




UNIVERSIDAD NACIONAL DE COLOMBIA

# **INTELIGENCIA EMBEBIDA: CÓMO EL HARDWARE EJECUTA LA IA SIN NECESIDAD DE LA NUBE**

Rosemberth Steeven Preciga Puentes  
Luis Guillermo Vaca Rincón

Verificación de sistemas digitales  
2025 - I



# TABLA DE CONTENIDO

- **Introducción**
- **Conceptos Clave**
- **Unidades EMM Y MAC: La base del hardware**
- **Cómo medir el rendimiento en IA local**
- **La precisión**
- **NPU Comerciales**
- **TinyML**
- **Conclusiones**



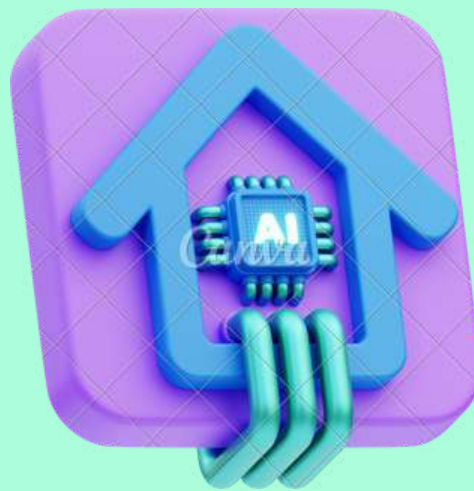


# ¿POR QUÉ IMPLEMENTAR IA LOCAL?

IA ya no es solo cosa de servidores: ahora la usamos en el bolsillo (celulares, relojes, sensores).



Mejor  
experiencia  
para el usuario



Mayor  
Privacidad



Menor  
Consumo  
Energetico



Menor  
Latencia

# CONCEPTOS CLAVE

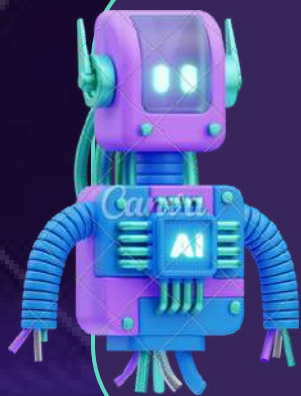
Fundamentales para comprender como funciona la IA tanto local como en la nube



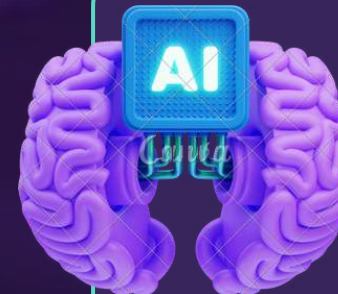
**Red Neuronal Artificial**



**Tensor**



**Nodo**



**Numeros y operaciones  
punto flotante**



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Red Neuronal Artificial

También llamadas redes neuronales artificiales, son un modelo computacional inspirado en la estructura del cerebro humano. están constituidas a partir de capas de nodos (neuronas) interconectadas entre si [1] .

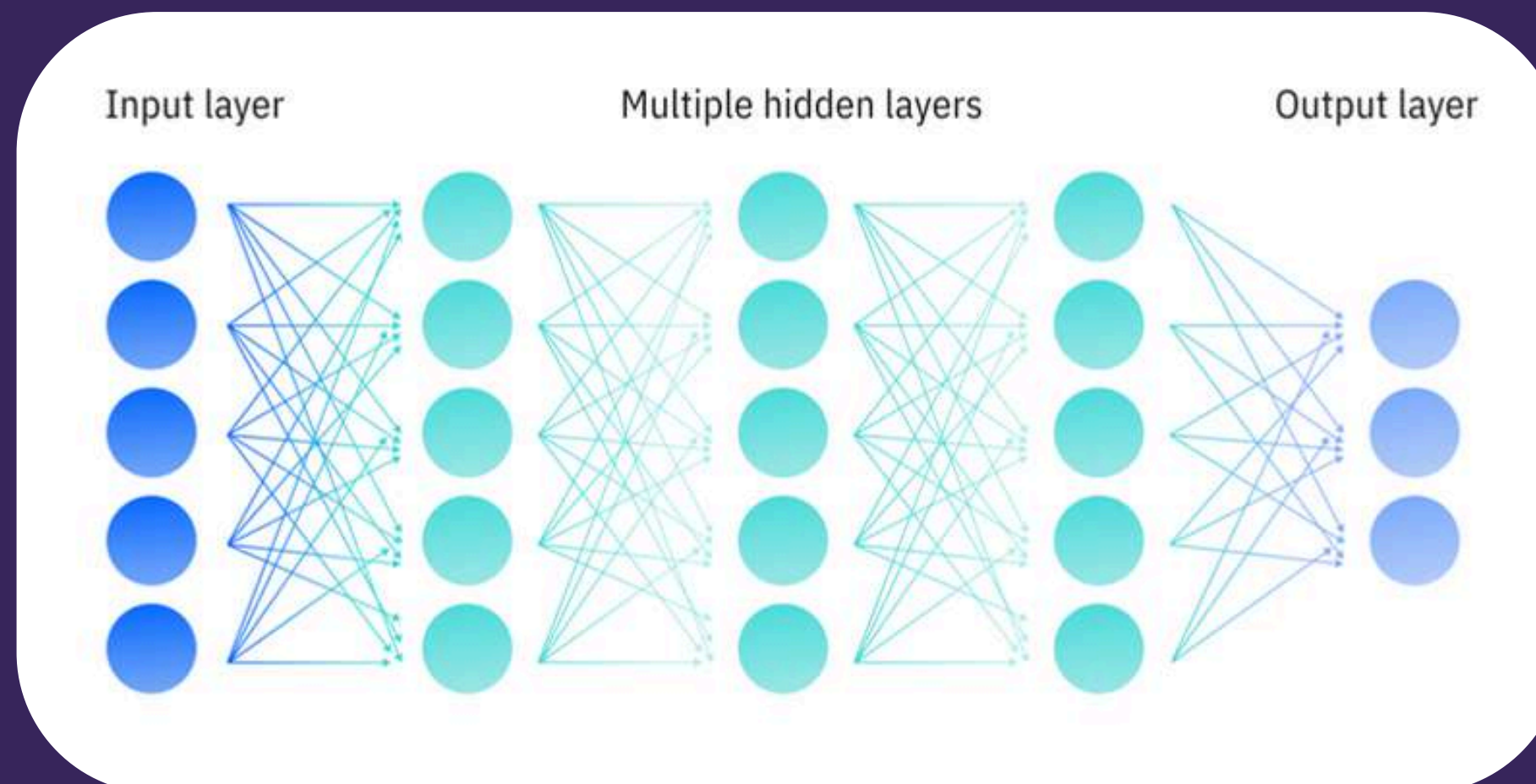


Imagen tomada de: <https://www.ibm.com/think/topics/artificial-intelligence>



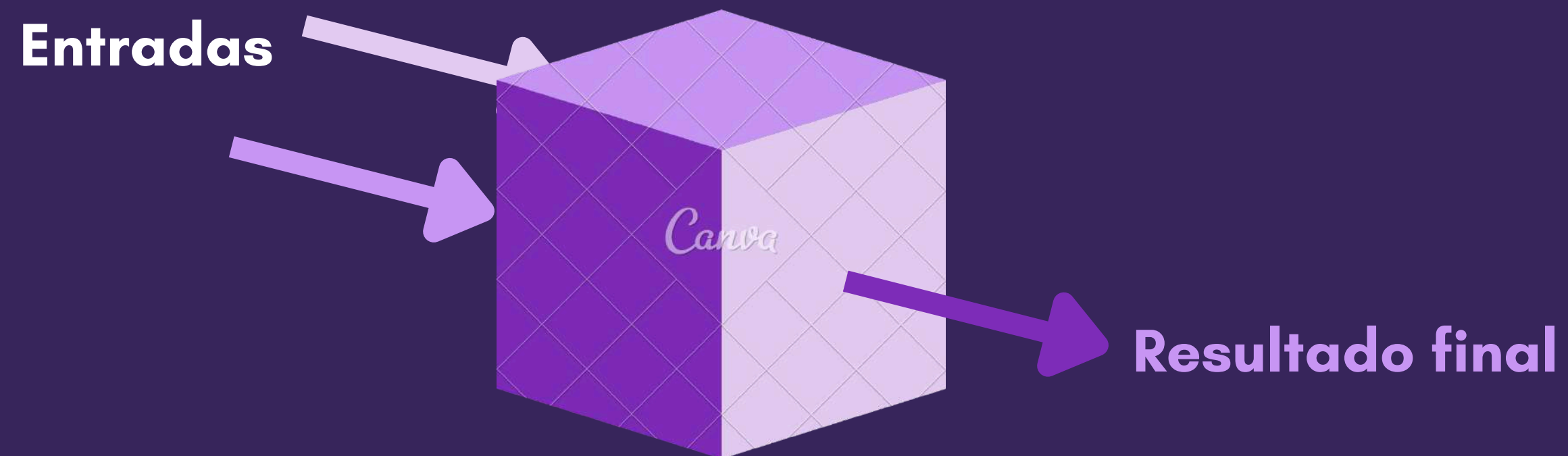


# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Nodo

Es una unidad básica de procesamiento en una red neuronal. Cada nodo recibe entradas, las procesa utilizando una función de activación y envía el resultado a otros Nodos [2].



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube



## Tensor

Los tensores son objetos matemáticos que almacenan valores numéricos y que pueden tener distintas dimensiones. Así, por ejemplo, un tensor de 1D es un vector, de 2D una matriz, de 3D un cubo incluso elementos 4D [3]. (En Python conocidos como NumPy Arrays)

**¿Como nos vemos nosotros en las redes sociales (a nivel de algoritmo) ?**

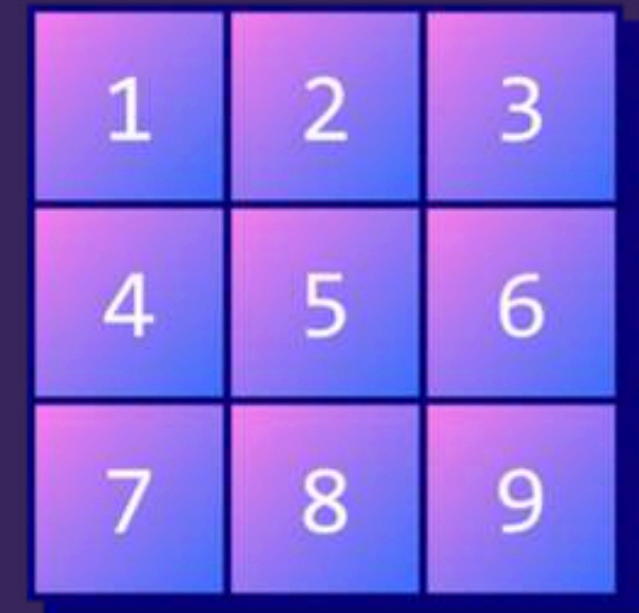


Imagen tomada de: <https://towardsdatascience.com/what-is-a-tensor-in-deep-learning-6dedd95d6507/>



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube



## Tensor



Tensor 750x750x3 (RGB Image)

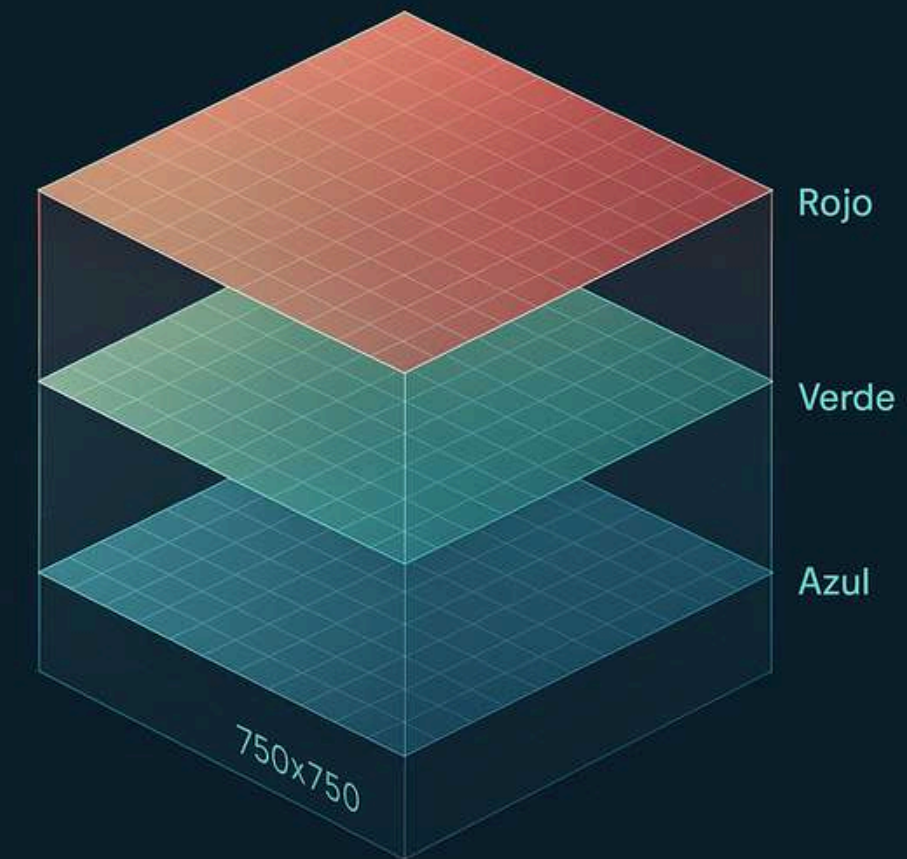
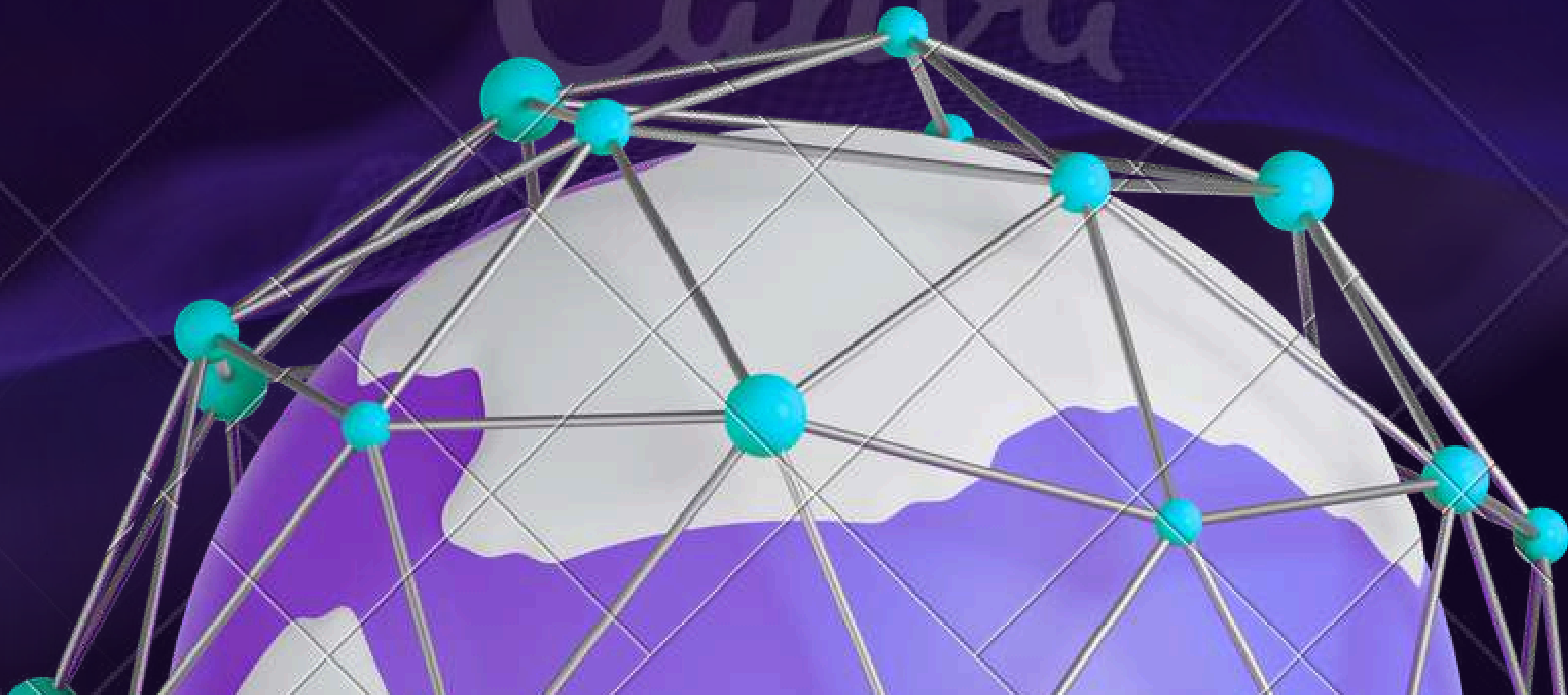


Imagen generada por ChatGPT



# ¿Y SI LO VEMOS SOBRE LA IMPLEMENTACIÓN FÍSICA?

EMM Y MACs: La base del hardware



# ¿QUE ES UNA MAC?

Una MAC unit es una unidad de hardware especializada para multiplicar y luego acumular. Lo requerido para emular lo que hacen nuestras neuronas. [5]

$$y = \sum_{i=0}^n (w_i \cdot x_i)$$

- $x_i$  Son los tensores de entrada
- $w_i$  Son los **pesos** aprendidos durante el entrenamiento
- $y$  Es la resultado/salida final

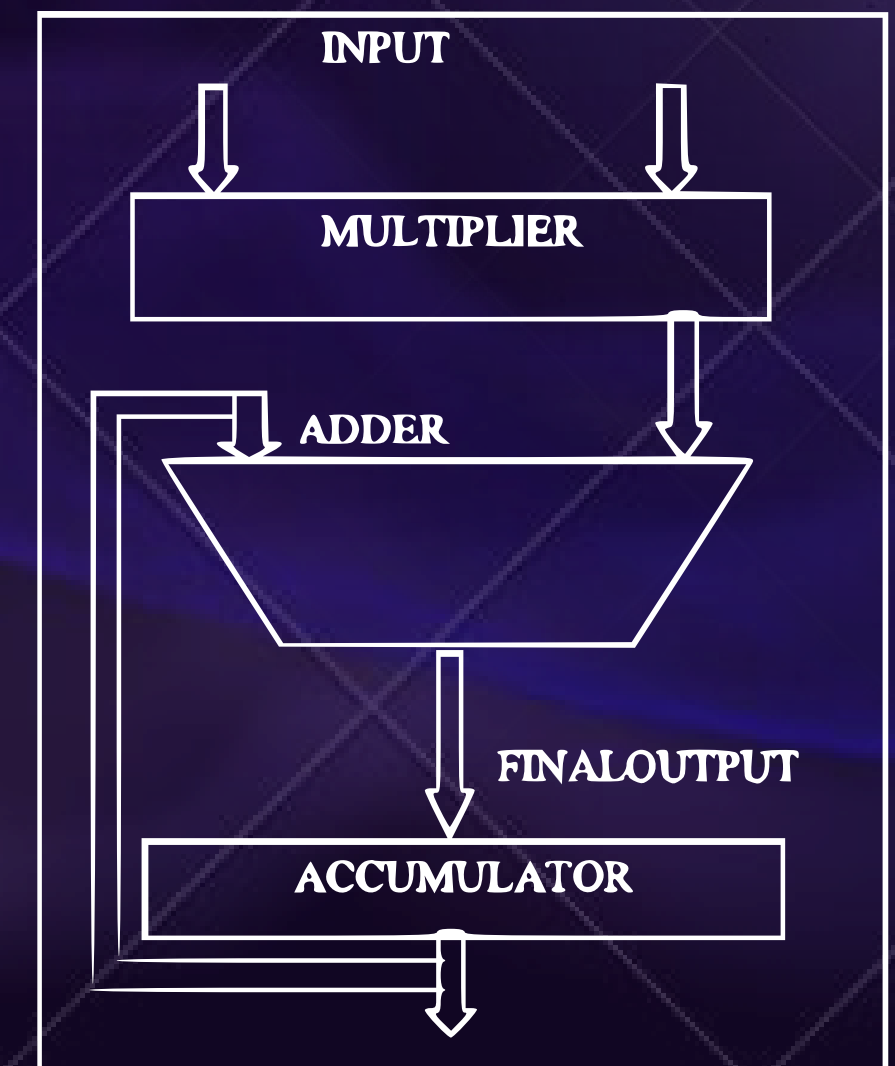
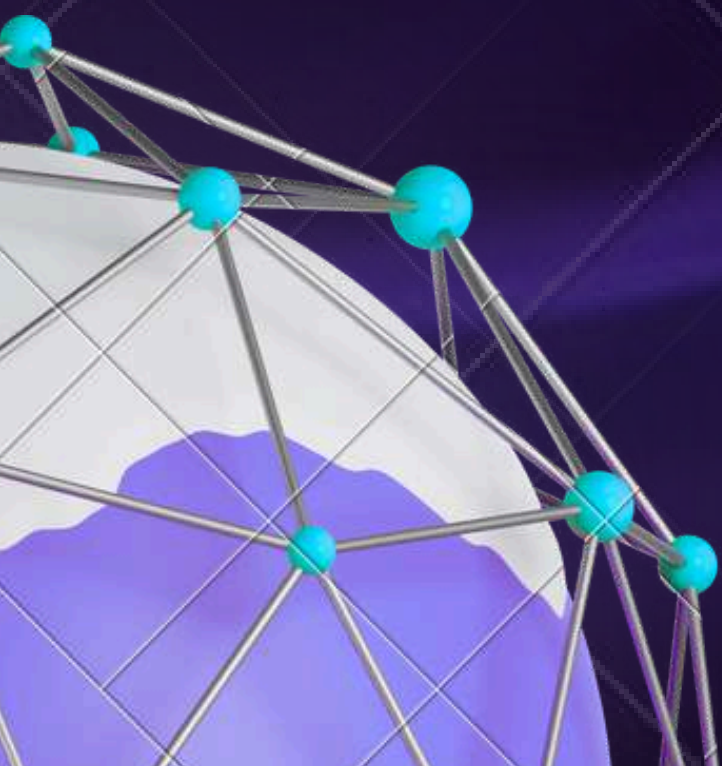


Diagrama tomado de: [https://www.researchgate.net/figure/Block-Diagram-of-MAC-architecture\\_fig1\\_353074076](https://www.researchgate.net/figure/Block-Diagram-of-MAC-architecture_fig1_353074076)



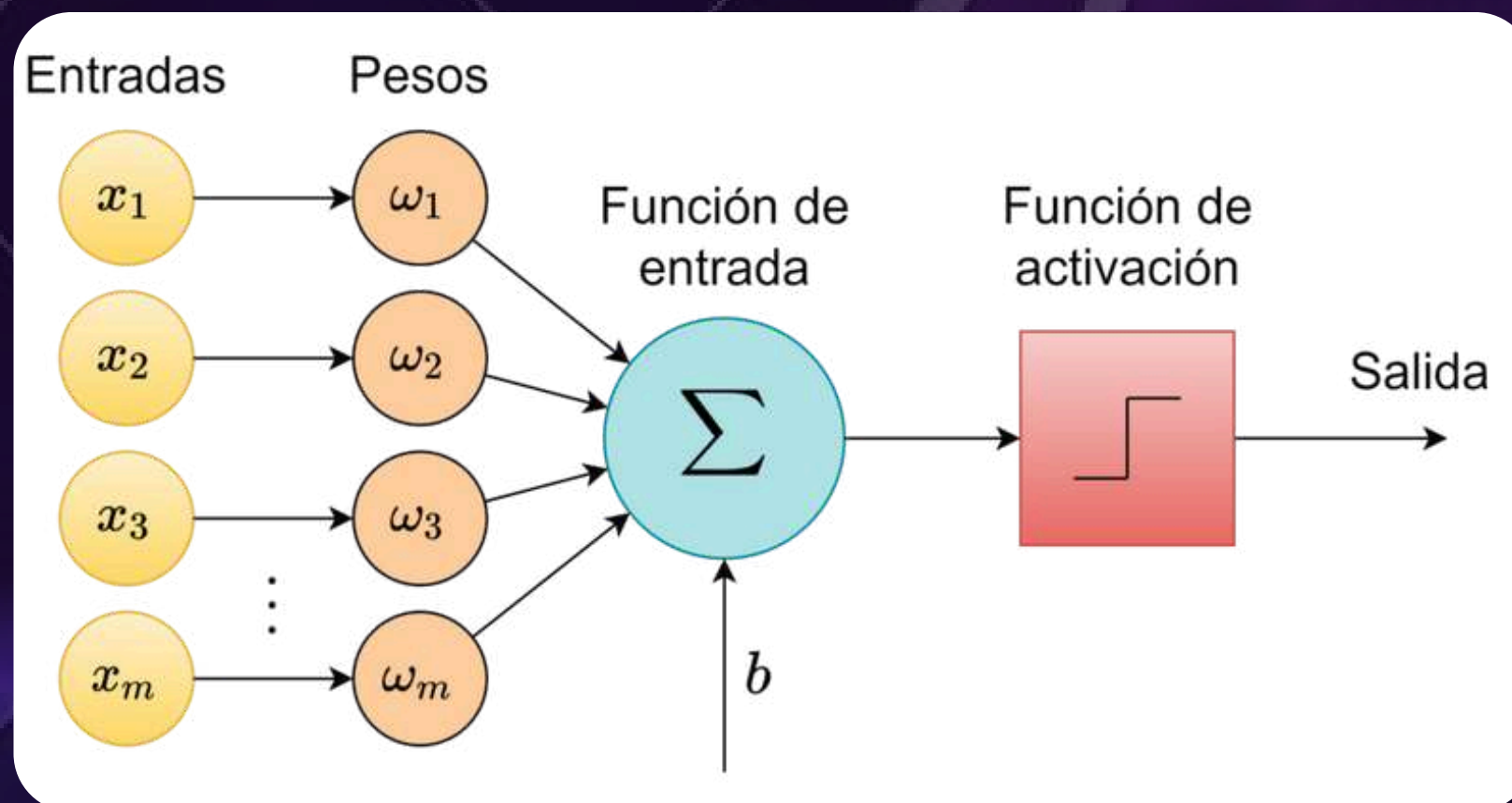
**PERO EN REALIDAD NADA ES  
LINEAL...**



# ¿COMO TRATAMOS LA NO LINEALIDAD?

## Función de Activación

Es una transformación matemática que determina la salida de un nodo. Influyen en el aprendizaje y procesamiento de una red al determinar cómo se transmiten las señales entre nodos. Introducen no linealidad, permitiendo a la red aprender patrones complejos.

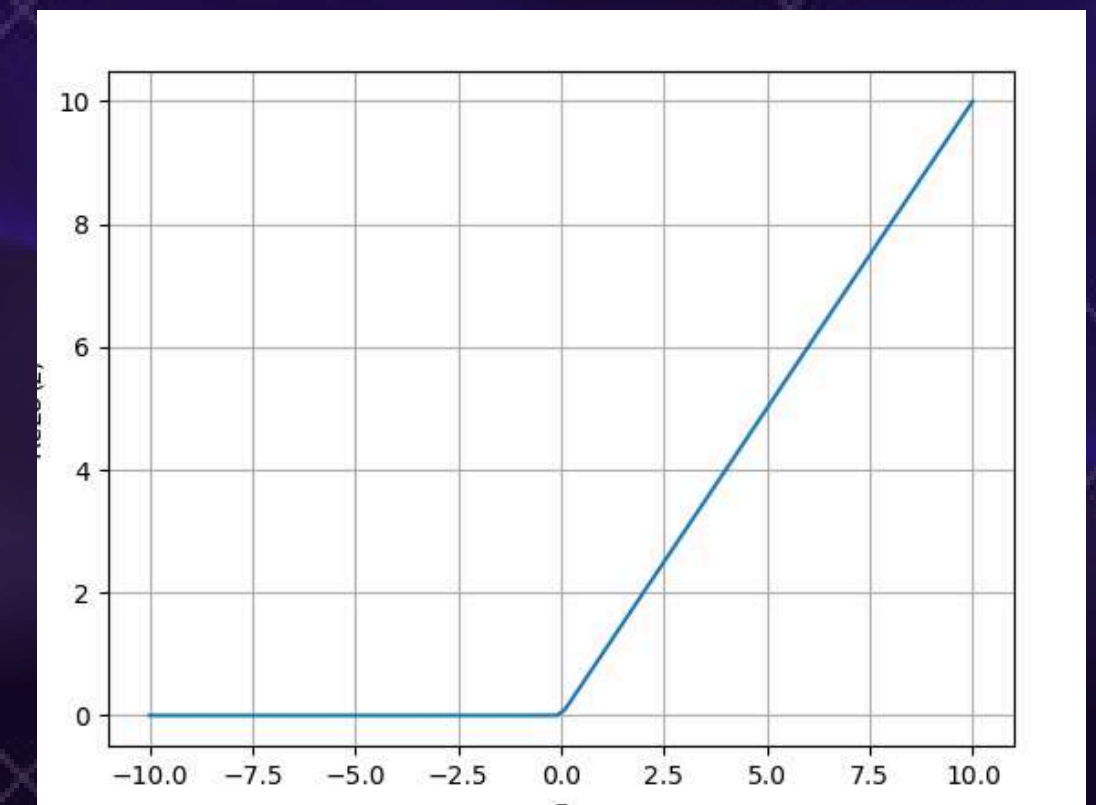
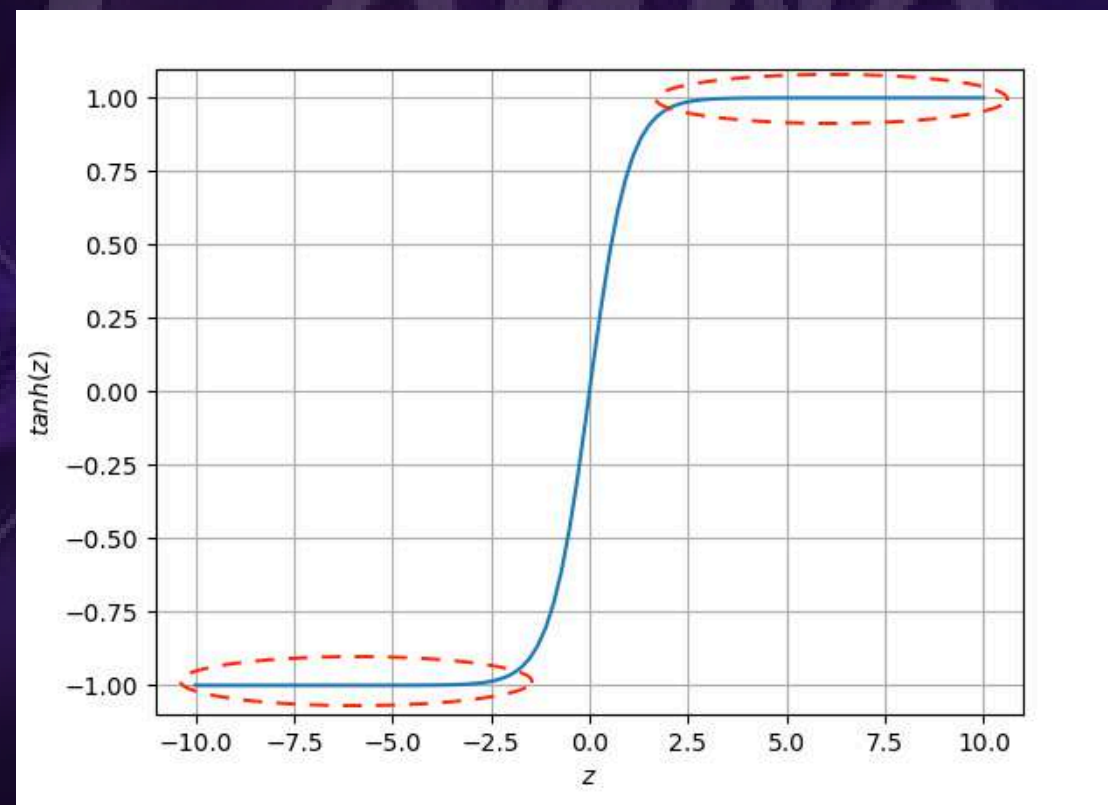
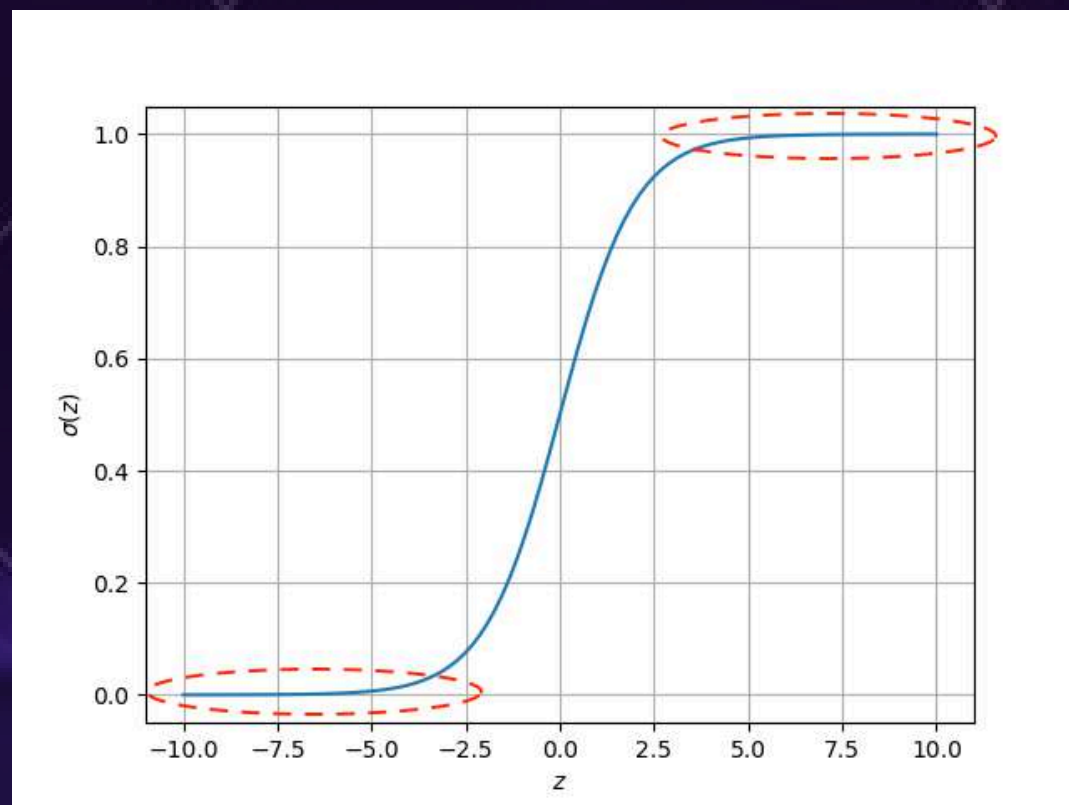


¿Porque existe? Problemas como clasificación y regresión no pueden resolverse solo con líneas rectas porque los datos están separados de forma mucho más compleja, entonces con estas funciones se puede construir curvas, superficies complejas para clasificar los datos



# Tipos de Función de Activación

- **Sigmoide:** Produce una salida entre 0 y 1, útil para modelos de clasificación binaria
- **Tanh (Tangente Hiperbólica):** Similar a la sigmoide, pero varía entre -1 y 1, ofreciendo una mejor eficiencia en ciertas aplicaciones.
- **ReLU (Unidad Lineal Rectificada):** Proporciona una salida lineal para entradas positivas y cero para negativas, es eficiente computacionalmente y reduce el desvanecimiento del gradiente, siendo popular en redes profundas.

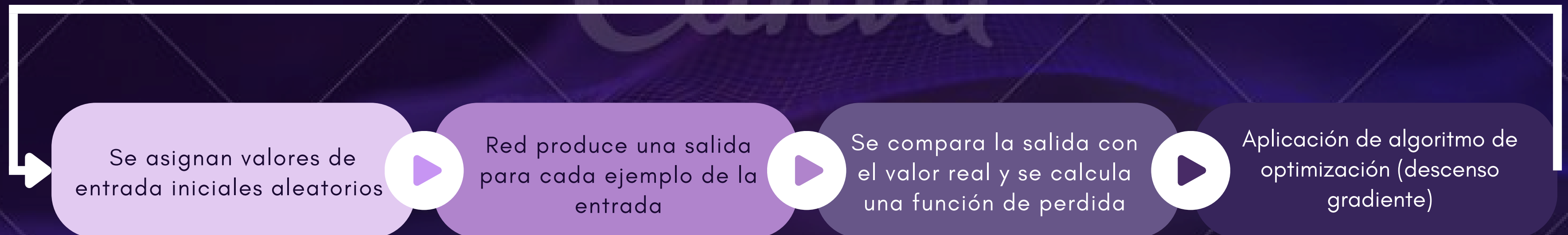


# ¿COMO TRATAMOS LA NO LINEALIDAD?

## Pesos

Un peso es un valor numérico que determina cuánta importancia le da la neurona a una entrada, es fundamental para determinar cómo la red interpreta y responde a los datos de entrada. En IA local estos se ajustan durante entrenamiento previo. [7]

## Repetición

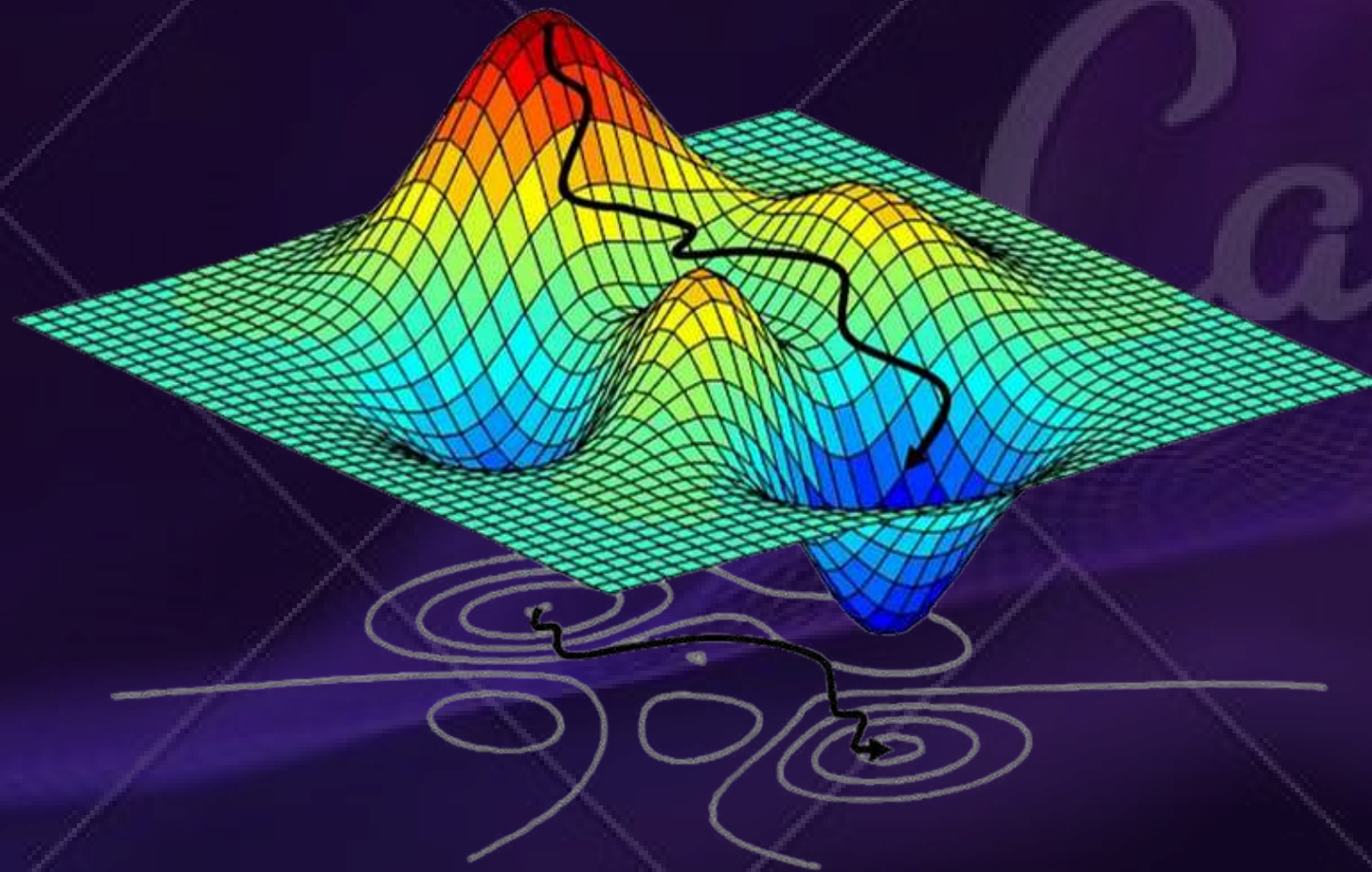




# ¿COMO TRATAMOS LA NO LINEALIDAD?

## Funcion de Perdidas

Es una función que cuantifica la discrepancia o el error entre la predicción de un modelo y el valor real.



Se ajustan los parámetros del modelo en la dirección opuesta al gradiente, es decir, “descendiendo” hacia el valor mínimo de la función de pérdida. La magnitud del ajuste se controla por la tasa de aprendizaje

Imagen obtenida de: <https://medium.com/data-science/an-introduction-to-surrogate-optimization-intuition-illustration-case-study-and-the-code-5d9364aed51b>

# DE MANERA ABSTRACTA:

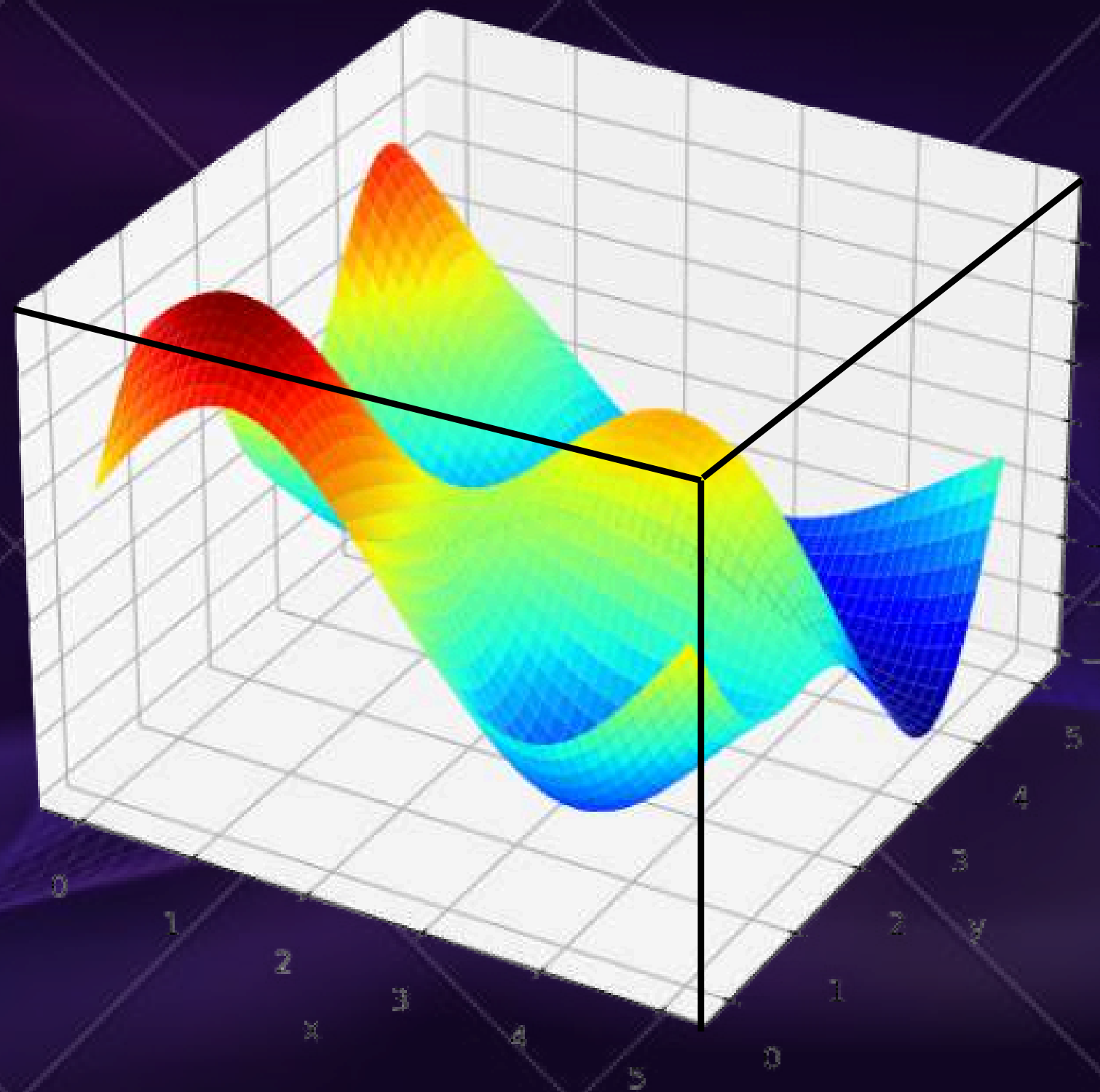


Imagen tomada de: <https://syelendrapramuditya.wordpress.com/2023/11/28/3d-surface-plot-of-xyz-data-from-file-with-python-numpy-matplotlib/>





# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## ¿Cómo se mide el rendimiento en IA?

TOPS (Tera Operations Per Second): Mide la capacidad de un procesador para realizar billones de operaciones IA por segundo.

$$TOPS = 2 \times MAC \text{ UNIT COUNT} \times FREQUENCY / 1 \text{ TRILLION.}$$

- 2: cada unidad MAC realiza 2 operaciones por ciclo.
- MAC unit count: número total de unidades MAC trabajando en paralelo.
- Frequency: velocidad a la que trabaja el hardware (en Hz).
- 1 trillion: Se divide por un billón para convertir las operaciones a tera



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

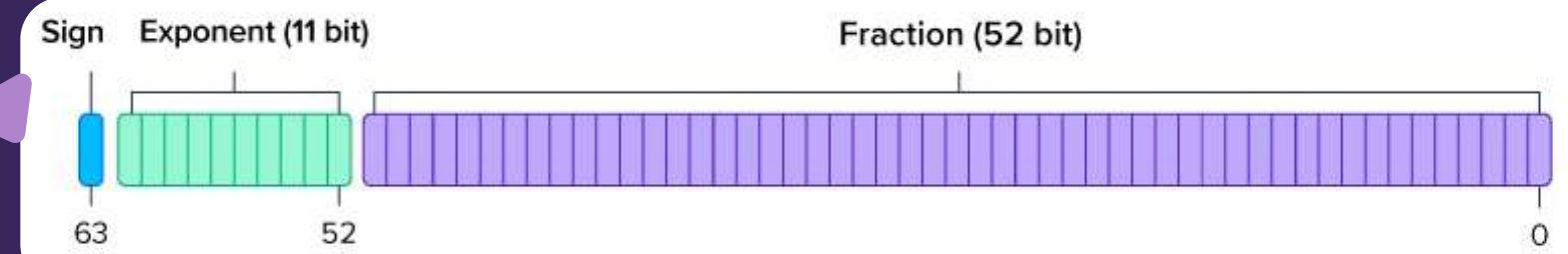
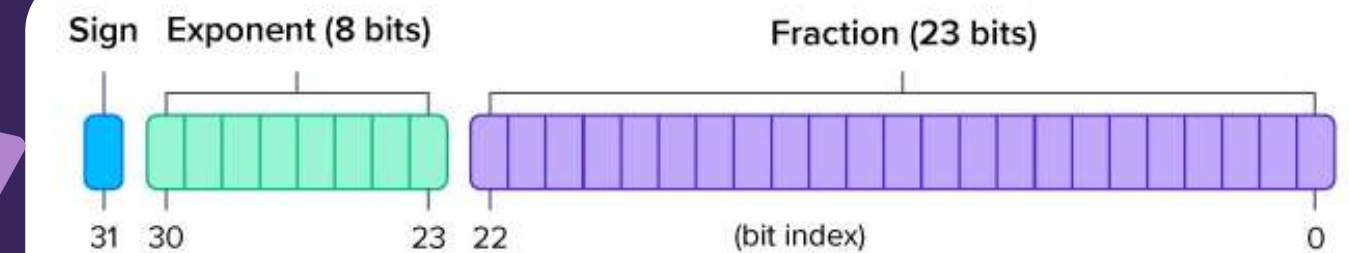
## Números y operaciones de punto flotante

Es una forma de notación científica usada en sistemas digitales para representar números reales extremadamente grandes o pequeños de una manera eficiente y compacta. El estándar actual para la representación en coma flotante esta regida bajo el estandar IEEE 754 [4]

Formato	Bits totales	Bits significativos	Bits del exponente	Número más pequeño	Número más grande
Precisión simple	32	23 + 1 signo	8	$\sim 1.2 \times 10^{-38}$	$\sim 3.4 \times 10^{38}$
Precisión doble	64	52 + 1 signo	11	$\sim 5.0 \times 10^{-324}$	$\sim 1.8 \times 10^{308}$

Tabla 2: estándar de IEEE 754

$$\begin{array}{cc} \text{signo} & \text{exponente} \\ \underbrace{+}_{\text{signo}} \underbrace{6.02}_{\text{mantisa}} \cdot \underbrace{10}_{\text{base}}^{-\underbrace{23}_{\text{exponente}}} & \underbrace{+}_{\text{signo}} \underbrace{1.01110}_{\text{mantisa}} \cdot \underbrace{2}_{\text{base}}^{-\underbrace{1101}_{\text{exponente}}} \end{array}$$



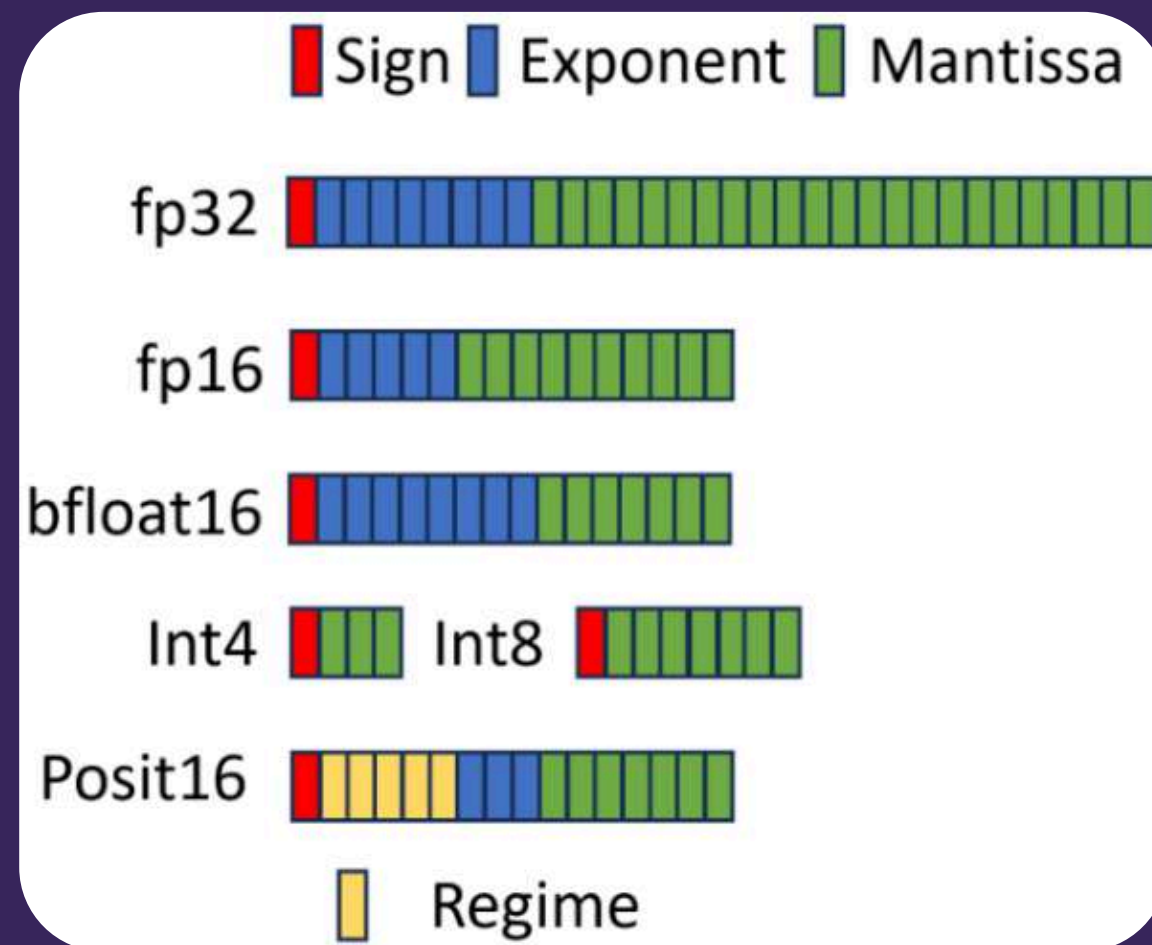
Imágenes tomadas de: <https://www.exxactcorp.com/blog/hpc/what-is-fp64-fp32-fp16>



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

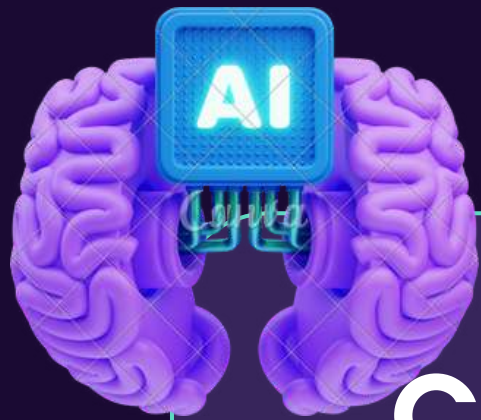
## Formato de datos



“Si observamos los pesos en una inferencia, estos pueden ser millones o incluso miles de millones de valores que deben transferirse a los elementos computacionales. Si podemos reducirlos de 32 bits a 16, o incluso a 8, obtenemos una mejora de 2 a 4 veces en nuestra capacidad para transferir esos datos. Generalmente, el movimiento y el almacenamiento de datos tienden a ser nuestros mayores cuellos de botella y nuestros mayores consumidores de energía.” – Russell Klein, director de programa del equipo Catapult HLS en Siemens EDA

Imagen tomada de: <https://semiengineering.com/data-formats-for-inference-on-the-edge/#:~:text=And%20that%20is%20only%20the,%E2%80%9D>





# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Consumo energético según la precisión

Operation	Energy (pJ)	Area (Normalized)
32-bit Floating-Point Multiplication	3.7	6×
32-bit Floating-Point Addition	0.9	4×
8-bit Fixed-Point Multiplication	0.03	1×
8-bit Fixed-Point Addition	0.03	1×

Reducir precisión ahorra:

- Consumo de energía
- Espacio físico (silicio)

Imagen tomada de: <https://semiengineering.com/data-formats-for-inference-on-the-edge/#:~:text=And%20that%20is%20only%20the,%E2%80%9D>

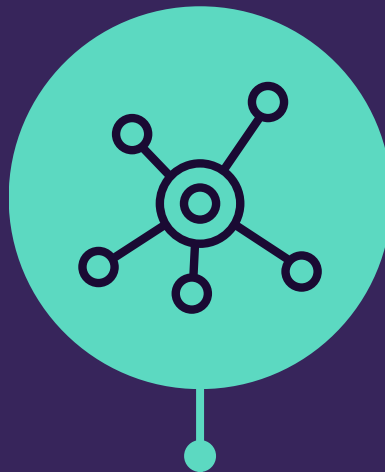


# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Cuantificación

“Quantization is the process of reducing the precision of a digital signal, typically from a higher-precision format to a lower-precision format. This technique is widely used in various fields, including signal processing, data compression and machine learning.” – Bryan Clark, Senior Technology Advocate at IBM



### INFERENCIA MÁS RÁPIDA

Los cálculos con enteros son más rápidos que con punto flotante. El modelo responde más rápido, ideal para aplicaciones en tiempo real.



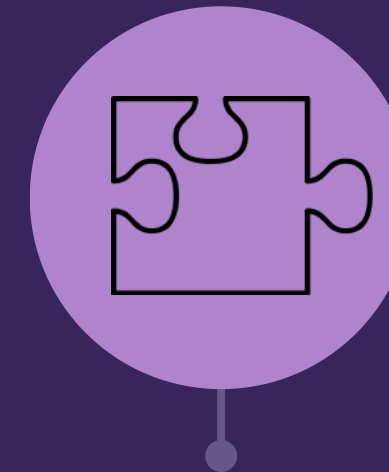
### EFICIENCIA

El modelo necesita menos recursos para funcionar. Puede ejecutarse en celulares, tablets y dispositivos pequeños.



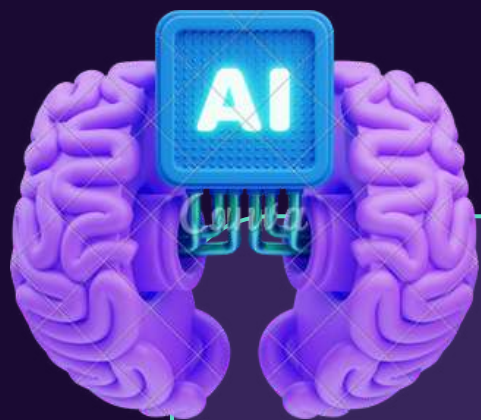
### MENOR CONSUMO DE ENERGÍA

Menos operaciones = menos gasto de batería. Ideal para móviles y dispositivos portátiles.



### COMPATIBILIDAD

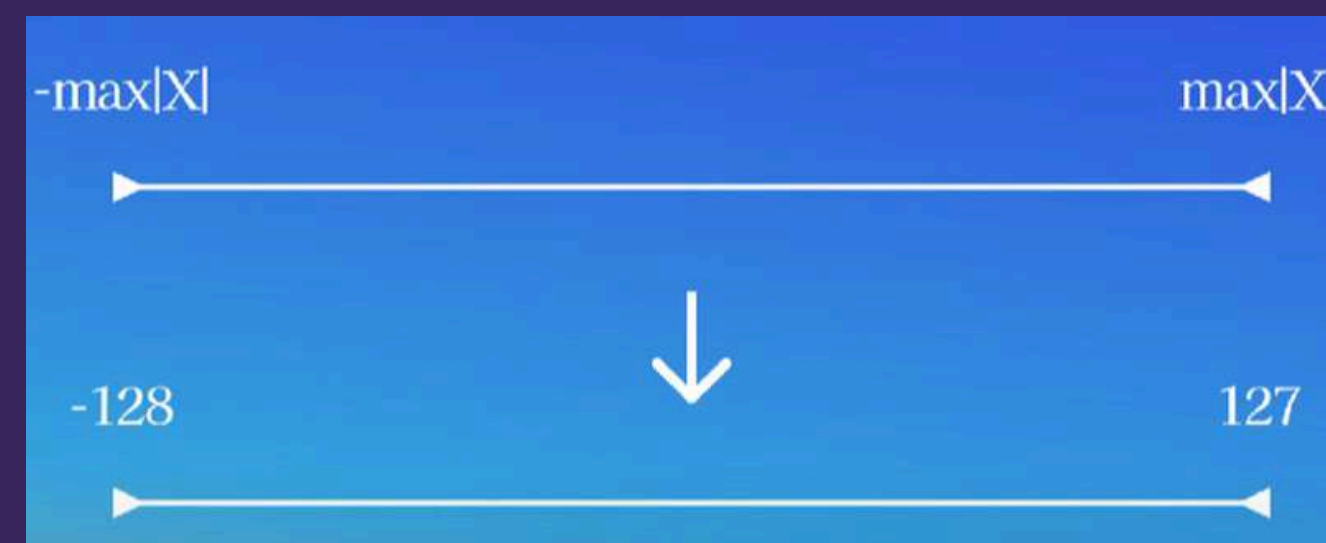
Permite usar hardware más simple y antiguo. Los modelos funcionan en más dispositivos.



# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Cuantificación



$$X_{\text{quant}} = \text{round} \left( \frac{127}{\max |X|} \cdot X \right)$$
$$X_{\text{dequant}} = \frac{\max |X|}{127} \cdot X_{\text{quant}}$$

Imágenes tomadas de: <https://www.youtube.com/watch?v=qqN63hbz1aI>





# CONCEPTOS CLAVE

Fundamentales para comprender como funciona la IA tanto local como en la nube

## Cuantificación

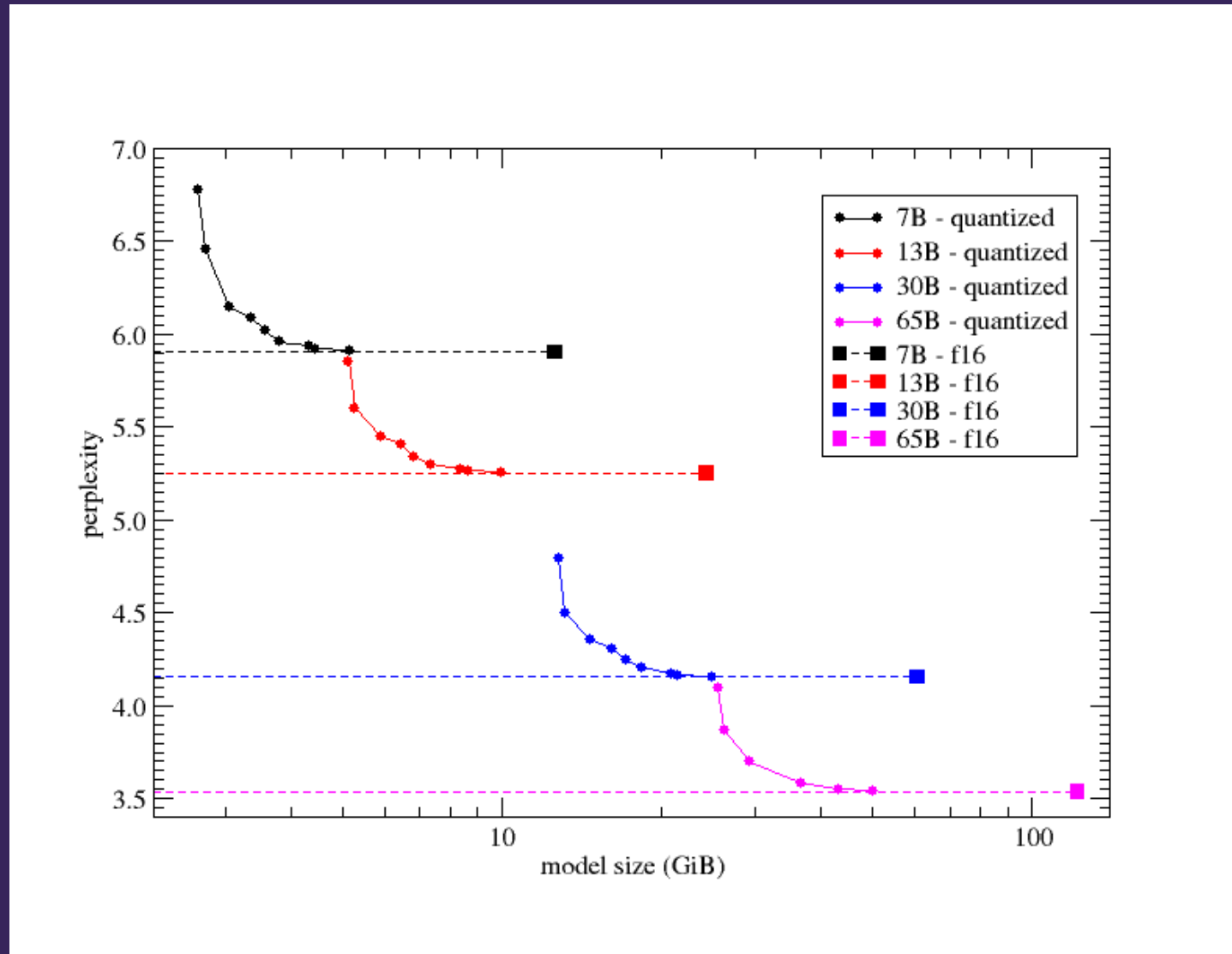


Imagen tomada de: <https://github.com/ggml-org/llama.cpp/pull/1684>

- PERPLEJIDAD = MÉTRICA QUE MIDE QUÉ TAN BUENO ES UN MODELO PARA PREDECIR.
- LOS CUADRADOS REPRESENTAN MODELOS SIN CUANTIFICAR (EN FP16).
- LOS CIRCULO REPRESENTAN MODELOS CUANTIFICADOS (MÁS PEQUEÑOS).
- "B" SIGNIFICA BILLONES DE PARÁMETROS
- REDUCIR EL TAMAÑO DEL MODELO, SE PIERDE ALGO DE PRECISIÓN, ESO SE REFLEJA COMO UN AUMENTO EN LA PERPLEJIDAD (EL MODELO ES MENOS PRECISO).

# DE MANERA FISICA:

## System Memory (Memoria del sistema)

- Guarda modelos, imágenes de entrada y datos grandes.
- Contiene los pesos entrenados que luego se cargan a la NPU.

## IOMMU (Input-Output Memory Management Unit)

- Traduce direcciones virtuales a físicas.
- Organiza el acceso de la NPU a la memoria del sistema.

## Global Control

- Coordina toda la NPU.
- Decide qué parte del modelo ejecutar y qué datos mover.

## MMU & DMA

- MMU: Gestiona direcciones locales internas.
- DMA: Mueve datos grandes (imágenes, pesos) entre memoria del sistema y memoria interna sin usar el procesador principal.

## Load / Store

- Carga datos (pesos, entradas).
- Guarda resultados parciales y finales de cada capa.

## Scratchpad SRAM

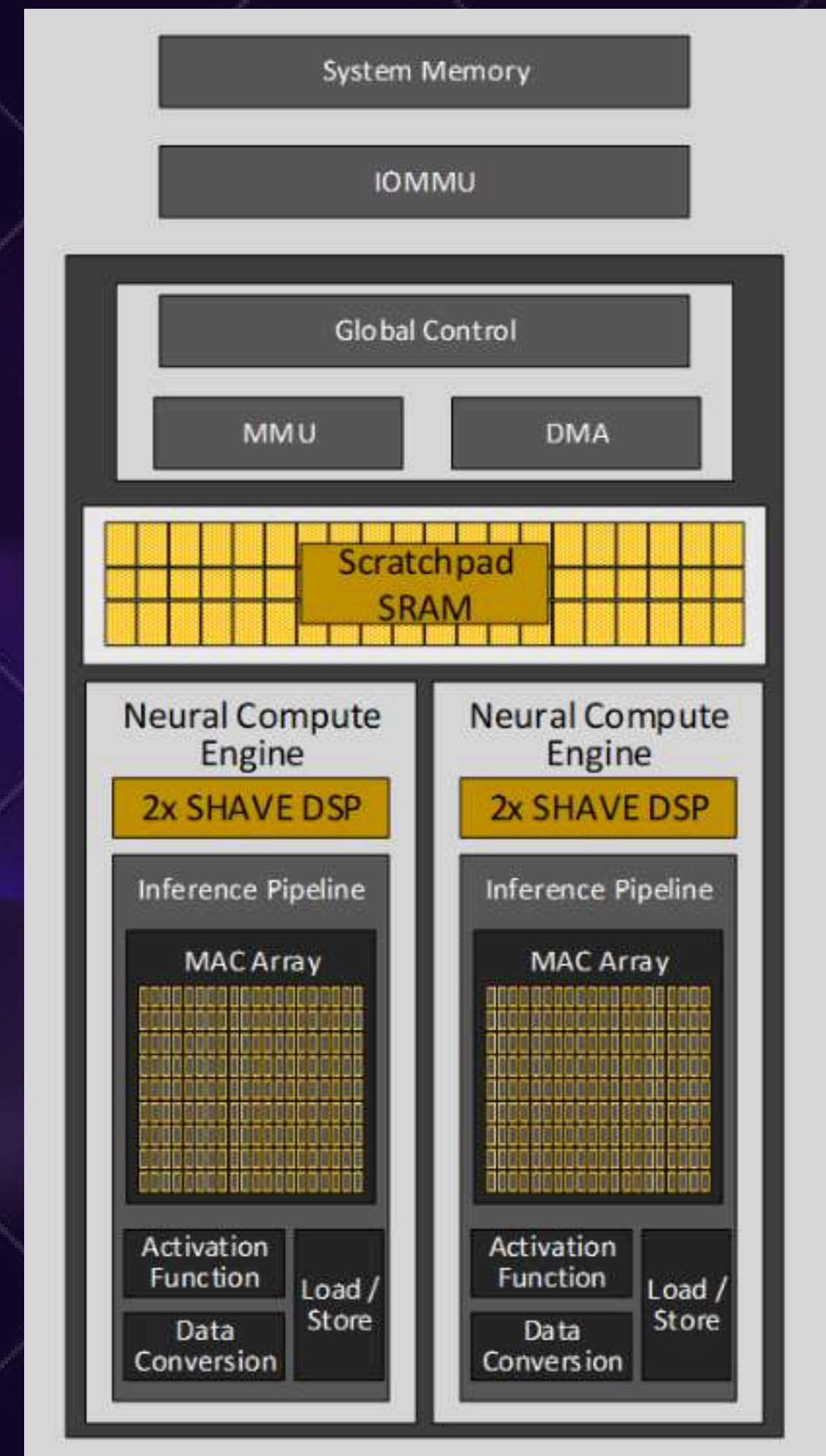
- Memoria interna rápida (SRAM).
- Guarda temporalmente pesos y datos que están por procesarse.

## Neural Compute Engine

- Núcleo que hace el cómputo pesado:
- MAC Array: multiplica y suma.
- 2x SHAVE DSP: procesadores para operaciones extra.
- Inference Pipeline: controla el flujo de datos.

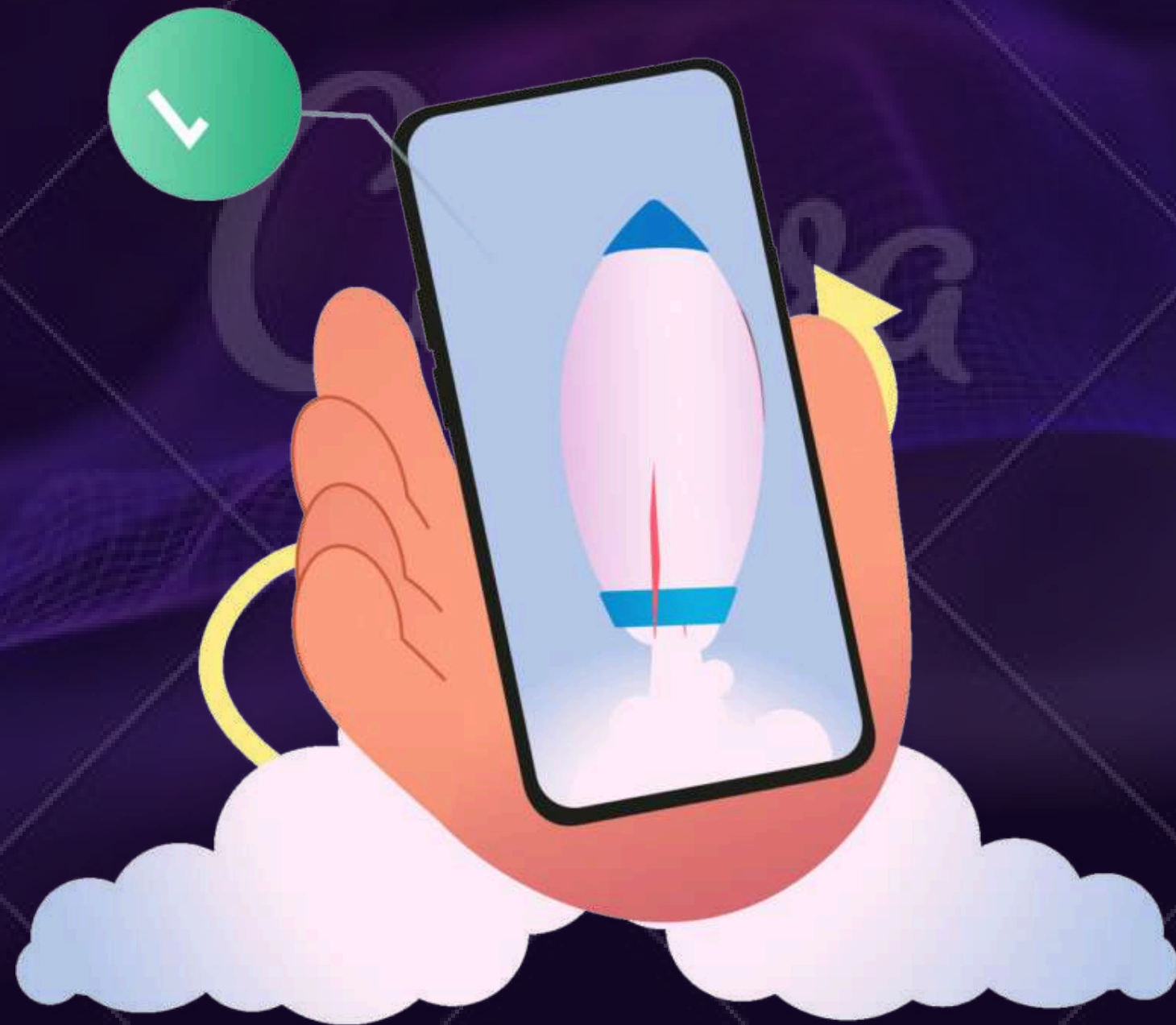
## Activation Functions.

## Data Conversion



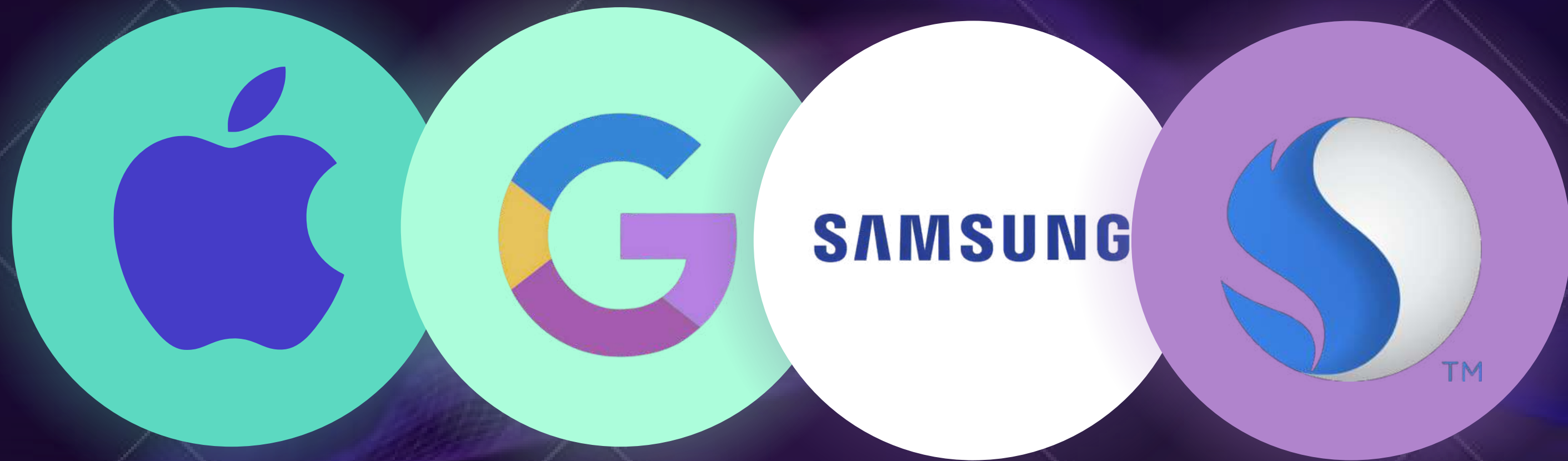
# IA AL ALCANCE DE LA MANO

Casos de IA local en Hardware Comercial





# LOS **PEQUEÑOS** GRANDES REFERENTES





 iPhone X — Animoji Yourself — Apple  Share



Watch on  YouTube

A video player interface with a white background. On the left is a portrait of a woman with blonde hair and a pink top. On the right is a 3D unicorn Animoji with a white face, pink mane, and a rainbow-colored horn. A red YouTube play button is centered between the two images. The top of the player has a header with a user avatar, the text 'iPhone X — Animoji Yourself — Apple', and a 'Share' button with a share icon. The bottom left corner features a dark bar with the text 'Watch on' and a YouTube logo.



# APPLE NEURAL ENGINE (APE)

El primer Neural Engine se presentó en septiembre de 2017 como parte del chip Apple A11 "Bionic" diseñado para el iPhone X

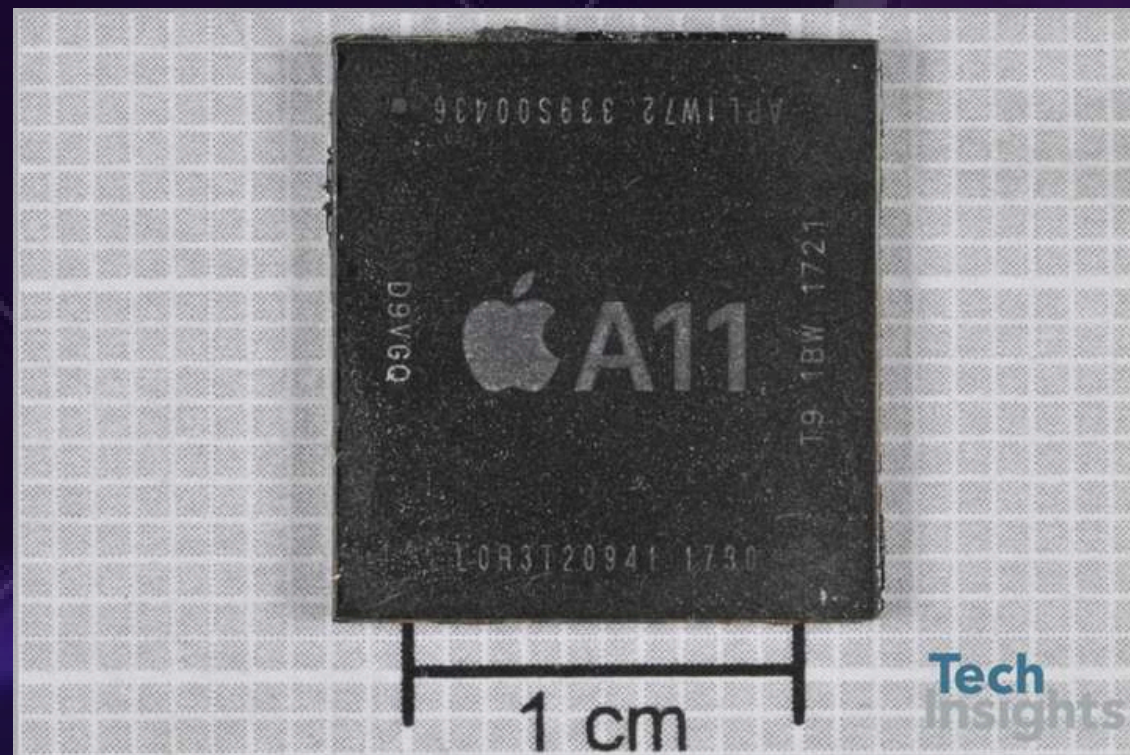


Imagen tomada de: <https://wccftech.com/apple-iphone-8-plus-a11-bionic-complete-teardown/>.

Consistía en dos núcleos capaces de realizar hasta 600 mil millones de operaciones por segundo para el procesamiento en tiempo real de algoritmos de aprendizaje automático dedicados a funciones como Animoji y Face ID.

Su rendimiento máximo era de 0,6 teraflops (TFlops) en formato de datos de punto flotante de precisión media (FP16).





# APPLE NEURAL ENGINE (APE)

¿Pero porque incluirlos?

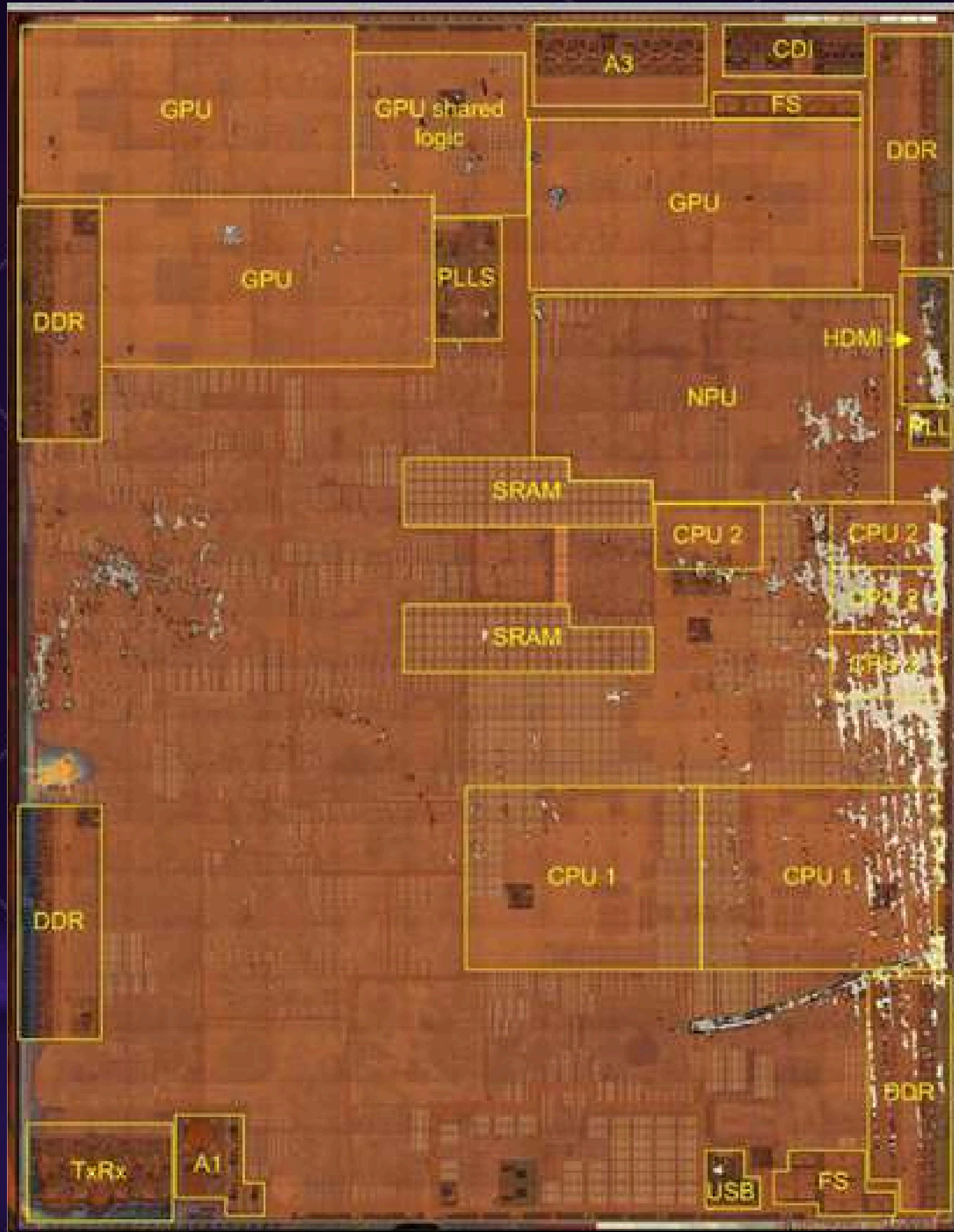
Para ellos crear una arquitectura de GPU completamente nueva no era lo “suficientemente innovador”

Entonces el Neural Engine se incorpora para dentro del Procesador de Señal de Imagen para resolver problemas muy específicos como la coincidencia, el análisis y el cálculo de miles de puntos de referencia de imagen que fluyen rápidamente desde el sensor de la cámara.



Imagen tomada de: <https://appleinsider.com/articles/17/09/23/inside-iphone-8-apples-a11-bionic-introduces-5-new-custom-silicon-engines>

# COMPARATIVA A11 VS A10



Imágenes tomadas de : <https://wccfttech.com/apple-iphone-8-plus-a11-bionic-complete-teardown/>



SoC	Introduced	Process	Neural cores	Peak ops/sec.	Note
Apple A11	Sep. 2017	10 nm	2	600 billion	First implementation.
Apple A12	Sep. 2018	7 nm	8	5 trillion	90% lower power consumption.
Apple A13	Sep. 2019	7 nm*	8	6 trillion	15% lower power. Enhanced 7 nm (N7P).
Apple A14	Oct. 2020	5 nm	16	11 trillion	First 5 nm process (N5).
Apple M1	Nov. 2020	5 nm	16	11 trillion	Comparable to Apple A14
Apple A15	Sep. 2021	5 nm*	16	15.8 trillion	Enhanced 5 nm process (N5P).
Apple M1 Pro Apple M1 Max	Oct. 2021	5 nm	16	11 trillion	Same as original Apple M1.
Apple M1 Ultra	Mar. 2022	5 nm	32	22 trillion	2x Apple M1 Max.
Apple M2	Jun. 2022	5 nm*	16	15.8 trillion	Comparable to Apple A15.
Apple A16	Sep. 2022	4 nm	16	17 trillion	Improved power efficiency.
Apple M2 Pro Apple M2 Max	Jan. 2023	5 nm*	16	15.8 trillion	Same as original Apple M2.
Apple M2 Ultra	Jun. 2023	5 nm*	32	31.6 trillion	2x Apple M2 Max.
Apple A17 Pro	Sep. 2023	3 nm	16	35 trillion	First 3 nm process (N3B).

# Procesadores especializados para Smartwatch



Apple S9	Sep. 2023	4 nm	4		Derived from Apple A16.
Apple M3	Oct. 2023	3 nm	16	18 trillion	Derived from Apple A17 Pro.
Apple M3 Pro Apple M3 Max	Nov. 2023	3 nm	16	18 trillion	Same as original Apple M3.
Apple M4	May 2024	3 nm*	16	38 trillion	Enhanced 3 nm process (N3E).
Apple A18	Sep. 2024	3 nm*	16	35 trillion	Optimized for Apple Intelligence.
Apple A18 Pro					Apple Intelligence 15% faster than A17 Pro
Apple S10	Sep. 2024	4 nm	4		Comparable to Apple S9
Apple M4 Pro Apple M4 Max	Nov. 2024	3 nm*	16	38 trillion	Same as original Apple M4.
Apple M3 Ultra	Mar. 2025	3 nm	32	36 trillion	2x Apple M3 Max.





# APPLE NEURAL ENGINE (APE) ¿Y todo este poder de computo que permite hacer?

01



## PROCESAMIENTO AVANZADO DE FOTOGRAFÍA COMPUTACIONAL

- Smart HDR, Deep Fusion, Photonic Engine: Optimizar detalles y reducir ruido.
- Desenfoque de retrato más preciso, segmentación en tiempo real de escenas.

02



## VIDEO INTELIGENTE

- Cinematic Mode (enfoco dinámico asistido por IA).
- Estabilización avanzada, detección de sujetos, grabación ProRes con ajustes en tiempo real.

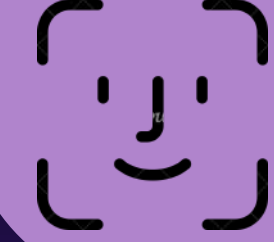
03



## RECONOCIMIENTO DE VOZ Y LENGUAJE NATURAL

- Dictado offline (desde A12).
- Siri más rápido, que puede procesar comandos comunes directamente en el dispositivo, mejorando privacidad y latencia.

04



## SEGURIDAD Y BIOMETRÍA

- Face ID más rápido y preciso gracias a redes neuronales optimizadas.
- Detección de rostros incluso con mascarilla (IA entrenada con datos locales, desde iOS 15).

05



## AR (REALIDAD AUMENTADA)

- Segmentación en tiempo real del usuario respecto al fondo.
- Estimación de profundidad y mapeo del entorno más natural.

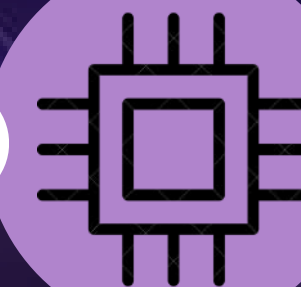
06



## TRADUCCIÓN LOCAL Y ACCESIBILIDAD

- App de traducción funcionando completamente sin conexión.
- Live Captions: subtítulos en tiempo real usando redes de reconocimiento de voz locales.

07



## LOS CHIPS MÁS RECIENTES SOPORTAN:

- Inferencia en INT8, FP16 y ahora BF16 (más precisión con menos coste energético).
- Redes Transformer y modelos de lenguaje (para predicción de texto, autocorrección más inteligente)



# PIXEL VISUAL CORE (2017)

- Acelerador de imagen (ISP avanzado) diseñado junto a Intel.
- Capaz de ejecutar modelos de visión por computadora directamente en el dispositivo:
  - HDR+ mejorado: combina varias fotos para lograr imágenes más nítidas y con mejor rango dinámico.
  - Reducción de ruido en fotos nocturnas.
- Permitía que apps de terceros (vía Android Camera API) usaran HDR+ de Google.
- Su activación de dio con el lanzamiento de Android 8.1



**Comparativa fotografía en aplicaciones de terceros vs fotografía con la cámara nativa y HDR+**

Imagen tomada de: <https://www.xataka.com/moviles/pixel-visual-core-el-primer-chip-propio-de-google-es-la-clave-de-la-camara-de-los-pixel-2>





# PIXEL VISUAL CORE (2017)

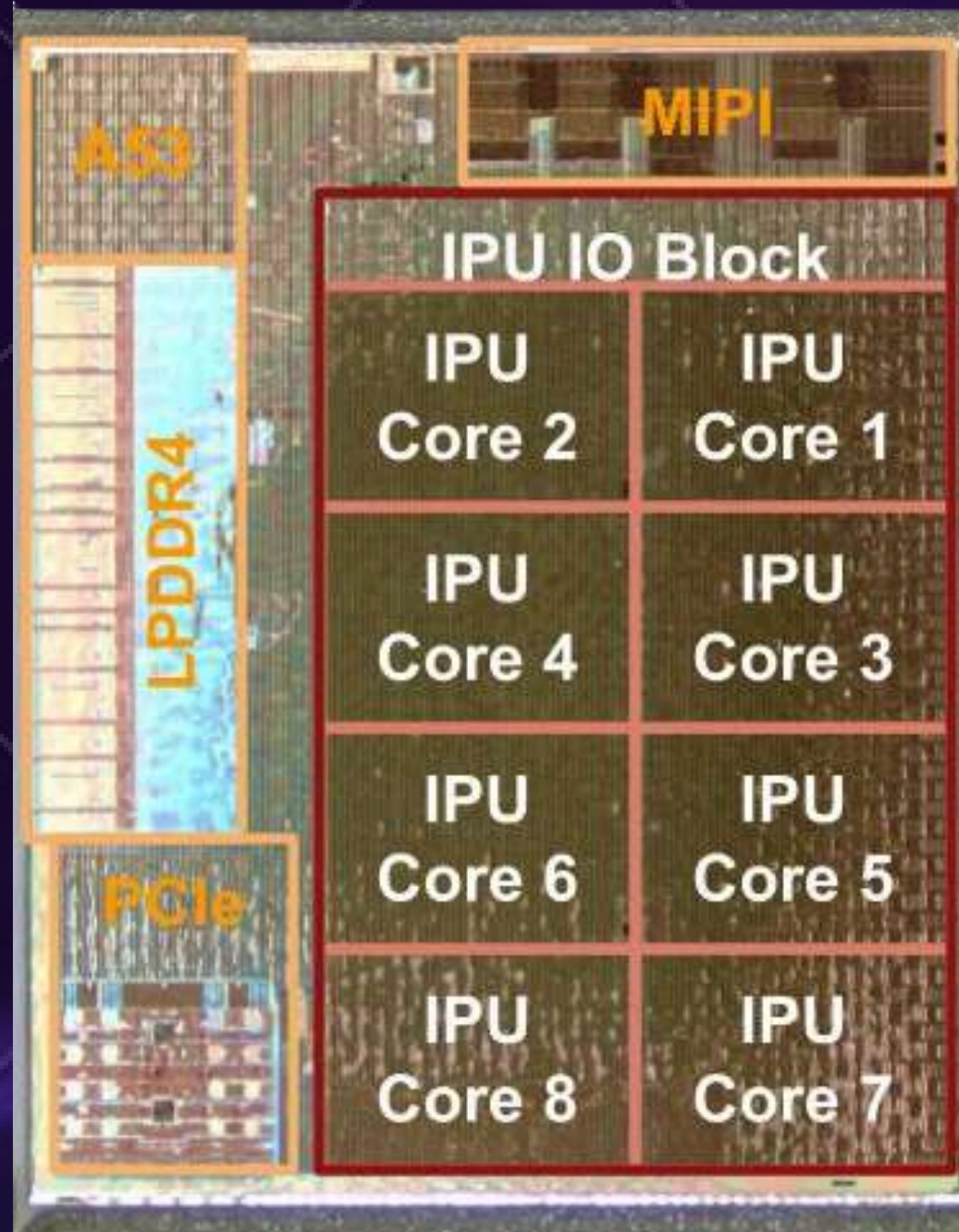


Imagen tomada de: <https://blog.google/products/pixel/pixel-visual-core-image-processing-and-machine-learning-pixel-2/>

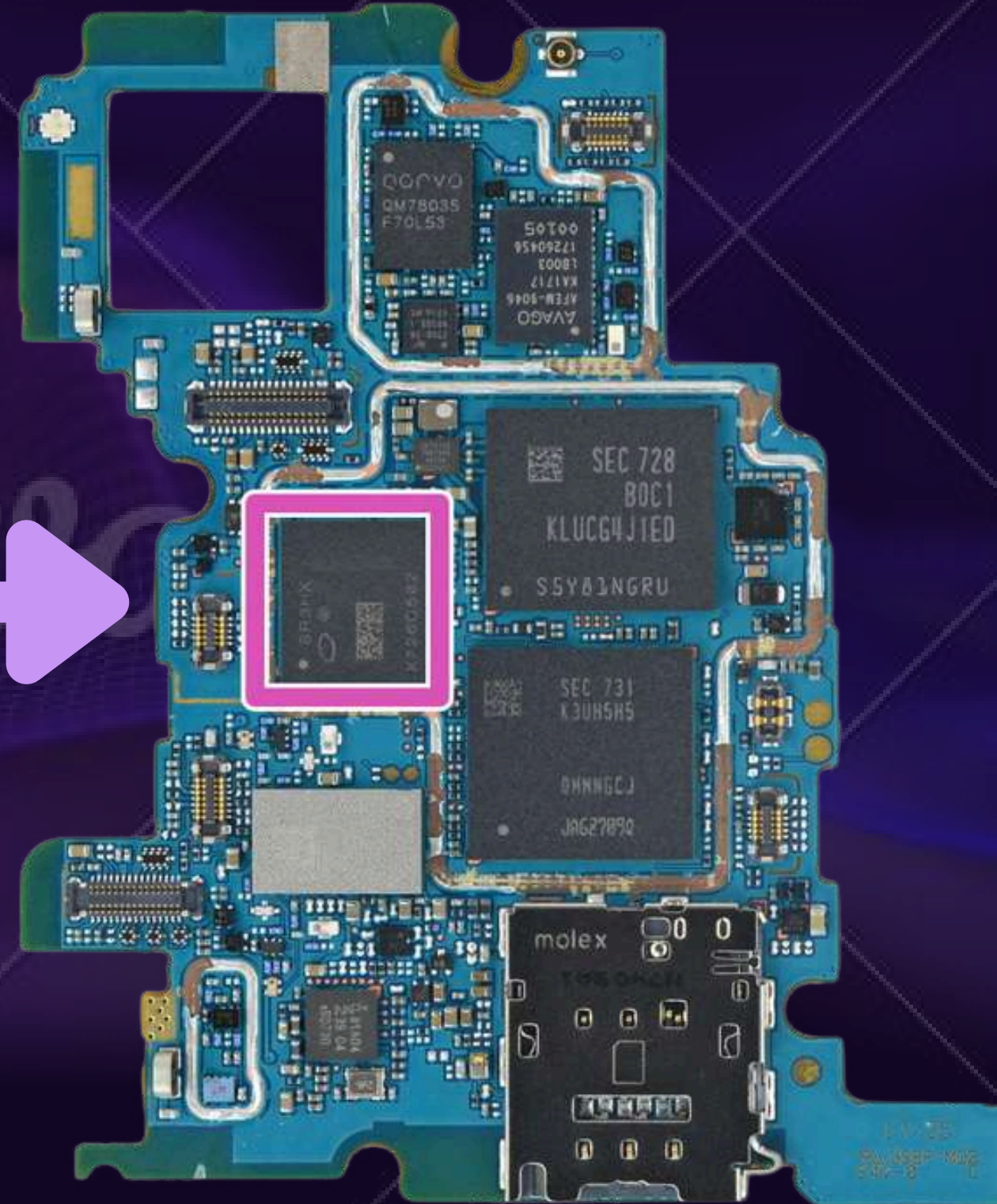
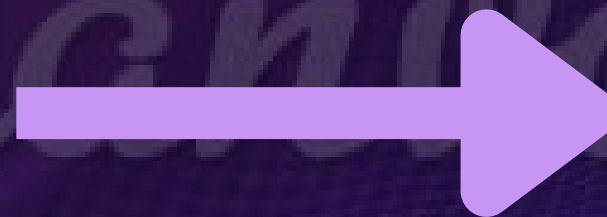


Imagen tomada de: [https://mobiltelefon.ru/post\\_1508939584.html](https://mobiltelefon.ru/post_1508939584.html)





# PIXEL NEURAL CORE (2019)

## Fotografía computacional en tiempo real:

- Live HDR+ preview: ver cómo quedará la foto antes de tomarla.
- Astrophotography mode: fotos de estrellas y cielo nocturno combinando múltiples capturas con IA.
- Frequent Faces: prioriza las caras que más fotografías para que salgan mejor enfocadas.

## IA para voz y lenguaje localmente:

- Google Assistant offline: responde rápido sin depender de la nube.
- Recorder app: transcribe voz a texto sin conexión.

## Desbloqueo facial 3D:

- Reconocimiento facial seguro, rápido y privado gracias a IA local.

**El Pixel 4 utilizaba el procesador Qualcomm Snapdragon 855**

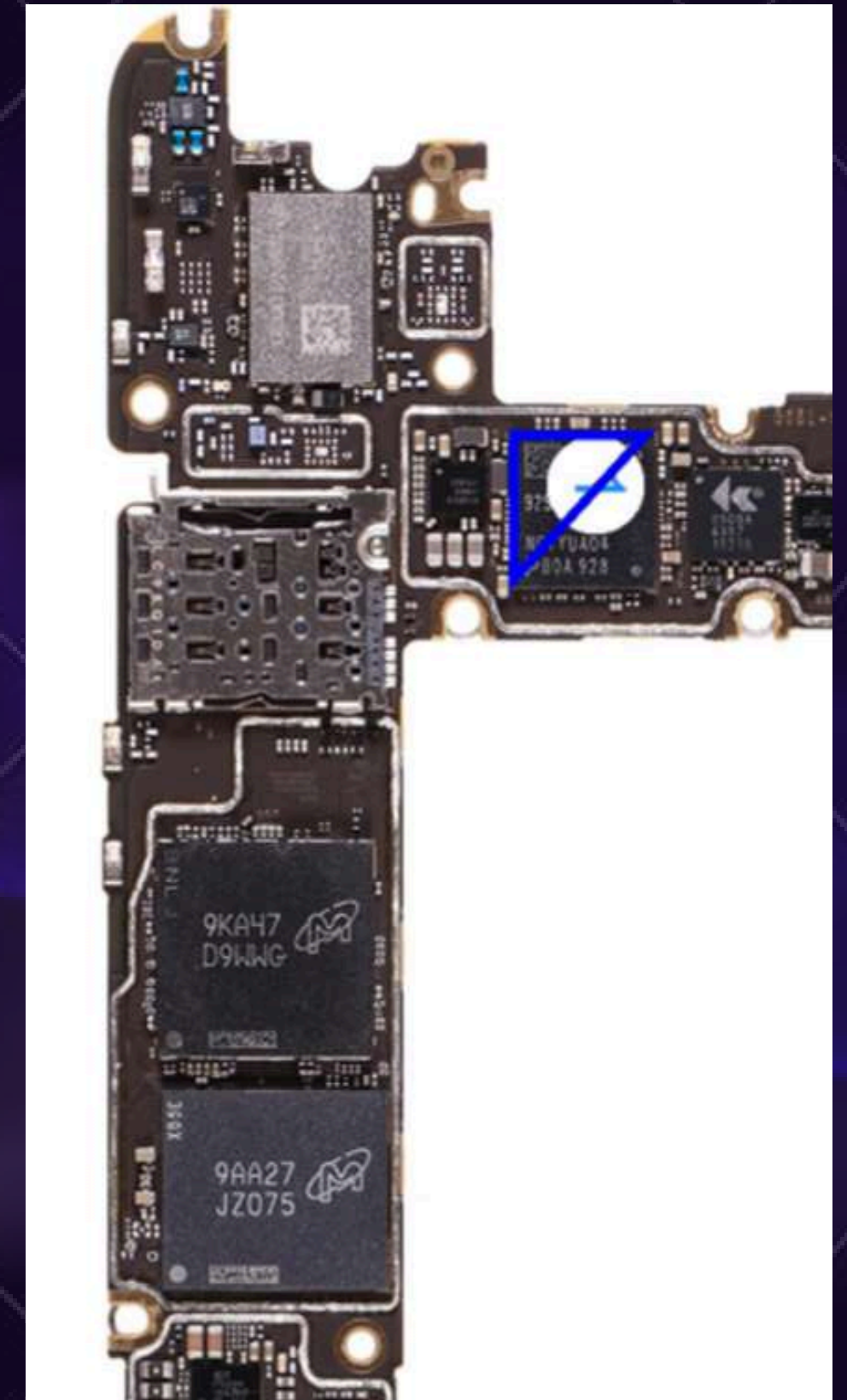


Imagen tomada de: <https://omdia.tech.informa.com/om006145/googles-custom-mobile-soc-ambitions-clears-up-mystery-behind-npu-chipsets-in-existing-pixels>



# GOOGLE TENSOR SoC (G1)

- **Primer SoC diseñado propiamente por Google**
- **TPU (Tensor Processing Unit)**
  - Acelerador especializado en inteligencia artificial.
  - Permite ejecutar modelos de visión, voz y lenguaje en el dispositivo, como:
    - Traducción en vivo sin conexión.
    - Mejoras en fotos y video (ej. Magic Eraser, Night Sight).
    - Dictado de voz más rápido y privado.
- **Tensor Security Core (Titan M2)**
- Núcleo dedicado a la seguridad:
  - Protege datos sensibles (como desbloqueo facial, contraseñas, llaves de cifrado).
  - Resistente a ataques físicos y malware.
- Trabaja junto al sistema operativo para mantener la integridad del dispositivo.
- **No se especifica la precisión de funcionamiento.**

Imagen tomada de: <https://mianjuger.com/google-pixel-6-comes-with-tensorcore/>



Imagen tomada de: <https://www.androidcentral.com/how-does-google-titan-m2-and-tensor-security-core-work>









# GOOGLE TENSOR (G4) & GEMINI NANO

Es el modelo más pequeño y eficiente de la nueva familia Gemini de Google (junto a Gemini Pro y Gemini Ultra).

## Ventajas clave

- Privacidad: los datos no salen del dispositivo.
- Disponibilidad: funciona incluso sin conexión a Internet.
- Rapidez: menor latencia porque el procesamiento es local.

## Existen dos versiones:

- Nano-1: 1.8B parámetros.
- Nano-2: 3.25B parámetros.
- Los modelos están cuantizados a 4 bits, lo que reduce consumo de memoria y permite ejecutarse en móviles.

**Google - Gemini Nano Preview**



## Gemini Nano

Imagenes tomadas de: <https://dig.watch/updates/google-opens-gemini-nano-ai-to-android-developers>

SAMSUNG

# NPU EN EXYNOS

## AI HW Evolution in Exynos



EXYNOS  
9820

EXYNOS  
990

EXYNOS  
2100

EXYNOS  
2200

En 2019, Samsung reveló el primer diseño de NPU en Exynos, que mejoró la eficiencia sobre las CPU y las GPU.

Samsung ha trabajado para mejorar tanto el área como la eficiencia energética.

En el 2022 ya incluía:

- Scatter: distribuir datos de una fuente hacia múltiples destinos no contiguos en memoria.
- Gather: recolectar datos dispersos en memoria y reunirlos en un único destino.
- Modo extremo de baja potencia permite que la NPU funcione sin DRAM para admitir escenarios siempre activos.
- Multi-precision ALU que proporciona el soporte FP16, INT8 e INT4, para lograr una mayor eficiencia y flexibilidad.

Imagen tomada de: [https://semiconductor.samsung.com/news-events/tech-blog/hyper-intelligence-ai-and-future-experiences/?utm\\_source=chatgpt.com](https://semiconductor.samsung.com/news-events/tech-blog/hyper-intelligence-ai-and-future-experiences/?utm_source=chatgpt.com)





- [illegible]

Imagem tomada de: <https://www.notebookcheck.net/Exynos-2400-2200-2100-die-shots-highlight-the-chips-evolution-over-the-years.832760.0.html>

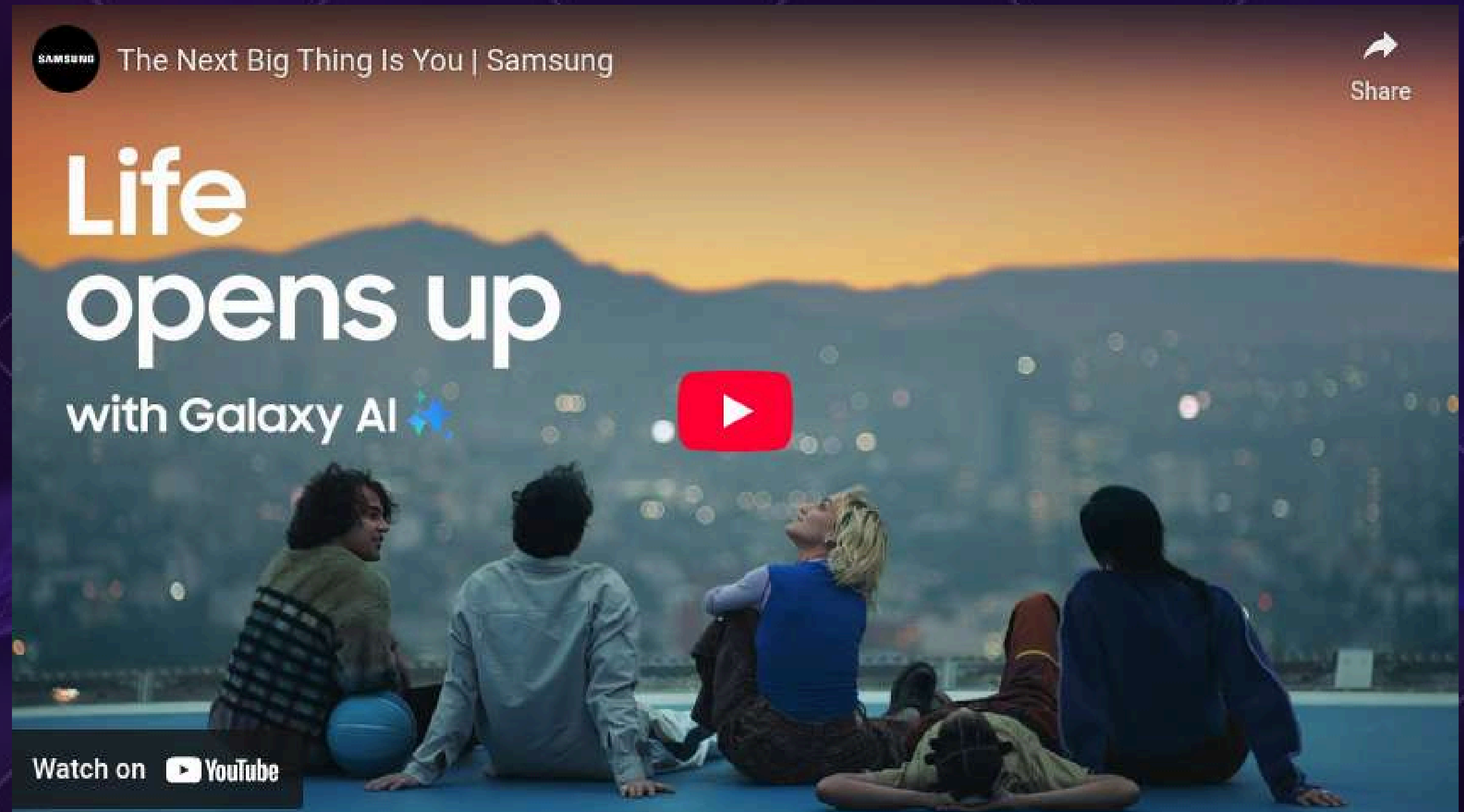


**SAMSUNG**

**GALAXY AI**

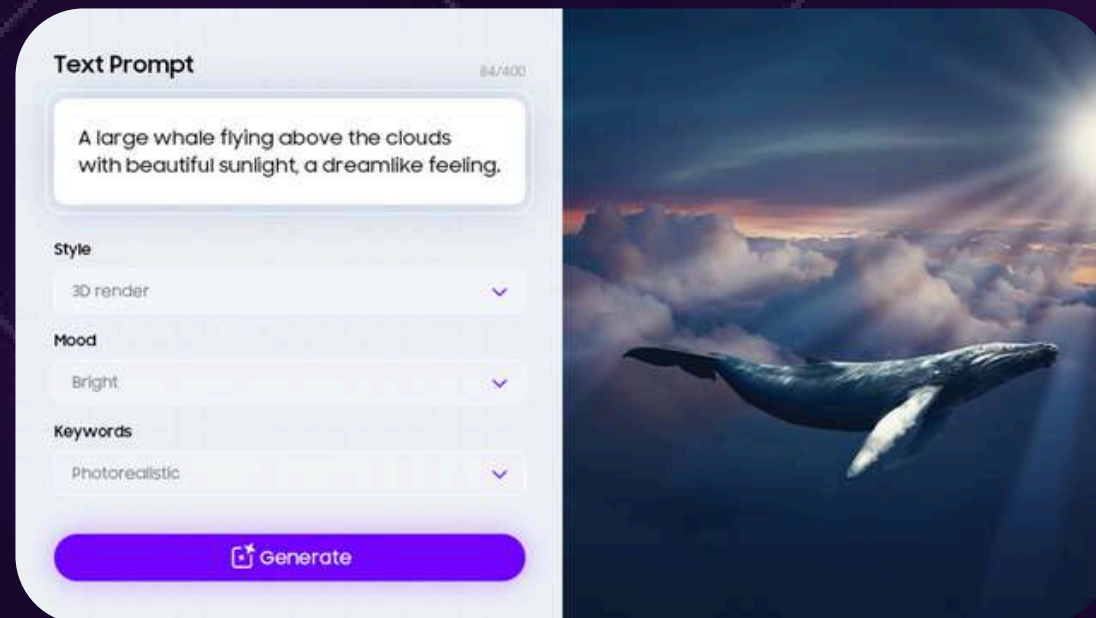
Imágenes tomadas de: <https://semiconductor.samsung.com/technologies/processor/on-device-ai/>

Galaxy AI busca convertirse en el compañero inteligente del usuario, actuando en múltiples contextos, para anticiparse a las necesidades de sus usuarios, mejorar productividad y facilitar interacciones diarias.



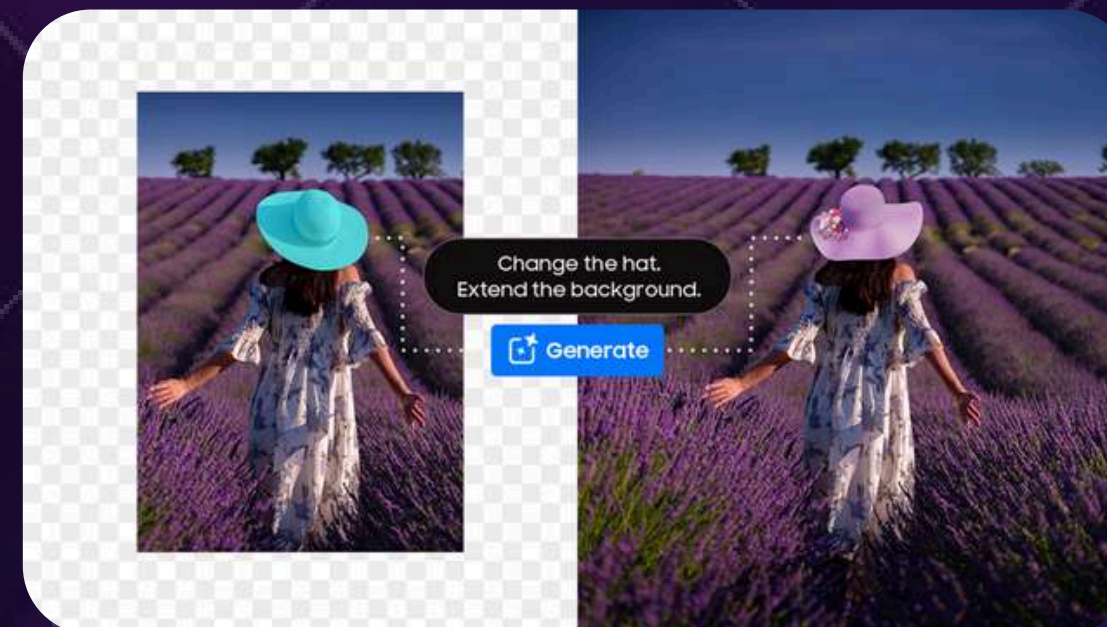


# GALAXY AI



## Generación de imágenes a partir de texto:

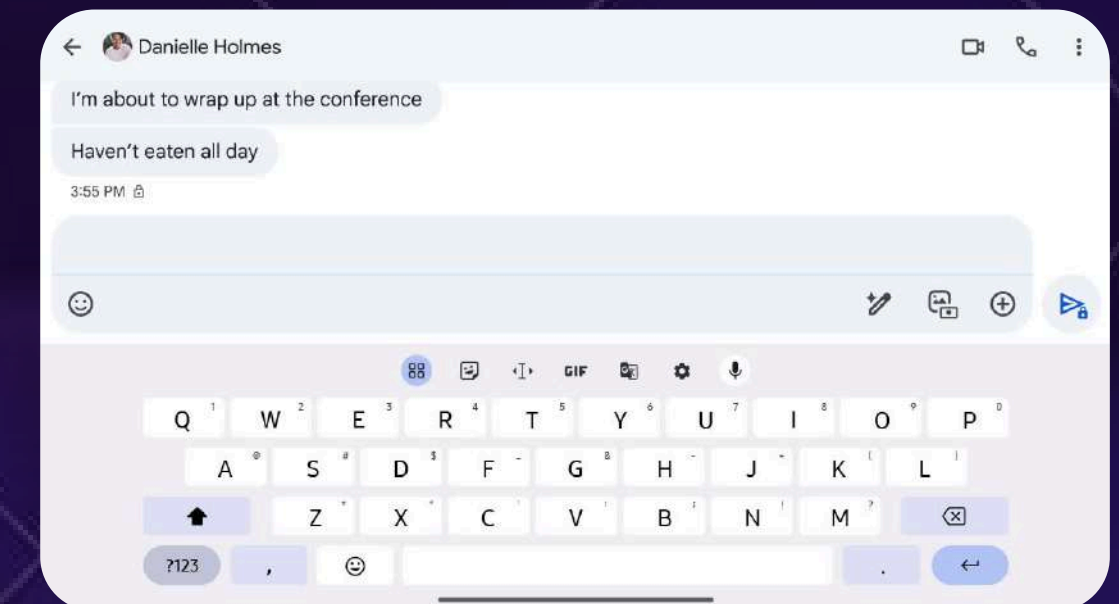
Crea imágenes desde una descripción escrita, directamente en tu dispositivo y en tiempo real. Solo escribes el texto describiendo la imagen que quieres, y el sistema la generará rápidamente.



## In-painting y Out-painting:

- In-painting: cambia o reemplaza elementos dentro de una imagen (por ejemplo, cambiar un sombrero).
- Out-painting: extiende una imagen hacia zonas fuera del marco original.

El resultado luce como si la foto siempre hubiera sido así.



## Corrección de tono y gramática:

Corrige textos y ajusta el tono de tus mensajes fácilmente, directamente desde tu teléfono. Funciones como el asistente de redacción permiten cambiar entre tono profesional, casual o creativo, según lo necesites.



# SNAPDRAGON 845 (2017)

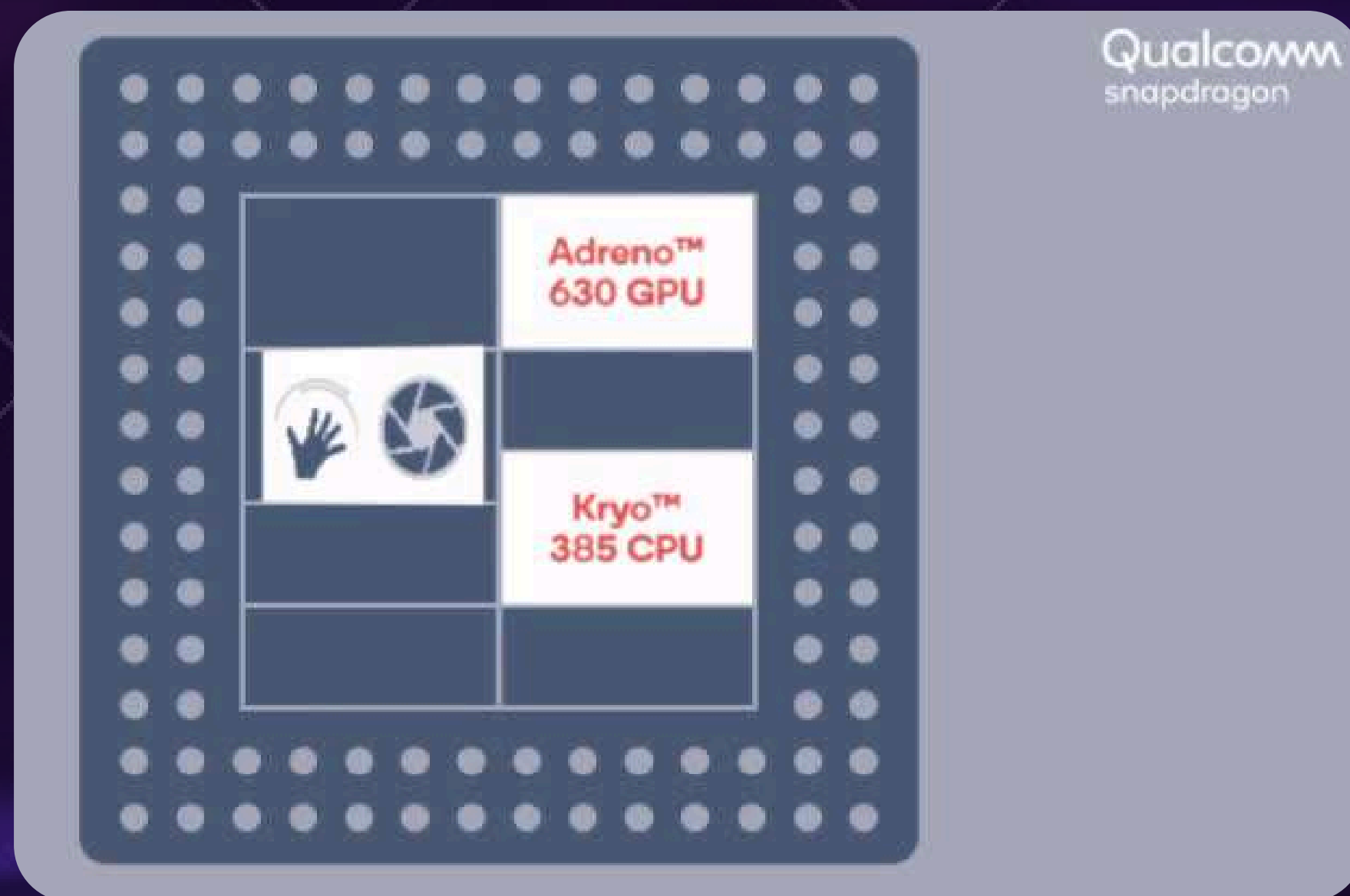


Imagen tomada de: <https://www.qualcomm.com/products/technology/processors/hexagon>

Primer procesador donde Qualcomm introduce oficialmente su "Hexagon 685 DSP" como acelerador para tareas de IA.

Qualcomm no usó desde el principio una "NPU" como Google o Apple, sino un enfoque basado en DSPs optimizados.

DSP (Digital Signal Processor) es un tipo especializado de procesador diseñado específicamente para procesar señales digitales de forma rápida y eficiente.

IA utilizada principalmente para:

- Mejora de imágenes (desenfoque, detección de escenas).
- Reconocimiento facial básico.
- Eficiencia energética mediante inferencias locales.

Qualcomm no reportó TOPS oficiales en esa época, pero diferentes fuentes estiman alrededor de 3 TOPS combinados entre DSP, CPU y GPU.





# SNAPDRAGON X ELITE

Integra una NPU Hexagon dedicada con 45 TOPS de rendimiento IA teórico

Esto lo posiciona como la NPU más potente en un chip para laptops, superando en rendimiento a Intel Core Ultra "Meteor Lake" (10 TOPS) y AMD Ryzen 8040 "Hawk Point" (16 TOPS)

Soporta modelos de hasta 13B de parámetros.

Arquitectura heterogénea. Hexagon NPU:

- Scalar Unit: gestiona operaciones simples y secuenciales.
- Vector Unit: acelera cálculos en paralelo, ideal para procesamiento de imágenes y audio.
- Tensor Accelerator: optimizado para operaciones matriciales y convolucionales, clave en redes neuronales profundas y modelos generativos.

Soporta INT4, que permite hasta 90% más rendimiento y 60% mejor eficiencia energética frente a INT8, sin sacrificar desempeño.







# SNAPDRAGON X ELITE

- **ResNet-50:** Red neuronal convolucional muy usada para clasificación de imágenes.
- **DeeplabV3:** Modelo de segmentación de imágenes. Delimita áreas específicas dentro de una imagen.
- **MobileNetV3:** Modelo optimizado para dispositivos móviles. Clasifica imágenes usando redes ligeras y eficientes.
- **InceptionV4:** Modelo más complejo para reconocimiento avanzado de imágenes, con estructuras profundas.
- **YoloV3:** Modelo especializado en detección de objetos en tiempo real.
- **ESRGAN:** Modelo de super resolución de imágenes (mejora calidad y detalle de imágenes pequeñas o borrosas).

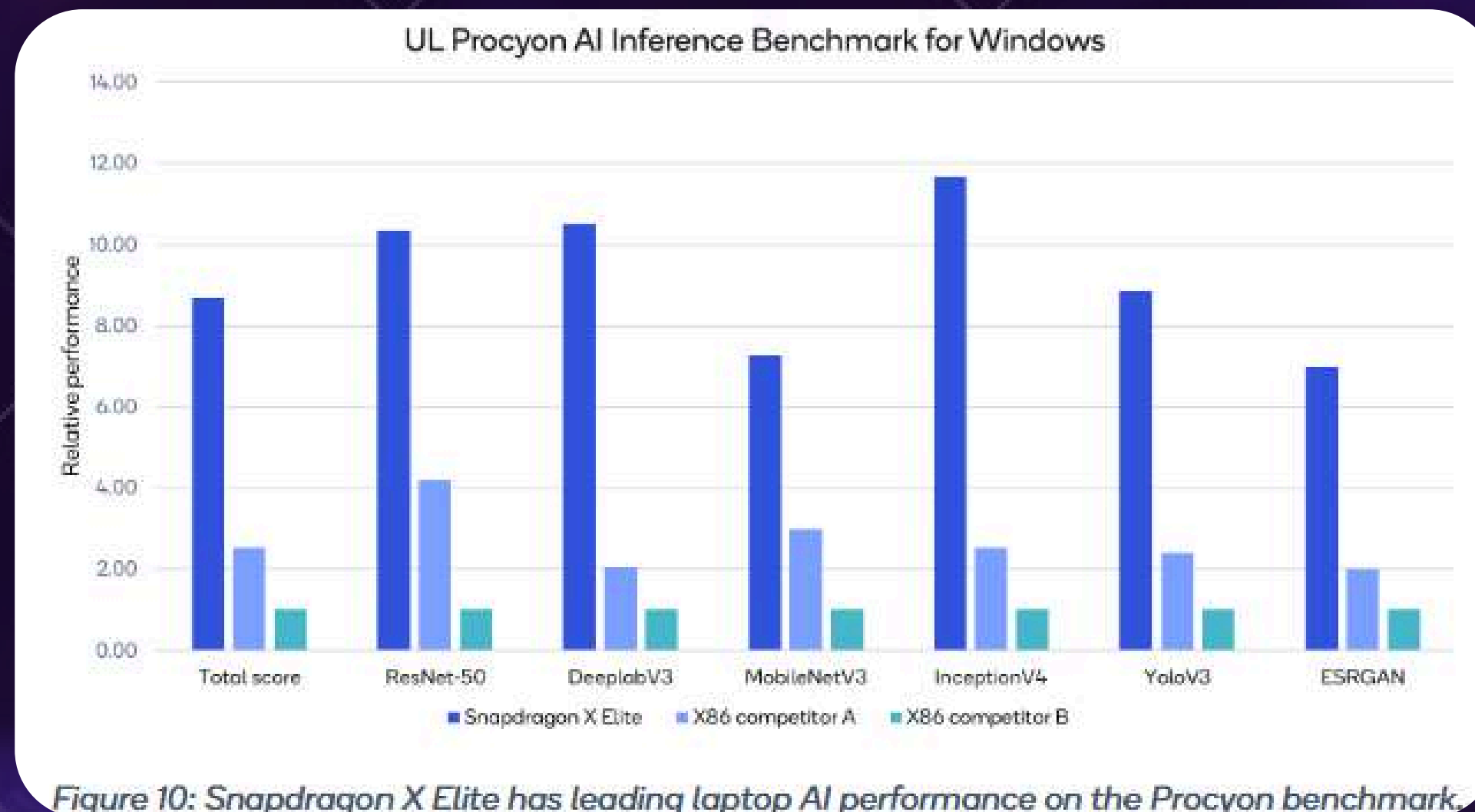
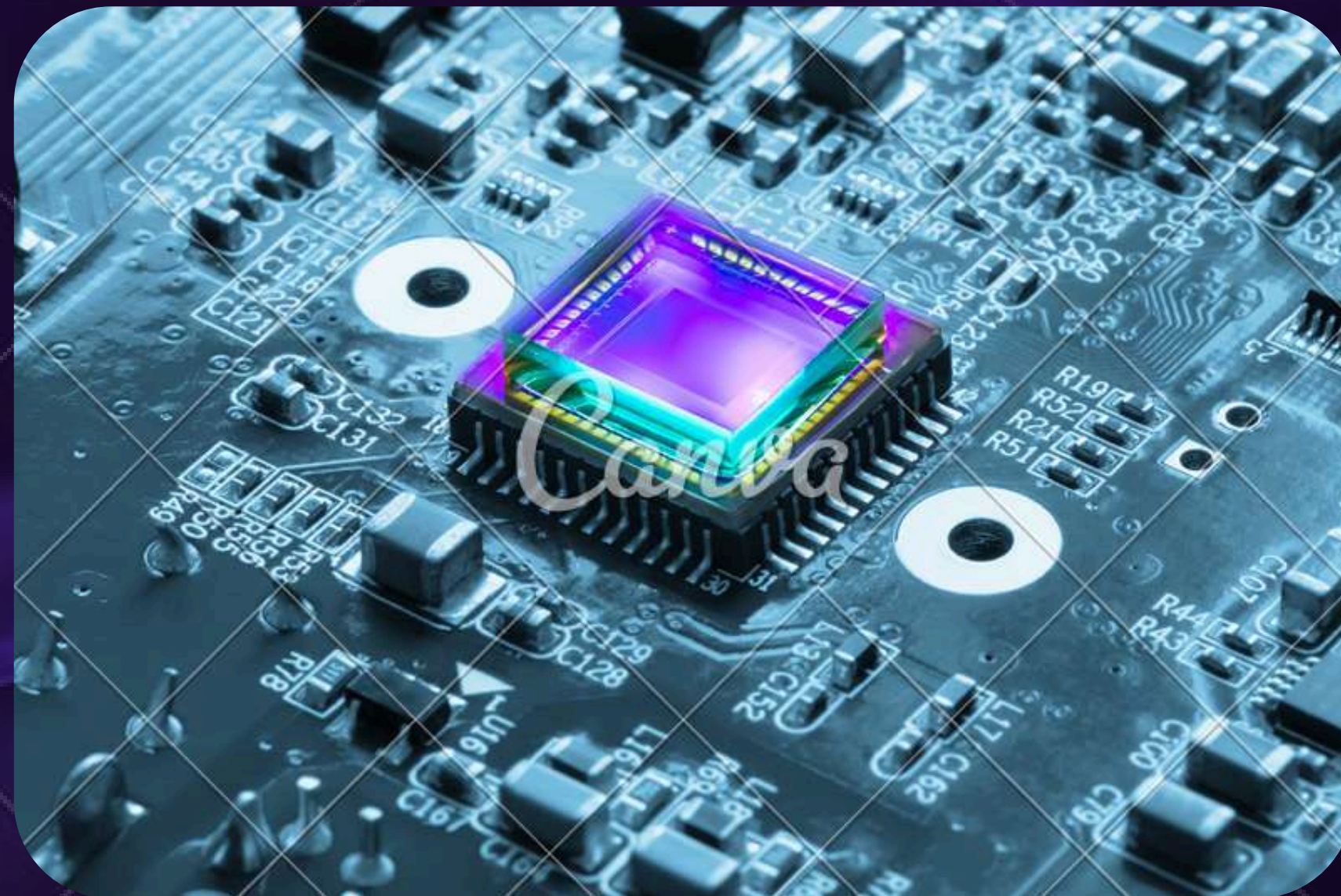


Imagen tomada de: <https://www.qualcomm.com/news/onq/2024/04/a-guide-to-ai-tops-and-npu-performance-metrics?>



# TINYML

Machine learning ejecutado en hardware mínimo





# ¿CÓMO CONSTRUIMOS TINY ML?

“The usual definition is running machine learning on embedded devices at an average of less than one milliwatt in power.” – Pete Warden, technical lead of the TensorFlow at Google.

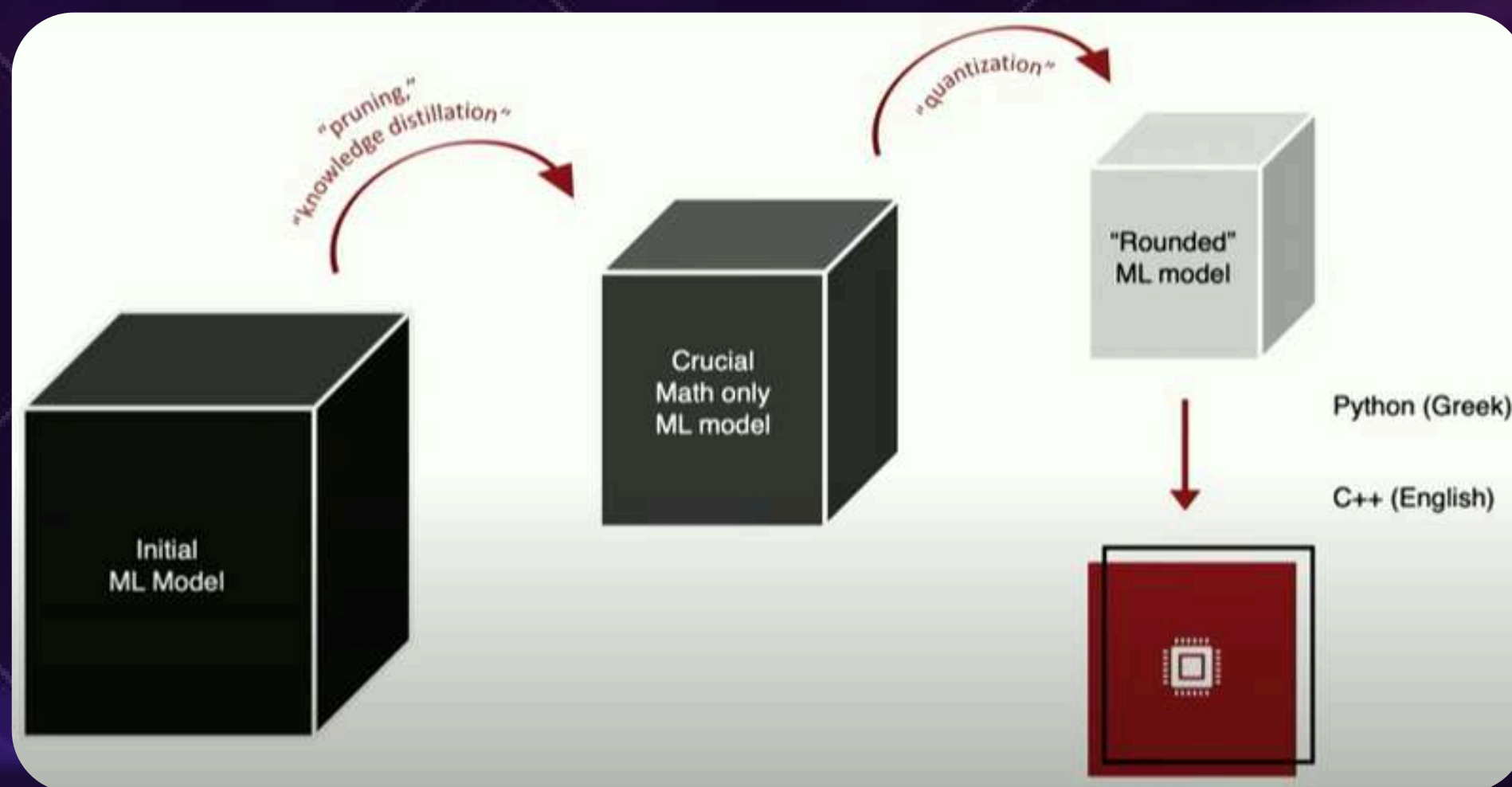


Imagen tomada de: <https://www.youtube.com/watch?v=rfFg1gLLaAo>

## Microcontroladores como procesadores de IA:

Gracias a TinyML, microcontroladores sencillos como los Cortex-M, que antes solo ejecutaban tareas básicas, ahora pueden realizar inferencias IA (reconocer sonidos, detectar movimientos o procesar imágenes simples).

## Modelos pequeños pero funcionales:

Se usan modelos optimizados, como MobileNet o detectores de palabras clave, que requieren apenas entre 20 y 200 kilobytes de memoria. Esto permite que funcionen incluso en dispositivos sin sistema operativo (Bare-metal).

## Recolección de energía como opción real

El consumo ultra bajo permite pensar en sistemas sin batería, usando recolección de energía ambiental (energy harvesting).

## IA distribuida a gran escala:

Al ser tan económicos y eficientes, podemos desplegar miles o millones de estos dispositivos inteligentes en objetos cotidianos: desde electrodomésticos hasta juguetes o dispositivos médicos.

# PUNTOS CLAVE DEL TINY ML

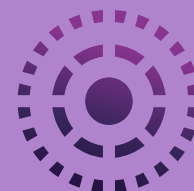
IA extremadamente eficiente, ejecutada en sensores y microcontroladores de muy bajo consumo.



## PRIVACIDAD

En el IoT tradicional, los dispositivos envían datos sin procesar a servidores externos (la nube) para su análisis.

Con TinyML, todo el procesamiento ocurre dentro del dispositivo. Los datos sensibles nunca abandonan el dispositivo, garantizando mayor seguridad y privacidad.



## VELOCIDAD

En sistemas IoT clásicos, los datos deben viajar hasta la nube, ser procesados y luego volver, generando latencia.

En TinyML, el modelo ejecuta directamente en el microcontrolador, junto al sensor. Obtienes respuestas instantáneas y en tiempo real, sin depender de internet ni de servidores.



## CONSUMO Y COSTO

Los dispositivos IoT tradicionales consumen energía constantemente para enviar y recibir datos.

TinyML reduce el costo total de operación al evitar transmisiones de datos continuas y el uso de servidores en la nube.



# GRANDES POSIBILIDADES

## Factores impulsores detrás del Tiny ML



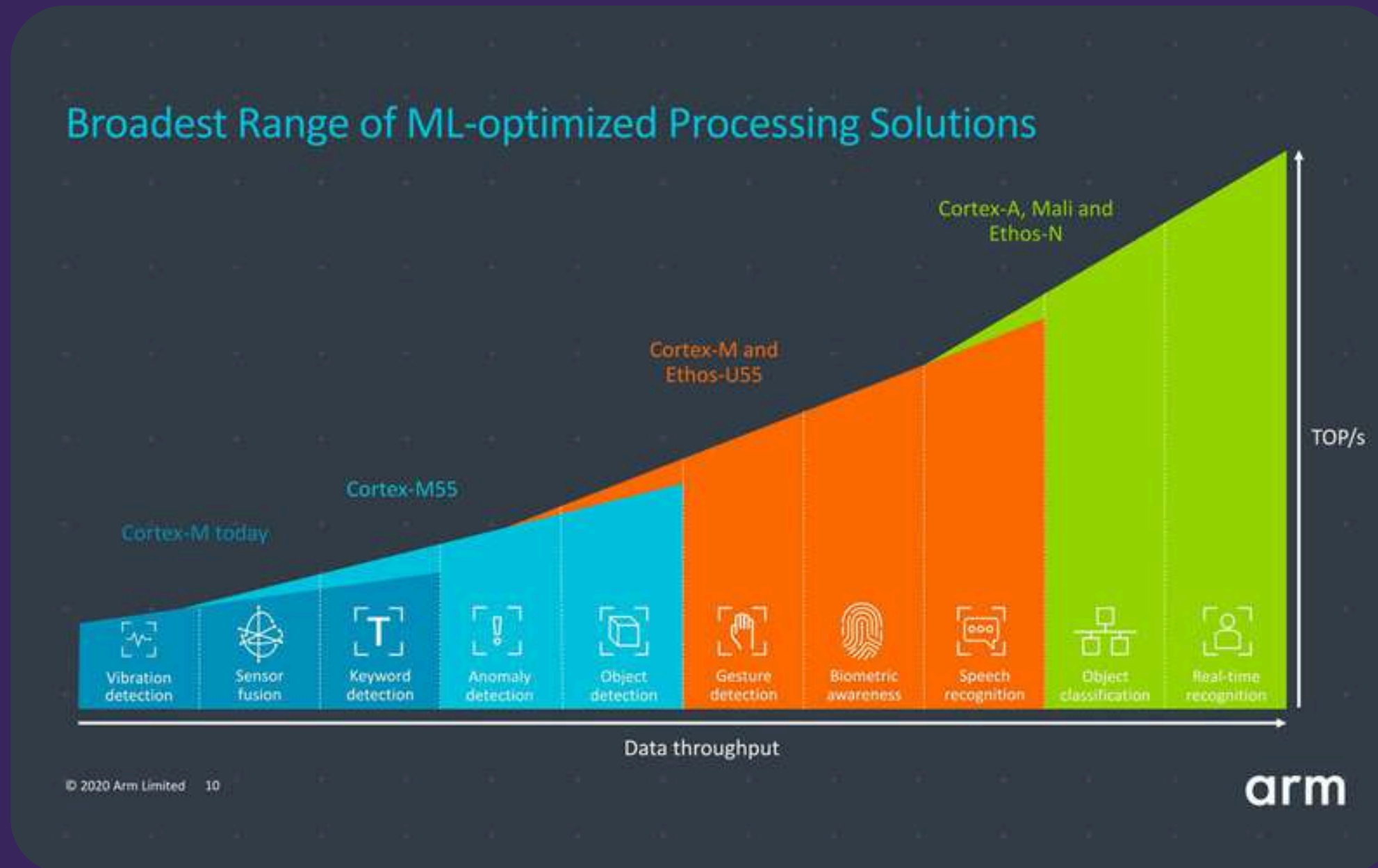
Video e imagen tomados de: <https://www.arm.com/campaigns/arm-tinyml>





# PROCESADORES ARM

## Diferentes procesadores, diferentes aplicaciones



- **Cortex - A:** alto rendimiento, ejecutan sistema operativo completo.
- **Cortex - M:** microcontroladores pequeños y de bajo consumo.
- **Ethos:** NPUs para machine learning y aplicaciones de redes neuronales.
- **Mali:** GPUs para procesamiento gráfico.

Imagen tomada de: <https://news.mynavi.jp/techplus/article/20200212-971390/>



# TENSOR FLOW LITE MICRO

Framework open source de Google

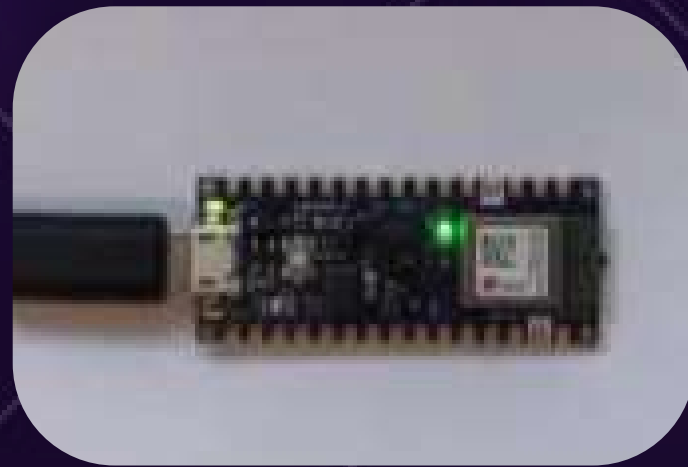


Imágenes tomadas de: <https://www.youtube.com/watch?v=lcX77I9bLJo&t=412s>

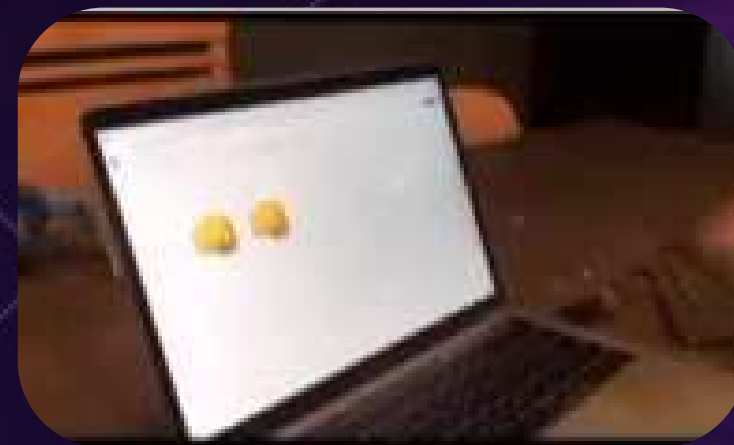
- Permite ejecutar modelos de aprendizaje automático en microcontroladores y dispositivos con recursos muy limitados.
- Optimizado para funcionar sin sistema operativo.
- El entorno de ejecución ocupa solo 16 KB.
- No depende de bibliotecas externas estándar (como las de C o C++). El framework es completamente autónomo, lo que evita problemas de compatibilidad

# EJEMPLOS TINY ML

Recursos disponibles en <https://developer.arm.com/IoT>



**RECONOCIMIENTO DE  
VOZ SIMPLE**



**RECONOCIMIENTO DE  
GESTOS MEDIANTE  
ACELERÓMETRO**



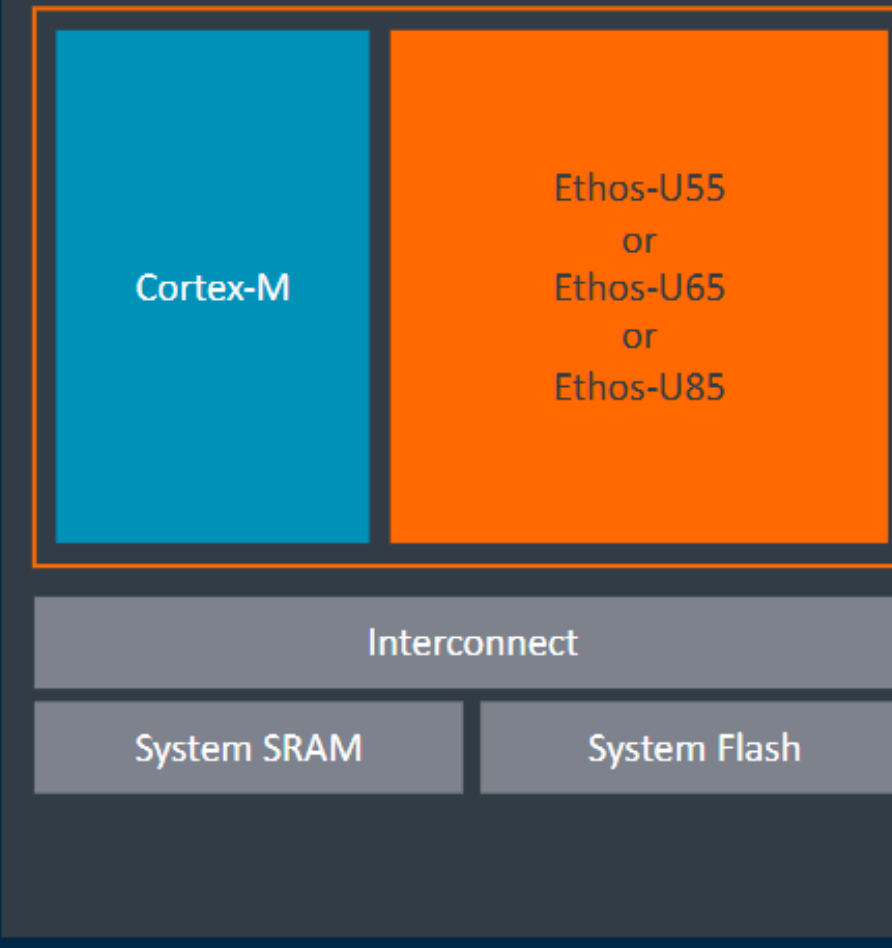
**DETECCIÓN DE  
PERSONAS**



**RECONOCIMIENTO DE  
OBJETOS CON OPEN  
MV**



## Cortex-M Based System



## ¿CUÁL ES SU RENDIMIENTO?

El ARM Ethos-U85 es un acelerador IA diseñado para funcionar junto a microcontroladores Cortex-M, permitiendo ejecutar redes neuronales de manera eficiente en dispositivos de muy bajo consumo.

Ofrece un rendimiento de hasta 4 TOPS (Tera Operations Per Second), según la configuración de hardware:

- Desde 128 MACs hasta 2.048 MACs.
- Mejora de 4x en desempeño IA respecto a generación anterior.

### Increased Performance


Configurations from 128 MACs to 2048 MACs (4 TOPs)

### Higher Power Efficiency

20% more energy efficient than previous generation

### Extended Operator Support

Transformer network support for faster customization



**¡GRACIAS POR  
LA ATENCIÓN!**

# REFERENCIAS:

- <https://www.ibm.com/think/topics/machine-learning> <https://foqum.io/blog/termino/nodo/> }
- <https://telefonicatech.com/blog/deep-learning-para-todos-los-publicos>
- [https://es.wikipedia.org/wiki/Coma\\_flotante](https://es.wikipedia.org/wiki/Coma_flotante) <https://foqum.io/blog/termino/nodo/>
- <https://foqum.io/blog/termino/funcion-de-activacion/>
- <https://foqum.io/blog/termino/peso/#:~:text=%C2%BFQu%C3%A9%20es%20un%20Peso?,optimizar%20el%20rendimiento%20del%20modelo> <https://blog.google/products/pixel/pixel-visual-core-image-processing-and-machine-learning-pixel-2/> [https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Faulasvirtuales.udistrital.edu.co%2Fpluginfile.php%2F819294%2Fmod\\_resource%2Fcontent%2F1%2F04%2520coma%2520flotante.pptx&wdOrigin=BROWSELINK](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Faulasvirtuales.udistrital.edu.co%2Fpluginfile.php%2F819294%2Fmod_resource%2Fcontent%2F1%2F04%2520coma%2520flotante.pptx&wdOrigin=BROWSELINK)
- <https://standards.ieee.org/ieee/754/6210/> [https://apple.fandom.com/wiki/Neural\\_Engine](https://apple.fandom.com/wiki/Neural_Engine)
- [http://m.datasheets.pl/elementy\\_czynne/IC/A11-BIONIC.pdf](http://m.datasheets.pl/elementy_czynne/IC/A11-BIONIC.pdf)
- <https://appleinsider.com/articles/17/09/23/inside-iphone-8-apples-a11-bionic-introduces-5-new-custom-silicon-engines> [https://machinelearning.apple.com/research/neural-engine-transformers?utm\\_source=chatgpt.com](https://machinelearning.apple.com/research/neural-engine-transformers?utm_source=chatgpt.com)



# REFERENCIAS:

- [¿Por qué las GPUs son buenas para la IA? | Data Coffee #12](#)
- [Tiny Machine Learning: Progress and Futures](#)
- <https://arxiv.org/abs/2301.03904>
- [\[1711.10374\] A Transprecision Floating-Point Platform for Ultra-Low Power Computing](#)
- <http://proceedings.mlr.press/v37/gupta15.pdf>
- [What is FP64, FP32, FP16? Defining Floating Point | Exxact Blog](#)
- [¿Qué es la cuantificación? | IBM](#)
- [Data Formats For Inference On The Edge](#)
- [Energy and Area Comparison for Floating-Point and Fixed-Point | Download Scientific Diagram](#)
- [\(130\) How TinyML Gives us Spider-Man Powers | Emelie Eldracher | TEDxMIT - YouTube](#)
- [\(130\) What is LLM quantization? - YouTube](#)
- [\(130\) Introduction to TinyML - Alessandro Grande - YouTube](#)
- [A guide to AI TOPS and NPU performance metrics | Qualcomm](#)
- [\[Tech Day 2022\] Hyper-intelligence: AI and future experiences | Samsung Semiconductor Global](#)
- [Samsung Exynos 2400 SoC fully detailed](#)
- [Samsung Exynos 2400: especificaciones en CPU, GPU e IA](#)
- [Los procesadores Exynos vuelven a la gama alta de Samsung, tras un año de sólo Qualcomm. Así es el Exynos 2400 de los S24 y S24+](#)
- [AI Dominates Qualcomm Snapdragon Summit With New Snapdragon Products](#)
- [Qualcomm details its Snapdragon X Elite: all SKUs have NPU with 45 TOPS for AI workloads](#)
- [Información de la NPU para Apple y Snapdragon : r/LocalLLaMA](#)
- [Qualcomm details its Snapdragon X Elite: all SKUs have NPU with 45 TOPS for AI workloads](#)
- [Inteligencia artificial, machine learning, deep learning | Grupo Bancolombia](#)
- [AI vs. Machine Learning vs. Deep Learning vs. Neural Networks | IBM](#)
- [What Is Artificial Intelligence \(AI\)? | IBM](#)
- <https://www-formal.stanford.edu/jmc/whatisai.pdf>
- [What Is Machine Learning \(ML\)? | IBM](#)
- [What Is Deep Learning? | IBM](#)
- [What is a Neural Network? | IBM](#)