UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
FACULTAD DE INGENIERÍA ELECTRÓNICA

Tarea #2

Arquitecturas de Computación en la Era Digital

Presentado a: Ing. Diego Alejandro Barragan Vargas

Juan Diego Báez Guerrero, Cód.: 2336781.

Resumen*— *En la actualidad, el desarrollo de arquitecturas de computación ha permitido avances significativos en procesamiento de datos, inteligencia artificial y computación en la nube. Este informe analiza en detalle las arquitecturas de GPU, TPU, neuromórficas, cuánticas, cloud computing, computación heterogénea y arquitecturas distribuidas en la nube. Se presenta un estudio sobre su evolución, características principales y aplicaciones en diversas industrias. Además, se discute su impacto en el rendimiento y eficiencia energética de los sistemas computacionales modernos.

Abstract*— *Currently, the development of computing architectures has enabled significant advances in data processing, artificial intelligence, and cloud computing. This report provides a detailed analysis of GPU, TPU, neuromorphic, quantum, cloud computing, heterogeneous computing, and distributed cloud architectures. It presents a study on their evolution, key characteristics, and applications across various industries. Additionally, their impact on the performance and energy efficiency of modern computing systems is discussed.*

I INTRODUCCIÓN

La evolución de la computación ha llevado al desarrollo de múltiples arquitecturas optimizadas para tareas específicas. Desde los procesadores tradicionales de propósito general hasta unidades de procesamiento gráfico (GPU) y aceleradores especializados como las TPU, las innovaciones en hardware han transformado el manejo y análisis de grandes volúmenes de datos.

En este informe, se presentan las principales arquitecturas utilizadas en la computación actual, incluyendo sistemas diseñados para inteligencia artificial, computación en la nube y procesamiento paralelo. Se exploran sus características fundamentales, ventajas y limitaciones, así como su impacto en la eficiencia energética y la escalabilidad de los sistemas modernos.

II MARCO TEÓRICO

A. *Arquitectura de GPU*

Las Unidades de Procesamiento Gráfico (GPU) fueron diseñadas originalmente para el renderizado de gráficos en videojuegos. Su capacidad para ejecutar múltiples operaciones en paralelo las ha convertido en herramientas fundamentales para inteligencia artificial, modelado 3D y simulaciones científicas.

Las GPU han revolucionado la manera en que se realizan cálculos de señales y procesamiento de imágenes en tiempo real. Además, en sistemas digitales, su arquitectura SIMD (Single Instruction, Multiple Data) permite paralelismo masivo, lo que acelera procesos computacionales intensivos.

Desde el punto de vista de sistemas operativos, las GPU requieren controladores especializados que gestionan la asignación de memoria y los procesos de cómputo paralelo.

Ejemplos de arquitecturas optimizadas incluyen CUDA de NVIDIA y ROCm de AMD [1], [2], [3].

B. *Arquitectura de TPU*

Las Unidades de Procesamiento Tensorial (TPU) fueron introducidas por Google como aceleradores especializados para aprendizaje automático. A diferencia de las GPU, están diseñadas específicamente para operaciones de redes neuronales, optimizando la eficiencia energética y reduciendo la latencia en inferencias de IA.

Han permitido el desarrollo de dispositivos inteligentes con capacidad de aprendizaje en el borde (edge computing). La arquitectura de las TPU se basa en unidades matriciales de alto rendimiento, optimizando operaciones con grandes volúmenes de datos.

Los sistemas operativos modernos han desarrollado marcos como TensorFlow y PyTorch para aprovechar estas capacidades de hardware[4], [5], [6].

C. *Arquitectura Neuromórfica*

Inspiradas en el funcionamiento del cerebro humano, las arquitecturas neuromórficas buscan replicar el comportamiento de las neuronas y sinapsis. Tecnologías como Loihi de Intel o TrueNorth de IBM representan avances en hardware que permiten un procesamiento altamente paralelo con bajo consumo energético.

Este tipo de arquitectura abre nuevas posibilidades en el diseño de chips inteligentes y dispositivos de IA embebidos. Se implementan redes neuronales hardware que eliminan la necesidad de simulación en software.

Desde el punto de vista de los sistemas operativos, estos chips requieren modelos de programación innovadores que imiten los procesos neuronales del cerebro [7], [8], [9].

D. Computación Cuántica

La computación cuántica se basa en principios de la mecánica cuántica, utilizando cúbits en lugar de bits tradicionales. Empresas como IBM, Google y D-Wave han liderado el desarrollo de hardware cuántico.

La computación cuántica representa un reto en la fabricación de circuitos superconductores para cúbits estables. Este paradigma requiere nuevas formas de codificación y procesamiento de información.

Han surgido plataformas como Qiskit e IBM Quantum Experience para la gestión y simulación de algoritmos cuánticos [10], [11], [12].

E. Cloud Computing

La computación en la nube permite el acceso remoto a recursos computacionales mediante servidores distribuidos. Gracias a la virtualización y escalabilidad, servicios como AWS, Google Cloud y Microsoft Azure han transformado la gestión de infraestructuras digitales.

La computación en la nube ha impulsado el Internet de las Cosas (IoT), permitiendo la conexión y gestión remota de dispositivos embebidos. Se optimiza la transferencia y almacenamiento de grandes volúmenes de datos.

Evolucionado con tecnologías de virtualización y contenedorización como Docker y Kubernetes para optimizar la infraestructura en la nube.

F. Computación Heterogénea

La computación heterogénea combina diferentes tipos de procesadores, como CPU, GPU y FPGA, para optimizar el rendimiento en aplicaciones específicas.

Esto ha llevado al desarrollo de SoCs (System on Chip) avanzados. Permite la asignación óptima de cargas de trabajo según la capacidad del hardware.

Los sistemas operativos modernos implementan estrategias de gestión de recursos heterogéneos para mejorar la eficiencia de cómputo.

G. Arquitecturas Distribuidas en la Nube

Los sistemas distribuidos en la nube aprovechan múltiples servidores conectados en red para procesar grandes volúmenes de datos de manera escalable y redundante. Tecnologías como Kubernetes y Docker han facilitado la gestión de microservicios.

Esta arquitectura ha impulsado el desarrollo de infraestructuras escalables en IoT. Permite la integración eficiente de redes de procesamiento masivo.

La gestión de contenedores y máquinas virtuales permite mejorar el rendimiento y disponibilidad de servicios en la nube.

REFERENCIAS

[1] NVIDIA, ÇUDA Programming Guide,"2021.
[2] AMD, ROCm Platform Overview,"2022.
[3] J. Sanders and E. Kandrot, ÇUDA by Example,.ªddison-Wesley, 2020.
[4] Google, Ïntroduction to Tensor Processing Units,"2018.
[5] N. Jouppi et al., "TPU: Machine Learning Accelerator,ÏEEE Micro, 2019.
[6] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning,"MIT Press, 2020.
[7] IBM, "TrueNorth Architecture Overview,"2017.
[8] Intel, "Loihi: Neuromorphic Computing Architecture,"2021.
[9] Y. Chua et al., "Trends in Neuromorphic Computing,ÏEEE Transactions, 2022.
[10] IBM, ÏBM Quantum Experience Overview,"2019.
[11] Google, "Quantum Computing Advances,"2020.
[12] IBM, "Qiskit: Open-Source Quantum Computing Framework,"2021.