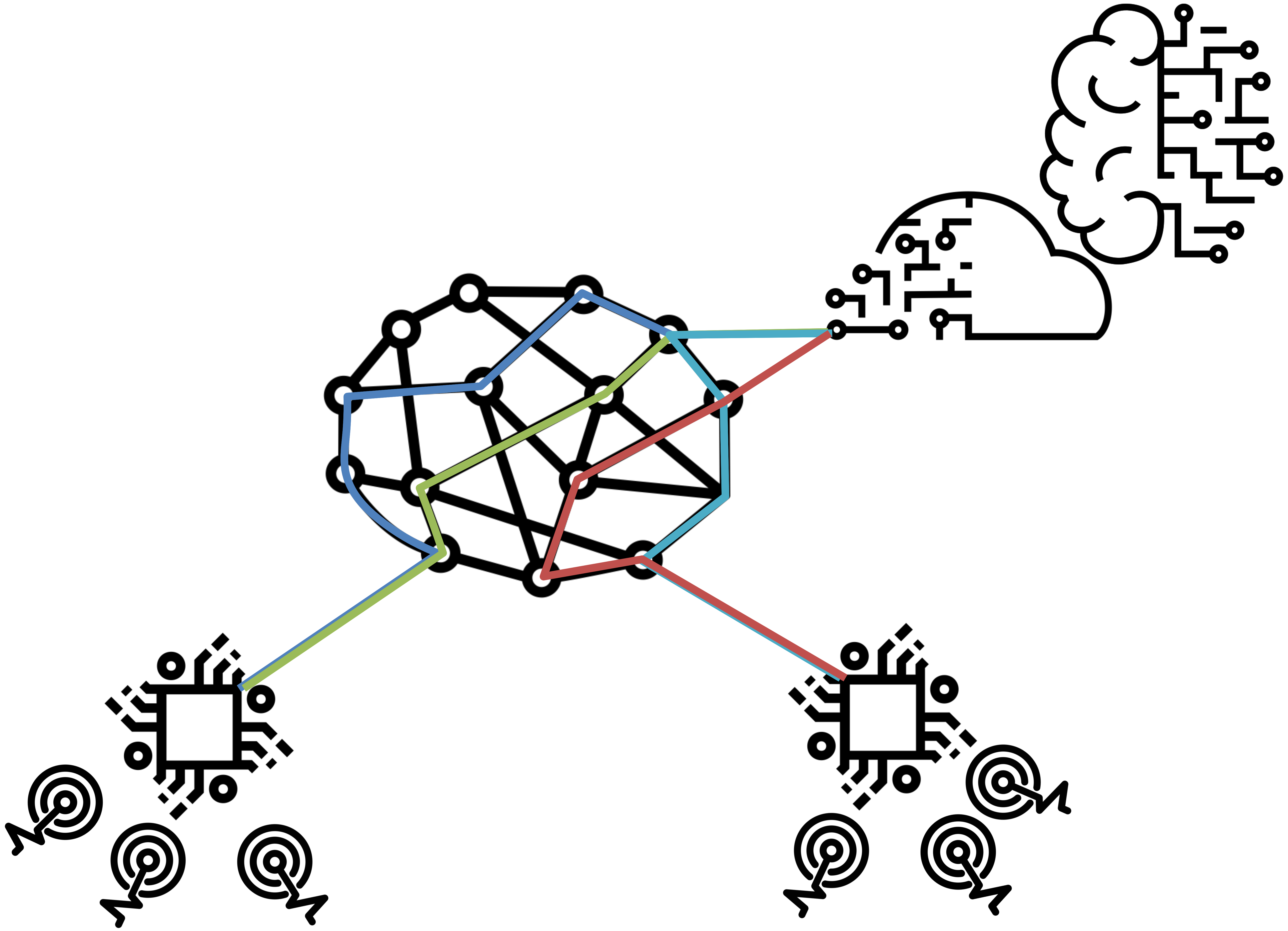


Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction



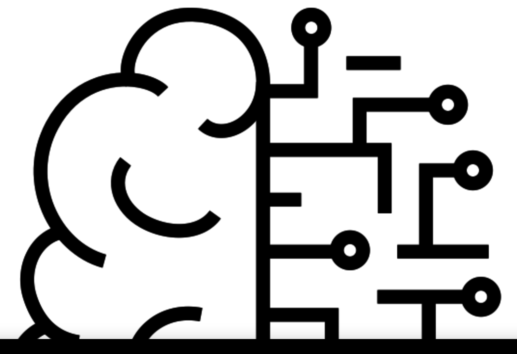
Artificial Intelligence of Things



Sensor by Lagot Design from NounProject.com
Iot by Souvik Bhattacharjee from NounProject.com
Network by ibrandify from NounProject.com
Brain by Andri Graphic from NounProject.com
Cloud by Andri Graphic from NounProject.com

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Artificial Intelligence of Things



hardly predictable, jittery, communication latency [1]
concerns about information privacy [2]
complex central management of heterogeneity [3]

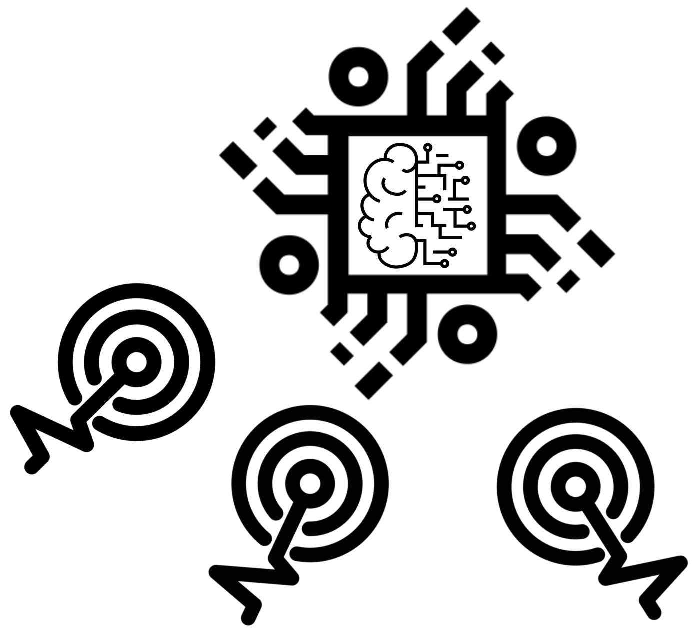
[1] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High- reliability and low-latency wireless communication for internet of things: Challenges, fundamentals, and enabling technologies," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7946–7970, 2019. [2] J. Daubert, A. Wiesmaier, and P. Kikiras, "A view on privacy amp; trust in iot," in 2015 IEEE International Conference on Communication Workshop (ICCW), 2015, pp. 2665–2670. [3] R. C. Motta, K. M. de Oliveira, and G. H. Travassos, "On challenges in engineering iot software systems," in Proceedings of the XXXII Brazilian Symposium on Software Engineering, ser. SBES '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 42–51. [Online]. Available: https://doi.org/10.1145/3266237.3266263

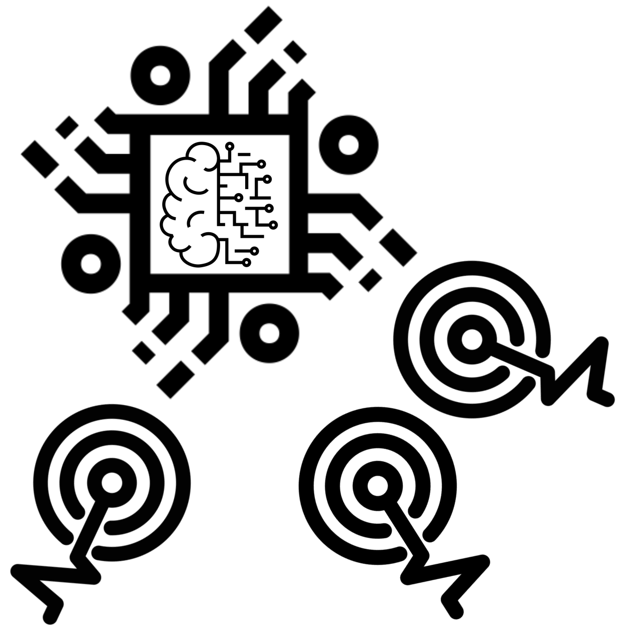
Artificial Intelligence of Things

Moving intelligence onboard
requires full usage of the few available resources

⇓

hardware co-processors

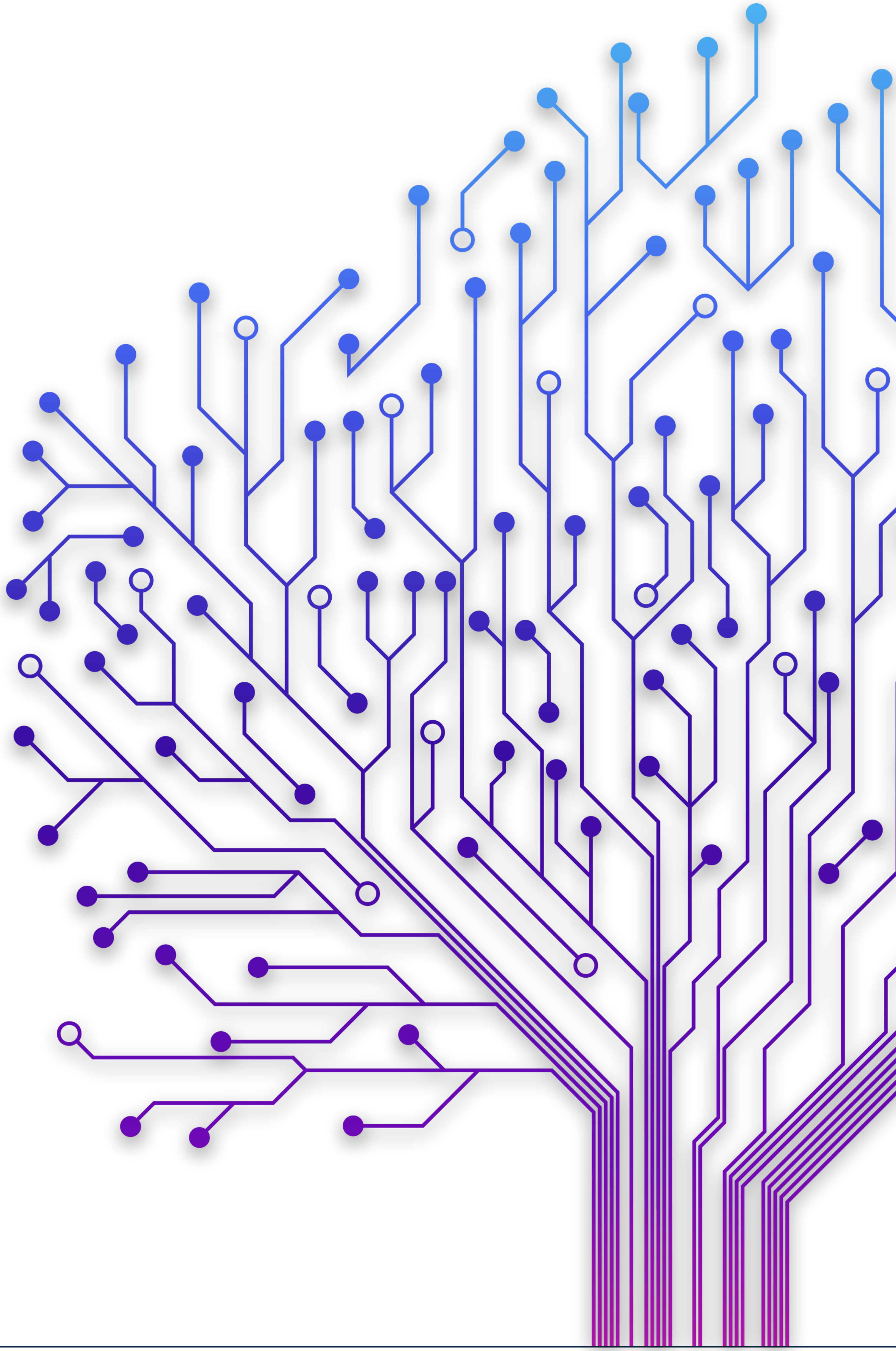




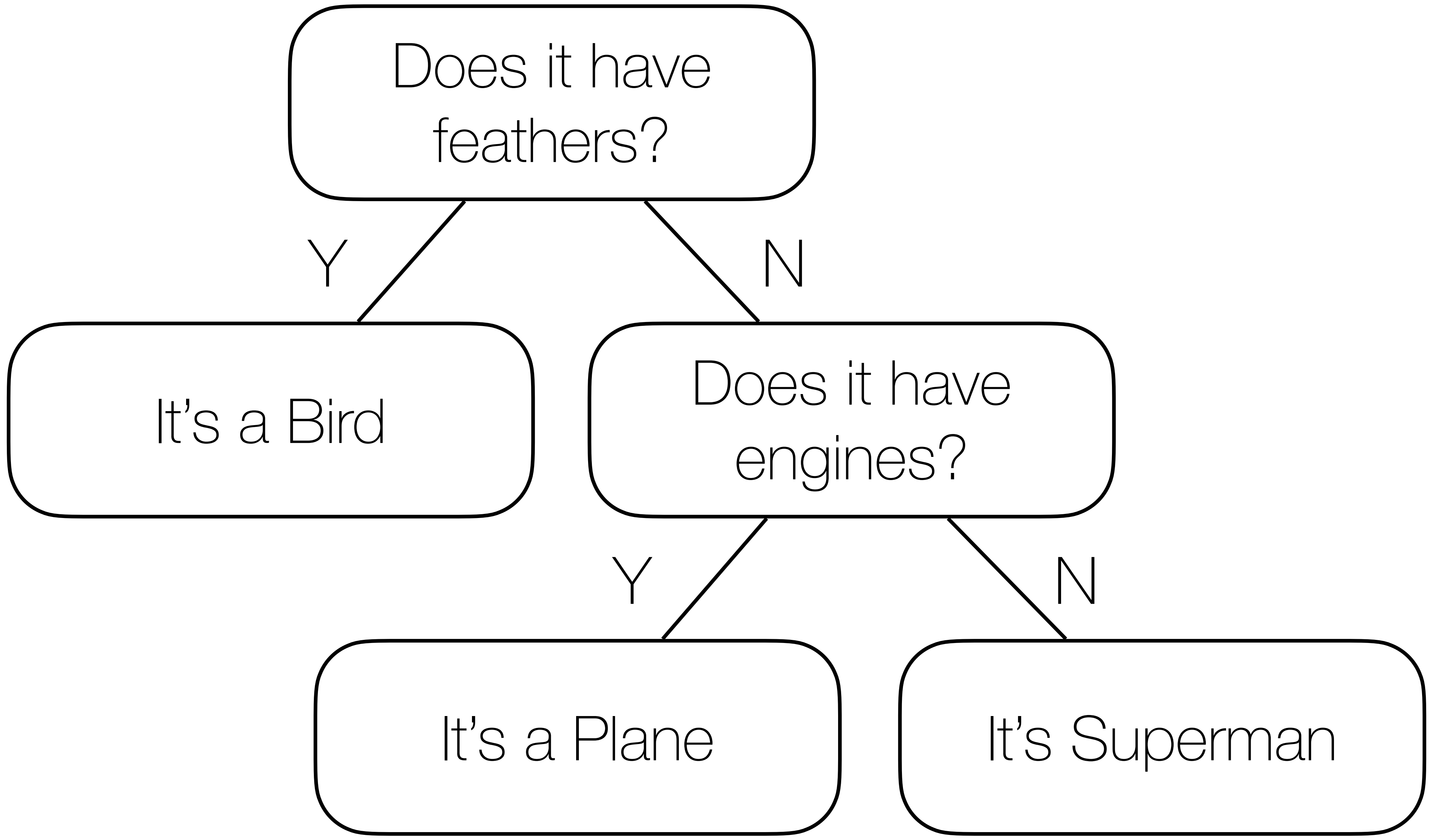
A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction



Decision Trees



Does it have feathers?

Y

N

It’s a Bird

Does it have engines?

Y

N

It’s a Plane

It’s Superman

“Tree ensembles are arguably among the most accurate ML models in use nowadays”

DTs do not require any “Post-hoc” analysis for obtain model transparency, which is fundamental for exploiting their explainability.

source [4]

[4] A. B. Arrietta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, 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.

co-processors for DT ensemble

Conifer[5]

Generation of static HLS/HDL code given scikit-learn DT ensemble

Image © Xilinx

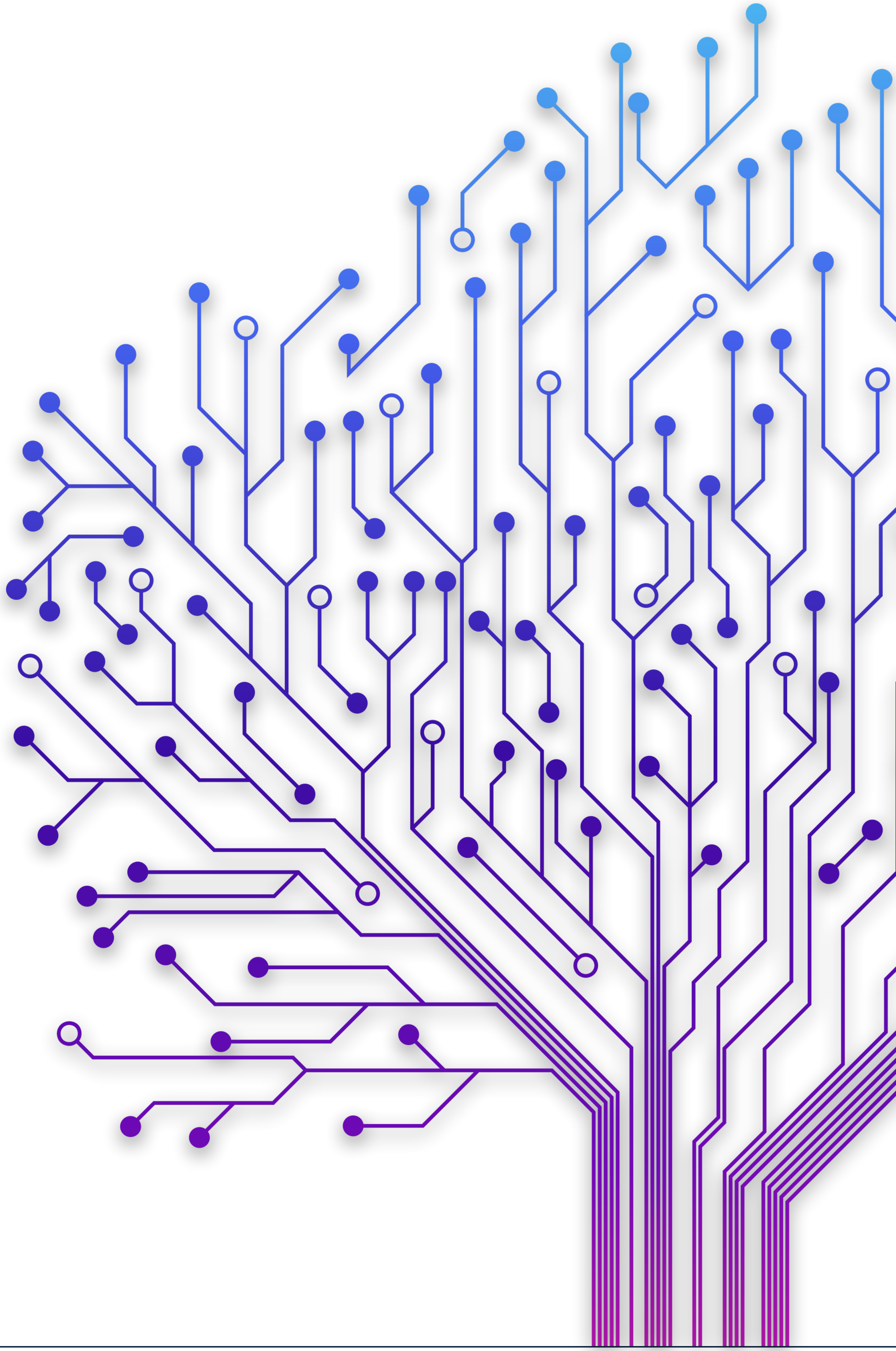
Used in particle physics
Thought for datacenter with FPGA

Extremely resource consuming
Rarely fits on embedded FPGAs

[5] S. Summers, G. Di Guglielmo, J. Duarte, P. Harris, D. Hoang, S. Jindariani, 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.

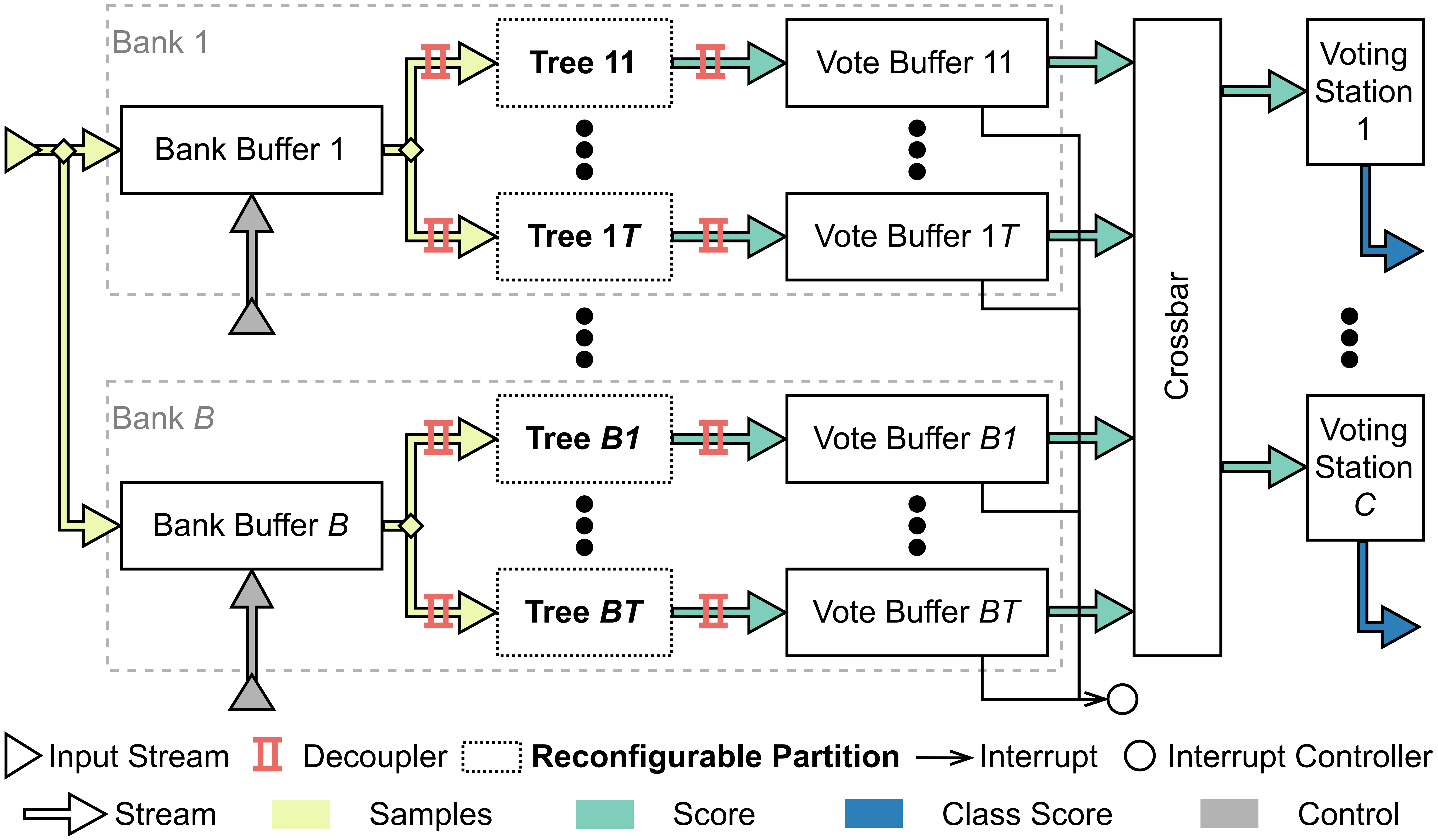
Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction



Entree

Template Architecture



Bank 1

Bank Buffer 1

Tree 11

Tree 1*T*

Vote Buffer 11

Vote Buffer 1*T*

Bank *B*

Bank Buffer *B*

Tree *B*1

Tree *BT*

Vote Buffer *B*1

Vote Buffer *BT*

Crossbar

Voting Station 1

Voting Station *C*

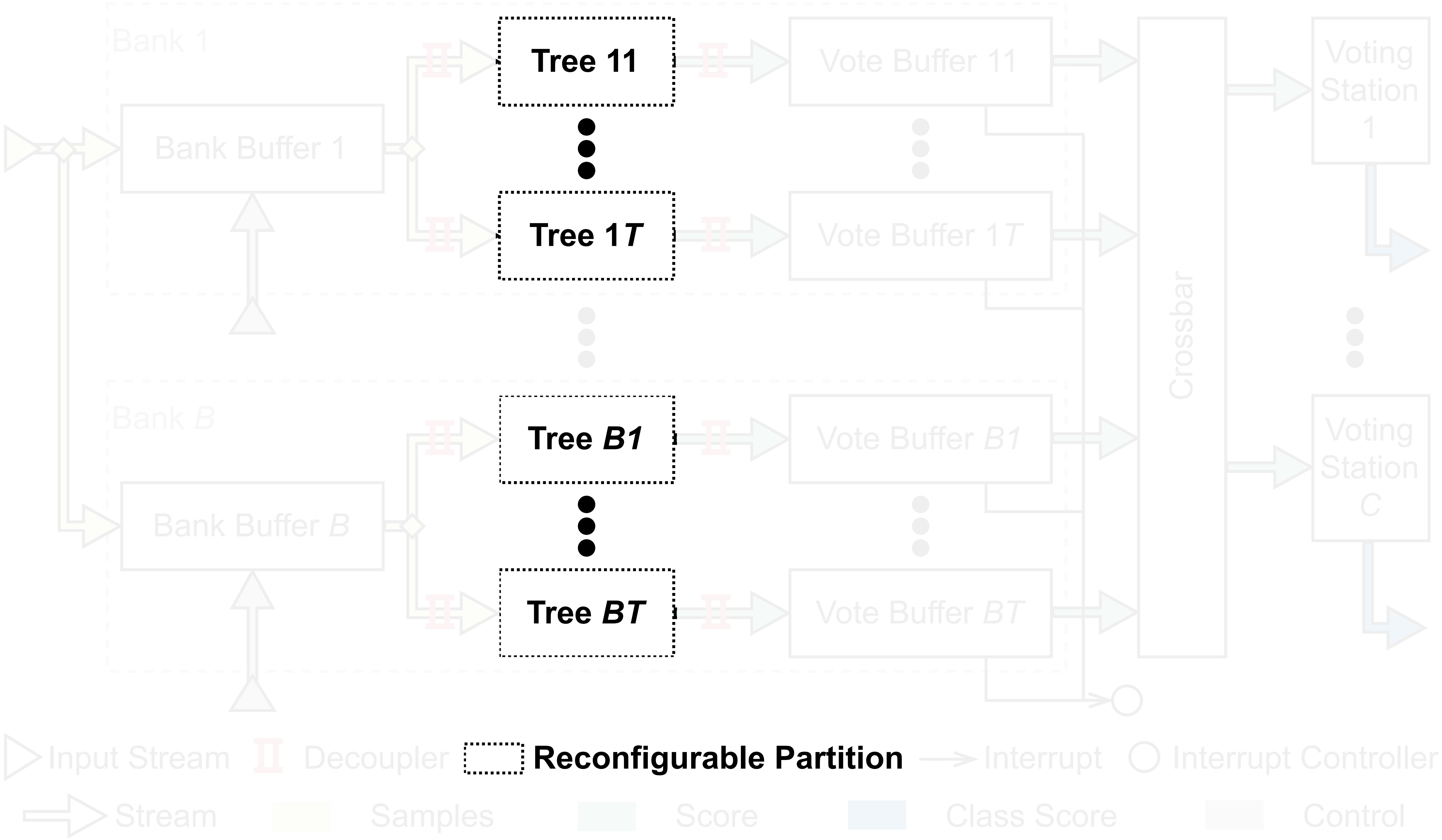
▷ Input Stream Ⅱ Decoupler ⬚ **Reconfigurable Partition** ⟶ Interrupt ○ Interrupt Controller

⇨ Stream ▨ Samples ▨ Score ▨ Class Score ▨ Control

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

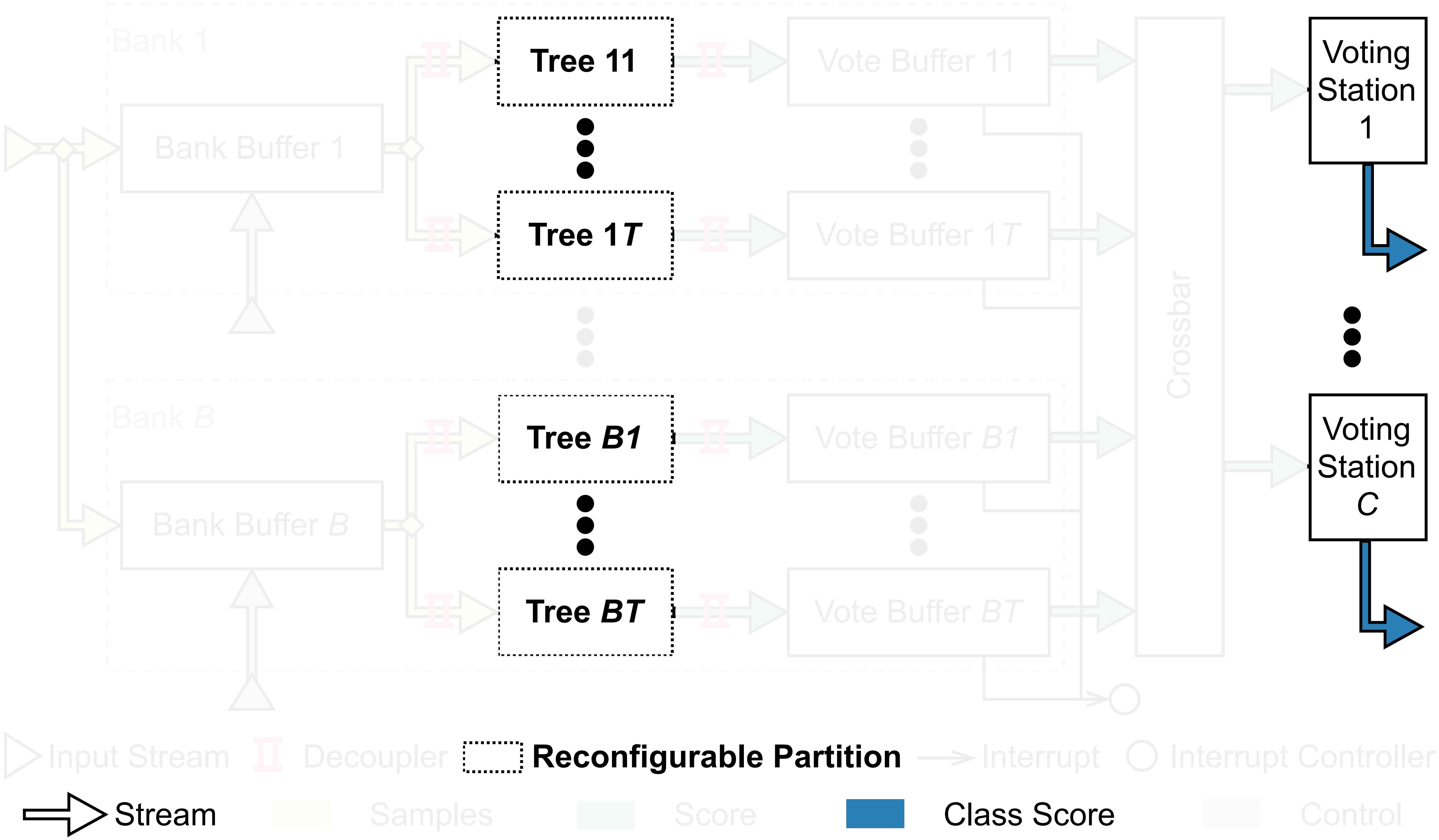
Template Architecture



A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

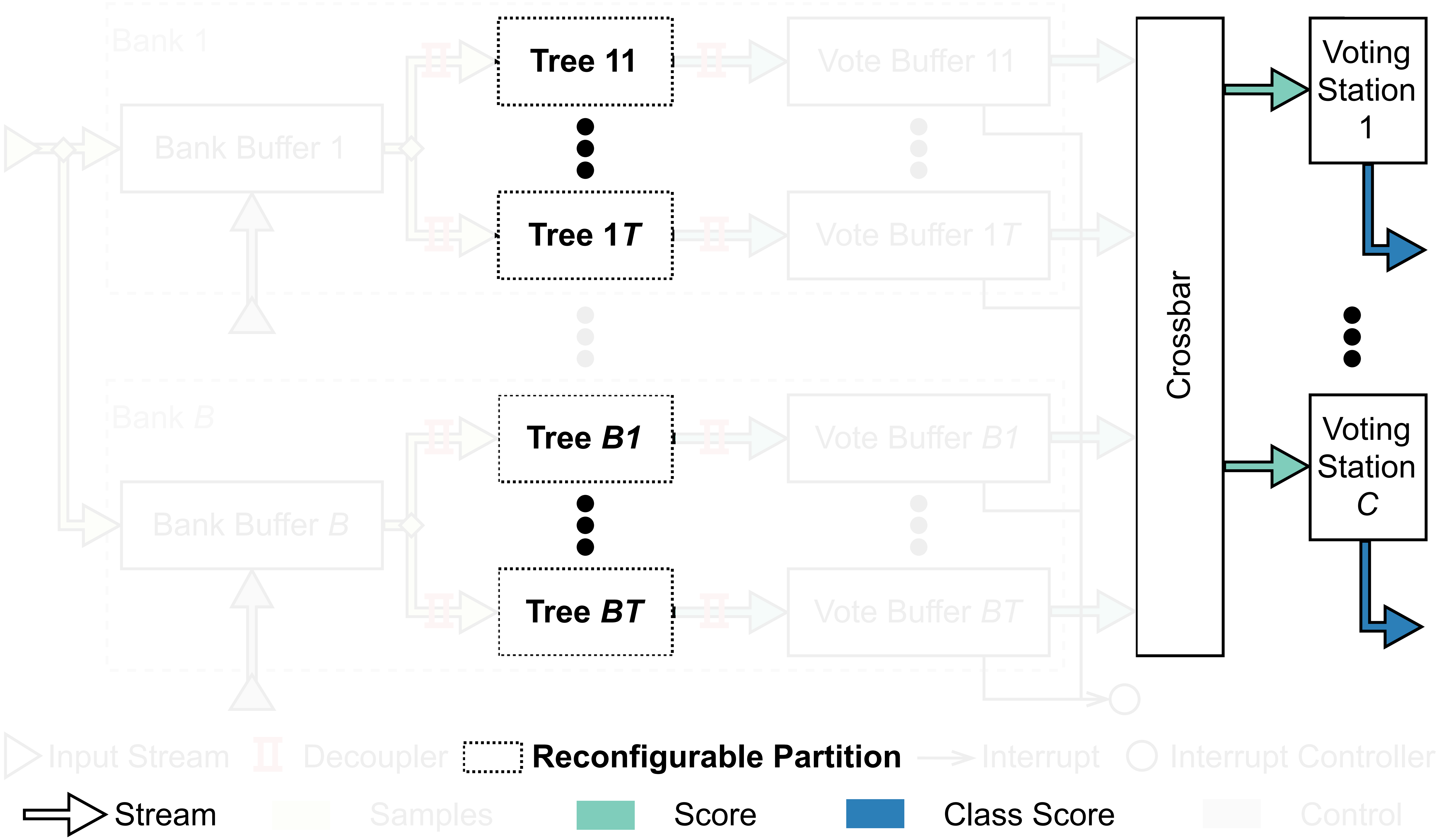
Template Architecture



A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

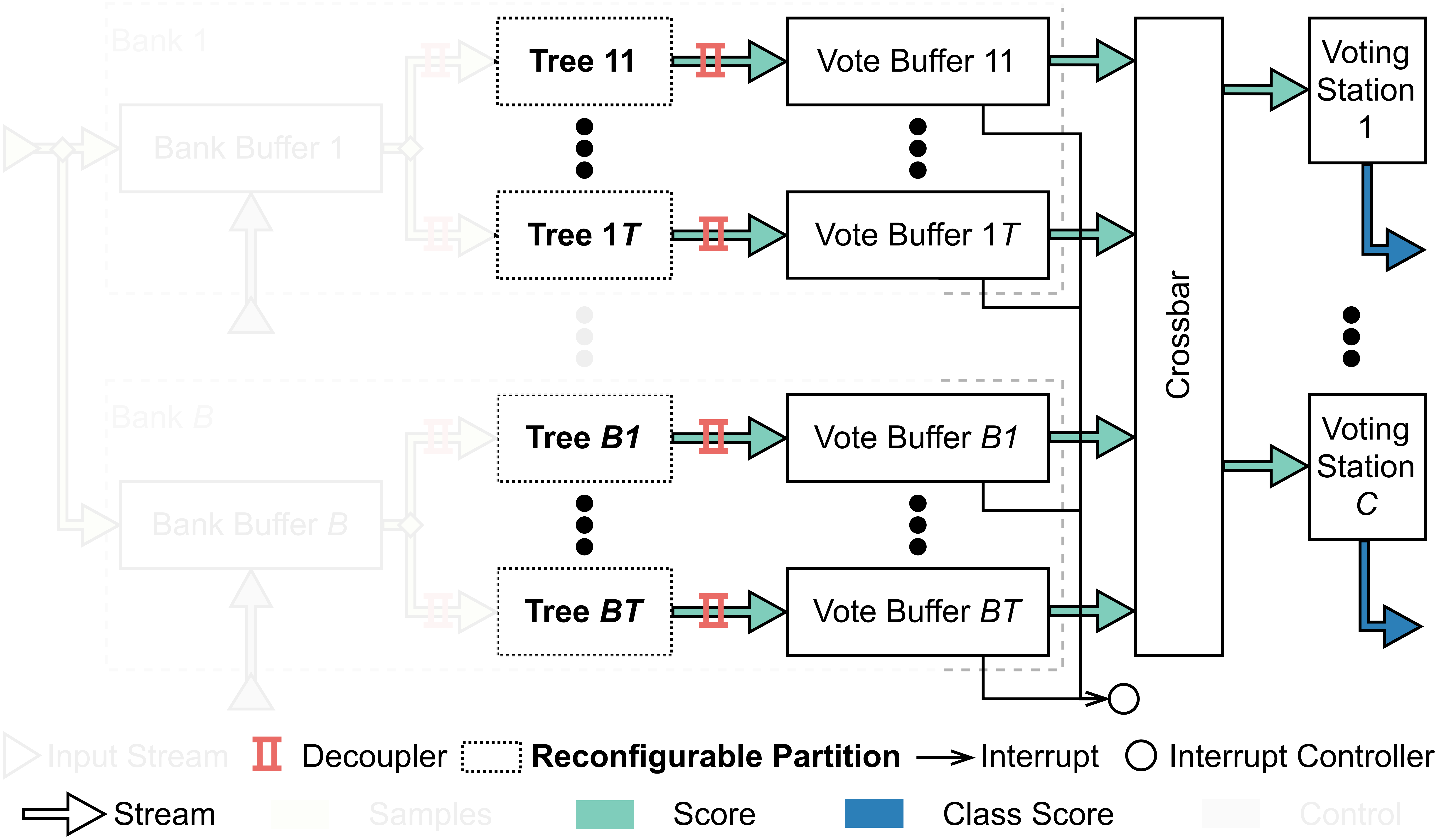
Template Architecture



A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

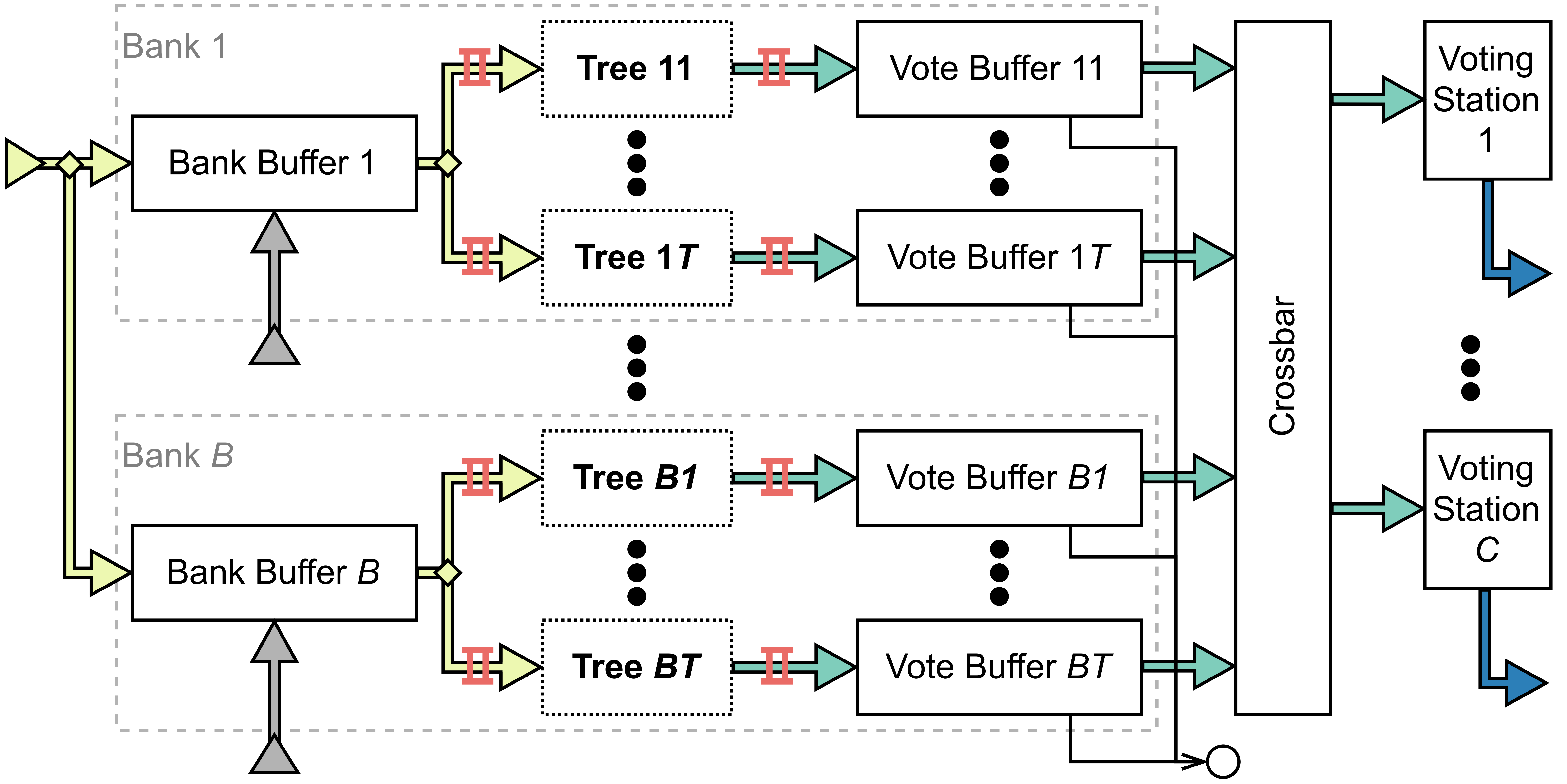
Template Architecture



A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

Template Architecture



Bank 1

Bank Buffer 1

Tree 11

Tree 1*T*

Vote Buffer 11

Vote Buffer 1*T*

Bank *B*

Bank Buffer *B*

Tree *B*1

Tree *BT*

Vote Buffer *B*1

Vote Buffer *BT*

Crossbar

Voting Station 1

Voting Station *C*

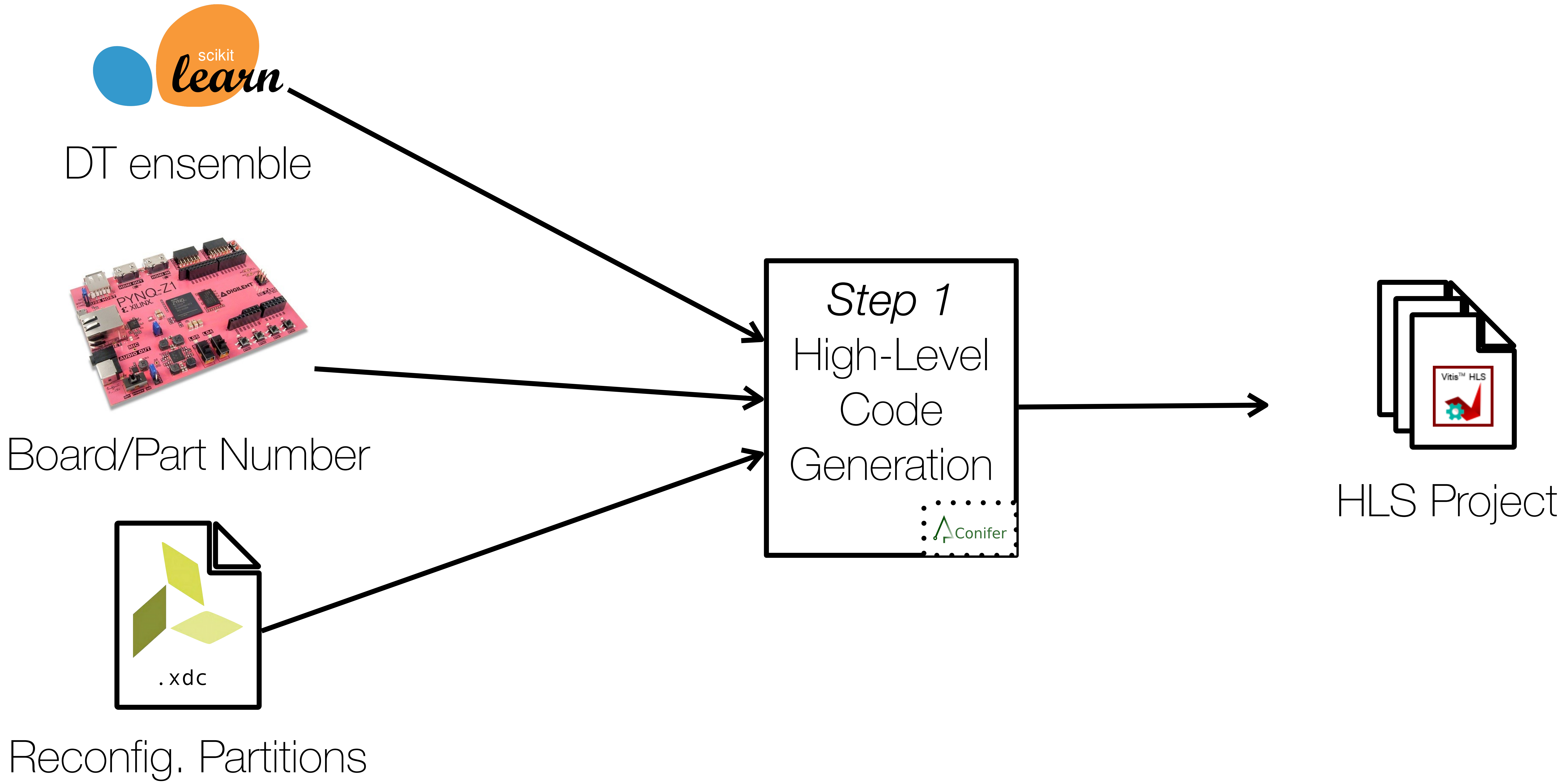
▷ Input Stream Ⅱ Decoupler ⬚ **Reconfigurable Partition** ⟶ Interrupt ○ Interrupt Controller

⇨ Stream 🟨 Samples 🟩 Score 🟦 Class Score ⬜ Control

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
"Large Forests and Where to "Partially" Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

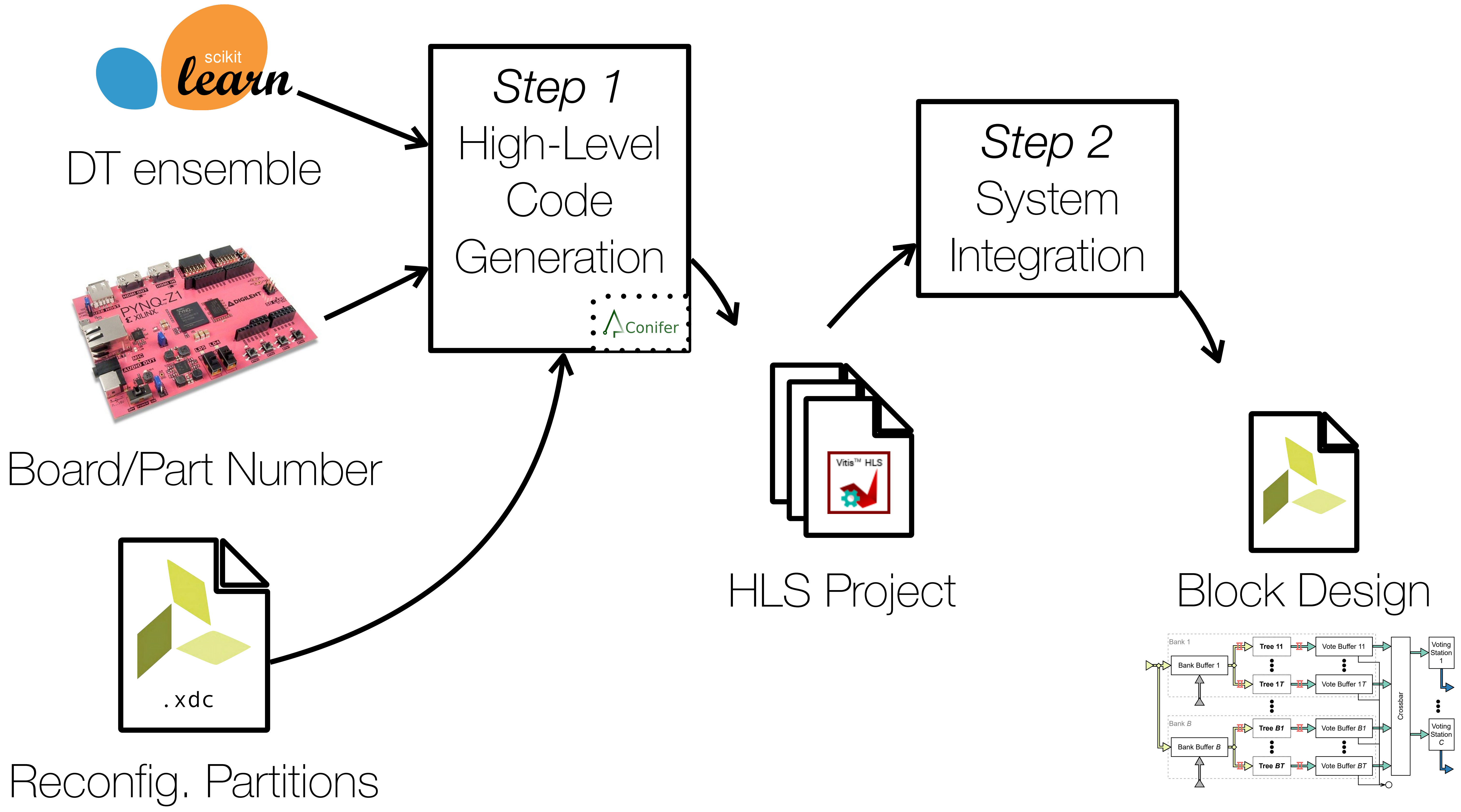
Automatic Design Flow



A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

Automatic Design Flow



DT ensemble

Board/Part Number

Reconfig. Partitions

Step 1
High-Level Code Generation

Step 2
System Integration

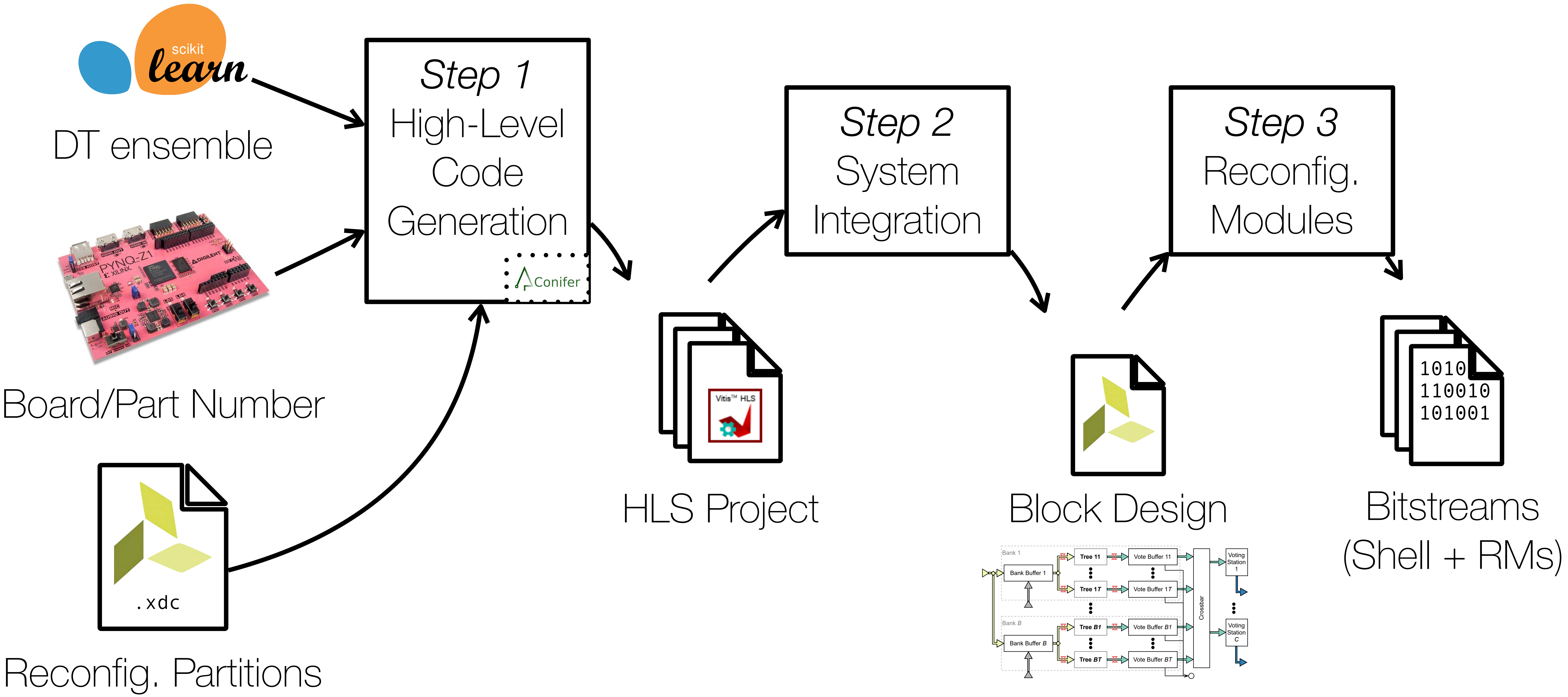
HLS Project

Block Design

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Entree

Automatic Design Flow



DT ensemble

Board/Part Number

Reconfig. Partitions

Step 1 High-Level Code Generation

Step 2 System Integration

Step 3 Reconfig. Modules

HLS Project

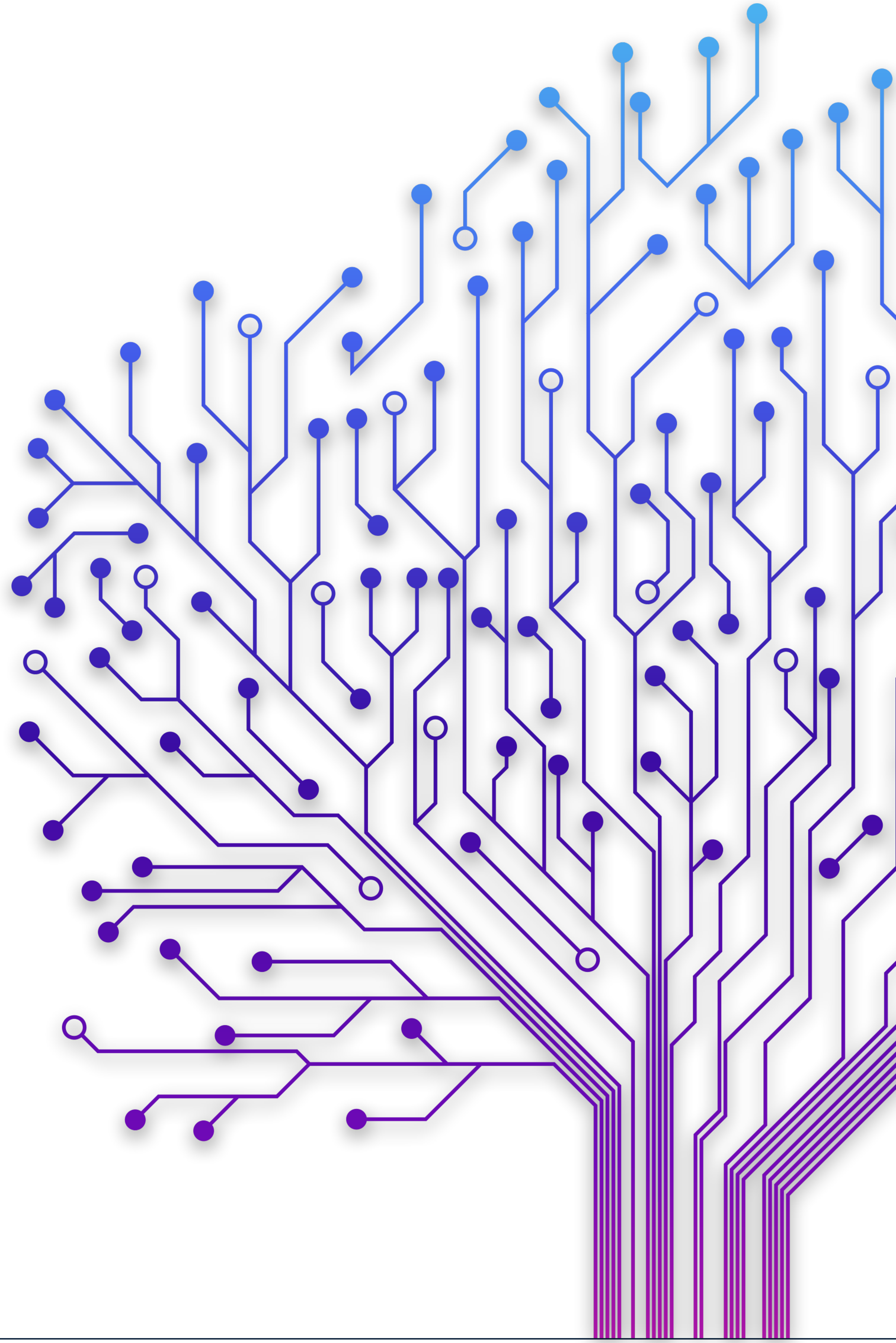
Block Design

Bitstreams (Shell + RMs)

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction



ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Mean Relative Latency Jitter

Measuring latency stability

Example

Which is better to have in a system to design?

— Edge-to-Cloud —

absolute latency: 39.55±9.96ms

— On-site computation —

absolute latency: 3.95±3.82ms

Mean Relative Latency Jitter

Measuring latency stability

$$J = \frac{\sqrt{\frac{\sum_{s'=1}^{S}\left(t_{s'} - \frac{\sum_{s=1}^{S} t_s}{S}\right)^2}{S}}}{\frac{\sum_{s=1}^{S} t_s}{S}}$$

S: number of samples

t_s: computation time of the s-th sample

Example
Which is better to have in a system to design?

— Edge-to-Cloud —
absolute latency: 39.55±9.96ms
J: 0.25

— On-site computation —
absolute latency: 3.95±3.82ms
J: 0.97

ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Mean Relative Latency Jitter

The Experiment

Classification

Optical Recognition of Handwritten Digits Data Set

8x8px grayscale images

=

64 floating point features

identify the digit

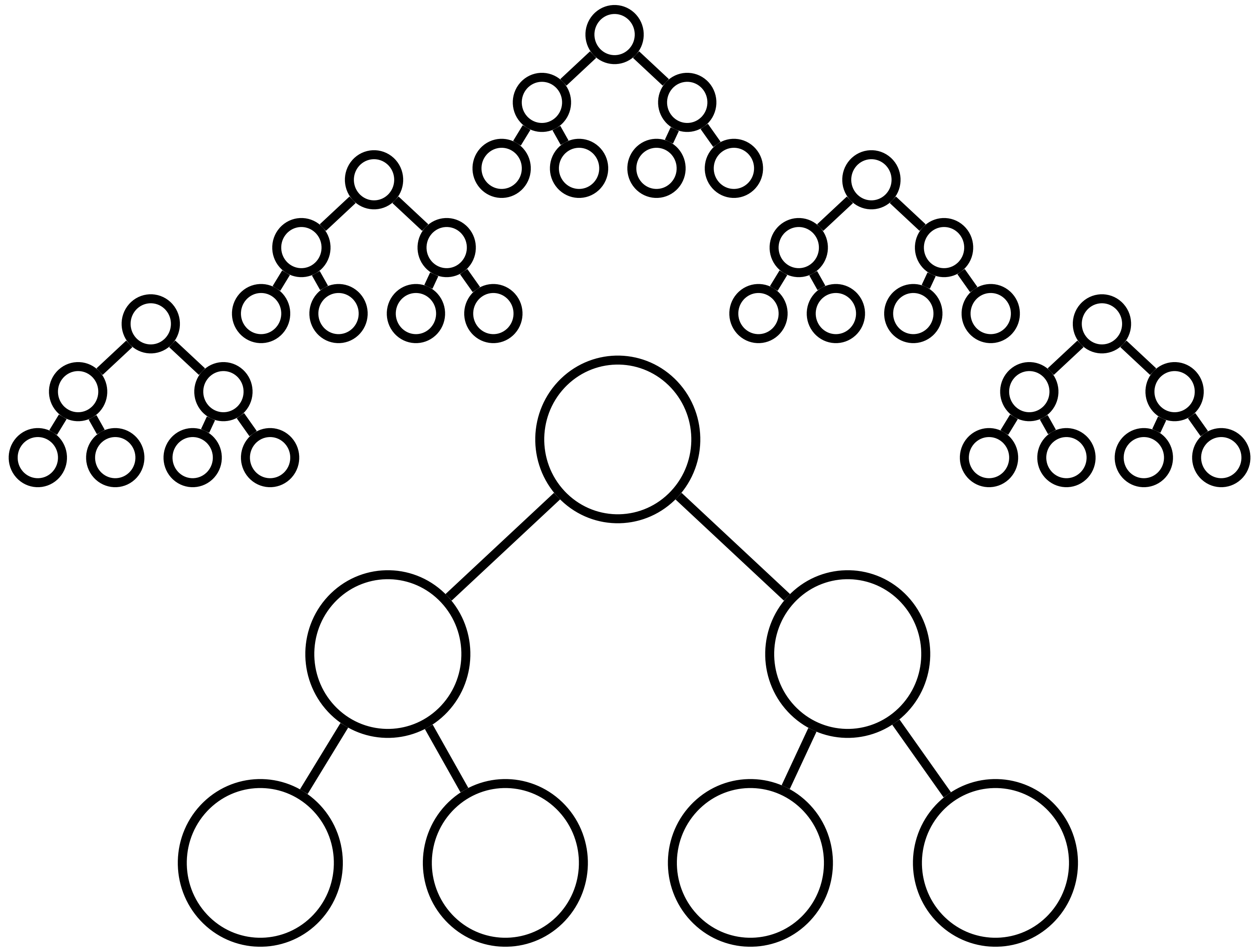
=

10 classes

A. Damiani, E. Del Sozzo and M. D. Santambrogio,
“Large Forests and Where to “Partially” Fit Them," 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), 2022.

Mean Relative Latency Jitter

The Experiment



Gradient Boosting
With different
number of estimators and *maximum depth*

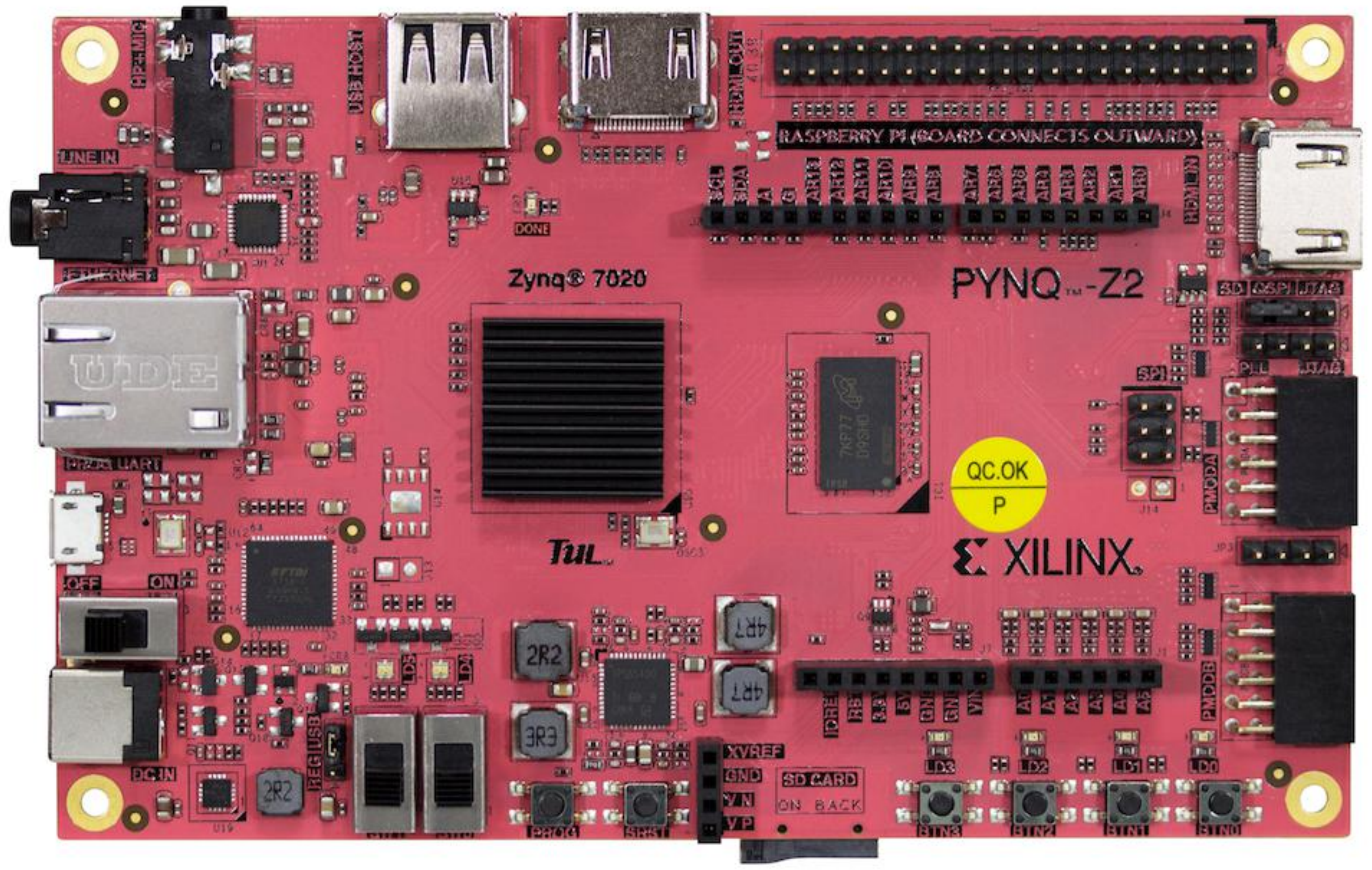
Classification
Optical Recognition of Handwritten Digits
Data Set

8x8px grayscale images
=
64 floating point features

identify the digit
=
10 classes

Mean Relative Latency Jitter

Experimental Setup



Tul PYNQ-Z2 development board

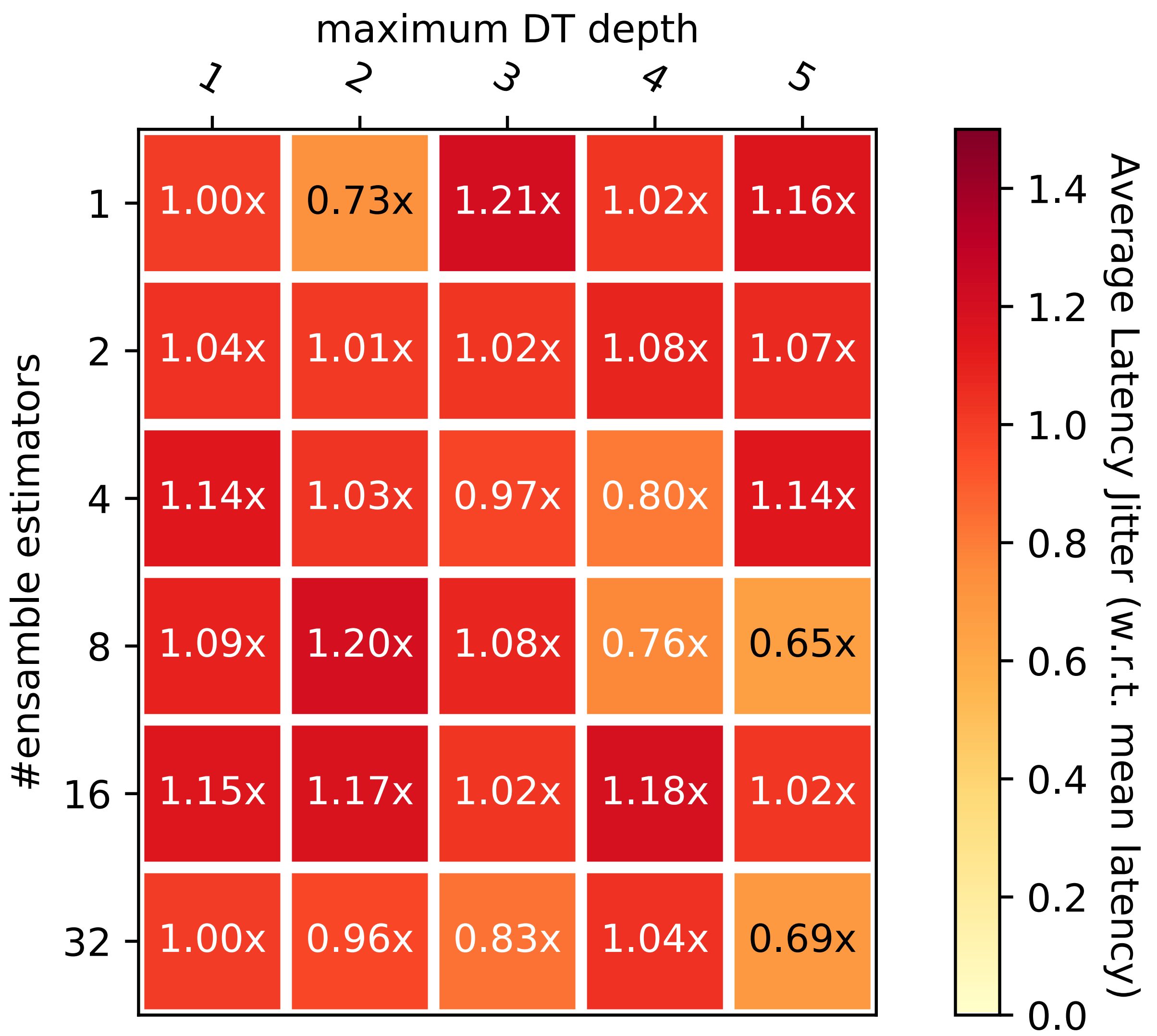
Mounting a Xilinx Zynq-7000
xc7z020clg400-1

Dual-Core Cortex-A9 @ 866MHz

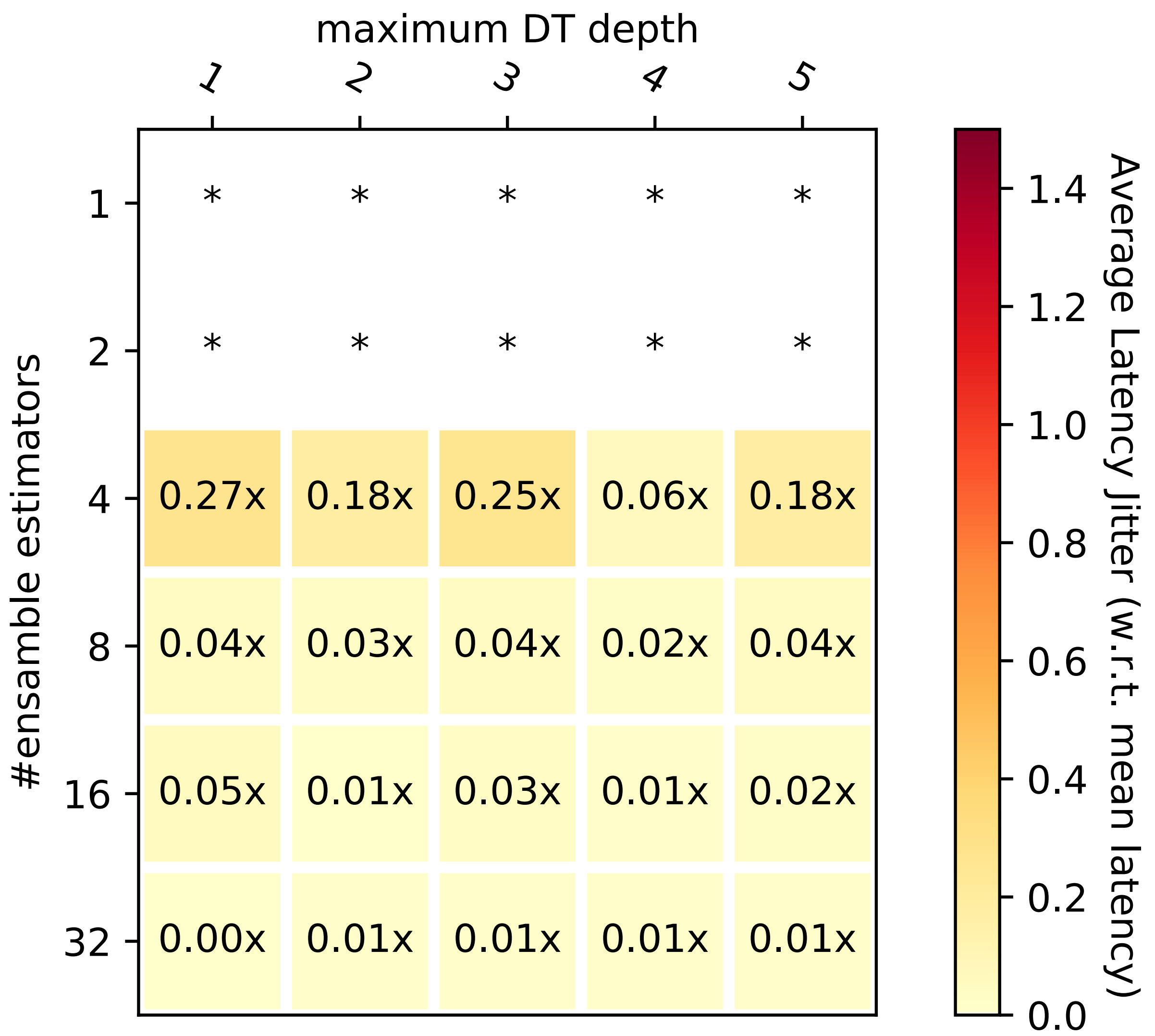
Artix-7 (equivalent) FPGA
~53k LUTs ~106k FFs

Mean Relative Latency Jitter

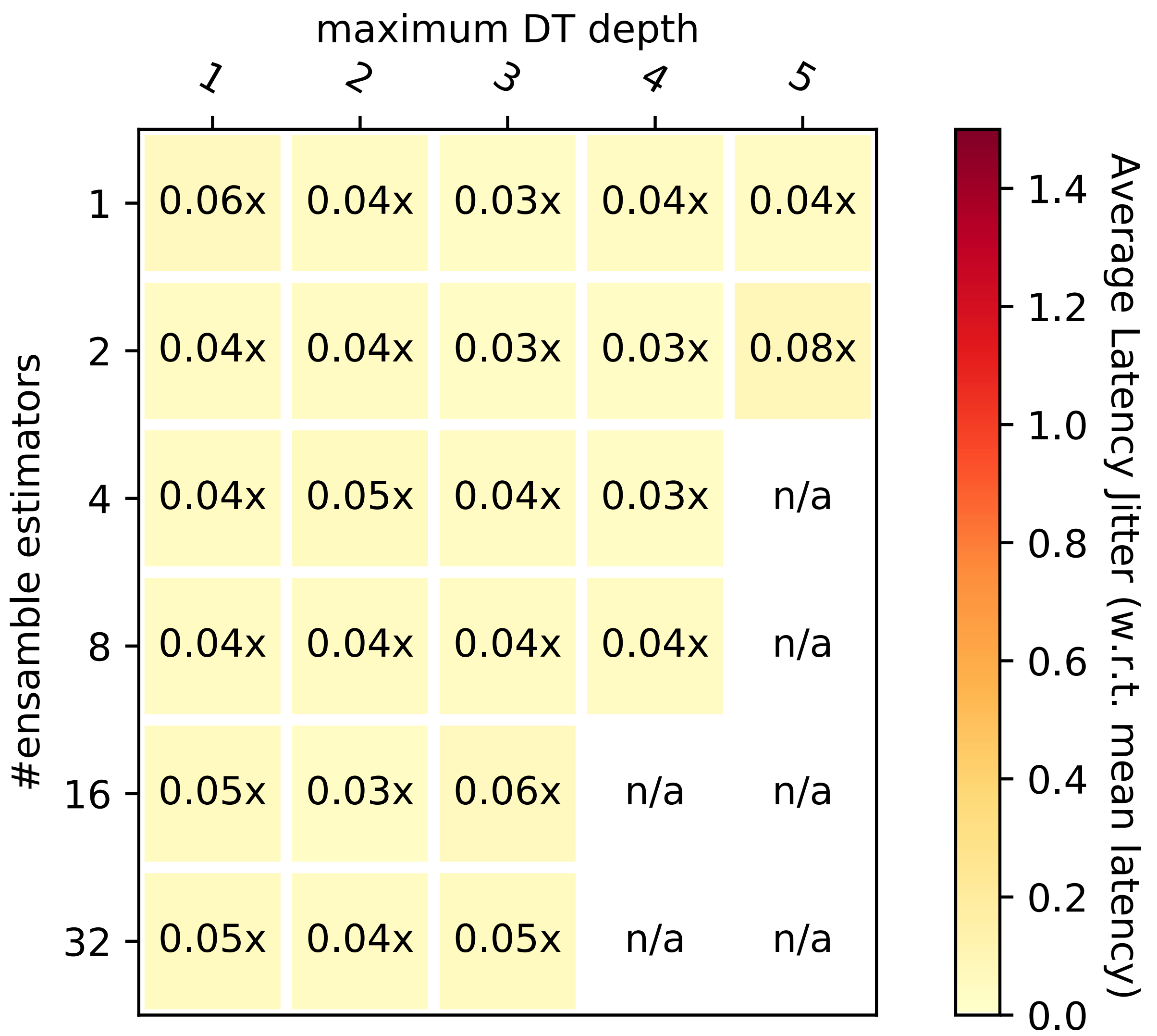
Results



scikit learn software execution



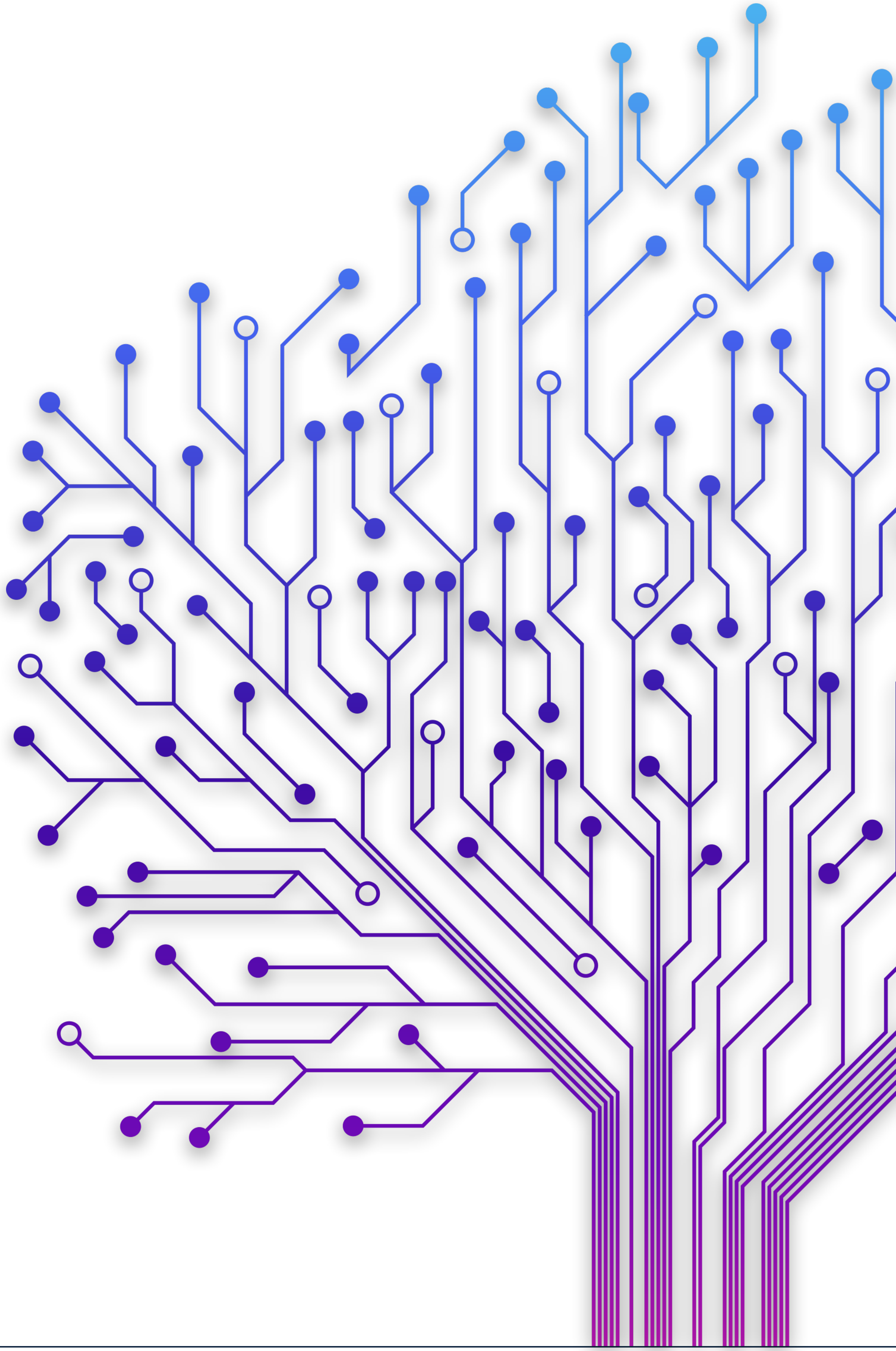
Entree



static Conifer co-processor

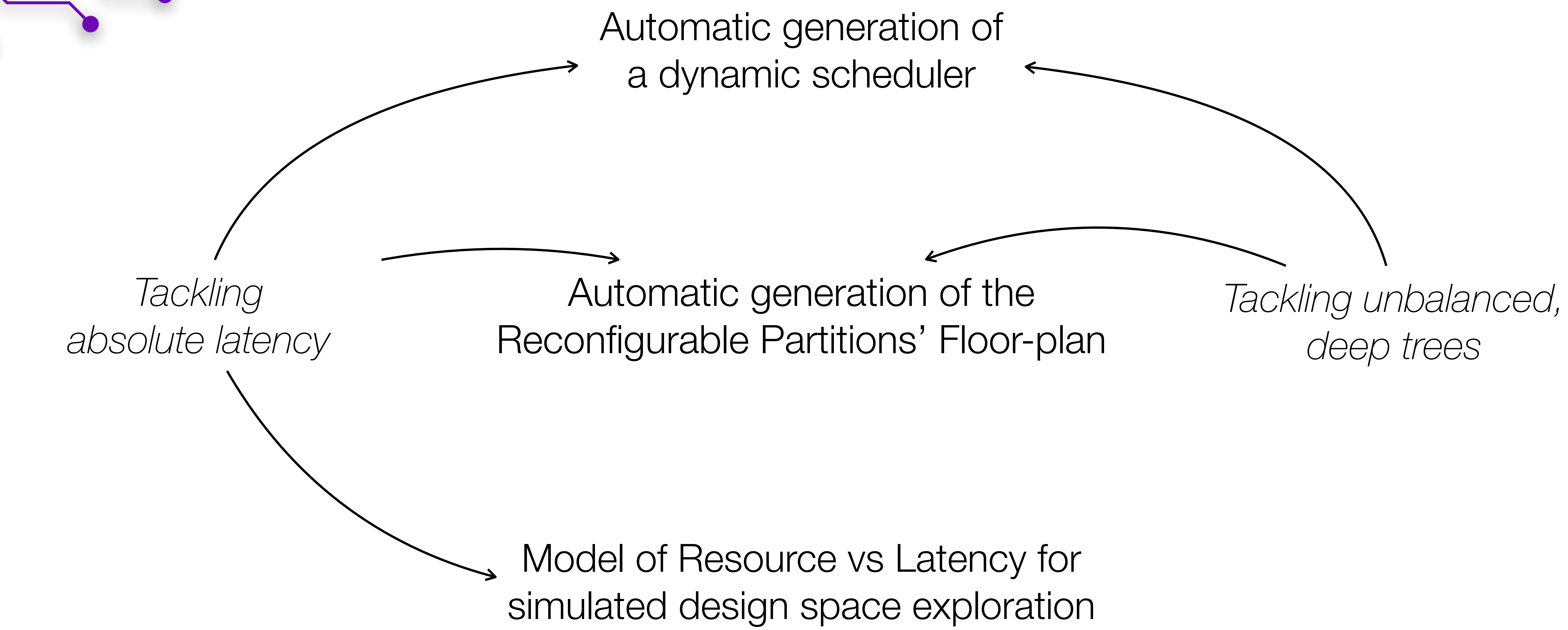
Outline

- Context definition: AIoT
- Decision Tree ensembles
- Entree: automatic design flow
- Experiments on latency jitter
- Future direction



J … is not enough

Future direction



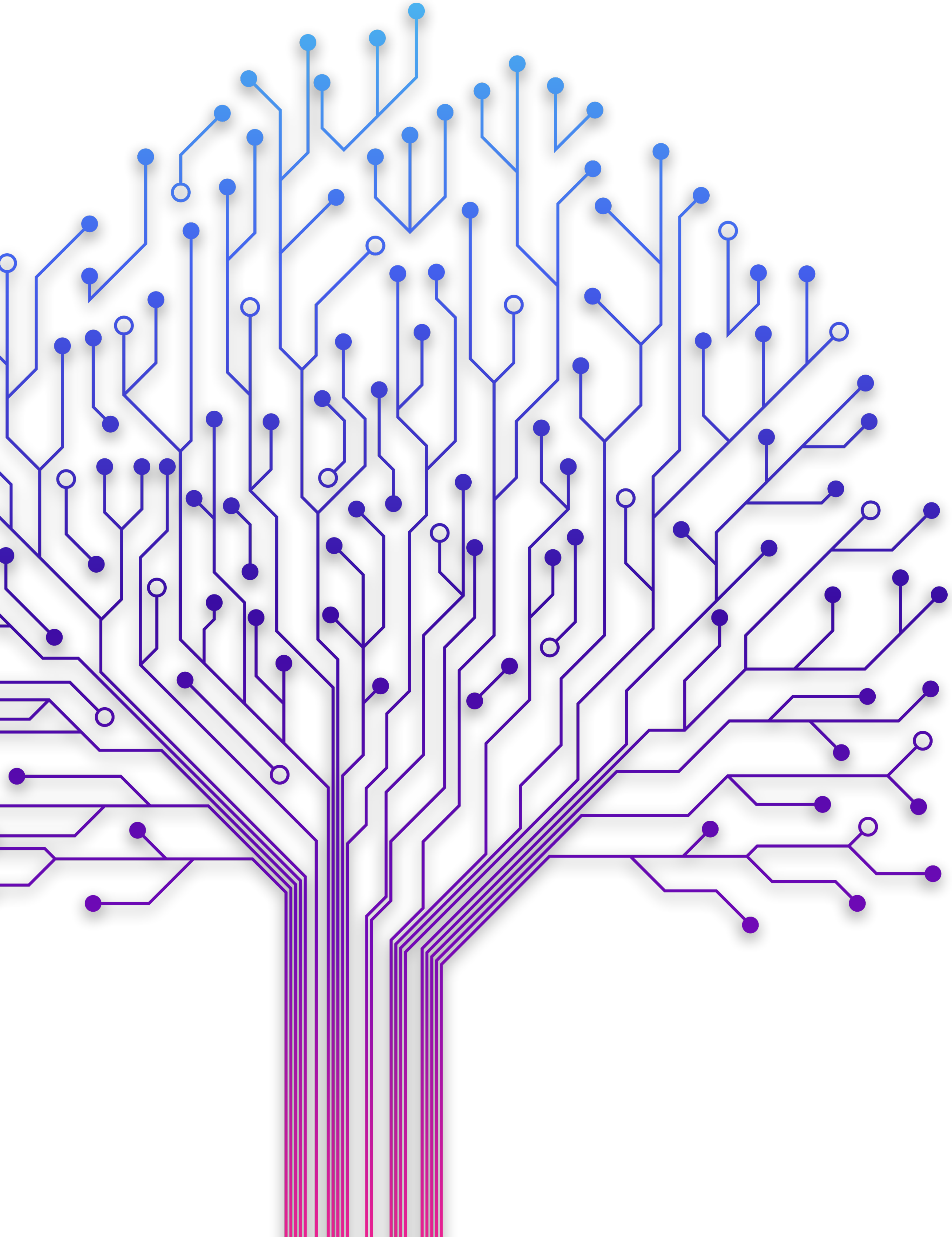
Automatic generation of a dynamic scheduler

Tackling absolute latency

Automatic generation of the Reconfigurable Partitions' Floor-plan

Tackling unbalanced, deep trees

Model of Resource vs Latency for simulated design space exploration



Large Forests and Where to “Partially” Fit Them

https://github.com/necst/entree_aspdac22

Thank you for your interest and attention

Andrea Damiani - andrea.damiani@polimi.it
Emanuele Del Sozzo - emanuele.delsozzo@polimi.it
Marco D. Santambrogio - marco.santambrogio@polimi.it

January, 2022 - ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

POLITECNICO MILANO 1863

POLITECNICO MILANO 1863 NECST laboratory