Impacto e Infraestructura tecnológica

CPU GPU ASIC FPGA NPU QC

J. Daniel Britos

UNC

June 16, 2023

Introduccion

En esta exposición mostraremos el hardware disponible para ejecutar los algoritmos necesarios y realizar los cálculos en las aplicaciones de Inteligencia artificial (IA). El tipo de hardware necesario para la IA depende de la aplicación específica y la complejidad de la tarea.

Introduccion

En AI existen dos tareas fundamentales:

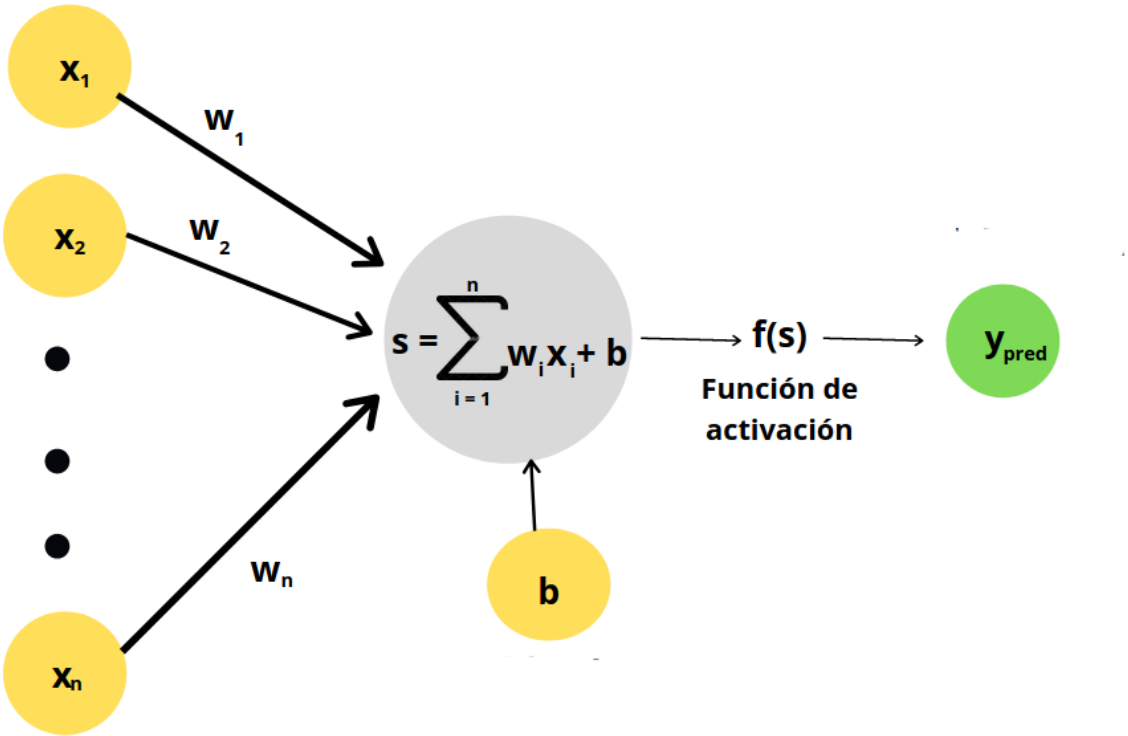
- ✓ Entrenamiento de la Red Neuronal
  El entrenamiento de la red generalmente no se requiere que sea en tiempo real y demanda una gran cantidad de recursos, generalmente se usan supercomputadoras.
- ✓ Ejecución de la Red Neuronal
  No requiere una gran cantidad de computo y generalmente se requiere la ejecución en tiempo real.

Introduccion

Algunos ejemplos de hardware utilizado para IA incluyen unidades de procesamiento de gráficos (GPU), matrices de puertas programables en campo (FPGA), circuitos integrados específicos de la aplicación (ASIC) y unidades de procesamiento tensorial (TPU). Las GPU se usan comúnmente para tareas de aprendizaje profundo debido a sus capacidades de procesamiento paralelo. Los FPGA y ASIC se utilizan para acelerar algoritmos específicos, y los TPU están diseñados para acelerar las cargas de trabajo de aprendizaje automático.
El desarrollo de hardware especializado para IA, como las TPU, ha mejorado significativamente el rendimiento de los algoritmos de IA y ha hecho posible entrenar y ejecutar modelos que antes no eran factibles. A medida que la IA continúa avanzando, el hardware seguirá desempeñando un papel fundamental para habilitar aplicaciones de IA más complejas y potentes.

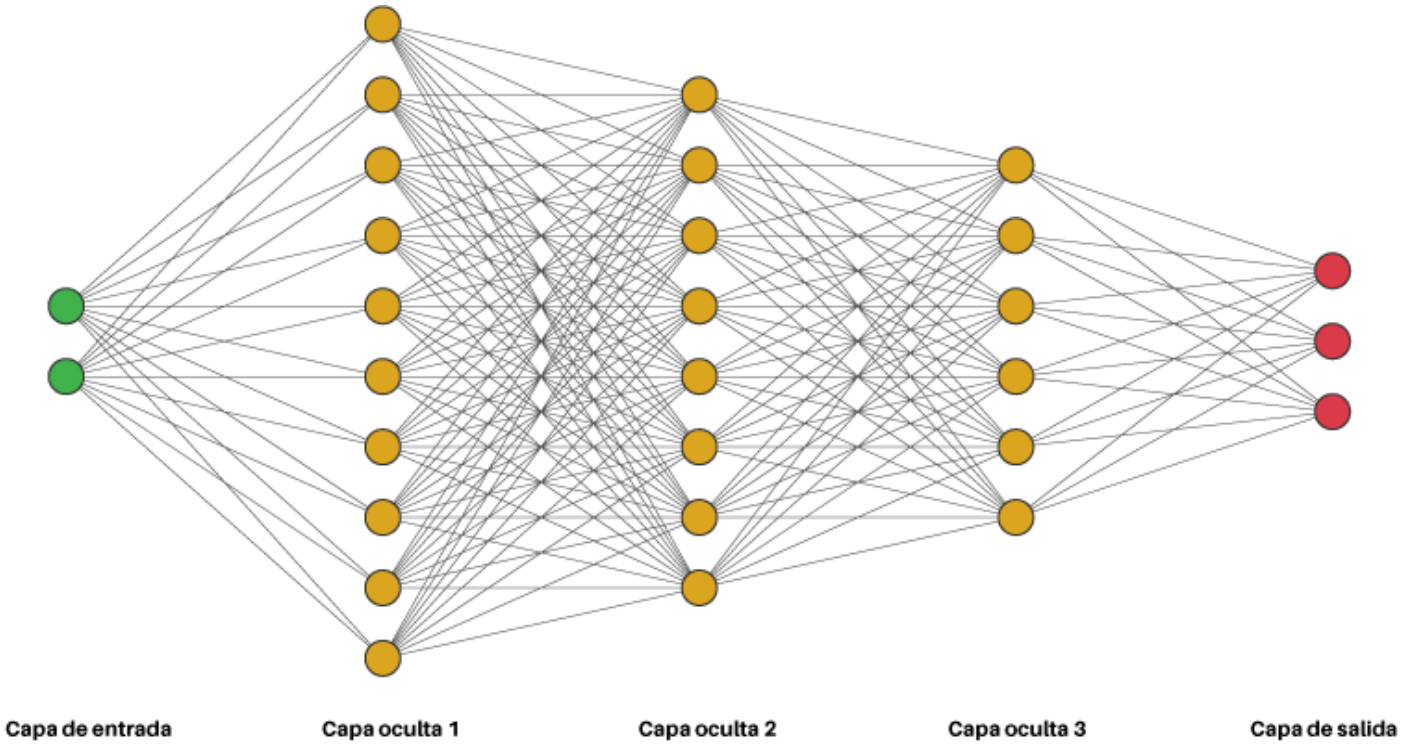
Introduccion

A continuación presentaremos el cálculo mas comúnmente usado en IA. En la figura representamos una neurona.



Introduccion

Normalmente una red neuronal tiene multiples neuronas por capas y multiples capas.



Introduccion

Como vimos en la primera figura.

$$S = \sum_{i=1}^{n} W_i X_i$$

Pero si tenemos multiple neuronas:

$$S_1 = \sum_{i=1}^{n} W_{i1} X_i$$

$$\dots \tag{1}$$

$$S_j = \sum_{i=1}^{n} W_{ij} X_i$$

$$\dots \tag{2}$$

$$S_m = \sum_{i=1}^{n} W_{im} X_i$$

Introducción

Este conjunto de ecuaciones lo podemos expresar en forma Matricial como.

$$S = W \times X$$

Donde W es :

$$W = \begin{bmatrix} W_{11} & \dots W_{i1} & \dots & W_{n1} \\ \dots & & & \\ W_{1j} & \dots W_{ij} & \dots & W_{nj} \\ \dots & & & \\ W_{1m} & \dots W_{im} & \dots & W_{nm} \end{bmatrix}$$

Donde podemos ver que se trata de multiplicación de matrices, el esquema de multiprocesadores como las GPU y TPU permite acelerar el proceso. A continuación veremos los elementos de harware que permiten la ejecución de estos algoritmos.

Introducción

$$S = W \times X$$

Básicamente la resolución de esta ecuación matricial involucra el mayor tiempo de maquina. Las placas gráficas (GPU) son buenas para resolver matrices, pero poseen longitud de palabras grandes mas de lo que se necesita en el calculo de redes y poseen una memoria centralizada, lo cual lleva tiempo de acceso, los chips dedicados a redes neuronales (NPU) tratan de tener longitud de palabras cortas y memoria localizada para almacenar los elemento de la matrix de pesos..

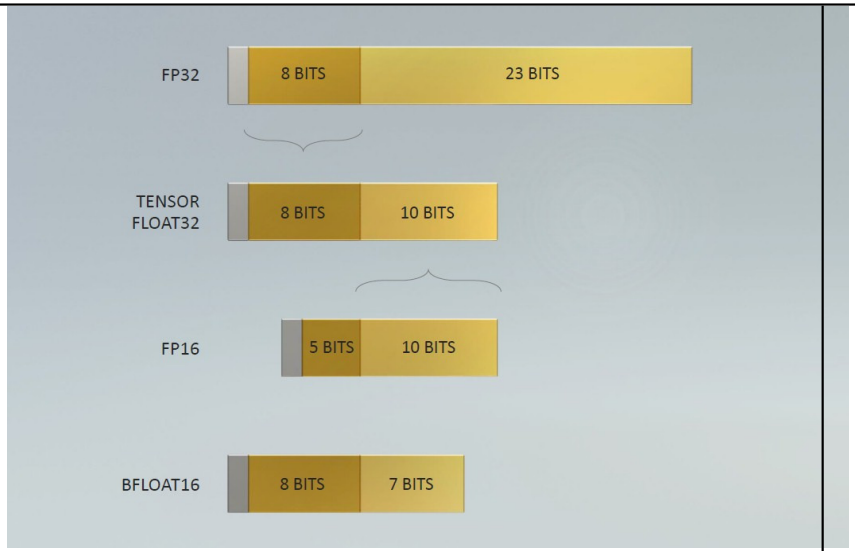
Introducción

Los flujos de trabajo relacionados con IA se basan en multiplicaciones de enteros por lo que obtener un algoritmo de multiplicación eficiente es fundamental.

$$C = O(n^2)$$

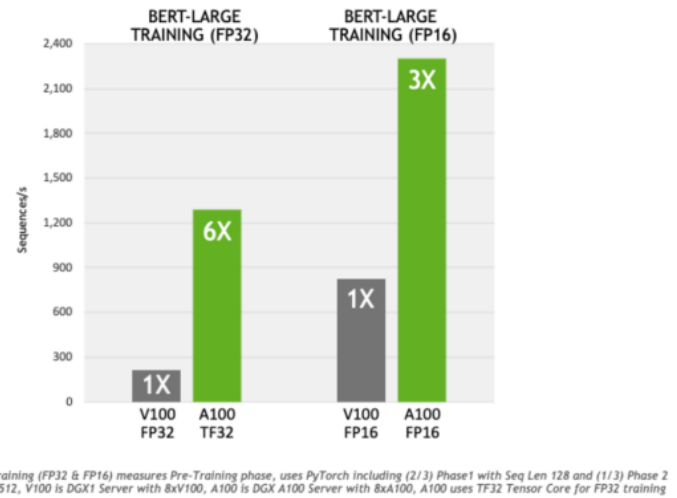
Por eso no es de extrañar que en 2022 se presentara un nuevo algoritmo en el Journal de Computación de Alta Eficiencia con una perfomance de $O(n \log(n) 2^{\theta(log(n))})$ [1].
Donde n es la longitud del numero en bits.
Para amplir este tema se puede leer el trabajo de Multiplicacion Entera

Introducción



La longitud de palabra influye en la velocidad de Calculo.

Introducción

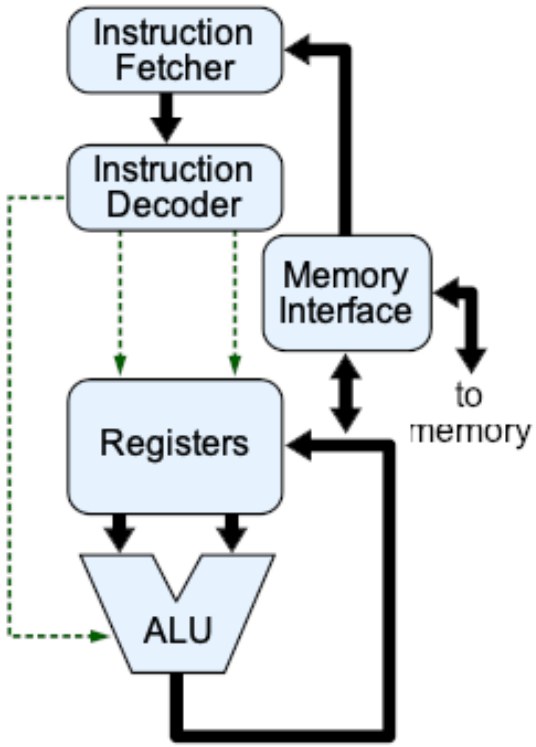


En este caso la empresa NVIDIA muestra la diferencia entre la longitud de palabra entre 16 y 32 bits.

Tipos de Hardware

- ✓ CPU
- ✓ GPU
- ✓ ASIC
- ✓ FPGA
- ✓ NPU/TPU
- ✓ QC

CPU Unidad Central de proceso



La unidad Central de proceso es el componente principal de las computadoras actuales

GPU

CPU Unidad Central de proceso

Arquitecturas de CPU

- ✓ RISC (Conjunto de Instrucciones reducido)
- ✓ CISC (Conjunto de instrucciones complejo)

Procesadores RISC

- ✓ Caracteristicas Se caracterizan por tener una arquitectura simple con un conjunto de instrucciones reducidas , esto hace que tengan pocos transistores por lo tanto menor disipación. Los compiladores son mas sencillos.
- ✓ Ejemplo La Arquitectura ARM por ejemplo un Cortex A15 tiene alrededor de 30 Millones de transistores

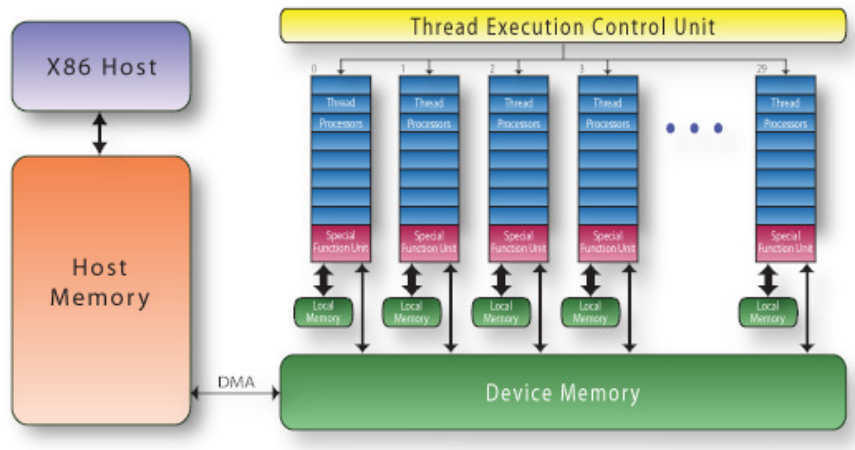
Implementaciones RISC

- ✓ ARM
- ✓ RISC V

Procesadores CISC

- ✓ Caracteristicas Se caracterizan por tener una arquitectura compleja con instrucciones complejas, esto hace que tengan mucho mas transistores por lo tanto mayor disipación. Los compiladores son mas complejos.
- ✓ Ejemplo La Arquitectura Intel X86 a partir del DX386. 2.950 Millones de transistores

GPU Unidad de proceso Grafico



Una unidad de procesamiento gráfico es un coprocesador dedicado al procesamiento de gráficos u operaciones de coma flotante, para aligerar la carga de trabajo del procesador [2].

▸▸ ASIC

GPU Unidad de proceso Grafico

Introduccion

Los algoritmos de aprendizaje profundo implican el procesamiento de grandes conjuntos de datos y la realización de una gran cantidad de operaciones matriciales. Las GPU pueden acelerar estas operaciones, lo que permite entrenar redes neuronales profundas mucho más rápido que con las CPU tradicionales. Debido a que las GPU pueden realizar varias tareas en paralelo, lo que permite un procesamiento más rápido de varias tareas simultáneamente.
En los últimos años, los fabricantes de GPU como NVIDIA han desarrollado GPU especializadas para IA, como NVIDIA Tesla V100, que ofrece un alto rendimiento y una gran capacidad de memoria adecuada para modelos complejos de aprendizaje profundo. Muchos proveedores de la nube también ofrecen instancias de GPU para aplicaciones de IA, como Amazon AWS, Microsoft Azure y Google Cloud Platform.

GPU Unidad de proceso Gráfico

GPU para IA

- ✓ Nvidia Tesla
- ✓ AMD Radeon Instinct
- ✓ Intel Xe

GPU Unidad de proceso Gráfico

GPU software para IA NVIDIA

NVIDIA CUDA-X AI es una librería de software de aprendizaje profundo para crear aplicaciones aceleradas por GPU de alto rendimiento para inteligencia artificial. Las bibliotecas CUDA-X AI ofrecen un gran rendimiento tanto para el entrenamiento como para la inferencia en los puntos de referencia de la industria, como MLPerf.
Todos los marcos de aprendizaje profundo, incluidos PyTorch, TensorFlow y JAX, se aceleran en GPU individuales, así como también se escalan a configuraciones de múltiples GPU y múltiples nodos.

GPU Unidad de proceso Gráfico

GPU software para IA AMD ROCm

- ✓ ROCm
- ✓ MIOpen
- ✓ MIGraphX
- ✓ MIVisionX

GPU Unidad de proceso Gráfico

GPU software para IA INTEL Xe

Scikit-learn es un módulo de Python para el aprendizaje automático. Intel® Extension para Scikit-learn acelera sin problemas sus aplicaciones de scikit-learn para CPU y GPU Intel en configuraciones de uno o varios nodos. Este paquete de extensión parcha dinámicamente los estimadores de scikit-learn mientras mejora el rendimiento de sus algoritmos de aprendizaje automático.

GPU Unidad de proceso Gráfico

GPU software para IA YOLO

Tal vez YOLO merezca un capitulo aparte ya que YOLO V8 se puede exportar a una gran variedad de formatos como puede verse en el grafico siguiente.

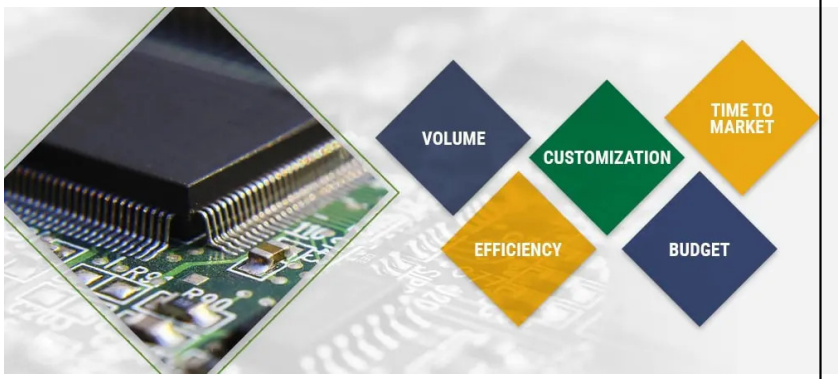
GPU Unidad de proceso Gráfico

GPU software para IA YOLO

Formato	Argumento	Modelo	Metadata
PyTorch	-	yolov8n.pt	✔
TorchScript	torchscript	yolov8n.torchscript	✔
ONNX	onnx	yolov8n.onnx	✔
OpenVINO	openvino	yolov8n_openvino_model/	✔
TensorRT	engine	yolov8n.engine	✔
CoreML	coreml	yolov8n.mlmodel	✔
TF SavedModel	saved_model	yolov8n_saved_model/	✔
TF GraphDef	pb	yolov8n.pb	✖
TF Lite	tflite	yolov8n.tflite	✔
TF Edge TPU	edgetpu	yolov8n_edgetpu.tflite	✔
TF.js	tfjs	yolov8n_web_model/	✔
PaddlePaddle	paddle	yolov8n_paddle_model/	✔

ASIC Circuito integrado de aplicación especifica

Introduccion



Un circuito integrado de aplicación específica (ASIC) es un chip de circuito integrado (IC) personalizado para un uso particular, en lugar de estar diseñado para un uso general, como un chip diseñado para ejecutarse en una grabadora de voz digital o un códec de video de alta eficiencia [3].

FPGA

ASIC Circuito integrado de aplicación especifica

Ventajas y Desventajas

Ventajas

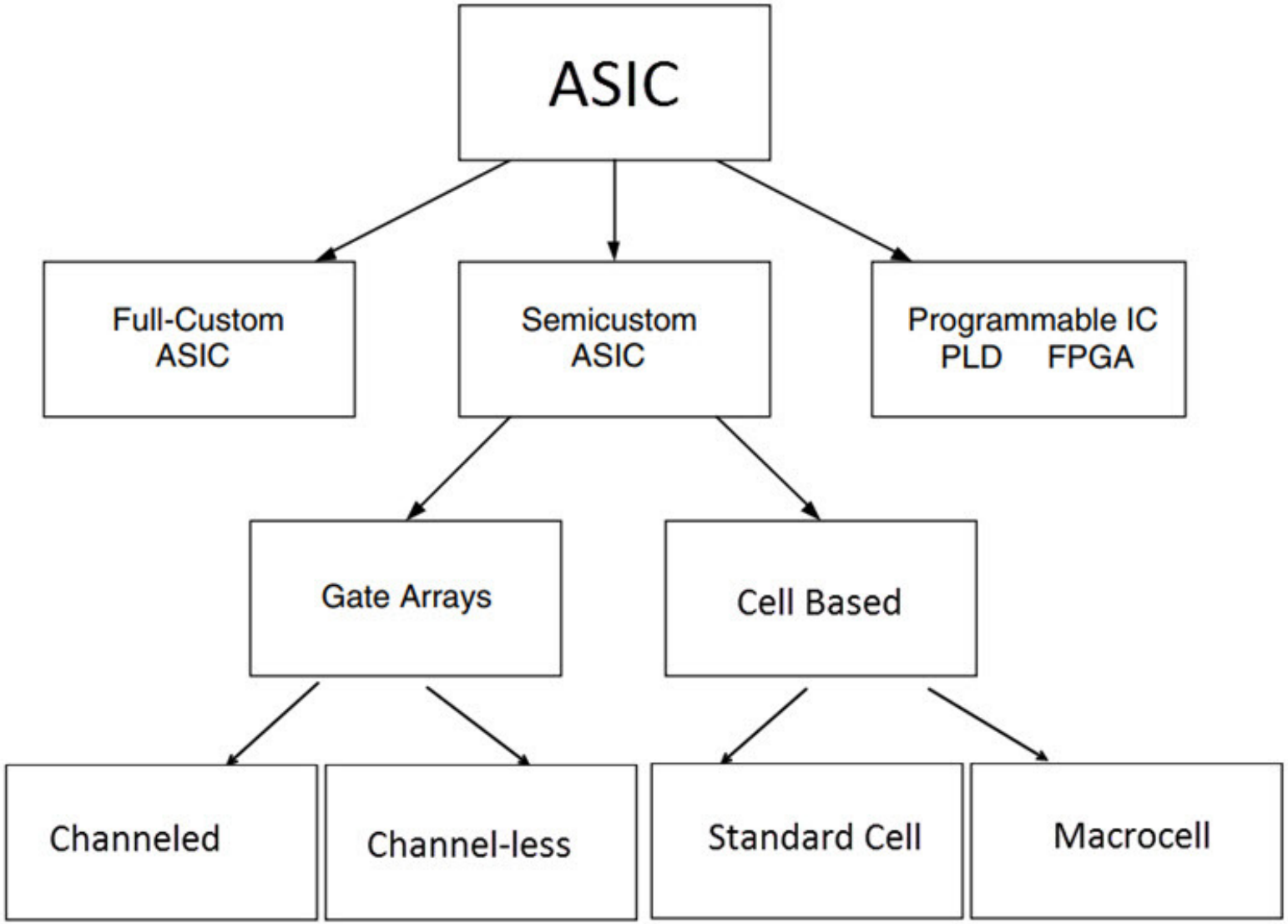
- ✓ Se puede realizar el circuito a medida.
- ✓ Alta eficiencia.
- ✓ No posee componentes innecesarios.
- ✓ Bajo Consumo

Desventajas

- ✓ Se deben realizar los drivers especificos.
- ✓ Mayor tiempo de desarrollo.

ASIC Circuito integrado de aplicación especifica

Tipos de ASIC



ASIC Circuito integrado de aplicación especifica

Servicio de obleas multiproyecto (Multi-project wafer service)

Muchas empresas ofrecen servicion en los cuales en una misma oblea se silicio imprimen integrados de diferentes proyectos.

- EUROPRACTICE
- CMC
- MOUSE
- GLOBAL FOUNDRIES
- XFAB
- SMIC

FPGA

Arquitectura de la FGPA

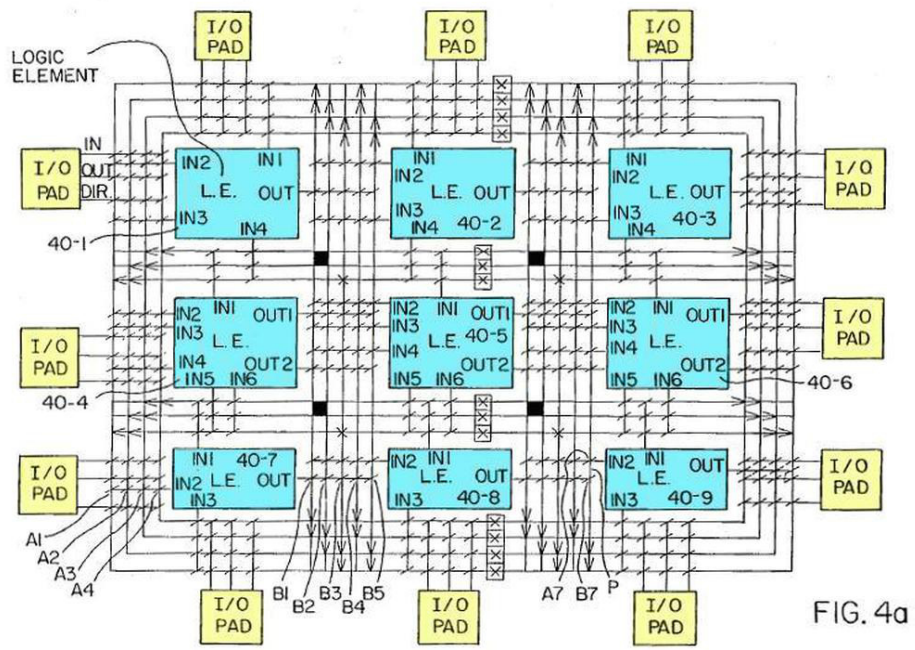


FIG. 4a

FPGA (Matriz de puertas lógicas programable en campo) es un dispositivo programable que contiene bloques de lógica cuya interconexión y funcionalidad puede ser configurada en el momento, mediante un lenguaje de descripción especializado [4].

NPU/TPU

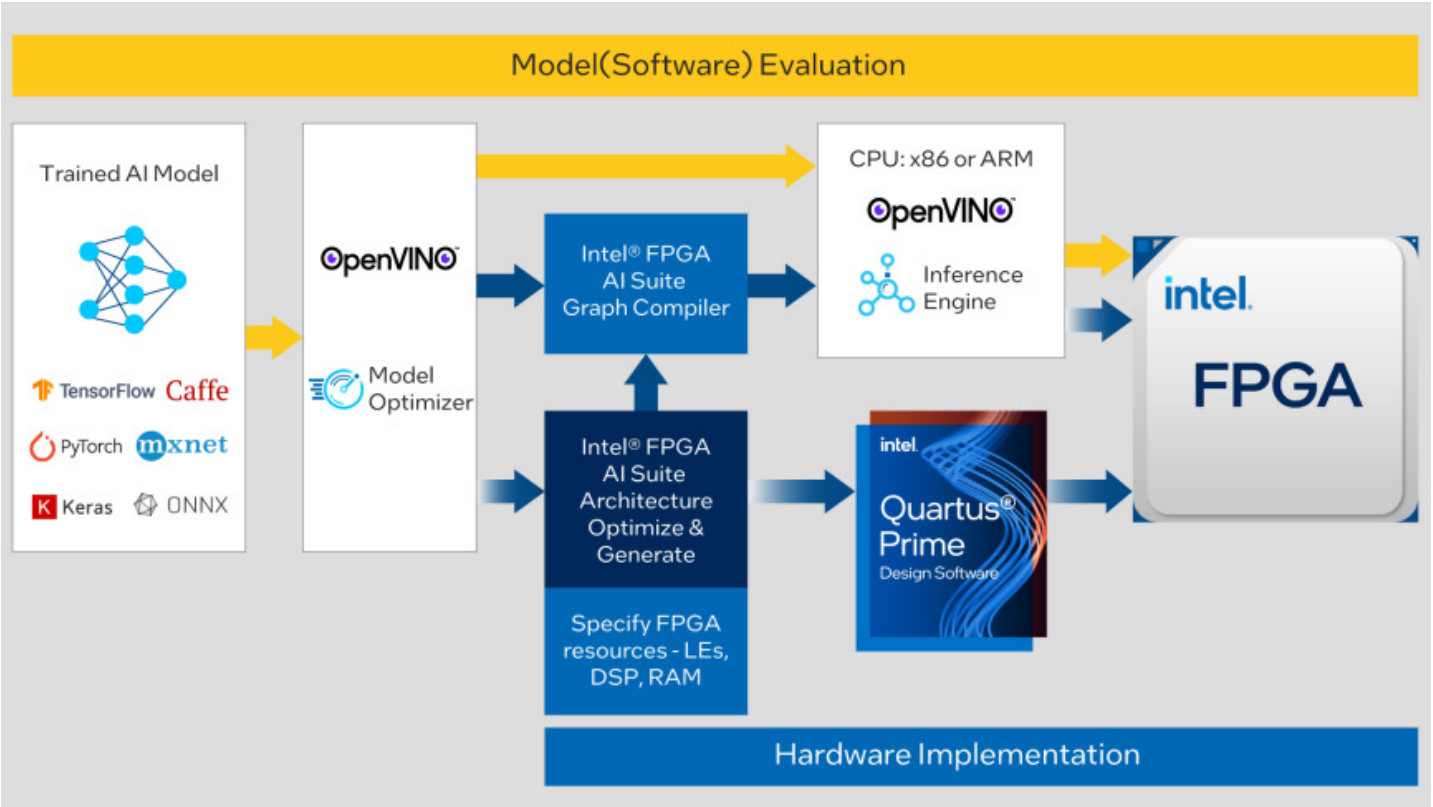
FPGA

Fabricantes FPGA

- ✓ Fabricantes: Xilinx , Altera, Intel.
- ✓ Herramientas de diseño: VHDL, Verilog, Vitis, Altera's Pro, FPGA AI Suite.

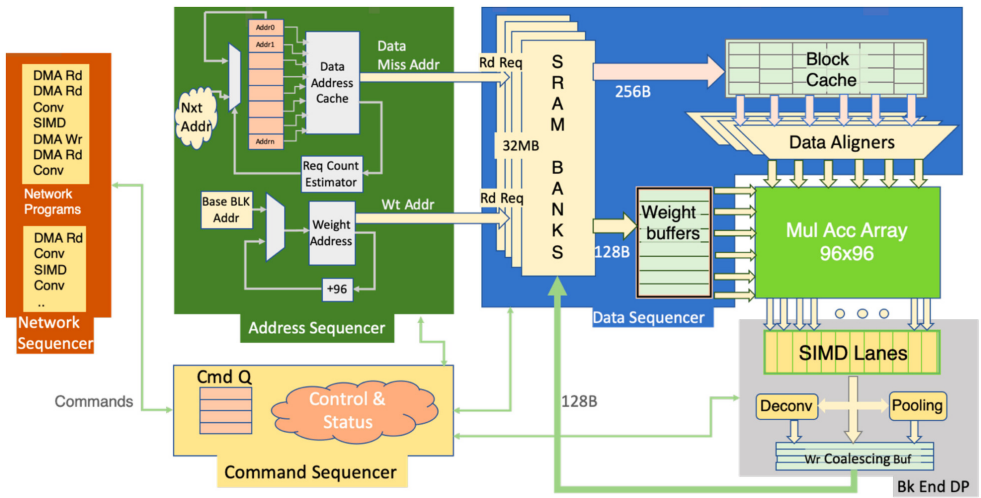
FPGA

Software de diseño FPGA



NPU Unidad de proceso Neuronal

Arquitectura de la NPU/TPU



La unidad procesadora Neuronal están optimizado para las operaciones matemáticas que se usan comúnmente en el aprendizaje automático, como multiplicaciones de matrices y convoluciones, y aceleran tareas de aprendizaje automático, como la clasificación de imágenes, la detección de objetos, el procesamiento del lenguaje natural y el reconocimiento de voz [5].

QC

NPU Unidad de proceso Neuronal

Tipos NPU

- ✓ NPU [5]
- ✓ TPU [5]
- ✓ Analogicas [6]

NPU Unidad de proceso Neuronal

Fabricantes NPU

- ✓ FSD Chip - Tesla (73,73 TOPS) [7]
- ✓ NVIDIA BlueField-3 DPU [8]
- ✓ Apple N1 Neural Engine [9]
- ✓ Intel Nervana processor [10]
- ✓ Dojo AI Supercomputer [11]
- ✓ Coral USB Acceleratorrr [12]
- ✓ Habana Gaudi2r [13]

NPU Unidad de proceso Neuronal

Fabricantes NPU

- ✓ Axelera Thetis (14.1 TOPS [14])
- ✓ Graphcore Bow (350 TF [15])
- ✓ GrAI Matter Lab [16]
- ✓ TensTorren Wormhole [17]
- ✓ Syntiang np [18]

NPU Unidad de proceso Neuronal

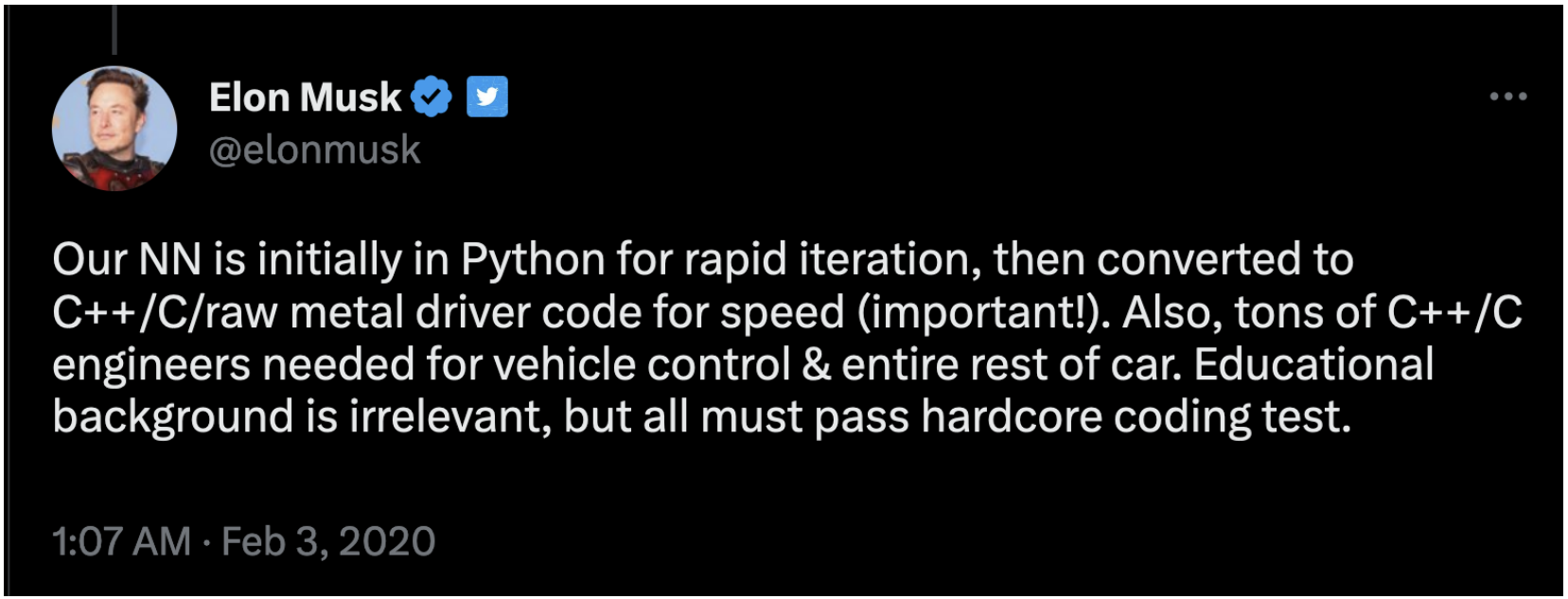
Fabricantes NPU

- ✓ Google TPU [17]
- ✓ Biren BR100 [19]
- ✓ Cerebras [20]
- ✓ Mythic Analog Matrix Processor [21]
- ✓ Untether AI Speed [22]

NPU Unidad de proceso Neuronal

Software NPU

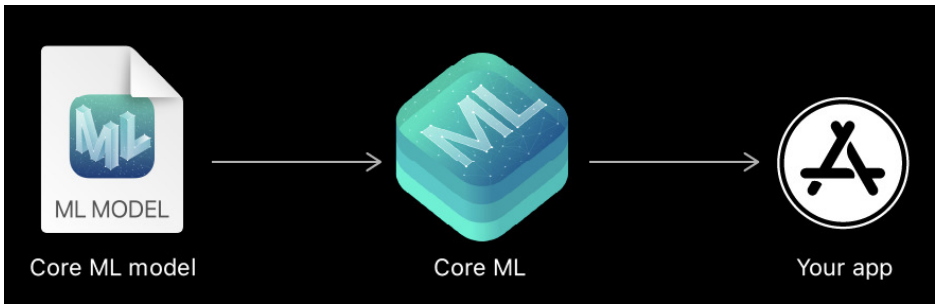
FSD Chip - Tesla (73,73 TOPS



NPU Unidad de proceso Neuronal

Software NPU

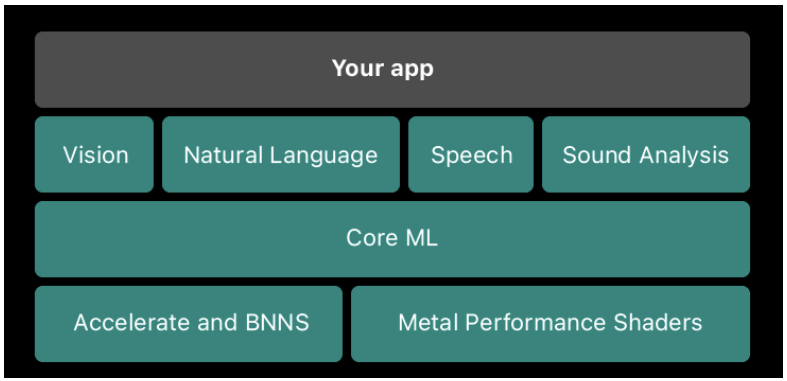
Apple N1 Neural Engine



NPU Unidad de proceso Neuronal

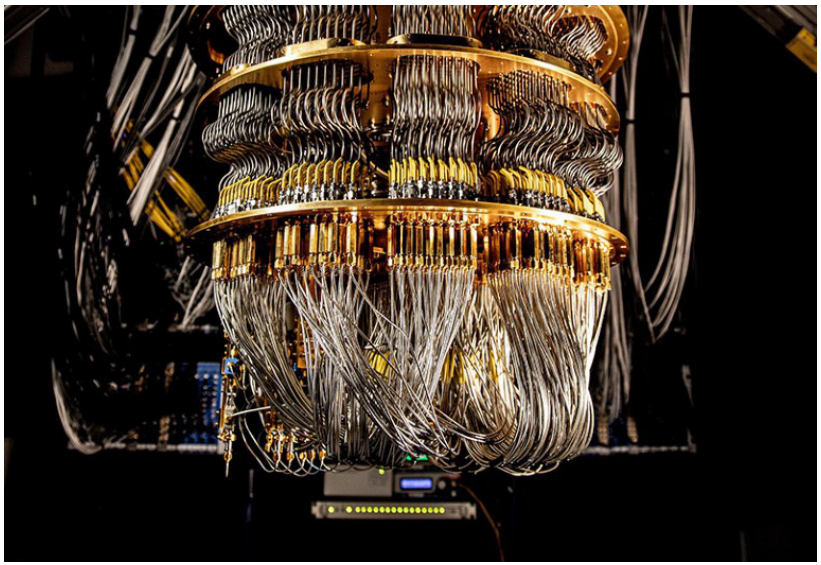
Software NPU

Apple N1 Neural Engine



QC Computadora Cuántica

Arquitecturas de QC



Una computadora cuantica trabaja con qbits en vez de bits estos pueden tomar simultanemente todos los valores posibles hasta que son observados o medidos. [23].

QC Computadora Cuántica

QC Estado del Arte

- ✓ System Two IBM 400 qb [24]
- ✓ Borealis Xanadu Cloud 200 qb [25]
- ✓ Rigetti Aspen M2 80 qb [26]
- ✓ Spinq Gemini 2 qb [27]
- ✓ DWAVE [28]
- ✓ Quantinuum [29]

Resumen I

Resumen II

- Tipos de ASIC
- MPWS

5. FPGA
   - Arquitectura de la FGPA
   - Fabricantes de FPGA
   - Evaluacion de Software
6. NPU/TPU
   - Arquitecturas NPU/TPU
   - Tipos NPU
   - Fabricantes NPU
   - Fabricantes NPU
   - Fabricantes NPU
   - Software NPU
   - Software NPU
   - Software NPU
7. QC

Resumen III

- Arquitecturas de QC
- Estado del arte de QC

8 Bibliografia

Bibliografia I

[1] A. P. Dieguez, M. Amor, R. Doallo, A. Nukada, and S. Matsuoka, "Efficient high-precision integer multiplication on the GPU," *The International Journal of High Performance Computing Applications*, vol. 36, no. 3, pp. 356–369, May 2022, publisher: SAGE Publications Ltd STM. [Online]. Available: https://doi.org/10.1177/10943420221077964

[2] "Unidad de procesamiento gráfico," Mar. 2023, page Version ID: 149613847. [Online]. Available: https://es.wikipedia.org/w/index.php?title=Unidad_de_procesamiento_gr%C3%A1fico&oldid=149613847

[3] Administrator, "Introduction to ASIC Technology | Different Types, Design Flow, Applications," Jan. 2020. [Online]. Available: https://www.electronicshub.org/introduction-to-asic-technology/

Bibliografia II

[4] "Field-programmable gate array," May 2022, page Version ID: 143711160. [Online]. Available: https://es.wikipedia.org/w/index.php?title=Field-programmable_gate_array&oldid=143711160

[5] "AI Chips: NPU vs. TPU - Bizety: Research & Consulting." [Online]. Available: https://www.bizety.com/2023/01/03/ai-chips-npu-vs-tpu/

[6] R. Khaddam-Aljameh, M. Stanisavljevic, J. Fornt Mas, G. Karunaratne, M. Brändli, F. Liu, A. Singh, S. M. Müller, U. Egger, A. Petropoulos, T. Antonakopoulos, K. Brew, S. Choi, I. Ok, F. L. Lie, N. Saulnier, V. Chan, I. Ahsan, V. Narayanan, S. R. Nandakumar, M. Le Gallo, P. A. Francese, A. Sebastian, and E. Eleftheriou, "HERMES-Core—A 1.59-TOPS/mm2 PCM on 14-nm CMOS In-Memory Compute Core Using 300-ps/LSB Linearized CCO-Based ADCs," *IEEE Journal of Solid-State Circuits*, vol. 57,

Bibliografia III

no. 4, pp. 1027–1038, Apr. 2022, conference Name: IEEE Journal of Solid-State Circuits.

[7] "FSD Chip - Tesla - WikiChip." [Online]. Available: https://en.wikichip.org/wiki/tesla_(car_company)/fsd_chip

[8] "NVIDIA BlueField Data Processing Units(DPUs)." [Online]. Available: https://www.nvidia.com/en-us/networking/products/data-processing-unit/

[9] "Neural Engine." [Online]. Available: https://apple.fandom.com/wiki/Neural_Engine

[10] "https://www.intel.com/content/www/us/en/artificial-intelligence/nnpi.html." [Online]. Available: https://www.intel.com/content/www/us/en/artificial-intelligence/nnpi.html

Bibliografia IV

[11] T. P. Morgan, "Inside Tesla's Innovative And Homegrown "Dojo" AI Supercomputer," Aug. 2022. [Online]. Available: https://www.nextplatform.com/2022/08/23/inside-teslas-innovative-and-homegrown-dojo-ai-supercomputer/

[12] "USB Accelerator." [Online]. Available: https://coral.ai/products/accelerator/

[13] "Habana Gaudi2." [Online]. Available: https://habana.ai/products/gaudi2/

[14] "Axelera AI Announces Successful Testing of Thetis Core Chip - Axelera AI," May 2022. [Online]. Available: https://www.axelera.ai/axelera-ai-announces-successful-testing-of-thetis-core-chip/

[15] G. Ltd, "Bow IPU Processors." [Online]. Available: https://www.graphcore.ai/bow-processors

Bibliografia V

[16] "GrAI Matter Labs | Fastest Edge AI Processor." [Online]. Available: https://www.graimatterlabs.ai//

[17] "AICloud." [Online]. Available: https://tenstorrent.com/aicloud/

[18] "Hardware Solutions." [Online]. Available: https://www.syntiant.com/hardware-solutions

[19] "La GPGPU china Biren BR100 no puede ser fabricada por TSMC debido a la normativa estadounidense y a la competencia con NVIDIA," Oct. 2022. [Online]. Available: https://www.notebookcheck.org/La-GPGPU-china-Biren-BR100-no-puede-ser-fabricada-por-TSMC-debic 663733.0.html

Bibliografia VI

[20] N. Dey, "Cerebras-GPT: A Family of Open, Compute-efficient, Large Language Models," Mar. 2023. [Online]. Available: https://www.cerebras.net/blog/cerebras-gpt-a-family-of-open-compute-efficient-large-language-models/

[21] "Technology." [Online]. Available: https://mythic.ai/technology/

[22] "Untether AI." [Online]. Available: https://www.untether.ai

[23] D. Castelvecchi, "Google's quantum computer hits key milestone by reducing errors," *Nature*, Feb. 2023, bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Computer science, Quantum information. [Online]. Available: https://www.nature.com/articles/d41586-023-00536-w

Bibliografia VII

[24] "IBM Unveils 400 Qubit-Plus Quantum Processor and Next-Generation IBM Quantum System Two." [Online]. Available: https://newsroom.ibm.com/2022-11-09-IBM-Unveils-400-Qubit-Plus-Quantum-Processor-and-Next-

[25] "Borealis." [Online]. Available: https://www.xanadu.ai/products/borealis

[26] "Quantum Computing." [Online]. Available: https://www.rigetti.com/

[27] "Gemini·2-qubit desktop NMR quantum computer - SpinQ." [Online]. Available: https://www.spinquanta.com/products-solutions/gemini

[28] "D-Wave Systems | The Practical Quantum Computing Company." [Online]. Available: https://www.dwavesys.com/

Bibliografia VIII

[29] “Quantinuum | Hardware | System Model H2.” [Online]. Available: https://www.quantinuum.com/hardware/h2