A Comparative Study of Digital Hardware Acceleration Techniques for AI Workloads

Karthik Wali

ASIC Design Engineer
ikarthikw@gmail.com

Abstract

To be more precise, remarkable progress in the development of AI and ML has resulted in high requirements for modern hardware architecture. The traditional CPUs are not efficient enough to handle these performance characteristics of artificial intelligence; thus, hardware accelerators such as GPUS FPGAs and ASICs have come to the forefront. In terms of power efficiency, processing ability, and flexibility, all of these technologies come with some advantages, which are listed below. This paper aims to discuss and compare different techniques in the acceleration of AI-based workloads in hardware. We explain how they are different from one another architecturally, their performances, and how the different methods can be useful in different architectures of AI. The most important figure of merit evaluated include computation throughput, power consumption, latency, and scalability. The paper also covers tendencies that have not been established yet, like neuromorphic computing and acceleration with quantum computers and their role in creating the future of AI processing. It can be beneficial to AI researchers and engineers to consider and choose the proper acceleration technique according to the application needs.

Keywords: AI acceleration, GPUs, FPGAs, ASICs, Machine learning, Neuromorphic computing, Quantum acceleration

1. Introduction

Artificial Intelligence (AI) has been at its peak of development regarding growth during the past few years. It is gradually becoming part of numerous sectors like healthcare finance, along with an increase in the use of autonomous systems. However, with the progressive development of AI models, their computational complexity has grown enormously high and needs appropriate hardware solutions. [1-4] CPUs have been observed to perform poorly in deep learning tasks because of the lack of capabilities in parallelism. This has given rise to the use of hardware acceleration techniques such as the use of GPUs, FPGAs, and ASICs where each possesses its benefits in the processing of AI. It is important to weigh between each of these parameters as they make significant impacts on AI workloads. The goal of this work is to compare these accelerators in order to determine which of the hardware options is the most appropriate for various applications of AI.

1.1. Need for Hardware Acceleration in AI

With the current developments in Artificial Intelligence (AI), there has been an increase in the number of computations needed to be done and thus the need for faster computations. Even regular processors, including the CPUs, are unable to process the vast parallelism required for machine learning tasks. It has given rise to the creationof specific hardware accelerators such as GPU, FPGA, and ASIC that are well-

suited for AI applications. The following are the major factors that make hardware acceleration crucial for artificial intelligence.

- **Limitations of General-Purpose CPUs:** The CPUs are basic computing cores that are optimized and specialized for non-parallel data processing and are thus not ideal for AI applications that involve a lot of matrix computations as well as deep learning training. Their base count is low and has low parallelism,making them slow when processing information. The energy consumed to power them is relatively high, hence unfit for modern AI uses.
- **Parallel Processing Requirements in AI:** Parallel operations like convolutions, matrix multiplications, and backpropagation are the key components deep learning networks use since they are AI models. Data processing or Matrix operations like multiplication, addition, etc., are much faster on Hardware accelerators like GPUs and TPUs due to the thousands of cores present in them.
- **Power Efficiency and Energy Constraints:** AI functions such as edge computing and data centers need chips that have high computational performance and power efficiency. FPGAs and ASICs propose a more efficient utilization of power per computation as they cut costs on power consumption and, therefore, on the battery consumption on AI-based devices like self-driving cars, smartphones, and more.

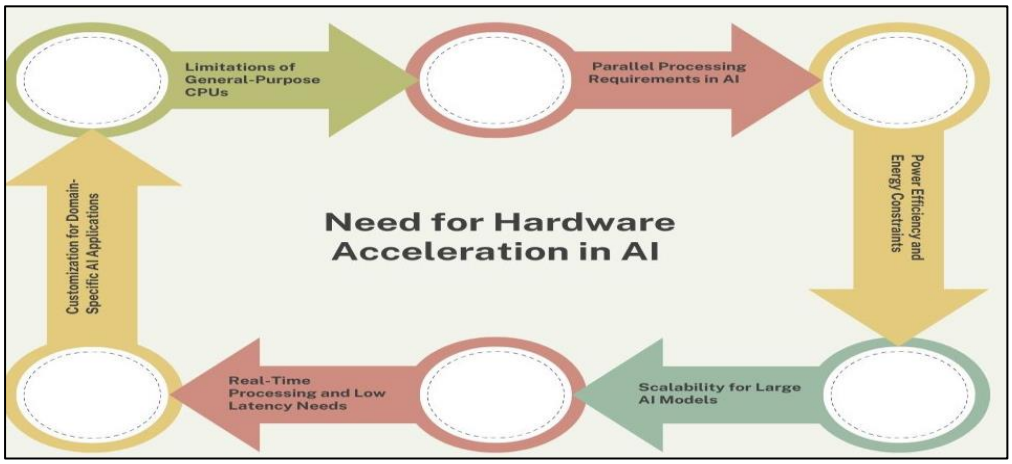

Figure 1: Need for Hardware Acceleration in AI

- **Scalability for Large AI Models:** Modern models, including GPT-4, BERT, and DALL·E, for example, consume a significant amount of data during processing beyond the capacities of a single CPU. Multi-chip and multi-node scalability are supported by hardware accelerators so that the AI workloads can be run on multiple devices to offer improved performance for training and inference.
- **Real-Time Processing and Low Latency Needs:** Some examples of such applications that use AI include autonomous cars, robots and even stock exchange trading. On Topic 2,between ASICs and FPGAs, the dedicated and reconfigurable architectures of the devices that are employed result in high processing speed, hence a shorter delay towards inference and decision-making in AI.
- **Customization for Domain-Specific AI Applications:** AI workloads differ by area of application, and thus, the accelerations are optimized suits. FPGA is reprogrammable and useful for implementing logic that changes during the operation of the hardware, while ASIC is a single-purpose chip designed to execute HE HLL tasks such as image or speech recognition, and in diagnosing medical conditions. To this end, these problems allow AI models to perform optimally and with a high level of accuracy for their respective purposes.

1.2. Evolution of AI Hardware

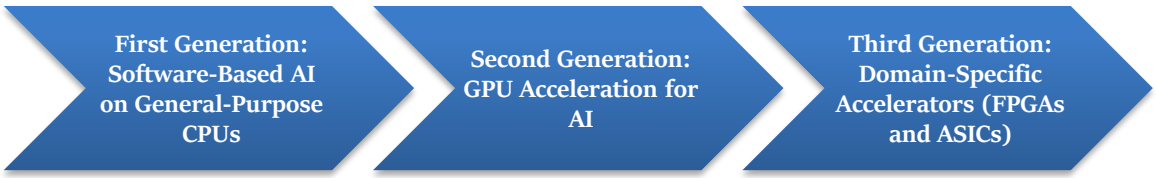

Figure 2: Evolution of AI Hardware

- **First Generation: Software-Based AI on General-Purpose CPUs:** The first generation of AI hardware was derived from general-purpose CPUs where AI models were executed in accustomed software solutions. [5,6] Then, CPUs, good for sequentially handling tasks, were inadequate to handle the computational needs of AI, especially in scenarios that required many matrix operations and deep learning. Although they had these shortcomings, CPUs found a great deal of use because of their accessibility, the comparative ease with which they could be programmed, and their compatibility with the existing software. However, as the models of AI become more complicated, then it was realized that there is a need for better hardware.
- **Second Generation: GPU Acceleration for AI:** The second generation was more revolutionary than evolutionary since, for the first time, it was using GPUs for AI acceleration. While CPUs have a single or few cores, GPUs have thousands of small cores that are ideal for AI workloads such as deep learning, computer vision, and NLP. Frameworks such as NVIDIA CUDA and AMD ROCm gave the possibility to use the computing power of GPUs for AI and enhanced training speed. This has paved this generation towards new advancements and innovations in AI and impacting areas such as the development of self-driving cars, healthcare, and big data.
- **Third Generation: Domain-Specific Accelerators (FPGAs and ASICs):** The third generation of AI hardware included specific FPGAs and ASICs to enhance the efficiency of AI computing. FPGAs are also reprogrammable; hence, the platforms can be designed to match the particular AI model being developed, aiding in cutting down on power and increasing efficiency. Whereas, ASICs are completely optimized AI processors like Google TPU and Huawei's Ascend, which are built for the highest performance and also power consumption. These accelerators perform better than GPUs, especially in the AI inference use case, and are applicable for the data centre, cloud, and edge AI. This generation is the peak of optimizing AI hardware and opening the next level of computing.

2. Literature Survey

2.1. General-Purpose CPUs for AI

Legacy processors are as effective as they were before, with substantial versatility and overall utility in computing. However, there is a large inconsistency in regard to their efficiency when it comes to AI-focused workloads and, in particular, those applications that entail parallelism and heavy matrix computations, including matrix multiplications and deep learning model training. Modern CPUs are epitomized by a small number of cores optimized for sequential computing, which restricts the ability to involve a large number of cores for AI tasks that require tremendous parallelism. [7-11] Nevertheless, current mainstream CPU architectures such as the Intel Xeon Scalable Processors equipped with the Intel DL Boost and the AMD EPYC second generation supporting the AMD AVX-512 cannot yet compete with domain-specific

hardware including GPUs, FPGAs, and ASICs the way accelerators are defined in this paper. Cores have been observed trailing in terms of AI performance, especially in areas such as training neural networks and real-time inference, as well as, making cores less suitable in large-scale AI developments.

2.2. GPU Acceleration in AI

GPUs are the primary accelerators for AI solutions and workloads since they have an impressive level of parallelism. Unlike its counterpart CPUs, GPUs involve a large number of smaller high-speed cores that work parallelly and, thus, are very suitable for complex operations like deep learning and image and language analysis. NVIDIA has provided a CUDA framework, and AMD has offered ROCm,which is the set of libraries and tools to accelerate AI models on GPU computing. These allow major rate enhancements in training large neural networks with tensor cores, memory structures, and high computation rates. Therefore, they remain the prevalent hardware in the field of artificial intelligence, both at cloud computing services and HPC environments.

2.3. FPGA-based AI Acceleration

The usage of FPGAs is beneficial in AI acceleration since FPGAs are hardware devices that can be reconfigured to suit certain tasks. It is worth mentioning that in configuring FPGAs, the approach to the arrangement of the data flow is flexible, unlike fixed-function processors, including GPUs and CPUs. This flexibility is very beneficial for large performance improvements, especially in the field of Artificial intelligence Inference, where latency and energy usage greatly impact the feasibility of the application. He further said that more formidable corporations like Intel and Xilinx (which is now a part of the Advanced Micro Device) have unveiled FPGA solutions, for instance, Intel's Stratix and Xilinx's Versal in the current past that provide higher performance in the execution of AI Models coupled with the minimal power intensity as compared to processors. However, FPGAs call for unique hardware programming skills. Nevertheless, frameworks, including the Vitis AI or OpenCL, promote FPGA-based AI acceleration among developers.

2.4. ASIC-based AI Acceleration

Application Specific Integrated Circuits (ASICs) provide the maximum density of optimized utilization for specific AI tasks since they are tailored for them. ASIC stands for application-specific integrated circuits, and they are different from conventional types of processors as they are basically intended to carry out specific functions with high efficiency and power-saving capacity. TPU, used by Google, and Ascend chips, developed by Huawei, can be considered examples of AI-oriented ASICs that have revolutionized deep learning acceleration. These chips perform matrix multiplications and tensor operations more effectively and efficiently than general-purpose CPU or graphic processor units. They are highly beneficial for high-volume AI inferencing that is recurrently used in programs like real-time voice sensitivity and recommend ability systems. Still, the high development cost and low configurability compared to FPGAs or GPUs make the ASICs applicable mainly to large-scale implementation and specific applications.

2.5. Emerging Technologies

Apart from the current wave of accelerators, the next-generation prospective accelerators like neuromorphic computing and quantum acceleration are in the process of development in order to enhance AI hardware. Neuromorphic computing is the technique in machine learning where circuits, as well as their functions, mimic that of the biological neural networks and, therefore, bring about efficient artificial intelligence processing with less energy consumption. Devices like Intel Loihi and IBM TrueNorth are example programs of neuromorphic processors that rely on spiking neural networks for real-time learning and

inference operations that consume less power. Quantum computing, on the other hand is another paradigm shift that proposes to use quantum mechanics in computation to solve problems by orders of magnitude faster than in classical computers. Quantum AI has the possibility of performing optimization and probabilistic tasks essentially beyond present-day Information technology frameworks. Despite the fact that all these technologies are at their burgeoning stage, they hold substantial potential for further development, and the discovery of possible paradigm shifts in driving artificial intelligence could trigger better advancements more prominently in areas of pharmaceutical discoveries, cryptography, and modeling of complicated systems.

3. Methodology

3.1. Comparative Framework

- **Computational Throughput (TFLOPS):** In some cases, based on the number of floating-point operations per second, which is measured in TeraFLOPS, it is called computational throughput. As discussed earlier, deep learning models and other delicate AI workloads are CPU-intensive and involve numerous matrix multiplications and tensor computations, meaning that throughput will be the most critical metric. [12-16] GPUs and ASICs provide the highest TFLOPS since the main architectures are highly parallel and consist of the AI processing cores, including tensor and systolic. On the other hand, general-purpose CPUs provide lower computational throughput because they are oriented mainly on configurability as compared to parallelism. Although providing field-programmable architecture that can be reconfigured dynamically, FPGAs can be a little less powerful than GPUs or other ASICs. Still, they can be very effective when applied to particular AI tasks.

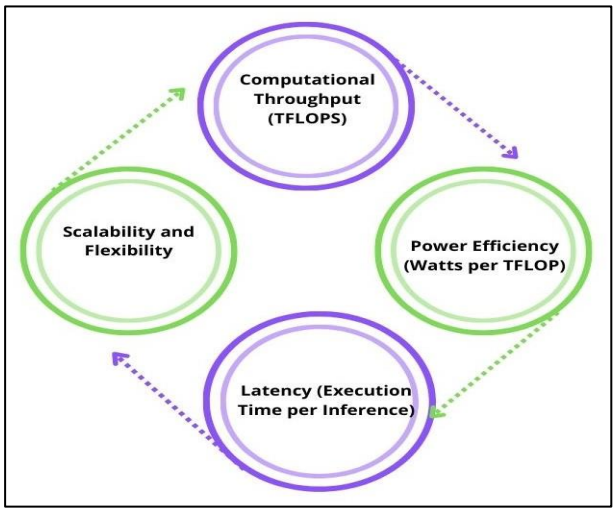

Figure 3: Comparative Framework

- **Power Efficiency (Watts per TFLOP):** Energy consumption per operation measured in watts per TFLOP is of high importance for AI workloads to improve throughput in data centers and edge devices. Less amount of power consumed per computing process means reduced costs and more sustainability for the system. Google TPU and Huawei's Ascend follow the same ideology and are really power-efficient as these ASICs are custom-built for AI calculations with no extra hardware besides an I/O interface. However, existing GPUs are generally-purpose and, as such consume much power despite their high performance. As a result, current CPUs are generally less power efficient than specialized processors for AI loads because they don't contain parallelism functions. FPGAs,

hence, provide ways of customizing a specific application and performance as well as power and energy efficiency.

- **Latency (Execution Time per Inference):** Latency is very important in AI-based systems, especially in applications involving real-time flows like self-driving cars, robots, and conversational AI. It captures the amount of time for an inference procedure starting from processing the input through the inference process to generation of an output. There is a great demand to have low latency to process the data because information- based applications need quick responses like speech recognition and fraud detection. They have lower latency than GPUs since ASICS are designed to perform small specific operations associated with AI. While GPUs are powerful, they might introduce some delays due to data transfer and scheduling costs. FPGAs can be optimized in a way that they have minimized latency for specific tasks and are thus great for real-time AI applications. CPUs, however, require higher latency as the tasks run sequentially as opposed to in parallel.
- **Scalability and Flexibility:** Scalability and flexibility define the ability of the hardware solution to increase the imposed load and accommodate multiple AI models. These models are also scalable since they support distributed computing and are well suited for the multi-GPU clusters for large-scale deep learning. Although ASICS provide high efficiency, they are not programmable since they are built for particular workloads and need redesigning for the new models of AI. FPGAs mean that they are reusable, hence allowing the application of reconfiguration in AI to deal with the continually changing AI algorithms; however, they are challenging to program. CPUs are still the most versatile because they are intended for any type of computing, but they don't efficiently expand for AI applications. In most cases, it can be concluded that the choice between the computer hardware is determined by such factors as: scalability capacity and flexibility and the computational requirements of the specific program.

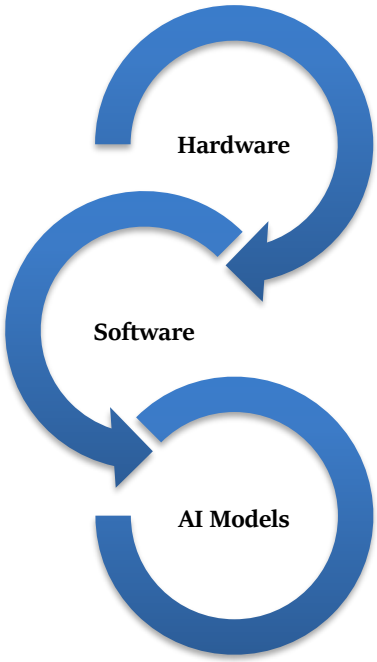

Figure 4: Experimental Setup

3.2. Experimental Setup

- **Hardware:** The configuration of experiments involves three different AI accelerators to assess their performance in various AI tasks. NVIDIA RTX 3090 is a powerful GPU having 10496 CUDA cores with 24GB GDDR6X memory and is designed for deep learning through tensor cores and mixed precision. It sets the standard of what consumers should expect of GPU when training deeper neural networks and inferencing them. The Intel Stratix 10 FPGA gives an opportunity to deploy the system with Turnkey AI acceleration pipelines on programmable hardware at runtime. As a flexible architecture with large memory bandwidth, it is evaluated on the inference task for real-time and low power consumption. The Google TPU v4 is built to perform especially for deep learning tasks. It provides great performance for both training and inference and ultimately improves the scale of matrix multiplication through the systolic array configuration. These three types of hardware platforms give good coverage for testing the efficiency of AI acceleration based on different architectural styles.
- **Software:** These software tools are available and widely used in the industry among other FPGA AI frameworks among other FPGA development tools. TensorFlow and PyTorch are deep learning frameworks that contain pre-optimized versions of AI models and offer hardware accelerators. Tensorflow has an excellent compatibility with TPUs to execute the model on Google's AI hardware while Pytorch is preferred for research purposes and using GPUs and CUDA. For making use of FPGA acceleration, the selected tool is Xilinx Vitis,which supports high-level synthesis tools and hardware-based AI model deployment. This design allows for a fair comparison of the AI models since the software optimizations are taken into consideration in their best forms for the architectures of the selected hardware.
- **AI Models:** To evaluate the above-mentioned selected hardware, three different AI models are chosen that cover some of the most important application areas in the field of deep learning. In this research, ResNet-50, which is a deep convolutional neural network, is utilized for image classification tasks and measures the throughput and inference speed on various accelerators. BERT (Bidirectional Encoder Representations from Transformers) is an NLP model based on the transformer that measures how each of the platforms under tests handles sequential data and large model training. YOLOv4 is an object detection model that is intended to be used in real-time applications, and therefore, it needs high throughput for real-time applications to match its low power consumption. These models mean a balanced and diverse set of benchmarks as well as vision and language models, together with real-life inference cases, to evaluate the pros and cons of each accelerator.

3.3. Benchmarking Metrics

As for the difference in AI accelerators' performance, three assessments are made for each accelerator Architecture: Computational Throughput (TFLOPS), Power Consumption (W), and Latency (ms). These metrics help to understand the power, performance, and energy consumption of GPUs, FPGAs as well as ASICs in terms of applying AI computations. Specifically, computational throughput is a measure of the number of teraflops (TFLOPS) that a specific accelerator can perform in realtime. A higher number of TFLOPS denotes the capability of training and performing inferences of artificial intelligence problems more efficiently. Currently, NVIDIA has released the new RTX 3090 GPU, which operates at 35 TFLOPS due to the thousands of CUDA cores and tensor cores. Indeed, the Intel Stratix 10 FPGA offers 10 TFLOPS, although it is a bit behind the GPU in terms of purely scientific computation real estate; this has been made deliberately to offer the FPGA a lot of flexibility since it is, after all, reconfigurable. Google TPU v4, once

again, is ahead with 120 TFLOPS, much ahead of both the GPU and FPGA, because it's designed for deep learning and makes use of matrix multiplication units for throughput. Power consumed in watts (W) is another aspect important to consider in order to decide the efficiency of an AI accelerator.

Actual energy consumption per millions of operations calculated in TFLOPS is preferably lower, the better because significant portions of such systems are often employed in large-scale data centers increasing operational costs through the consumed energy. Thus, the NVIDIA RTX 3090 GPU has a power consumption of 320W, which is not the highest for the given category since general-purpose GPUs are not designed with power efficiency in mind. The actual power consumption of Intel Stratix 10 FPGA is 80 W, which is much lower than that of the GPU, as FPGAs are intended to provide flexible and efficient processing capabilities. The design of Google's TPU v4 co-processor shows that its power draw of 40W is highly efficient since the chip was designed specifically for AI applications and does not feature redundant components that are unnecessary for this type of task.Latency represents the time required to perform a single operation, and it has units in milliseconds (ms only). This is because low latency is critical for the application of AI for real-time use cases such as self-driving cars, speech recognition, and edge computing. The latency of NVIDIA RTX 3090 is approximately 12ms high, which is caused by memory access time, compute operations, and others. This comes from the optimized flow of data and the design of hardware, which the Intel Stratix 10 FPGA boasts of 8ms. The Google TPU v4 stands out with a low latency of 3ms due to the fact that the chip is optimized for AI processing, thereby causing fewer delays and being the most suitable variant when it comes to real-time operations.

4. Results and Discussion

4.1. Performance Analysis

The performance of GPUs, FPGAs, and ASICs is compared based on Computational Throughput (TFLOPS), Power Consumption (W), and Latency (ms).

Table 1: Performance Comparison of AI Accelerators

Metric	GPU	FPGA	ASIC
Computational Throughput (TFLOPS)	35	10	120
Power Consumption (W)	320	80	40
Latency (ms)	12	8	3

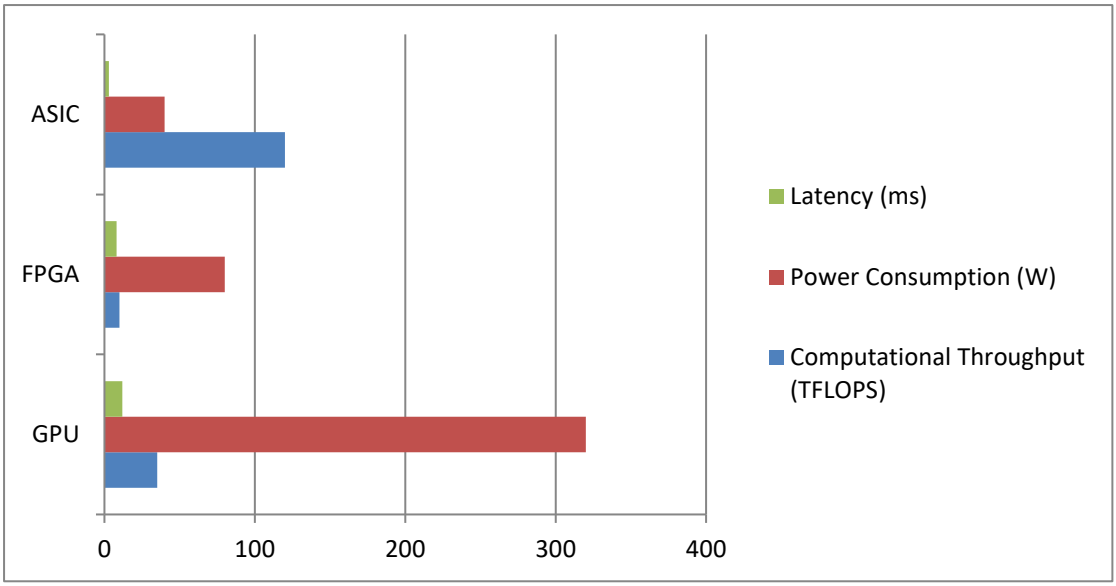

Figure 5: Graph representing Performance Comparison of AI Accelerators

- **Computational Throughput (TFLOPS):** Computational throughput counts the number of floating-point operations per second; it defines the capacity for artificial intelligence calculations. The latest Google TPU version, TPU v4, has a performance level of 120 TFLOPS, thereby leaving behind two other popular computational processing devices, which are the NVIDIA RTX 3090 GPU that has 35 TFLOPS and the Intel Stratix 10 FPGA that has 10 TFLOPS. This demonstrates that ASICs perform well for AI computations as compared to GPUs, which are still competitive for general AI computations and FPGAs are moderate and reconfigurable.
- **Power Consumption (W):** In this regard, power consumption remains a major consideration when it comes to the choice of AI hardware to be implemented. The Google TPU v4 consumes the least amount of power at 40W as opposed to 80W for the Intel Stratix 10 FPGA and 320W for NVIDIA RTX 309 0GPU. Nevertheless, continuous usage increases the power needs many folds. Hence, ASICs are preferred when energy consumption is a concern in AI applications; FPGAs are moderate in their power efficiency needs.
- **Latency (ms):** Latency is the amount of time it takes to complete an AI inference, and the time is described in terms of lower time. In this release, Google TPU v4 sets the record of only 3ms of latency making it suitable for real-time AI. The next one is called Intel Stratix 10 FPGA with latencies at 8ms, and this is because it is easy to modify its hardware logic. While GPUs show excellent promotion in parallelism, they slow down when it comes to inference tasks, in contrast to ASICs that reduce latency.

4.2. Power Efficiency Comparison

Table 2: Power Efficiency (Percentage Comparison)

Accelerator	Percentage Efficiency (%)
GPU	3.67%
FPGA	4.17%
ASIC	100%

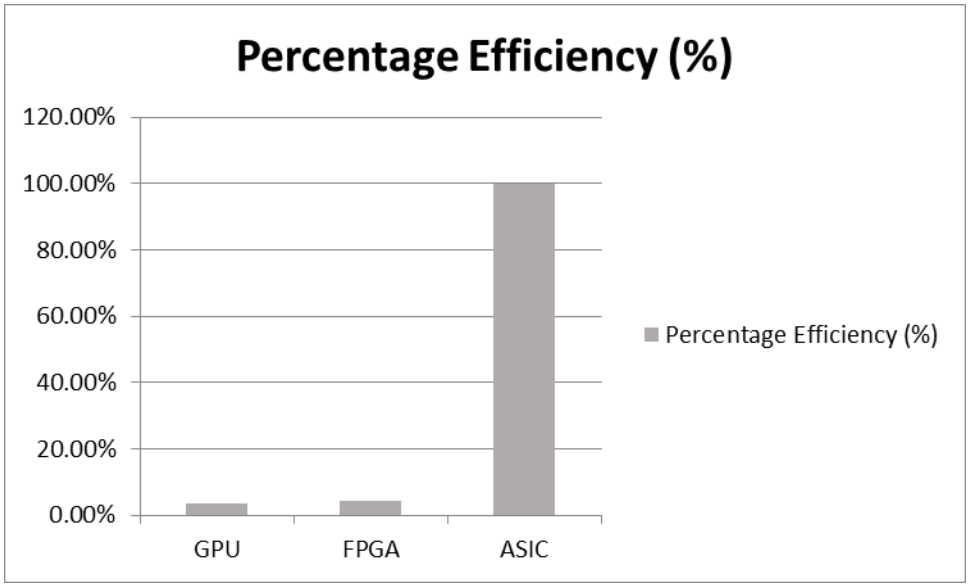

Figure 6: Graph representing Power Efficiency (Percentage Comparison)

- **GPU:** The NVIDIA RTX 3090 was only able to achieve a power efficiency of 3.67 % to that of the ASIC baseline. While being rather efficient in terms of single core and overall computational performance, it currently consumes as much as 320W of power. This is due to the fact that GPUs are made for general computation while being less efficient in terms of energy consumption for AI tasks.
- **FPGA:** The Intel Stratix 10 is also 4.17% more efficient than other FPGA chips, primarily because it is designed for AI acceleration with a power-saving chip architecture. Although, the FPGAs have less power consumption (80W), they are less efficient than the ASICs in terms of Computational Throughput and are also associated with overhead of reconfiguration.
- **ASIC:** The most efficient one is the Google TPU v4 ASIC, in this case, with the perfect efficiency ratio of 100%. ASICs are specifically designed for AI operations and require only 40W while attaining 120 TFLOPS; no more computational efficiency is built in needlessly. This makes them the most suitable for massive AI implementations, as well as they provide high results with low power input.

4.3. Scalability and Adaptability Comparison

- **Scalability:** Scalability then basically speaks of the measure to which an accelerator is capable of holding increasing demands. GPUs tend to be highly scalable and thus, multiple GPUs can be easily interconnected into a cluster for training of an AI model. FPGAs are somewhat scalable, especially in that it is possible to load FPGA with varying content in order to enable it to better handle large volumes of work); however, scaling an FPGA design takes extra time. ASICs have low scalability since they are application-specific processors that cannot be used for diverse purposes since they are optimized for general AI-related tasks.
- **Software Flexibility:** Another important aspect constitutes the flexibility that an accelerator has in its ability to be implemented crosswise with various AI frameworks and applications. GPUs are proficient in this regard as they have standardized libraries such as TensorFlow and PyTorch which helps in easy deployment as well as incorporation of newer models. In contrast, FPGAs used in space have the disadvantage of low software flexibility due to their programming reconfiguration at the hardware level, which is not an easy task. Another disadvantage is the low flexibility: ASIs are designed to work on specific operations where the use of AI is feasible.

- **Adaptability:** Flexibility looks at the capability of the accelerator to transform with AI workloads as they experience changes in their environment. FPGAs have the most flexibility and reconfigurable nature derived from the hardware and bring in the idea that the architecture can be altered according to different modes of AI. Moreover, that flexibility is afforded by the fact that GPUs are sometimes improved through software and driver updates, altering their requirements for AI tasks. ASICs' main drawback is low flexibility, as their hardware is tailored to particular AI tasks. Therefore, they are perfect for certain tasks but ineffective when it comes to various AI applications.

5. Conclusion

In this research work, we discuss how three major types of accelerators, namely GPUs, FPGAs, and ASICs, are fit for various AI applications. Some of the analyzed KPI confirms were computational throughput in terms of TFLOPS as well as power efficiency or consumption in watts per TFLOP, latency, scalability, and flexibility. The outcomes also reveal that GPUs are generally compatible with most varieties of AI tasks and very suitable for training deep learning models because of the power of computation and the software versatility. However, it needs high power and has higher latency in comparison to the other options of accelerators. FPGAs are still the most suitable for AI inference tasks because they offer high performance as well as low power consumption. Due to the adaptability of their hardware design for specifically AI uses, they have higher efficiency than GPUs in power consumption. However, the usage of FPGAs is somewhat complicated and challenging since programming and optimization of FPGAs remains a difficult task.

On the other hand, ASICs outperform both GPUs and FPGAs in terms of computational efficiency and power consumption. ASIC solution provides the best efficiency by means of working on customized circuits depending on a particular AI job, thus consuming less power and providingthe least latency, which makes it the optimal equipment for AI algorithms used in inferenceąp. However, they do not possess flexibility compared to others; the development cost is quite high, and it takes a long time to be introduced in the market, limiting its applicability to large-scale, specific AI implementations. In summary, it can be realized that the workload characteristics inform the decision of the type of AI hardware. GPUs are best suited for several application AI-centered tasks. At the same time, FPGAs are tailored application-specific in addition to offering power efficiency for inference. In contrast, ASICs are best for high speed and low power in additional large-scale applications such as data centres and edge computing.

5.1. Future Research Directions

New trends in AI acceleration also include neuromorphic computing and quantum AI,which may become a breakthrough in AI development as they will offer considerable optimization and acceleration. Based on the structure of the human brain, that is, neuromorphic computing makes use of specialized Spiking Neural Networks (SNNs) and chips such as Intel's Loihi for low-power and real-time AI. This could let the AI models handle information flow as the brain handles as this technology aims at emulating the way the brain learns and transforms dynamically. On a related note, some of the research activities in this sphere are centered on optimizing neural designs, energy consumption, and practical implementations in robotics and edge computing. Another exciting concept is quantum acceleration,which aims to use quantum principles to conduct computations much faster than the existing classical counterparts. It has the capability for optimization, improvement of AI model training, and creation of enhanced algorithms in deep learning. Google, IBM, and D-Wave develop quantum processors for performing AI tasks that cannot be performed with the help of accelerators. Still, there are potential issues that have to be solved in order to make quantum AI a common future: how to stabilize the hardware, how to correct errors, and how to make quantum AI scalable.As such, future studies should focus on neural-accelerative AI, quantum-accelerative AI, and the

blend of neuromorphic and quantum, as well as classical AI accelerations that will form more efficient and adaptive AI systems. Also, the study of low-power AI hardware platforms for edge AI and embedded AI will extend the AI penetration into more smartphone, IoT, and auto systems.

References

1. Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017, June). In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th annual International Symposium on Computer Architecture (pp. 1-12).
2. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
3. Hennessy, J. L., & Patterson, D. A. (2011). Computer architecture: a quantitative approach. Elsevier.
4. Venkatesh, G., Sampson, J., Goulding, N., Garcia, S., Bryksin, V., Lugo-Martinez, J., ... & Taylor, M. B. (2010). Conservation cores: reducing the energy of mature computations. ACM Sigplan Notices, 45(3), 205-218.
5. Guide, D. (2020). Cuda c++ programming guide. NVIDIA, July.
6. Putnam, A., Caulfield, A. M., Chung, E. S., Chiou, D., Constantinides, K., Demme, J., ... & Burger, D. (2014). A reconfigurable fabric for accelerating large-scale datacenter services. ACM SIGARCH Computer Architecture News, 42(3), 13-24.
7. Kuon, I., & Rose, J. (2006, February). Measuring the gap between FPGAs and ASICs. In Proceedings of the 2006 ACM/SIGDA 14th international symposium on Field programmable gate arrays (pp. 21-30).
8. Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE Journal of solid-state circuits, 52(1), 127-138.
9. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.
10. Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., ... & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, *38*(1), 82-99.
11. Marković, D., Mizrahi, A., Querlioz, D., &Grollier, J. (2020). Physics for neuromorphic computing. Nature Reviews Physics, 2(9), 499-510.
12. Preskill, J. (2018). Quantum computing in the NISQ era and beyond. Quantum, 2, 79.
13. Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., ... & Martinis, J. M. (2019). Quantum supremacy using a programmable superconducting processor. Nature, 574(7779), 505-510.
14. Karras, K., Pallis, E., Mastorakis, G., Nikoloudakis, Y., Batalla, J. M., Mavromoustakis, C. X., & Markakis, E. (2020). A hardware acceleration platform for AI-based inference at the edge. Circuits, Systems, and Signal Processing, 39(2), 1059-1070.
15. Chang, H. Y., Narayanan, P., Lewis, S. C., Farinha, N. C., Hosokawa, K., Mackin, C., ... & Burr, G. W. (2019). AI hardware acceleration with analog memory: Microarchitectures for low energy at high speed. IBM Journal of Research and Development, 63(6), 8-1.
16. Welser, J., Pitera, J. W., & Goldberg, C. (2018, December). Future computing hardware for AI. In 2018 IEEE International Electron Devices Meeting (IEDM) (pp. 1-3). IEEE.
17. Milojicic, D. (2020). Accelerators for artificial intelligence and high-performance computing. Computer, 53(02), 14-22.
18. Wang, Y., Wang, Q., Shi, S., He, X., Tang, Z., Zhao, K., & Chu, X. (2020, May). Benchmarking the performance and energy efficiency of AI accelerators for AI training. In 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID) (pp. 744-751). IEEE.

19. Jelenkovic, P. R., Kang, X., & Tan, J. (2007). Adaptive and scalable comparison scheduling. ACM SIGMETRICS Performance Evaluation Review, 35(1), 215-226.
20. Roosta, S. H., &Roosta, S. H. (2000). Artificial intelligence and parallel processing. Parallel Processing and Parallel Algorithms: Theory and Computation, 501-534.
21. Weems, C. C. (2002). Architectural requirements of image understanding with respect to parallel processing. Proceedings of the IEEE, 79(4), 537-547.