

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

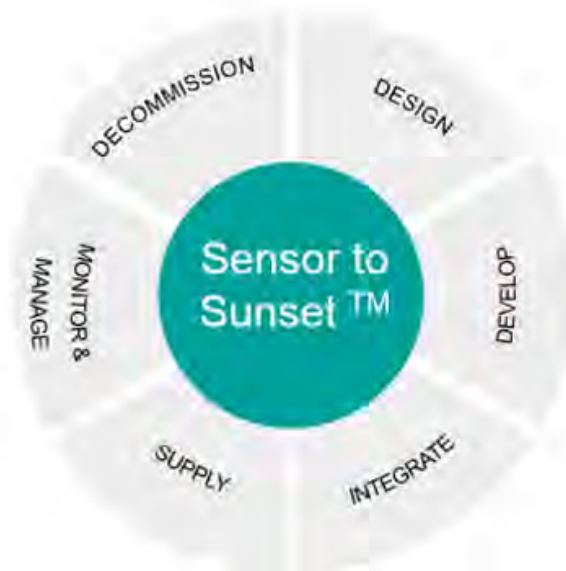
- SW Integration
- POS terminal, ATM,
- Embedded x86 apps
- Industrial handheld
- Kiosks & Gateways
- Digital Signage
- Security & IP Cameras



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™**

A hand is shown holding a glowing globe. The globe is covered in a network of white lines and nodes, with several nodes containing a person icon. The background is dark blue with faint binary code and a satellite-like graphic. The text 'IA' is in the top right, and 'MARCHE & APPLICATIONS' is in the center.

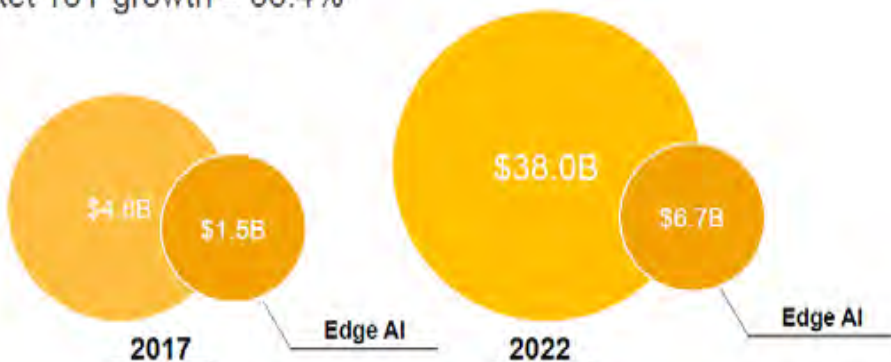
IA

MARCHE & APPLICATIONS

IA : MARCHE EN CROISSANCE

Global AI market YoY growth – 51%

Edge AI market YoY growth – 35.4%



IA : APPLICATIONS

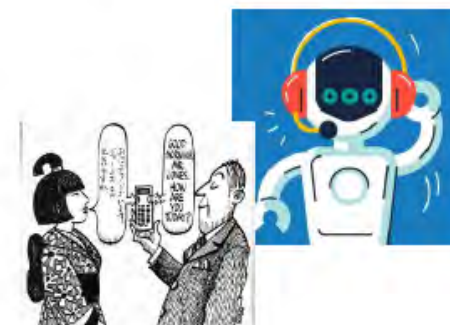
Image Recognition



Speech Recognition & Synthesis



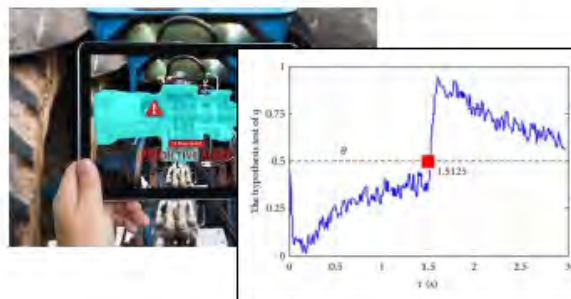
Natural Language Processing



Fintech & Insurance



Predictive Maintenance



Robotics



IA : FACTEURS DE CROISSANCE

DISPONIBILITE de DONNEES REELLES

Ex: Données des Réseaux Sociaux
Données des applications WEARABLE



Source : INTEL

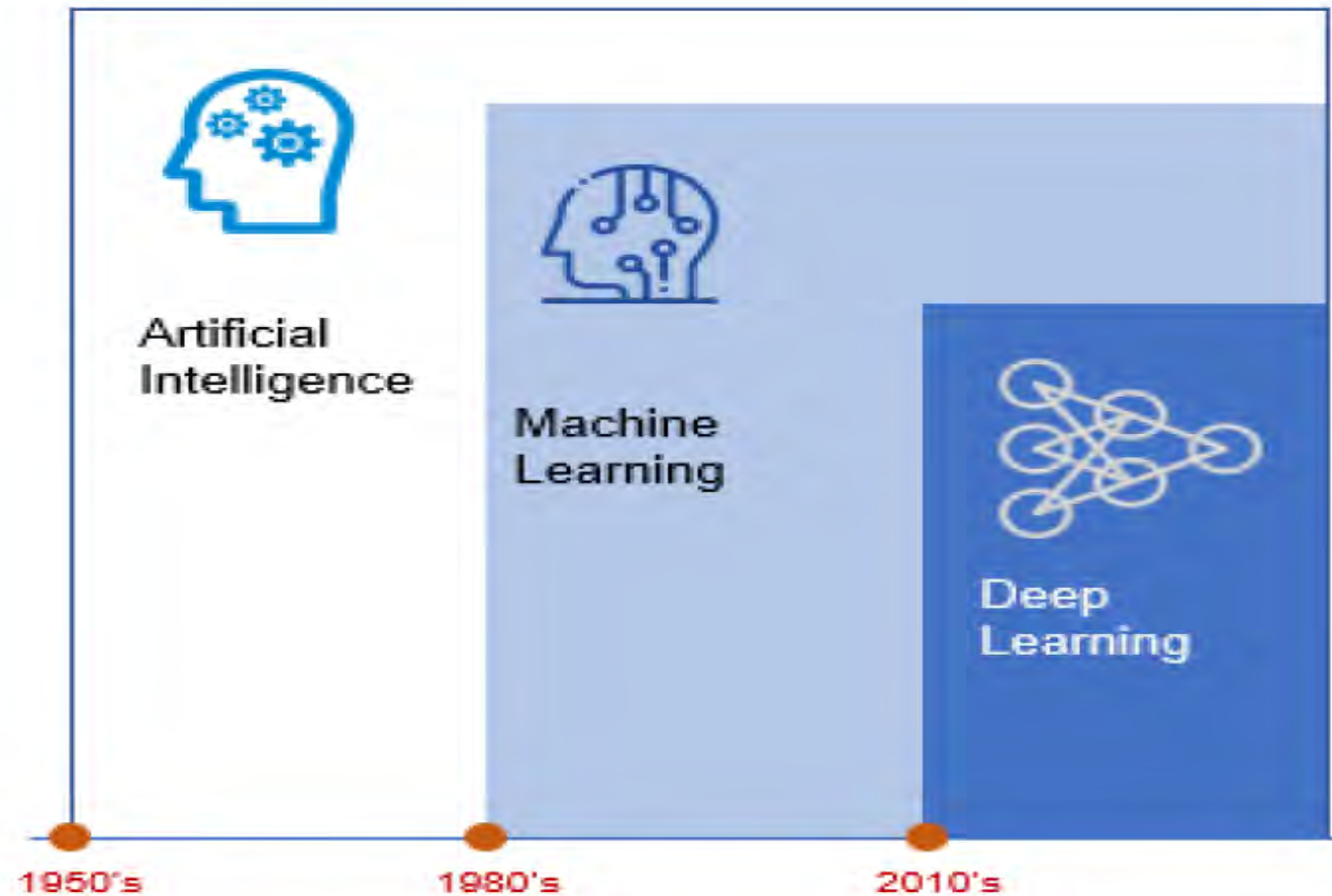
DISPONIBILITE de MACHINES DE TRAITEMENT HAUTE PERFORMANCES ABORDABLES

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

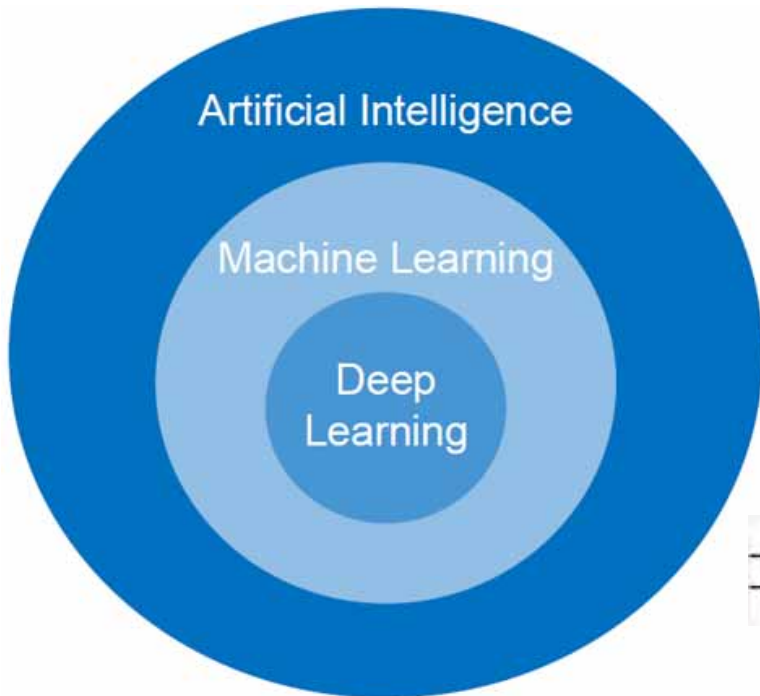
MACHINE LEARNING (ML)

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

DEEP LEARNING (DL)



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



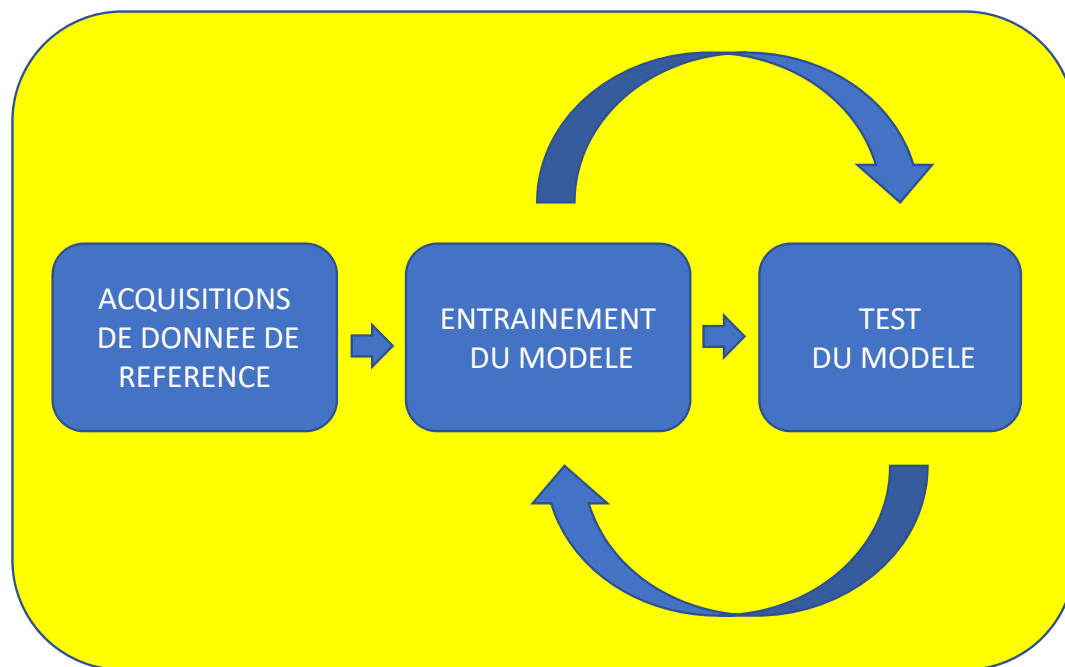
A hand is shown holding a glowing globe. The globe is covered in a network of white lines and nodes, with several nodes containing a person icon. The background is dark blue with faint binary code and a satellite-like structure. The text 'IA' is in the top right, and 'MISE EN OEUVRE' is in the center right.

IA

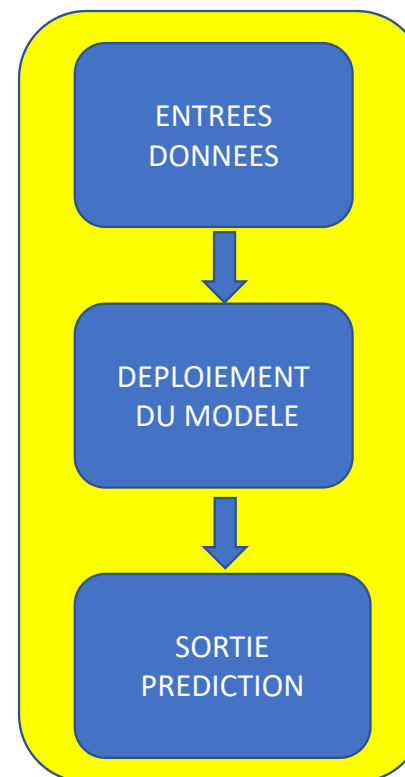
MISE EN OEUVRE

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

BANDWIDTH	LATENCY	PRIVACY	AVAILABILITY
			
1 billion cameras WW (2020) 10's of petabytes per day	30 images per second <u>200ms</u> latency	Confidentiality Private cloud or <u>on-premise</u> storage	50% of populated world < 8mbps Bulk of uninhabited world no 3G+

INFERENCE on CLOUD COMPUTING

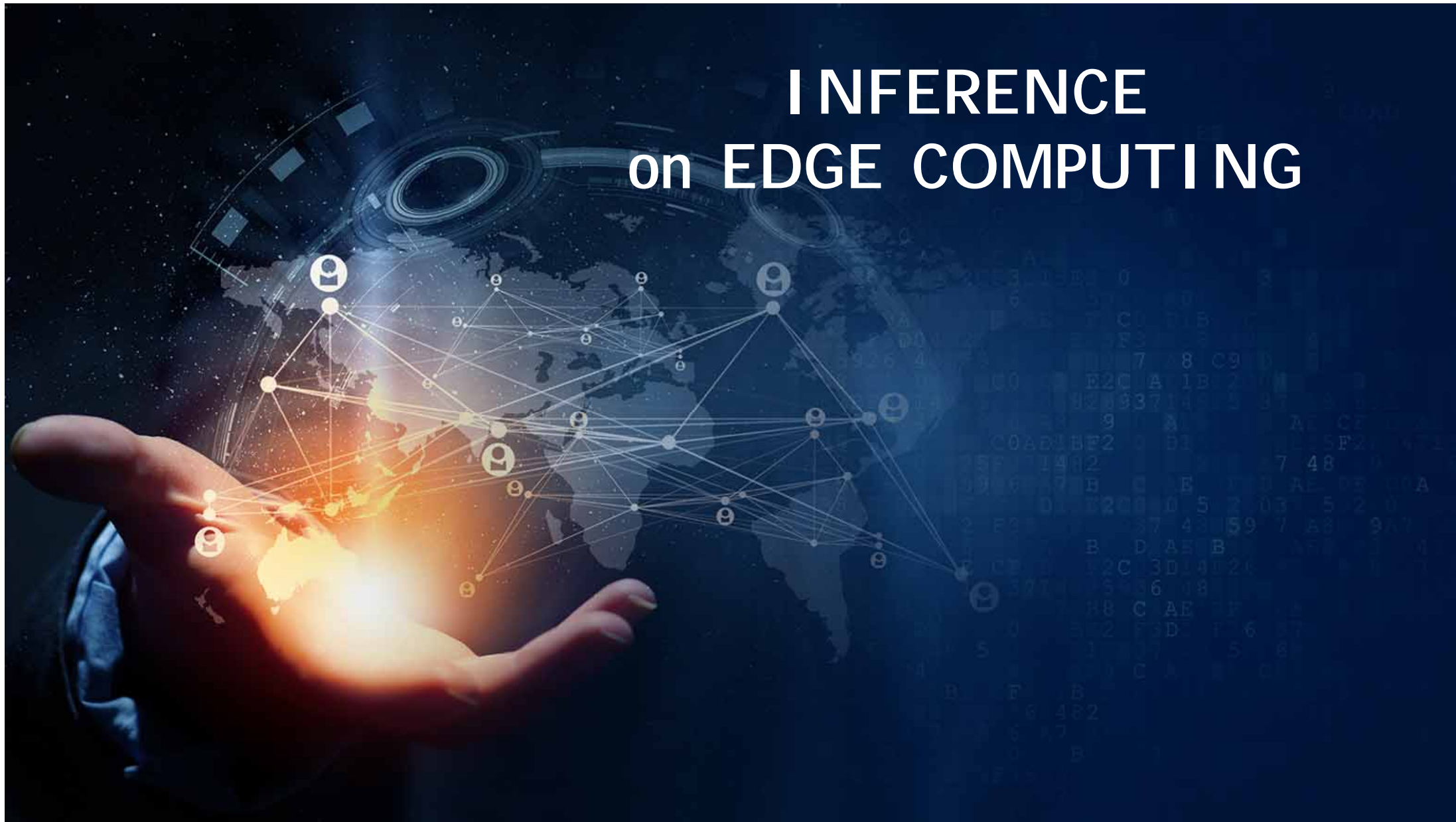
AVANTAGES

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

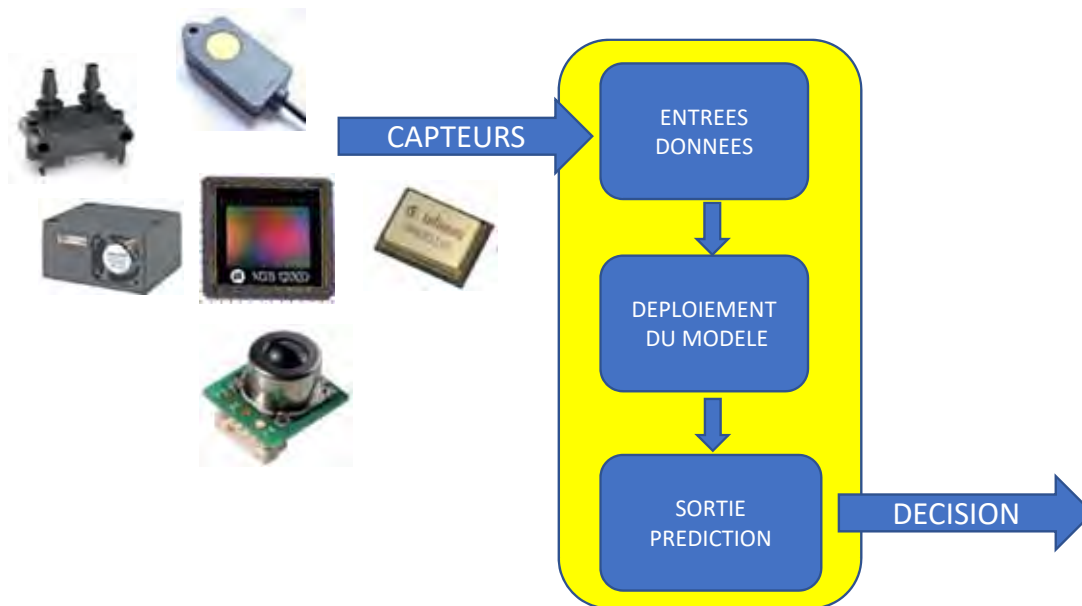
- Données Privées et Sécurisées
- Consommation Faible
- Coûts réduits
- Taille Réduite
- Temps de réponse court
- Pas de Connexion Internet

INCONVENIENTS

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

INFERENCE on EDGE COMPUTING :

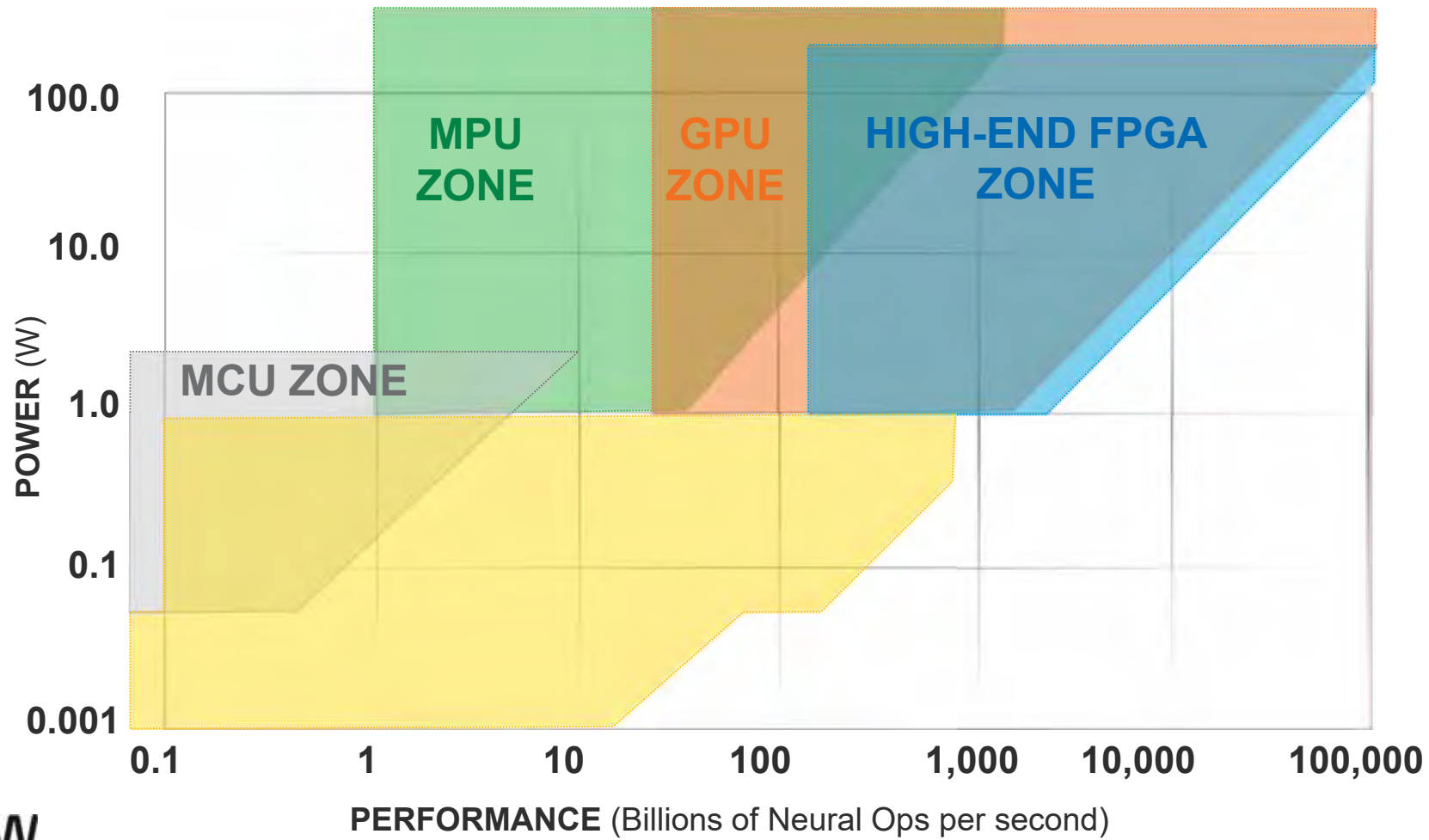
Principe



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

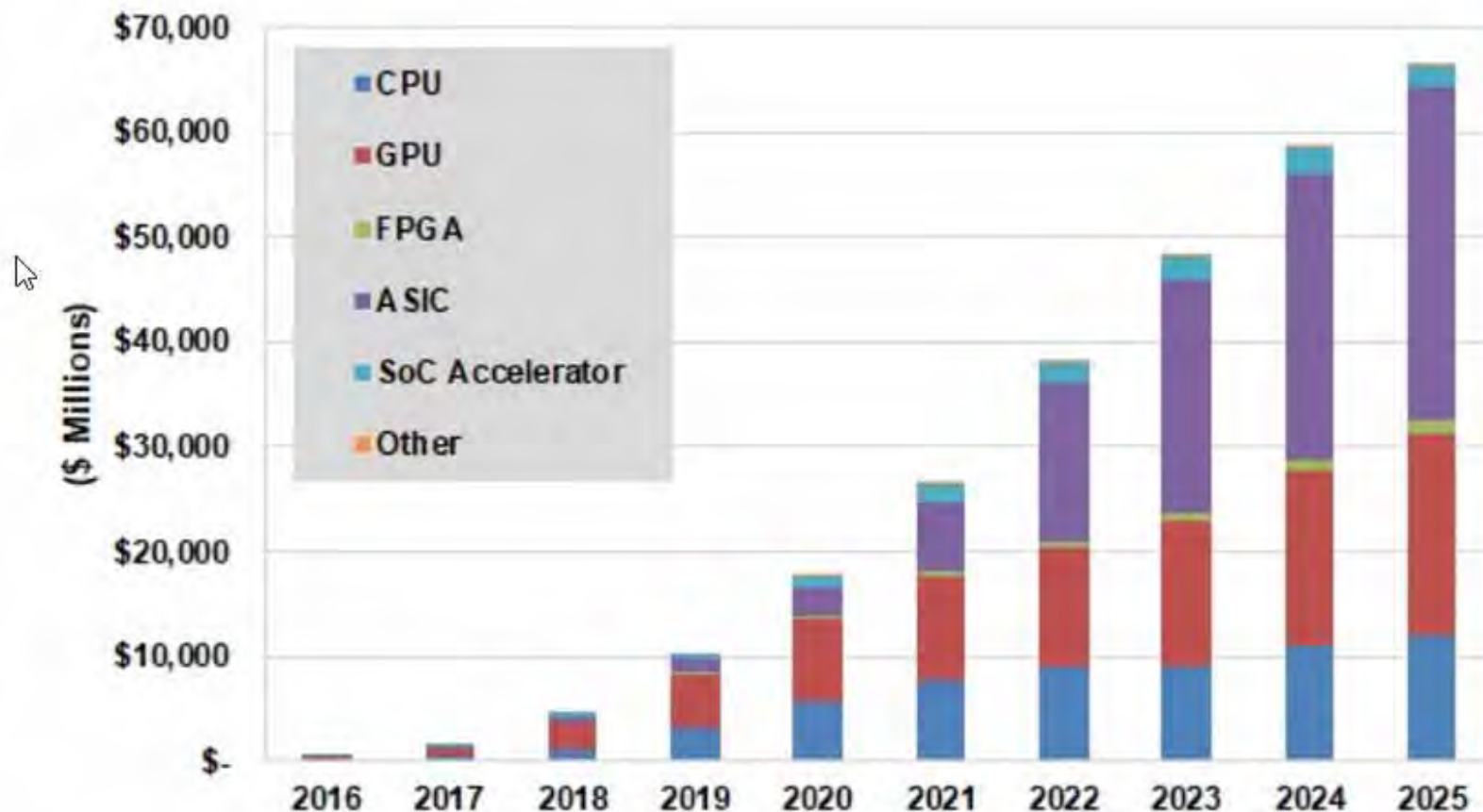
INFERENCE on EDGE COMPUTING

Consummation VS Performances



INFERENCE on EDGE COMPUTING

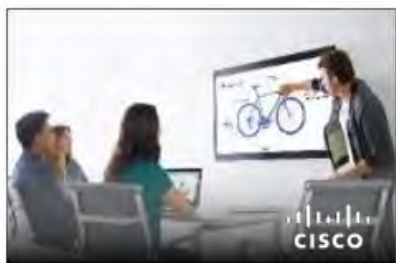
Tendances du marché



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

Michel VINEZ - 17 Oct. 2019

EDGE COMPUTING : Exemples



Enterprise Collaboration



Factory Automation



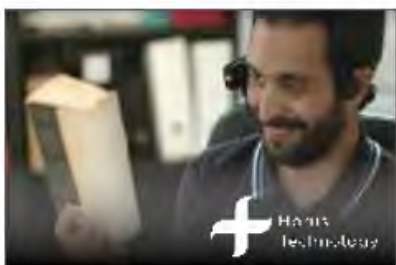
Service Robotics



Package Delivery



AI City



Personal Assist



Search and Rescue



Industrial Inspection



Portable Medical



Academia and Research

A hand is shown holding a glowing globe. The globe is covered with a network of white lines and nodes, representing a neural network or data flow. The background is dark blue with faint binary code (0s and 1s) and a circular pattern resembling a globe or a data visualization. The overall theme is artificial intelligence and global connectivity.

DEEP LEARNING

OFFRE ARROW

TRAINING : Offre GPU NVIDIA



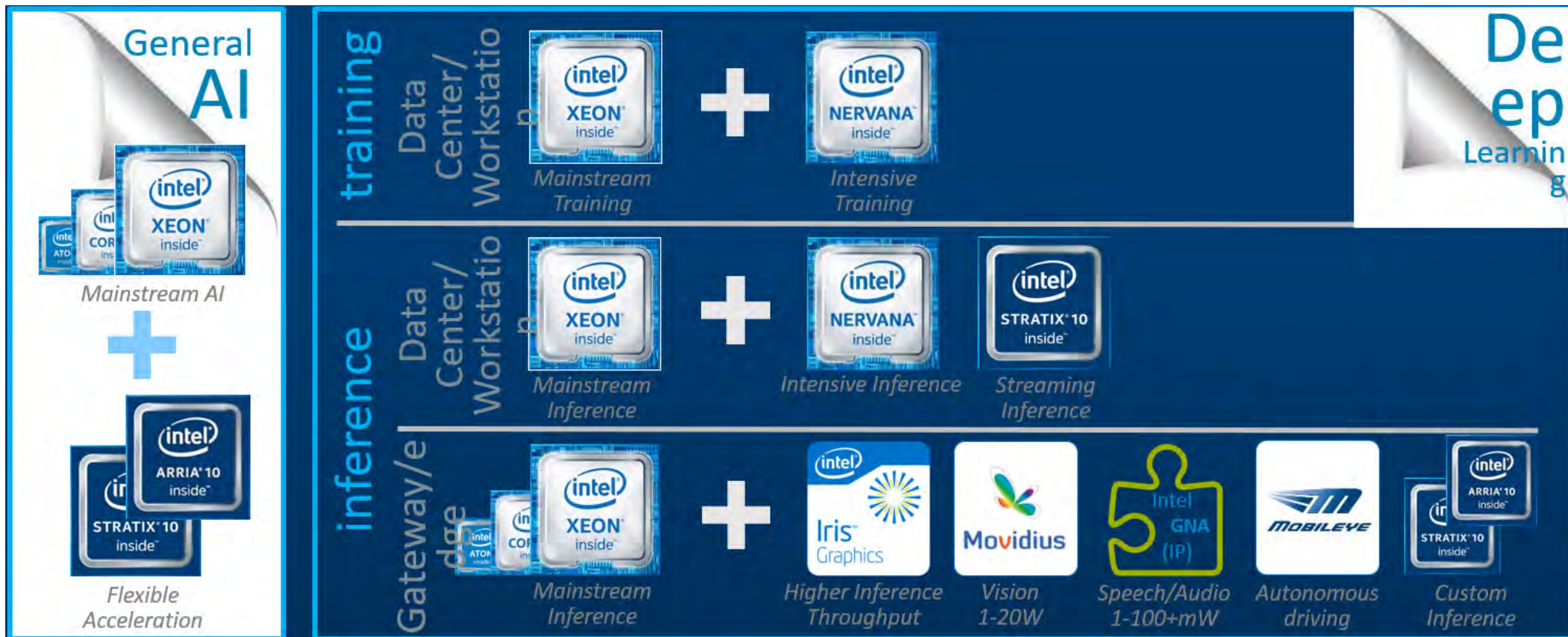
LOGITIELS : Offre GPU NVIDIA

World's Leading Data Center Platform for Accelerating HPC and AI

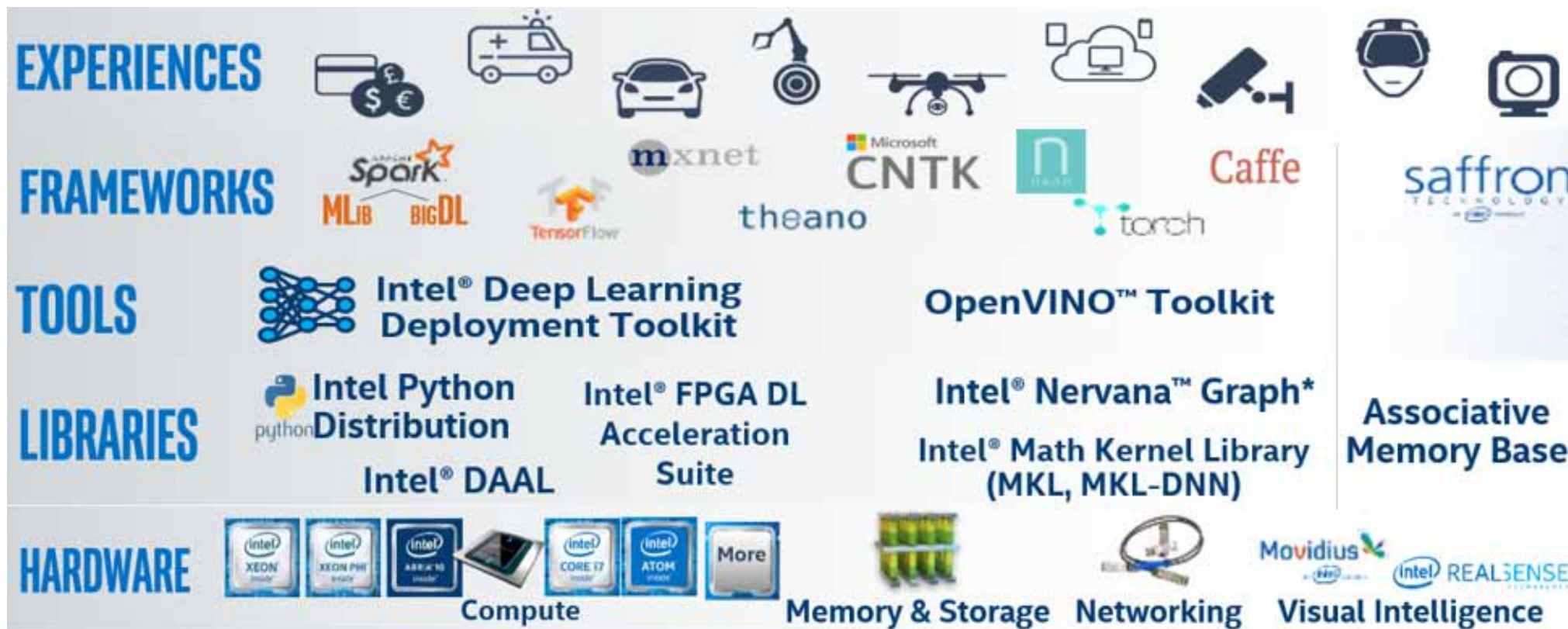
The diagram illustrates the NVIDIA ecosystem, organized into four horizontal layers:

- CUSTOMER USECASES:** This layer is divided into three categories: **CONSUMER INTERNET** (Speech, Translate, Recommender), **ENTERPRISE APPLICATIONS** (Healthcare, Manufacturing, Engineering), and **SUPERCOMPUTING** (Molecular Simulations, Weather Forecasting, Seismic Mapping).
- INDUSTRY FRAMEWORKS & APPLICATIONS:** This layer lists various frameworks and applications. On the left, it includes Caffe2, Chainer, KALDI, mxnet, PaddlePaddle, PYTORCH, and TensorFlow. On the right, it lists Amber, ANSYS, CHROMA, GROMACS, LAMMPS, NAMM, SIMULIA, and VASP, along with a note for "+550 Applications".
- NVIDIA SDK & LIBRARIES:** This layer lists the software development kits and libraries, including cuBLAS, cuDNN, cuFFT, cuRAND, cuSPARSE, DeepStream, NCCL, TensorRT, PGI, and OpenACC. A green bar labeled "CUDA" is positioned below these items.
- TESLA GPUs & SYSTEMS:** This layer shows the hardware and system partners. It includes Tesla GPU, NVIDIA DGX STATION, NVIDIA DGX, NVIDIA HGX-1, SYSTEM OEM (Dell, Hewlett Packard Enterprise, IBM), and CLOUD (AWS, Google Cloud Platform, Microsoft Azure).

TRAINING : Offre INTEL



LOGITIELS : Solutions INTEL



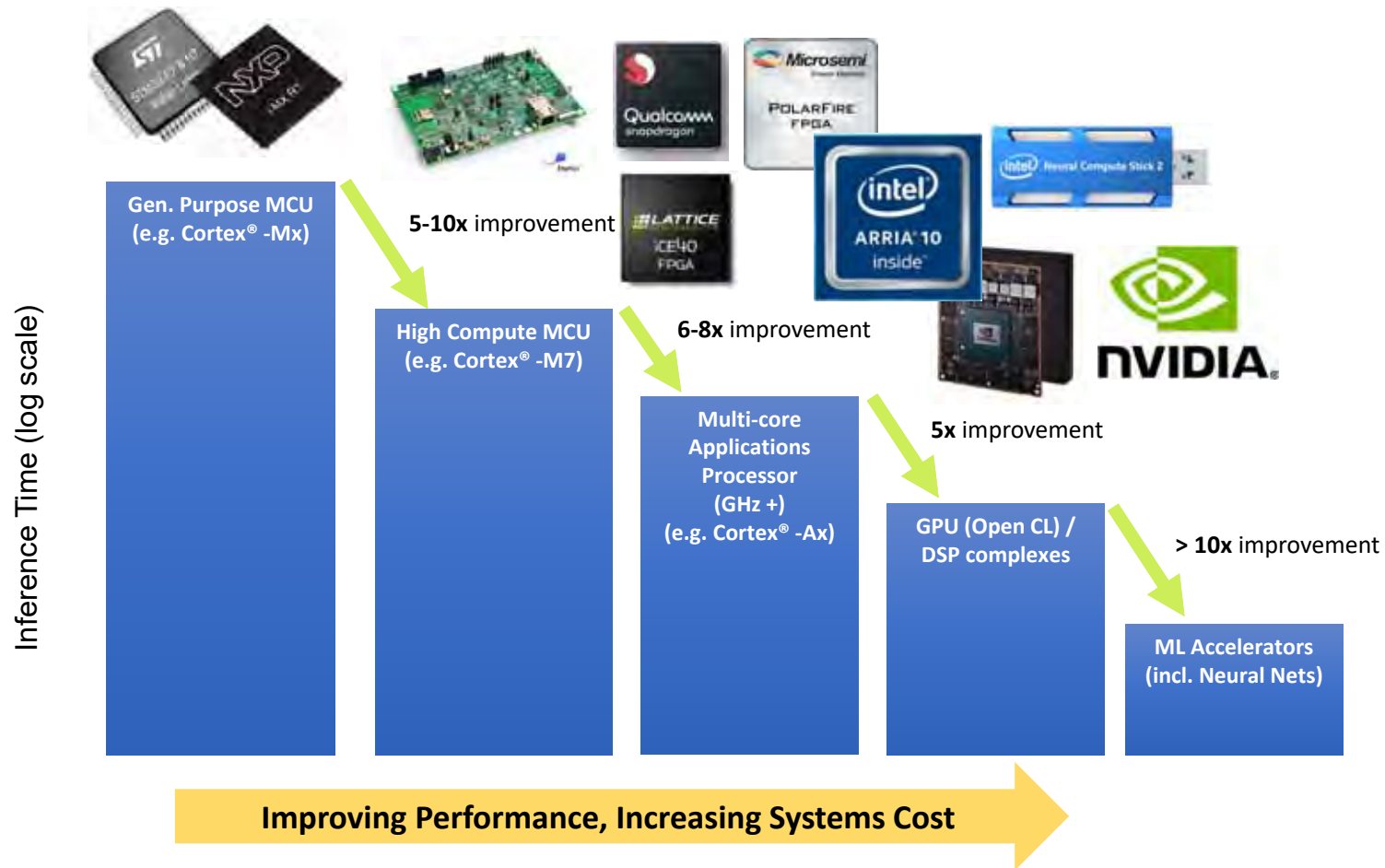
EMBEDDED EDGE COMPUTING

OFFRE ARROW

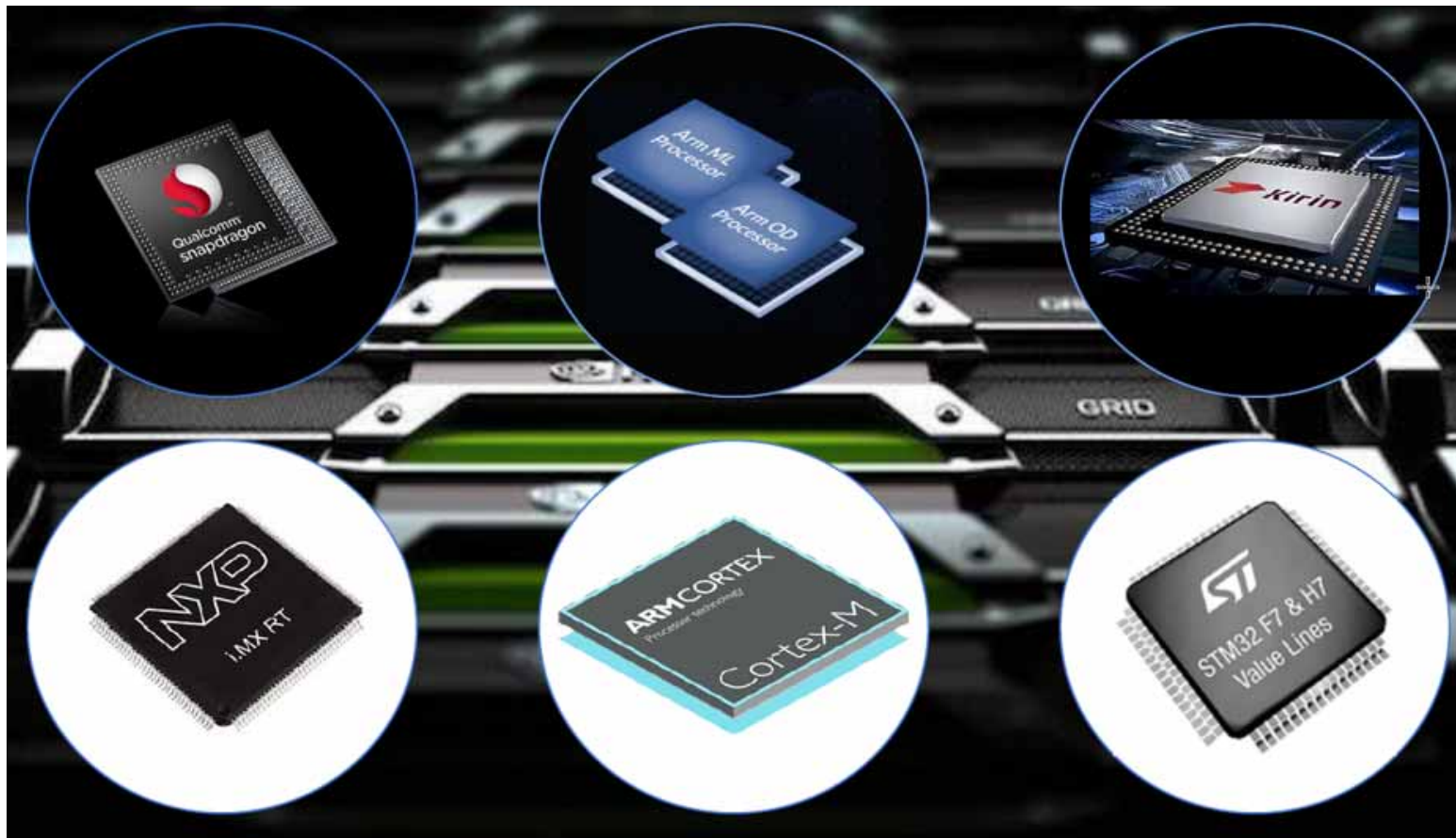


SOLUTIONS ARROW

INFERENCE on EDGE COMPUTING



INFERENCE OFFRE CPU / MCU



INFERENCE : Offre GPU NVIDIA



JETSON NANO
5 - 10W
0.5 TFLOPS (FP16)
45mm x 70mm
\$129



JETSON TX2 4GB
7 - 15W
1.3 TFLOPS (FP16)
50mm x 87mm
\$249



JETSON TX2
7 - 15W
1.3 TFLOPS (FP16)
50mm x 87mm
\$399 - \$749



JETSON AGX XAVIER 8GB
10 - 30W
5.5 TOPS (FP16)
11 TOPS (INT8)
100mm x 87mm
\$599



JETSON AGX XAVIER
10 - 30W
11 TOPS (FP16)
22 TOPS (INT8)
100mm x 87mm
\$899

AI at the edge

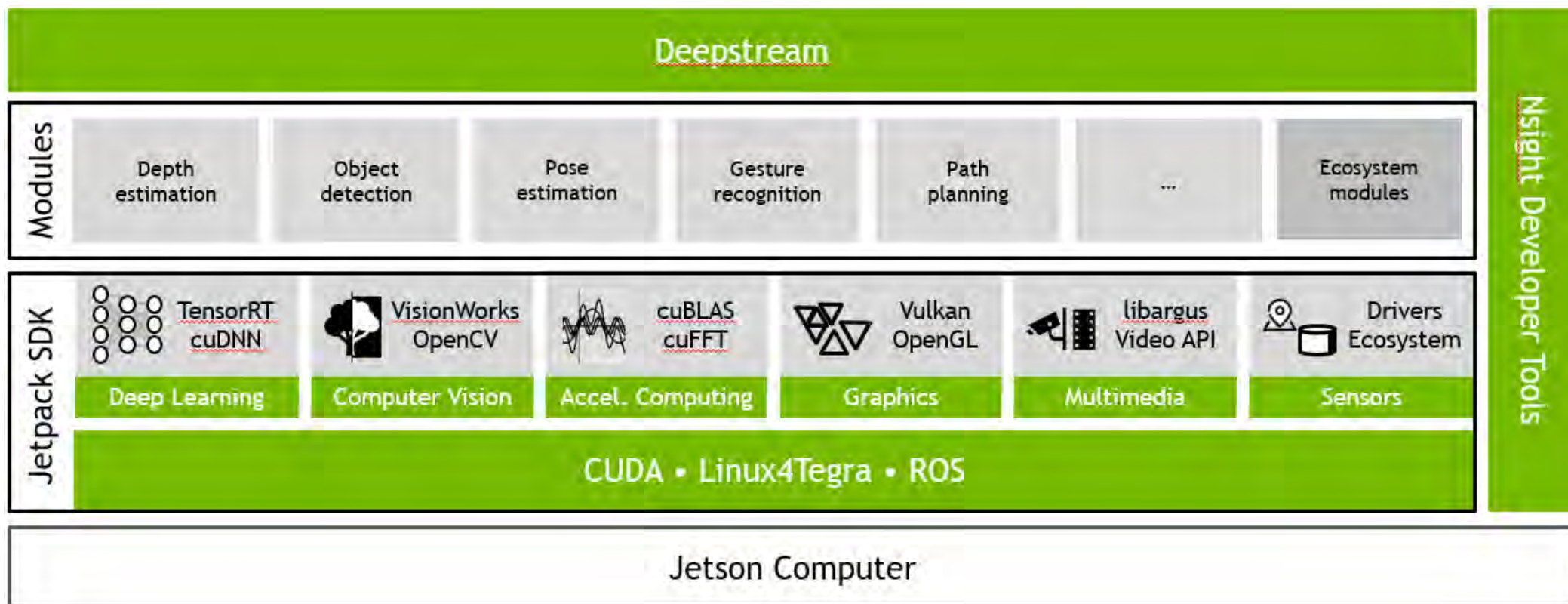
UAVs • AI subsystems • AI Cameras

Fully autonomous machines

Factory automation • Logistics • Delivery robots

Multiple devices • Unified software

LOGITIELS : Offre GPU NVIDIA



INFERENCE : Offre INTEL



INTEL® ARRIA® 10 DEVELOPMENT KIT

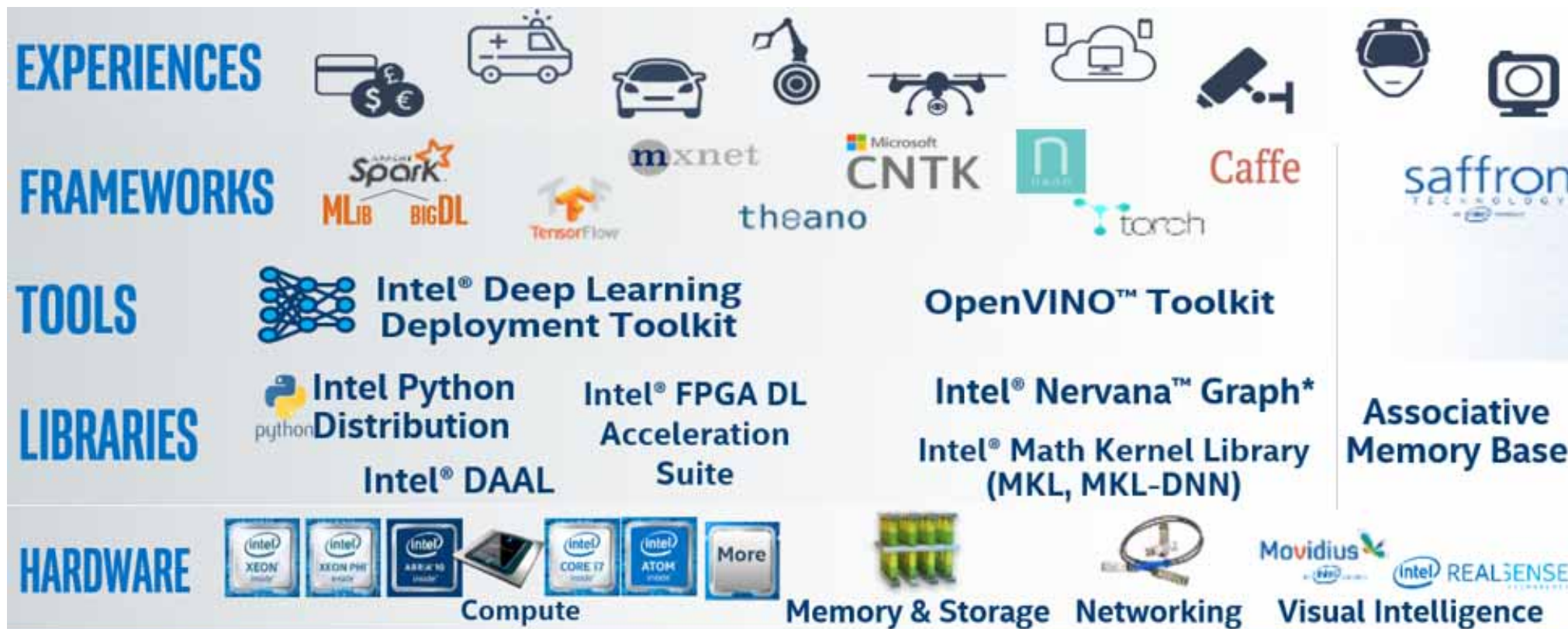
- Versatile FPGA development kit packaged for optimal performance at 20 nm*
- Native abilities for variable precision INT4, INT8, FP11, FP16, FP32
- Evaluation ready for taking your edge applications from prototype to production



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

NXP eIQ –Inference Engines & Libraries

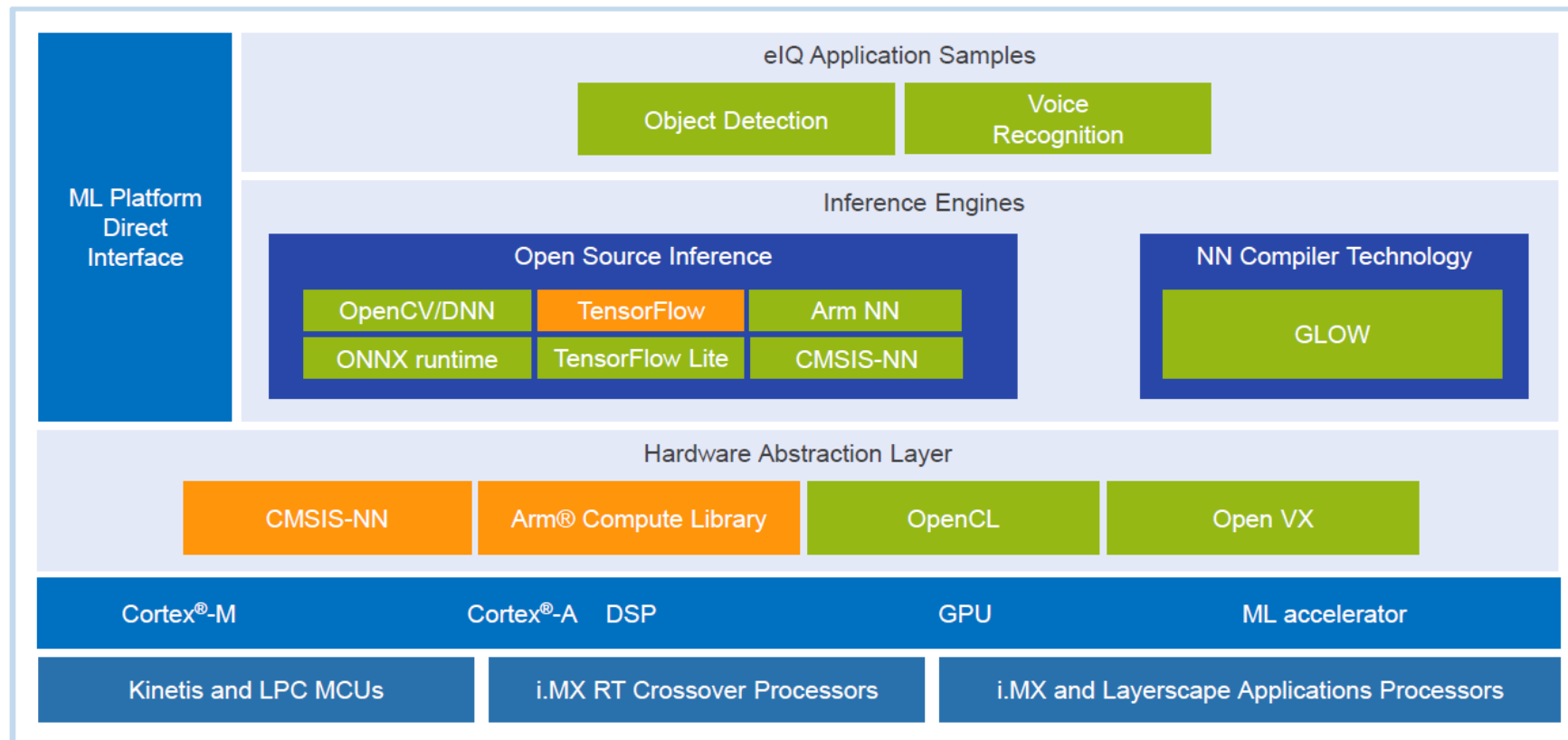


Embedded Compute Engines

	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



Available

In progress

INFERENCE : Offre STMicro

Microprocessors

Dual Cortex-A7 @ 650MHz (2470 DMIPS)
Cortex-M4 @ 200MHz (250 DMIPS)



Microcontrollers



Cortex-M3 @ 120MHz (150 DMIPS)



Cortex-M4 @ 180MHz (225 DMIPS)



Cortex-M7 @ 216MHz (462 DMIPS)



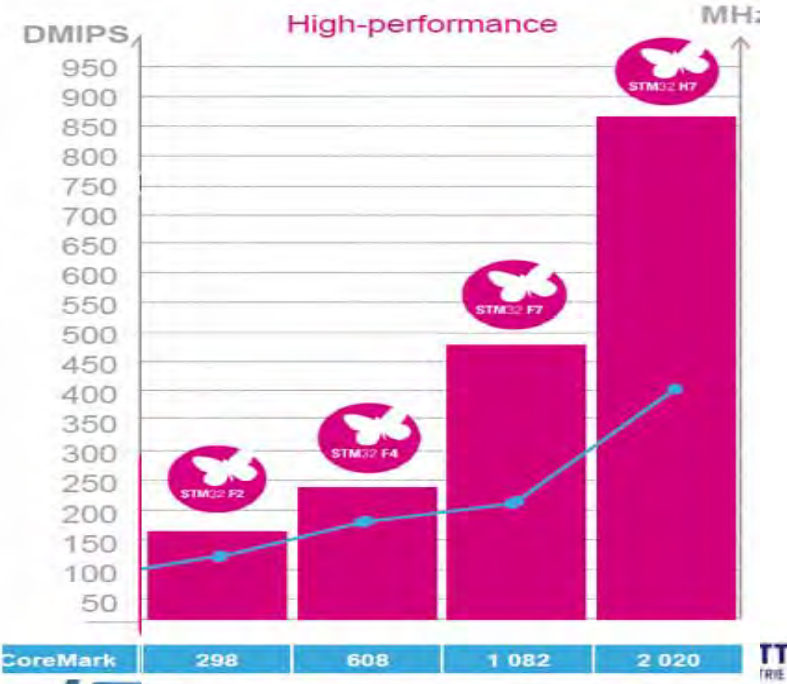
Cortex-M7 @ 400MHz (858 DMIPS)

STM32 High Performance Portfolio

ST Confidential

STM32MP157

System	Dual Cortex-A7 @ 650MHz	3D GPU OpenGL ES2.0 @ 533MHz
CPU & Internal Accelerators	Cortex-A7 @ 650MHz Cortex-M4 @ 200MHz	26Mbit/sec, 133Mpix/sec
Memory - DRAM	128KB SRAM 512KB SRAM	Connectivity
Memory - Flash	128KB SRAM	2x SPI (FIFO) (Slave)
Flash - Range (On-Chip)	128KB SRAM	SPI (FIFO) (Master)
Flash - Range (External)	128KB SRAM	I2C (Slave)
Cache - L1	128KB SRAM	I2C (Master)
Cache - L2	128KB SRAM	UART
Cache - L3	128KB SRAM	CAN
Cache - L4	128KB SRAM	Ethernet
Cache - L5	128KB SRAM	USB
Cache - L6	128KB SRAM	SDIO
Cache - L7	128KB SRAM	MMC
Cache - L8	128KB SRAM	FSMC
Cache - L9	128KB SRAM	QSPI
Cache - L10	128KB SRAM	I2S
Cache - L11	128KB SRAM	DAC
Cache - L12	128KB SRAM	ADC
Cache - L13	128KB SRAM	Comparator
Cache - L14	128KB SRAM	Timer
Cache - L15	128KB SRAM	DMA
Cache - L16	128KB SRAM	RTC
Cache - L17	128KB SRAM	Crypto
Cache - L18	128KB SRAM	Security
Cache - L19	128KB SRAM	Control
Cache - L20	128KB SRAM	Analog



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



Process & analyze new data using trained NN



Convert NN into optimized code for MCU

OFFRE ARROW : CAPTEURS



Pour vos Prototypes et Pré Série Pensez ARROW.COM

NOUVEAUX CLIENTS : ÉCONOMISEZ 20 \$ SUR VOTRE PREMIER ACHAT À PARTIR DE 100 \$ AVEC LE CODE 20NEW

ARROW ECS ARROW SERVICES IOT ARROWCLOUD INDIEGOGO 00-800-8000-1010 FR € EUR

ARROW Toutes les catégo... Rechercher plus de 1 million de produits et des milliers de fournisseurs

MyArrow™ S'IDENTIFIER

RECHERCHE AVANCÉE PAR PARAMÈTRES

229 260 NOUVEAUTÉS Produits Fabricants Fiches techniques Références Articles - Vidéos - Événements Centre de conception Outil BOM ArrowPlus powered by Freelancer

Présentation du transformateur planaire PulseR PL10201NL

Découvrez un transformateur de commutation planaire haute fréquence d'une puissance nominale allant jusqu'à 250 W et d'une plage de fréquence comprise entre 200 kHz et 700 kHz dans un format 29,5 mm x 26,7 mm.

[JE COMMANDE](#)

Livraison Express Gratuite

Arrow.com offre maintenant la livraison gratuite dès 50 \$ de commande. Sans code requis.

[EN SAVOIR PLUS](#)

Inscrivez-vous

Soyez parmi les premiers à être informés des nouveautés et offres spéciales.

[S'INSCRIRE](#)

NOUVEAUX PRODUITS [TOUT AFFICHER](#) FICHES TECHNIQUES [TOUT AFFICHER](#) Spark Virtual Assistant

A hand is shown holding a glowing globe. The globe is covered in a network of white lines and nodes, with several nodes highlighted in yellow. The background is dark blue with faint binary code and a satellite-like structure. The text 'ANNEXES' is written in white at the top right, and 'EXEMPLES de REALISATIONS' is written in white in the center.

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



* Available soon

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



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

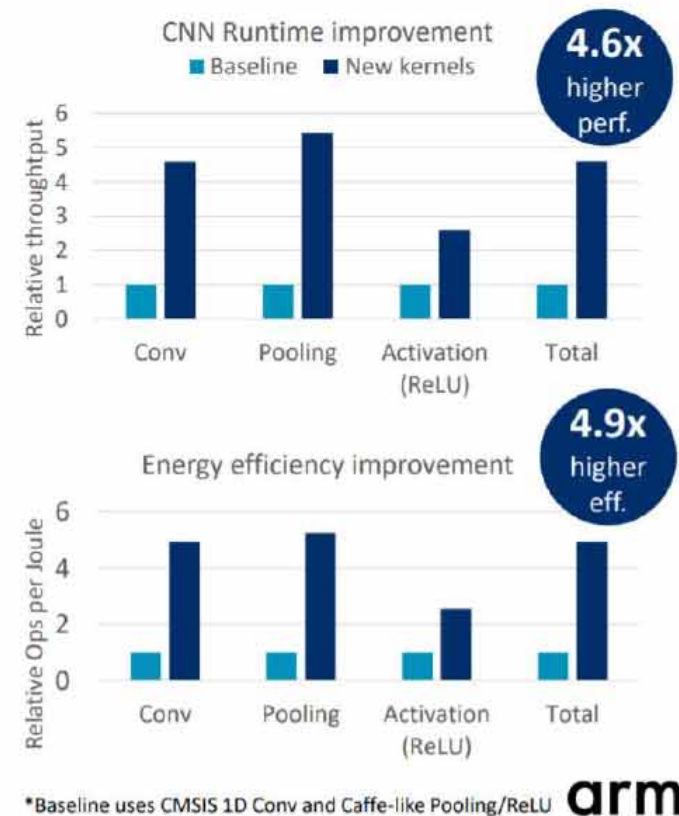
- Boost compute density with GEMM based implementation
- Reduce data movement overhead with depth-first data layout
- Interleave data movement and compute to minimize memory footprint

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.

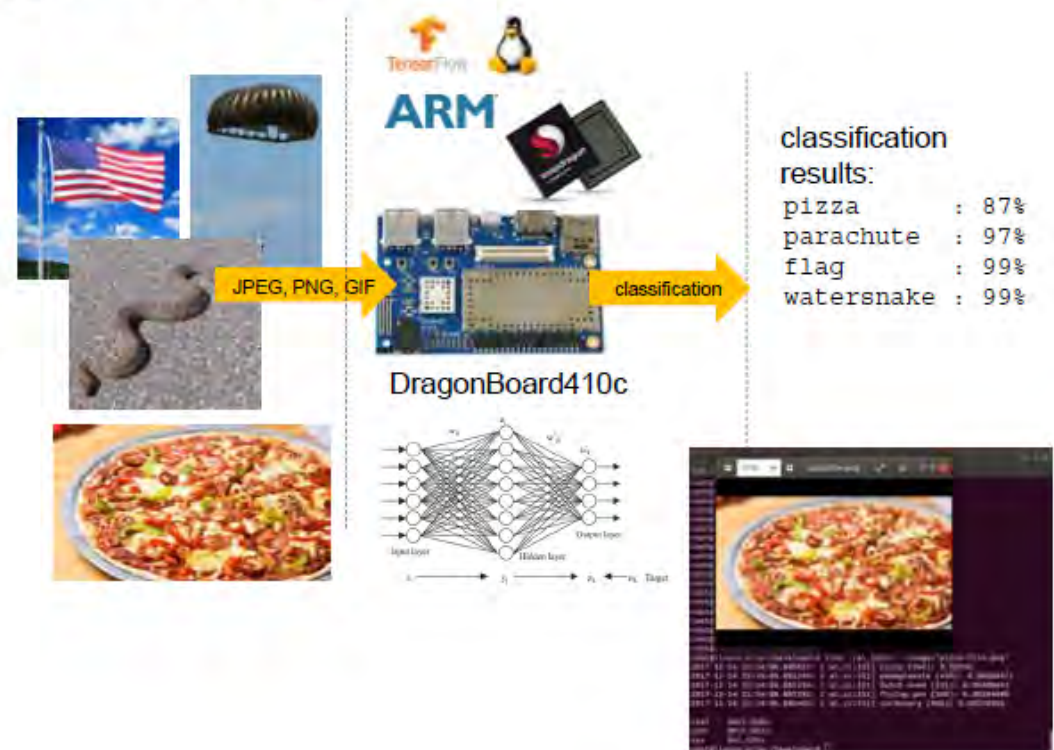
Device optimization: Significantly extends battery life by optimizing operating system scheduling for individual applications.

Tensorflow SDK on DragonBoard 410c

Problem statement:

How to simplify edge AI deployment?

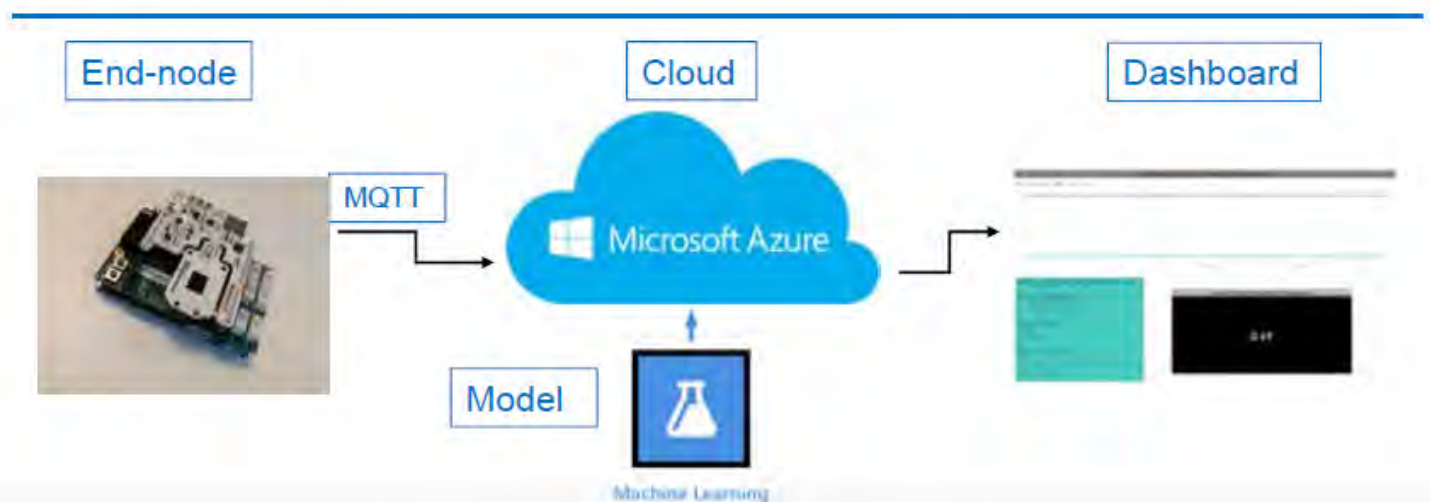
- Enable most widely AI framework on ARM-based systems
- Seamless transition – Cloud trained model directly to embedded device
- Benchmarking for evaluation expertise (board capabilities vs. use case)



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

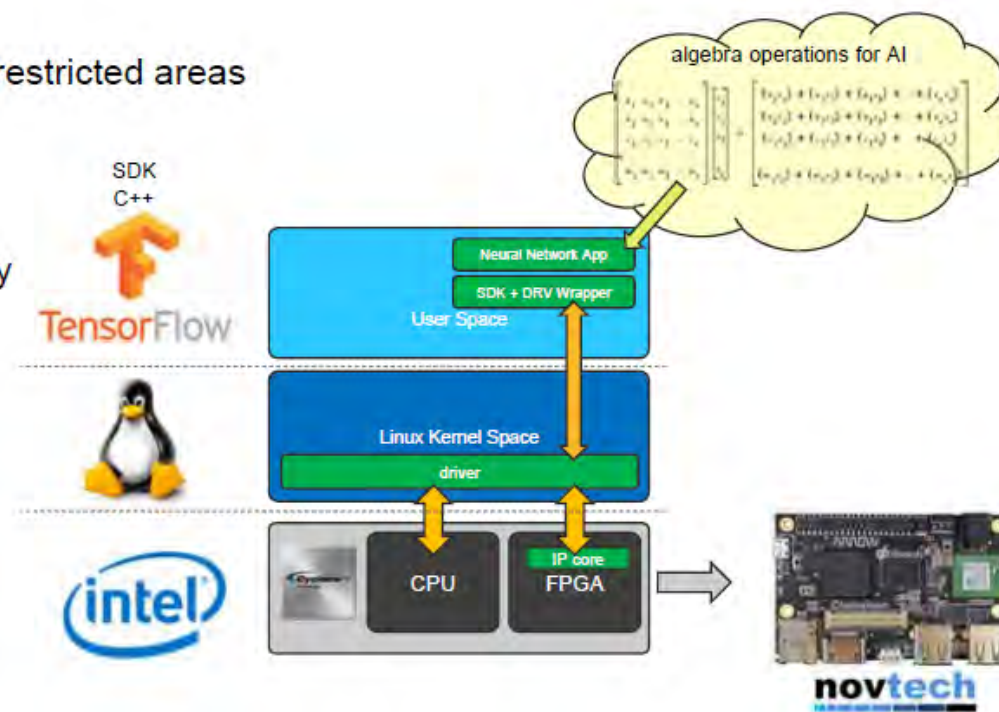


Face recognition

Problem statement:

Camera for people monitoring in disconnected or restricted areas

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

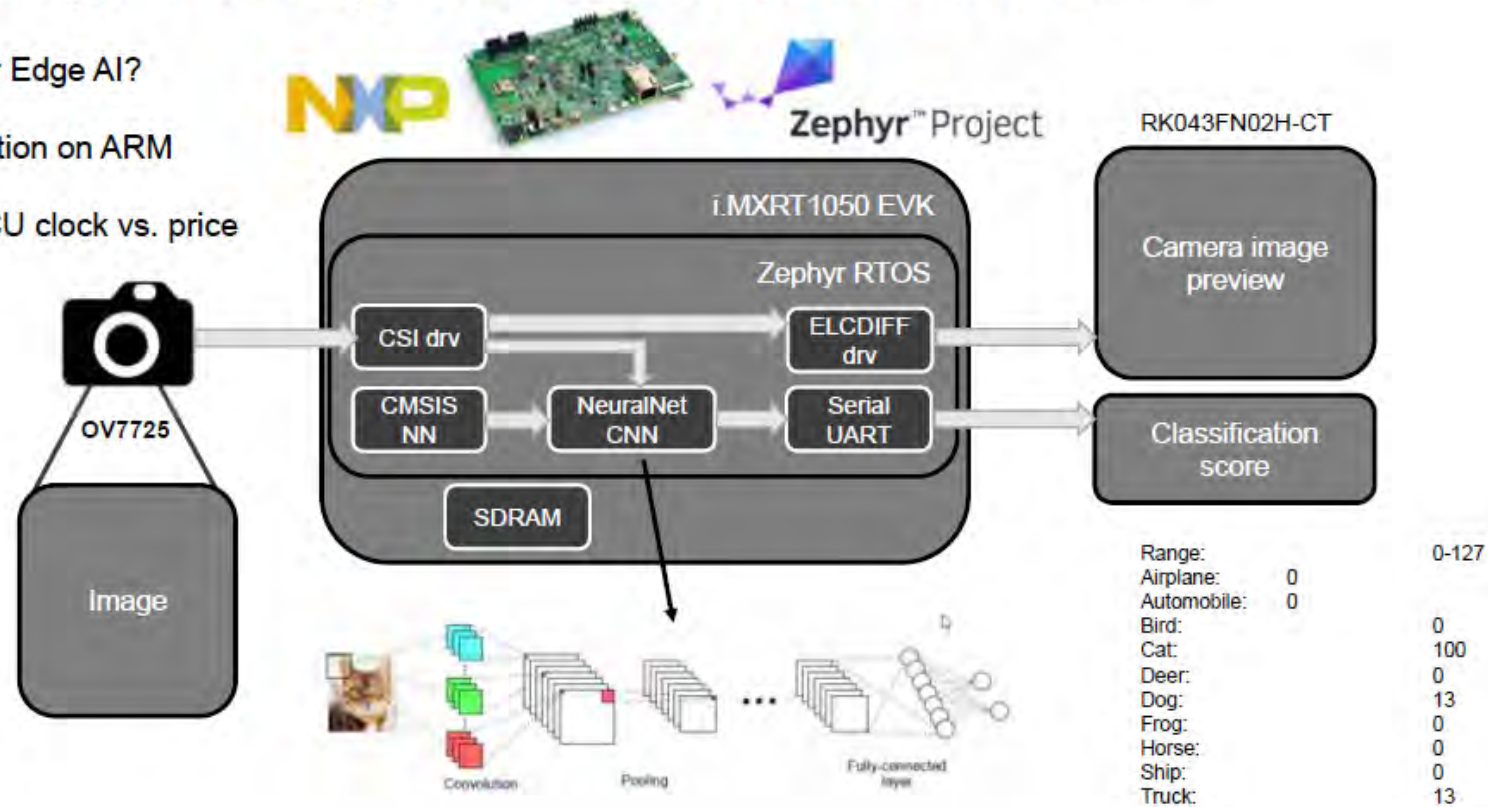


Cifar10 Image Recognition demo on RTOS

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

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



 **ANALOG
DEVICES**
AHEAD OF WHAT'S POSSIBLE™



MC27561-DRAGONFLY
SmartMesh IP

