IA Embarquée :

Enjeux et Technologies

Thierry Collette
2 Juillet 2024
PFIA, La Rochelle

www.thalesgroup.com

Artificial intelligence: ability to process information by artificial means

Inputs

Data

Information

Knowledge

AI Algorithm

Data-driven AI

"The world's most valuable resource is no longer oil, but **data.**" – The Economist (2017)

Knowledge-based AI

"A good decision is based on **knowledge**, not on numbers" – Plato

Hybrid AI

"A good decision is based **on knowledge and on data**"

Outputs

Cognitive Capacities

- **Perception / Recognition / Identification**
- **Understanding / Abstraction**
- **Reasoning / Decision**
- **Planning / Anticipation**

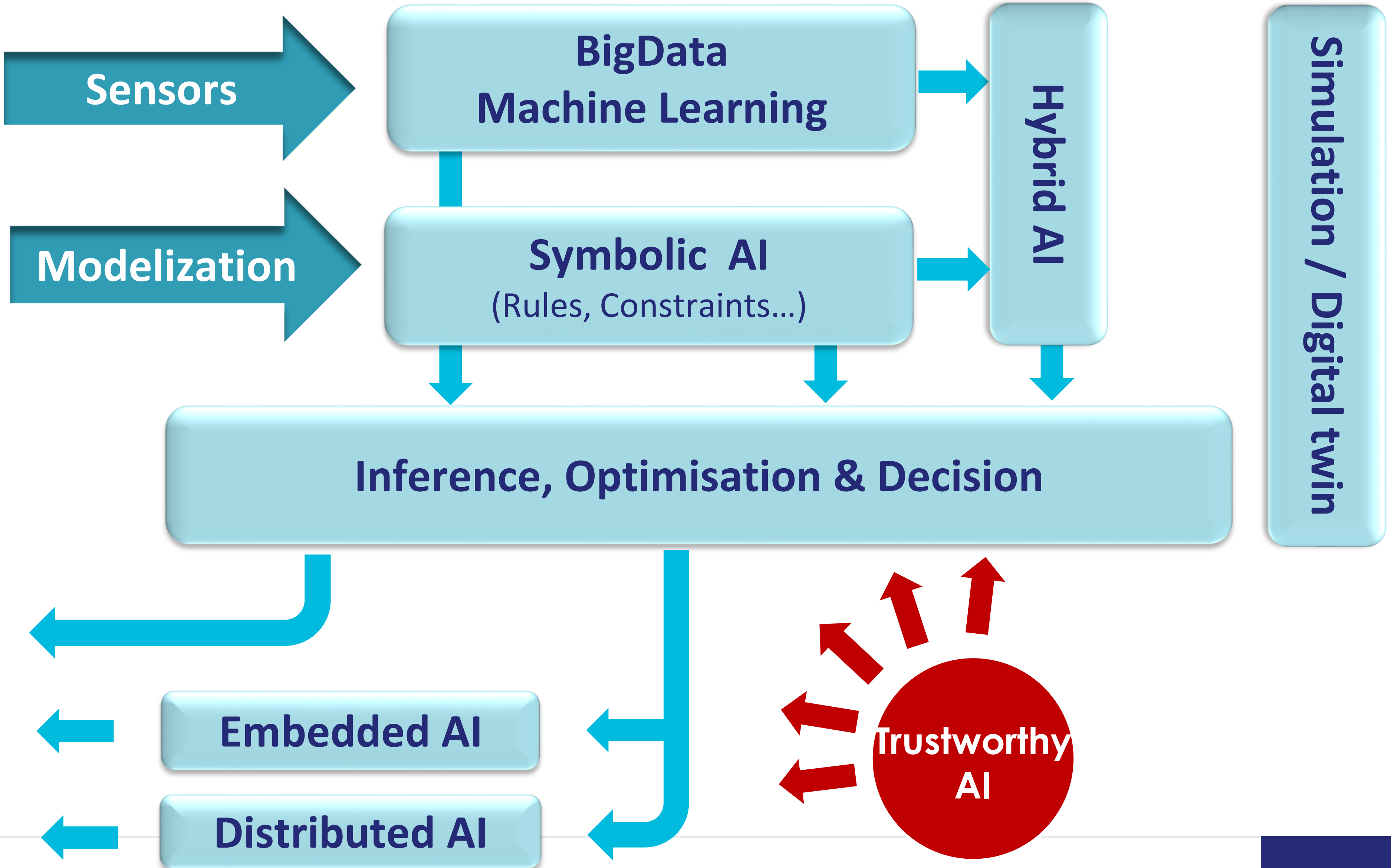
AI Research Activities

Maritime defense

Collaborative combat

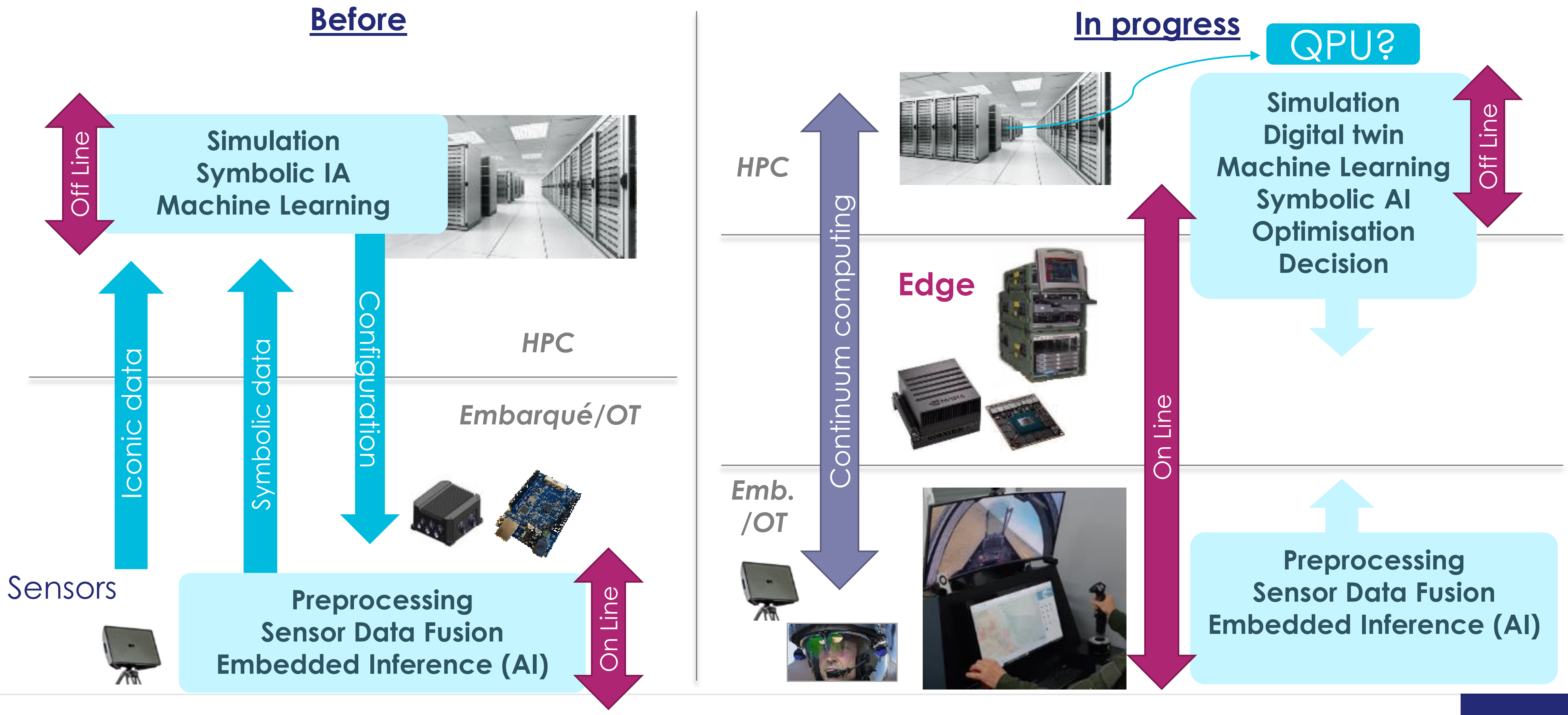
UxV

Aerian Traffic Control

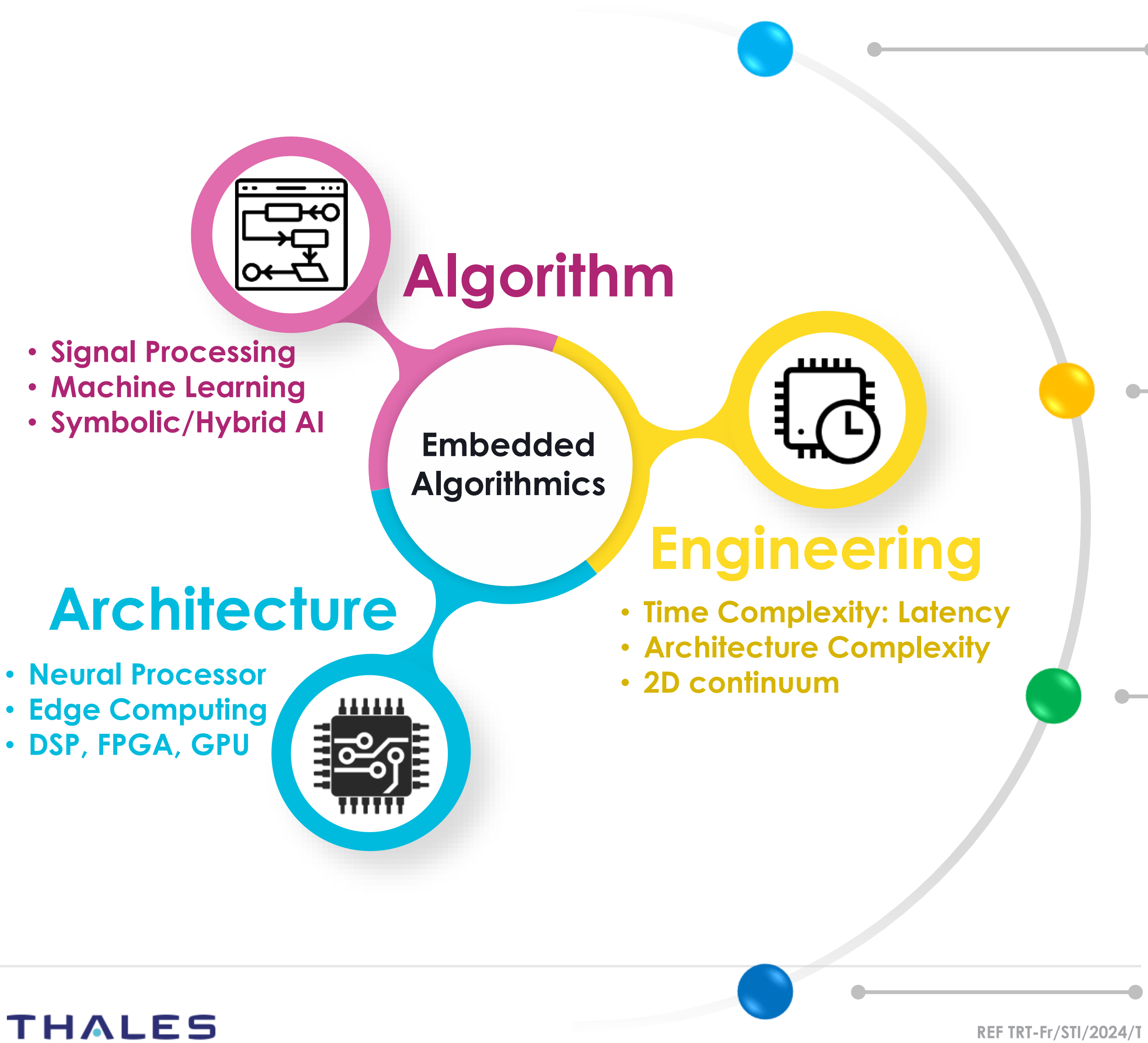


THALES
Building a future we can all trust

Towards the IT – OT convergence



Embedded AI



Algorithm paradigm vs. Algorithm complexity
Which paradigm is the most suitable with respect to technical constraints (Real Time, Space memory, …) and functional performances ?

Embedded AI
Tradeoff between functional and technical performances (e.g. accuracy vs. size)

Green AI
Reducing the carbon footprint of Thales AI based systems

Sovereignty (toward dis-ITAR-ization)
Licenses, tools and libraries

R&T Challenges for Mastering Embedded AI

Distributed AI & Computing

AI Applications & Algorithms

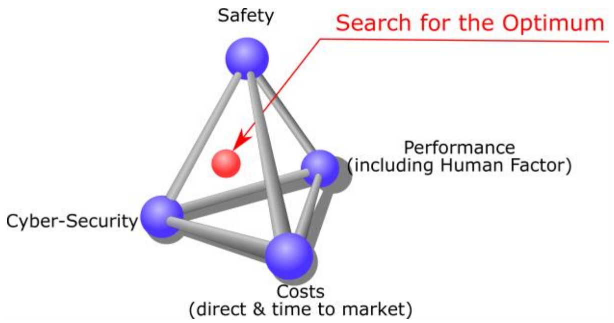
Embedded AI

Computing Platforms

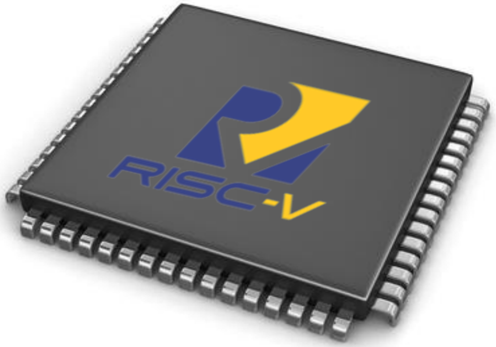
Tools & Methodo-logy

EDGE Computing

Open Hardware



Safety & Cyber-Security

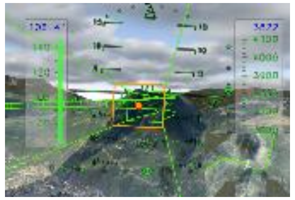


HW Accelerators

Calypso

e.g. Object detection

e.g. on-board satellite image processing



e.g. trajectory optimisation

THALES applications

Design constraints (perf/SWaP…)

Design and Optimisation

- AI algorithms optimisation and implementation
- Design and development of dedicated HW architectures
- Optimisation of applications for IoT and Edge computing
- Co-design with Thales Business Lines
- Selection of the best hardware target

Expertise

PoCs & derisking

Innovation transfers

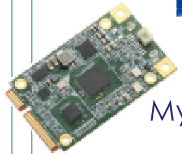
NVidia Tegra X2

NVidia Xavier

Xilinx VERSAL

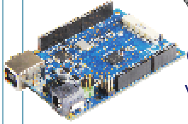
Microsemi PolarFire FPGA

Kalray MPPA



Intel MyriadX

STM32



Green waves GAP8

Embedded COTS

cnnGrid

Thales Neural Processor (TNP)

Thales IP

SNN, MRAM, Quantum computing …

Advanced technologies

THALES
Building a future we can all trust

Embedded AI workflow: towards automation

Automated Deployment workflow

- Pruning, layer fusion, decomposition, sparsity.
- Post Training Quantization.
- Quantization Aware Training.

DNN frameworks

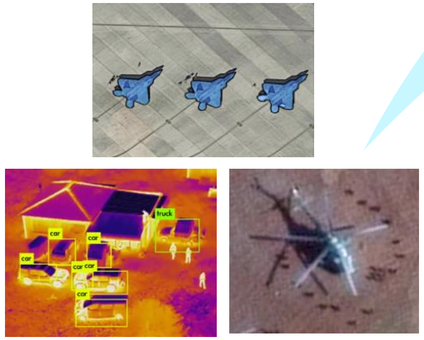
- Keras, TFLite, PyTorch
- TensorRT, TensorFlow Lite
- Apache TVM, AIDGE

Automated design flow

- Automatic NN optimization *(hyper-parameter search, HW-aware NAS methods, evolutionary algorithms…)*
- Algorithmic frugality *(few-shot learning, incremental learning).*

Implementation on embedded computing devices

Applications & annotated database



Data frugality & optimizations

- Data augmentation
- Domain adaptation / transfer learning

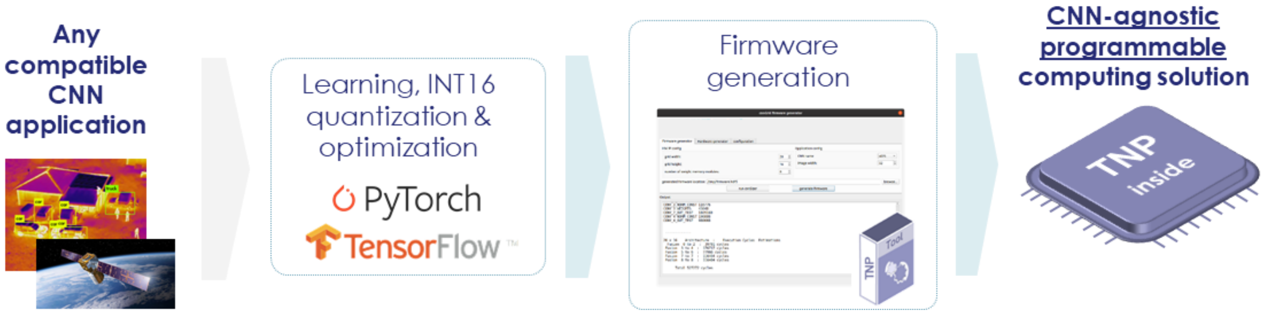
Thales Neural Processor

> Programmable processor in FPGA

- Fully programmable Convolutional Neural Network (CNN) processor.
- Support of a wide range of CNNs.
- No need of design expertise.

> Fully scalable architecture

- Exploiting 100s to 1000s of processing elements
  – Optimally used to deliver hundreds of GMAC/s
- Can be automatically tuned to any FPGA
  – Low-cost to high-performance devices
  – Automatic compiler generation for a given configuration



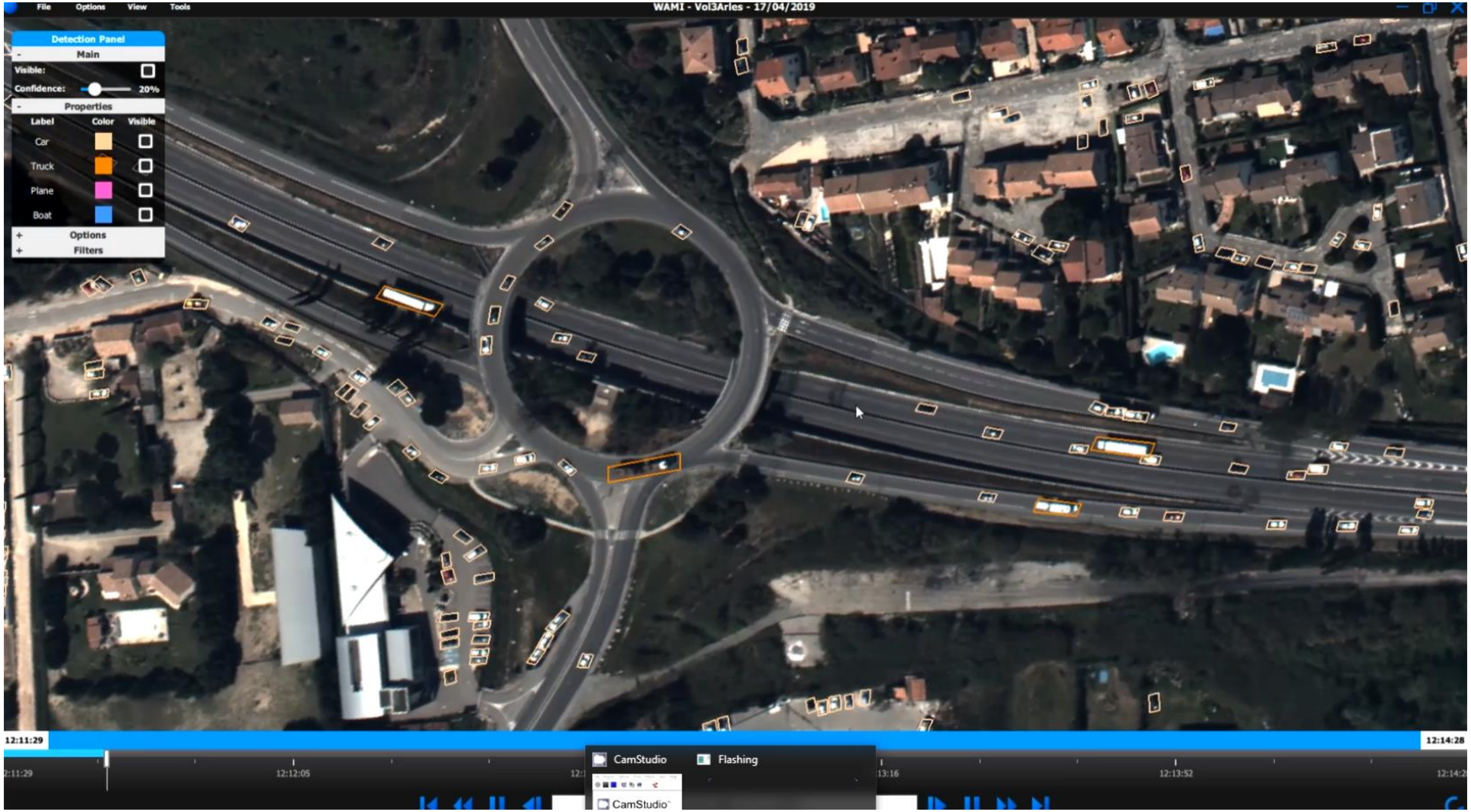
NanoXplore MICROCHIP XILINX intel FPGA

> Multi FPGA vendor support (ongoing work)

> Can be integrated in embedded smart sensors

Use-case example 1: oil spill detection

Use-case example 2: aerial vehicle detections



Edge Computing a new technology to master

THALES
Building a future we can all trust

Modular Open HW approach in order to optimize the energy efficiency

> Design of a modular software-defined Edge computer (ODYSSAI)

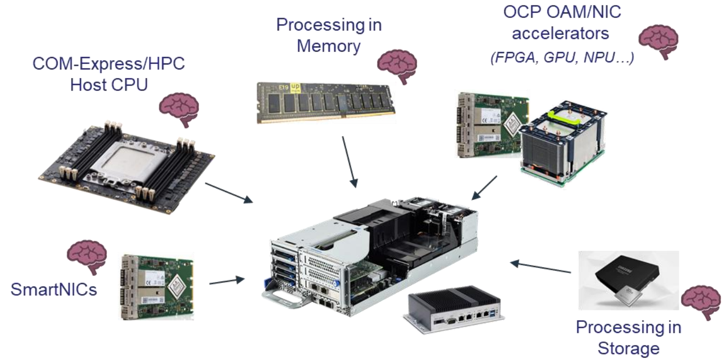
- Use of existing open standards (OCP, interfaces, interoperability, firmware, power, security, etc.),
- Integration of heterogeneous and energy-efficient computing solutions (either low cost),
- High modularity and scalability capabilities (sustainability),
- Leverage IT/OT convergence,
- Chiplet architectures with UCIe interfaces, …

OPEN Compute Project

ODYSSAI

> Leverage the compute-everywhere architecture

- Compute where the data are stored (NearMC, InMC),
- Compute during data transfers (DPU), RL during missions,…



Compute-everywhere Edge architecture

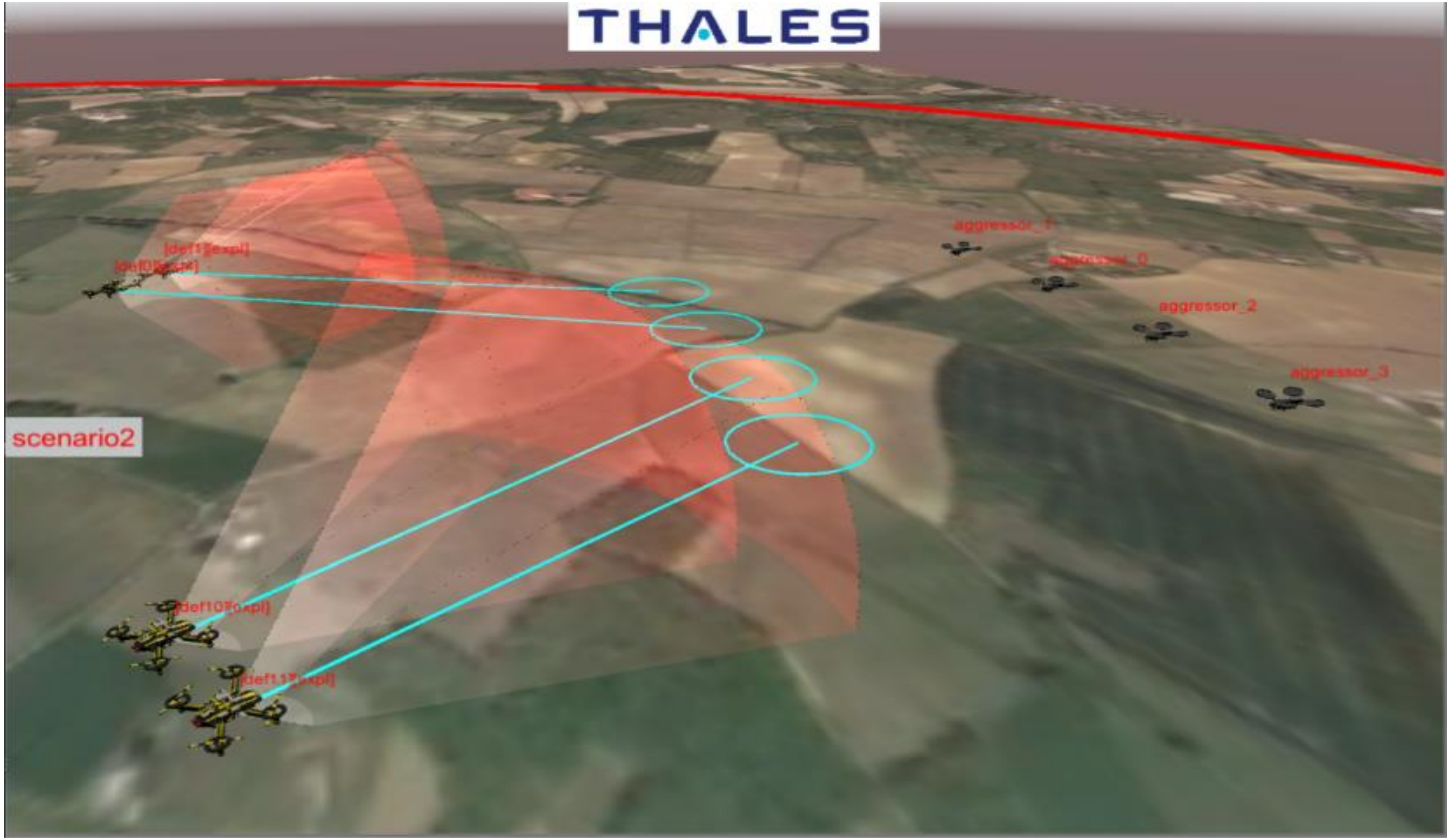
> Use of embedded computers instead of IT ones

- eGPUs vs IT GPU accelerators,
- ARM vs x86 cores, RISC-V tomorrow.

Opening up to European Suppliers (SMEs, start-ups, etc.) and then creating an European Edge Ecosystem

Autonomy = Distributed Embedded AI

REF TRT-Fr/STI/2024/TC0037 – July 2024

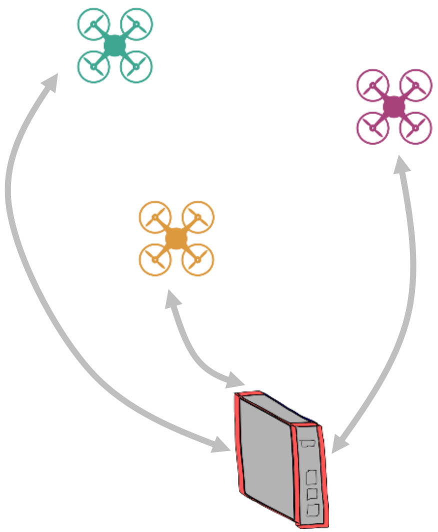

THALES
aggressor_2
aggressor_3
scenario2

What's next?

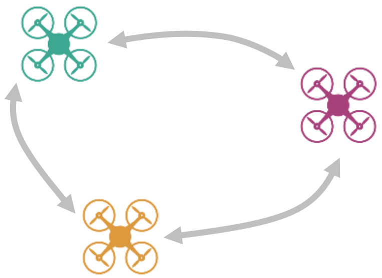
Optimality

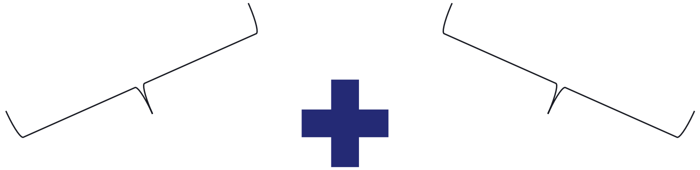
Resilience
Increased autonomy

Centralized



Decentralized





Hybrid

Best of both worlds

Take-away

IA EMBARQUÉE POUR LA DEFENSE

Le calcul « off line » devient de plus en plus « in line »

La maitrise des architectures edge est clé

L'IA distribuée n'est pas qu'un sujet algorithmique