Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 10, October-2021

Artificial Intelligence Chips: A Promising Future of Semiconductor Industry

Juilee Samir Kotnis
Dept. Electronics Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India.

Mrs. Nibha Desai
Dept. Electronics Engineering
Shah & Anchor Kutchhi Engineering College
Mumbai, Maharashtra, India.

***Abstract*— Technology is a steady and rapidly growing field that has produced colossal technological developments over the recent past. These developments have transformed the manner and speed at which the entire world operates. Hence, the advancements in technology have proved to be productive in all aspects of life. Semiconductors have played a critical role as enablers in powering several cutting-edge digital devices to achieve all these technological developments. The global semiconductor industries are expected to maintain their continuous growth because of the new technological discoveries such as Artificial-Intelligence Chips, AI Chips. The AI Chips are the recent creation specially designed to handle high-speed artificial intelligence commands with minimum power. Therefore, the AI Chips could play a pivotal role in global economic development by being featured in smartphones, autonomous cars, smart homes, robotics, and other technologies. Hence, this detailed literature survey extensively reviews the transformation from general-purpose chips to AI Chips, their techniques of operation, various types of AI Chips available such as GPU, FPGA, ASIC, and their applications.**

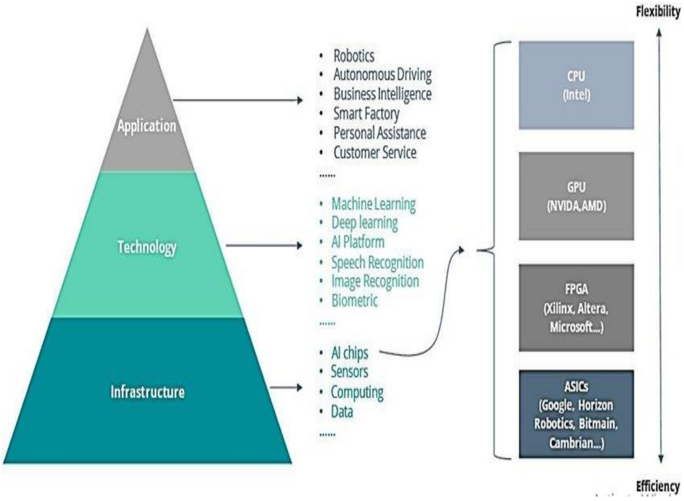
Keywords*— *Moore's Law, General Purpose Chips, Artificial Intelligence, AI Chips, Training, Inference, GPU, FPGA, ASIC.

I. INTRODUCTION

Computing devices have taken over the world and it has been challenging to operate without them. Computing tools are utilized in fields such as healthcare, travel, and entertainment. The devices can perform various tasks with the assistance of minute engines called semiconductors devices. These semiconductor tools called transistors are electrical switches that perform complex analysis. Moore's law, predicted by Gordon Moore in 1965, the co-founder of Intel, stated that transistors per silicon chip would double in every two years thus escalating computational power and reducing the cost of chips [1]. This resulted in the transistors to be oberved as only a few atoms wide. Recently, the process of shrinking transistors has lost momentum as it has started affecting the boundaries of physics thus making the pathway for specialized chips such as AI Chips. Artificial Intelligence is fundamental because it imparts the computing devices the propensity to employ prodigious quantum of information and employ their intelligence to take decisions and make discoveries within a fraction of a second comapred to humans. The AI Chips have enabled the computing devices to influence the modern world by delivering complex computations owing to their unique design.

II. ROLE OF AI CHIPS IN THE FRAMEWORK OF ARTIFICIAL INTELLIGENCE

The framework of Artificial Intelligence comprises of three layers as shown below:



A. The Infrastructure Layer

This layer consists of the AI Chips and sensors required to support the technology layer's decision-making, reasoning, and learning abilities.

B. The Technology Layer

This layer employs various technologies such as deep learning, machine learning and speech and image recognition and hence is responsible for driving this layer and is fundamental in AI algorithms processing.

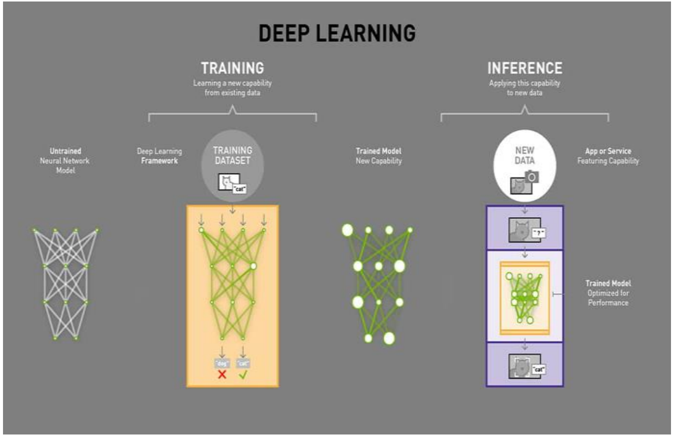
C. The Application Layer

This layer offers different domains where AI Chips can be utilized such as business intelligence, automotive industry, customer services, and individual assistance.

III. DEPLOYMENTS OF AI IN DEEP LEARNING

Two significant factors drive the demand for specialized AI Chips. The first one includes the essential enhancements in semiconductor abilities that have shifted from producing to designing and software. The second is the ever-rising demand for applications depending on AI where the complex computations can only be achieved by the AI Chips. Therefore, the AI algorithms that fulfill the requisites of both these factors are called Deep Neural Networks which utilize a specific machine learning process such as Supervised

Learning comprising two computational phases which are explained with the help of a pictorial representation below [2]:


DEEP LEARNING
TRAINING
Learning a new capability from existing data
INFERENCE
Applying this capability to new data
Untrained Neural Network Model
Deep Learning Framework
TRAINING DATASET
"cat"
"dog"
"cat"
Trained Model New Capability
NEW DATA
App or Service Featuring Capability
"?"
Trained Model Optimized for Performance
"cat"

A. Training

At this stage, AI uses a colossal amount of information to teach the Neural Network models to produce novel trained frameworks. In order to reach a deduction, these models are devised to procure information from new sets of information. Massive amount of computational power is utilized stepwise to present such colossal amount of information. Thus, to process such massive, divergent and parallel data sets, hardware equipment such as high-end encrypted servers are extremely important for the fulfillment of the tasks.

B. Inference

At this stage, the previously procured information is utilized to reach a deduction and then the trained AI algorithm is performed. This step critically analyzes power usage, delay and hardware required by this stage.

IV. VARIOUS TECHNIQUES OF AI CHIPS

The AI Chips use specific techniques to enhance their efficacy and speed than the general-purpose chips which are as given below:

A. Parallel Computing

Parallel computing is the significant development an AI chip enabling it to perform numerous computations simultaneously compared to the general-purpose chips. The teaching and inference phases require many distinct matrix multiplication processes that need enormous multiplications that are put together resulting in Multiply and Accumulate Operations. Thus, to perform the matrix multiplications effectively, AI Chips use Multiply and Accumulate circuits, making it easier to complete the calculations by utilizing two techniques such as data and model parallelism [3].

- Data Parallelism
  This technique is the most recurrent type that systematically divides the information set entered into various batches wherein computations are performed on each batch. The batches are then split further across different execution units or across execution units of AI Chips affixed parallelly. Data parallelism is recommendable for the various neural networks as it offers the advantage of using several batches at the initial training stage. At this stage, the techniques emphasize upon accomplishing consistent framework accuracy without proliferation of the total number of calculations required. However, this comes with the drawback of no reduction in time consumption [3].

- Model Parallelism
  This technique divides the framework into subsections on which calculations are performed simultaneously on specific units of AI Chips or across specific units of AI Chips affixed parallelly [3].

B. Low Precision Computing

Low Precision Computing is highly recommended for AI algorithms as it compensates numerical precision for efficiency and speed. For instance, let us consider a X bit processor which comprises of execution units devised to alter the information that is denoted by X bits. X bit would sanction 2^x distinct combinations as it follows the concept that a transistor can occupy one bit which could be either 0 or 1. Higher bit data types are able to denote a vast range of numbers or higher precision numbers within a restricted range. Lower bit data types are also suitable for the same task as higher bit data types as the training and inference stages of the AI algorithms of AI chips perform nearly well [3].

C. Memory Optimization

AI Chips can make the best of the size of the memory, the location and types when an AI algorithm's memory access patterns can be anticipated. An example is where few AI chips contain enough memory to complete AI algorithms on the chip. In comparison to off-chip memory, on-chip memory can give a reasonable frequency and raise the speed. An alternative becomes model parallelism in scenarios where a model is estimated to be too big to fit in a single AI chip. It separates a model enabling it to have various partitions for training on different AI chips affixed parallelly. Also, memory optimization may be done directly or indirectly and affects either two or more cost merits, including how it performs, area and the new implementation's power dissipation. There are different ways of enabling memory optimization, for instance, correcting memory leakage and corruption. Once these are fixed, the usage and usability is improved.

D. Domain Specific Languages

Domain Specific Languages prove to be vital for complex computations carried out on specialized chips such as AI chips. For instance, let us consider the example of two general purpose programming languages such as Python and C wherein both can be utilized for carrying out computations but are not efficient enough to carry out complex computations that specialized chips such as AI chips demand. In such a case Domain Specific Languages come to the aid. Examples of Domain-Specific Language include the HTML used for the web pages. Domain-specific Language serves in solving different problems by aiming at particular tasks and working on them [3].

V. DIFFERENT TYPES OF AI CHIPS

There are various types of hardware that make up artificial intelligence chips. They include Graphic Processing Unit, Field Programming Gate Array, and the Application Specific Integrated Circuit.

A. Graphic Processing Unit

When talking about Artificial Intelligence, the Graphics Processing Unit (GPU) serves an essential purpose. GPU was mainly invented for video and graphic productions, though they became popular in Artificial Intelligence with time. It can supply a large amount of workload and integrate a vast amount of computational capacity when looking at its nature. Therefore, this makes it suitable for AI algorithms that need many parallelisms. GPU's central area is broadly utilized in the cloud, data centers, security and automotive industries. Initially, GPU was designed to render 3D graphics, however, they have become flexible enhancing their performance with time.

B. Field Programmable Gate Array

The Field Programmable Gate Arrays offers the best work to the Artificial Intelligence chips compared to other hardware. Compared to GPU, the Field Programmable Array does not have the circuitry inside the chip hard etched; therefore, it can be reprogrammed. Field Programmable Gate Array has several advantages; it offers cost-effectiveness, efficiency in power and functional diversity. This combination can give a perfect performance in Artificial Intelligence operations that requires considerably low latency. The primary functions of the Field Programmable Gate Array are doing away with input and output bottlenecks, incorporating AI workloads, giving sensor amalgamation and giving speed to High-Performance Computing.

C. Application Specific Integrated Circuits

There are many functions of the Application Specific Integrated Circuits including executing a specific AI algorithm. They are majorly applied in situations which are disparate, intensive and workloads that require high effectiveness. The main advantage of using Application-Specific Integrated Circuit is that it uses little power. It has many variations, such as Tensor Processing Unit, Brain Processing Unit, Vector Processing Unit and Neutral Network Processing Unit.

VI. COMPARISON OF AI CHIPS

Given below is a table that provides a clear distinction of GPU, FPGA and ASIC [4].

	Graphic Processing Units	Field Programmable Gate Arrays	Application Specific Integrated Circuits
Power consumption	High	Medium	Low
Latency	More	Less	Less
Flexibility	Medium	High	Low
Efficiency	Low	Medium	High

VII. APPLICATIONS OF AI CHIPS

AI chips can be employed for a variety of applications for a variety of reasons which are explained as following. Price of smart devices will surge as they will employ AI chips with the target of enhancing performance and user experience. The surge in the price of AI chips will increase the price of smart devices thus providing increased revenue to the manufacturers of smart devices. Autonomous driving requires AI chips as training and inference is involved in all its stages which are sensing, modelling and decision making. Only the computational power of AI chips will be able to satisfy all these demands. Surveillance systems have been through two stages wherein they acquired the capabilities of recording clear videos and networking with interconnection. Now they are experiencing the third stage wherein with the help of AI chips they are able to conduct real time processing of video data which enables them to save a lot of storage space and thus generate a lot of data every single day.

VIII. CURRENT AI CHIPS IN USE

The table given below provides information of the various AI chips in use and their details:

Type	AI Chip	Use
GPU	Radeon Instinct	Deep Learning Artificial Neural Network High Performance Computing GPGPU
	Tesla V100	High Performance Computing Data Science Graphics
FPGA	Agilex	Data Center 5G Network Smart NICs
	Virtex	10G to 100G Networking Portable Radar
ASIC	FSD Computer	Tesla Self Driving Cars
	TPU v3	Custom developed for Google’s Machine Learning

IX. CONCLUSION

Artificial Intelligence is present everywhere around us. We can see how AI chips reside everywhere right from the processor of our smartphones to automotive technology. Artificial Intelligence is being utilized everywhere in the world today. For any system to be considered efficient, the developers must utilize an Artificial Intelligence chip. For an organization to produce a chip that can perform tasks faster and cheaper, they need to rely on AI. Through the review, the overview of AI chips is studied which suggests that soon AI Chips will be used in almost every application.

REFERENCES

[1] Marco Chiappetta on “Chips designed by AI are the future of future of semiconductor evolution beyond Moore’s Law” March 2021.
[2] Kim Dilmegani on “AI chips: In depth guide to cost-efficient AI training and inference” July 2021.
[3] Saif M. Khan and Alexander Mann on “AI chips: What-They-Are-and-Why-They-Matter” April 2020.
[4] Lynnette Reese on “Comparing hardware for Artificial Intelligence: GPUs VS FPGAs VS ASICs” July 2018.
[5] Deloitte on “Semiconductors: The Next Wave” April 2019