Créateur de réussites numériques

Global AI Bootcamp 2019

AI on Edge, la rencontre en l'intelligence artificielle et l'Internet des Objets

AZUGFR
Azure User Group France

Global AI Bootcamp

Paris, samedi 14 décembre 2019

Ecole Supérieure de Génie Informatique

GUSS

Jonathan Pacifico

AI Developer

@pacific0_o

École IA Microsoft

AZEO

talents & technology

Jean-Pierre Riehl

Innovation Director

@djeepy1

http://blog.djeepy1.net

Microsoft® Most Valuable Professional

MVP Data Platform

GUSS Club Power BI

French Data Community **Leader**


AI
AI EVERYWHERE
imgflip.com

HappyScore : 0,7964

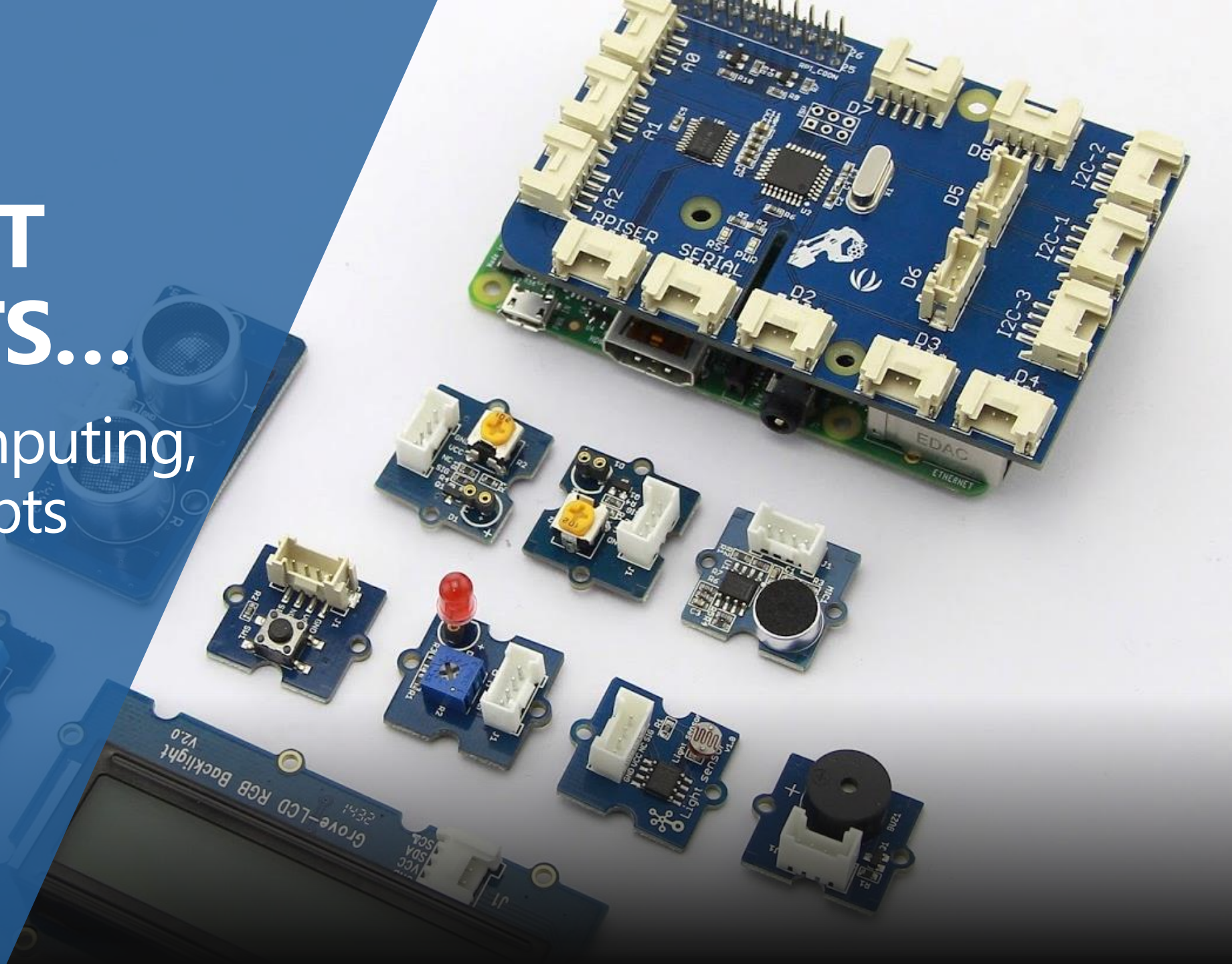
Hackathon 2019

Projet #2 : ***HappyFace vNext***

AZEO
talents & technology

L'INTERNET DES OBJETS…

IoT & Edge Computing, quelques concepts




"I BELIEVE OVER THE NEXT DECADE COMPUTING WILL BECOME EVEN MORE UBIQUITOUS AND INTELLIGENCE WILL BECOME AMBIENT...THIS WILL BE MADE POSSIBLE BY AN EVER-GROWING NETWORK OF CONNECTED DEVICES, INCREDIBLE COMPUTING CAPACITY FROM THE CLOUD, INSIGHTS FROM BIG DATA, AND INTELLIGENCE FROM MACHINE LEARNING"
SATYA NADELLA
[CEO, MICROSOFT]
AZEO
talents & technology

L'ANALOGIE DE LA BOUILLOIRE

AZEO
talents & technology

Edge Computing

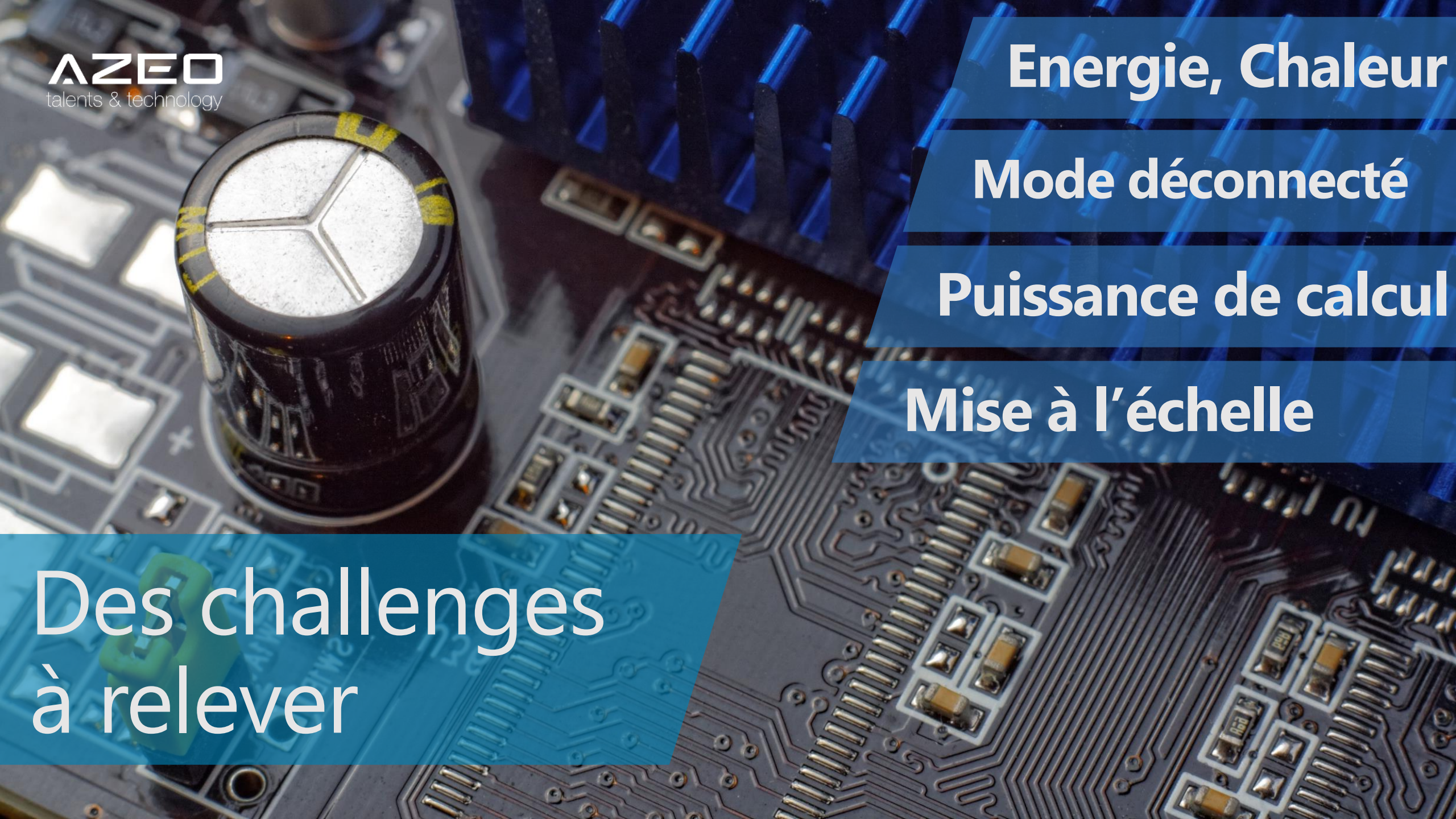
…traiter les données en périphérie du réseau directement où elles sont générées…

Source : https://www.lebigdata.fr/edge-computing-definition

FOG COMPUTING

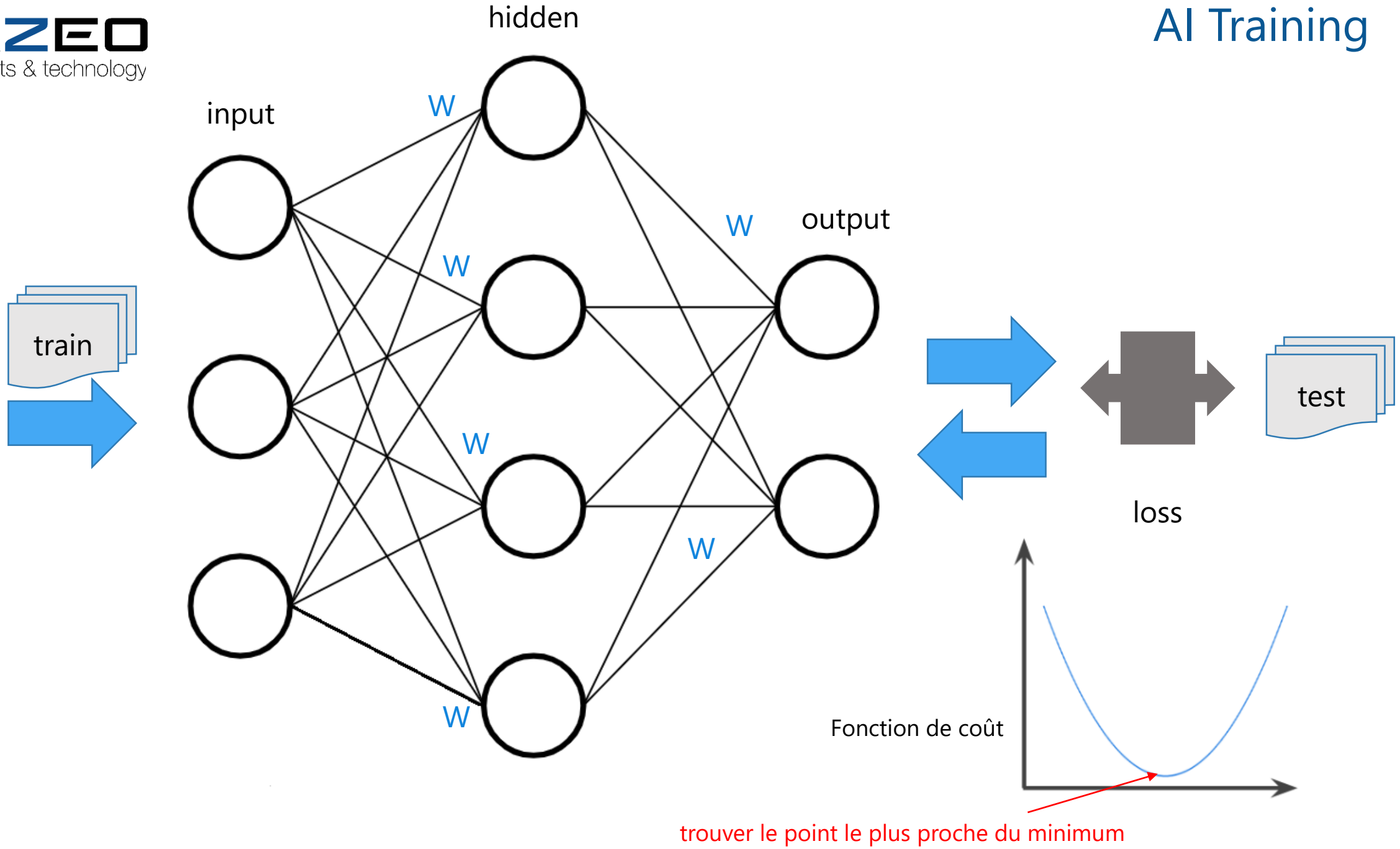
AZEO
talents & technology

Et si on mettait de l'IA sur le Edge ?

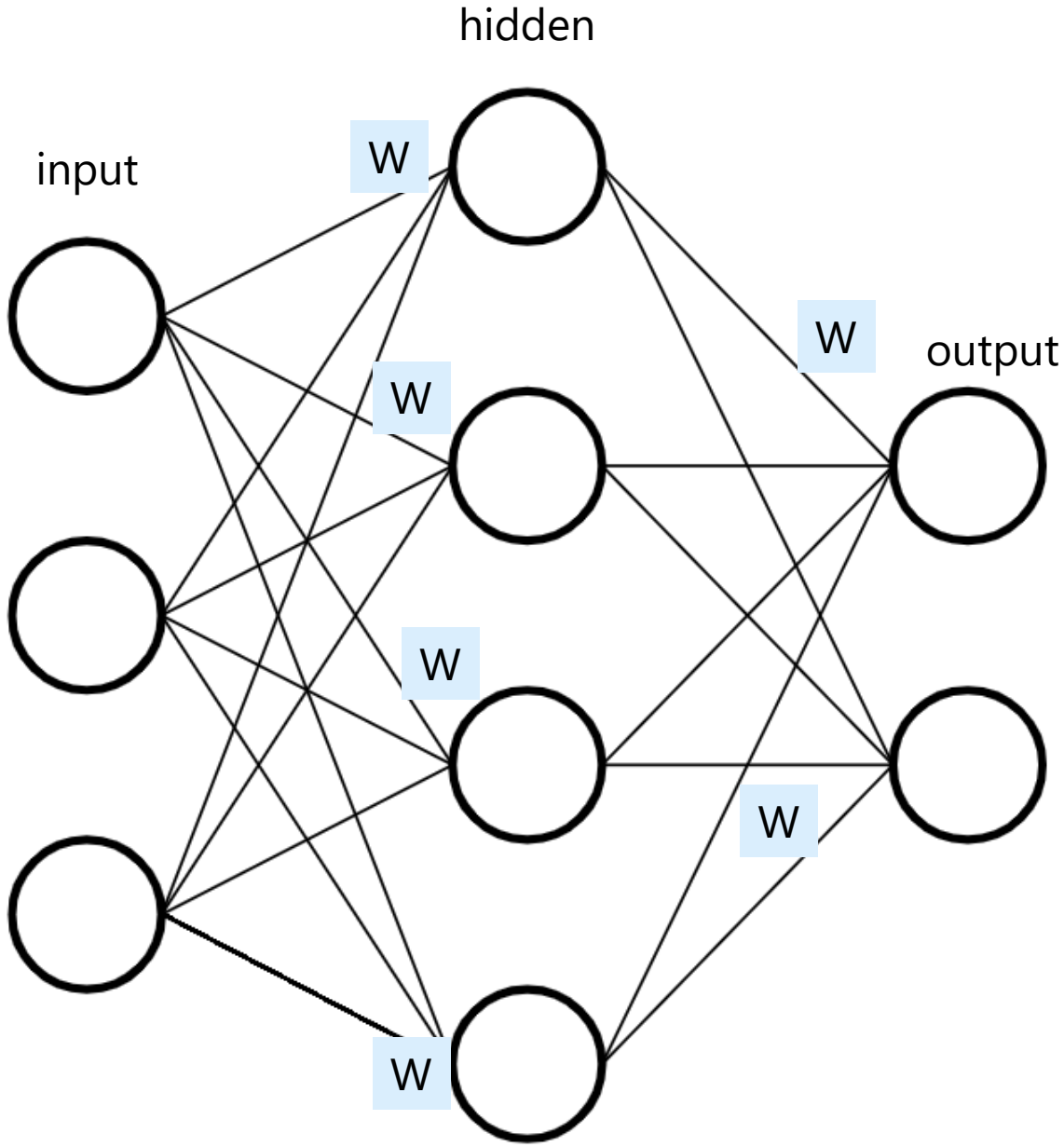

AZEO
talents & technology
Energie, Chaleur
Mode déconnecté
Puissance de calcul
Mise à l'échelle
Des challenges
à relever

Entrainer un modèle pour le Edge

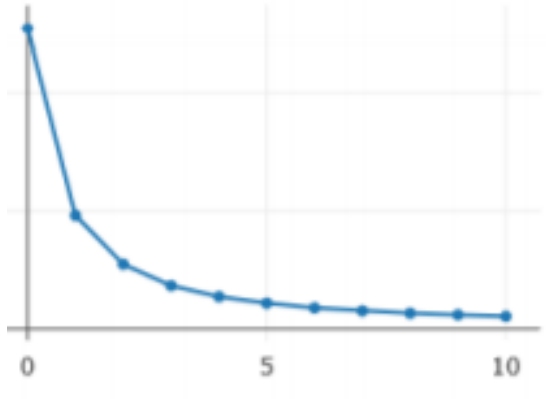
AI Training


hidden
input
W
W
W
W
W
W
output
train
test
loss
Fonction de coût
trouver le point le plus proche du minimum

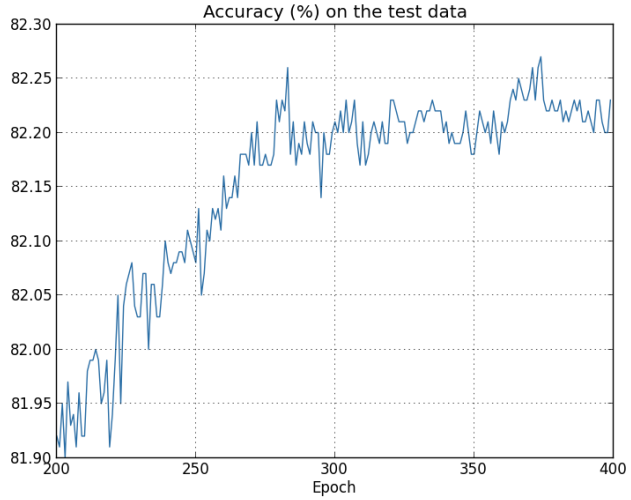
AZEO
talents & technology



loss

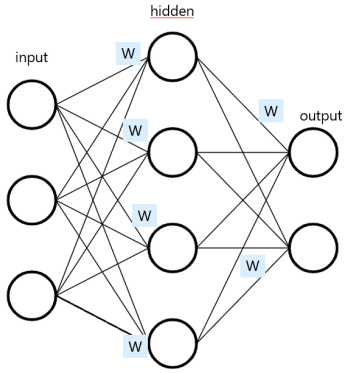


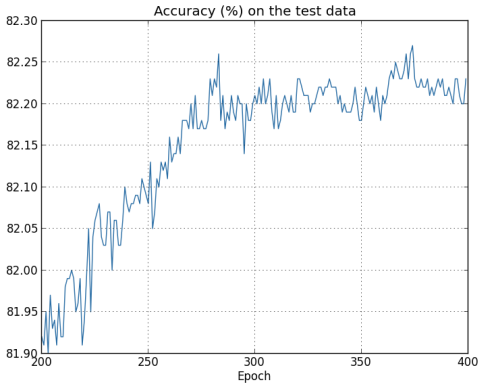
accuracy

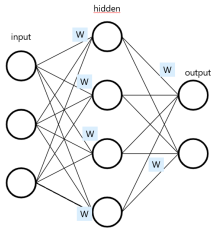


AZEO
talents & technology

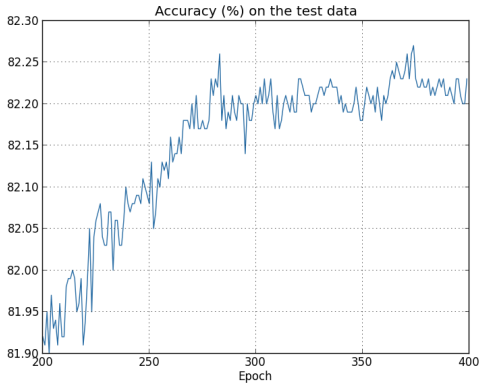
Challenge

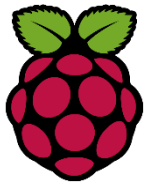






??

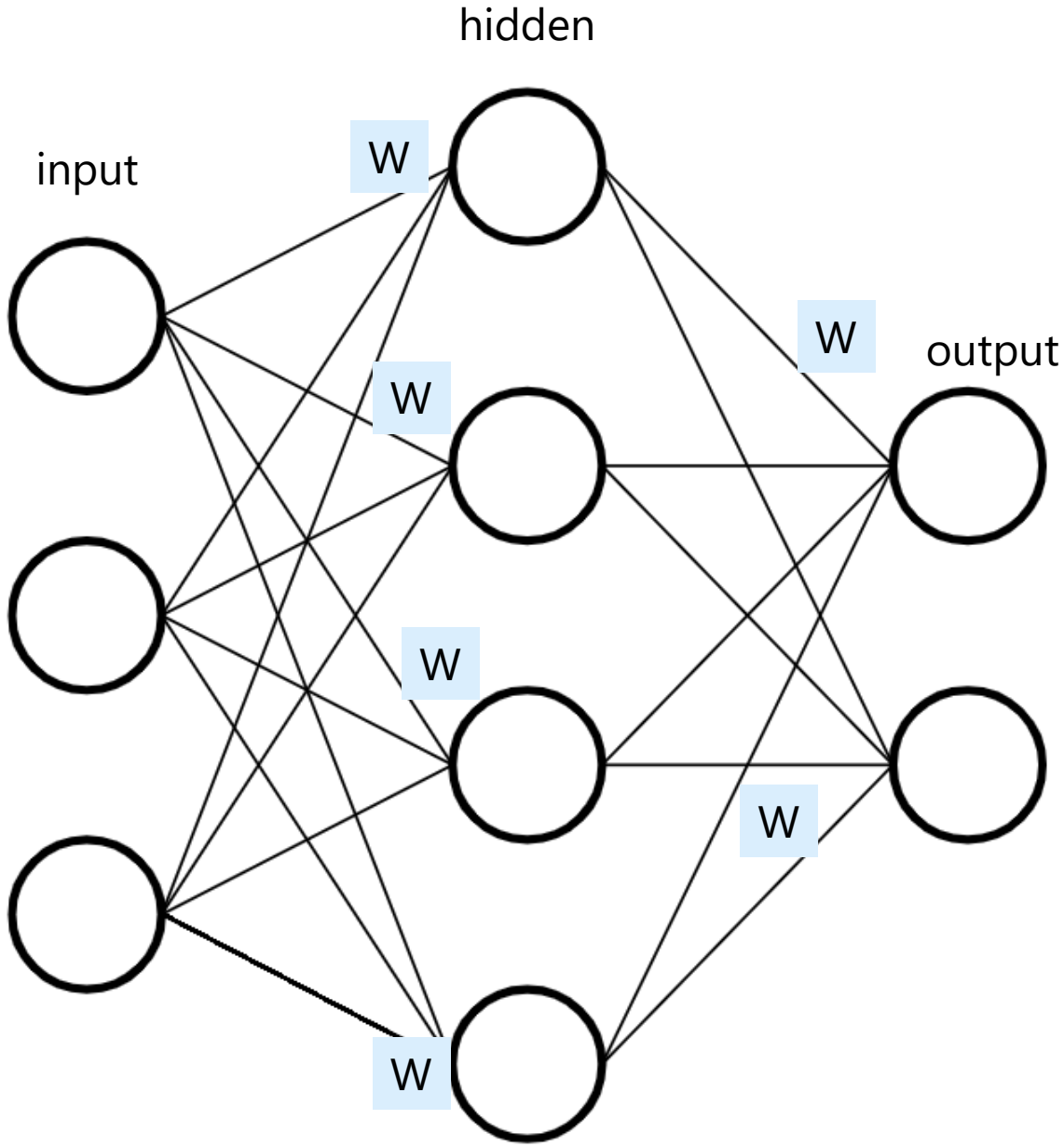




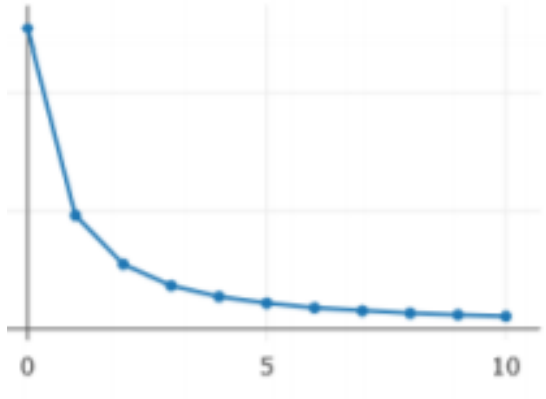
AZEO
talents & technology

3 pistes d’optimisation

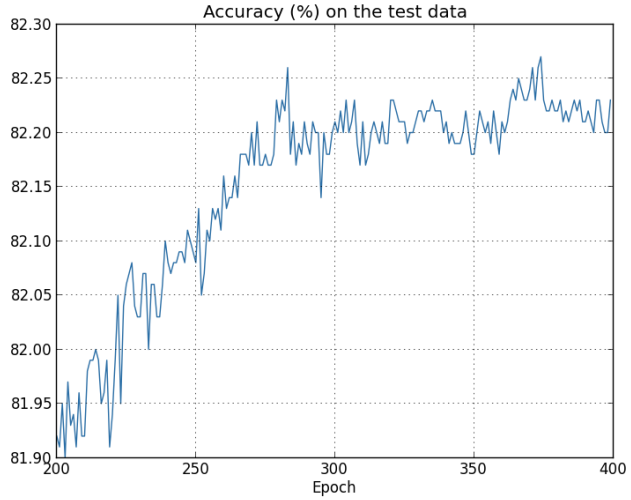
AZEO
talents & technology



loss

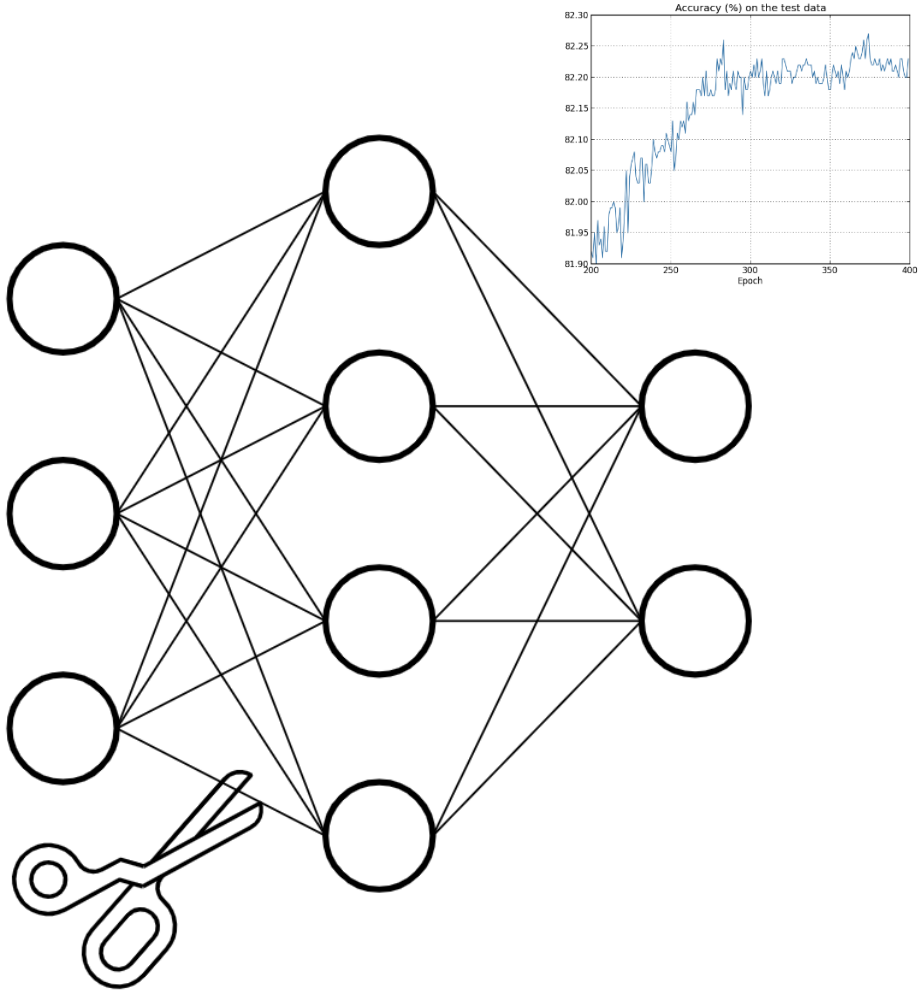


accuracy

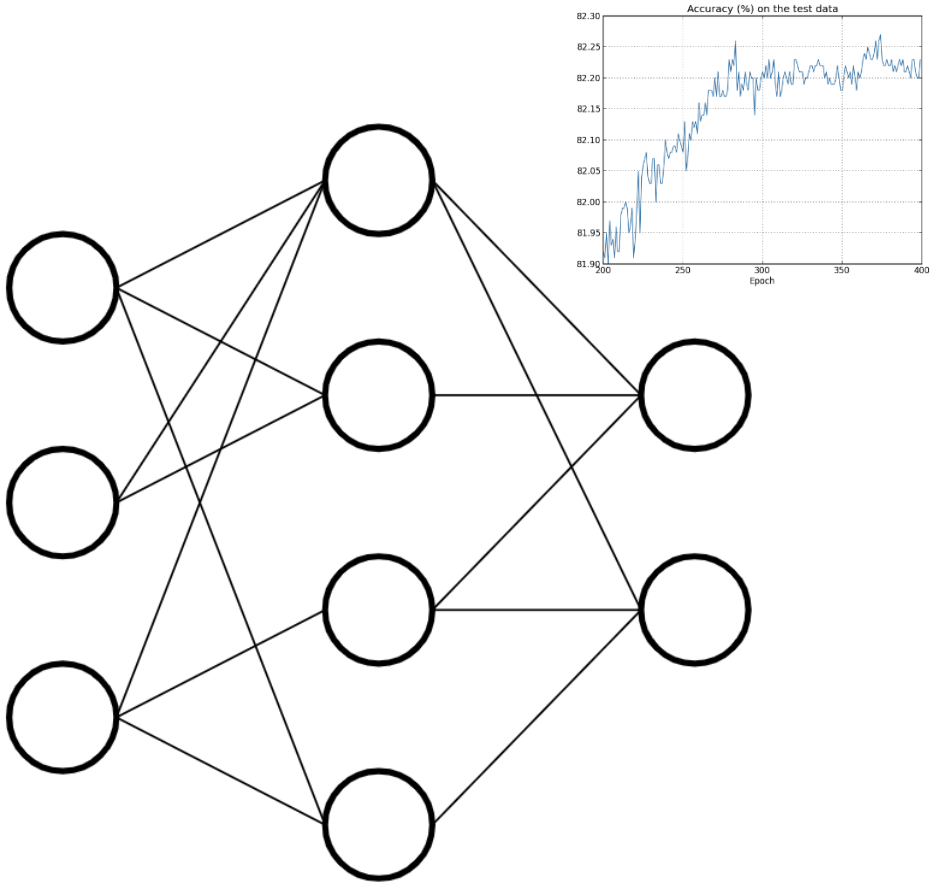


AZEO
talents & technology

Pruning



Before pruning

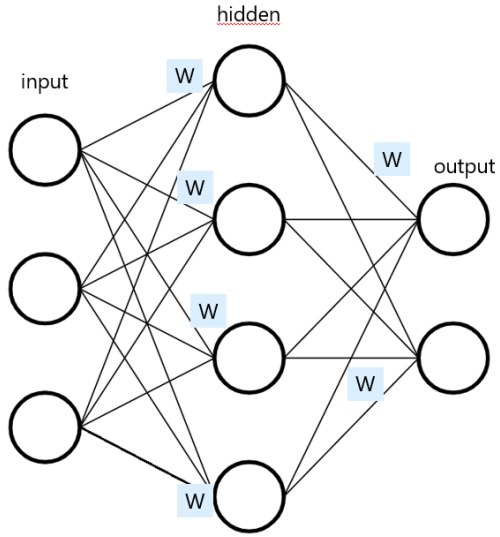


After pruning

Quantification

8 bits **W** 256 valeurs possibles

5 bits **W** 30 valeurs possibles

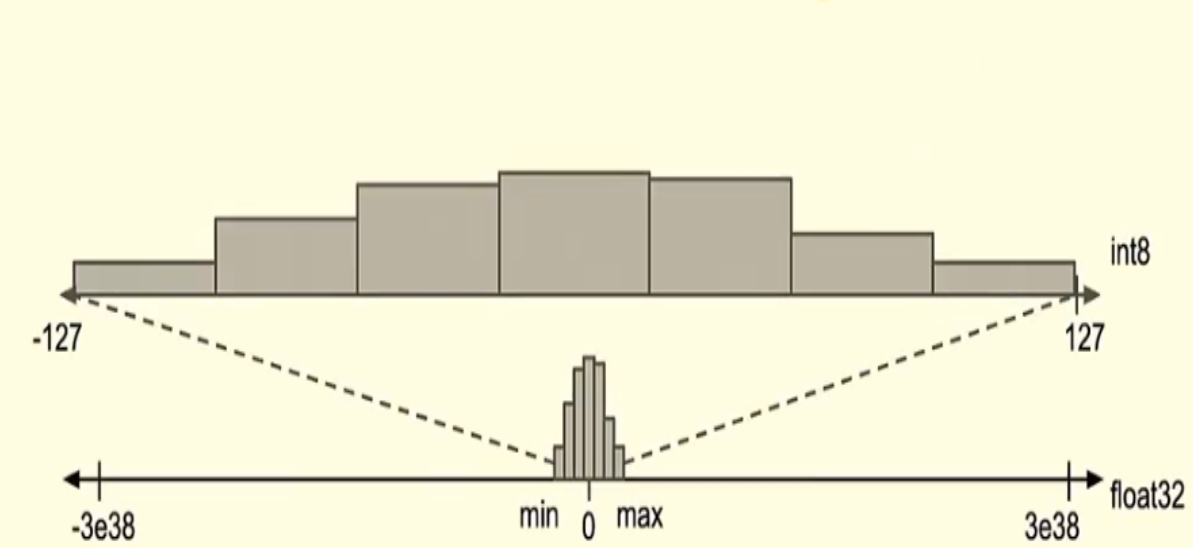


1000 poids de 8 bits = 8000 bits

1000 poids de 5 bits = 5000 bits

35% gain mémoire

Quantization



Source : Tensorflow

Quantification de valeurs flottantes en valeurs entières

```python
import tensorflow as tf

saved_model_dir = "/path/to/mobilenet_v1_1.0_224/"
converter = tf.lite.TFLiteConverter.
    from_saved_model(saved_model_dir)
converter.optimizations = [tf.lite.Optimize.DEFAULT]
def data_generator():
  for i in range(calibration_steps):
    # get sample input data
    yield [input_sample]
converter.representative_dataset = data_generator
tflite_model = converter.convert()
open("converted_model.tflite", "wb").write(tflite_model)
```

Post training API (Tensorflow)

Quantization

Résultats obtenus après quantification sur deux Réseaux CNN

Model		Time	Storage	Memory	Top-5 Err.
AlexNet	CNN	2.93s	232.56MB	264.74MB	19.74%
	Q-CNN	**0.95s**	**12.60MB**	**74.65MB**	**20.70%**
CNN-S	CNN	10.58s	392.57MB	468.90MB	15.82%
	Q-CNN	**2.61s**	**20.13MB**	**129.49MB**	**16.68%**

Source : Labo TIMA Université de Grenoble Alpes

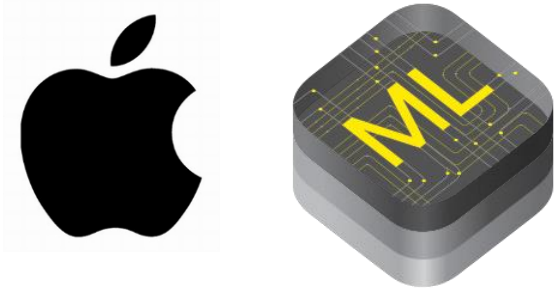
NB : le pruning est compatible avec la quantification

AZEO
talents & technology

Optimization Software-Hardware

AZEO
talents & technology

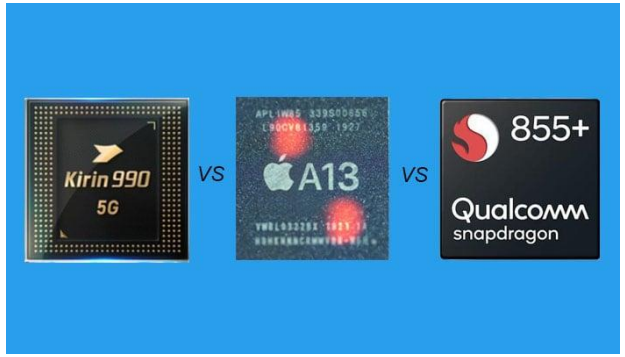
Optimization



CoreML

TFLite

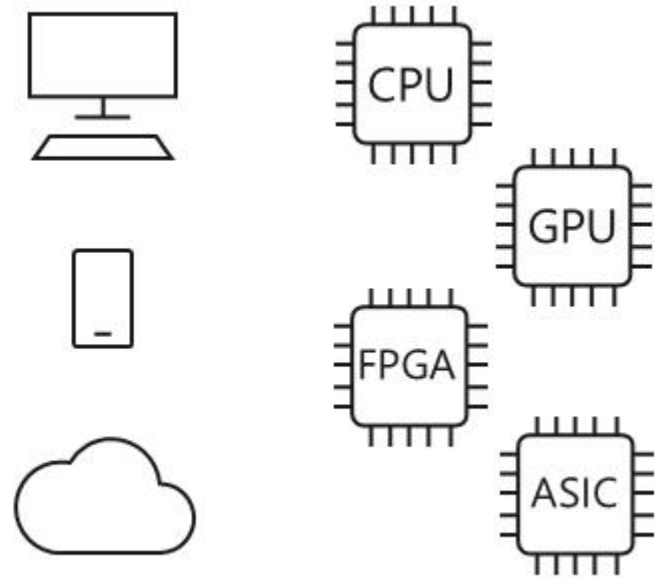
Conversion des modèles sous un format optimisé pour des devices mobiles eux-mêmes optimisés pour le ML



Puces neuronales 5 TOPS

Optimization

Inter-opérabilité entre frameworks d'IA



Optimisation matérielle

QUALCOMM intel NVIDIA

Le challenge du Hardware

Les types de chipsets

Hardware

Calculs 1 par 1 : modèles statistiques, régressions, ANN

Multiples calculs en parallèle : deep learning

Dédié au traitement IA plus économe en énergie , traitement environ 15 fois + rapide que GPU (notam. utilisés pour AlphaGo)

Haute performance, modèles lourds

CPU
- Small models
- Small datasets
- Useful for design space exploration

GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

TPU
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

AZEO
talents & technology

Hardware

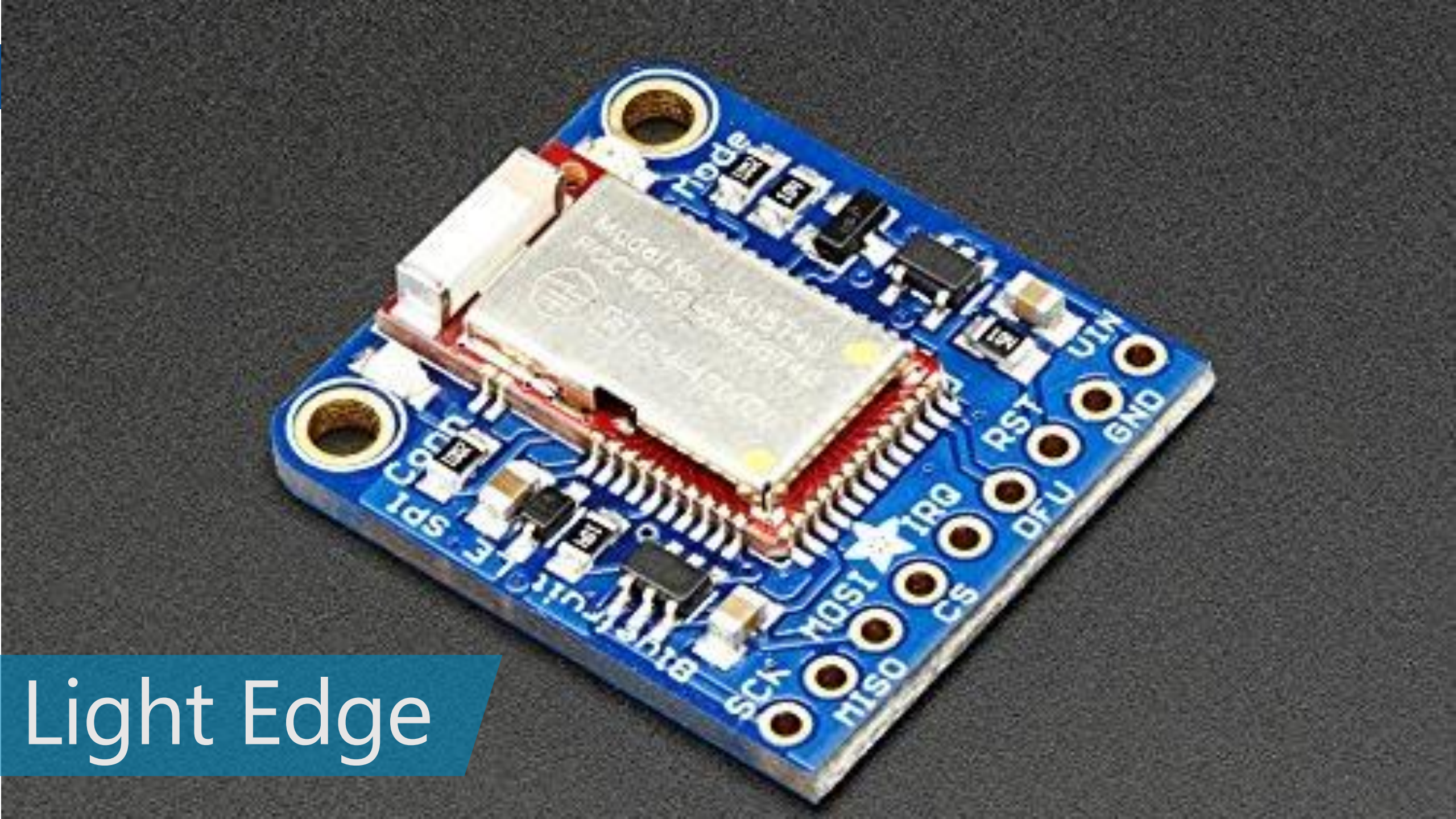
CPU

GPU

TPU

Le challenge du Hardware

Light Edge ou Heavy Edge



Light Edge

Heavy Edge

AZEO
talents & technology

Optimization

Vision AI Dev Kit

Intel Movidius

Nvidia Jetson
Nano Dev Kit

Nvidia Tesla

Databox Edge
(Intel FPGA)

Light Edge

Heavy Edge

QUELQUES DIFFÉRENCES

CONSOMMATEURS

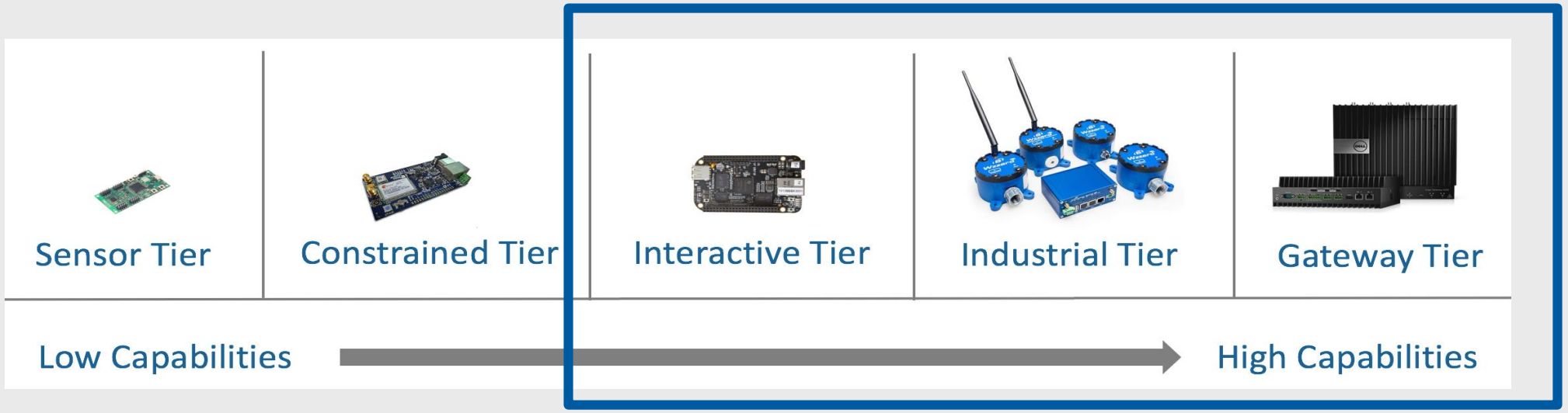
- Battery
- Wireless
- Embedded OS

INDUSTRIELS

- Plug
- Wired
- Full-Fledged OS

AZEO
talents & technology

Hardware for Azure IoT Edge



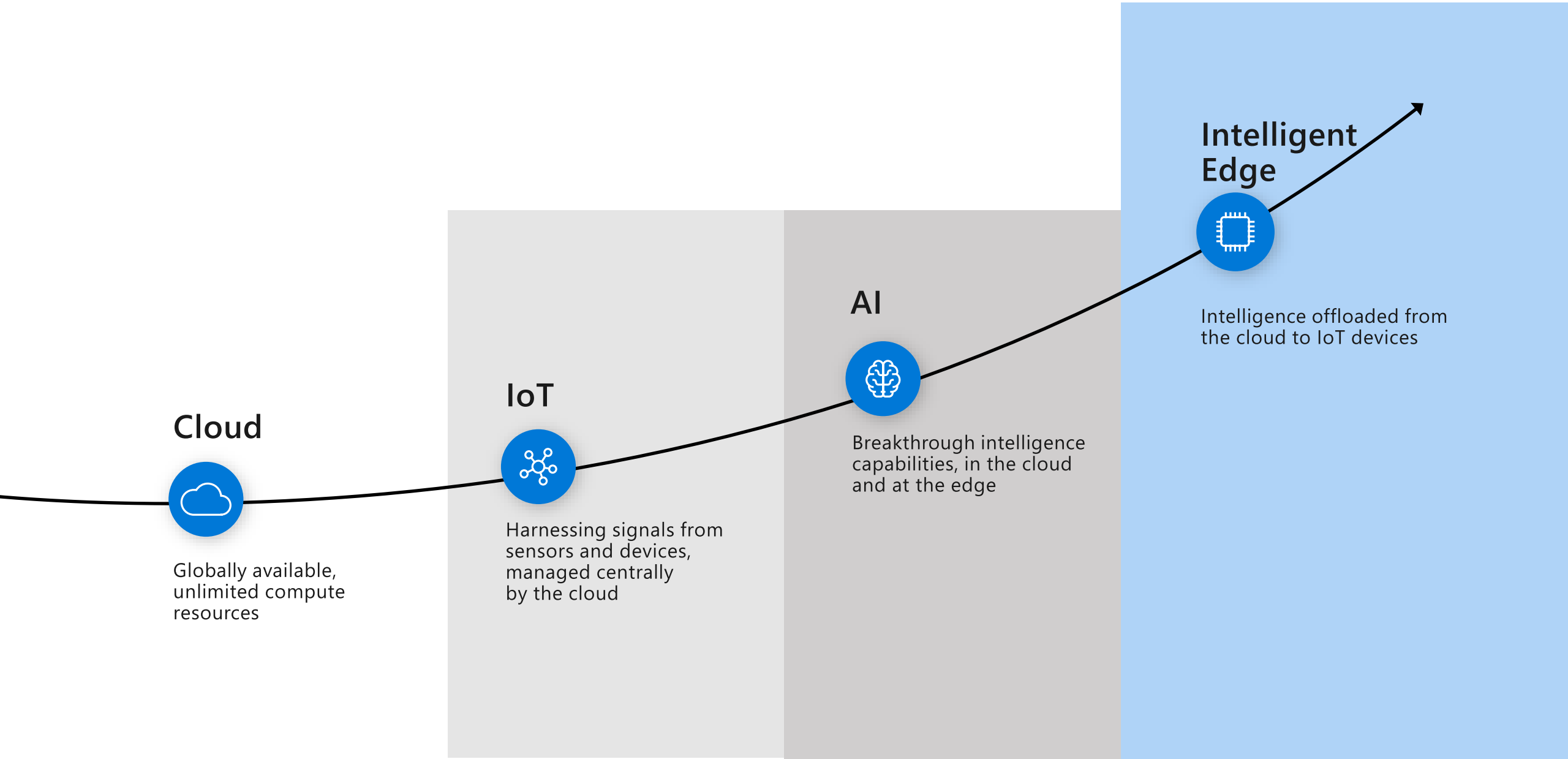
Ability to run on devices smaller than a Raspberry Pi

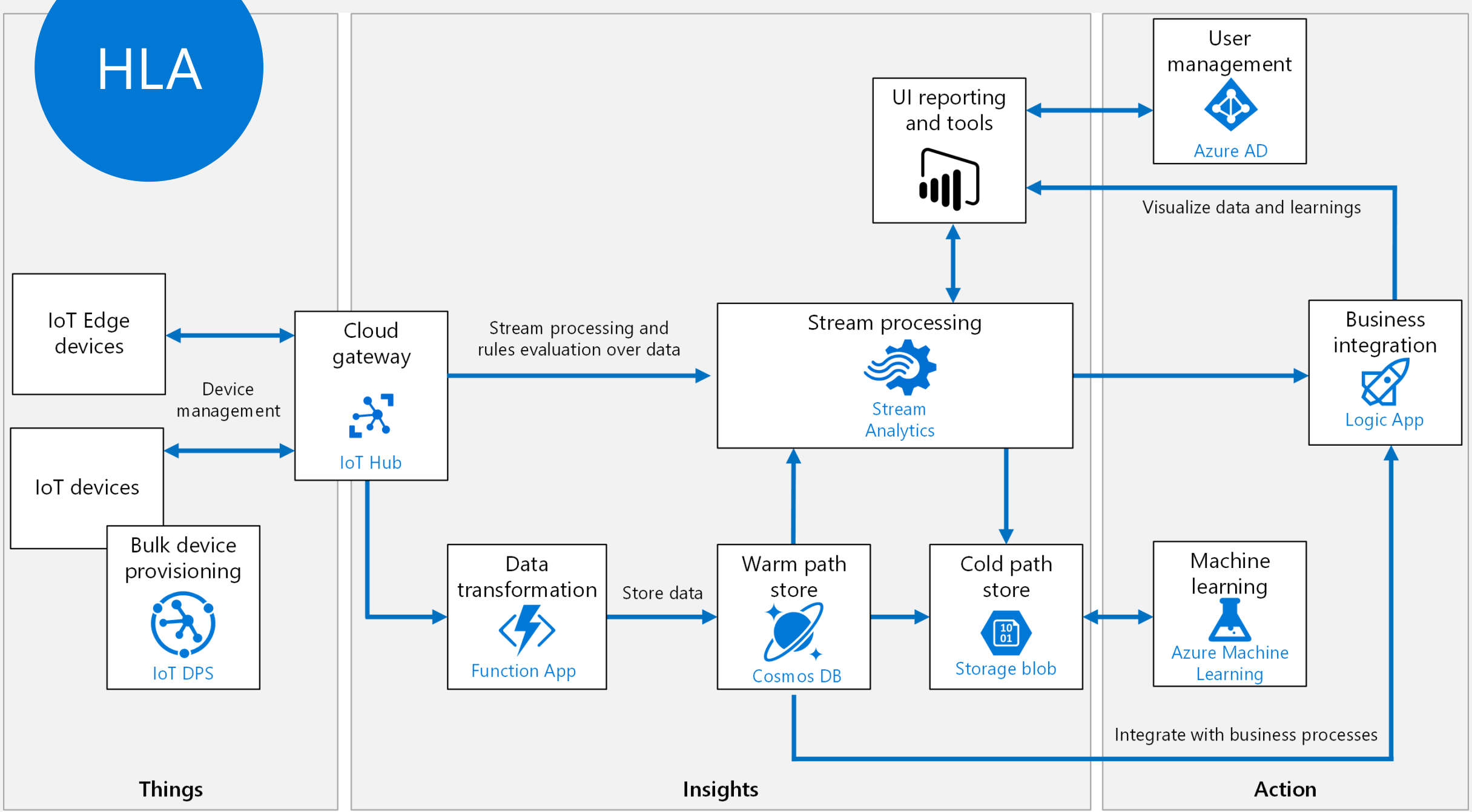
128MB memory

Support best in class operating systems such as Windows, and Linux

Et l’architecture

AZEO
talents & technology

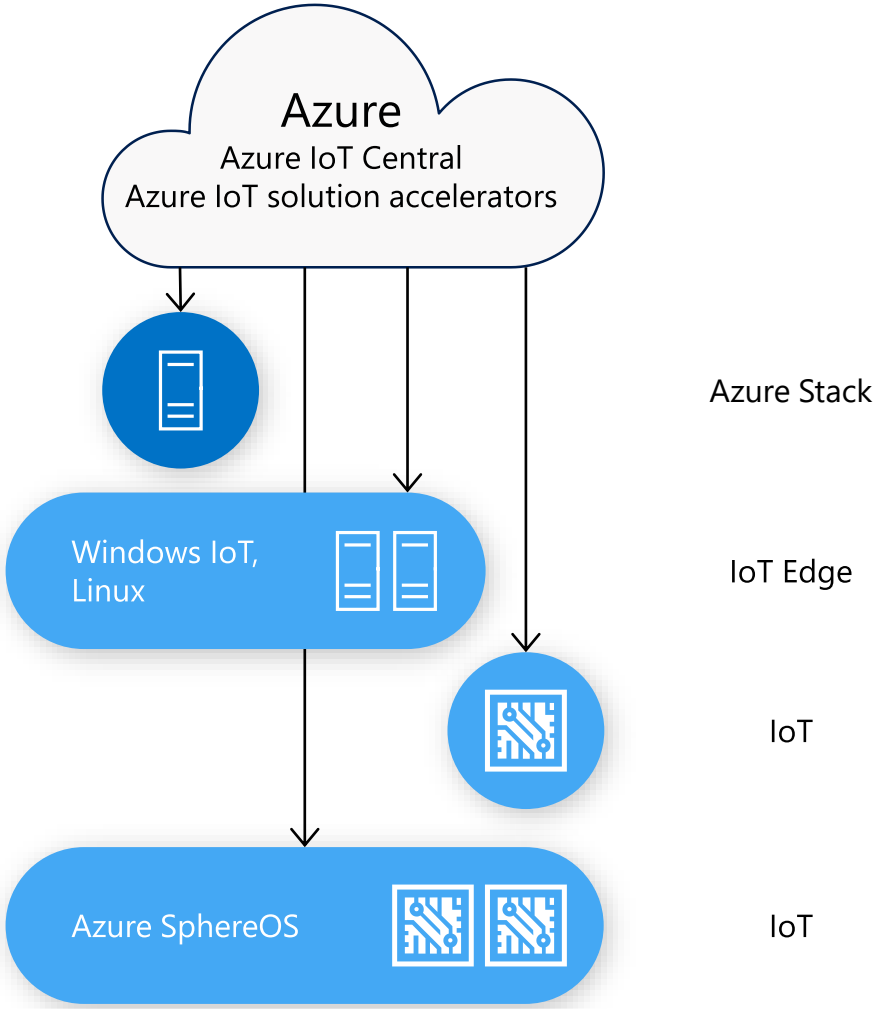


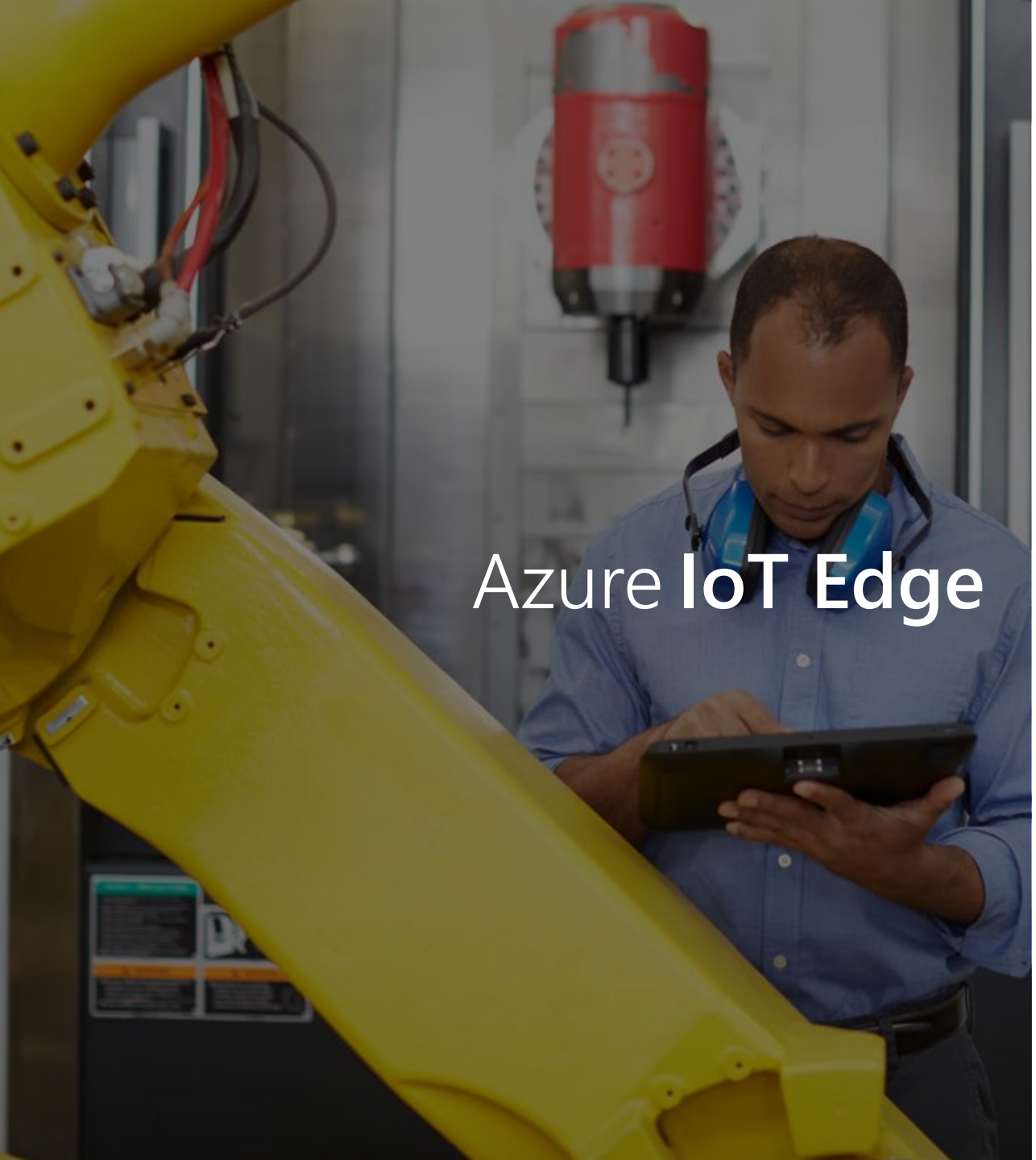

HLA
User management
Azure AD
UI reporting and tools
Visualize data and learnings
IoT Edge devices
Cloud gateway
Stream processing and rules evaluation over data
Stream processing
Stream Analytics
Business integration
Logic App
Device management
IoT Hub
IoT devices
Bulk device provisioning
IoT DPS
Data transformation
Function App
Store data
Warm path store
Cosmos DB
Cold path store
Storage blob
Machine learning
Azure Machine Learning
Integrate with business processes
Things
Insights
Action

AZEO talents & technology

Azure, Azure Stack, IoT Edge, and IoT

Azure	• Available in Azure Regions • Full functionality
Azure Stack	• Azure Services & Management on-prem • Managed by Azure or Locally
Azure IoT Edge	• Deploy and manage cloud services • Managed by Azure or Azure Stack
Windows IoT, Linux	• Azure IoT Edge runs on Windows and Linux
Azure IoT Device SDK	• Multi-device, multi-language, multi-OS • iOS, Android, Windows, Linux
Azure Sphere	• Peerless security for MCU devices • Connect directly to Azure or via Azure IoT Edge
Azure Sphere OS	• Linux Kernel that modernizes MCU devices

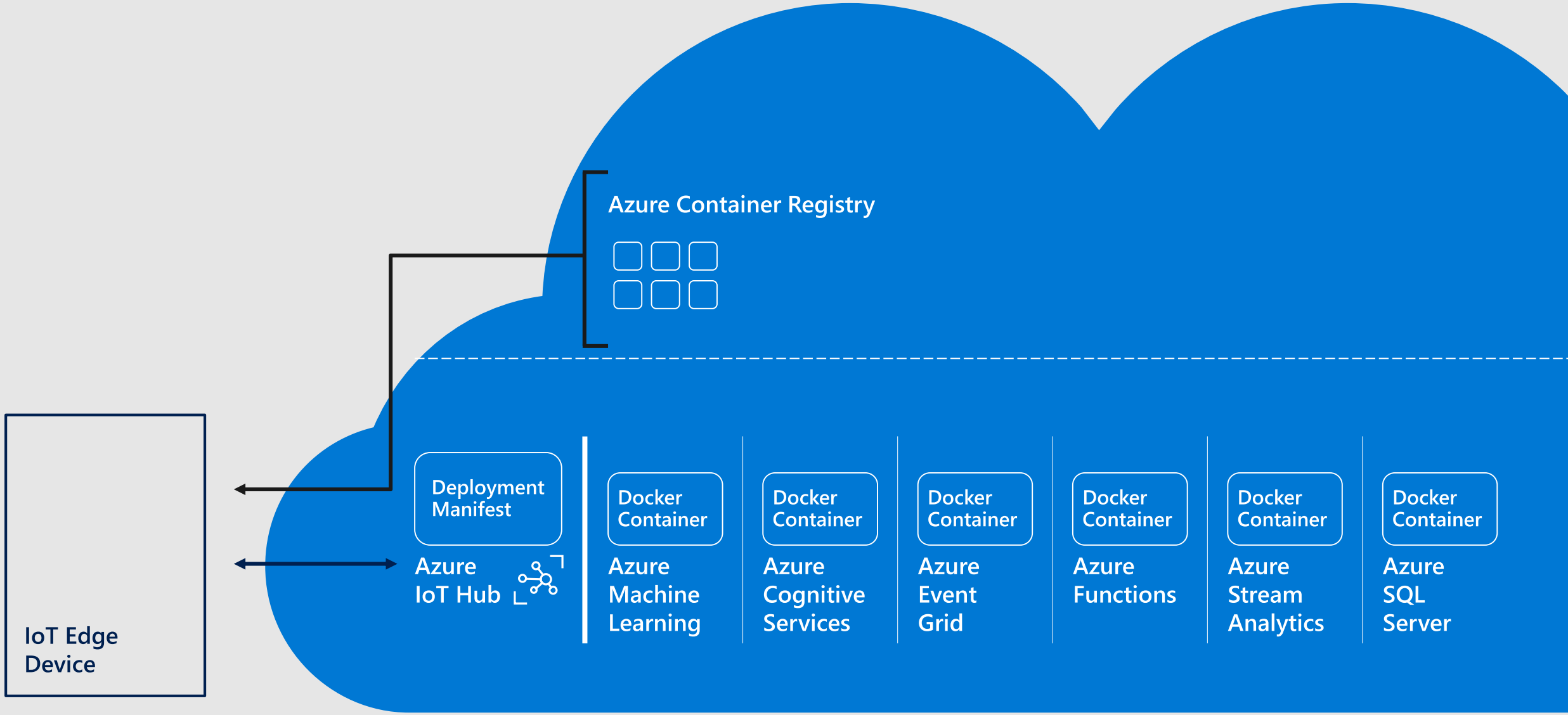




Azure **IoT Edge**

- Move cloud and custom workloads to the edge, securely
- Seamless deployment of AI and advanced analytics
- Configure, update and monitor from the cloud
- Compatible with popular operating systems
- Code symmetry between cloud and edge for easy development and testing
- Secure solution from chipset to cloud

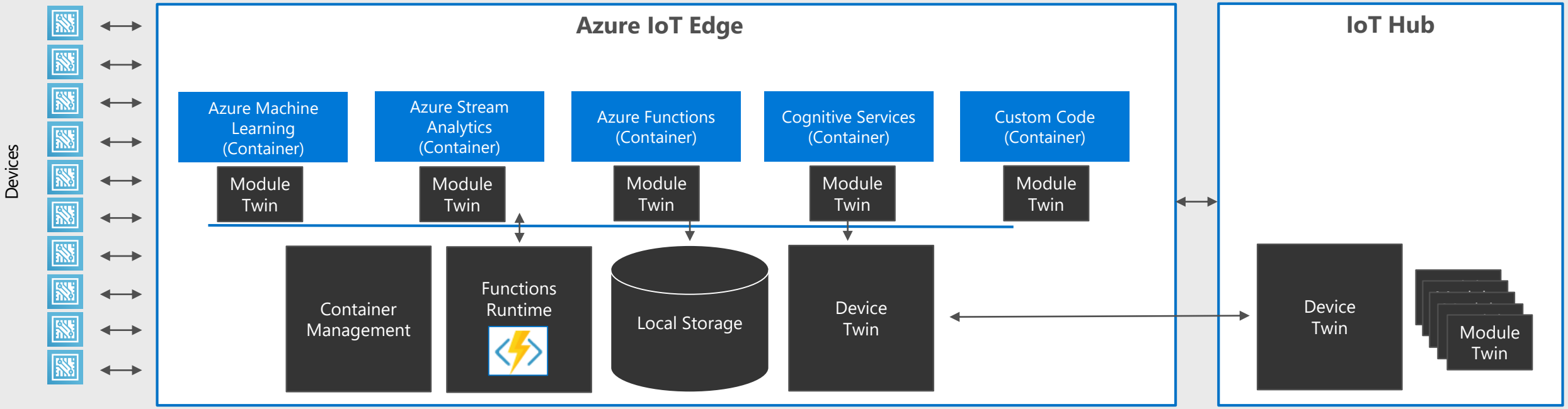
Azure IoT Edge Deployment



Azure IoT Edge

- Container based modules
- Azure Functions
- Azure Stream Analytics
- Azure Machine Learning
- Cognitive Services

- Offline / Synchronized Device Twins
- Local Storage
- Cloud Management & Deployment
- High Availability / Fault Tolerance
- Cloud Dev/Test Support



Azure **Sphere**

A new **Azure Sphere class of MCUs,** from silicon partners, with built-in Microsoft security technology provide connectivity and a dependable **hardware root of trust**

A new **Azure Sphere OS** secured by Microsoft for the devices 10-year lifetime to create **a trustworthy platform** for new IoT experiences

The **Azure Sphere Security Service** guards every Azure Sphere device; it **brokers trust** for device-to-device and device-to-cloud communication, **detects emerging threats,** and **renews device security**

TIME TO CONCLUDE

AZEO
talents & technology

Take Away : à retenir !

- **Design Hardware**
- **Optimisation du modèle**
- **Container**