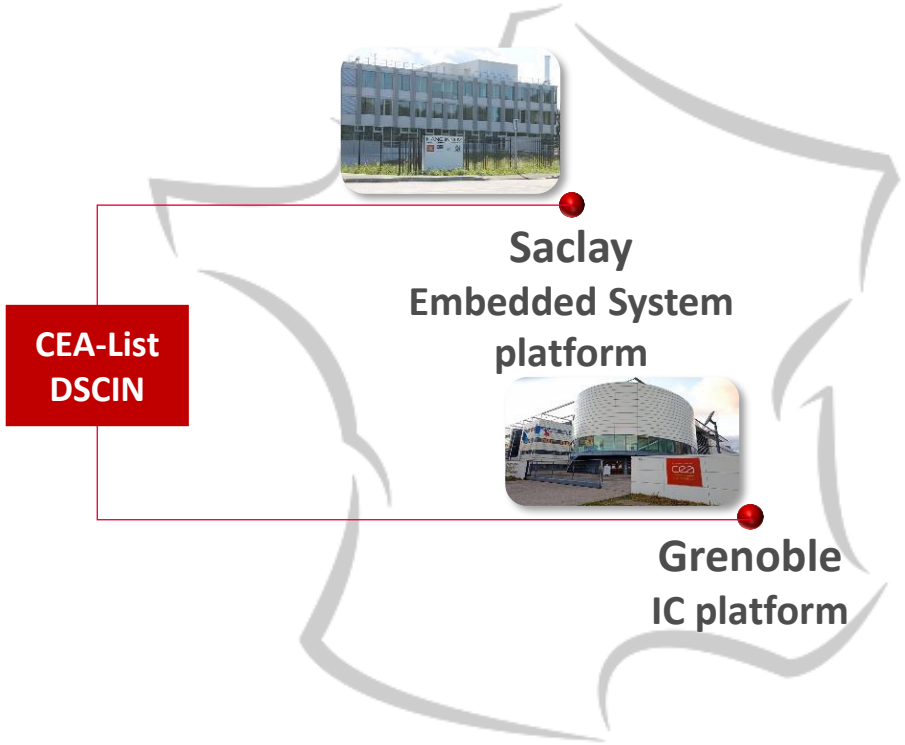
Les enjeux de l'IA embarquée

Olivier Bichler

olivier.bichler@cea.fr

Sommaire

DSCIN: Digital System & Integrated Circuit division



200 Staff members

2 500 m^2 of facilities

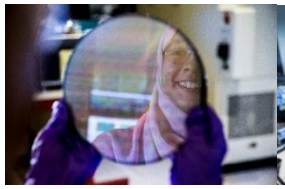
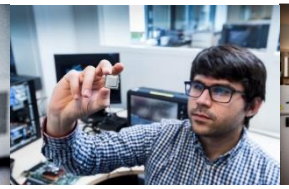
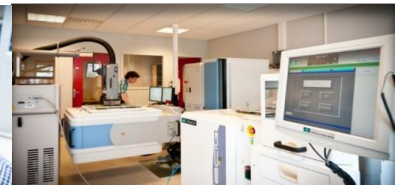
30 Industrial partners

10 IC designs per year

25 Patents granted per year

100 Publications per year

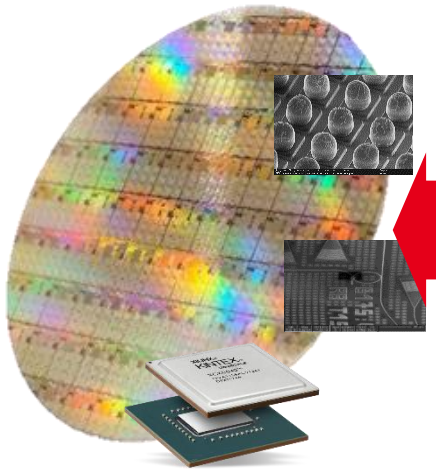


CEA Digital system design

Eco-innovation

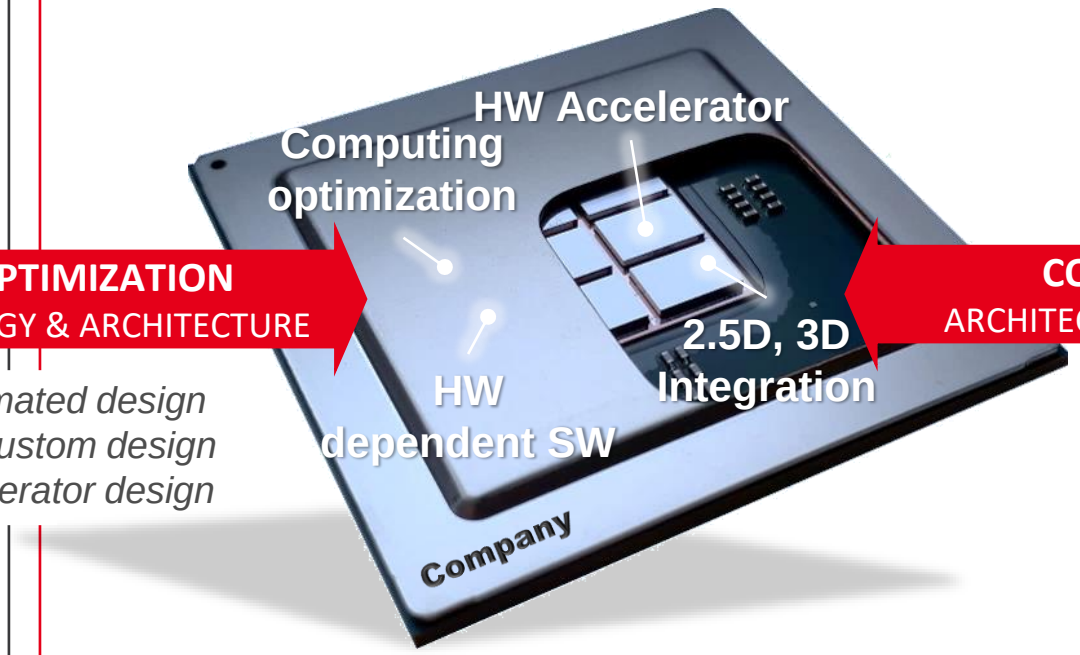
Semiconductor technologies



CO-OPTIMIZATION
TECHNOLOGY & ARCHITECTURE

- *Automated design*
- *Full custom design*
- *Accelerator design*

A 360° APPLICATION DESIGN



CO-DESIGN
ARCHITECTURE & SYSTEM

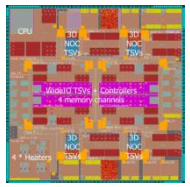
- *Algorithm*
- *Software stack*
- *Data coding*

Smart computing systems

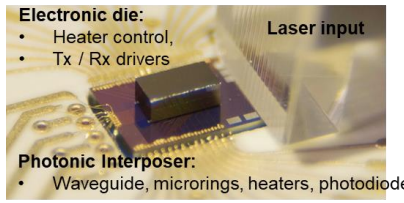
CHARACTERIZATION ← DESIGN ← EXPLORE ← MODEL & SIMULATE

10+ years of experience in Chip design

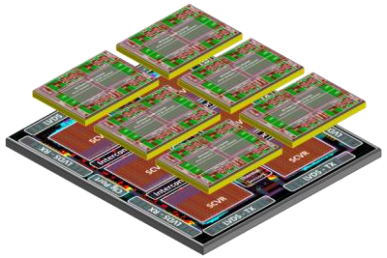
HIGH EFFICIENCY



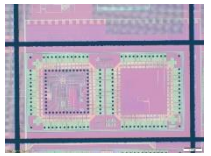
MAG-3D
3D Network-on-Chip



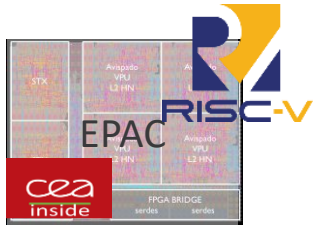
HUBEO
Photonic NoC interposer



INTACT
6 chiplets & 96 processors

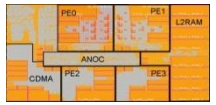


CRYOCMOS
Control for quantum computing

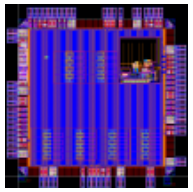


EPAC
HPC Variable Precision Accelerator

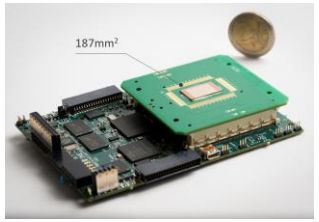
TRUST-WORTHY



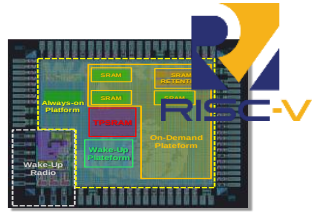
LOCOMOTIV
Adaptive Voltage & Frequency Scaling



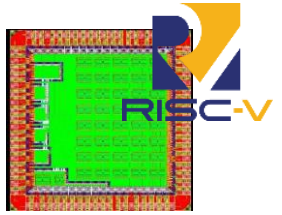
FRISBEE
ULP FDSOI



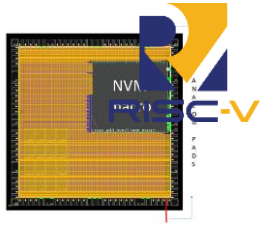
RETINE
Ultra-fast smart imager



WARRIOR
RISC-V IoT IC with wake-up

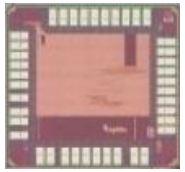


Cyber-VT
Test Vehicle for IoT security enhancement

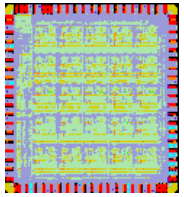


Non-Volatile-Memory
NVM subsystem for Microcontrolers

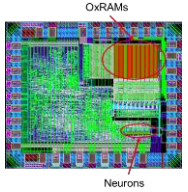
AI



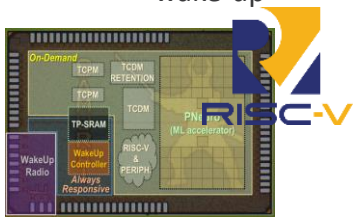
REPTILE
Analogue neuron



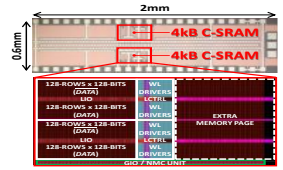
SPIDER
Neuromorphic DSP



SPIRIT
Spiking NN with eNVM



SAMURAI
IoT IC with NN accelerator

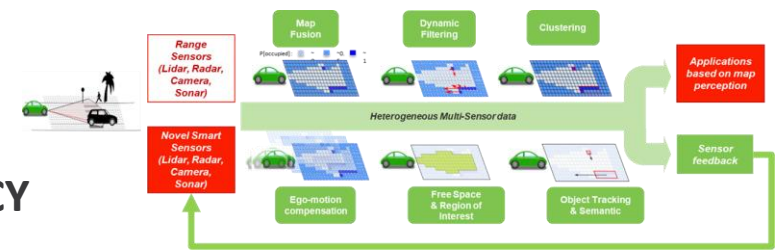


In-Memory-Computing
Compute-SRAM

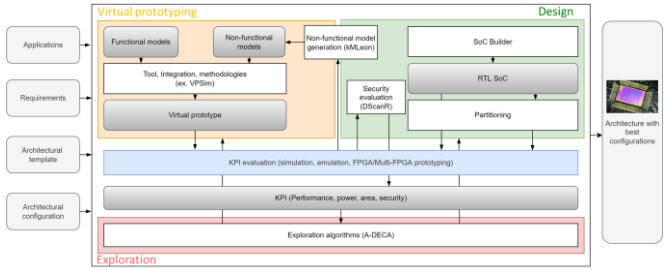
ESPERANTO
RNN with 50k synapses

10+ years of experience in Systems co-design

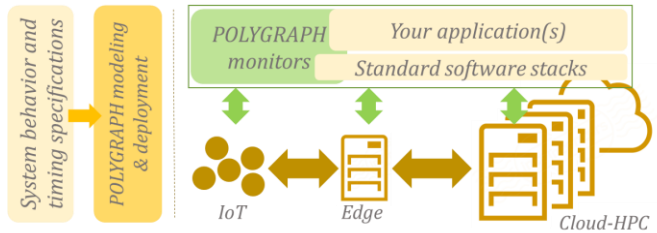
HIGH EFFICIENCY



SIGMAFUSION – ULP sensor fusion

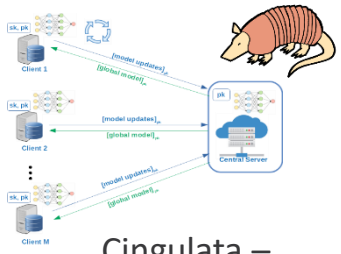


VPSIM – HW Digital Twin



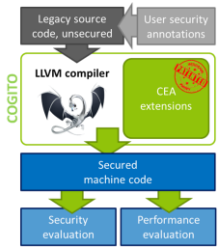
Polygraph- Edge2Cloud

TRUST-WORTHY

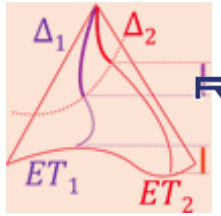


Cingulata – cryptocomputing

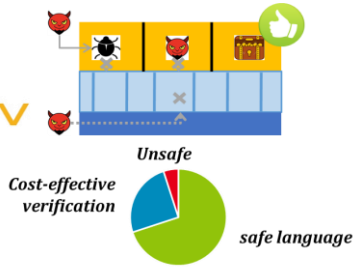
MoCC – Communication & computing QoS



Cogito – SW Cyber



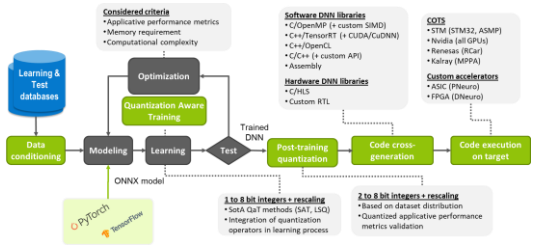
Leaf – real time reliability



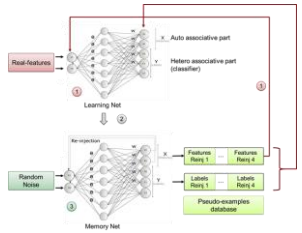
XANTHOS µKernel

AI

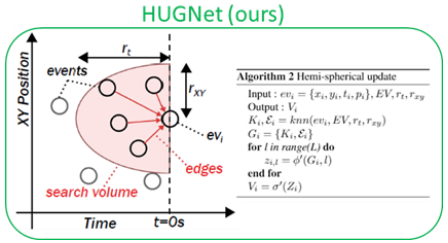
aidge



AIDGE embedded AI



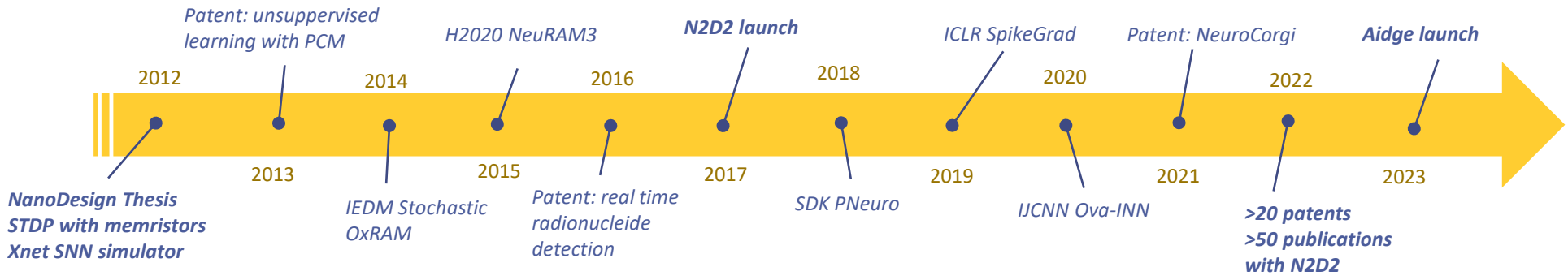
DreamNet - Incremental Learning



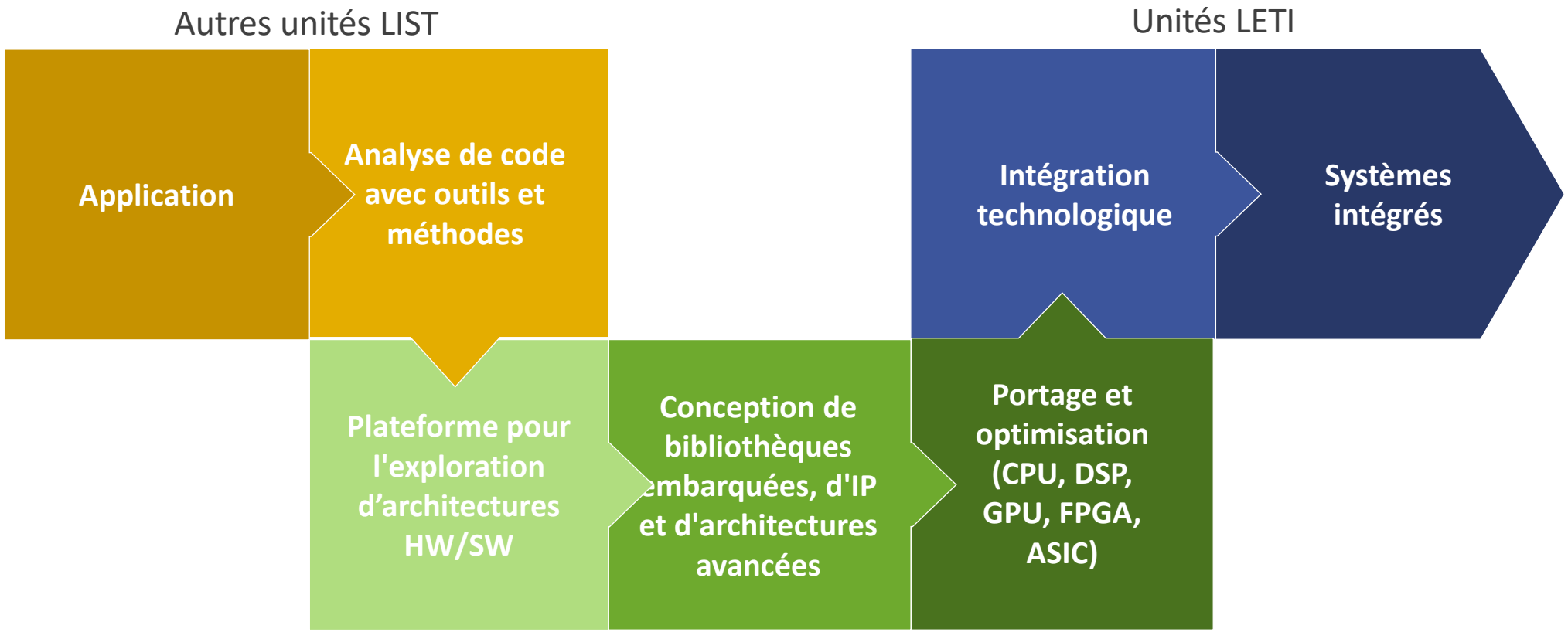
ULP GNN

Les origines du LIAE

Laboratoire Calcul Embarqué et Image	Laboratoire Calcul Embarqué	Laboratoire Adéquation Algorithme Architecture	Laboratoire Intelligence Artificielle Embarquée
Jusqu'en 2009	*2010 - 2014*	*2015 - 2019*	*Depuis 2020*
DTSI/SARC/LCEI	**DACLE/LCE**	**DACLE/SCSN/L3A**	**DSCIN/LIAE**
Mission : Concevoir des solutions HW/SW pour des systèmes embarqués complexes basés sur l'image. **Les solutions sont principalement "classiques" avec un peu de recherche sur les réseaux neuronaux profonds.**	**Mission : Concevoir des solutions HW/SW pour des systèmes embarqués complexes.** **Séparation de l'équipe de traitement d'images de haut niveau avec la création du DIASI/LVIC.**	**Mission : Concevoir des solutions HW/SW pour des systèmes embarqués spécifiques à une application complexe.** **Séparation de la conception de l'architecture HW "générique" qui a créé le nouveau LCE.**	**Mission : Concevoir des solutions HW/SW pour des systèmes embarqués d'IA complexes.** **Intégration de l'activité des réseaux neuronaux profonds à la mi-2019.**



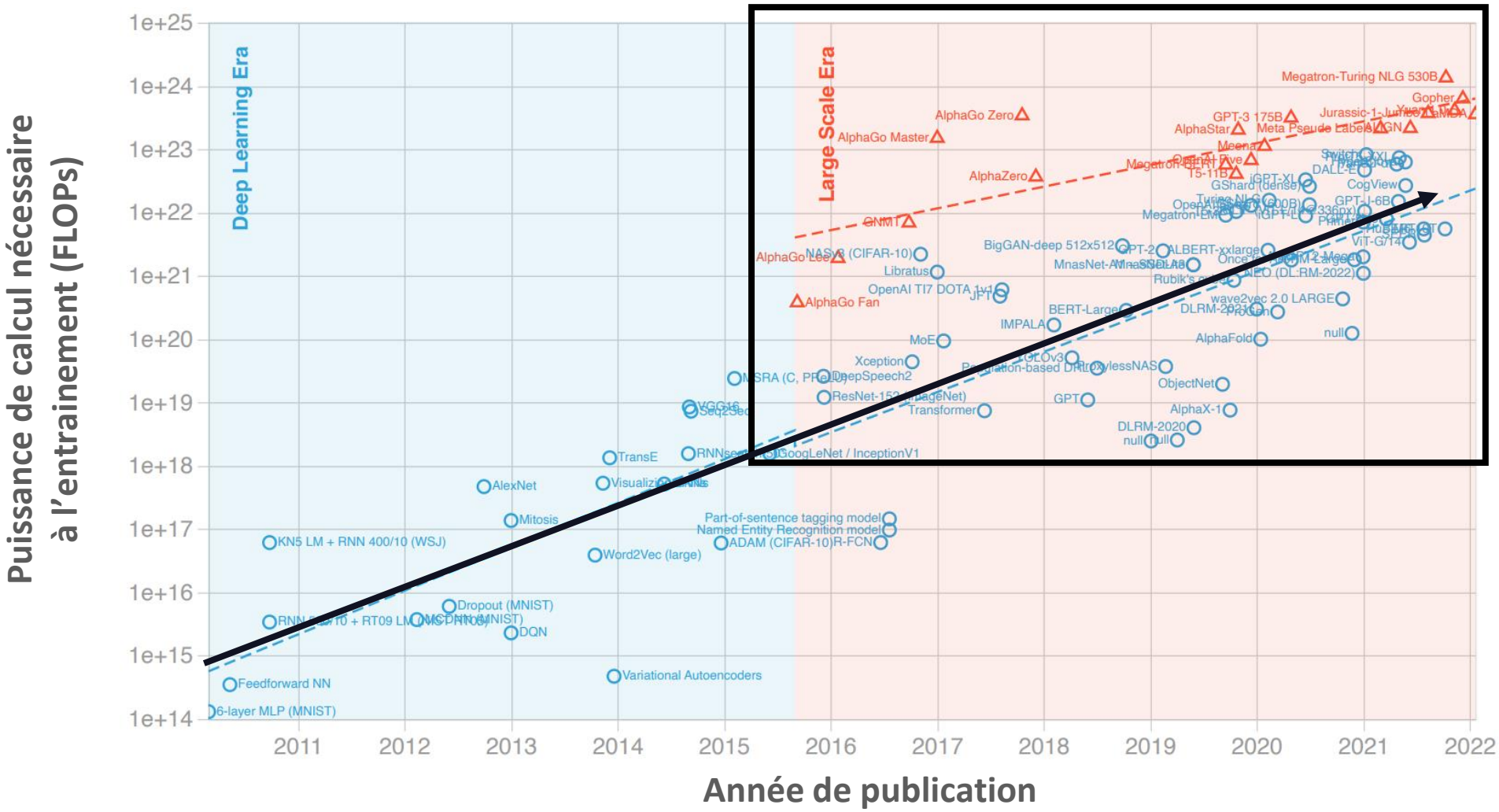
LIAE est un joueur d'équipe clé



LIAE : Combler le fossé entre le développement d'applications et l'intégration technologique

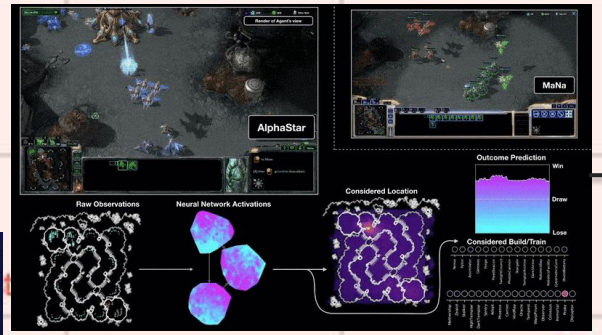
Sommaire

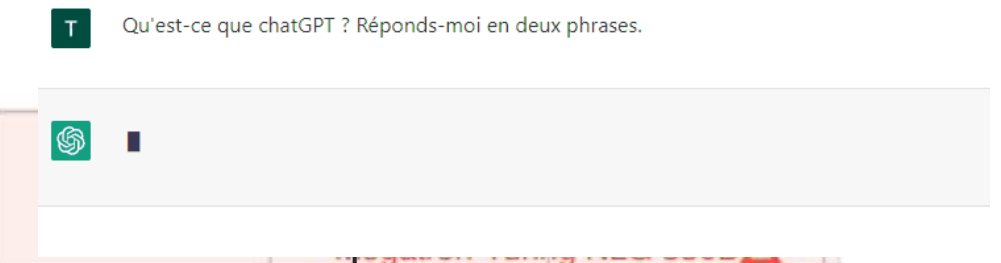
L'essor du *deep learning* et de l'IA

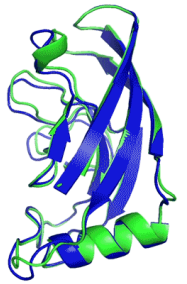


L'IA est largement dominé par le *deep learning*

L’essor de l’IA





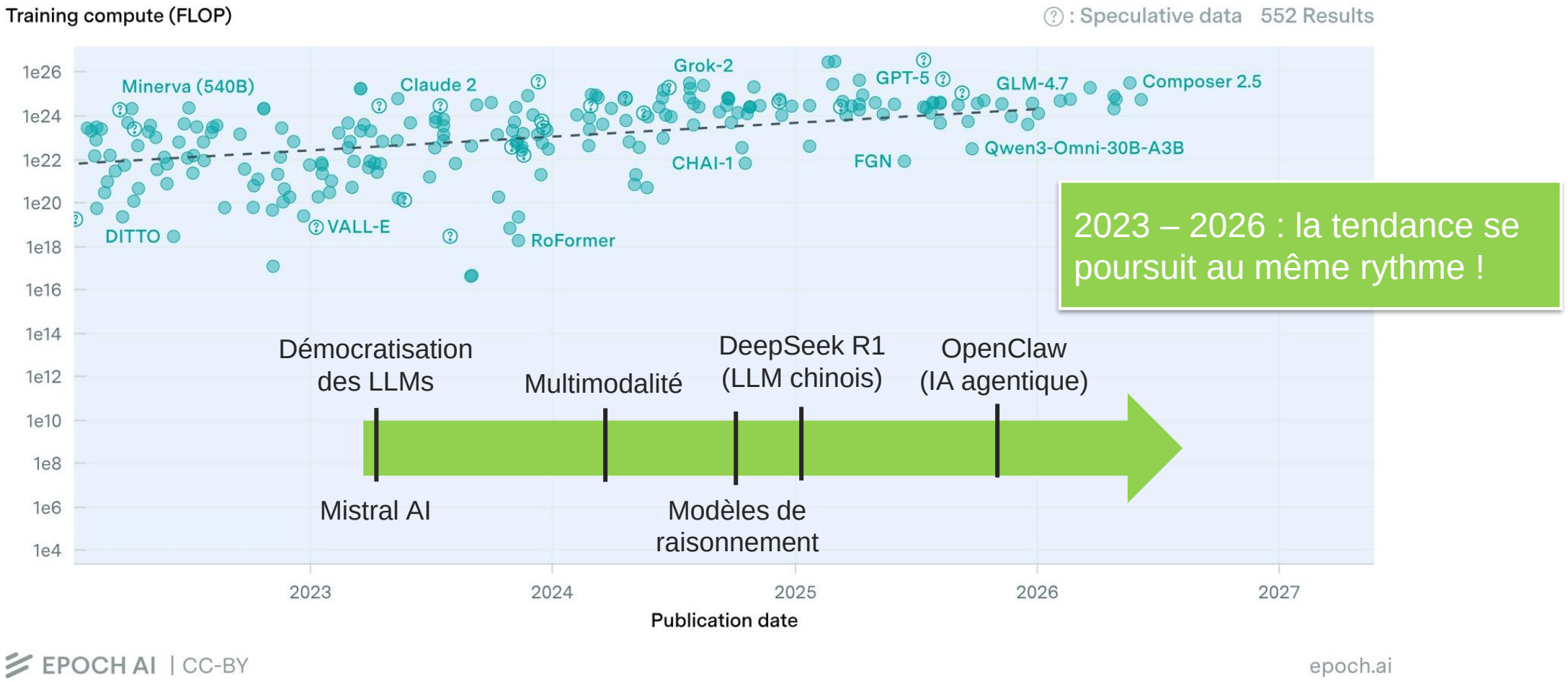


T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

L’essor de l’IA



2023 – 2026 : la tendance se poursuit au même rythme !

Un développement aujourd'hui exponentiel

Bank of America: "It took ChatGPT just 5 days to reach 1 million users, 1 billion cumulative visits in 3 months and an adoption rate which is 3x TikTok's and 10x Instagram's,".
"The technology is developing exponentially."

Marché IA adressable selon SpaceX d'ici 2030 : **26 000Md**
CA visé en 2030 : 322Md

Croissance globale du marché de l'IA : de 200Md en 2023 à 2 000Md[1]

- **de l'edge IA : de 10Md à 100Md[2] en 2030 ➔ x10**

Explosion des investissements

- Microsoft / OpenAI : 10Md en 2023, supercalculateurs à 10k-100k GPUs, entre 100k-1M GPUs déployés[4]
- Google : des douzaines de supercalculateurs à 4k TPUs v4[5]
- Twitter : 10k GPUs
- France/GENCI : Jean Zay (3k GPUs), CEA-List/Factory-IA (1k GPUs)…

xAI Colossus : 200k H100 GPUs

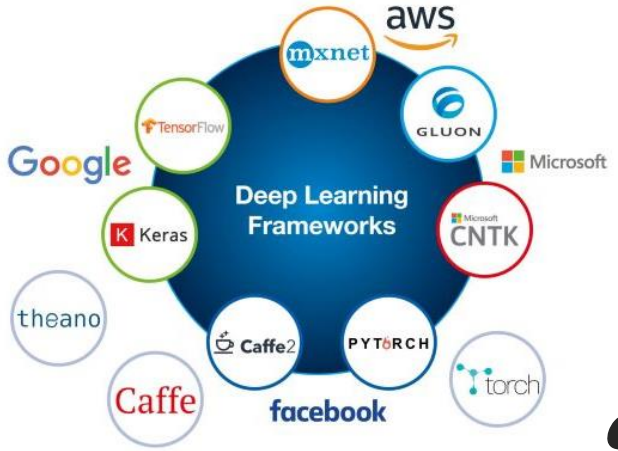
Mistral :
- 2026 : mise en service d'un centre de données à Bruyères-le-Châtel
- 1 GW de capacité en 2030

L'IA pourrait contribuer à 20% de croissance du PIB de l'Europe en 2030, soit 3 000Md[3]

Soutenu par des plateformes *open source*

Des outils performants et accessibles

Les US dominent le marché avec des plateformes de deep learning largement utilisées dans le monde <u>positionnées dans le cloud aujourd'hui</u>



GAFAM

INVESTISSEMENT PUBLIC :

- *National Science and Technology Council*
- *DARPA Defense Advanced Research Projects Agency*
- *Advanced Research Projects Agency-Energy*

La Chine entre dans la course avec ses propres plateformes de deep learning issues des BATX. Les autorités chinoises investissent aussi massivement sur le sujet

Baidu 百度

Plan AI+ lancé en 2024

- Modèles de fondation
- Infrastructure souveraine
- **Processeurs IA souverains**

15 000 robots humanoïdes produits en 2025 (90% de la production mondiale)

BATX

INVESTISSEMENT PUBLIC :

- *Les autorités chinoises ont annoncé qu'elles investiraient jusqu'à 59 milliards de $ d'ici à 2025 dans l'objectif de rattraper les États-Unis*

Solutions profondément orientés **déploiement GPU/Cloud** (quasi-monopole NVidia)

Jusqu'où ?

L'IA pourrait représenter plus de la moitié du *cloud computing* en 2030

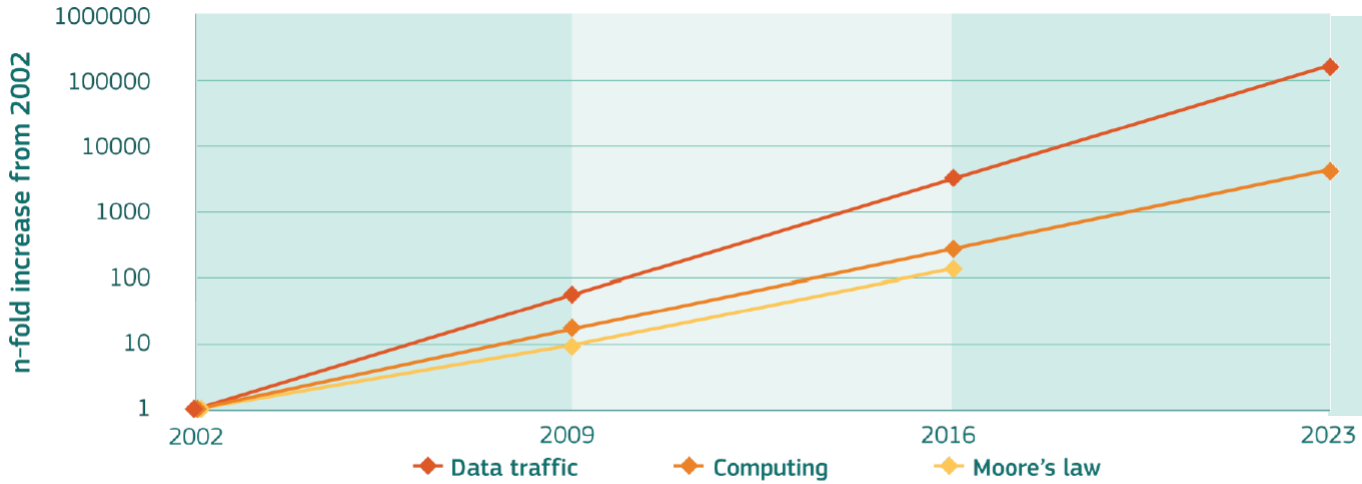


Figure 13. Projection of data and computing growths (logarithmic scale)
Source: JRC based on Kambatla et al., 2014

*Le Cloud = **3 à 4%** de la consommation d'énergie en Europe en 2023[1]*
(½ transmissions de données + ½ calcul)

5,7% en 2035

USA :
4,4% en 2023
9% en 2030 *(Electric Power Research Institute (EPRI))*

~2036
Limite quantique de la loi de Moore[2]

*The expected growth in data, computing processing, and energy consumption are changing the present paradigm "do everything in the cloud" to a new one that can be called "**bring computing to where the data are**".*

EUR 29245 - Artificial Intelligence: A European Perspective

L'edge IA, une nécessité pour **passer à l'échelle et diminuer le coût** !

Vers l'essor de l'IA embarquée

Le déploiement, du Cloud vers l'*edge* :

- Maturité des applications
- Passage à l'échelle
- Diminution du coût
- Interaction avec le monde physique

- Souveraineté
- Confidentialité

Taalas (start-up à Toronto)
Llama sur silicium atteint
17 000 jetons/sec !
10x meilleur que le SotA

Limite quantique de la loi de Moore[2]

Cloud IA non soutenable

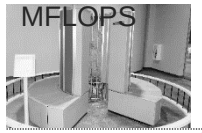
Cloud IA : développement
(apprentissages large échelle, génération de circuits)

Ere des circuits « câblés »

Edge IA : déploiement
(inférence & adaptation)

Smart Cities | Industrial Robotics | Autonomous Vehicles | Warehouse Robots
Drones | Robotic Cleaners | Augmented Reality | Smart Agriculture

Cray-1
160
MFLOPS



IBM ASCI White
7.23 TFLOPS

ARM Cortex-M4
250 MOPS

Apple A16
17 TOPS

1985 | 2000 | 2010 | 2022 | ~2036

Puissance de calcul (TOPS) – *échelle log*

Vers l’essor de l’IA embarquée

MINISTÈRE
DE L'ÉCONOMIE,
DES FINANCES
ET DE LA SOUVERAINETÉ
INDUSTRIELLE ET NUMÉRIQUE
Liberté
Égalité
Fraternité

La tendance de fond du développement du Edge computing, ou Intelligence Artificielle Embarquée, représente une valeur potentielle considérable estimée entre 175 à 215 milliards US$ en 2025 pour la seule composante hardware.

« L’IA embarquée promet donc
défense, agriculture…), et repre

Une prévision commode mais « trop » visionnaire ?

Des catalyseurs forts :

- Capacités autonomes dans la défense (guerre en Ukraine, instabilités…)
- Robotique humanoïde industrielle et domestique (Plan AI+ Chine, 1X, Tesla…)
- Banalisation des NPUs compatibles LLM/VLM *low cost* (notamment chinois : Sophon, SpaceMIT, RockChip, Huawei…)

L’USINE NOUVELLE

« L’edge computing est une plus grosse opportunité que le cloud », Thierry Breton, PDG d’Atos (2019)

« Le client nous demande de les traiter dans les datacenters […] on arrive à un volume de 40 zetaoctets, soit 40 000 milliards de milliards de données. Selon différents cabinets, 80% vont être traités de façon centralisée sur les datacenters et le cloud, et 20% ailleurs, à la périphérie ou l’edge […]. A l’horizon de cinq ans, ce seront 80% à traiter dans l’edge et 20% dans les datacenters et le cloud. »

Pourquoi embarquer l'IA ?

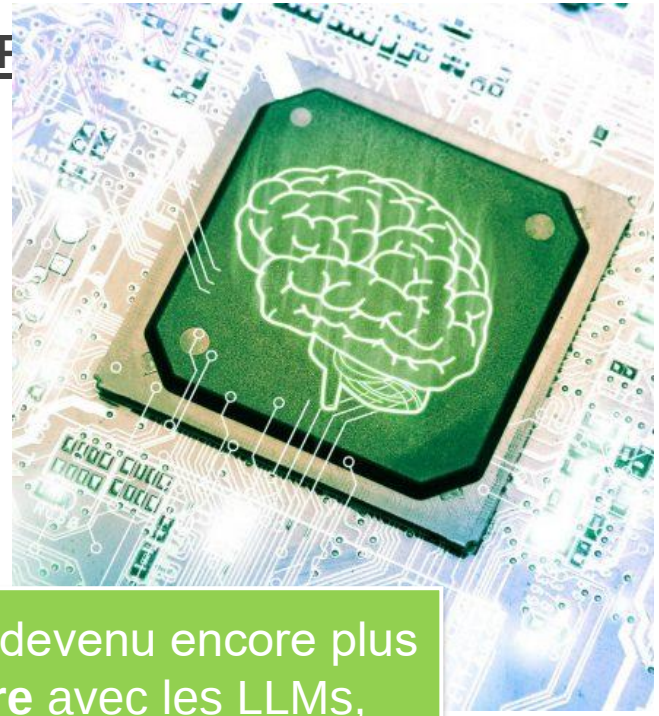
Motivateurs :

Passage à l'échelle & réduction des coûts (transmission, infrastructure…)

Temps réel « dur »
(temps de réponse garanti)

Sureté de fonctionnement
(disponibilité garantie)

Confidentialité des données
(si on est obligé…)



Fiabilité : capacité à garantir les résultats
(explicabilité, robustesse…)

Frugalité : capacité d'adaptation avec peu de données
(*zero-/one-short learning, continual learning…*)

Oui, c'est devenu encore plus **nécessaire** avec les LLMs, et ce à tous les niveaux (états, industries, individus)

Importance de la mémorisation du contexte pour les grands modèles

Quelles solutions pour l'IA embarquée ?

Des outils aux orientations et dépendances fortes

Hexagon

Qualcomm snapdragon

android

NNAPI

SAMSUNG Exynos

iOS

TensorFlow Lite

TensorFlow Lite *for Microcontrollers*

Adreno GPU

Qualcomm snapdragon

android

AMD XILINX

Vitis AI

TensorRT

Accelerator

- *…veau opérateur*
- *…ilage de boucles*
- *Backend LLVM*

ARM Cortex-M

micro tvm

Montée en puissance d'**ExecuTorch** (Meta), lancé en août 2023
➔ aujourd'hui le projet avec la plus forte croissance !

CoreML

android

NNAPI

intel

OpenVINO, oneDNN

TensorRT

AMD XILINX

Vitis AI

Qualcomm snapdragon

SNPE SDK

arm

ARM NN

AMD ROCm

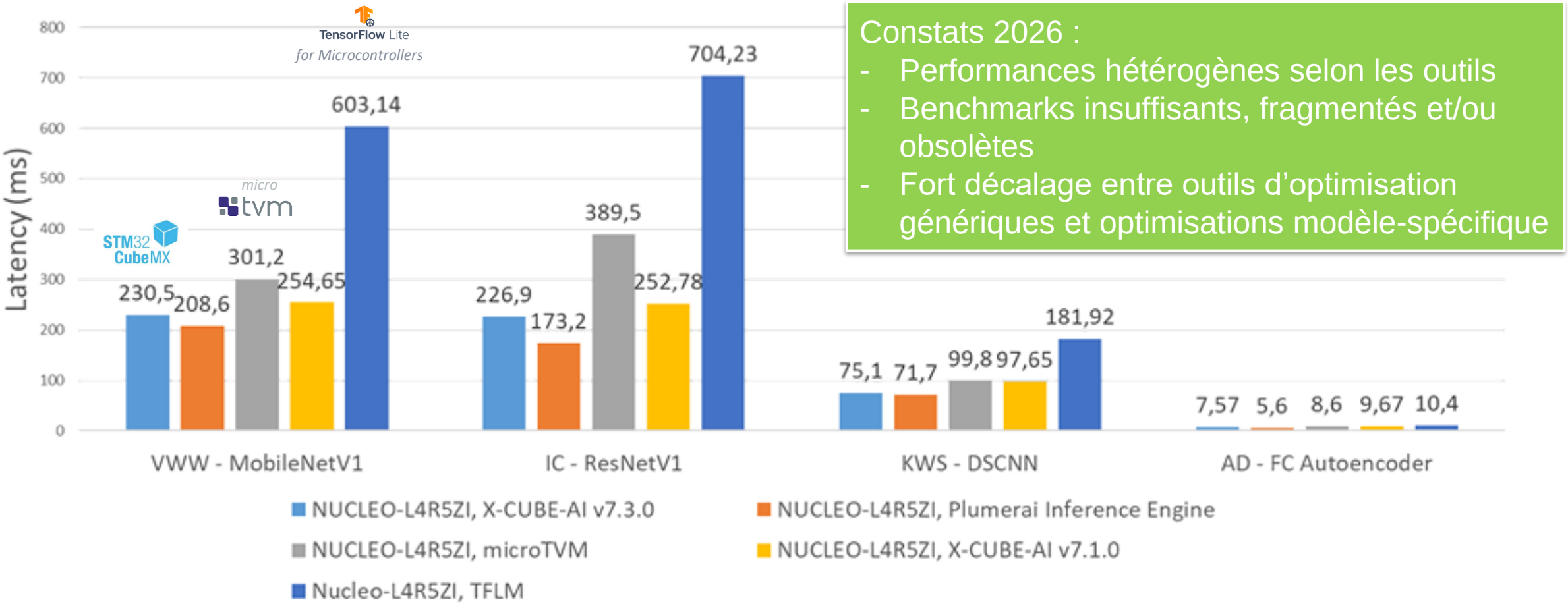
ONNX RUNTIME

- *Wrapper sur librairies propriétaires*

Développement continu d'ONNX et ONNX Runtime

Quelles solutions pour l'IA embarquée ?

Pas forcément optimisés : benchmark 2023 **MLPerf Tiny** sur **Cortex-M4**



Constats 2026 :

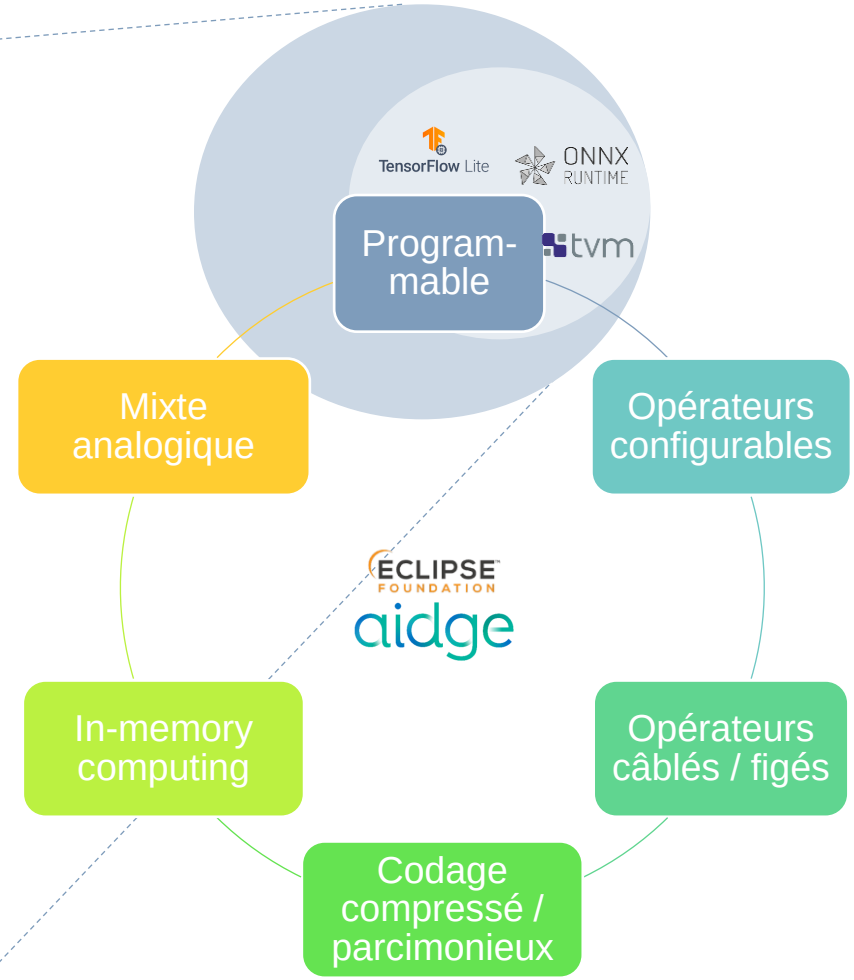
- Performances hétérogènes selon les outils
- Benchmarks insuffisants, fragmentés et/ou obsolètes
- Fort décalage entre outils d'optimisation génériques et optimisations modèle-spécifique

Quelles solutions pour l'IA embarquée ?

Contrastant avec la diversité des solutions et architectures matérielles

Processor	Company	Year	Process
ERGO	Perceive	4Q20	22nm
Recogni	Recogni Inc.	4Q22	7 nm
ARC NPX6-96K	Synopsys		
AiWare 4	AiMotive		
AiWare 3	AiMotive		
NeuPro-M	CEVA		
NeuPro-S NPS4000	CEVA		
Tyr 3	Vsora		
Tyr 1	Vsora		
AD1028	Vsora		
Origin E4	Expedera		
RAPID	IBM		
TruStream	Cornami		
NNE 110	Cadence		
NEOX A	Think Silicon		
SAKURA DNA-A800	EdgeCortix		
Journey 6	Horizon Robotics		

Processor	Company	Year	Process
Journey 5	Horizon Robotics	4Q21	
Ethos-N78	ARM	2Q21	7 nm
Hailo-8	Hailo	2Q20	16 nm
NMP 700	AiM Future	1H23	
Tiny Raptor v2	Dolphin	1H22	22 nm
Chimera QB4	Quadric	1Q23	7 nm
Q16	Quadric	3Q21	16 nm
Myriad X	Intel Movidus	4Q18	16 TSMC
Infer X X1	Flex Logix	1Q23	16 nm TSMC
WARBOY	FuriosaAI	1Q23	14 nm
ARC EV74	Synopsys	1Q20	16 nm

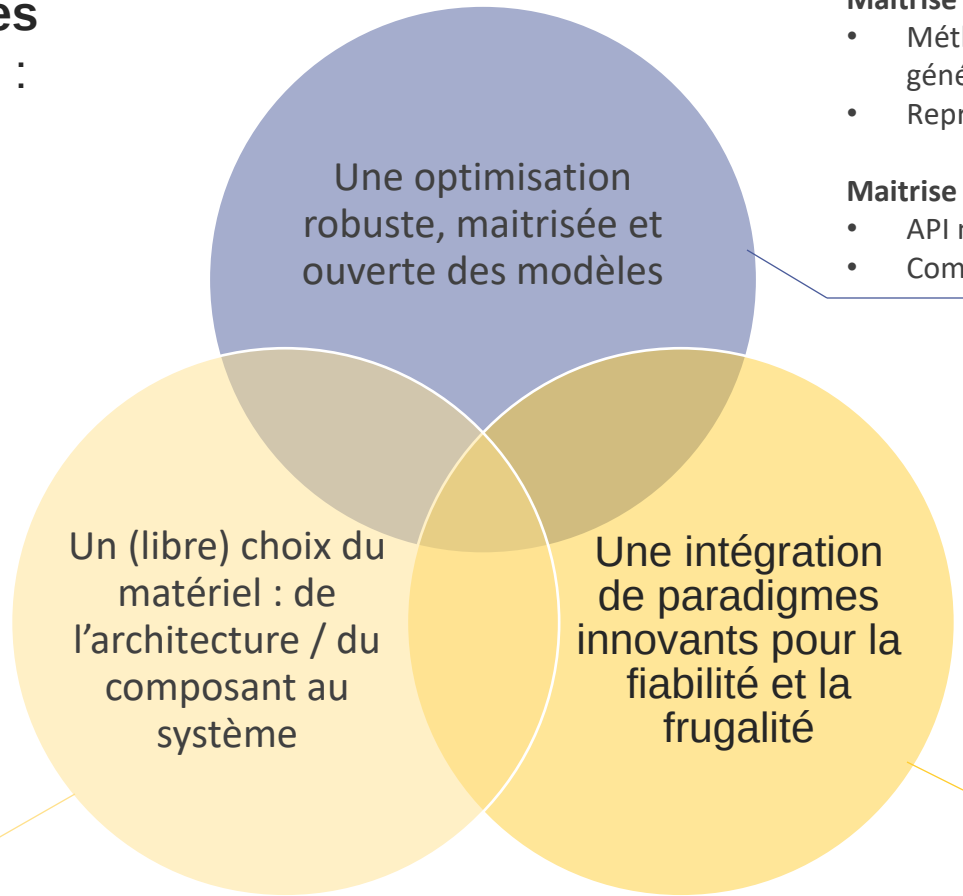


Processor	Company	Year	Process
Spiking neural Processor			
Loihi 2	Intel	4Q21	Intel 7
In memory DLA			
AI-X Lightspeeur 2803S	Gyrfalcon Technology	4Q20	28nm
MX3	MemryX	2H22	
SpeedAI240	Untether AI	3Q23	7 nm TSMC
RunAI200	Untether AI	2Q21	16 nm
Analog DLA			
AML100	Aspinity	4Q22	
M1076	Mythic	1Q22	40 nm
GPX-10	Ambient	2Q21	40nm
Passage	Lightmatter	4Q22	
Mars	Lightmatter	4Q21	90nm (photo die) 14nm (dig. die)
Aurora DeepLight 1.0	Cognifiber	2H23	

Les enjeux pour l'IA embarquée

Au-delà des **performances, indispensables**,
de **multiples enjeux techniques**
pour assurer notre compétitivité :

- *Maitrise*
- *Souveraineté*
- *Innovation*



Maitrise des algorithmes :
- Méthodes ouvertes, documentés, génériques et réutilisables
- Reproductibilité des modèles optimisés

Maitrise du déploiement :
- API modulaire et unifiée pré- et post-déploiement
- Compatible avec des objectifs de certifiabilité

Hétérogénéité :
- Multi-cibles matérielles
- Gestion de systèmes hétérogènes
- Multi-paradigmes de calcul

Souveraineté :
- Noyaux de calcul ouverts et indépendants
- Architectures et composants souverains
- Intégration de technologies souveraines (FDSOI, 3D-chiplet, mémoires NVM)

Fiabilité :
- Intégration de connaissances « innées »
- Maitrise du « domaine de vol »
- Robustesse aux fautes et aux attaques

Frugalité :
- Transférabilité de modèles « maitres »
- Adaptation hors ligne et en ligne

Sommaire

Convolutional neural networks overview

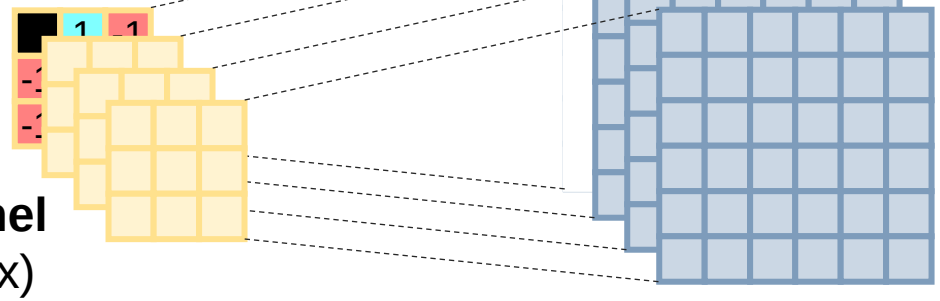
CNN layer:

Input map
($I_{i,j}$ matrix)

$$O_{i,j} = \tanh\left(\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} I_{i+k,j+l} \cdot K_{k,l}\right)$$

Output feature map
($O_{i,j}$ matrix)

0	0	0	0	0	0	1	-1
0	0	0	0	0	0	1	-1
0	0	0	0	0	0	1	-1
0	0	0	0	0	0	1	-1
0	0	0	0	0	0	1	-1
0	0	0	0	0	0	1	-1
-1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	1	0



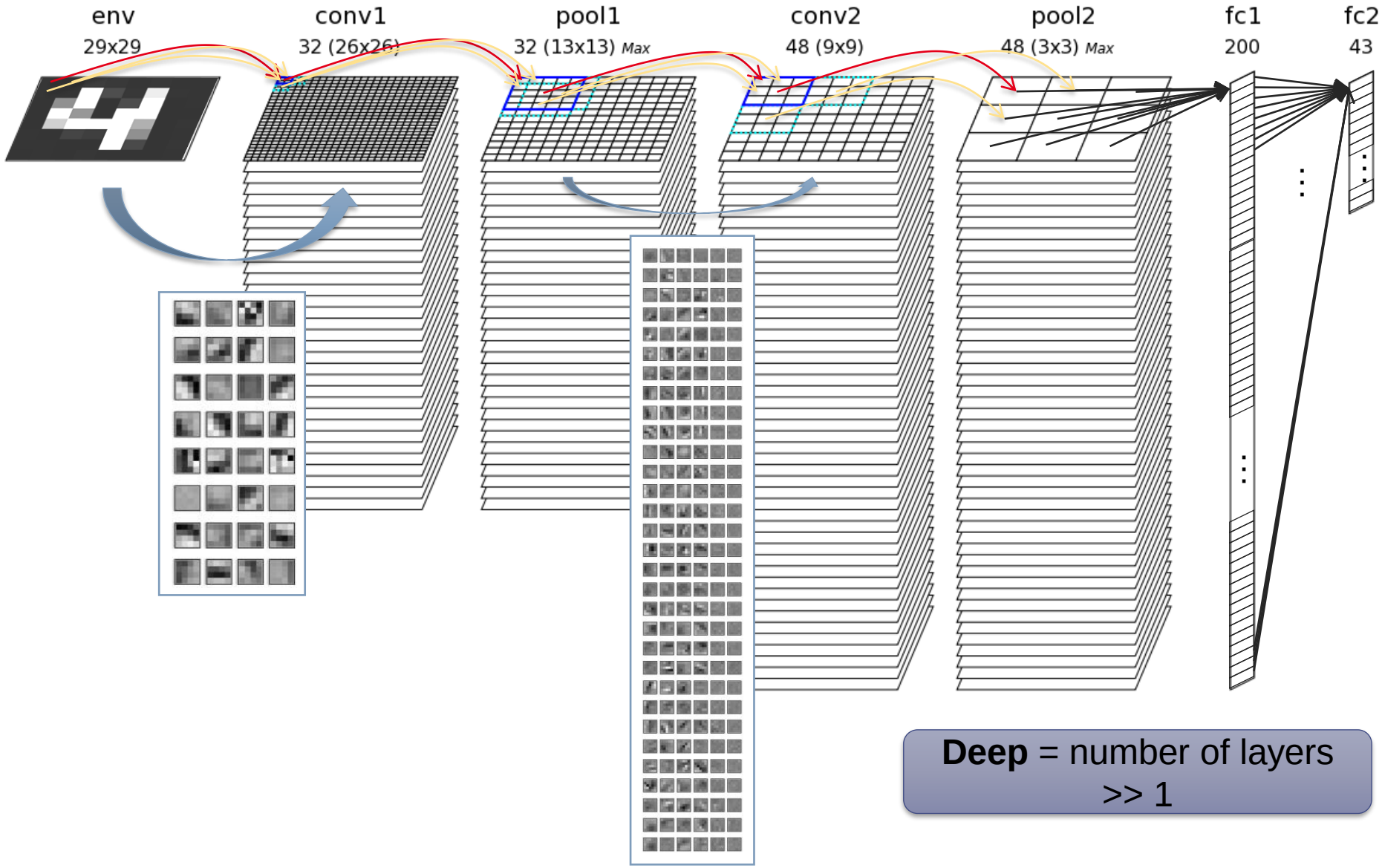
$n \times n$ kernel
($K_{k,l}$ matrix)

Each kernel generates ≠ output feature maps

Convolution operation: $O_{i,j} = \tanh\left(\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} I_{i+k,j+l} \cdot K_{k,l}\right)$

Kernels are learned with gradient-descent algorithms
(classical back-propagation is very efficient!)

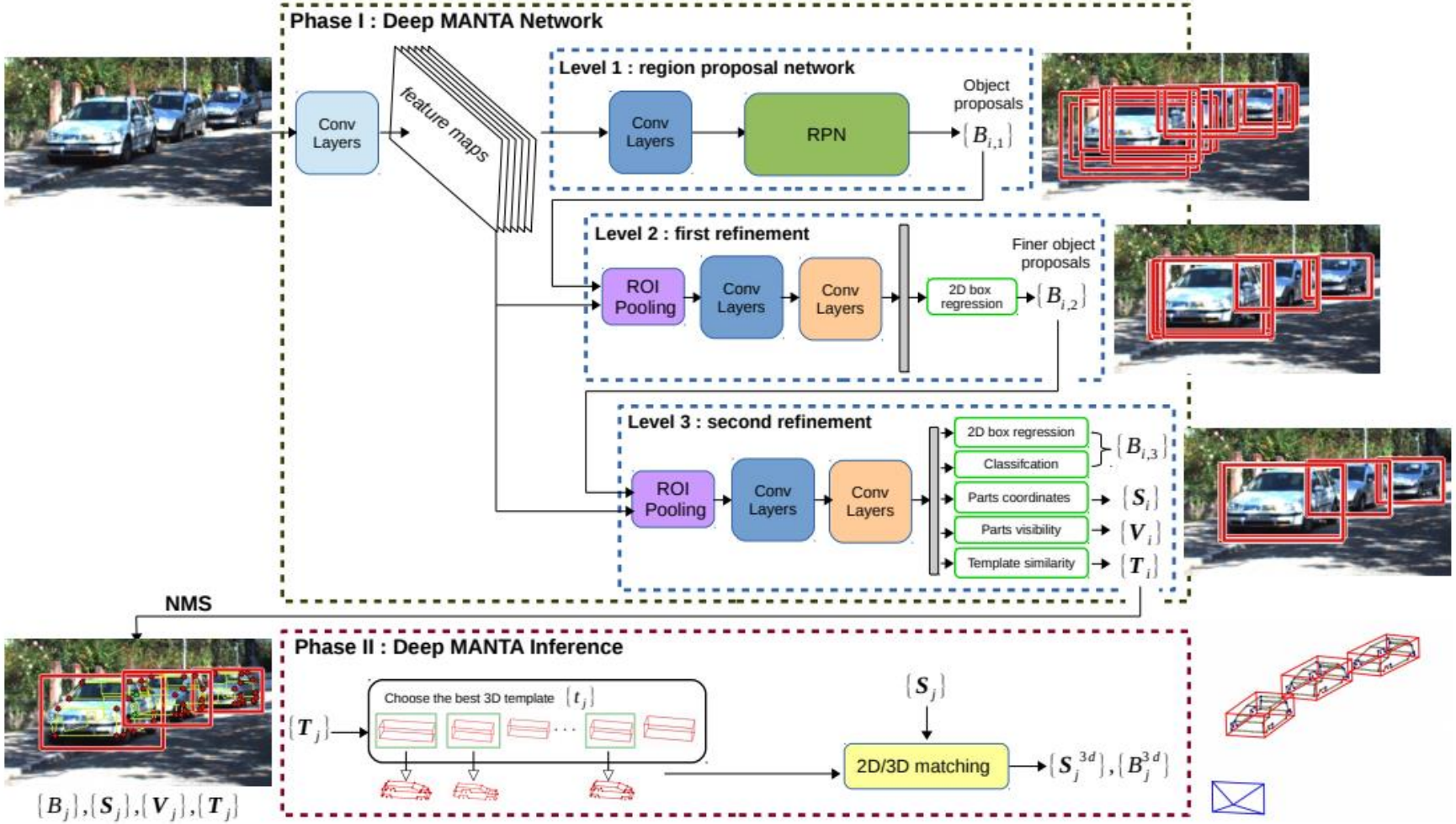
Convolutional neural networks overview



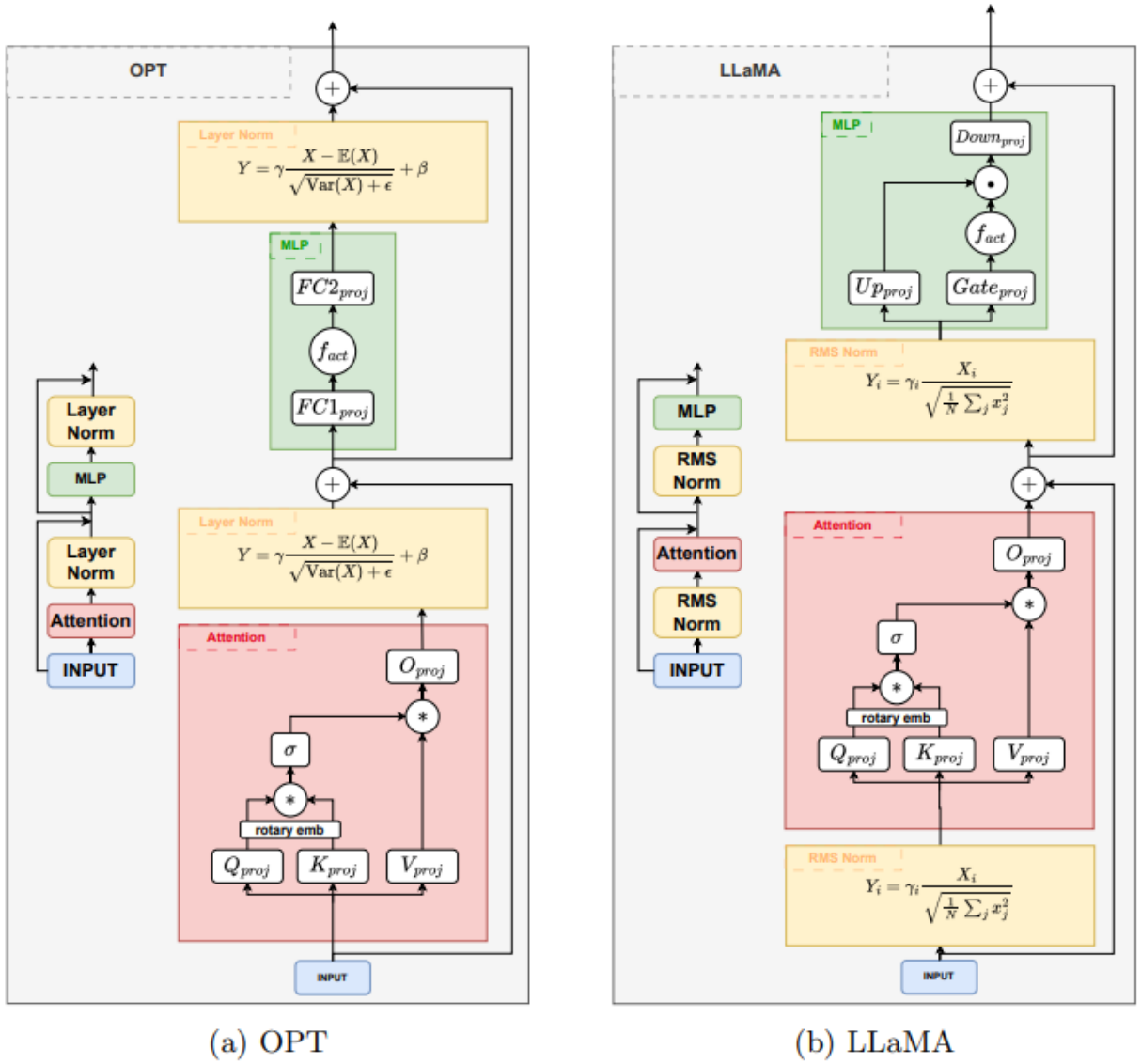
Deep = number of layers >> 1

Complex neural network systems

Faster-RCNN applied to ADAS:



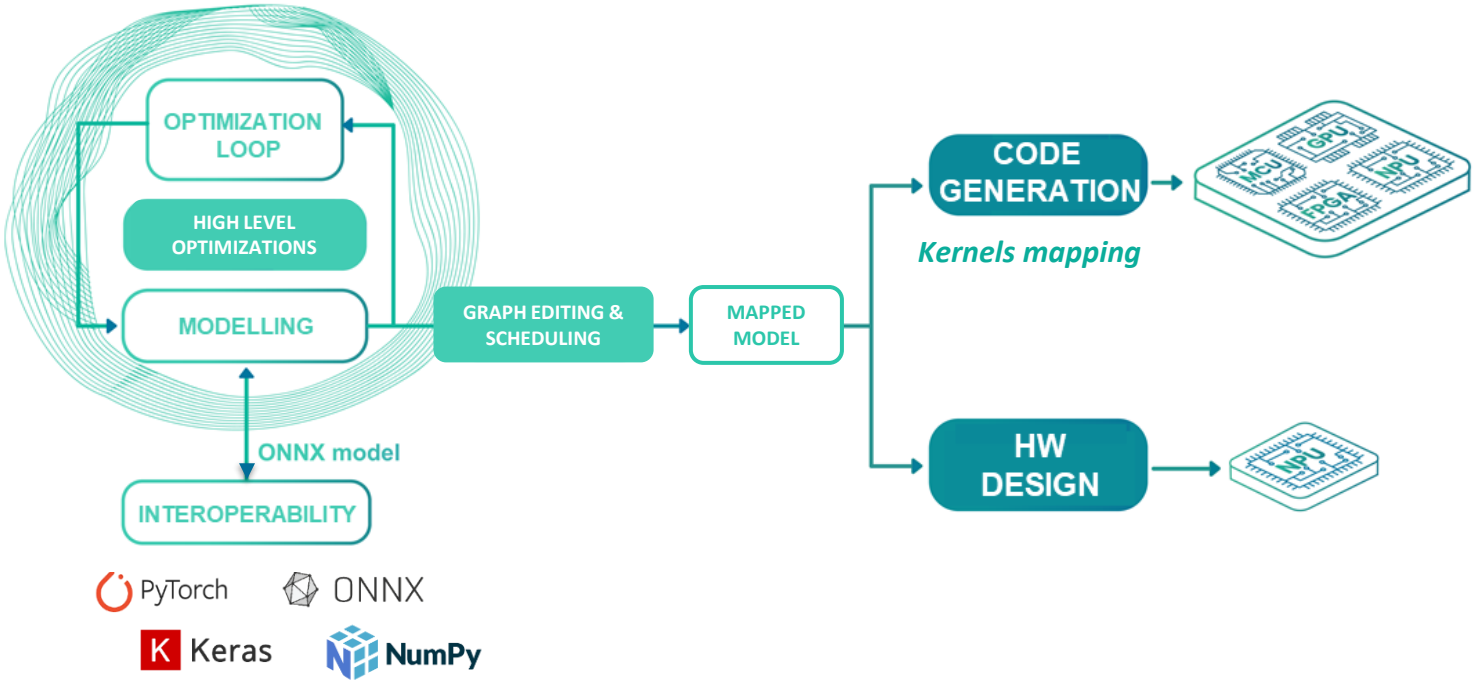
Transformers: ViT and LLM



(a) OPT (b) LLaMA

- LLM are causal models
- They only predict the next token
- 3 main blocs:
  - Attention
  - Feed-Forward Network
  - Normalisation

Aidge: a DL framework dedicated to embedded AI



Remember: the targeted hardware does not necessarily have a compiler!

End-to-end interoperable framework including optimization and deployment

Modular library with light core, plugins system and few dependencies

Open source code hosted by ECLIPSE FOUNDATION with adapted license to create proprietary software (EPLv2)

+40 industrial and academics partners

Modular deployment framework: beyond fragmentation & black-box

Interoperability

- Broad ONNX support (import and export)
- Intelligent **fallback mechanism**
- Automatic **operator registration**
- ONNX runtime **test suite** for validation

Intermediate Representation

- **Hierarchical and adaptative IR**
- **Expansion or fusion** using meta-operators
- Unique **graph-matching engine**
- Rich **analysis tools** (Aidge Model Explorer)

Model Optimization

- Guaranteed **reproducibility**
- **Quantization**: post-training or aware training (SAT, LSQ)
- **Compression**: tensor decomposition

Graph Edition & Transformation

- **Model simplification**
- **Model transformation** to match hardware capabilities (**tiling**)
- **Model adaptation** to available kernels

Code Generation

- Transparent code generation using **templates**
- Support **multiple programming models**
- Orchestration using a **producer–consumer model**
- Static memory allocation

API

Multi-Platform and Packaging

Windows Mac ubuntu PIP

The right IR is everything

Graph matching and rewrite is **the single most important operation** for optimizing and mapping high level algorithms to hardware compute resources:

.

Strong ONNX interoperability require graph adaptation for **import and export**

- **ONNX simplifiers** consist of matching sub-graph pattern and replacing them
- **Quantization** consists of inserting quantizer nodes, scaling nodes and performing nodes fusion
- Tensor-decomposition **compression** consists of replacing Conv with several smaller Conv
- Mapping to memory-limited physical computing resources require operators **tiling**
- Mapping to available compute kernels require **graph adaptation** (casting, transposition, packing…)
- Multi-modal optimization with data augmentation require synchronized data processing pipelines

… all these can be described extremely efficiently with the **right intermediate representation** (IR)!

Key features of Aidge's IR

Key features:

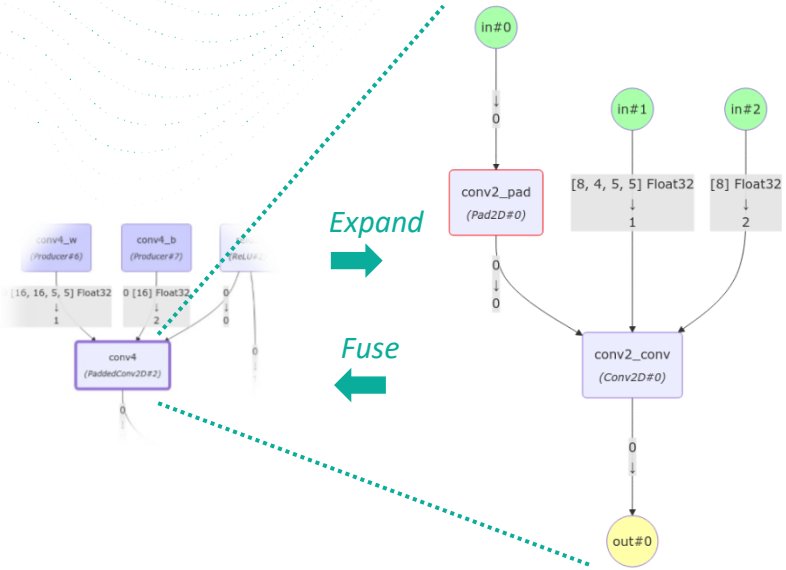
A **transparent** representation with a graph

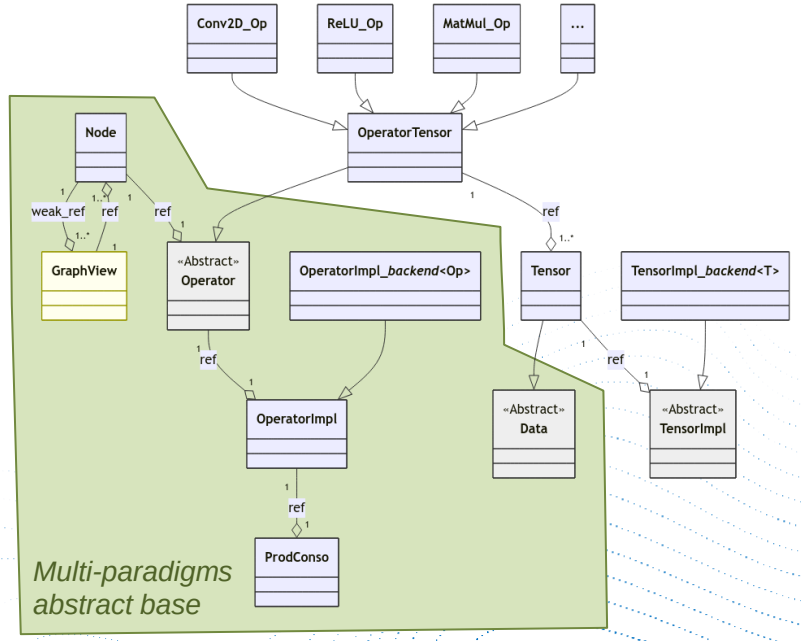
A **multi-paradigm** abstraction model

A **hierarchical** graph model ➔ *Meta Operator*

A **generic** graph model ➔ *Generic Operator*

An **accessible** graph model ➔ *Graph Matching*

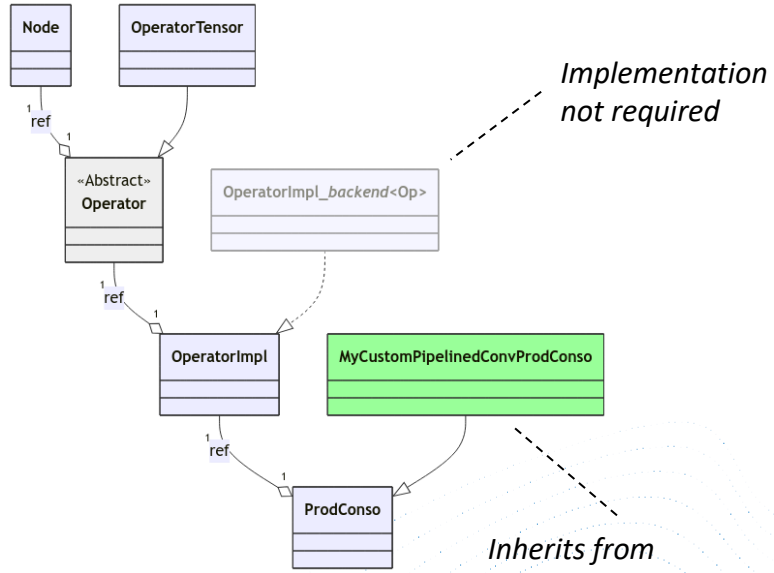





```
dinov2_model = aidge_onnx.load_onnx("dinov2.onnx")
aidge_core.fuse_to_metaops(dinov2_model, "MatMul-*>Add", "Linear")
aidge_core.fuse_to_metaops(dinov2_model, "MatMul->Div#1->Softmax-*>MatMul;"
                                         "Div#1<1~Producer", "ScaledDotProductAttention")

aidge_core.fuse_to_metaops(dinov2_model, "ScaledDotProductAttention#1->Transpose->Reshape#1->Linear;"
                                         "Reshape#1<1~Producer;"
                                         "ScaledDotProductAttention#1<0-(Transpose<-Reshape#2<-Add#1);"
                                         "ScaledDotProductAttention#1<1-(Transpose<-Reshape#3<-Add#2);"
                                         "ScaledDotProductAttention#1<2-(Transpose<-Reshape#4<-Add#3);"
                                         "Reshape#2<1~Producer;"
                                         "Add#1<*-0-Split#1;"
                                         "Add#2<*-1-Split#1;"
                                         "Add#3<*-2-Split#1;"
                                         "Split#1<-MatMul;"
                                         "Split#1<1~Producer", "MultiHeadAttention")
```

Key features of Aidge's IR

- Aidge introduces a well-defined consumer-producer (C-P) model
  - Similar to transaction-level modeling (TLM) for electronic design



C-P model is ***state full***

```
class Conv2D_DataFlow_CP(aidge_core.ProdConso):
    def __init__(self, op: aidge_core.Operator):
        aidge_core.ProdConso.__init__(self, op)
        self.state_begin = True

    def get_nb_required_data(self, input_idx):
        input = self.get_operator().get_input(input_idx)
        if input:
            if self.state_begin:
                # Require 3 lines for 3x3 kernel
                return aidge_core.Elts_t.data_elts(3 * input.dims()[2])
            else:
                return aidge_core.Elts_t.data_elts(input.dims()[2])
        else:
            return aidge_core.Elts_t.none_elts()

    def get_required_memory(self, output_idx, inputs_size):
        output = self.get_operator().get_output(output_idx)
        self.state_begin = False
        return aidge_core.Elts_t.data_elts(output.dims()[2])
```

C-P model	Specification
Token-based	Arbitrary data quantity (FSM logic / Petri net modeling) *Standard Tensor-based AI frameworks scheduling*
Data-based	**Precise amount of data elements consumed and produced at each execution step**
Hybrid	Data-based to token-based

Fine grain (pipelined) dataflow modeling

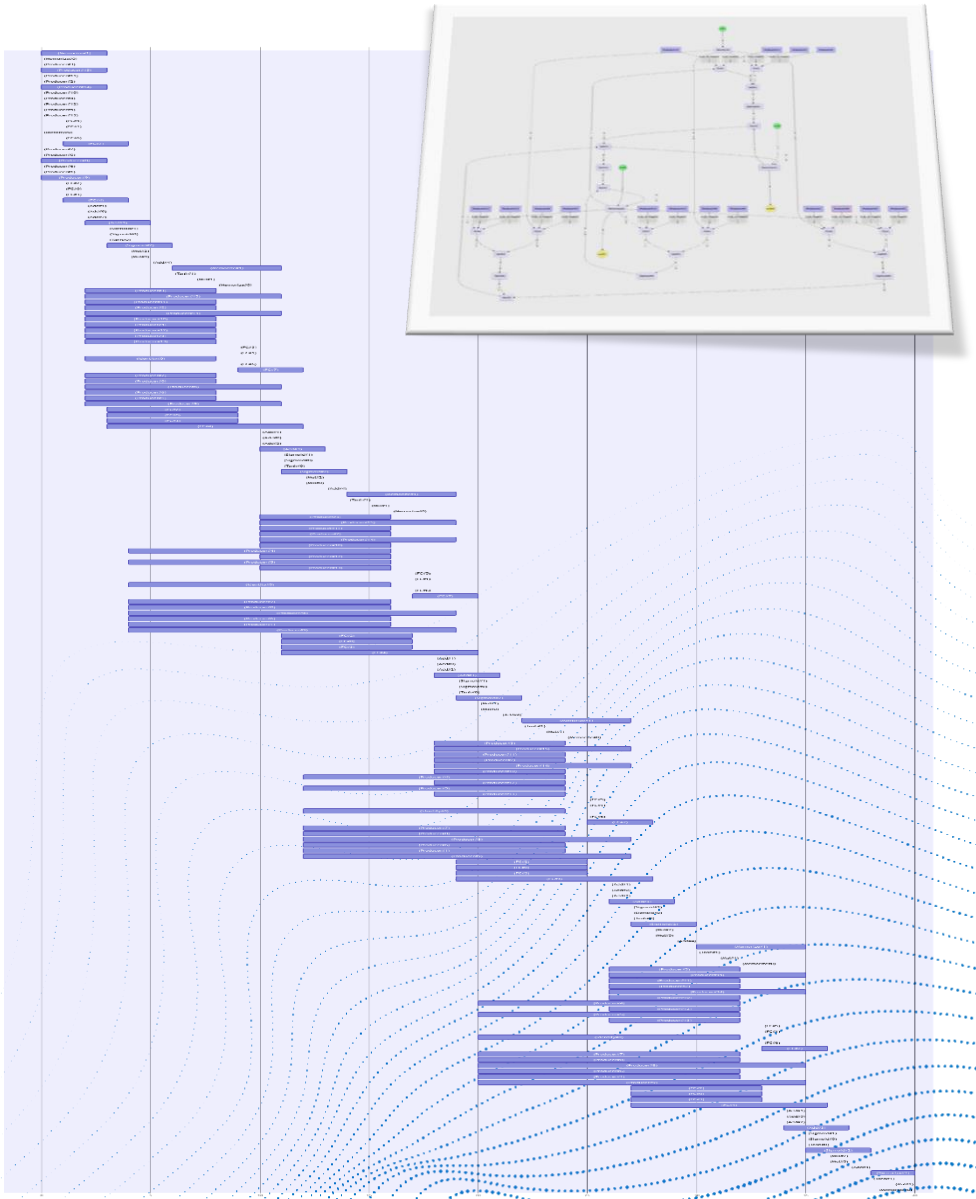
Key features of Aidge's IR

- Aidge's scheduling is **static**
  - Get early and late logical static scheduling for any graph, cyclic or acyclic

```
lstm = aidge_core.LSTM(in_channels=4, hidden_channels=8, seq_length=5)

# Flatten the graph:
lstm_model = aidge_core.get_connected_graph_view(lstm)
aidge_core.expand_metaops(lstm_model)
lstm_model.set_backend("cpu")

# Create the Scheduler
lstm_scheduler = aidge_core.SequentialScheduler(lstm_model)
lstm_scheduler.generate_scheduling()

# Display static scheduling
lstm_scheduler.save_static_scheduling_diagram("lstm_scheduling")
```

Based on the specified C-P model

Agnostic of the programming model
An associated operator implementation is not required
(the scheduling is decoupled from the execution path)



Key features of Aidge's IR

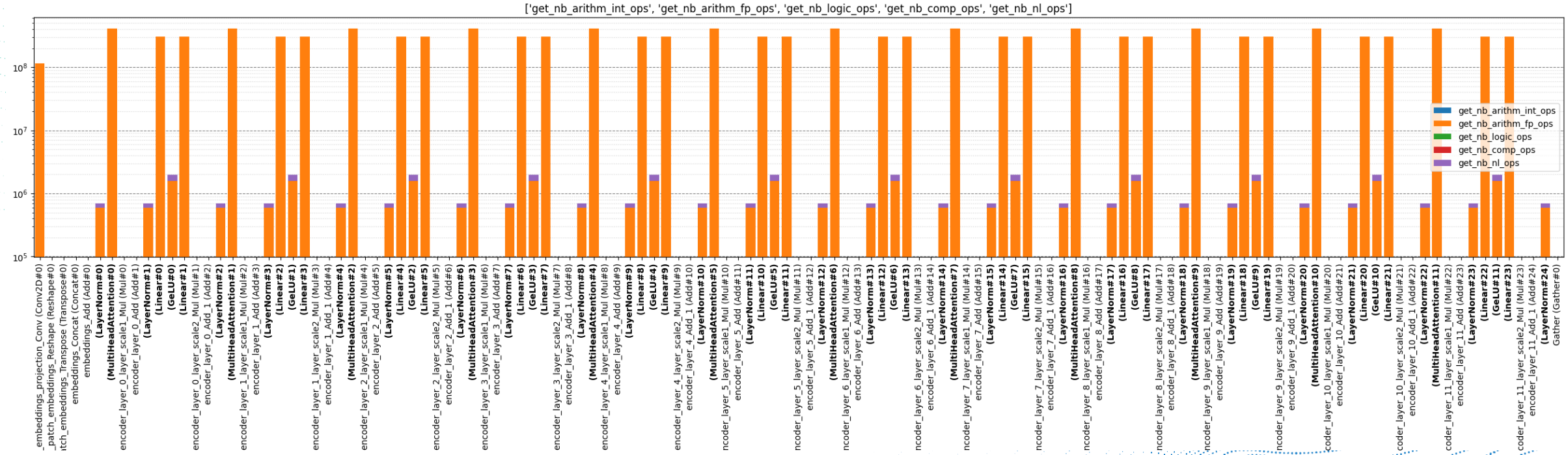
- Comprehensive and simple static analysis

```
stats = aidge_core.static_analysis.StaticAnalysisExt(dinov2_model)
stats.summary() # Simple text output similar to Keras model.summary()

stats.log_nb_ops_by_type("stats_ops.png", log_scale=True)
```

Rich low-level & high-level APIs:

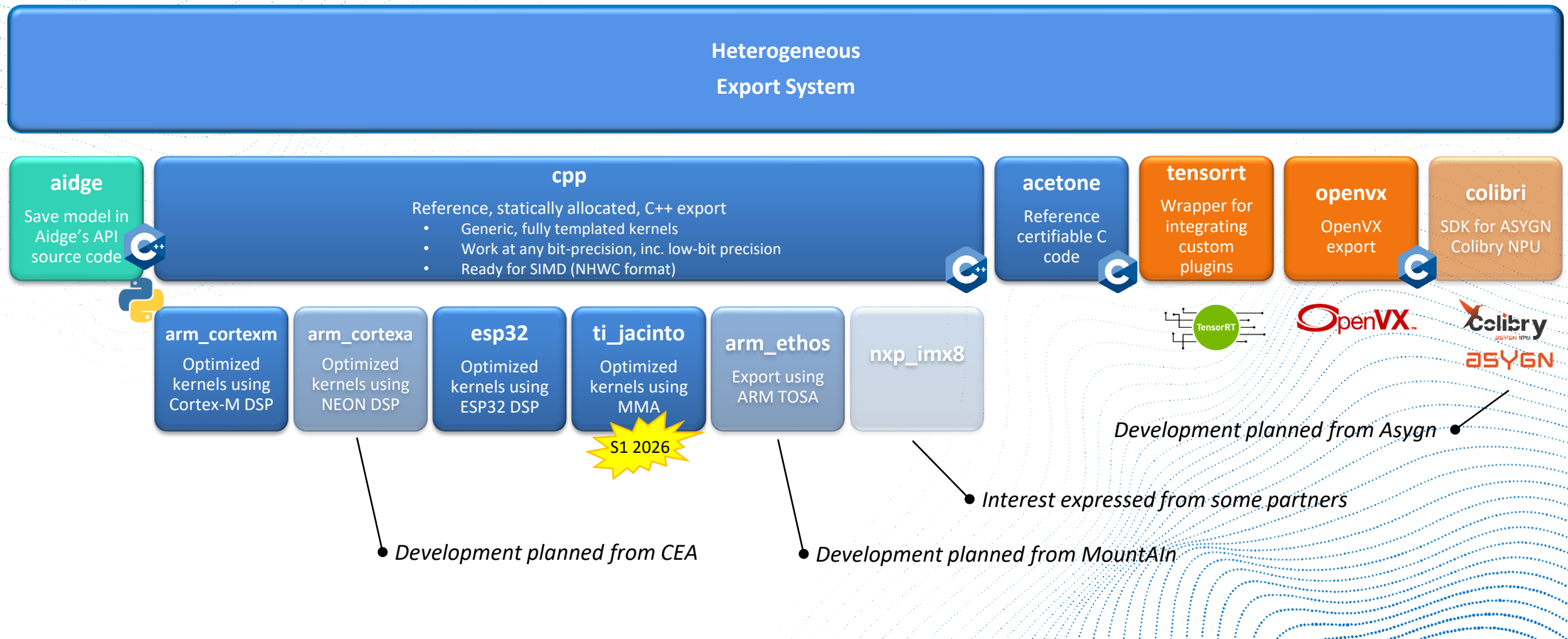
- *Base C++: for lower-level heuristics integration*
- *Extended Python: with additional visualization methods*

Auto aggregation for Meta operators, specifiable for Generic operators



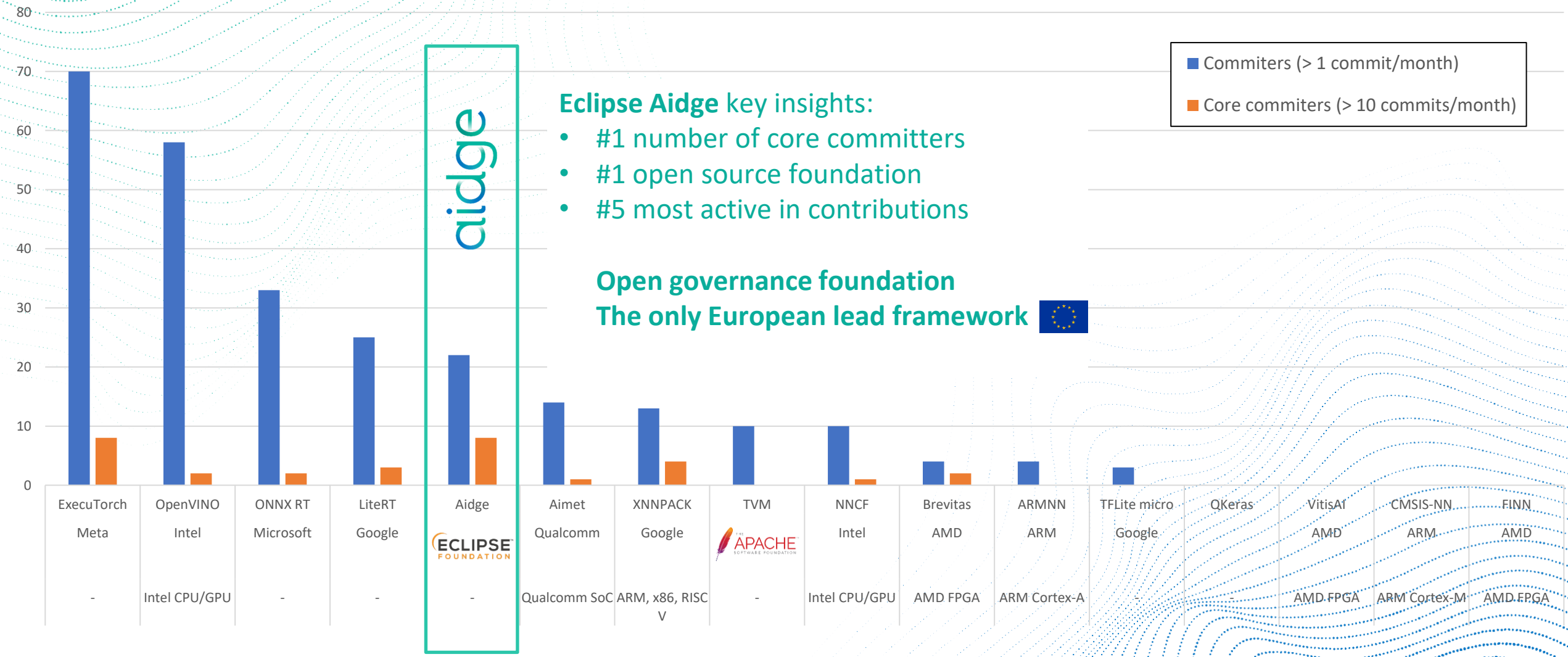
Aidge's export modules overview

Code generation exports



Aidge in the open source landscape

Project activity in the last 12 months (sept. 24 – sept. 25)



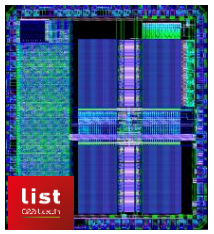
Eclipse Aidge key insights:

- #1 number of core committers
- #1 open source foundation
- #5 most active in contributions

Open governance foundation
The only European lead framework

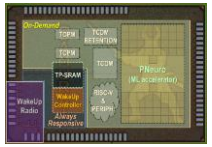
AI Circuits & Architectures

Smart Sensors



Esperanto
Gesture detection
130nm + OxRAM + PMUT,

CNN



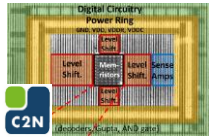
PNEURO
Raptor HW/SW IP
FD28nm

DOLPHIN DESIGN

Transformers & multimodality

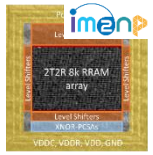
INSTITUT CARNOT CEA LIST

BUMBLEBEE
Attention based AI
Circuit Architecture

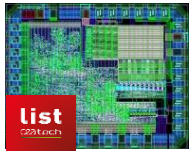


Bayesian machine

Classification



BNN
Digital popcount
130nm, OxRAM



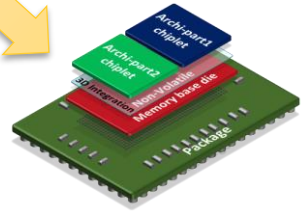
SPIRIT
Spiking SNN
130nm, OxRAM

Spiking SNN

LARGO
28nm, OxRAM

CNN backbone

NeuroCORGI
Feature Extractor (HD images)
GF22nm, OxRAM



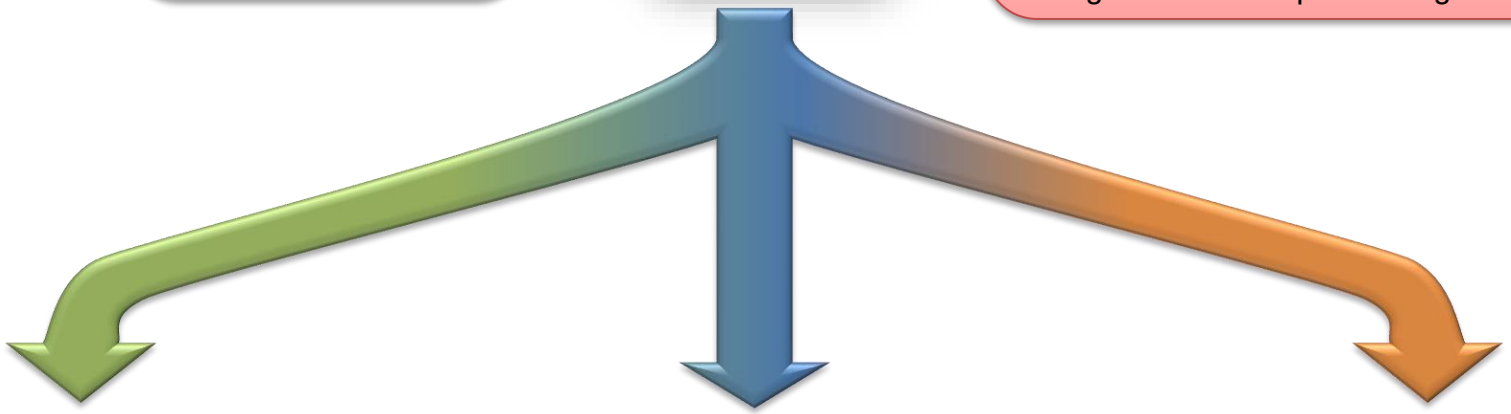
MOSAIC
Modular AI systems composed of heterogeneous components

2018 2020 2022 2023 2026

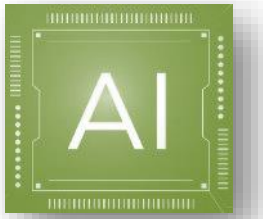
Towards hardware *generation*

AI Applications:
- Classification
- Segmentation
- Objects detection
- Language model

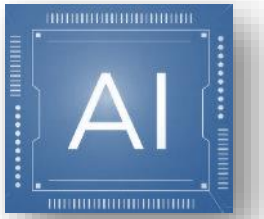
- Generic architecture
- Require the latest technological process node to reach the best energy efficiency
  - Huge costs of 500M$
- High information processing latency



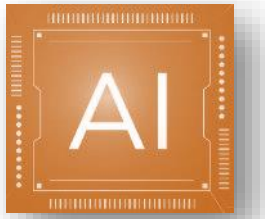
Classification
Segmentation



Objects detection



Language model



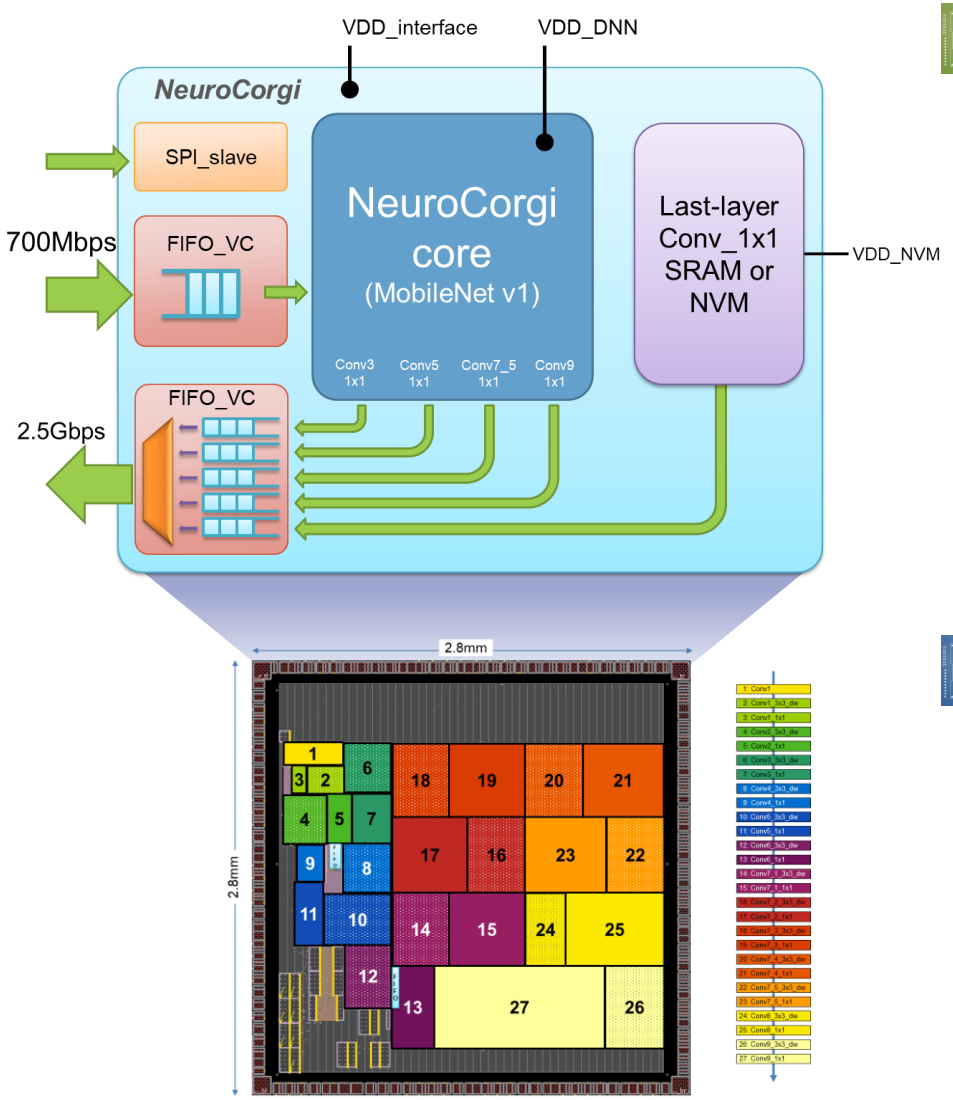
- Per domain specialized architecture
- High energy efficiency
- Low processing latency
- Do not require the latest process node

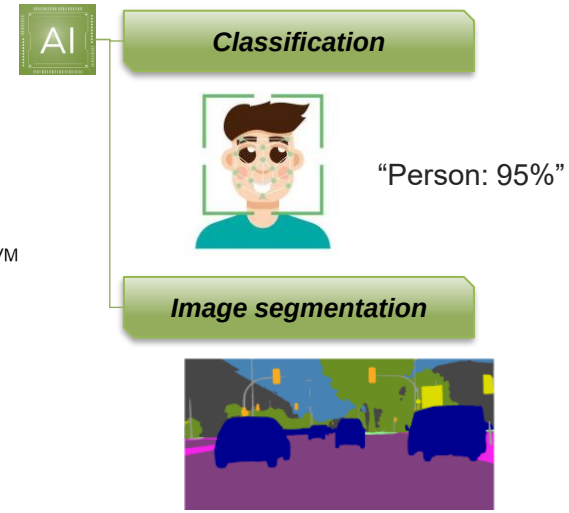
The NeuroCORGI circuit

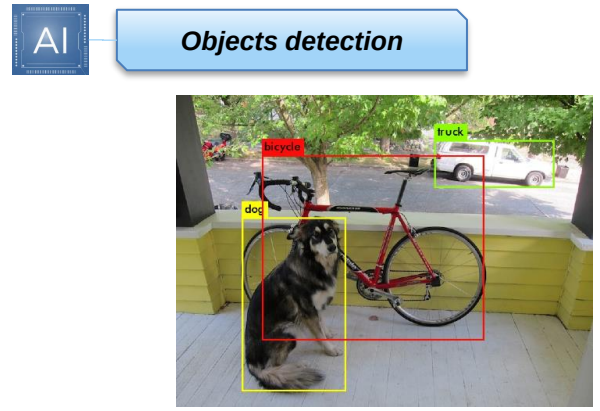
- Fix topology: MobileNet v1
- Parameters freeze at design
- 4-bits quantization
- Very low clock frequency
- NVM-Leti memory
- 3 different circuit versions:

AI ; AI + NVM ; AI .

⇒ *Ultra low energy consumption*
⇒ *Ultra low processing latency*







NeuroCORGI benchmark results

Ultra low power: Active = **26mW ➔ 26%** of imager power (>100mW) ; Leakage = **218µW**

Best in class energy / inference / pixel: **0,84nJ ➔ 8x** to **1000x** better compared to SoTA

Best in class latency with batch size 1: **14ms** ➔ 42% of 30fps imager latency

	This work	Samsung (ISSCC 2022)	Vega (ISSCC 2021)	MediaTek (ISSCC2020)	Alibaba (ISSCC 2020)	KAIST	Eyeriss v2	Dolphin RAPTOR	GAP8
Technology	GF22FDX	4nm	GF22FDX	7nm	TSMC 12nm	65nm	TSMC 65nm	GF22FDX	
AI model	MobileNet v1	MobileNet TPU	MobileNet v2	MobileNet v1	ResNet v1	MobileNet	MobileNet v1 (alpha 0.5)	MobileNet v2	MobileNet v1
Energy/inference (224x224)	0.042mJ	0.35mJ	1.19mJ	0.17mJ	2.00 – 3.51 mJ	1.2mJ	1.55mJ*	2.6mJ	49.7mJ
Energy/inference/ pixel	0.84nJ	6.90nJ	23.72nJ	3.46nJ	39.92 – 70.02 nJ	23.92nJ	30.99nJ	51.82nJ	990.05nJ
Energy/inference/ pixel W.r.t. this work	1x	8.19x	28.15x	4.11x	47.31x – 83.10x	36.78x	36.78x	61.5x	1175x

Approach proven on several applications

Drones/USV

Underwater Acoustic Signal Classification

Robust Autonomous Landing

3D Object Detection and Classification of Road Users

Autonomous Weeding System

Tomato pests and diseases forecast

Olivier Bichler