UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
FACULTAD INGENIERÍA ELECTRÓNICA

TAREA 2

Integrantes: J. Alvarado Monroy,

facultad de Ingeniería Electrónica, Universidad Santo Tomás, Bogotá

I.INTRODUCCIÓN

Las arquitecturas de procesamiento actuales abarcan una amplia gama de tecnologías diseñadas para optimizar tareas específicas en el ámbito de la computación y la inteligencia artificial. Este informe presenta un análisis detallado de las principales arquitecturas utilizadas en la actualidad.

II. MARCO TEÓRICO

Procesador (CPU, GPU, TPU, etc.)
Un procesador es el componente de hardware encargado de ejecutar instrucciones y realizar cálculos en un sistema informático. Existen diferentes tipos de procesadores, como:

- CPU (Unidad Central de Procesamiento): Es el procesador principal de un sistema y se encarga de realizar tareas generales.
- GPU (Unidad de Procesamiento Gráfico): Diseñada originalmente para el procesamiento gráfico, ahora se usa ampliamente en cálculos paralelos y aprendizaje profundo.
- TPU (Unidad de Procesamiento Tensorial): Procesador optimizado para el aprendizaje automático y operaciones de inteligencia artificial.

Arquitectura Von Neumann: Usa una memoria unificada para datos e instrucciones, lo que puede generar cuellos de botella.
Arquitectura Harvard: Separa la memoria de datos y de instrucciones, permitiendo mayor velocidad.

Computación en la Nube y Distribuida

- Computación en la Nube: Uso de servidores remotos para almacenar y procesar datos en lugar de hacerlo localmente en un solo dispositivo.
- Computación Distribuida: Un sistema en el que múltiples dispositivos o servidores trabajan juntos para resolver un problema.

Inteligencia Artificial y Aprendizaje Automático

- Inteligencia Artificial (IA): Campo de la computación que busca desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana.
- Aprendizaje Automático (Machine Learning): Rama de la IA donde las computadoras aprenden de datos en lugar de ser programadas explícitamente.
- Redes Neuronales: Algoritmos inspirados en el cerebro humano que se usan en el aprendizaje profundo.

Computación Cuántica y Neuromórfica

- Computación Cuántica: Modelo de computación basado en los principios de la

mecánica cuántica, utilizando cúbits en lugar de bits tradicionales.

- **Computación Neuromórfica**: Diseño de hardware inspirado en el funcionamiento del cerebro, utilizando redes neuronales para procesamiento eficiente.

III. PROCEDIMIENTO Y RESULTADOS

1) Arquitecturas de Procesamiento

Unidades de Procesamiento Gráfico (GPU)

Originalmente diseñadas para renderizar gráficos, las GPUs han evolucionado para manejar operaciones paralelas masivas. Son ideales para tareas de aprendizaje profundo y procesamiento de grandes volúmenes de datos. Su arquitectura permite ejecutar múltiples operaciones simultáneamente, acelerando significativamente el entrenamiento e inferencia de modelos de inteligencia artificial.

Unidades de Procesamiento Tensorial (TPU)

Desarrolladas por Google, las TPUs son circuitos integrados específicos para aplicaciones (ASIC) optimizados para operaciones tensoriales. A diferencia de las GPUs, las TPUs están diseñadas específicamente para acelerar cargas de trabajo de aprendizaje automático, ofreciendo tiempos de entrenamiento e inferencia más rápidos gracias a su arquitectura personalizada.

Computación Neuromórfica

Esta tecnología se inspira en la estructura y funcionamiento del cerebro humano, utilizando redes neuronales de impulsos (SNN) que procesan y almacenan información simultáneamente. Los sistemas neuromórficos destacan por su adaptabilidad, eficiencia energética y capacidad de procesamiento paralelo, lo que los hace prometedores para aplicaciones que requieren aprendizaje en tiempo real y adaptación continua.

Computación Cuántica

Basada en los principios de la mecánica cuántica, esta tecnología utiliza cúbit que pueden representar múltiples estados simultáneamente, permitiendo realizar cálculos complejos a velocidades inalcanzables para las computadoras clásicas. Un ejemplo destacado es "Willow", el chip cuántico de Google capaz de resolver en cinco minutos tareas que los superordenadores actuales tardarían cuatrillones de años en completar.

2) Computación en la Nube y Arquitecturas Distribuidas

Computación en la Nube (Cloud Computing)

Esta arquitectura permite el acceso remoto a recursos informáticos a través de internet, ofreciendo servicios como almacenamiento, procesamiento y software bajo demanda. Facilita la escalabilidad, flexibilidad y reducción de costos, permitiendo a las organizaciones adaptarse rápidamente a las necesidades cambiantes sin invertir en infraestructura física propia.

Computación Heterogénea

Consiste en la integración de diferentes tipos de unidades de procesamiento (CPU, GPU, TPU, entre otras) en un mismo sistema para optimizar el rendimiento y la eficiencia energética. Al asignar tareas específicas al hardware más adecuado, se mejora la eficiencia y se reducen los tiempos de procesamiento en aplicaciones complejas.

Arquitecturas Distribuidas en la Nube

Estas arquitecturas implican la distribución de tareas y datos a través de múltiples servidores y centros de datos en la nube, mejorando la disponibilidad, escalabilidad y resiliencia de las aplicaciones. Permiten gestionar grandes

volúmenes de datos y tráfico, garantizando un rendimiento consistente incluso ante fallos en componentes individuales.

IV. CONCLUSIONES

Las arquitecturas de computación modernas, como las GPU, TPU, neuromórficas, cuánticas, y las basadas en cloud computing, computación heterogénea y sistemas distribuidos, representan un ecosistema diverso y complementario que está impulsando la innovación tecnológica a un ritmo sin precedentes. Cada una de estas arquitecturas tiene un propósito específico y se adapta a diferentes necesidades, desde el procesamiento masivo de datos hasta la simulación de sistemas complejos y la ejecución de algoritmos de inteligencia artificial

V. REFERENCIAS

VI. [1] D. B. Kirk and W. W. Hwu, *Programming Massively Parallel Processors: A Hands-on Approach*, 3rd ed. Amsterdam, Netherlands: Elsevier, 2016.

VII. [2] N. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, Toronto, ON, Canada, 2017, pp. 1-12.

VIII. [3] W. S. McCuistion, J. T. W. Kuo, and E. Culurciello, "Neuromorphic Computing and Spiking Neural Networks: An Overview," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 6, pp. 1121-1133, Dec. 2020.

IX. [4] J. Preskill, "Quantum Computing in the NISQ Era and Beyond," *Quantum*, vol. 2, p. 79, 2018.

X. [5] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," *National Institute of Standards and Technology*, Gaithersburg, MD, USA, Tech. Rep. NIST SP 800-145, 2011.

XI. [6] H. Esmaeilzadeh, T. Cao, X. Yang, S. Blackburn, and K. Sankaralingam, "Architecting for Performance and Energy Efficiency in Heterogeneous Computing," *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Cambridge, MA, USA, 2017, pp. 121-134.

XII. [7] I. Foster, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of High Performance Computing Applications*, vol. 15, no. 3, pp. 200-222, 200