Arquitecturas y aceleradores: innovaciones clave en la computación moderna

Sonia Angélica Muñoz Sandoval
Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga,Santander
soniaangelicam.s@gmail.com

Karen Juliana Mora Jaimes
Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Santander
karen2202027@correo.uis.edu.co

Santiago Andrés Delgado Quiceno
Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Santander
santiago2211799@correo.uis.edu.co

RESUMEN

Con el desarrollo de la computación actual, tecnologías como las arquitecturas especializadas de la mano de los aceleradores de hardware o coprocesadores, han sido objeto de estudio debido a su innovación y los beneficios que trae la implementación de los mismos, principalmente en cuanto a rendimiento, eficiencia energética y flexibilidad. El siguiente paper académico abarca la manera en la que estas tecnologías han evolucionado, además del impacto que han tenido en el contexto actual de la informática, haciendo mención de la aparición de las primeras computadoras electrónicas hasta las tendencias en el momento. A través de una revisión detallada de la literatura y casos de estudio relevantes para la investigación, se analizan los campos de aplicación, las ventajas y desventajas, los resultados obtenidos y por último las implicaciones y tendencias en el futuro, contribuyendo no solamente al otorgar una vista amplia de lo que significa el avance de las arquitecturas y los aceleradores, sino que también identificando las áreas y principios clave para la investigación futura y el desarrollo continuo y óptimo de la informática.

***Abstract*—With the advancement of modern computing, technologies such as specialized architectures, in conjunction with hardware accelerators or coprocessors, have become subjects of study due to their innovation and the benefits their implementation brings, primarily in terms of performance, energy efficiency, and flexibility. This academic paper covers these technologies' evolution and impact on the current computing landscape, tracing the emergence of the first electronic computers to present-day trends. The paper analyzes application fields, advantages and disadvantages, obtained results, and future implications and trends through a detailed literature review and relevant case studies. This study contributes by providing a comprehensive overview of the advancements in architectures and accelerators, while also identifying key areas and principles for future research and the continuous, optimal development of computer science."**

***Key words*—Specialized architectures, hardware accelerators, performance, energy efficiency, computing flexibility, computer science development**

I. INTRODUCCIÓN

La historia de la evolución de las arquitecturas de computadores y aceleradores pueden describirse como un viaje de constante innovación, donde la primera de estas se remonta desde la década de los 40, en la cual aparecen las primeras computadoras electrónicas UNIVAC y ENIAC, basadas en la arquitectura de Von Neumann y caracterizadas por utilizar tubos al vacío para circuitos y tambores magnéticos para la memoria, hasta pasar por la era de los transistores, los circuitos integrados, los microprocesadores, las computadoras personales, los procesadores superescalares y multinúcleo, la computación heterogénea y finalmente, la computación cuántica y demás tendencias actuales, han buscado satisfacer la demanda que se ha venido presentando con el pasar de los años en cuanto a rendimiento, eficiencia energética, flexibilidad, conectividad, entre otros (Instituto Nacional de Aprendizaje, s.f).

Paralelamente a lo anterior, cabe resaltar también el papel que han tenido los aceleradores, desde los primeros coprocesadores hasta las TPUs o unidades sofisticadas de procesamiento tensorial, que también se han encargado de resolver ciertas problemáticas frente a la optimización de tareas específicas y cuellos de botella en el rendimiento computacional. ¿Pero qué es un acelerador? "Un acelerador es una subestructura arquitectónica separada (en el mismo chip o en un dado diferente) que está diseñado utilizando un conjunto de objetivos diferentes a los del procesador base, donde estos objetivos se derivan de las necesidades de una clase especial de aplicaciones" (Sanjay Patel y Wen-mei W. Hwu, 2007, p. 4). En general, el acelerador se encarga de proporcionar un mayor rendimiento a menor costo, menor consumo de energía o menor esfuerzo de desarrollo. Las GPUs que inicialmente fueron diseñadas para renderizar gráficos 3D, terminaron convirtiéndose en herramientas esenciales para la computación de alto rendimiento, así como los demás tipos comunes de aceleradores de hardware: FPGAs y ASICs.

Como los aceleradores, la arquitectura de computadores se define como la organización lógica del hardware de los ordenadores, en otras palabras, el conjunto de los principios que detallan las características del hardware del ordenador y su interacción mutua, incluyendo componentes como la CPU, los dispositivos de entrada/salida, la memoria, entre otros (Isaac, 2022), y la relación entre ambos conceptos es que en los computadores actuales se está realizando la implementación de aceleradores como parte de su diseño, en donde trabajan en conjunto con el procesador principal (computación heterogénea), lo que permite que la eficiencia de

la computadora sea mayor y que la capacidad y la velocidad para manejar tareas complejas aumente.

La importancia de las arquitecturas modernas y el porqué de indagar más sobre este tema yace en que los aceleradores cada vez tienen un mayor impacto en la computación actual, puesto a que están presentes en áreas clave de la tecnología que están en constante evolución, como en el caso de la inteligencia artificial, en donde se ha hecho indispensable contar con hardware especializado que pueda acelerar los procesos necesarios y mejorar la eficiencia y el rendimiento, como por ejemplo el uso de GPU para la manipulación de imágenes y el cálculo de propiedades locales de imágenes, el uso de FPGAs que facilitan la evolución de hardware, marcos y software de manera conjunta y por último, el circuito integrado ASIC para optimizar el uso de memoria y la aritmética de menor precisión para aumentar el rendimiento y acelerar el cálculo (ICCSI, s.f). "Dentro del deep learning, los aceleradores permiten realizar interferencias aceleradas mediante GPU y VPU que combinan el alto rendimiento con la eficiencia energética y el soporte duradero requerido para las aplicaciones de inteligencia artificial edge" (QNV, s.f). Esta y otras aplicaciones como el procesamiento de señales y comunicaciones y la computación en la nube se tratarán más adelante.

Esta investigación tiene como objetivo abordar el impacto que tienen las arquitecturas modernas de computadores y los aceleradores en cuanto al rendimiento, la eficiencia y demás factores primordiales en los sistemas computacionales, con un enfoque en las aplicaciones de la computación actual y las funciones que cumplen dentro de las mismas, la influencia de la computación heterogénea, entre otros. Para lograr este propósito, se establecieron los siguientes objetivos específicos a manera de resumen:

1) Analizar la evolución histórica de las arquitecturas de computadores y los aceleradores, priorizando las innovaciones llevadas a cabo en los diseños actuales.
2) Investigar los desafíos actuales en el diseño e implementación de arquitecturas modernas y aceleradores.
3) Examinar el impacto de los aceleradores en áreas de aplicación como la inteligencia artificial, el procesamiento de big data y la computación de alto rendimiento.
4) Explorar las tendencias emergentes en el campo de las arquitecturas y aceleradores y su potencial de aporte en el futuro de la computación

El documento está dividido en nueve secciones compuestas por resumen, abstract, introducción, fundamentos teóricos, estado del arte, metodología, análisis y discusión, casos de estudio, y finalmente conclusiones.

II. Fundamentos Teóricos

A. Definición de Arquitectura de Computadores

La arquitectura de computadores se refiere a la manera en la que se organizan los equipos informáticos de forma lógica. Por otra parte, abarca la microarquitectura de la CPU, determina el rendimiento, las capacidades y los límites que posea el ordenador (Isacc, 2022). Según algorino (s.f) "La arquitectura de una computadora establece la forma en que los componentes de hardware, como el procesador, la memoria RAM, la tarjeta gráfica y la placa base, se interconectan y comunican para cumplir con los requerimientos de rendimiento y eficiencia, manteniendo un balance entre coste y funcionalidad."

De igual manera, cabe enunciar que la arquitectura de computadores no solamente garantiza el diseño de sistemas óptimos y eficientes que incrementan el rendimiento y reducen los recursos que se requieren, sino que permite marcar la diferencia en cuanto a velocidad de procesamiento, capacidad de respuesta y consumo de energía si se encuentra bien estructurada respecto a otras arquitecturas (Tiffin University, 2023).

Entre los principales componentes que debe tener una arquitectura de computadoras se encuentran la CPU encargada de ejecutar las instrucciones de los programas y cumplir la función del cerebro de la computadora, la memoria principal, la unidad de entrada/salida, los buses, entre otros.

B. Definición de Aceleradores

Los aceleradorers de hardware o coprocesadores son diseños específicos creados para trabajar de manera conjunta a un procesador y acelerar una función o carga de trabajo específica. Es debido a esto que los sistemas compuestos tanto por procesadores como por aceleradores, ofrecen una gran cantidad de beneficios en la programabilidad de software que se ejecuta en el procesador, rendimiento y eficiencia energética (Cadence, s.f).

Es aquí cuando es importante introducir el concepto de la computación heterogénea, que principalmente emplea aceleradores CPU, GPU, APU, y DSP para mejorar el rendimiento y la eficiencia energética. Se combina la CPU y la GPU para asignar tareas a los recursos diseñados para manejar dicha tarea (PhoenixNAP, 2022). Según PhoenixNAP (2022) "El beneficio de la computación heterogénea es que aumenta el rendimiento al procesar diversas tareas, como el cálculo avanzado y el procesamiento de imágenes, en paralelo en hardware especializado en lugar de agregar más potencia de procesamiento en bruto."

En general, es correcto afirmar que la computación heterogénea seguirá cumpliendo un papel primordial en la informática o arquitectura de los diseños a futuro, puesto que permite hacer que la computación se eleve constantemente y se resuelvan problemáticas que solían ocurrir en los sistemas homogéneos (Silva, Rueda, Rondón et al. s.f).

C. Historia y Evolución

Comenzando en 1950 con las primeras computadoras ENIAC Y UVIVAC usada con tubos vacío usadas mayormente para el procesamiento de datos cientificos , pasando a 1970 donde se crearon los primeros microprocesadores como intel 4004 el primer microprocesador comercialmente disponible, marcandoel inicio de la era de los microprocesadores y IBM PC la introducción de la computadora personal IBM PC popularizó el uso de computadoras en hogares y oficinas, impulsando la industria del software y hardware.

Desde 1990 hasta la actualidad para complemento de los microprocesadores se crearon los aceleradores graficos y GPUS como NVIDIA lanzó la primera GPU (GeForce 256), diseñada específicamente para acelerar el procesamiento gráfico, lo que revolucionó los videojuegos y aplicaciones gráficas , y posteriormente aceleradores mas personalizados como FPGAs: Los FPGAs (Field-Programmable Gate Arrays) han sido utilizados para tareas específicas que requieren alta velocidad y flexibilidad, como en telecomunicaciones y procesamiento de señales.

Con estos grandes aportes del pasado se han creado estructuras híbridas como los sistemas heterogéneos que son la integración de CPUs, GPUs, FPGAs y otros aceleradores en un solo sistema permite aprovechar las ventajas de cada tipo de procesador para tareas específicas.

D. Clasificación de Aceleradores

Los aceleradores son dispositivos o componentes diseñados para mejorar el rendimiento de ciertas tareas específicas en los sistemas de cómputo. Estos pueden clasificarse de acuerdo con su propósito, flexibilidad, y eficiencia en el procesamiento de datos.

Las CPUs son procesadores de propósito general diseñados para ejecutar una amplia variedad de tareas y programas.

Las GPUs están diseñadas para manejar grandes volúmenes de datos y realizar cálculos en paralelo, lo que las hace ideales para tareas que requieren procesamiento gráfico y computación intensiva, como la renderización de gráficos y el aprendizaje profundo.

Las TPUs están optimizadas para el procesamiento de redes neuronales y tareas de aprendizaje automático. Son diseñadas específicamente para mejorar el rendimiento en aplicaciones de inteligencia artificial

Los ASICs son chips diseñados para realizar una tarea específica de manera muy eficiente. Son altamente optimizados para una función particular y no se pueden reconfigurar.

Los NNAs están específicamente diseñados para acelerar el procesamiento de redes neuronales.

III. Estado del Arte

A. Investigaciones y Desarrollos Recientes

Las arquitecturas específicas de dominio han ganado popularidad debido a la disminución de los retornos de la escalabilidad tecnológica. Estas arquitecturas están diseñadas para aplicaciones emergentes como aprendizaje automático, robótica y centros de datos, y buscan integrar de manera eficiente el hardware y el software para mejorar el rendimiento.En cuanto a los aceleradores exiten tres tipos de hardware dedicados al aprendizaje profundo, como las unidades de procesamiento de tensor (TPUs) y las unidades de procesamiento de gráficos (GPUs),los aceleradores de memoria que es una tendencia emergente que integra capacidades de procesamiento directamente en los chips de memoria para reducir la latencia y el consumo de energía al procesar datos localmente y los aceleradores de sistemas embebidos que están siendo integrados en sistemas embebidos, como smartphones y vehículos autónomos, para mejorar la eficiencia energética y el rendimiento.

B. Aplicaciones Prácticas

- Aprendizaje Automático y Deep Learning:Las GPUs y TPUs se utilizan ampliamente para entrenar modelos de deep learning, como redes neuronales convolucionales (CNNs) para la visión por computadora y redes neuronales recurrentes (RNNs) para procesamiento de lenguaje natural
- Centros de Datos y Computación en la Nube: Los centros de datos modernos utilizan aceleradores específicos de dominio para tareas como análisis de datos, cifrado y descifrado, y procesamiento de transacciones financieras.
- Procesamiento de Señales y Comunicaciones: Los sistemas de telecomunicaciones utilizan aceleradores de hardware para el procesamiento de señales, como la modulación y demodulación de señales, filtrado y compresión de datos.

IV. Metodología

A. Métodos de Investigación

Para recopilar y analizar la información sobre la clasificación de aceleradores en arquitectura de computadores, se ha seguido una metodología estructurada que abarca varios métodos y enfoques. Esto haciendo uso de artículos y papers publicados en fuentes tales como IEEE Xplore, Google Scholar, entre otros.

B. Fuentes de Datos

- Google Scholar: Utilizado para acceder a artículos científicos revisados por pares en diversas áreas de estudio, proporcionando una amplia gama de literatura académica.
- IEEE: Consultado para acceder a conferencias, revistas y reportes técnicos en el campo de la ingeniería eléctrica y electrónica, incluyendo avances significativos en arquitecturas de hardware y aceleradores de procesamiento.
- ScienceDirect: Empleado para conseguir artículos de revistas científicas y contenido académico sobre los aceleradores de hardware y las arquitecturas de aceleradores.
- Biblioteca virtual UIS: Consultado para realizar búsquedas de libros, investigaciones científicas, revistas, artículos en línea, entre otros, sobre las arquitecturas y los aceleradores.
- Navegadores web: Utilizado para acceder a definiciones, características y detalles sobre las arquitecturas de computadores, aceleradores y sus áreas de aplicación.
- NVIDIA blackwell: Consultado para obtener más información de nuevos avances en desarrollo de arquitectura y aceleradores.
- Ashutosh Mishra: Consultado para tener datos sobre las nuevas tecnologias con inteligencia artificial y aceleradores de hardware en SpringerLink, 2019.

V. Análisis y Discusión

A. Comparación de Tecnologías

La elección de la tecnología de acelerador adecuada depende en gran medida de las necesidades específicas de la aplicación y de las prioridades en términos de rendimiento, consumo de energía, costo y flexibilidad. Mientras que las CPUs y GPUs ofrecen una gran versatilidad y capacidad para manejar una amplia gama de tareas, los ASICs y TPUs proporcionan un rendimiento superior para aplicaciones especializadas. Los FPGAs y DSPs ofrecen una combinación de eficiencia y flexibilidad para tareas específicas, aunque con ciertas limitaciones en cuanto a la generalidad y la facilidad de desarrollo.

TABLE I
Comparación de Tecnologías de Aceleradores

Acelerador	Arquitectura	Flexibilidad	Paralelismo
CPU	Propósito general	Alta	Limitado
GPU	paralelismo masivo	Media	Muy alto
TPU	Optimizada para IA	Baja	Alto
FPGA	Reconfigurable	Alta	Variable
ASIC	No reconfigurable	Muy baja	Alto
NNa	Para redes neuronales	Baja	Muy alto
DSP	Para señales digitales	Media	Alto

B. Ventajas y Desventajas

CPU: Tiene una gran versatilidad para manejar una amplia variedad de tareas. Desventaja: Rendimiento limitado en procesamiento paralelo intensivo.

GPU: Excelente rendimiento en tareas de procesamiento paralelo, ideal para gráficos y aprendizaje profundo. Desventaja: Alto consumo de energía.

TPU: Rendimiento superior y eficiencia energética en tareas de inteligencia artificial. Desventaja: Limitada a aplicaciones específicas de IA.

FPGA: Alta flexibilidad y reconfigurabilidad para tareas específicas. Desventaja: Complejidad en la programación y configuración.

ASIC: Máximo rendimiento y eficiencia energética para tareas especializadas. Desventaja: No es flexible y tiene un alto costo de desarrollo.

NNa: Alta eficiencia en el procesamiento de redes neuronales. Desventaja: Limitada a tareas de redes neuronales y no adecuada para otras aplicaciones.

DSP: Alta eficiencia en el procesamiento de señales digitales. Desventaja: Flexibilidad limitada a aplicaciones de procesamiento de señales.

C. Rendimiento y Eficiencia

Los aceleradodes se sometieron a pruebas de aprendizaje automatico y redes neuronales, se midió tanto la energía consumida, cómo la velocidad de procesamiento:

Tensor Processing Units (TPUs) Rendimiento: Hasta 30 veces más rápido que GPUs en algunas tareas de aprendizaje automático. Eficiencia: Eficiente en energía debido a su arquitectura optimizada para tareas específicas de redes neuronales.

Graphics Processing Units (GPUs) Rendimiento: Alta capacidad de procesamiento paralelo, con hasta 5120 CUDA cores y 640 tensor cores. Eficiencia: Consumen más energía que TPUs, pero ofrecen alto rendimiento en aplicaciones de aprendizaje profundo y cálculos paralelos.

Central Processing Units (CPUs) Rendimiento: Procesamiento general y versátil, con velocidades que varían significativamente según la carga de trabajo y configuración. Eficiencia: Menos eficiente en términos de rendimiento por vatio comparado con TPUs y GPUs especializados para tareas específicas como IA y aprendizaje profundo.

Que ciertos aceleradores sean mejores que otros en ciertos casos específicos solo demuestra la especificidad de cada uno de ellos.

VI. Casos de Estudio

A. Ejemplos Prácticos

Empresas como Microsoft utilizan FPGA para acelerar el procesamiento de grandes volúmenes de datos en sus centros de datos.

Microsoft ha integrado FPGA en sus centros de datos para acelerar el procesamiento de bases de datos, especialmente para consultas complejas y análisis de datos en tiempo real. Los FPGA se programan para ejecutar algoritmos específicos de manera paralela y con alta eficiencia energética, uno de los beneficios mas visibles que ha tenido Microsoft es mejorar el rendimiento de sus servicios en la nube, reducir los tiempos de respuesta en consultas de datos complejas y aumentar la eficiencia operativa de sus centros de datos.

B. Resultados Observados

- Mejora en la Precisión y Eficiencia: La capacidad de procesamiento paralelo de las GPUs ha permitido a Google mejorar la precisión de sus modelos de IA al entrenar con conjuntos de datos más grandes y complejos. Esto se traduce en una mejor capacidad para entender y traducir lenguajes con mayor precisión y contexto.
- Escalabilidad: La escalabilidad de las soluciones basadas en GPU permite a Google manejar grandes volúmenes de datos y servicios a millones de usuarios simultáneamente, manteniendo tiempos de respuesta rápidos y eficiencia operativa.
- Reducción del Tiempo de Entrenamiento: Gracias al uso de GPUs Tesla V100, Google ha logrado reducir significativamente el tiempo necesario para entrenar modelos de redes neuronales. Por ejemplo, el tiempo de entrenamiento de modelos complejos para traducción automática en Google Translate se ha reducido de semanas a días o incluso horas.

VII. Conclusiones

A. Resumen de hallazgos

En conclusión, se realizó una investigación furtiva sobre el tema principal de arquitecturas y aceleradores enfocado en la arquitectura de computadores, donde inicialmente se empezó por otorgar una definición, al igual que una contextualización

a través de la historia, evidenciando la evolución de ambos conceptos.

Además del análisis de las investigaciones realizadas por diversos autores y compañías, que han ido ganando cierta popularidad con el desarrollo de la computación actual, donde se destaca que es gracias a la integración de aceleradores en el diseño de las arquitecturas modernas, en donde la CPU trabaja en simultaneo con un acelerador, la mejora en variables como el rendimiento, la eficiencia energética, la flexibilidad y la adaptabilidad, cabe añadir que dentro del análisis y discusión en donde se comparan las tecnologías de aceleradores, sin contar el sinfín de ventajas que tienen para el desarrollo de tareas y el procesamiento, también se presentan desventajas en cuanto al rendimiento limitado, alto consumo energético, limitación en su aplicación en áreas de interés, complejidad en la programación, entre otros.

No obstante, el uso de aceleradores es un tema que aún se encuentra en desarrollo, es decir, que poco a poco las empresas y usuarios interesados en emplear dichas tecnologías deberán enfrentarse a las adversidades que trae consigo el futuro de la computación, optimizando y mejorando las técnicas y estudios que empleen.

Teniendo en cuenta esto, la elección de la tecnología de acelerador adecuada depende en gran medida de las necesidades específicas de la aplicación y de las prioridades en términos de rendimiento, consumo de energía, costo y flexibilidad.

B. Implicaciones Prácticas

Como se evidenció en los casos de estudio, mejora la escalabilidad , rendimiento y velocidad, es decir que en conjunto, las arquitecturas y aceleradores representan herramientas fundamentales para mejorar el rendimiento, la eficiencia y la capacidad de innovación en el ámbito de la informática. Su adopción y uso adecuados pueden proporcionar ventajas competitivas significativas a las organizaciones que buscan aprovechar al máximo las capacidades tecnológicas modernas.

C. Futuras Direcciones

Finalemente, dentro de las futuras direcciones se destacan las siguientes:

Optimización de Eficiencia Energética: Investigar métodos para mejorar la eficiencia energética de GPUs, TPUs y otros aceleradores sin comprometer el rendimiento. Esto incluye técnicas de administración de energía, diseño de circuitos de bajo consumo y uso de materiales y procesos más eficientes.

Mejora de la Flexibilidad en ASICs y FPGAs: Explorar técnicas para aumentar la flexibilidad y reconfigurabilidad de ASICs y FPGAs sin perder el rendimiento especializado. Esto podría incluir el desarrollo de herramientas de diseño más accesibles y la exploración de arquitecturas híbridas que combinen ASICs con capacidades de reconfiguración parcial.

Integración de Aceleradores en Sistemas Embebidos: Investigar cómo integrar de manera eficiente aceleradores como TPUs y DSPs en sistemas embebidos para aplicaciones IoT y dispositivos móviles. Esto podría implicar miniaturización, optimización de consumo de energía y diseño de interfaces de programación más amigables.

Seguridad en Aceleradores de IA: Explorar técnicas para mejorar la seguridad y la privacidad en el procesamiento de datos en aceleradores utilizados en aplicaciones de inteligencia artificial. Esto incluye la protección contra ataques de adversarios y el desarrollo de algoritmos y protocolos robustos de seguridad.

REFERENCES

[1] John W. Mauchly y J. Presper Eckert, *Electronic Numerical Integrator and Computer (ENIAC)*, 1945.

[2] J. Presper Eckert y John Mauchly, *UNIVAC I (Universal Automatic Computer I)*, Remington Rand, 1951.

[3] John Bardeen, Walter Brattain y William Shockley, *The Invention of the Transistor*, Bell Labs, 1947.

[4] Jack Kilby, *Miniaturized Electronic Circuits*, Texas Instruments, 1958.

[5] Federico Faggin, Marcian Hoff, Stanley Mazor y Masatoshi Shima, *Intel 4004 Microprocessor*, Intel, 1971.

[6] Mark Dean, Philip Don Estridge y William C. Lowe, *IBM Personal Computer*, IBM, 1981.

[7] NVIDIA Corporation, *GeForce 256 GPU*, 1999.

[8] NVIDIA Corporation, *Compute Unified Device Architecture (CUDA)*, 2006.

[9] Stephen Brown y Jonathan Rose, *Field-Programmable Gate Arrays*, Kluwer Academic Publishers, 1992.

[10] Norman P. Jouppi et al., *In-Datacenter Performance Analysis of a Tensor Processing Unit*, Google, 2017.

[11] Ashutosh Mishra, et al., *Artificial Intelligence and Hardware Accelerators*, SpringerLink, 2019.

[12] Yeji Kang, *Understanding the Implications of Current AI Hardware*, Harvard University, 2023.

[13] NVIDIA, *NVIDIA Blackwell Platform Arrives to Power a New Era of Computing*, NVIDIA Newsroom, 2024.

[14] A. Shahid and M. Mushtaq, "A Survey Comparing Specialized Hardware And Evolution In TPUs For Neural Networks," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Bahawalpur, Pakistan, 2020, pp. 1-6.

[15] Instituto Nacional de Aprendizaje, *Generaciones de las computadoras (1.a ed.)*.

[16] Inteligencia, *Aceleradores de ia: rendimiento y eficiencia en tecnología*, ICCSI, 2023, 21 mayo.

[17] M, G., *Acelerador de Deep Learning*, QNV Solutions, 2019, 8 octubre.

[18] Isaac, *Arquitectura de computadoras: ¿Qué son? ¿Cómo funcionan?*, Profesional Review, 2022, 24 agosto.

[19] Arquitectura de computadoras: modelos esenciales, Tiffin University, 2023, 8 marzo.

[20] Fundamentos de la arquitectura de computadoras, Algor Cards, s.f.

[21] Cadence, *Hardware Accelerator*.

[22] Silva Garces, J. F. S. G., Rueda Serrano, J. C. R. S., & Rondon Arango, J. S. R. A., *Computación Heterogénea Y su Gran Auge en los Últimas Décadas*, Universidad Industrial de Santander, s.f.

[23] Patel, S. J. P., & Hwu, W. W. H., *Accelerator Architectures*, IEEE Micro, 2008, agosto.

[24] Valdez, R. V., & Maldonado, Y. M., *Computación heterogénea y FPGAs como aceleradores eficientes*, Memorias de la Décima Segunda Conferencia Iberoamericana de Complejidad, Informática y Cibernética, 2022.

[25] TecnoDigital, & TecnoDigital, *Arquitectura Computadoras: Introducción a su Evolución y Diseño*, Informática y Tecnología Digital, 2024, 27 abril.

[26] Sheldon, *Una visión de la computación acelerada*, Comunidad FS, s.f.