INTELLIGENCE ARTIFICIELLE
et
ELECTRONIQUE EMBARQUEE

CRESITT Orleans 17 Oct. 2019
Michel VINEZ ARROW

AGENDA

ARROW

MARCHES & APPLICATIONS

MISE EN ŒUVRE

EDGE COMPUTING

DEEP LEARNING : OFFRE ARROW

INFERENCE : OFFRE ARROW

EXEMPLES DE REALISATIONS

Key Metrics

"Our strategic roadmap includes advancing our leadership in IoT and providing the engineering services and support that are integral for our customers to bring new, compelling products and services rapidly to the market"

Mike Long
President, CEO, Chairman
Arrow Electronics

Arrow Electronics

Core Business Units

Global Components

- Semiconductor
- IoT Sensors, Wireless, Gateways
- Embedded Products
- Lighting
- Power management
- Electro mechanical
- Design Services
- Technical Support
- Kitting & Programming
- Supply Chain Services

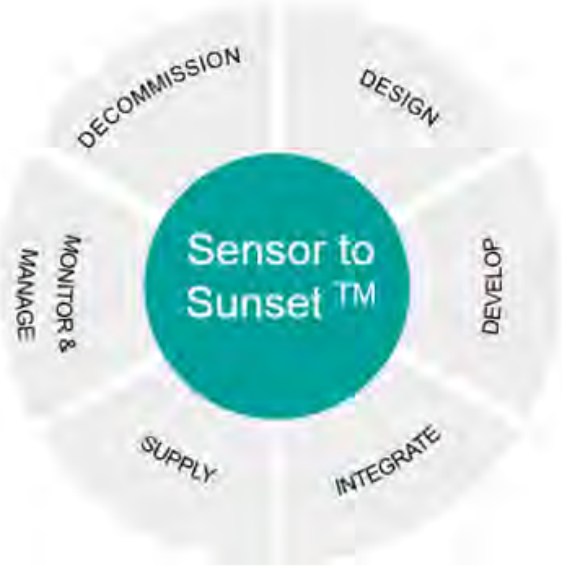
Global Services

- SW Integration
- POS terminal, ATM,
- Embedded x86 apps
- Industrial handheld
- Kiosks & Gateways
- Digital Signage
- Security & IP Cameras

Enterprise Computing Solutions

- Server & storage
- Virtualisation
- Security
- Networking
- Cloud Services
- Enterprise SW
- Data & Analytics

Arrow IoT



Mission

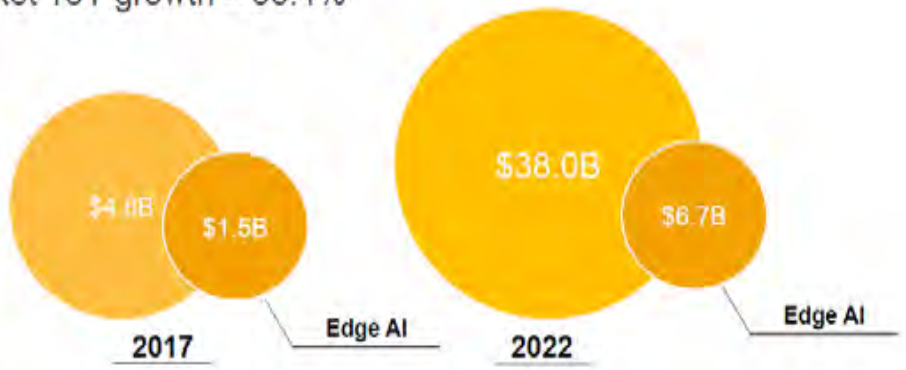
To offer a **complete** solution enabling businesses to deploy, manage, monitor, analyze and monetize secure connected devices throughout their entire lifecycle globally. **From Sensor to Sunset ™**


IA
MARCHE & APPLICATIONS

IA : MARCHE EN CROISSANCE

Global AI market YoY growth – 51%

Edge AI market YoY growth – 35.4%



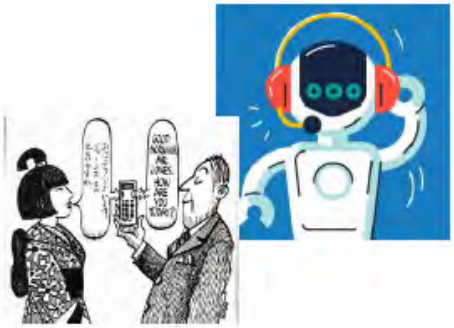
Source: MarketsandMarkets, statista

IA : APPLICATIONS

Image Recognition

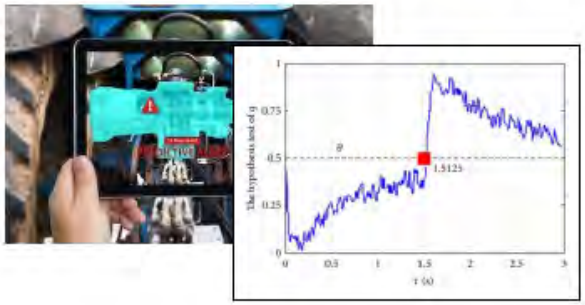
Speech Recognition & Synthesis

Natural Language Processing



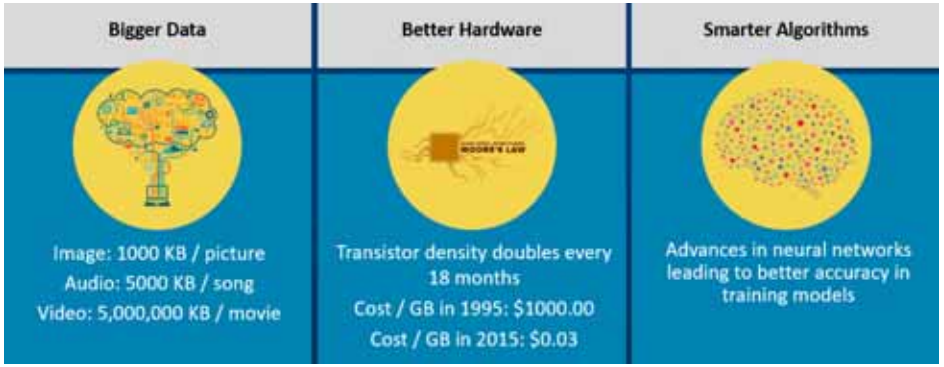
Fintech & Insurance

Predictive Maintenance



Robotics

IA : FACTEURS DE CROISSANCE



Source : INTEL

DISPONIBILITE de DONNEES REELLES

Ex:Données des Réseaux Sociaux
Données des applications WEARABLE

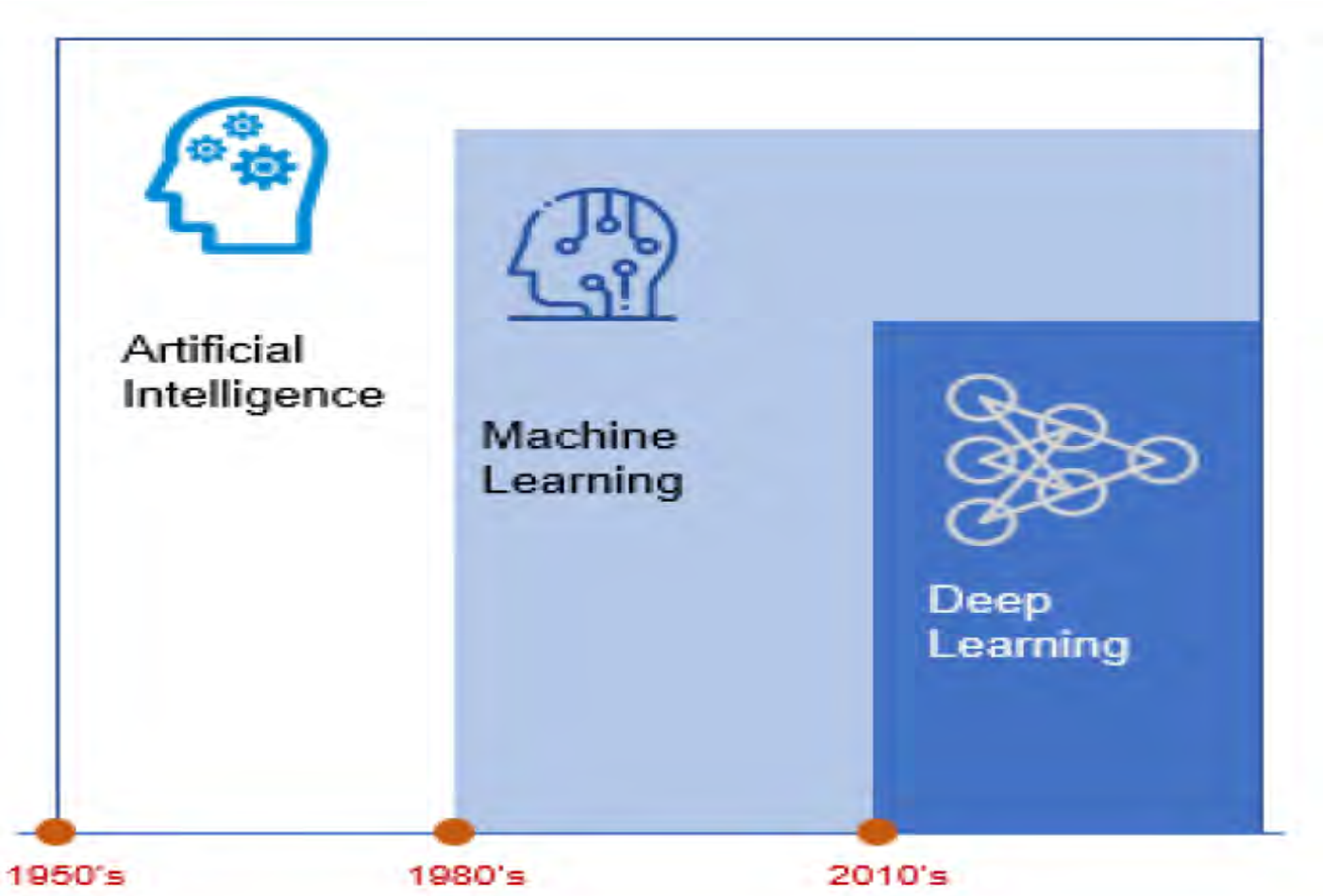
DISPONIBILITE de MACHINES DE TRAITEMENT HAUTE PERFORMANCES ABORDABLES

Ex: Modules JETSON NVIDIA, FPGA Intel, Cortex A/M

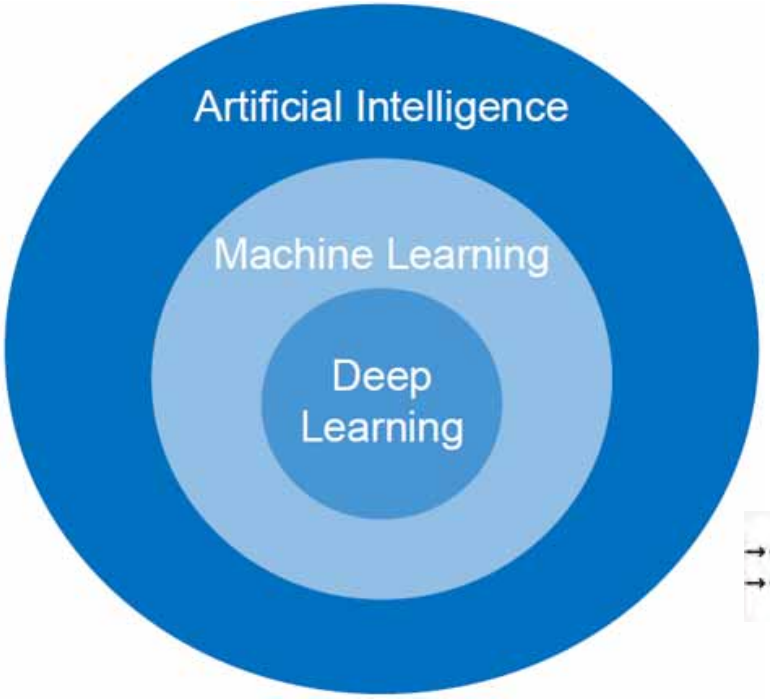
NOUVEAUX DEVELOPPEMENTS ET OPTIMISATIONS des RESEAUX DE NEURONE DANS LE DOMAINE DE L'APPRENTISSAGE

Indispensable pour le EDGE COMPUTING (IA EMBEDDED)

IA : VUE TEMPORELLE



IA : DEFINITIONS



INTELLIGENCE ARTIFICIELLE (IA)

Englobe l'ensemble MATERIEL et LOGITIEL servant à imiter l'intelligence Humaine

MACHINE LEARNING (ML)

Une des solutions de faire de l'IA
Basée sur des Techniques STATISTIQUES :
Auto-Apprentissage
Auto-amelioration

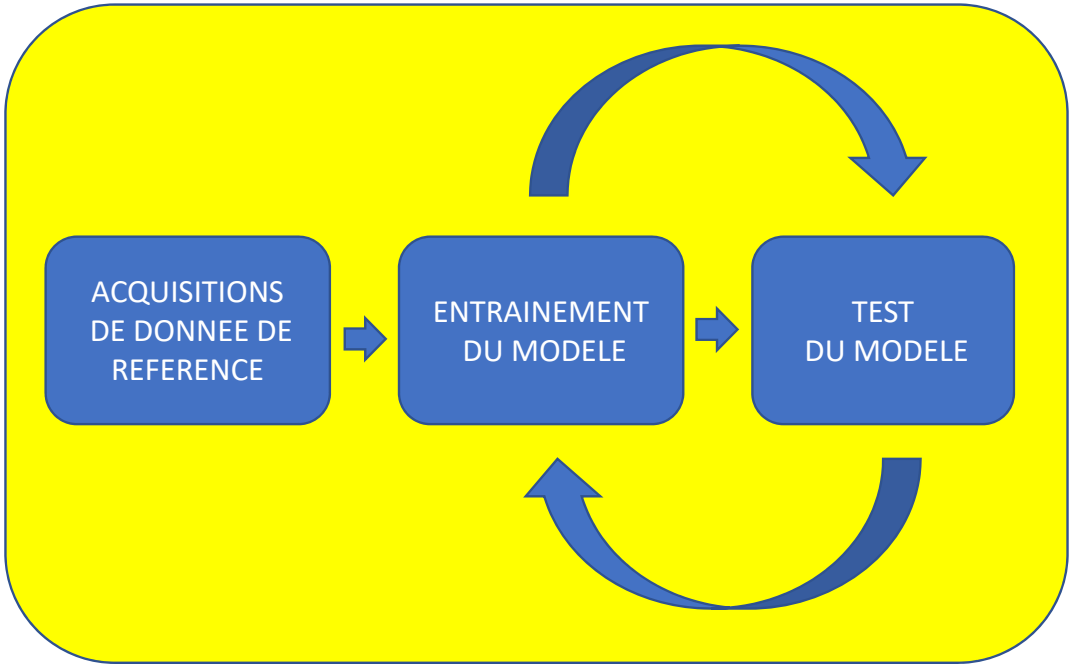
DEEP LEARNING (DL)

Une manières de faire du Machine Learning
Utilise les Réseaux Neuronaux Artificiels Multi-Niveaux
Apprentissage Autonomes
Nécessite une Puissance de Traitement très importante

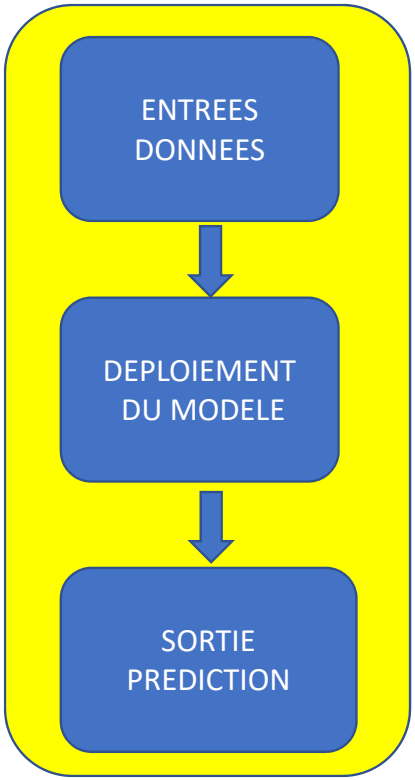

IA
MISE EN OEUVRE

IA: PRINCIPE DE FONCTIONNEMENT

Deux grandes Phases :



DEEP LEARNING (APPRENTISSAGE)



INFERENCE (PREDICTION)

DEEP LEARNING :

CONSTRUCTION BASE DE DONNES DE REFERENCE.
DIFFRENTS TYPES D'APPRENTISSAGE :

SUPERVISE

ETIQUETAGE DES DONNEES (LABELLING)
DES EXPERTS FONT LE TRAVAIL

NON SUPERVISE

PAS DE DONNEES DE REFERENCE

CLASSIFICATION PAR LE SYSTÈME LUI-MÊME
REGROUPEMENTS PAR CARACTERISTIQUES

DEEP LEARNING

NECESSITE PUISSANCE DE TRAITEMENTS IMPORTANTES

- **HPC**
- **SERVEURS MULTI PROCESSEURS**
- **CLOUD COMPUTING**
- **FARM DE GPU**

UTILISE DES BASES DE DONNEES ENORMES

- **Alimentées par :**
- **BIG DATA**
- **RESEAUX SOCIAUX**
- **DATA IOT**

UTILISATION de FRAMEWORKS DEDIES

- –TensorFlow – Google framework
- –Keras – higher level API, usually built on top of TensorFlow
- –Caffe2 / PyTorch – Facebook framework
- ….

INFERENCE : Deux Modèles

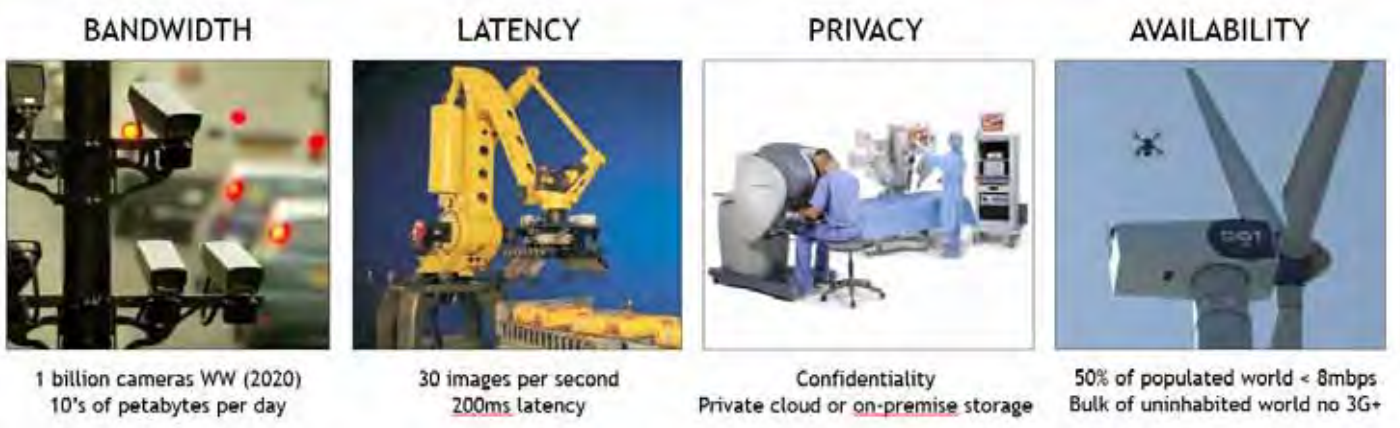
CLOUD COMPUTING

NECESSITE UNE CONNECTION INTERNET HAUT Débits

EDGE COMPUTING

NE NECESSITE PAS DE CONNECTION INTERNET

CONTRAINTES COMMUNES



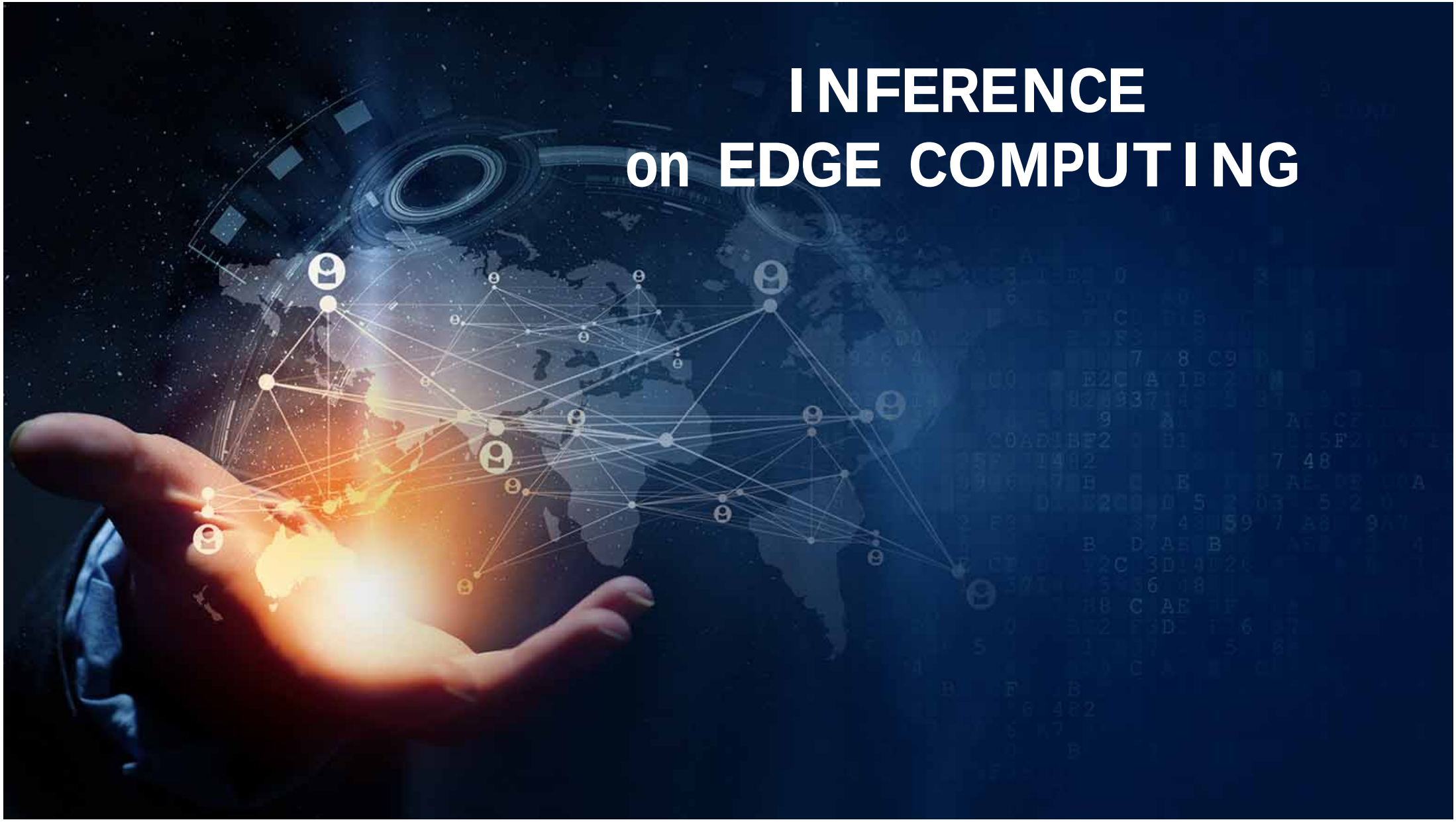
INFERENCE on CLOUD COMPUTING

AVANTAGES

DONNEES de Référence de Grande Qualité
Puissance de Calcul sans limite

INCONVENIENTS

Connectivité Cloud Haut Débits
Temps de Réponse non maitrisable
Energie consommée
Couts Fixes et récurrents élevés
Sécurité des Données privées non garantie (RGPD)


INFERENCE
on EDGE COMPUTING

INFERENCE on EDGE COMPUTING

AVANTAGES

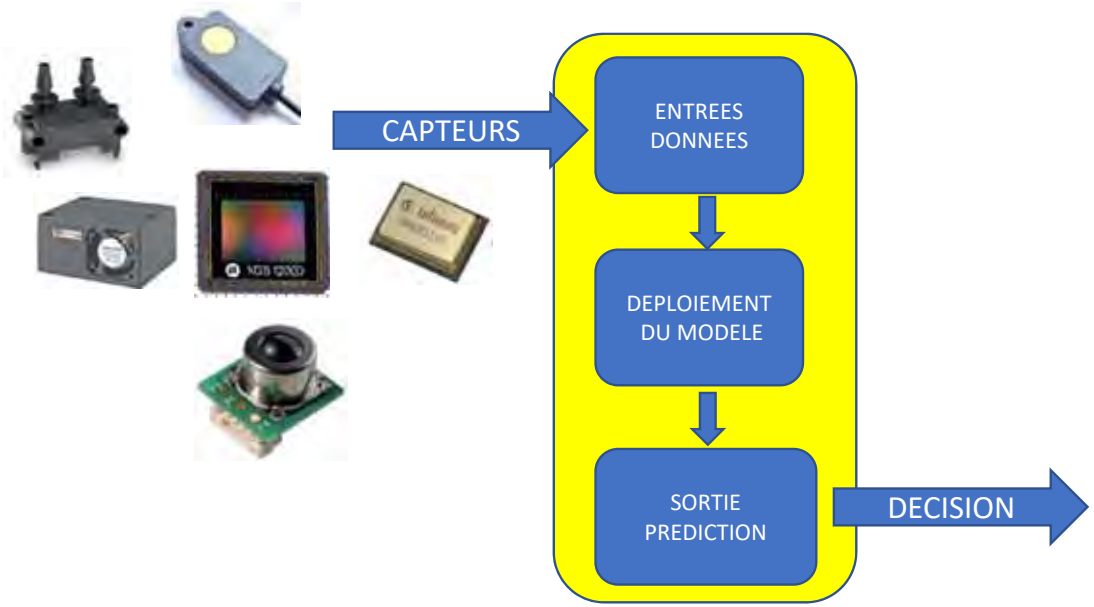
Données Privées et Sécurisées
Consommation Faible
Couts réduits
Taille Réduite
Temps de réponse court
Pas de Connexion Internet

INCONVENIENTS

Mémoire disponible réduite
Puissance de Calcul limitée
Nécessite un travail d'OPTIMISATION des Modèles

INFERENCE on EDGE COMPUTING :

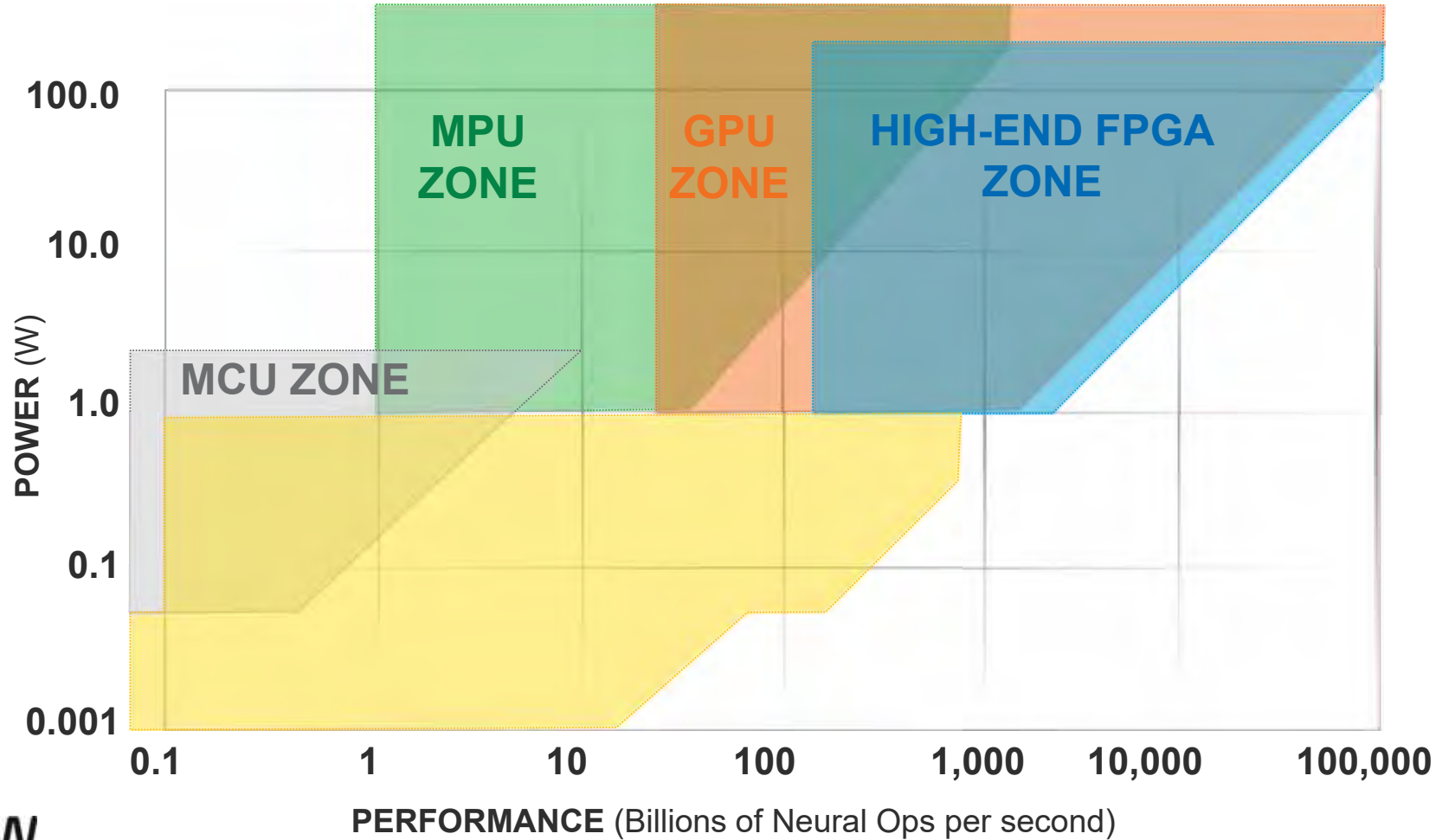
Principe



CPU/MCU	GPU	NPU	IP
Arm CORTEX A Arm Cortex M X86 INTEL	Arm MALI NVIDIA	Machine Learning Processeur	FPGA ASIC

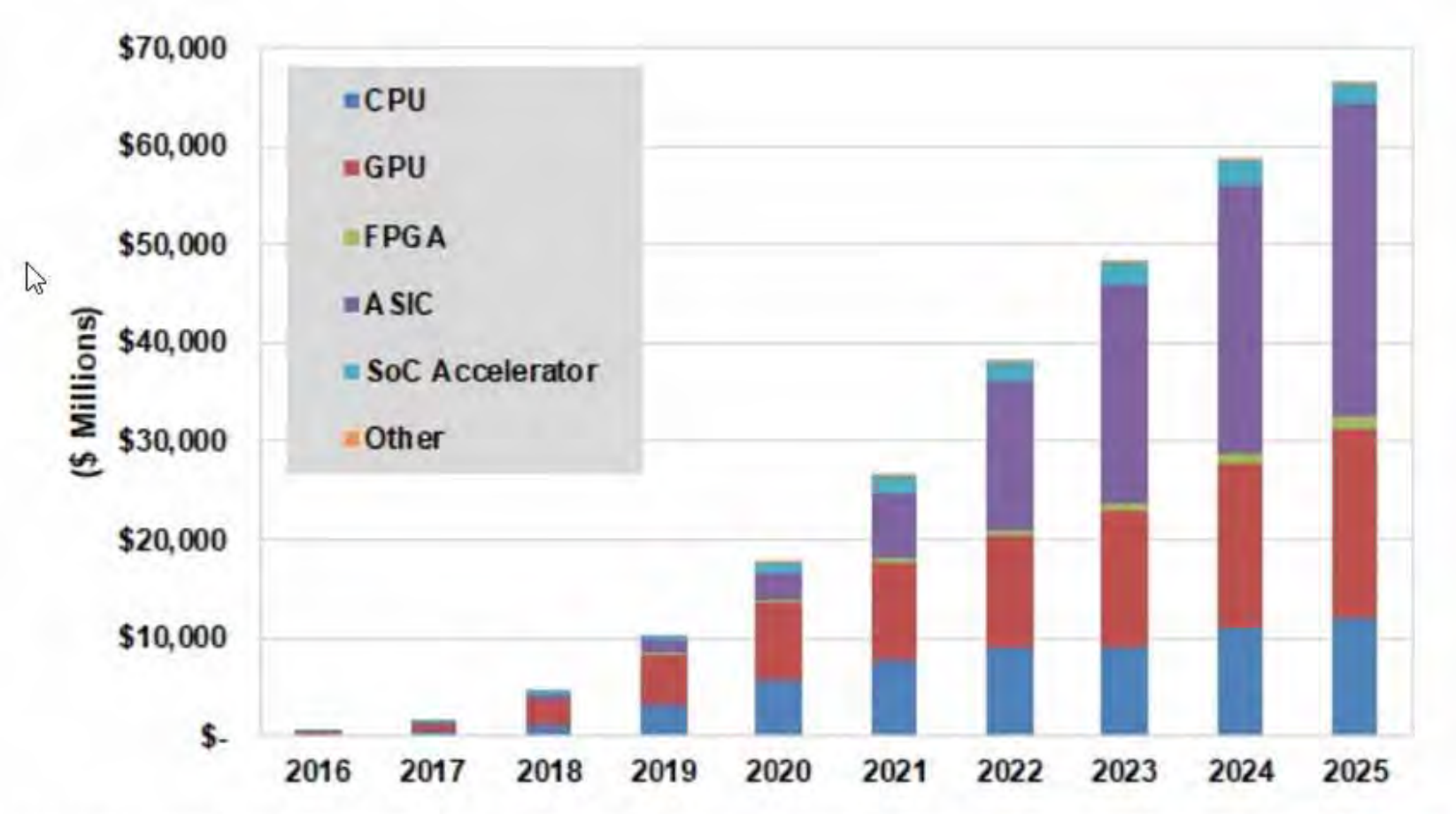
INFERENCE on EDGE COMPUTING

Consommation VS Performances



INFERENCE on EDGE COMPUTING

Tendances du marché



Les GPU suivent une croissance soutenue, mais perdent en part de marché relative (prévisions selon Tractica)

EDGE COMPUTING : Exemples

Enterprise Collaboration

Factory Automation

Service Robotics

Package Delivery

AI City

Personal Assist

Search and Rescue

Industrial Inspection

Portable Medical

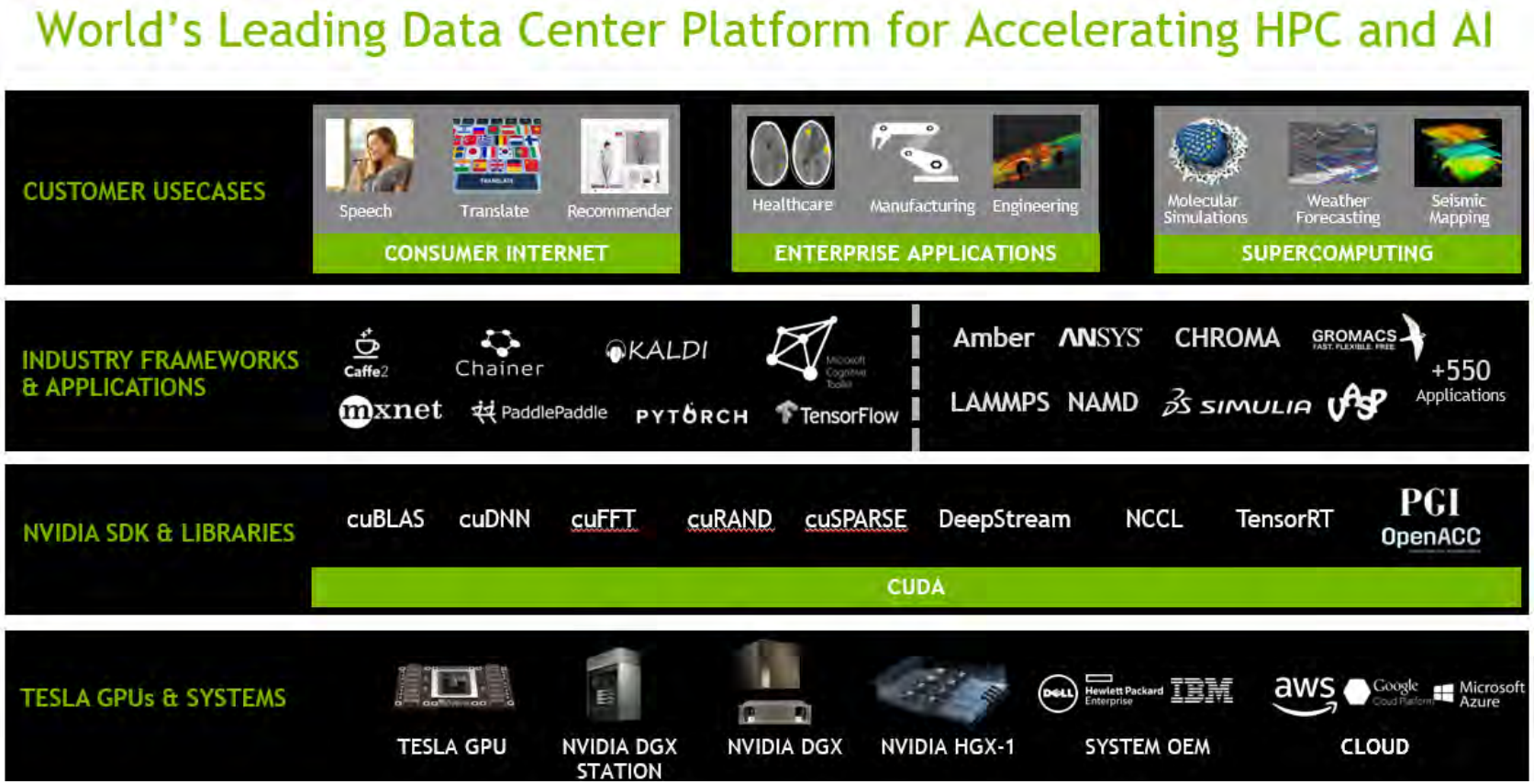
Academia and Research

Source : NVIDIA

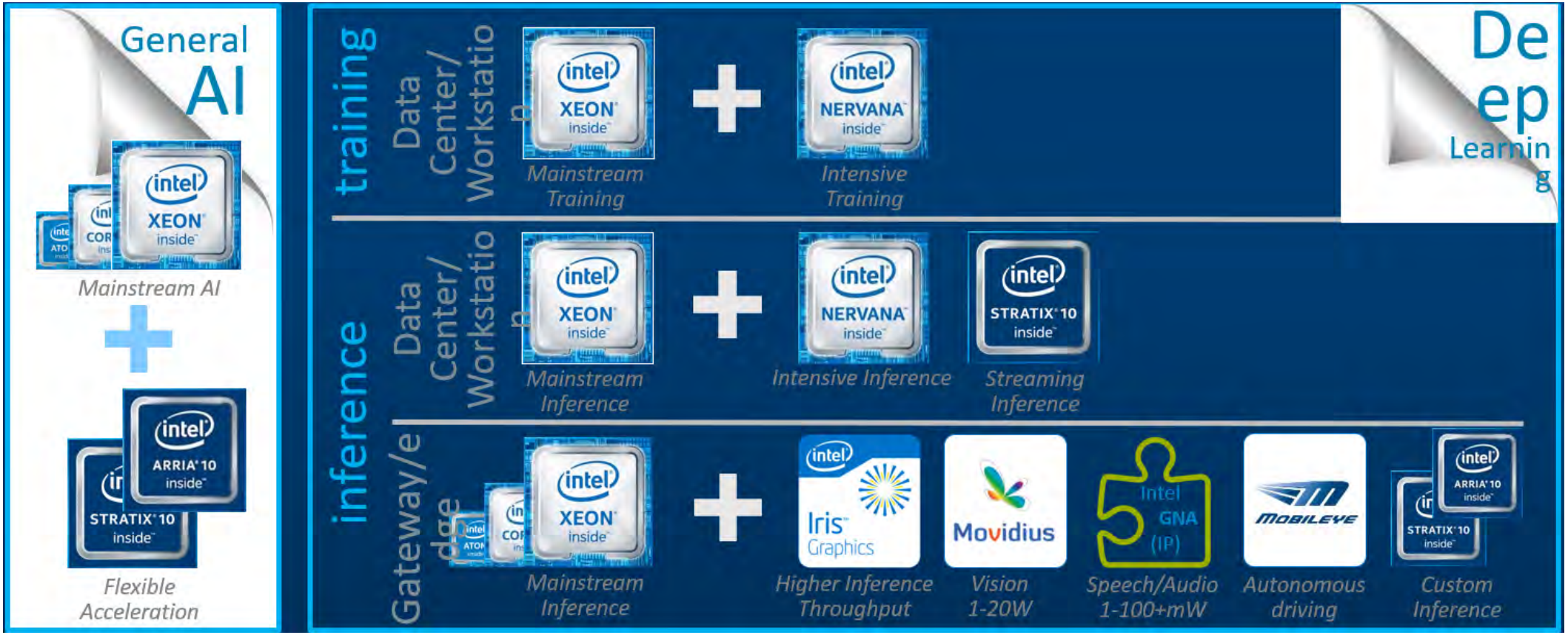

DEEP LEARNING
OFFRE ARROW

TRAINING : Offre GPU NVIDIA

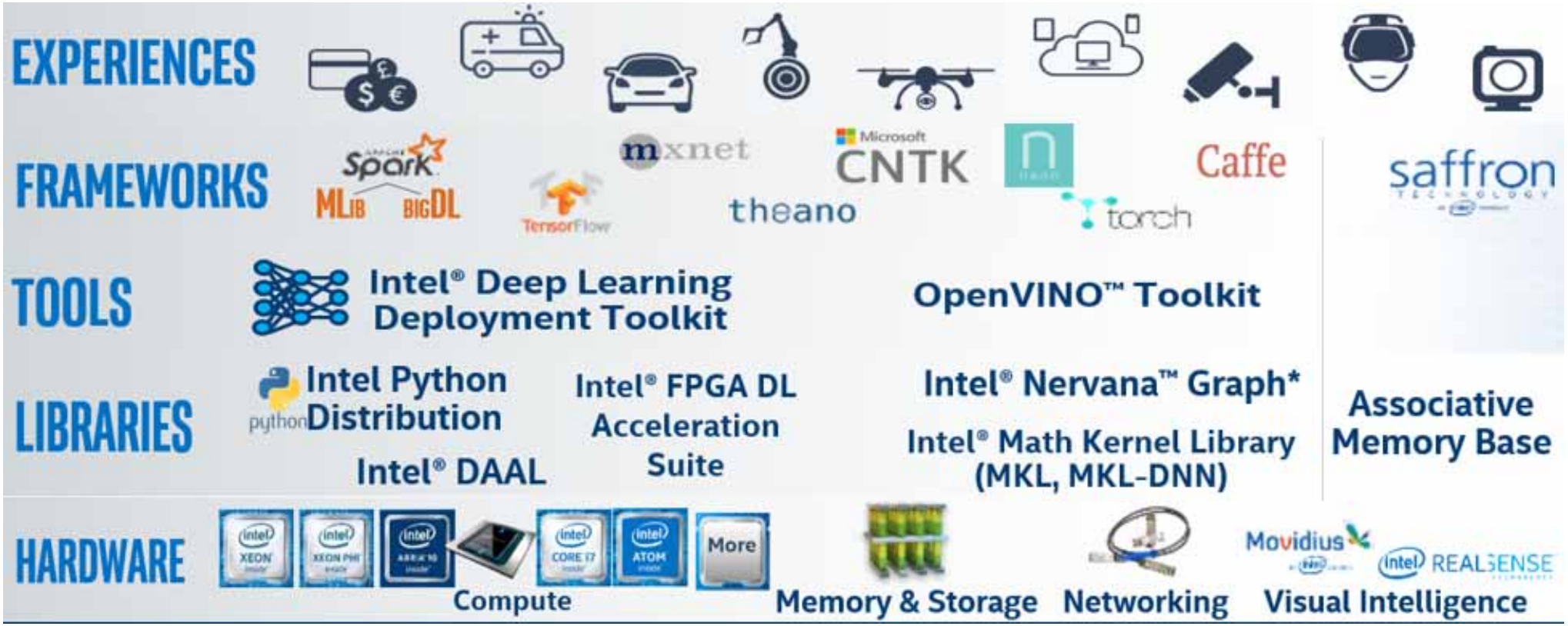
LOGITIELS : Offre GPU NVIDIA



TRAINING : Offre INTEL

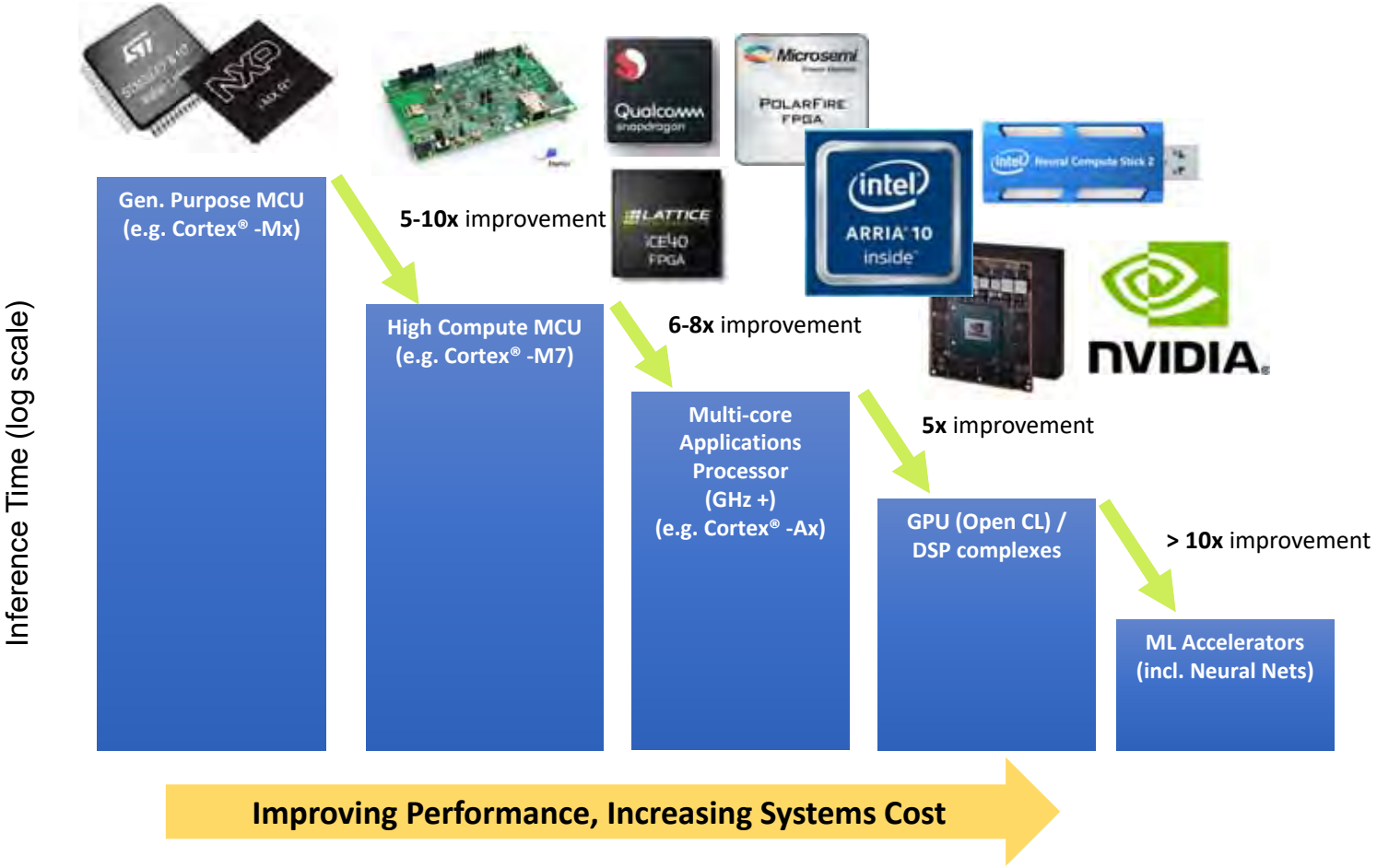


LOGITIELS : Solutions INTEL

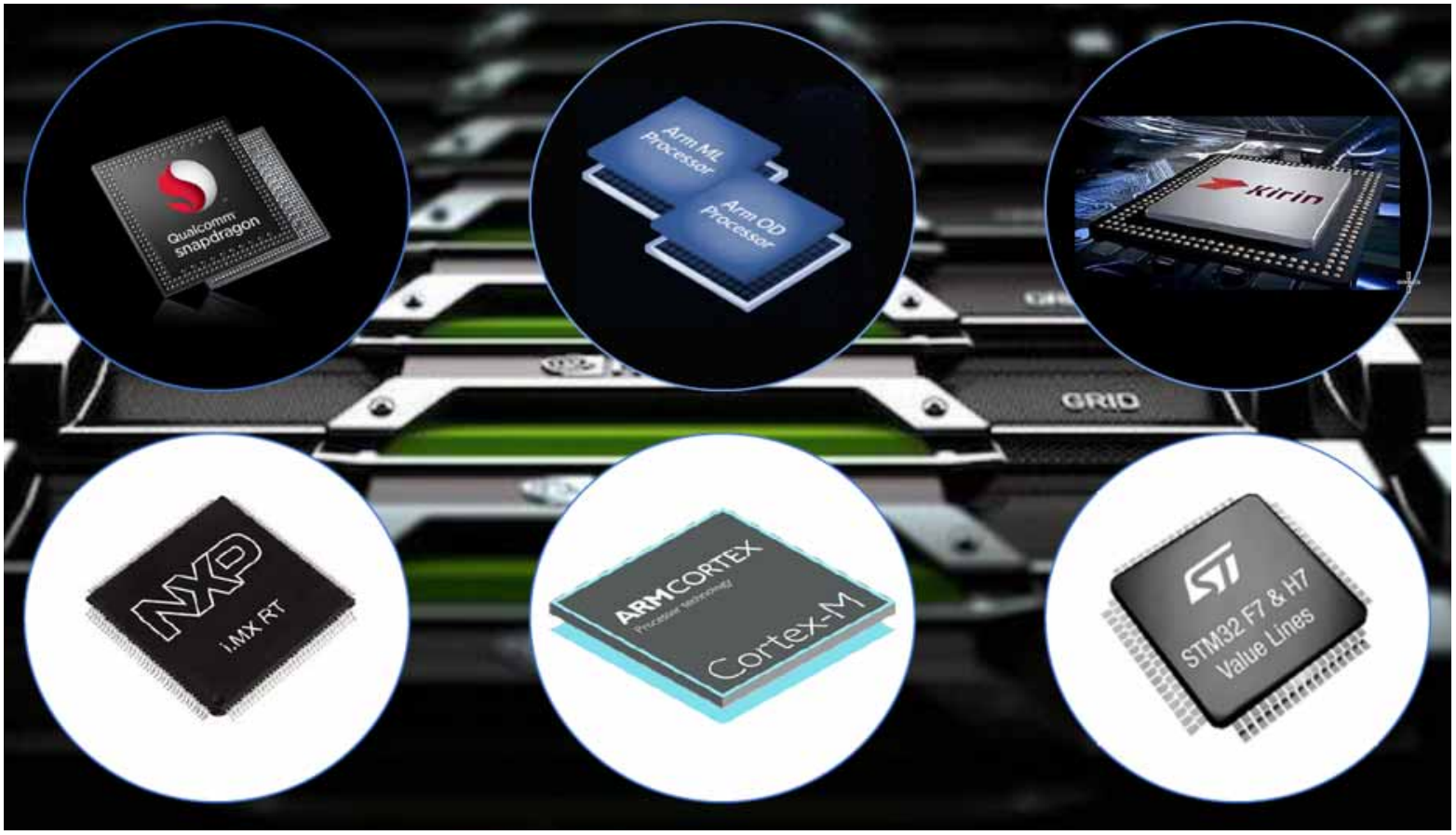



EMBEDDED EDGE COMPUTING
OFFRE ARROW

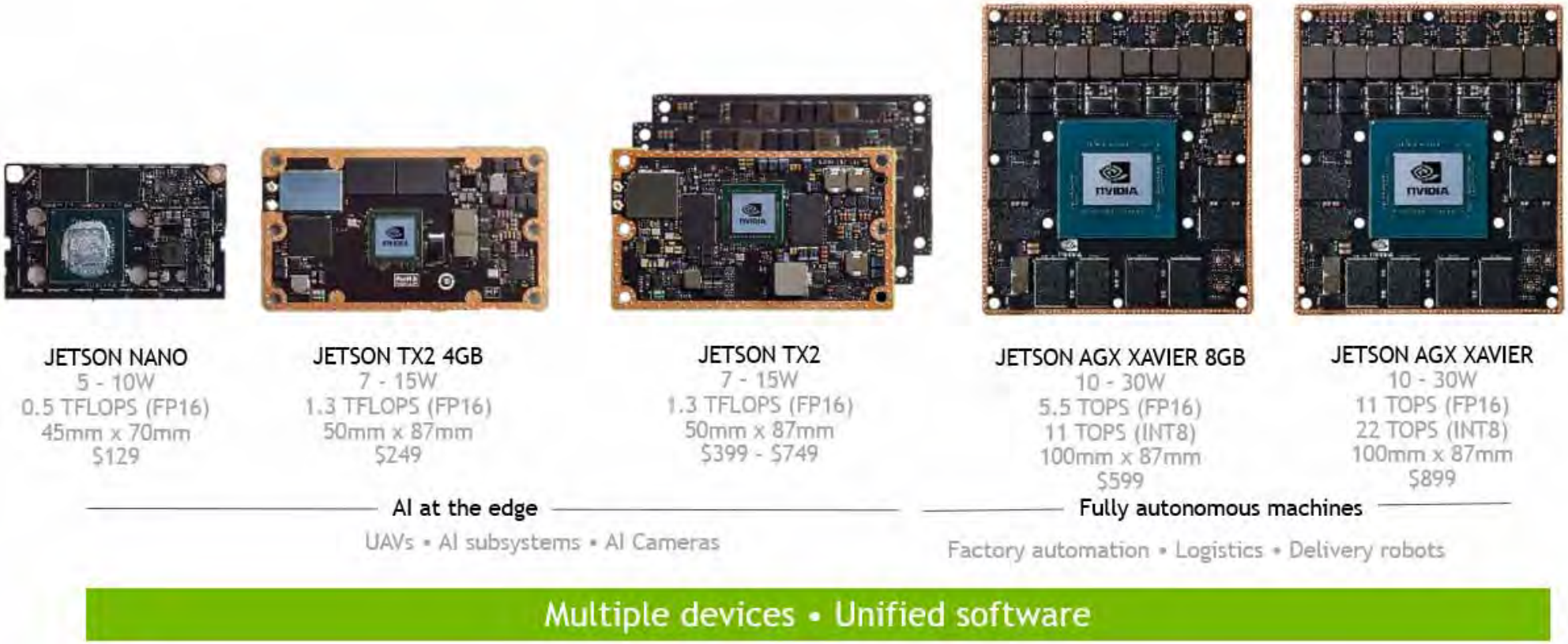
SOLUTIONS ARROW
INFERENCE on EDGE COMPUTING



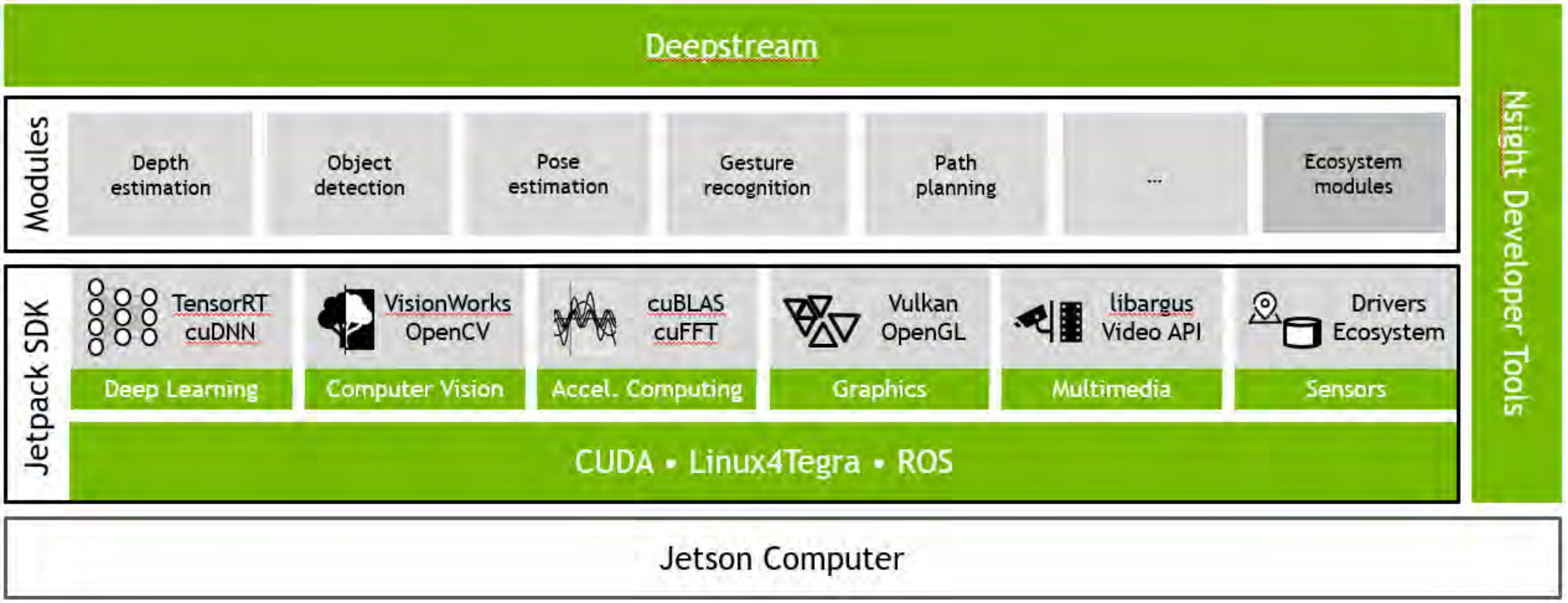
INFERENCE OFFRE CPU / MCU



INFERENCE : Offre GPU NVIDIA



LOGITIELS : Offre GPU NVIDIA



INFERENCE : Offre INTEL

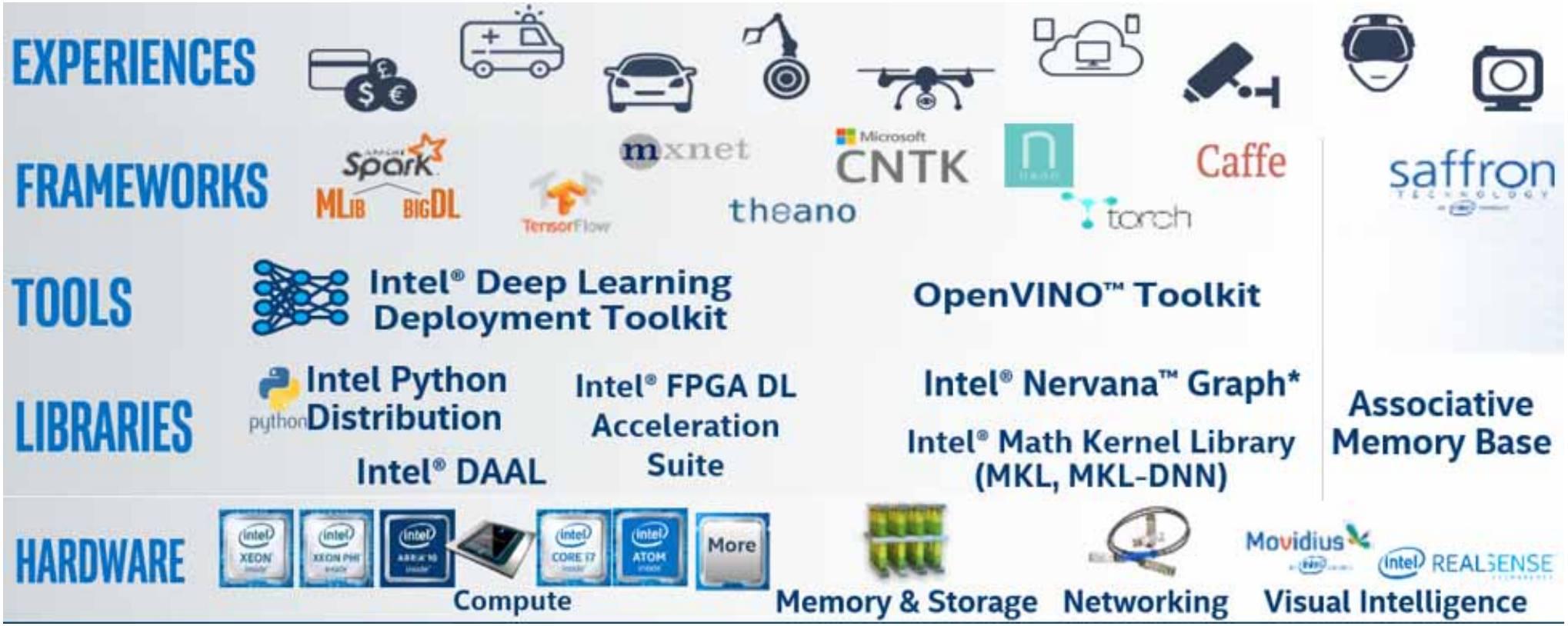
INTEL® ARRIA® 10 DEVELOPMENT KIT

- Versatile FPGA development kit packaged for optimal performance at 20 nm*
- Native abilities for variable precision INT4, INT8, FP11, FP16, FP32
- Evaluation ready for taking your edge applications from prototype to production

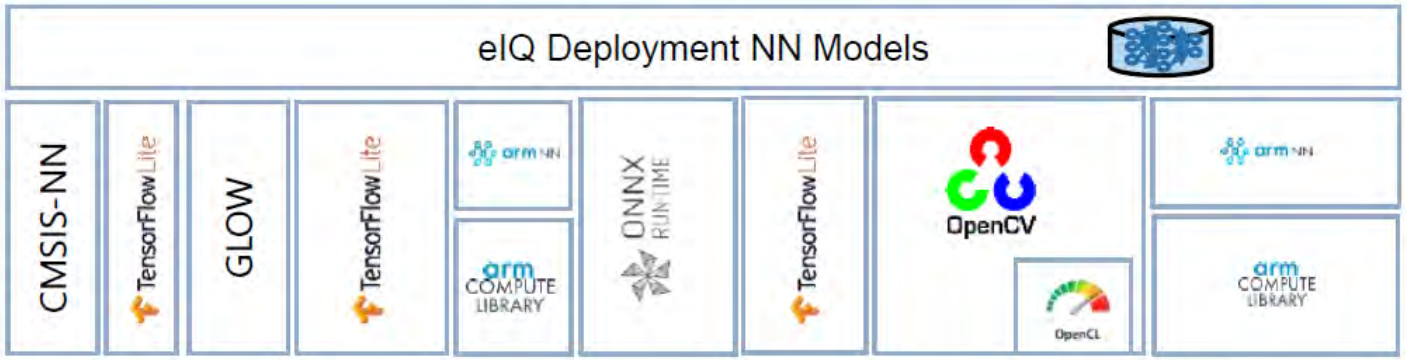
Intel PROGRAMMABLE ACCELERATION CARD WITH INTEL ARRIA 10

- Simplifying FPGA use for servers
- Supports Acceleration Stack for Intel Xeon CPU with FPGAs for easy app deployment
- Plug and play architecture for simple insertion and configuration in minutes
- FPGA Interface Manager pairs with Intel Xeon Processor over a PCIe Express bus

LOGITIELS : Solutions INTEL



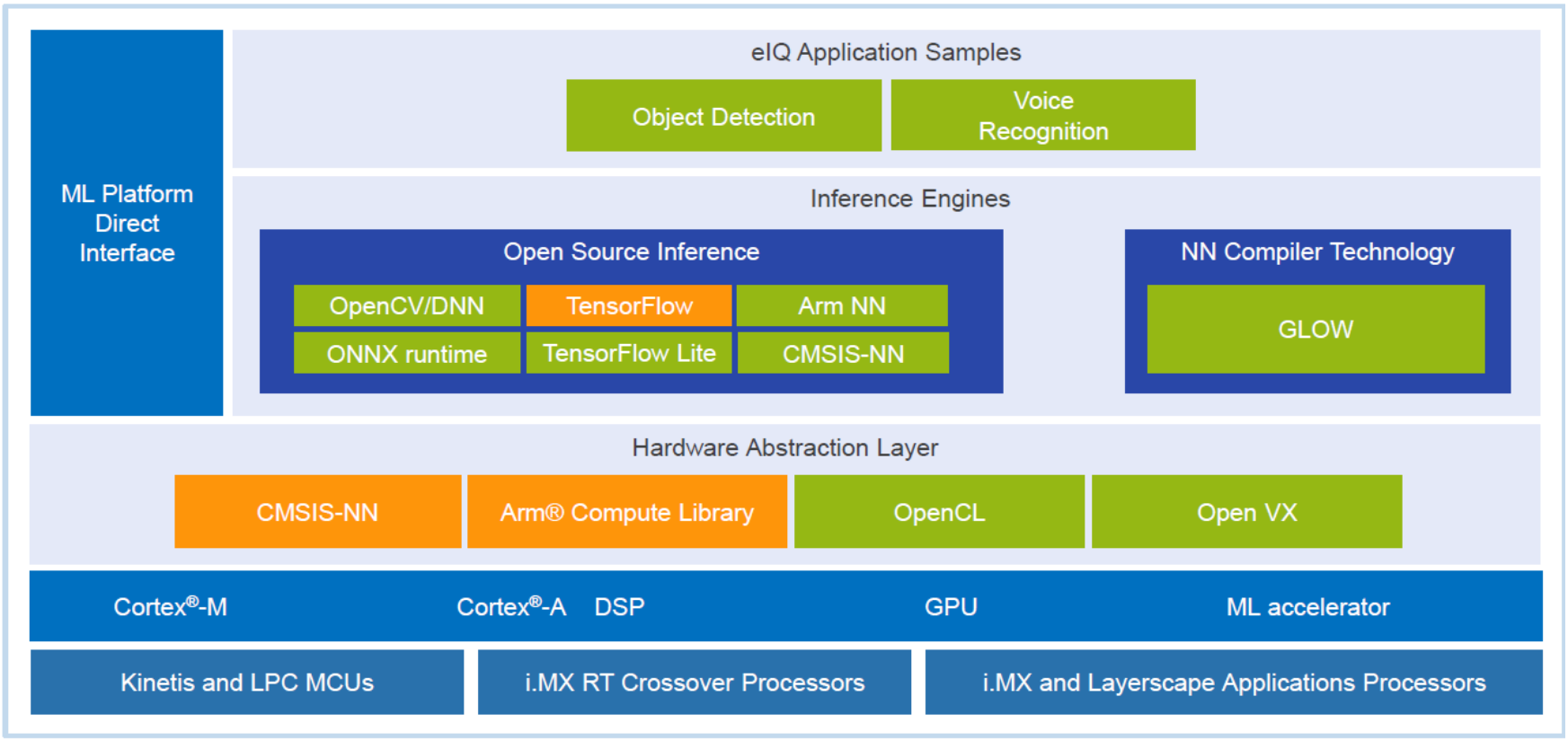
INFERENCE : Offre NXP



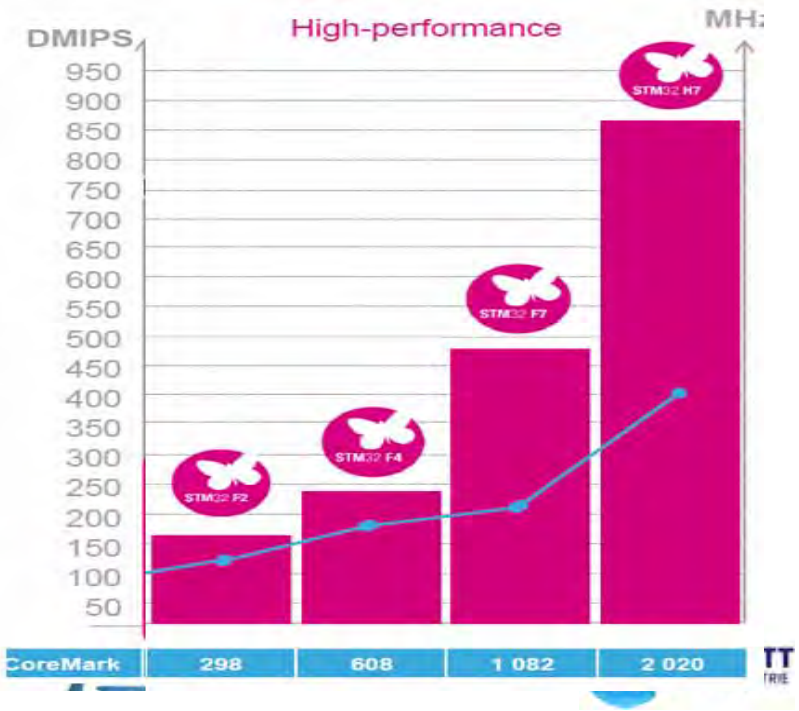
Embedded Compute Engines	Cortex-M		DSP	Cortex-A			GPU		
i.MX 8QM	*	*		NOW	May '19	May '19	NOW	July '19	July '19
i.MX 8QXP	*	*		NOW	*	*	NOW	July '19	July '19
i.MX 8M Quad	*	*		NOW	*	*	NOW	July '19	July '19
i.MX 8M Mini	*	*		NOW	*	*	NOW		
i.MX 6 and 7	* (only some models)	* (only some models)		*	*	*	*		
LS1, LS2, LX2	---	---		*	*	*	*		
i.MX RT600	TBD	TBD		---	---	---	---		
i.MX RT1050/1060	NOW	May '19		---	---	---	---		

LOGITIEL : Solutions NXP

eIQ-Core Machine Learning Software Development Environme



INFERENCE : Offre STMicro



Five Years Out

LOGITIEL : Offre STMicro

Input your framework-dependent, pre-trained Neural Network into the **STM32Cube**.AI conversion tool

Automatic and fast generation of an STM32-optimized library

STM32Cube.AI offers interoperability with state-of-the-art Deep Learning design frameworks

Train NN Model

ConvNetJS
Deep Learning in your browser
Lasagne
Caffe

STM32 Cube.AI

Process & analyze new data using trained NN

STM32

Convert NN into optimized code for MCU

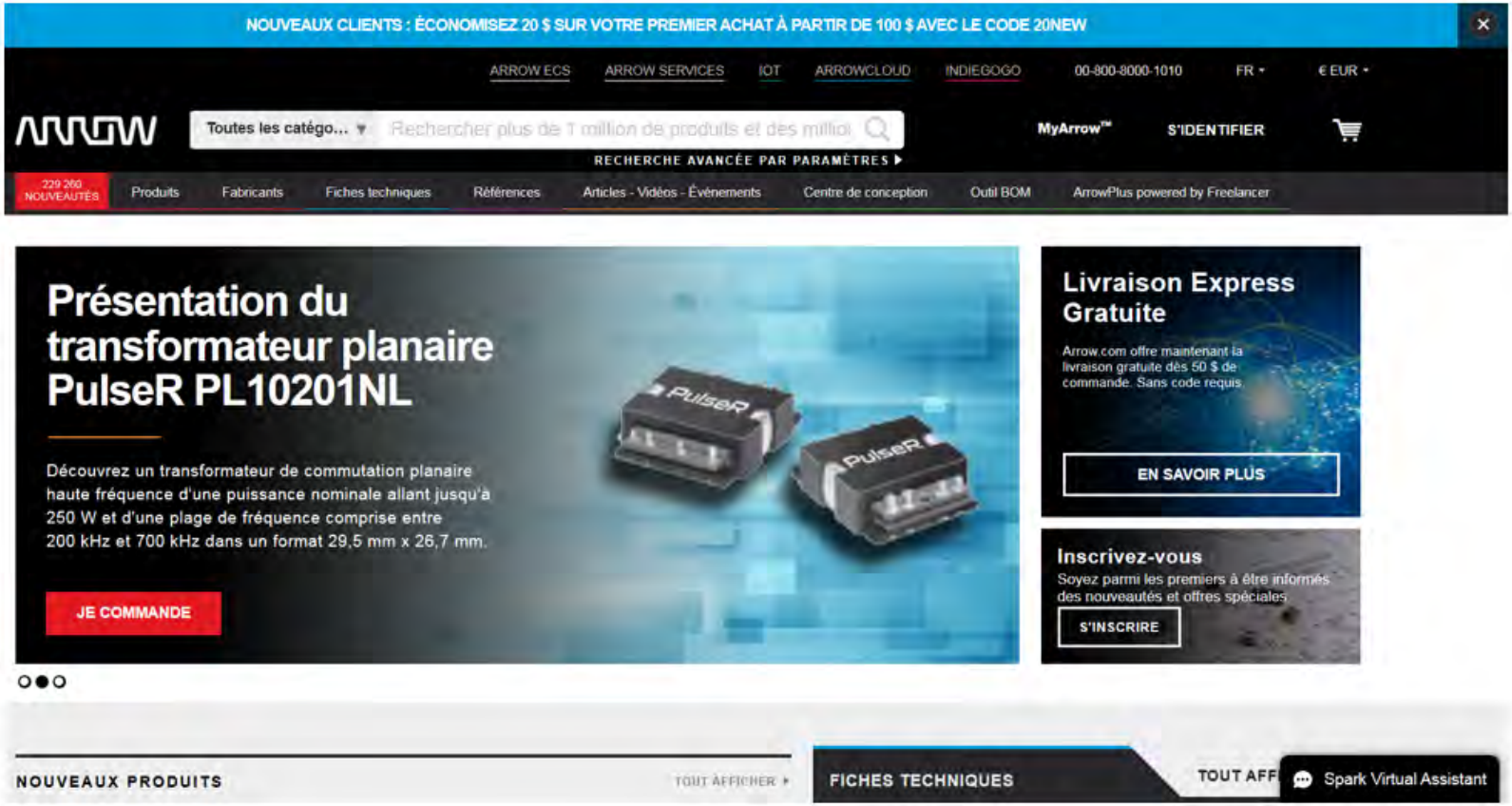
life.augmented

Five Years Out

CRESITT INDUSTRIE

OFFRE ARROW : CAPTEURS

Pour vos Prototypes et Pré Série
Pensez ARROW.COM




ANNEXES
EXEMPLES de REALISATIONS

Today's ST Motion Sensor Offer

Consumer, Industrial, Automotive

MOTION SENSORS	FEATURES	APPLICATIONS	PRODUCTS
Accelerometer Movement, Shock, Vibration, Wakeup, Tilt/Inclination Free fall	Movement detection		❖ Consumer, Movement detection • LIS2DE12, LIS2DH12, LIS2DW12, LIS2DTW12 ❖ Industrial, Tilt, Vibration • IIS2DH, IIS2DLPC, IIS2ICLX*, IIS3DWB* ❖ Automotive, Alarm, shock • AISS328DQ, AISS3624DQ, AIS2DW12*
6-axis IMU Combo gyroscopes and accelerometer sensors	Rotation for high accuracy movement monitoring		❖ Consumer, Movement recognition • LSM6DSO, LSM6DSOX, LSM6DSR ❖ Industrial, Robot • ISM330DLC, ISM330DHCX ❖ Automotive, Telematics • ASM330LHH
Compass Standalone magnetometer for magnetic field measurement, combo with accelerometer	Magnetic field + acceleration		❖ Alarm, E-compass • LIS2MDL, LSM303AGR , LSM303AH ❖ Industrial, Anti-tamper • IIS2MDC, ISM303DAC

life.augmented

BUT THERE'S MORE THAT CAN BE DONE WITH ST SENSORS…

* Available soon

Inference – ARM MCU

Open-source neural network implementation on ARM utilizing DSP/SIMD functions

Fixed-point arithmetic (no floating point)

4.6x speed improvement over baseline ARM implementation

Example: Cifar-10 image classification, trained with Caffe

Accuracy of 80.3% with 10 samples/s on STM NUCLEO-F746ZG Cortex-M7 with 216MHz, 320KB SRAM

- Model trained using Caffe - must be converted to run effectively on ARM
- Memory and performance constrains

	Layer Type	Filter Shape	Output Shape	Ops	Runtime
Layer 1	Convolution	5x5x3x32 (2.3 KB)	32x32x32 (32 KB)	4.9 M	31.4 ms
Layer 2	Max Pooling	N.A.	16x16x32 (8 KB)	73.7 K	1.6 ms
Layer 3	Convolution	5x5x32x32 (25 KB)	16x16x32 (8 KB)	13.1 M	42.8 ms
Layer 4	Max Pooling	N.A.	8x8x32 (2 KB)	18.4 K	0.4 ms
Layer 5	Convolution	5x5x32x64 (50 KB)	8x8x64 (4 KB)	6.6 M	22.6 ms
Layer 6	Max Pooling	N.A.	4x4x64 (1 KB)	9.2 K	0.2 ms
Layer 7	Fully-connected	4x4x64x10 (10 KB)	10	20 K	0.1 ms
Total		87 KB weights	55 KB activations	24.7 M	99.1 ms

Layer type	Baseline runtime	New kernel runtime	Improvement	
			Throughput	Energy Efficiency
Convolution	443.4 ms	96.4 ms	4.6X	4.9X
Pooling	11.83 ms	2.2 ms	5.4X	5.2X
ReLU	1.06 ms	0.4 ms	2.6X	2.6X
Total	456.4ms	99.1 ms	4.6X	4.9X

CMSIS-NN – Efficient NN Kernels for Cortex-M CPUs

Convolution

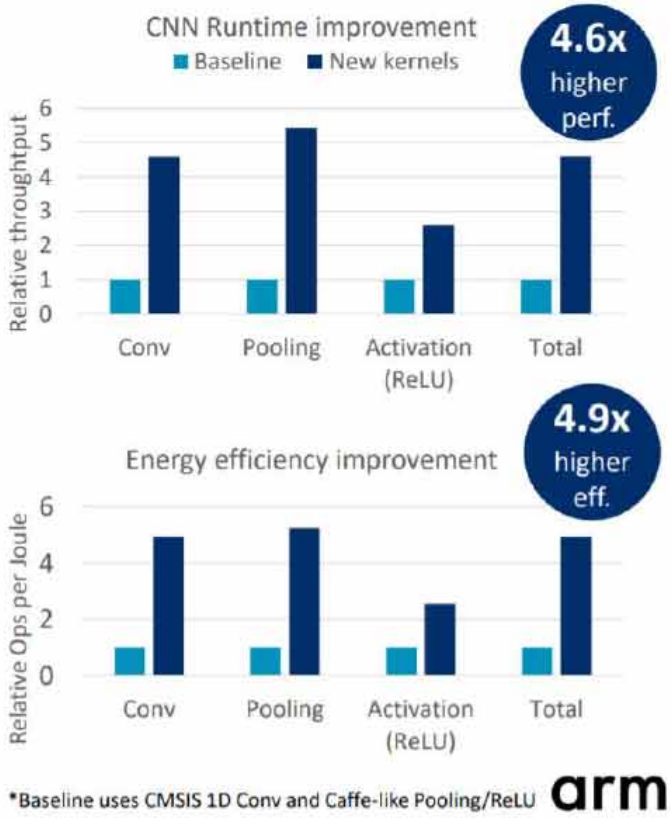
- Boost compute density with GEMM based implementation
- Reduce data movement overhead with depth-first data layout
- Interleave data movement and compute to minimize memory footprint

Pooling

- Improve performance by splitting pooling into x-y directions
- Improve memory access and footprint with in-situ updates

Activation

- ReLU: Improve parallelism by branch-free implementation
- Sigmoid/Tanh: fast table-lookup instead of exponent computation



IA EDGE COMPUTING : Exemples

Secure, low-power smart home security: Uses on-device, always-on, motion, person, and sound detection to identify family members or intruders. Starts recording only when it detects motion or sound, sending a notification to the user's smartphone.

Hospital staff/visitor tracking: Alerts receptionists to unknown people or unauthorized access. Edge recognition means images of visitors and patients are never stored or transmitted.

Plant disease detection: Uses an image recognition smartphone app to detect disease with near 100 percent accuracy — even off-network.

Faster produce selection: Classifies fruit and vegetables by camera, automatically identifying different categories and improving production-line efficiency.

IA EDGE COMPUTING :

Drone avionics: Recognizes and follows a target while avoiding obstacles, via camera-based vision and movement prediction.

No-latency driver assistance: Helps to reduce collisions using cameras, motion sensors, and GPS to understand and guide driver behavior in real time.

Improved human-machine interaction: Streamlines interactions, boosts productivity, and creates a smoother user experience across devices.

On-device translation: Makes communication possible — however remote the location — and avoids costly roaming charges.

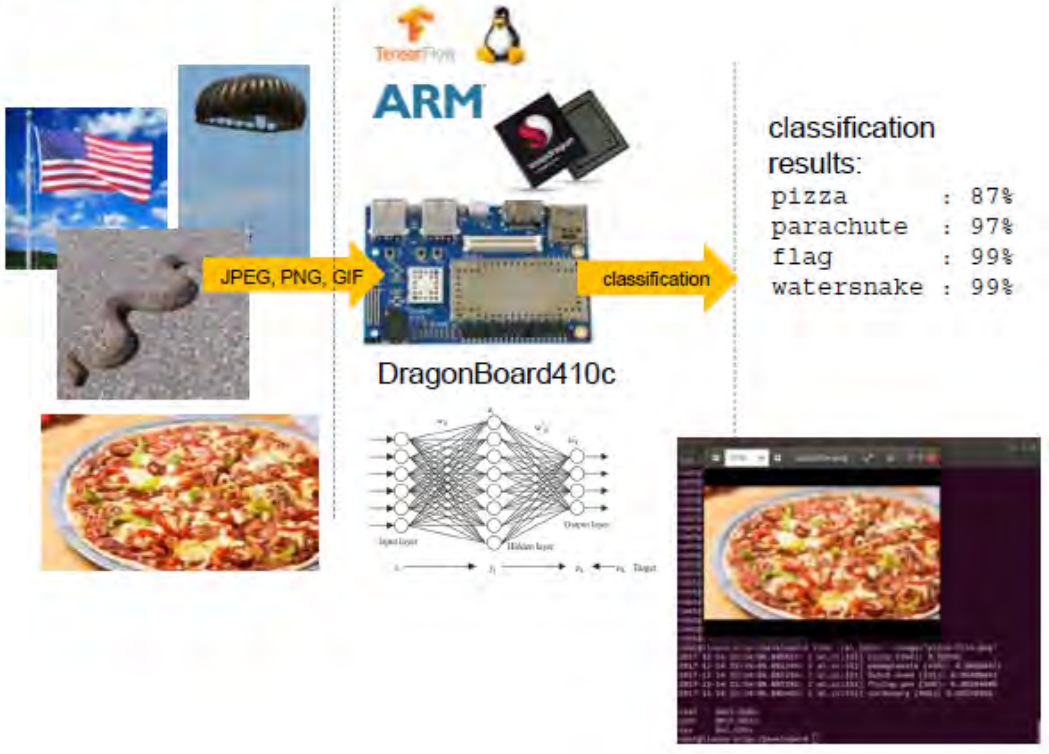
Device optimization: Significantly extends battery life by optimizing operating system scheduling for individual applications.

Tensorflow SDK on DragonBoard 410c

Problem statement:

How to simplify edge AI deployment?

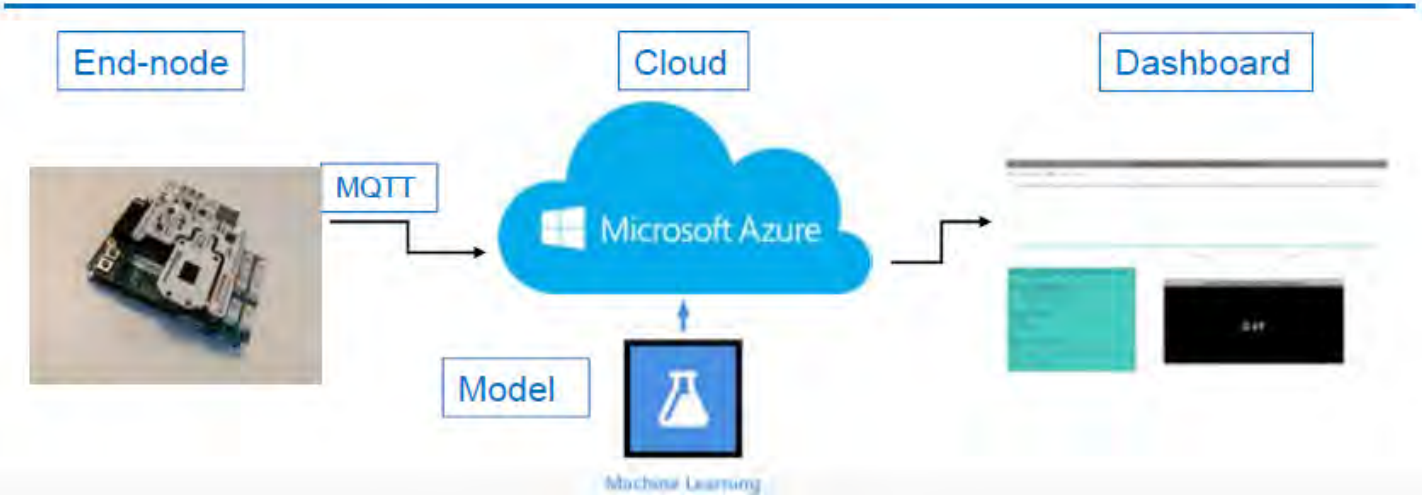
- Enable most widely AI framework on ARM-based systems
- Seamless transition – Cloud trained model directly to embedded device
- Benchmarking for evaluation expertise (board capabilities vs. use case)



AI in IoT & Cloud

Most AI deployments use Cloud-based Deep Neural Networks

- Telemetry is collected by end-nodes (cameras, sensors)
- Data is streamed to the cloud (MQTT / HTTPS)
- Processing and inference (classification) by model in the cloud

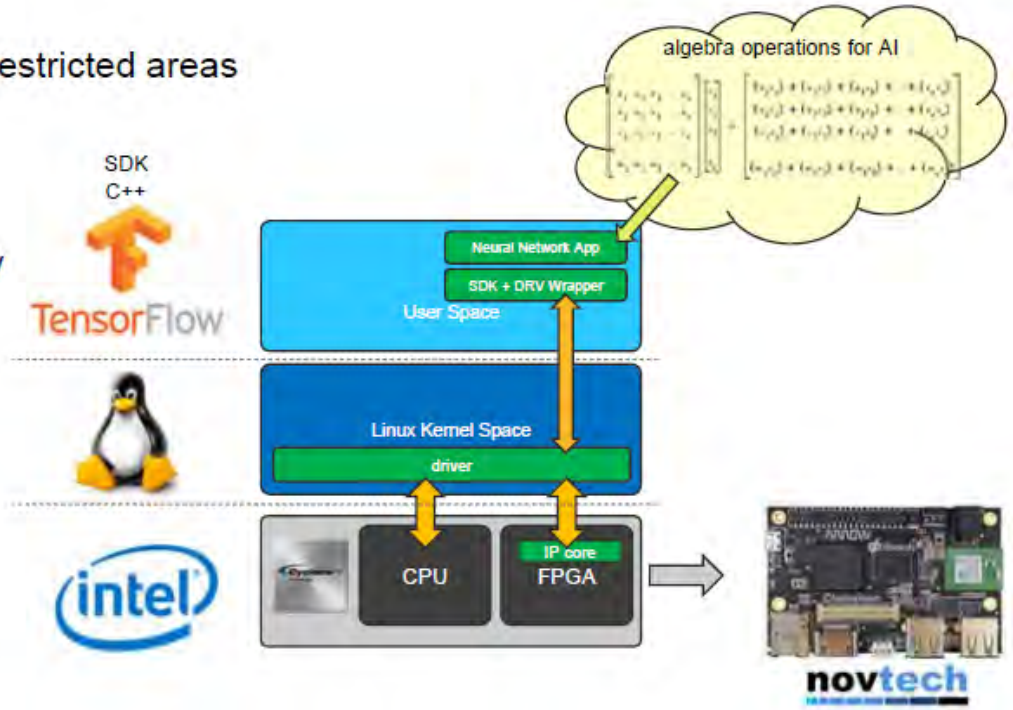


Face recognition

Problem statement:

Camera for people monitoring in disconnected or restricted areas

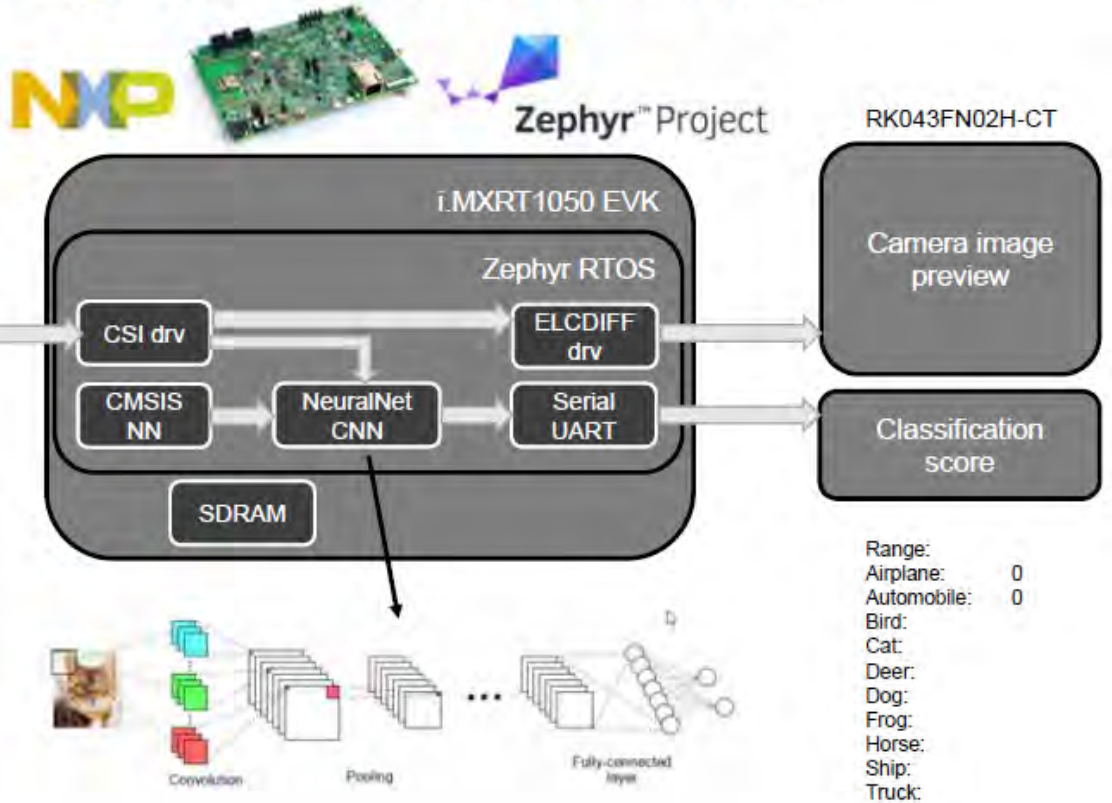
- Model training with data of personnel
- Sensitive data and access prohibits connectivity



Cifar10 Image Recognition demo on RTOS

Problem statement:
Can MCUs be utilized for Edge AI?

- Full open-source solution on ARM Cortex-M MCUs
- Inference time vs. MCU clock vs. price



Food detection

Dedicated Machine Learning software stack developed for embedded devices

System recognizes type of food and displays relevant information

A proof of concept: how to improve the comfort of the user?

Tune heating time, allergen information, integrate with fitness apps, …

- Inference time depends on core frequency & GPU availability
- More complex networks require more memory footprint

Qualcomm DragonBoard 410C / NXP i.MX 8QMax	NXP i.MX RT
Quad Cortex-A53 1.4/1.5 GHz	Cortex-M7 600MHz
Food ID (20 classifier)	Food ID (5 classifier)
6ms inference	66ms inference
99% accuracy	99% accuracy

Industrial IoT

Autonomous decisions in time-critical tasks improve efficiency

Predictive maintenance reduces costs and ensures production continuity

Quick abnormality detection minimizes possible damage

AI in the edge - sensitive data is processed locally

High reliability is key to responsiveness

Introducing AI can be done with low impact to production

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

MC27561-DRAGONFLY
SmartMesh IP

- Improved response time, no latency
- No factory network bandwidth utilization
- Sensitive data not exposed to external network
- Constrained performance