Forum TERATEC 24

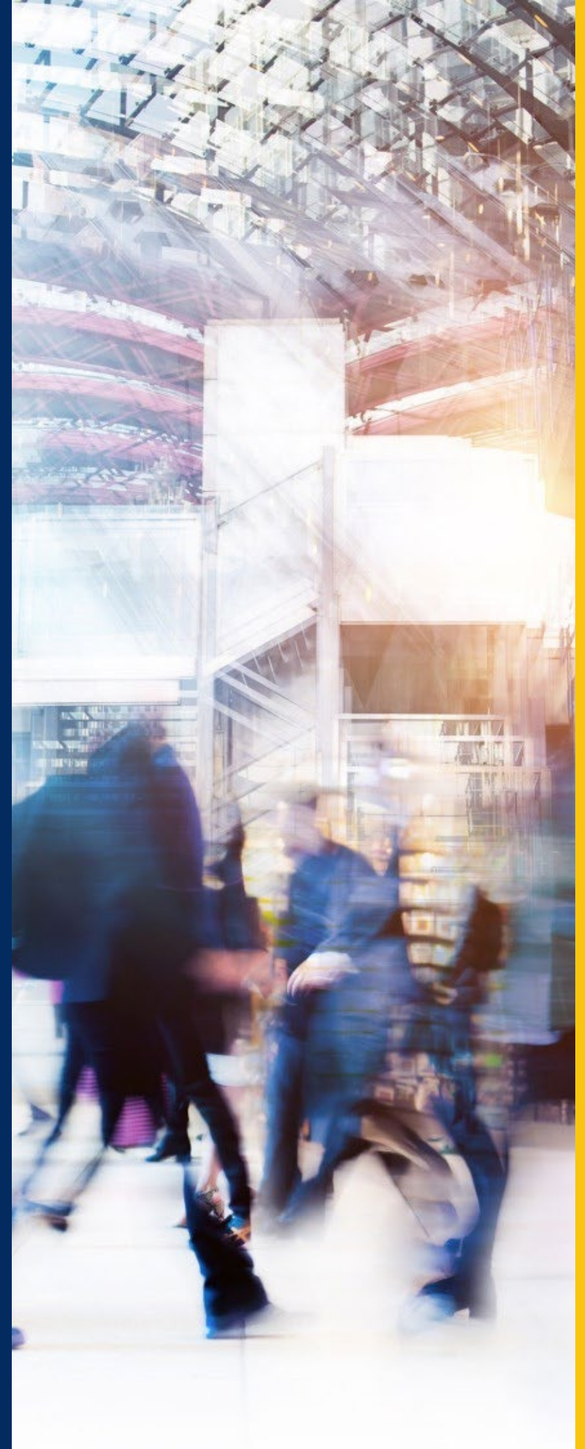
Unlock the future

Atelier 8 - : Technologies et usages du futur
30 mai 2024 — 09:00 - 12:30
PARC FLORAL PARIS

Evolution of computing and bringing Artificial Intelligence closer to the user

Marc Duranton (CEA)

Evolution of computing…



Exponential increase of performances in 33 years

Production car of 1985
Lamborghini Countach 5000QV
Max speed 300 Km/h

X 100 000 000

27 times the speed of light
Warp 3 ?
Star Trek Enterprise
(Year: about 2290)

Peta = 10^{15} = million of milliard

Exponential increase of performances in 33 years

Cray 2 – 1985
2 GFLOPS (2×10^9 FLOPS)

X 100 000 000

Summit – 2018
200 PFLOPS (2×10^{17} FLOPS)

200 kW

X 49

9 783 kW

Peta = 10^{15} = million of milliard

Energy efficiency x 2 000 000 in 33 years

Still increasing playing on specialization, architecture, data coding, …

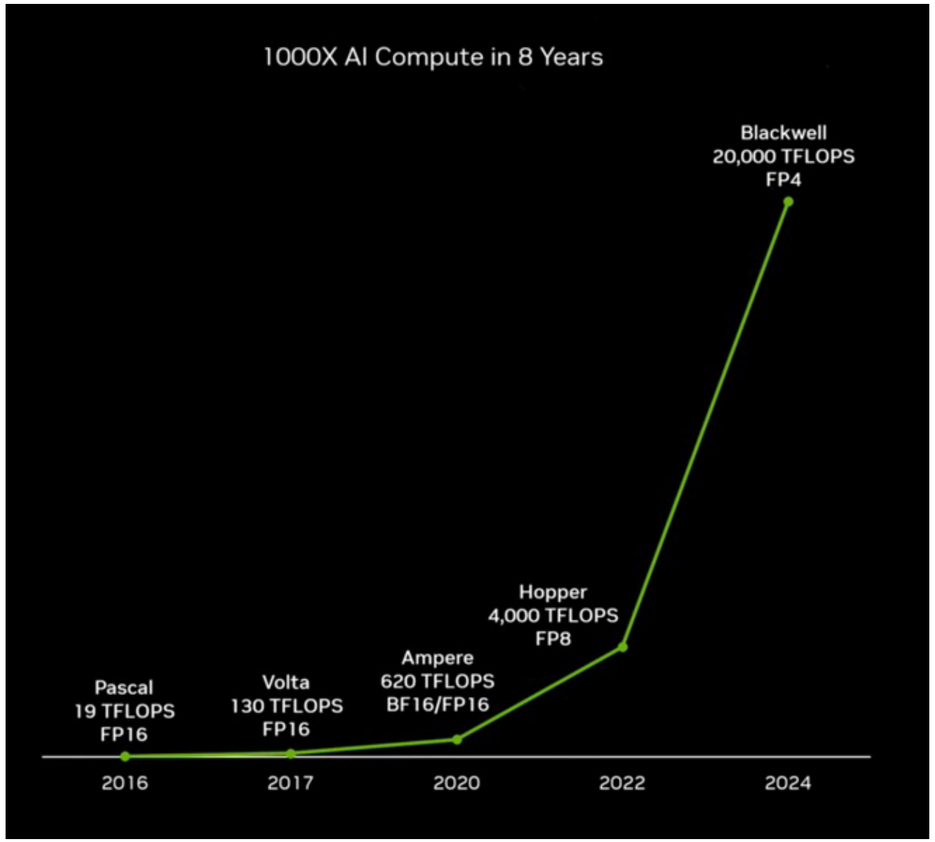
Even better increase for the AI accelerators

Exponential increase of performances in 33 years

Cray 2 – 1985
2 GFLOPS ($2x10^9$ FLOPS)

X 100 000 000
in 33 years

Summit – 2018
200 PFLOPS ($2x10^{17}$ FLOPS)

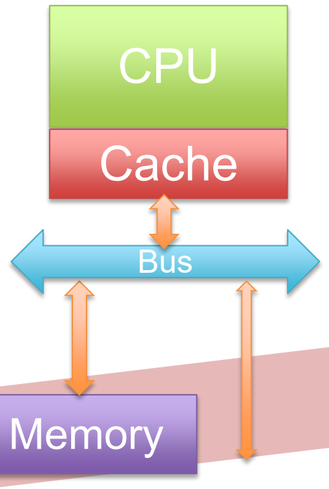


From Nvidia, J. Huang keynote 2024

Evolution of processing architectures

End of Dennard's scaling

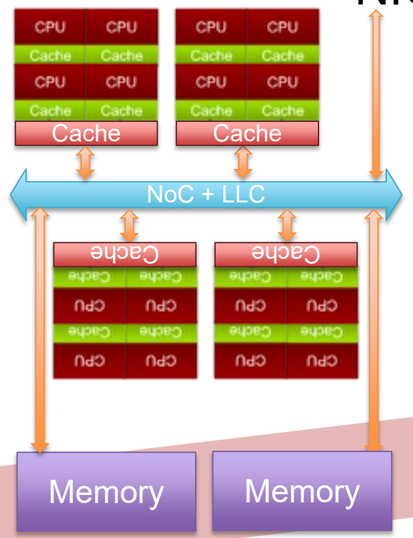
Mono-core architecture for single thread performance



NIC
(Network InterConnect)

~2006

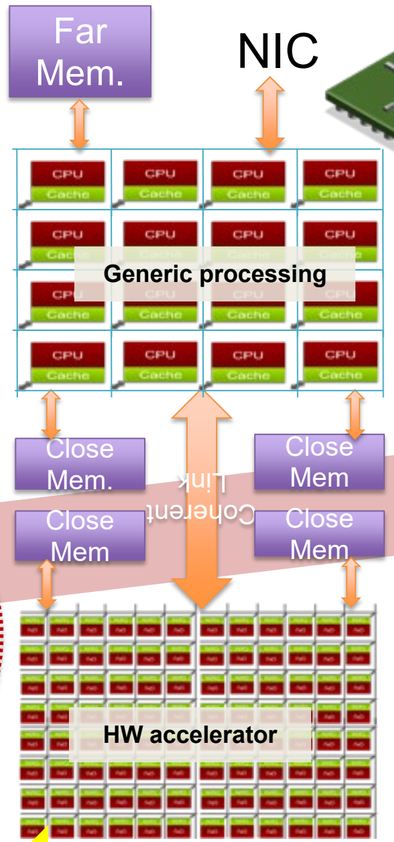
Many-core architecture for parallelism



NIC

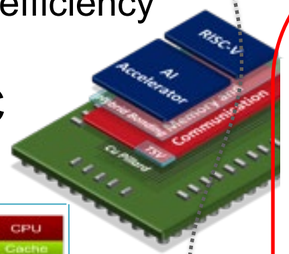
Heterogeneous architecture for energy efficiency

~2016



Reticule size wall

Heterogeneous integration for cost efficiency

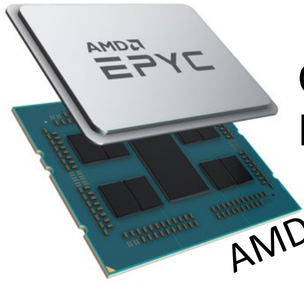


~2023...

Disaggregated SoC w/ functional chiplets

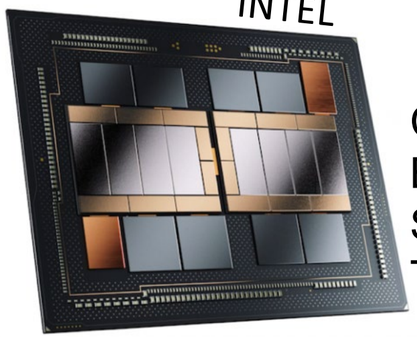
- Flexibility
- Perf scalability

Chiplet based architecture for higher density



Compute chiplet
IO chiplet

Chiplet invades High Performance Processors



Compute chiplet
Base die
Switch chiplet
Thermal chiplet

Increasing frequency (time)

Increasing parallelism (~ 2D space)

Increasing specialization (architecture)

Computing fabric (approximate computing?)

No more electrons?

Slide originally from Denis Dutoit (CEA)

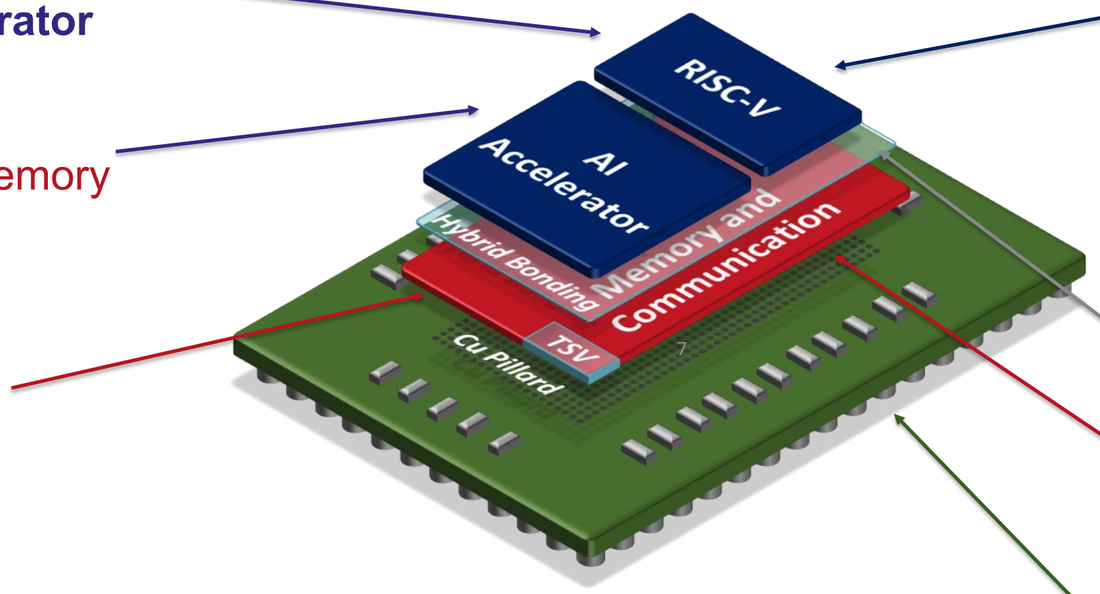
A complete computing system in a package

Technology and architecture co-optimization towards a modular approach for heterogeneous integration

Architecture

- **Generic Computing**
- **Low power accelerator**

- **Chassis die** with peripherals, IOs, memory and communication infrastructure

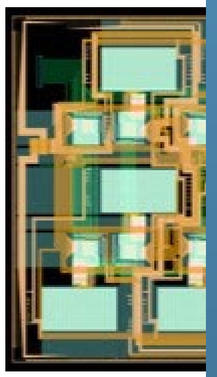


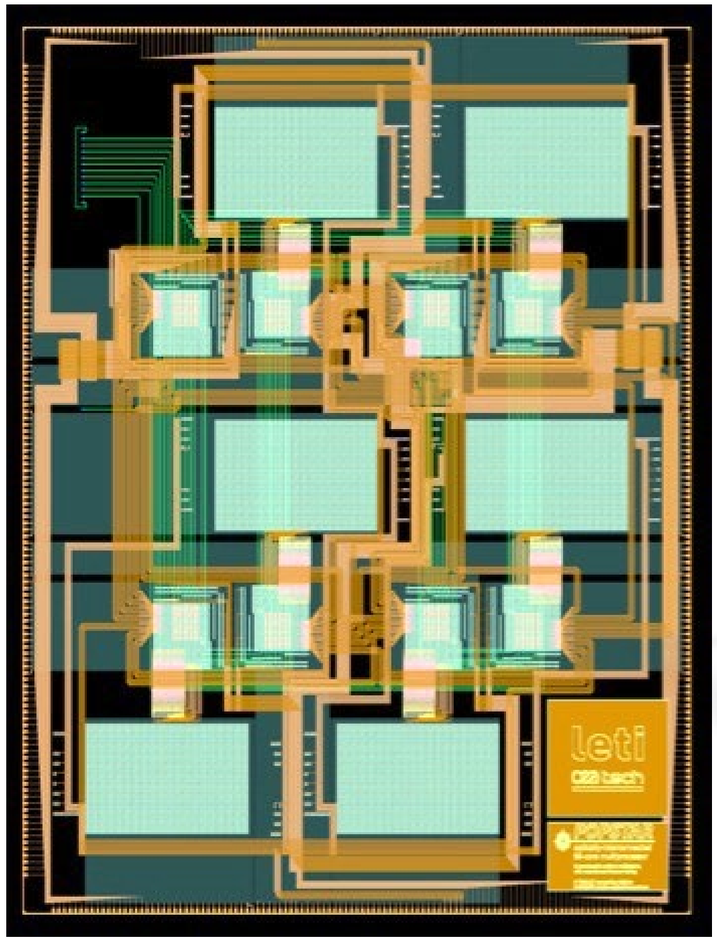
Technology

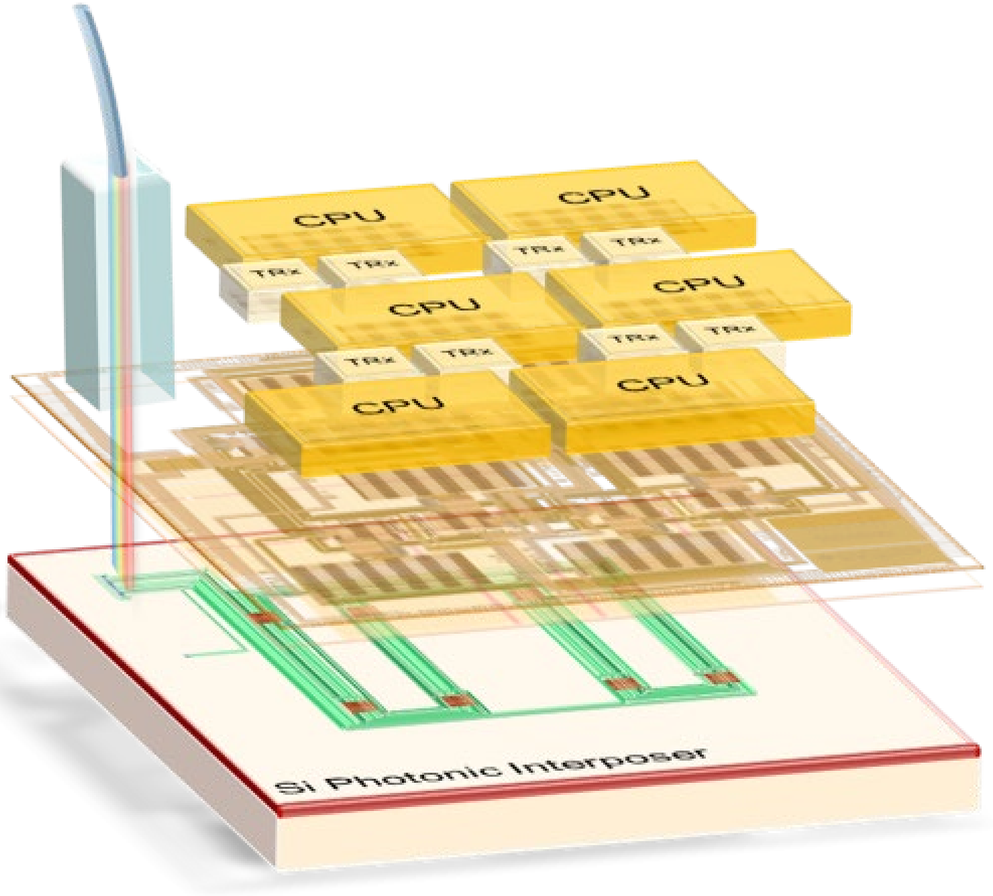
- **Multi-chiplets:**
  - Advanced technology node
  - Heterogeneous (size, technology node, pitch)
  - Face down
  - No TSV
  - Full digital compute chiplet
- **Hybrid bonding:**
  - Die-to-wafer, Face-to-face
- **Base die:**
  - Mature technology node
  - TSVs for power delivery and IOs
  - Face-up
- **Package**

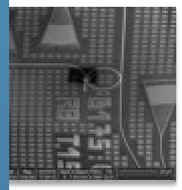
PHOTONIC INTERPOSER: OPTICAL COMMUNICATION ON INTERPOSER

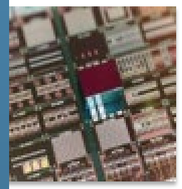
- Aims to valida
  interposer

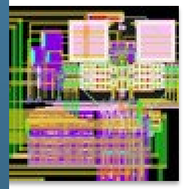


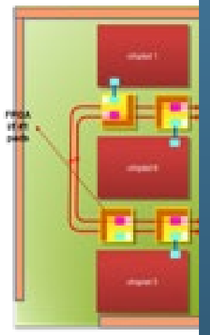










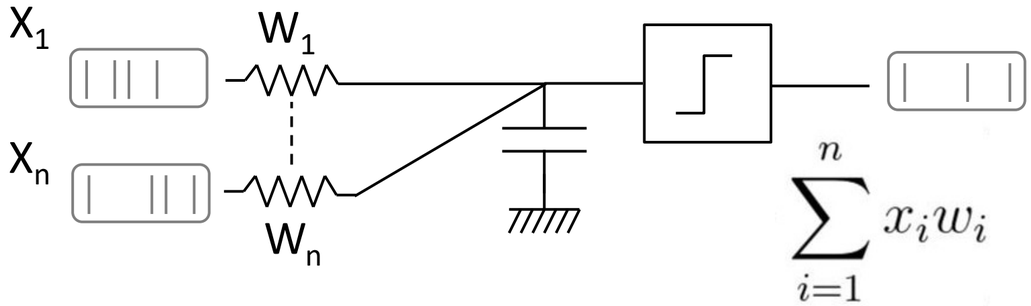


96 cores pr
interposer

Few points for innovative and new hardware

New paradigms are stimulating interdisciplinary research (materials, information theory, complexity etc.):

- quantum computing (and impact on algorithms)
- Accelerators not using “bits”
  - neuromorphic computing
    - Information coded in time (“spikes”)
    - Using physical phenomenon to compute
  - Ising based accelerators
  - Computing using photons



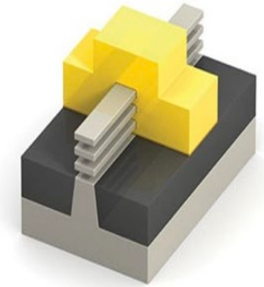
New technologies are emerging:

- memory e.g. MRAM, spintronics
- processors e.g. stacked nanosheet FET (GAA), photonics for computing
- organic and flexible electronics

Innovative architectures, often driven by the performance demands of data-intensive computing, tailored to the application:

- e.g. in-memory computing
- e.g. 3D stacking
- e.g. using adapted precision (accuracy)

- Programmability of the new approaches? Ease of use for the developers?



Stacked nanosheet FET
The gate completely surrounds the channel regions to give even better control than the FinFET.

Changing computing paradigm: from "precise" to « approximate » computing

Until now, computers and other computational systems have been used for ***reproducible***, relatively ***precise computation*** (with the exception of errors caused by floating-point representation => Extended Precision Computation, CEA's VXP coprocessor, - and overflows (integer computation)).

A Computing Accelerator Designed for High Precision Computation (up to 512-bit mantissa)

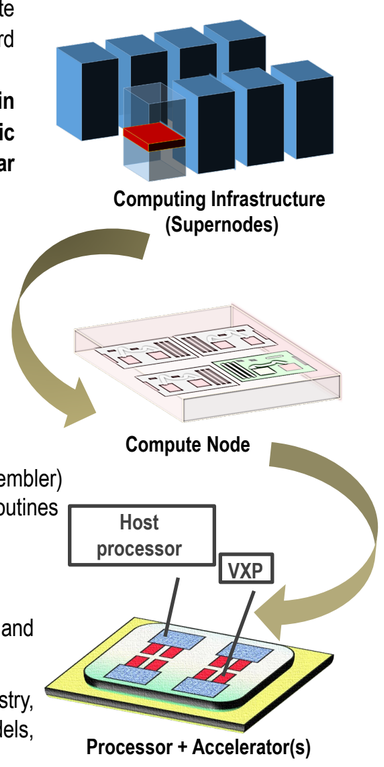
OVERVIEW

The variable extended precision processor (VXP) is a dedicated hardware/software **accelerator suitable for the resolution of large ill-conditioned systems of equations.** Its **tunable, dynamic precision speeds up convergence** and **improves memory usage** and **computational efficiency.**

BENEFIT : HIGHER PRECISION FOR IMPROVED EFFICIENCY

Increased precision greatly reduces rounding errors, and improves the computing efficiency of algebraic computations at the compute node level. Certain problems do not even converge with standard double precision.

The VXP accelerator supports **arithmetic operations in hardware with up to 512 bits of mantissa.** Its **dynamic precision** is **fine grain tunable** for **optimal use of near processor memory.**



KEY FEATURES

The VXP is a complete **hardware** and **software** solution with:

- Dedicated hardware :
  - ✓ Silicon proven in GF 22nm FDX and new design in TSMC 7nm (**E**uropean **P**rocessor **I**nitiative)
  - ✓ FPGA board for early access
- Software stack :
  - ✓ C-like programming environment (compiler and assembler)
  - ✓ Library for mathematic and low-level algebraic subroutines
  - ✓ Runtime environment

APPLICATIONS

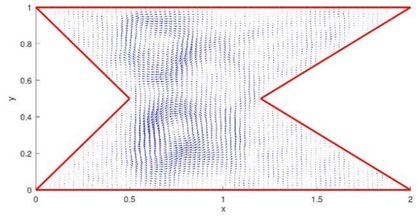
Improve the efficiency of computing for algebraic solvers and eigensolvers :

- Scientific computing : computational physics and chemistry, electronic simulation, structural computation, climate models, weather prediction, fluid dynamics.
- Model order reduction : learning for AI, large dynamic systems.

FIRST RESULTS ON AN ILLUSTRATIVE EXAMPLE

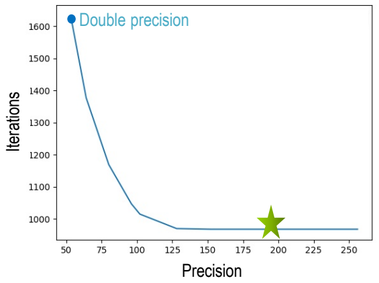
Modern linear algebra kernels (solvers, eigensolvers…) are highly sensitive to numerical pitfalls (cancellation, absorption...) which are a source of computational instability. This may alter or even occult some physical phenomena, whereas augmenting precision restores the numerical consistency of the model.

By modelling a classical laminar flow problem over a cavity with **192 bits** of mantissa**, turbulence details appear** whereas they are lost in the noise with double precision (53 bits of mantissa).



Difference in Laminar Flow between Solutions with 53- and 192-bit Precision.

In this application, the variable extended precision processor (VXP) allows us to :
- model non-observable variations in standard precision.
- reduce the number of iterations until convergence by 40%



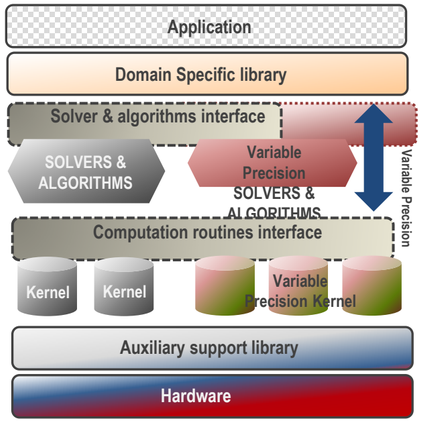
Number of Required Iterations versus Numeric Precision (bits of mantissa)

HOW DOES IT WORK ?

A conventional scientific application can smoothly be integrated with the VXP unit, just as a plug-in for scientific software applications. Whenever the compute node host cannot achieve the expected accuracy with standard precision, the VXP takes over and continues with extended precision until the error tolerance constraint is met.

In the current version, the VXP is embedded as a functional unit in a 64-bit RISC-V processor pipeline. The VXP extends the standard RISC-V instruction set with basic arithmetic operations and specific instructions in variable precision.

The VXP relies on the RTEMS software for communication with the host and global synchronization.



See https://list.cea.fr/en/vxp-extended-precision-processor/

Changing computing paradigm: from "precise" to « approximate » computing

Until now, computers and other computational systems have been used for ***reproducible***, relatively ***precise computation*** (with the exception of errors caused by floating-point representation => Extended Precision Computation, CEA's VXP coprocessor, - and overflows (integer computation)).

- The "new" generation of computational approaches make computing more "***approximate***" :
  - Neural network-based approaches, including generative AI (hallucinations, ...), low-precision coding (FP4 [1, 3, 0])
  - ***Ising-based coprocessors*** (Fujitsu Digital Annealer, Hitachi machine, Dwave): find a minimum of a function, not necessarily the global minimum => Quadratic unconstrained binary optimization => optimization problems.
  - Quantum computing (stochastic measurement of results)

- We are going ***from (parallel) Turing machines* (1936) to universal approximators*** of Mc Culloch & Pitts (1943)[+]

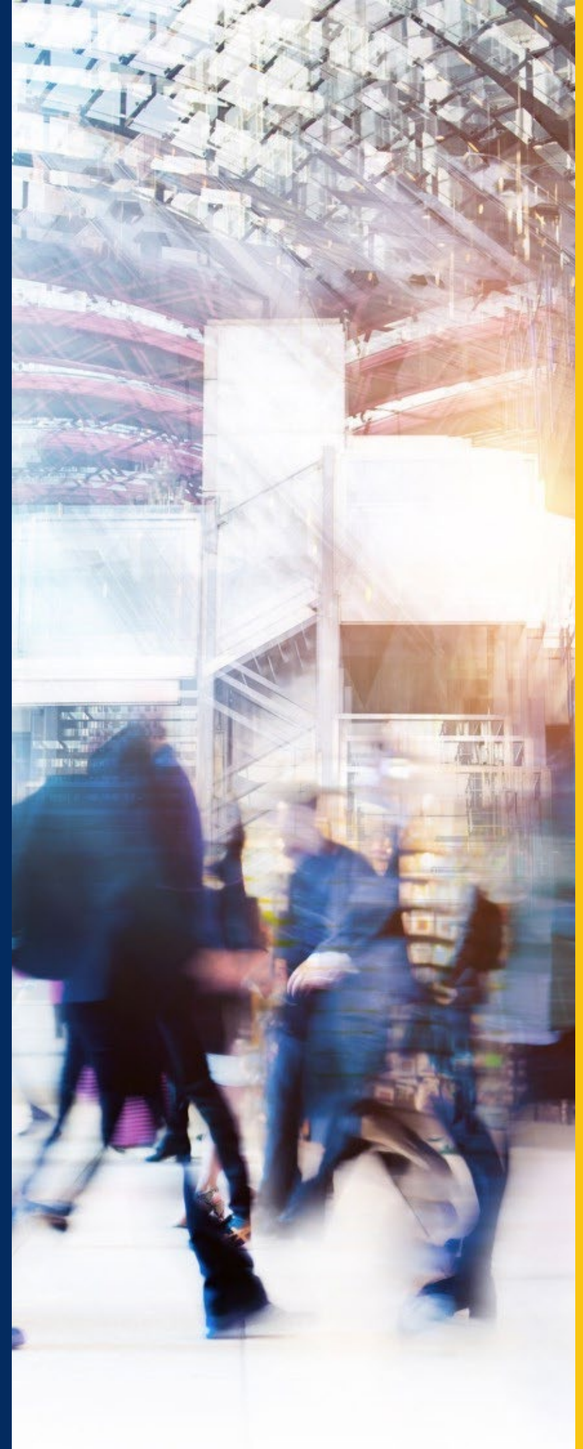
- Tomorrow's systems ***will have to combine the two**** types in "loop", "reinforcement" systems

*Turing, 1942: *"Any form of mathematical reasoning can be made by a machine".*

[+] Mc Culloch & Pitts, 1943, A finite size neural network can approximate any function to a desired degree of precision

** *Like in « Thinking, Fast and Slow »* a 2011 popular science book by psychologist Daniel Kahneman

The “boom” of Artificial Intelligence



Hype Cycle for Artificial Intelligence, 2023

Plateau will be reached:

○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ⊗ obsolete before plateau

As of July 2023

gartner.com

Source: Gartner

Gartner

Smaller LLM models get more powerful, ready for on-premise processing

Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
GPT-4-0314	1186	+3/-3	53597	OpenAI	Proprietary	2021/9
Qwen-Max-0428	1184	+4/-4	21973	Alibaba	Proprietary	Unknown
GLM-4-0116	1184	+6/-6	7585	Zhipu AI	Proprietary	Unknown
Claude 3 Haiku	1178	+4/-2	82998	Anthropic	Proprietary	2023/8
Qwen1.5-110B-Chat	1164	+4/-4	19369	Alibaba	Qianwen LICENSE	2024/4
GPT-4-0613	1161	+3/-3	75182	OpenAI	Proprietary	2021/9
Reka-Flash-21B-online	1156	+4/-4	16039	Reka AI	Proprietary	Online
Mistral-Large-2402	1156	+3/-3	53756	Mistral	Proprietary	Unknown
Llama-3-8b-Instruct	1153	+2/-2	79064	Meta	Llama 3 Community	2023/3
Claude-1	1149	+5/-4	21216	Anthropic	Proprietary	Unknown
Reka-Flash-21B	1148	+3/-4	23182	Reka AI	Proprietary	2023/11
Mistral Medium	1148	+3/-3	35600	Mistral	Proprietary	Unknown
Command R	1147	+4/-3	44680	Cohere	CC-BY-NC-4.0	2024/3
Qwen1.5-72B-Chat	1147	+3/-3	38871	Alibaba	Qianwen LICENSE	2024/2
Mixtral-8x22b-Instruct-v0.1	1146	+4/-3	34799	Mistral	Apache 2.0	2024/4
Claude-2.0	1131	+5/-5	12789	Anthropic	Proprietary	Unknown
Gemini Pro (Dev API)	1131	+6/-5	18839	Google	Proprietary	2023/4

From: https://chat.lmsys.org/?leaderboard on 24/05/29

LLM running locally on your smartphone in 2024

From https://www.mediatek.com/products/smartphones-2/mediatek-dimensity-9300

November 6th, 2023

« the APU 790 can run a 7 billion parameter LLM at 20 tokens per second, which is fast enough for real-time use. …For comparison, **Qualcomm** says its **Snapdragon 8 Gen 3** can run a 10 billion parameter LLM at almost 15 tokens per second, which seems fairly comparable. The Dimensity 9300 can extend this to run a 13 billion LLM within 16GB of RAM, right up to 33 billion parameters with 24GB RAM, albeit with a much slower 3-4 tokens per second processing rate. »*

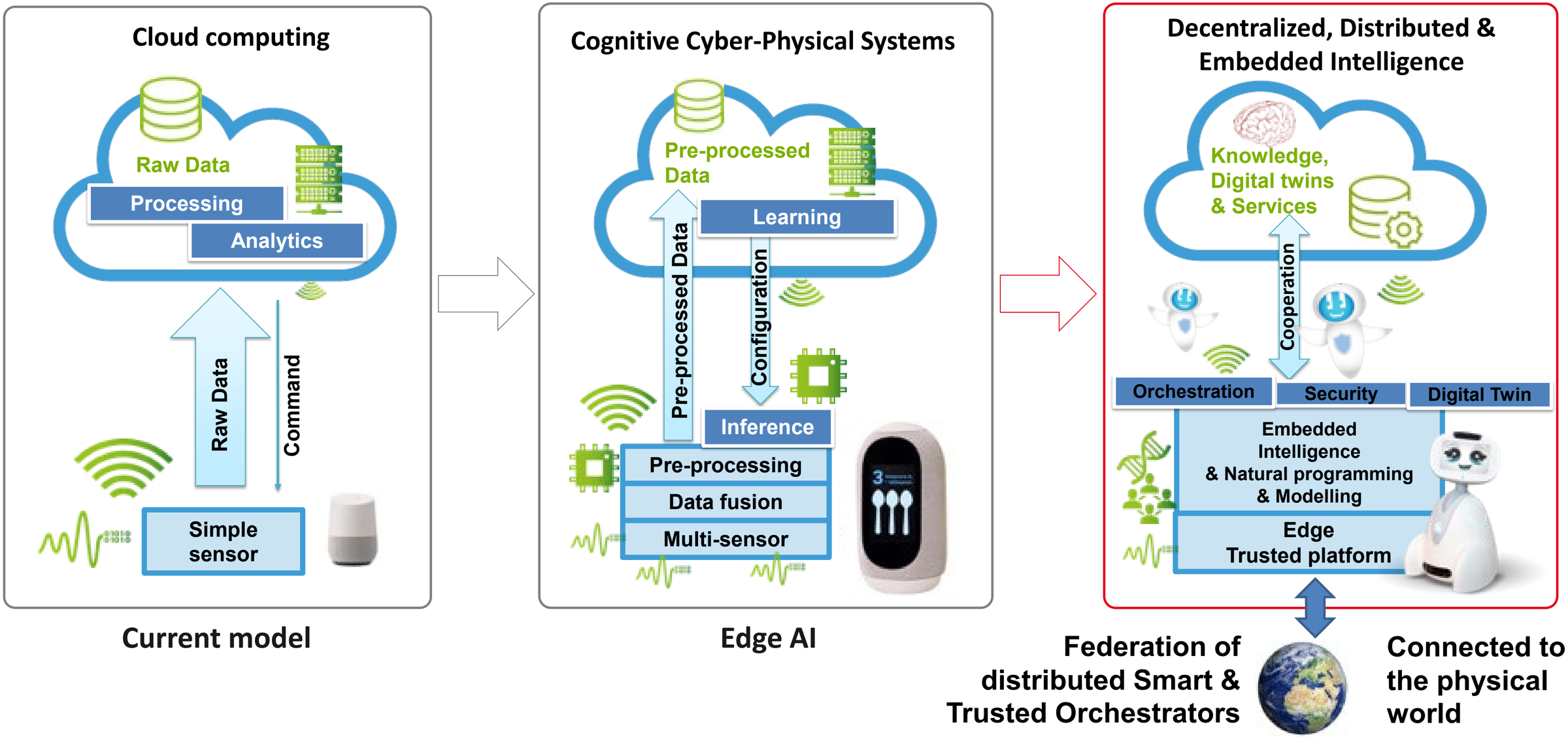
« the APU 790 supports INT4 (A16W4) to run smaller quantized models and a dedicated hardware memory decompression block that feeds the APU. In MediaTek's example, a 13GB INT8 model can be pre-compressed to just 5GB to fit into RAM and then decompressed in hardware on its way to the APU. »*

Support for NeuroPilot Fusion, which can continuously perform LoRA low-rank adaptation

* From https://www.androidauthority.com/mediatek-dimensity-9300-explained-3381678/

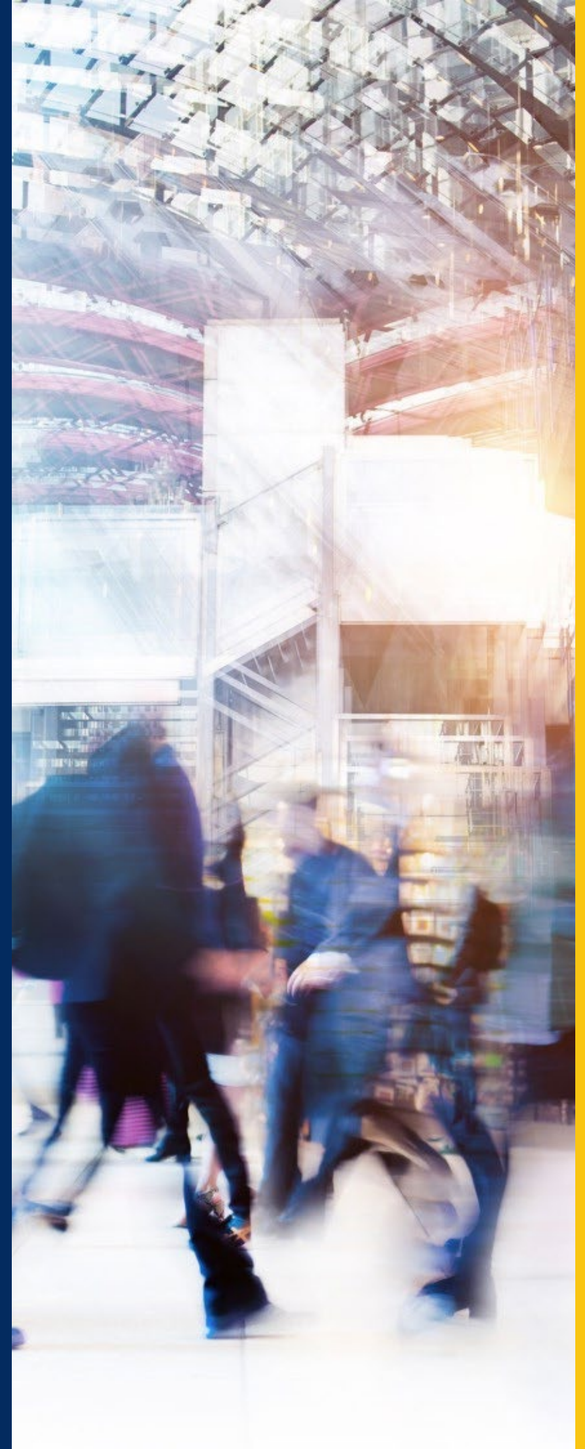
Evolution of computing:

Cloud computing ➔ Continuum of Computing

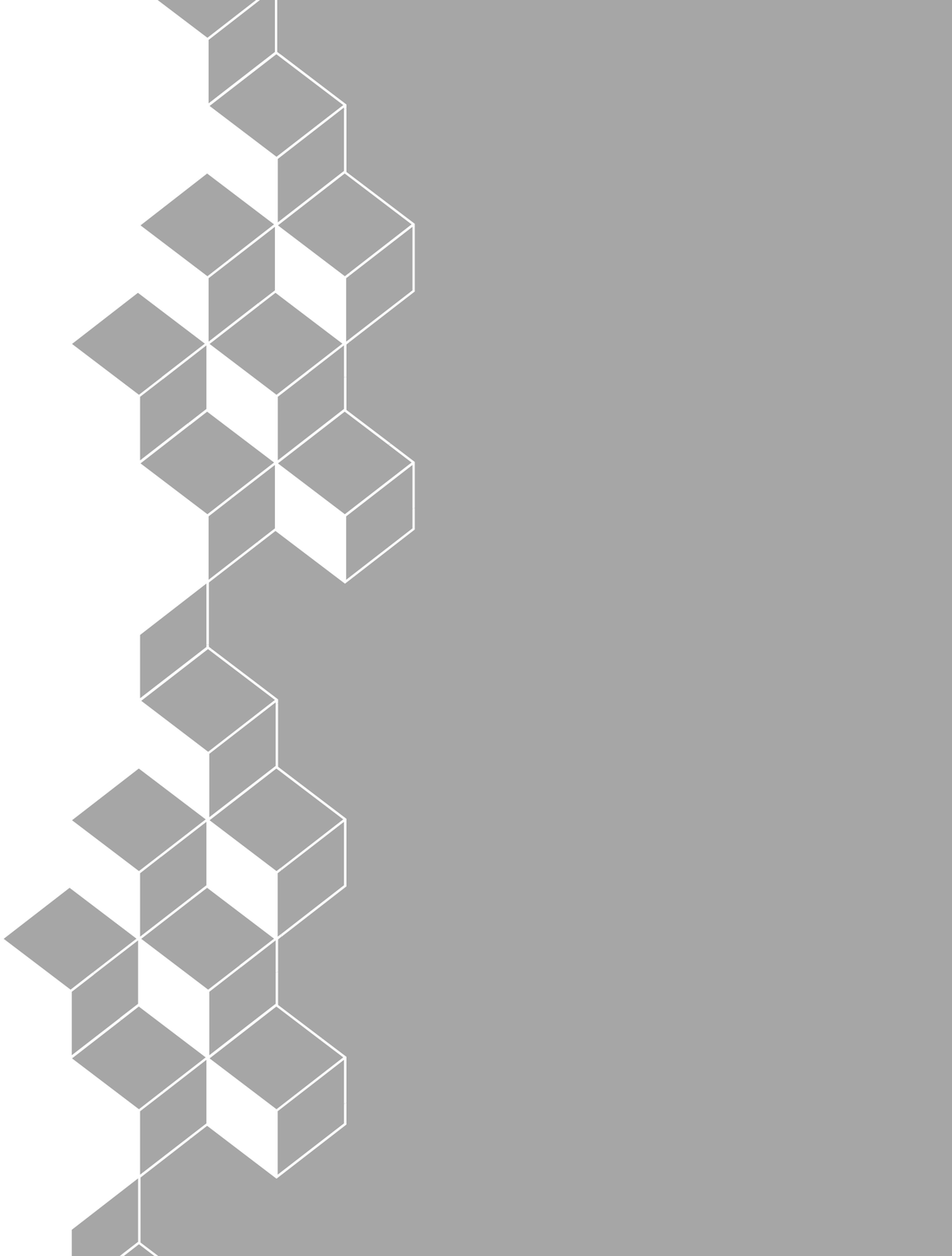


Slide originally from Denis Dutoit (CEA)

How to have efficient AI devices at the edge?

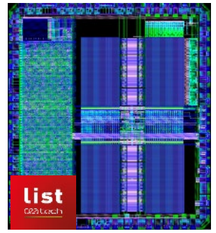


NEURAL NETWORKS HARDWARE ACCELERATORS

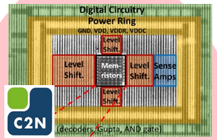


Efficient Computing @Edge: Examples of AI Circuits & Architectures

Smart Sensors



Esperanto
Gesture detection
130nm + OxRAM + PMUT,



Bayesian machine

BNN
Digital popcount
130nm, OxRAM

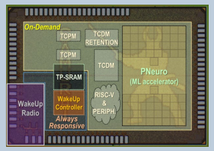
Classification

SPIRIT
Spiking SNN
130nm, OxRAM

Spiking SNN

LARGO
28nm, OxRAM

CNN



PNEURO
Raptor HW/SW IP
FD28nm

DOLPHIN DESIGN

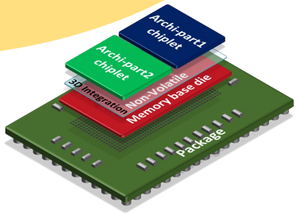
CNN backbone

NeuroCORGI
Feature Extractor (HD images)
GF22nm, OxRAM

Transformers & multimodality

INSTITUT CARNOT CEA LIST

BUMBLEBEE
Attention based AI Circuit Architecture



MOSAIC
Modular AI systems composed of heterogeneous components

Zoom on PNEURO: A vision/audio AI accelerator IP
Bringing AI computing into energy constrained sensors

Object classification | Object detection | Image segmentation | Energy efficient | Fully-prog. | Scalable

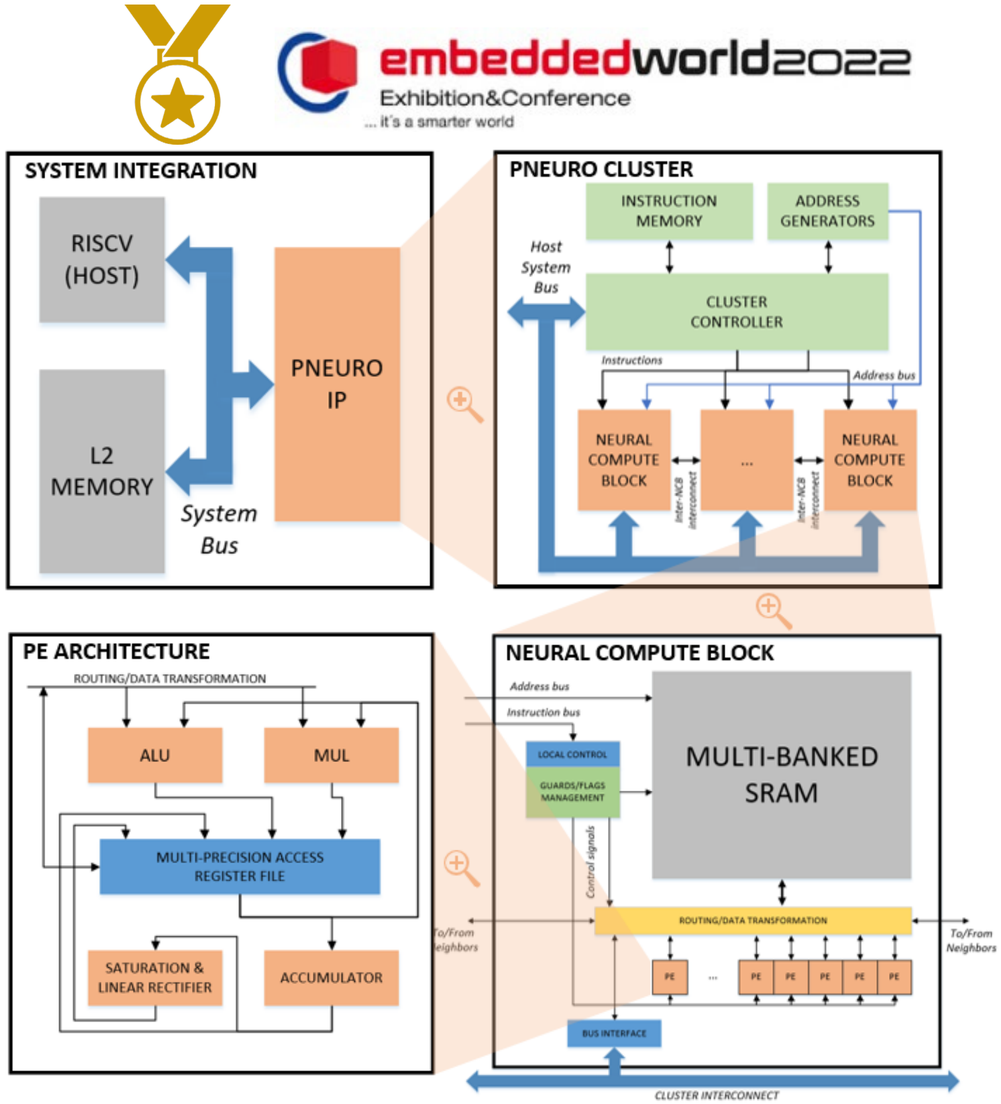
- An efficient and configurable tiny ML accelerator
- Designed to execute Deep Neural Network (Pre/post-processing phases, CNN)
- Scalable thanks to a clustered SIMD (Single Instruction Multiple Data) architecture
- Energy efficient thanks to strong coupling between PE and memory (NMC)

IP — Hardware IP

SDK — Programming user interface

ISS — Model for virtual programming

embeddedworld2022 Exhibition&Conference ... it's a smarter world



Tiny Raptor/P-Neuro – Benchmark results

Fully programmable Neural Processing Unit, designed to execute **Deep Neural Networks (DNN)** in an energy-efficient way thanks to **Near-Memory Computing** architecture

DOLPHIN DESIGN

RAPTOR

KeyWord Spotting results

Data: Google Speech Commands Model: DS-CNN Accuracy: 90% (top1)

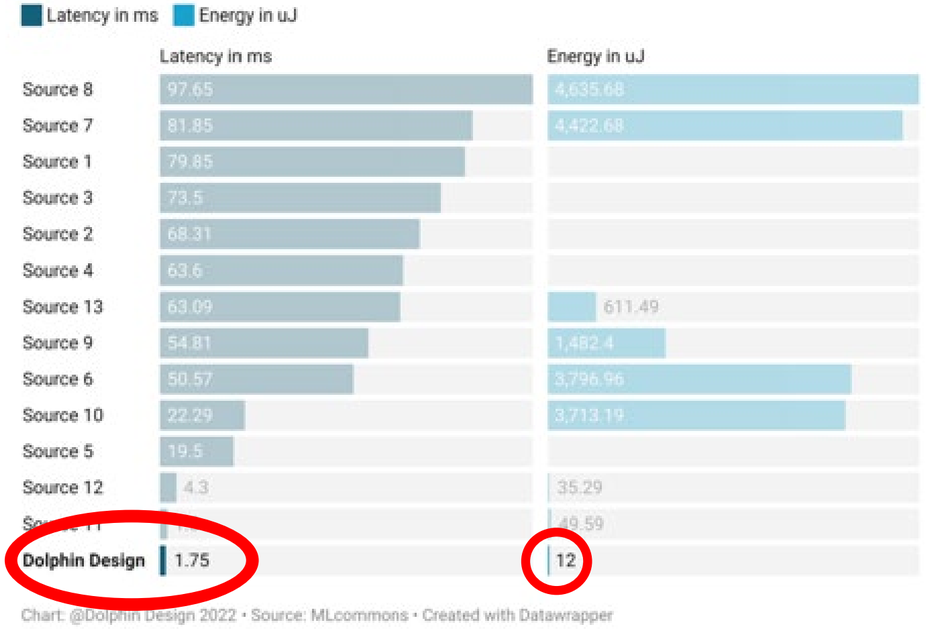


Chart: @Dolphin Design 2022 • Source: MLcommons • Created with Datawrapper

Visual Wake Words results

Data: Visual Wake Words Dataset Model: Mobilenetv1 (0.25x) Accuracy: 80% (top1)

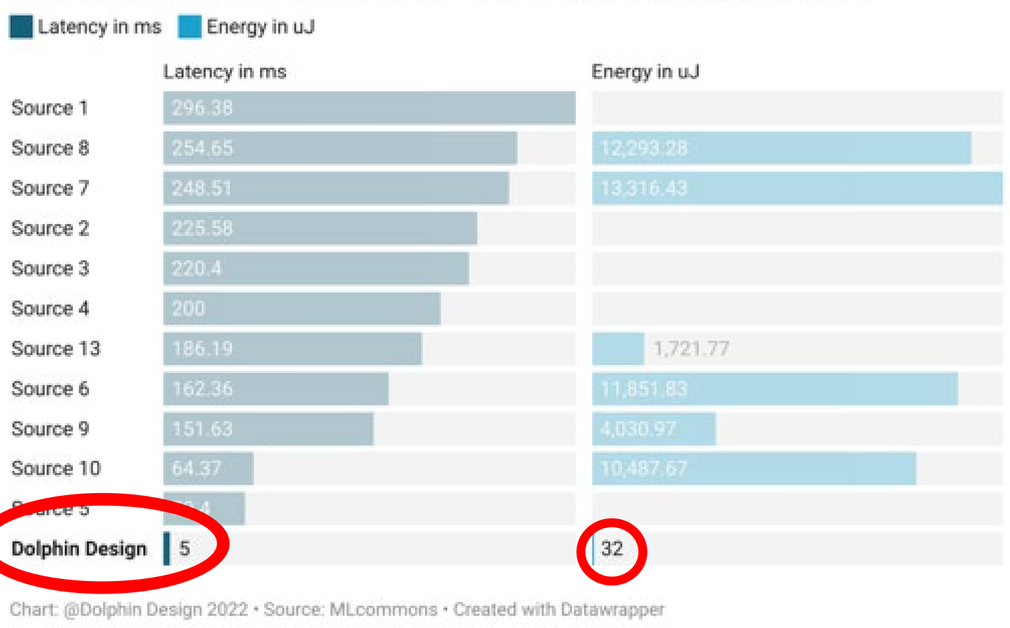
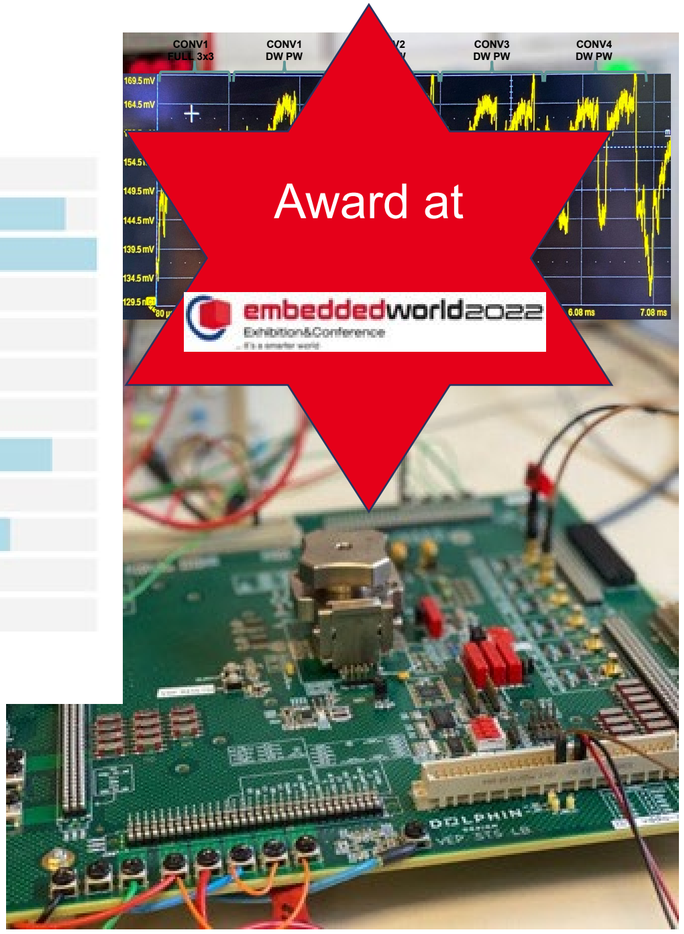


Chart: @Dolphin Design 2022 • Source: MLcommons • Created with Datawrapper

[1] https://mlcommons.org/en/inference-tiny-07/



embeddedworld2022
Exhibition&Conference
... it´s a smarter world

Embedded World Award 2022 in the start-up category for Tiny Raptor

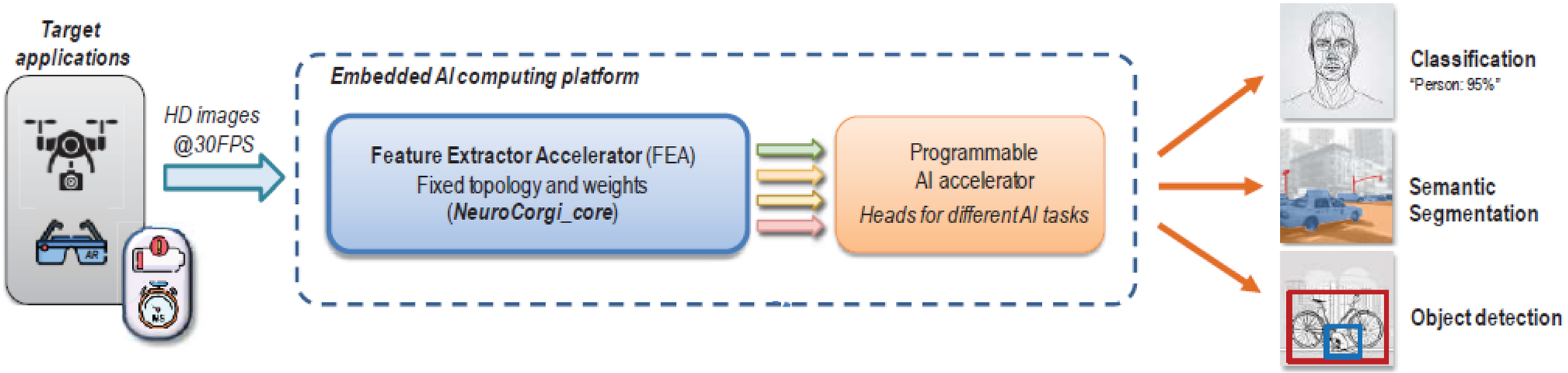
cea

Tiny Raptor – Demonstration CamCube at CES 2023

RAPTOR

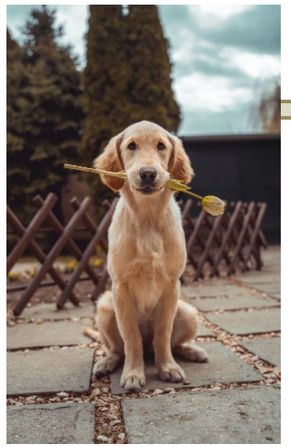
NeuroCorgi – A Feature Extraction Accelerator

- Key Idea : Dedicated hardware **Feature Extraction Accelerator** (FEA)
  - Weights and topology of the network are frozen
  - Fixed weights allows hardware optimization : MCM=Multi-Constant Multiplier
  - Highly tuned quantization (4-bits, same accuracy as original FP32)



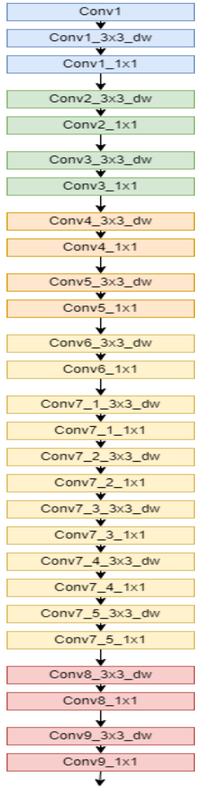
- Same FEA used for different tasks : Classification, Semantic Segmentation, Object Detection
  - With transfer learning (TL) → same features can be used for completely different applications

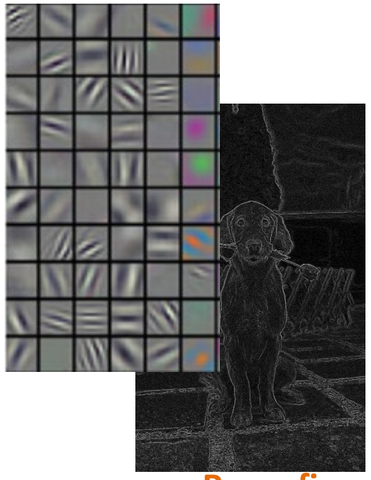
NeuroCorgi concept and demonstration at CES 2024



Input

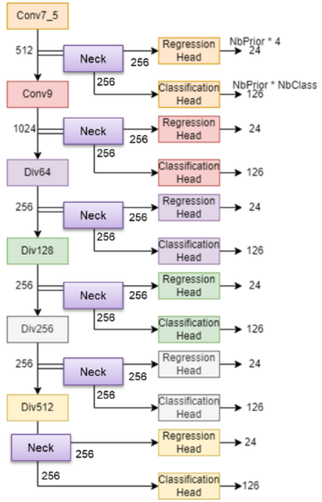
Fixed



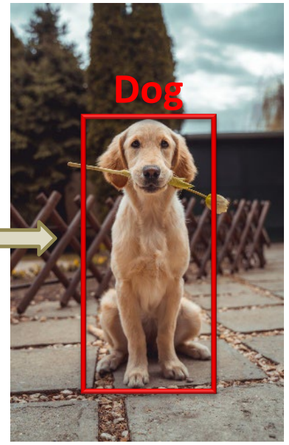


Backbone
(features extractor)

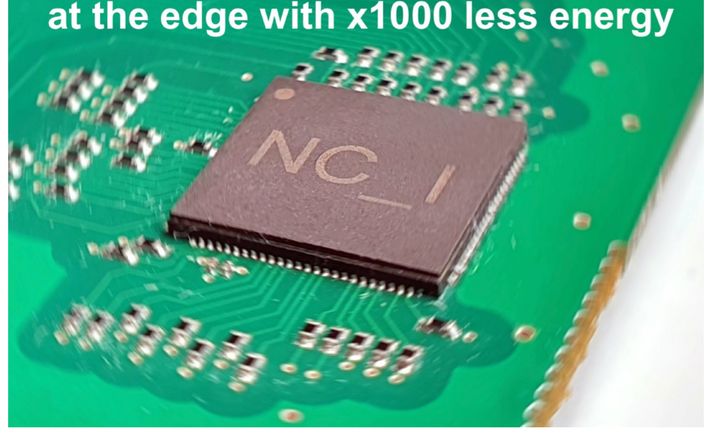
Reconfigurable



NeuroCorgi



Output



Use Case 2.1:
Autonomous
Weeding System

Use Case 3.1:
Drones/USV

Use Case 2.2:
Tomato pests and
diseases forecast

Use Case 3.2:
Underwater Acoust
Signal Classification

Use Case 3.3:
3D Object Detection
and Classification of
Road Users

NeuroCorgi – Circuit Performance

- Based on MobileNetV1 topology trained with ImageNet (27 CNN layers) [1]

Parameter	Value
Frame Rate	30 FPS
Image Format	Up to 1280x720 pixels
Technology	GF 22FDX
FEA Area	4.45 mm^2
Main Clock	59 MHz

Parameter	Value
Power	<100 mW
Latency (1280x720)	<10 msec

[1] I. Miro-Panades et al., "Meeting the Latency and Energy Constraints on Timing-critical Edge-AI Systems," book Embedded Artificial Intelligence, River Publishers, 2023.

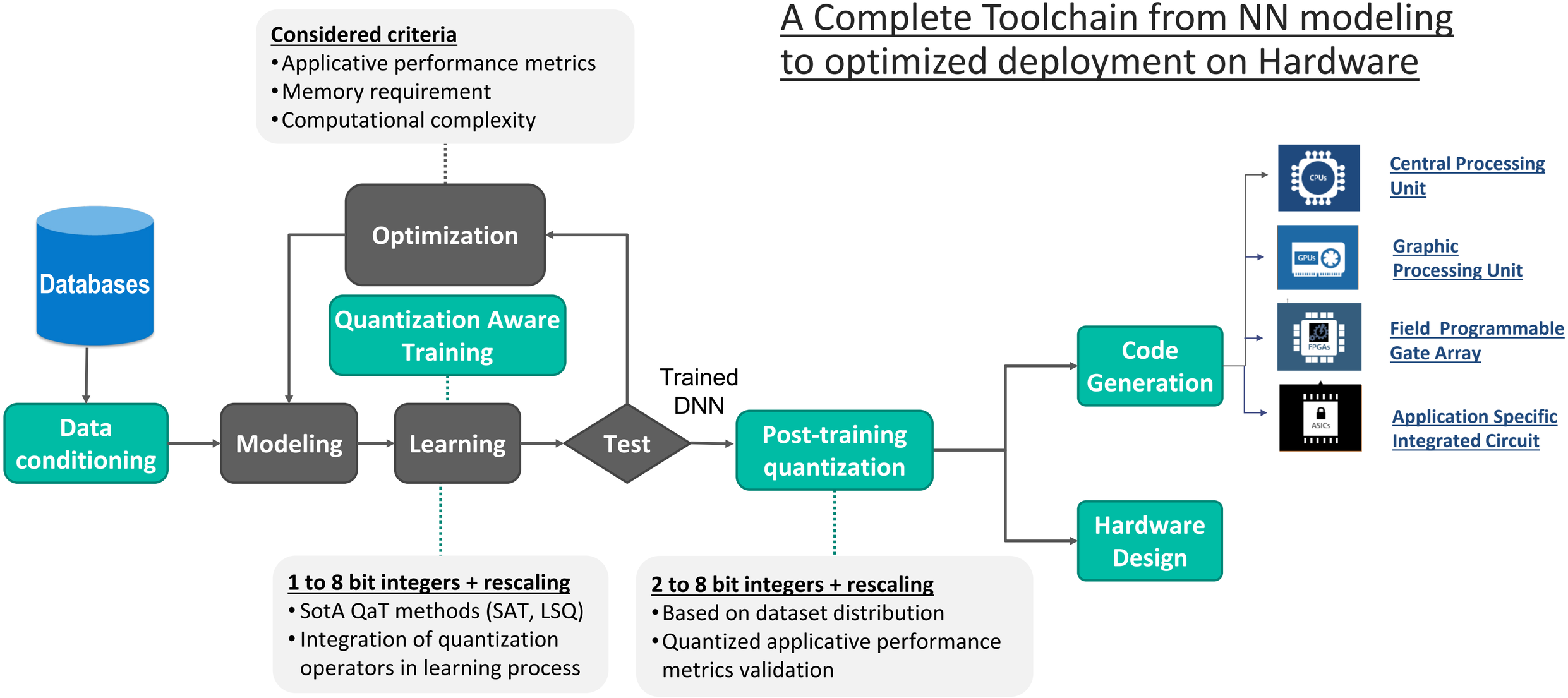
INDEPENDANT DEEP LEARNING PLATFORM FOR EMBEDDED AI

aidge

Deep Learning Platform for Embedded

aidge

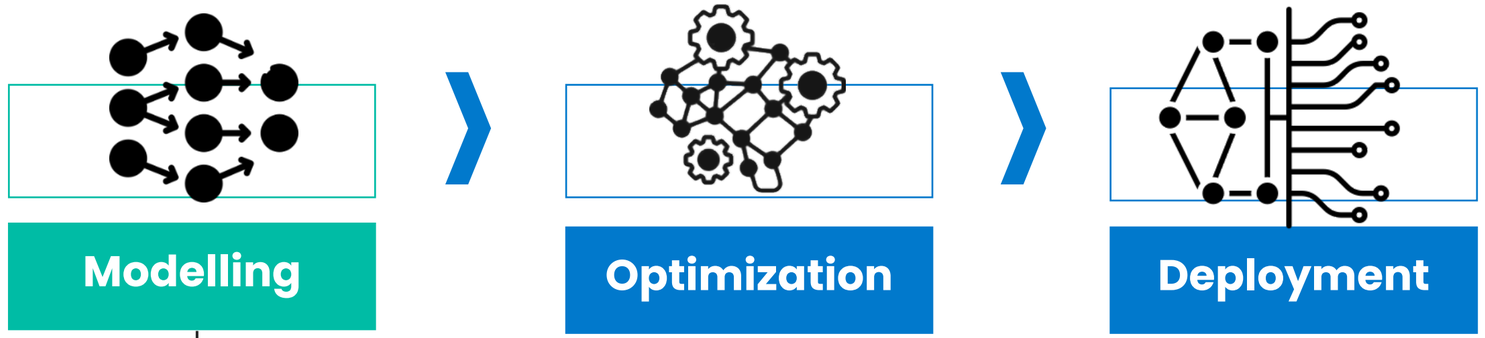
A Complete Toolchain from NN modeling to optimized deployment on Hardware



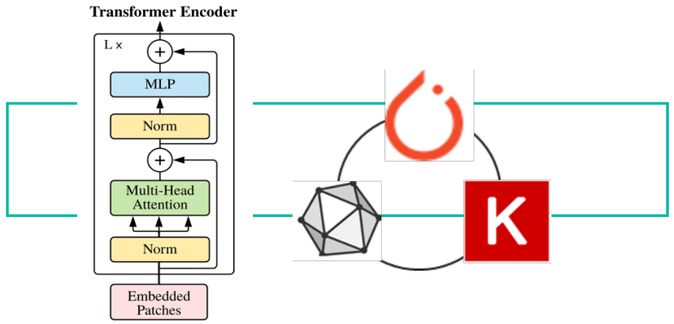
cea

Deep Learning Platform for Embedded

aidge

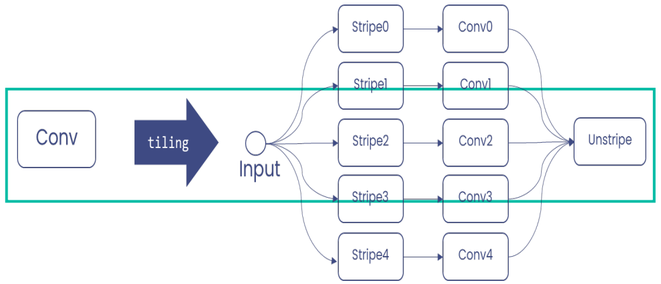


Interoperable data flow graph



State of the art models : CNN
RNN, GAN, Attention
ONNX import

Powerful graph manipulation



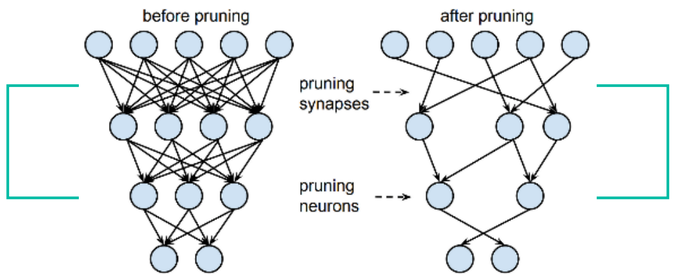
Tiling, Graph search and
replace engine

cea

Deep Learning Platform for Embedded

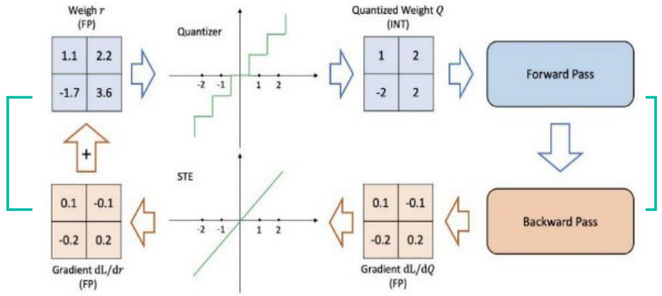
aidge

Post Training Optimization



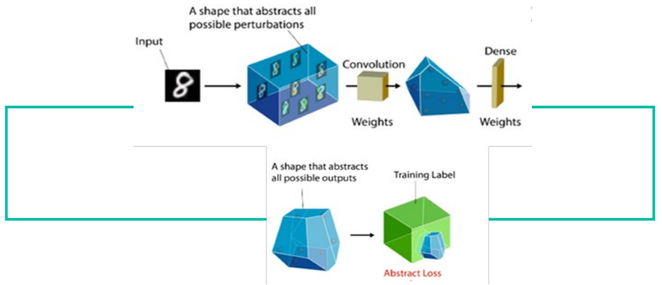
Quantization, Pruning, Compression

Quantization Aware Training



Innovative SotA QAT based on SAT and LSQ

Robust approaches for learning and inference

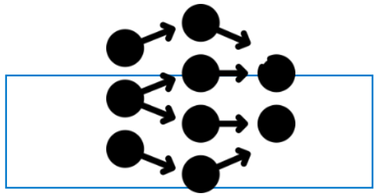


Adversarial attack, Incremental learning

cea

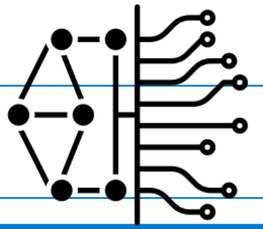
Deep Learning Platform for Embedded

aidge



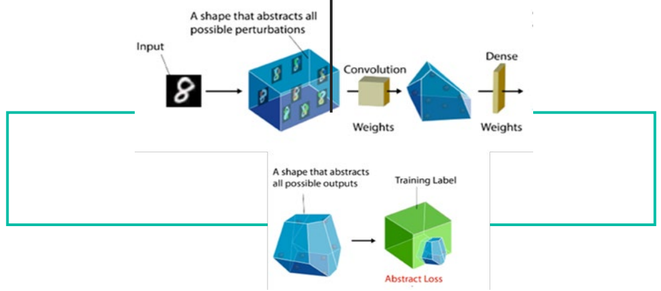
Modelling

Optimization



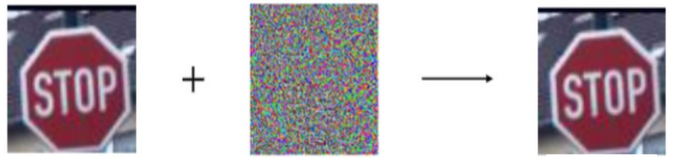
Deployment

Robust approaches for learning and inference



Adversarial attack,
Incremental learning

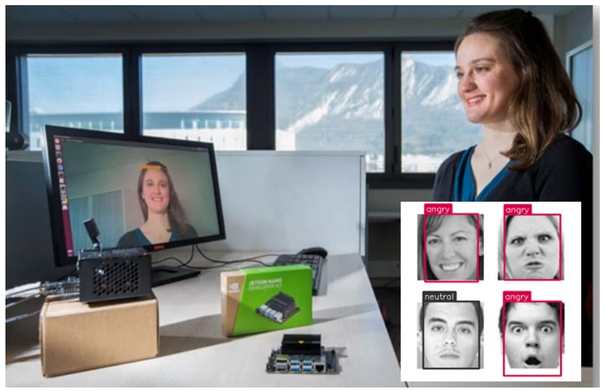
Adversarial Attack



Stop noise

Combination of Sota Methods without latency loss

Incremental learning



Publication

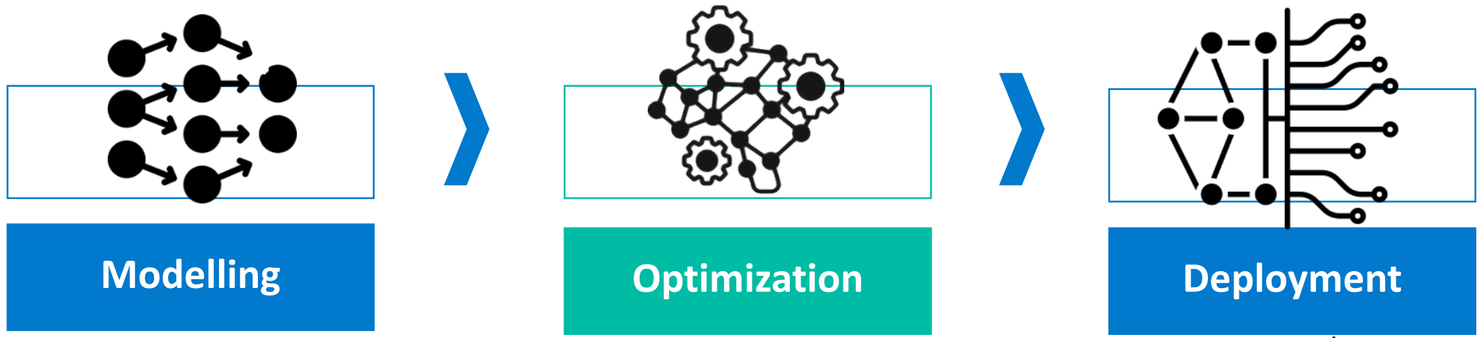
An Embedded Continual Learning System for Facial Emotion Recognition

O. Antoni, M. Mainsant, C. Godin, M. Mermillod, and M. Reyboz **@ Demo track ECML 2022**

cea

Deep Learning Platform for Embedded

aidge



Efficient cross code generation

C++ NVIDIA CUDA cuDNN OpenMP HLS

Generation engine based on template

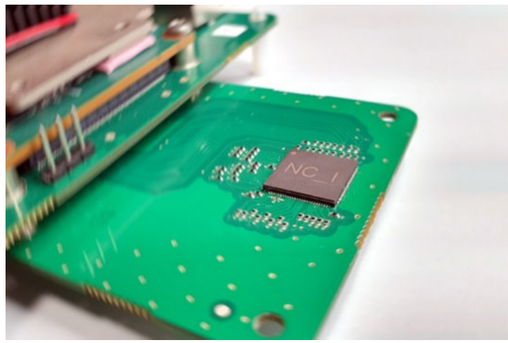
Code execution for multiple hardware

MCU, CPU, GPU, NPU, ASIC, FPGA

etc ...

COTS and sovereign hardware targets

Hardware design



AI-ASIC Neurocorgi

cea

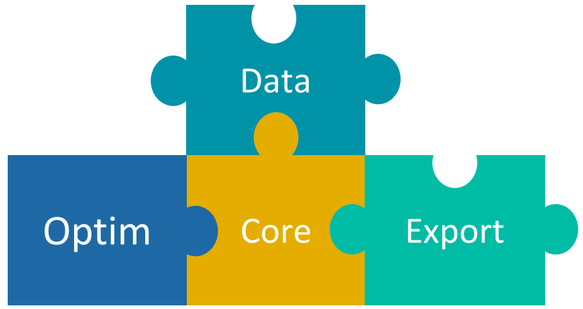
Deep Learning Platform for Embedded

aidge

Open source software platform for collaborative dynamic, code transparency – hosted by Eclipse Foundation

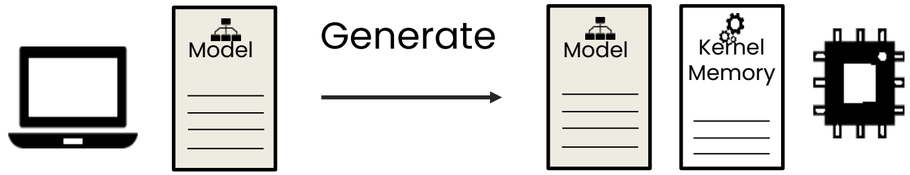
ECLIPSE FOUNDATION

Modular and extensible framework
with minimal set of dependencies and appropriate programming language (C++/Python)



Supporting **multiple hardware targets** and envision heterogeneous architectures

Producing **human readable code** with unified representation wrt to the design



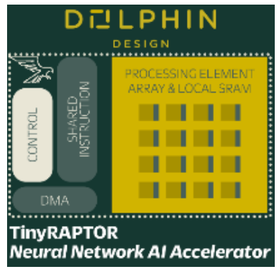
cea

Zoom on optimized deployment

aidge

SDK entirely based on aidge (prev. N2D2)

- CEA's specialized circuit for neural networks
- Data reuse >90%



- Tiny MLPerf benchmark results
  - 32µJ and 10ms latency on *Visual Wake Words*
  - 12µJ and 3,5ms latency on *Mlcommons* keywords

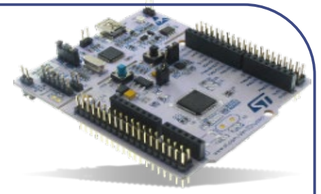
Embedded World Award 2022

embeddedworld2022
Exhibition&Conference
... it's a smarter world

HW export: Tiny Raptor NPU

Multi precision quantization

- 8-bit, 4-bit, 1-bit
- 2x lower memory usage with 4-bit export vs 8-bit (+10% inference latency)
- 1-bit export prototypes (+5% latency)



Mixed precision export

- Best results with layer-wise precision optimization

HW export: STM32 MCU

Environment constrains

- High speed rolling at 20m/s
- Tiny defect (~mm), low contrast

Solutions

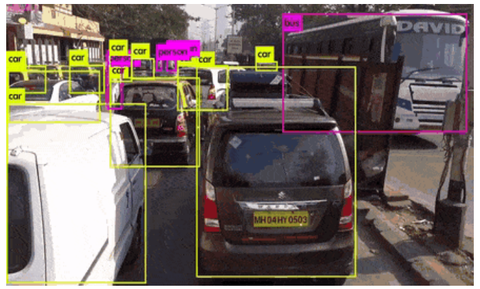
- Data augmentation
- Fast neural network exploration
- Performance vs complexity tradeoff analy

Currently tested on production line

HW export: NVidia TensorRT GPU

Use case: metal coil defect detection

Urban detection algorithm optimization from 8 frames per second with Tensorflow to **25 FPS after export**



HW export: NVidia TensorRT GPU

Usage case: urban detection

cea

aidge Ecosystem

NEUROKIT2E

STAKEHOLDERS

DEEPGREEN

Forum TERATEC 24

Thank you!