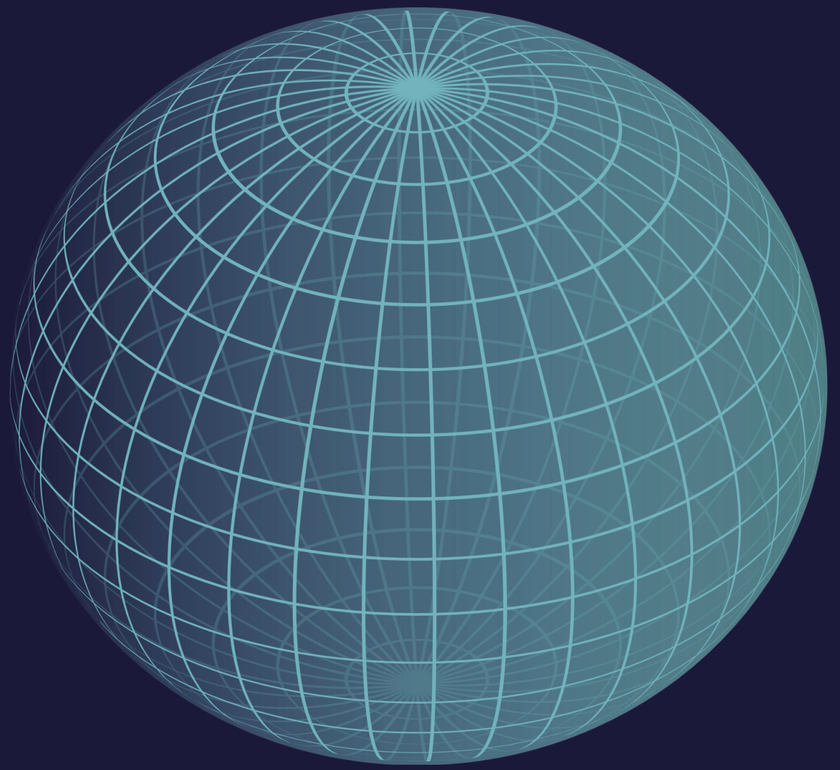
ML en Sistemas Embebidos



Alejandro Arias

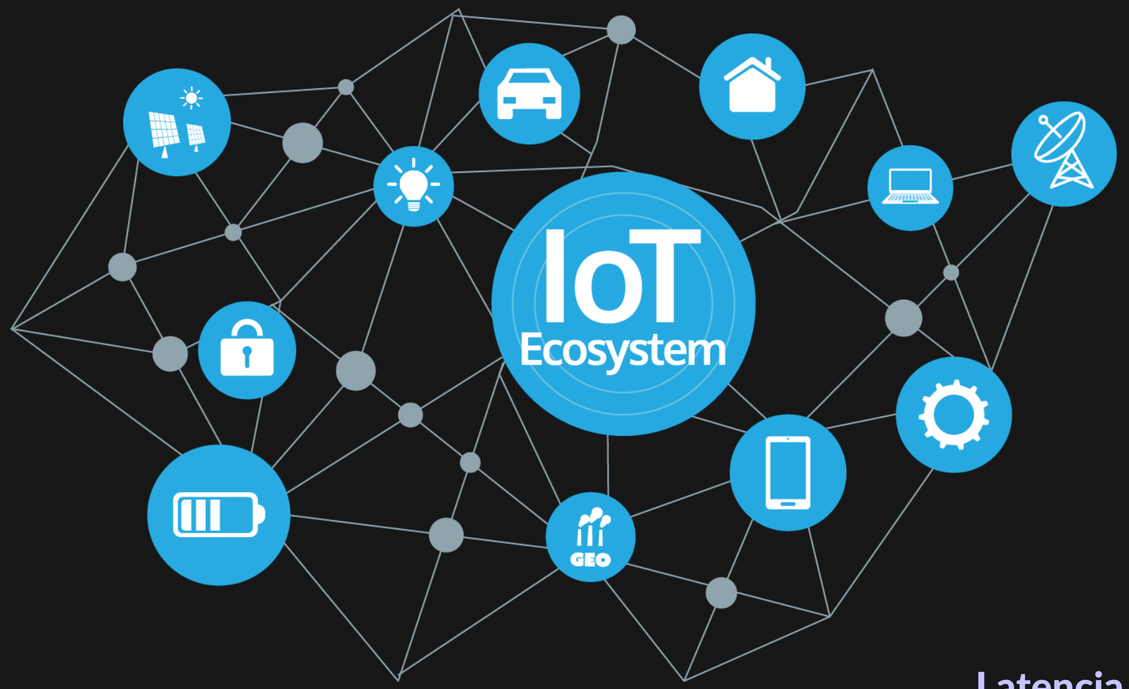
Plantilla: Sebastián Sierra

Motivación

Consumo energético

- Tesla K20 GPU: 225W ($139,99)
- Tesla T4 GPU: 70W ($2,199)
- Nvidia Jetson Nano: 5W ($99)
- STM32F7x7: 640mW ($14,91)

Privacidad

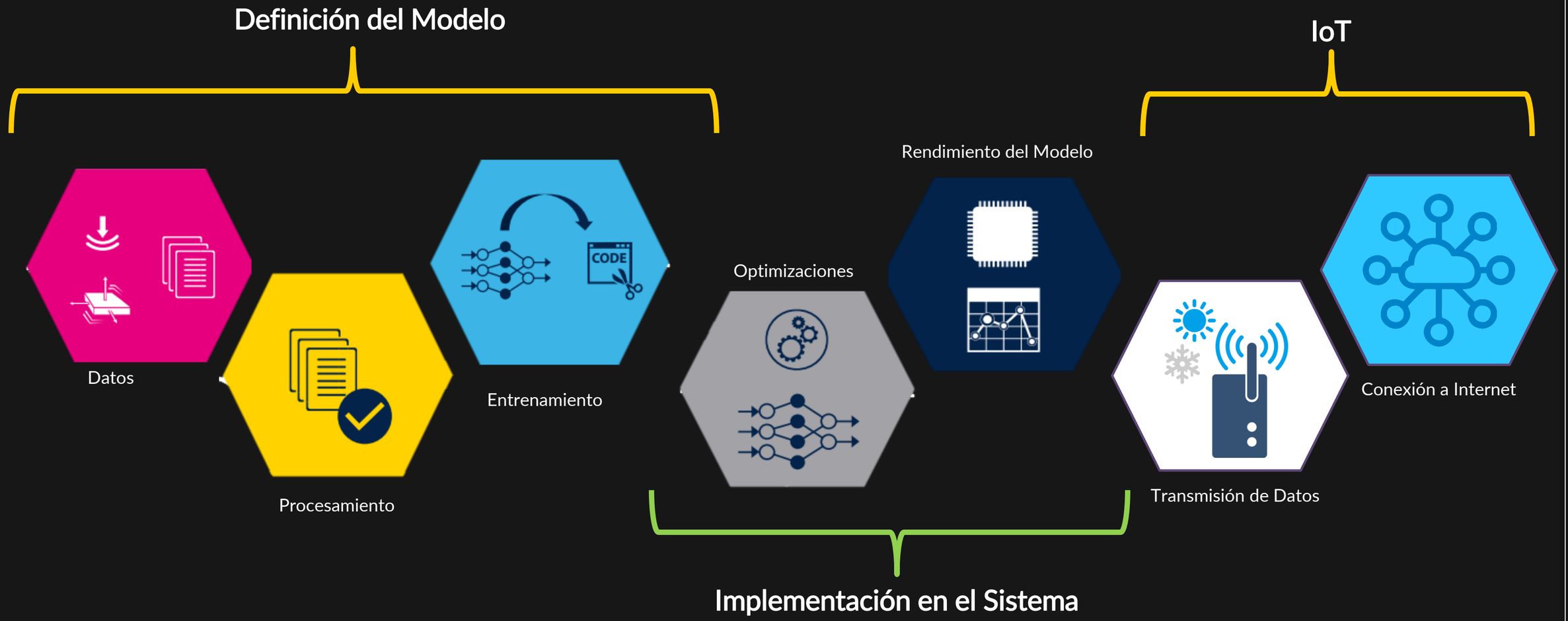


Almacenamiento

Latencia

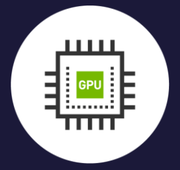
- Factores externos como: conexión a internet

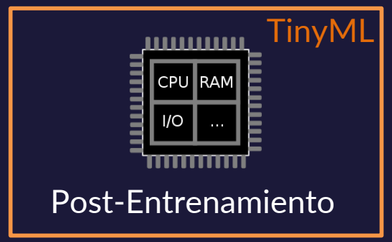
ML en Sistemas Embebidos



"Tiny machine learning is the intersection of machine learning and embedded internet of things devices. The field is an emerging engineering discipline that has the potential to revolutionize many industries."

Entrenamiento





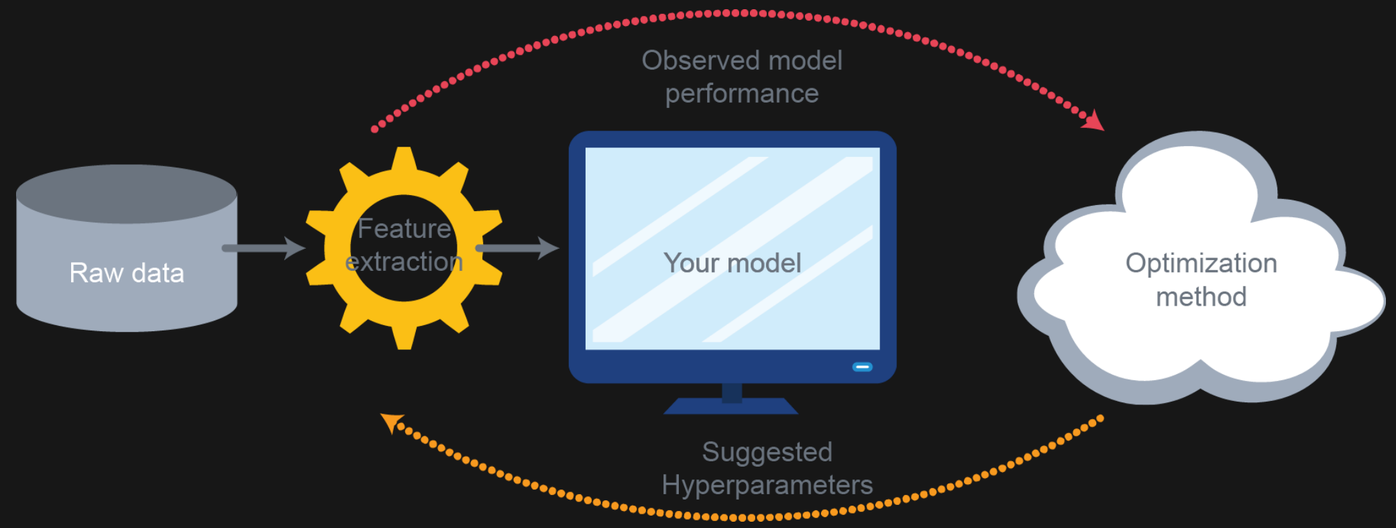
- ¿Cómo afecta la compresión de los modelos?
  - Desempeño (latencia, accuracy)

- ¿Cuál es objetivo final de sistema embebido?

≡ Optimizaciones

Redundancia presente en las redes neuronales

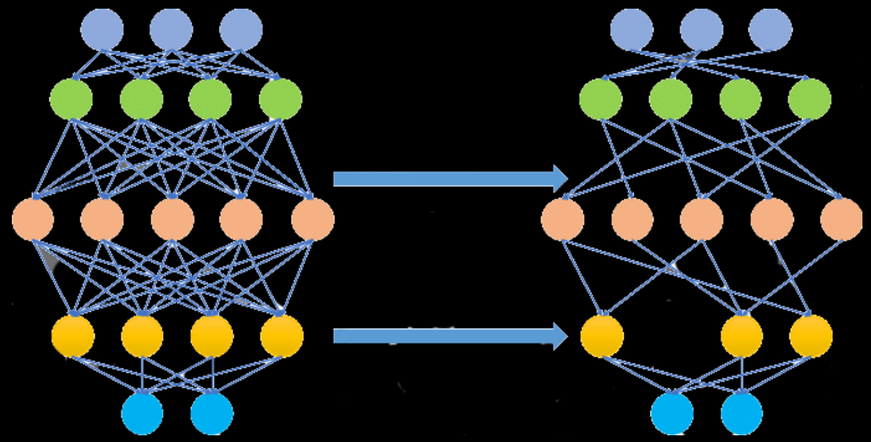
Redes más pequeñas permiten obtener un rendimiento similar.



≡ Quantization

Reducir tamaño de los pesos de la red: Float32 -> Float16/int8

≡ Pruning



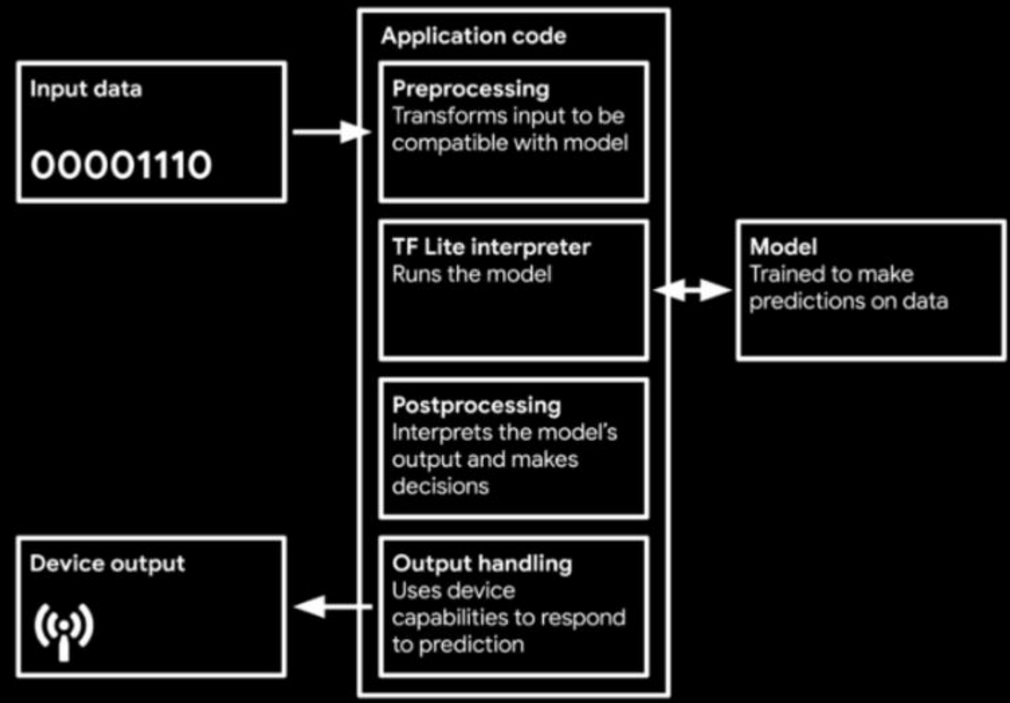
Eliminar conexiones/neuronas innecesarias

Compilación

TensorFlow Lite

TF lite - 500KB **TF micro - 20KB**

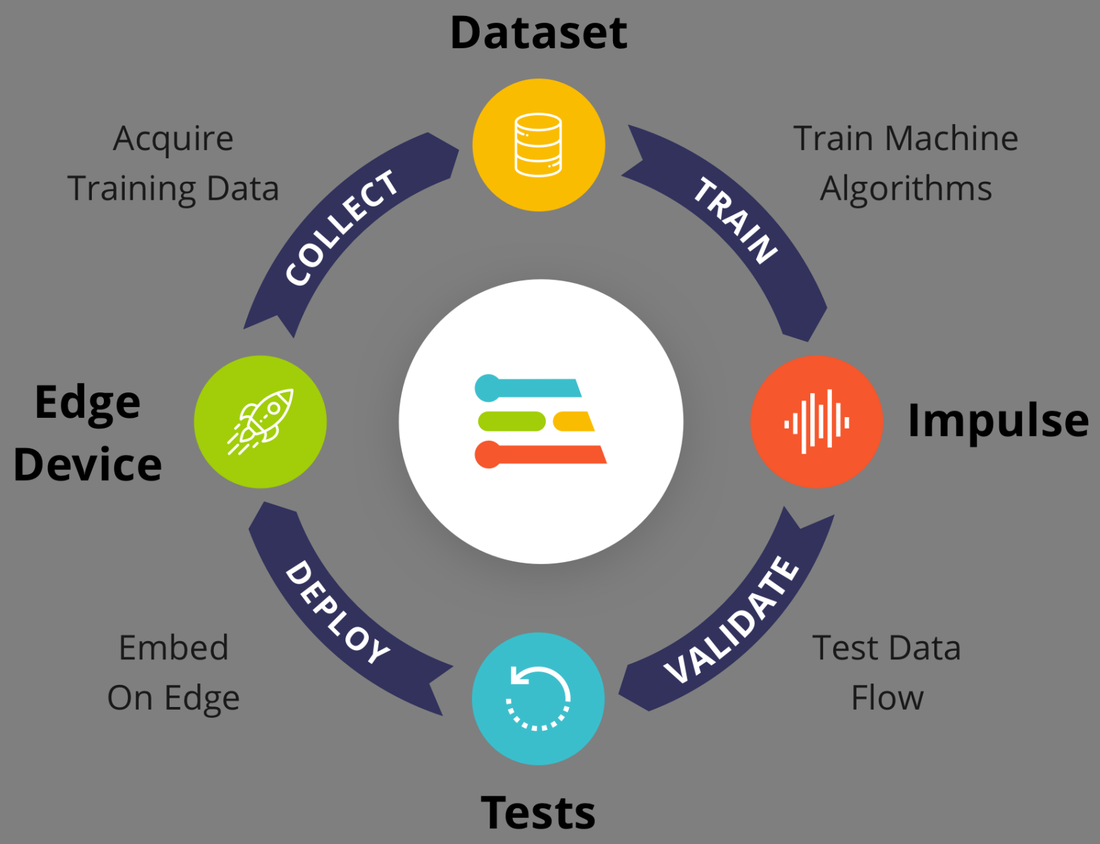
Se busca eliminar cualquier funcionalidad innecesaria, como la depuración y visualización de las redes



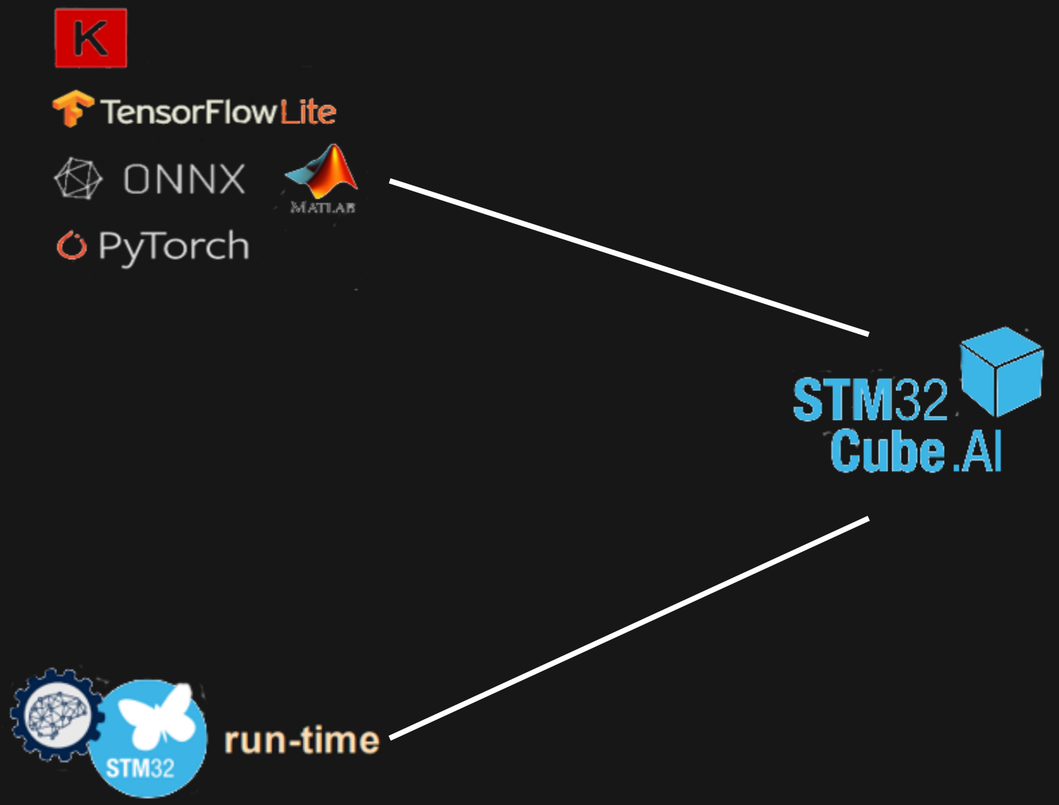
The workflow of TInyML application (Source: TinyML book by Pete Warden and Daniel Situnayake)

Implementación

EDGE IMPULSE



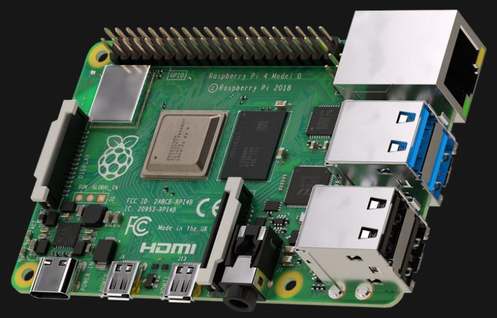
STM32



Alternativas de Hardware

“Alto” Rendimiento

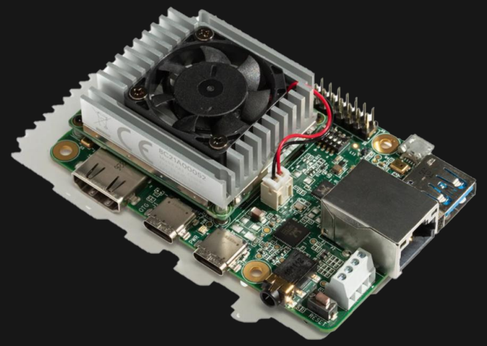
- Raspberry Pi 4:
  - Quad-core CPU 64-bit processor
  - RAM: 4GB



- Rock Pi N10:
  - Six-core CPU: Dual Cortex-A72 + Quad Cortex-A53
  - GPU of RK3399Pro is Mali T860MP4
  - NPU: 8/16 bit computing

- Nvidia Jetson Nano:
  - Quad-core ARM Cortex A57
  - Nvidia Maxwell arch. GPU (128 Cuda cores)
  - RAM: 4 GB 64-bit LPDDR4 1600MHz

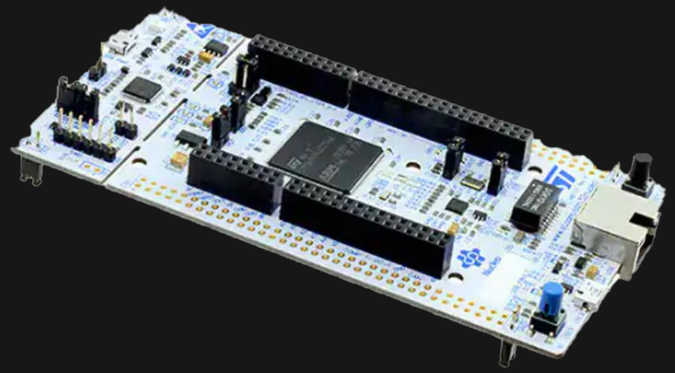
- Google Coral Dev Board:
  - CPU: Quad-core Cortex-A53, Cortex-M4F
  - GPU: Integrated GC7000 Lite Graphics
  - ML Accelerator: Google Edge TPU coprocessor
  - RAM: 1/4 GB LPDDR4
  - Flash Mem: 8 GB eMMC, MicroSD slot



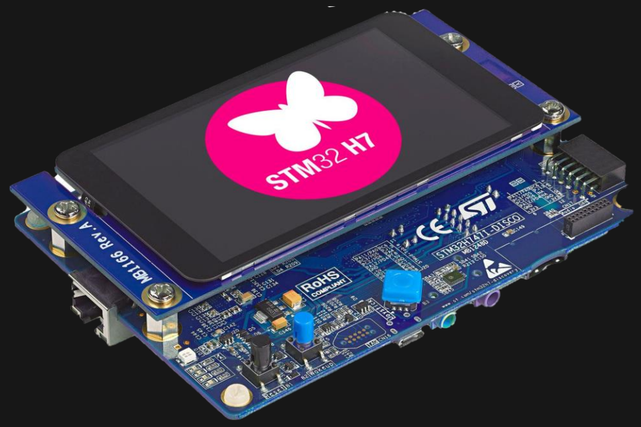
Alternativas de Hardware

Bajo Consumo

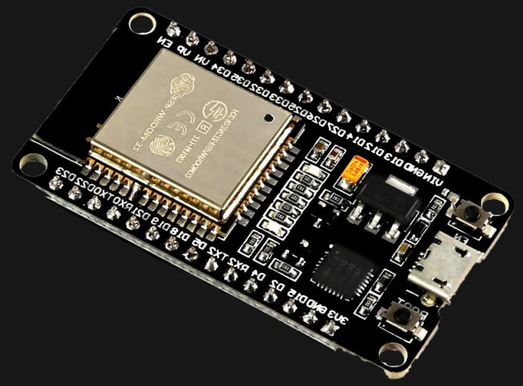
- STM32F767ZI:
  - Arm Cortex-M7 MCU
  - 2MB Flash Memory
  - 512kB RAM
  - 216 MHz CPU



- STM32H745XI:
  - Dual-core Arm Cortex-M7 + Cortex-M4 MCU
  - 2MB Flash Memory
  - 1MB RAM
  - 480 MHz CPU



- ESP32:
  - CPU: Xtensa dual-core (or single-core) 32-bit LX6 microprocessor, operating at 160 or 240 MHz
  - Memory: 520 KiB SRAM, 448 KiB ROM



IoT: Transmisión de Datos

Alternativas para comunicación inalámbrica:
LoRaWAN, Zigbee, SigFox, ...

Consideraciones:

- Tamaño de paquetes
- Distancia
- Potencia
- Ubicación geográfica

Data

Central (Gateway)

IoT: Transmisión de Datos

- Tamaño de paquetes
- Distancia
- Potencia

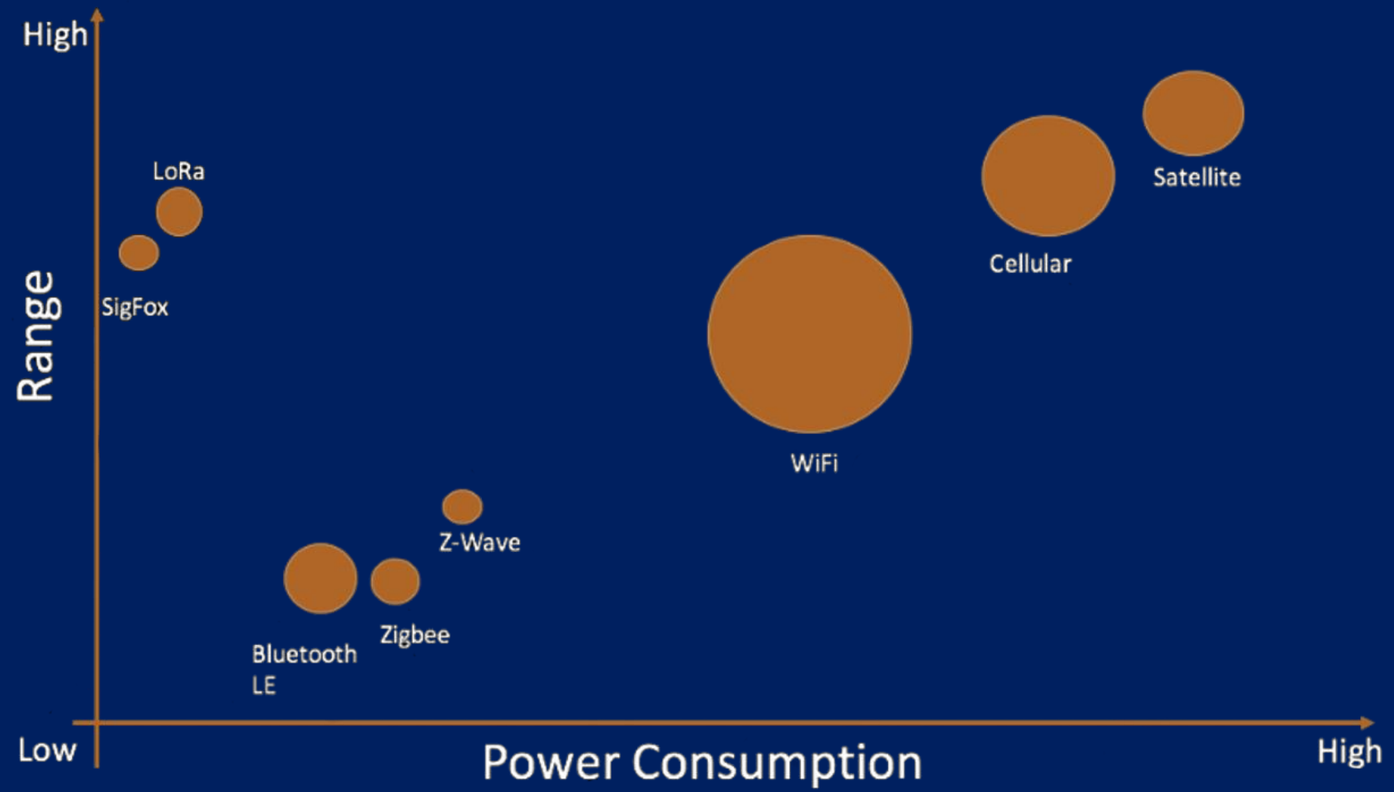
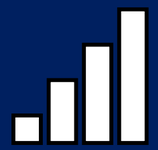


Imagen tomada de: Seeds of Silicon: Internet of Things for Smallholder Agriculture. Antony, Anish Paul Lu, Jennifer Sweeney, Daniel J.

IoT: Conexión a Internet

- Wi-Fi
- Celular
- Satelital (Swarm u otros)



SWARM