ISBN: 978-93-5779-832-7

Semiconductors on AI Technologies

[1]**Dr.P.Pavitra**, Asst. Professor of Chemistry, Dadi Institute of Engineering and Technology, Visakhapatnam, A.P., India
[2]**B.Roopa Lakshmi Tulasi**, Dadi Institute of Engineering and Technology, Visakhapatnam, A.P., India

Abstract
Now a day, semiconductors are significant technology enablers that control many of the cutting-edge digital devices. The comprehensive semiconductor industries are chosen to maintain its robust progress due to arriving technologies such as self-directed driving, artificial intelligence (AI), 5G and Internet of Things in the following decade. Many potential divisions especially in the automotive sector and AI will provide huge prospects for semiconductor companies. AI semiconductor has realized a sprint not just at the application level but also at the semiconductor chip level, commonly known as AI Chips. As the term recommends, AI chips refers to a recent generation of microprocessors which are particularly planned to process artificial intelligence tasks faster, using less power. AI chips could play a central function in economic growth moving forward because they will surely feature in cars which are becoming intentionally autonomous, smart homes where electronic devices are becoming more intelligent, robotics and many other services. This paper reviews about the competing technologies and the development trends of AI chips.

Key Words: Semiconductors, artificial intelligence (AI), automotive sector, AI chips, smart homes, and manufacturing

Artificial intelligence (AI) chips are inclusive silicon chips which integrate AI technology and are used for machine learning. AI helps to remove or diminish the risk to human life in many industry areas. The need for more productive systems to solve computational problems is becoming critical, owing to the increase the volume of data. Thus, on developing AI chips and applications, many the key players in the IT industry have enthusiastic them self. Moreover, the influx of quantum computing and increased application of AI chips in

ISBN: 978-93-5779-832-7

manufacturing the growth of the global artificial intelligence chip market. In addition, the initiation of autonomous robotics is predicted to provide potential growth predictions for the market. current years, most of the computations of AI are almost been done abstractedly in data centers or on secure essential appliances or on telecom edge supercomputers, because of AI computations are demanding hundreds of varying types of chips to tool and are suggestively processor-intensive. It is necessarily incredible to integrate AI divisions in anything smaller than a footlocker because of its size; cost and power channel of the hardware.

At present, all those have been transformed by AI chips. These AI chips are entirely small, reasonable cost, use less power and produce very less heat. These parameters are making AI chips possible to participate into hand-held devices such as smartphones Consequently, AI chips can deliver the data with high speed, security, and confidentiality by agreeing the above devices to execute processor-intensive AI computations locally there by reducing or eliminating the necessity to send large amount of data to a remote location.

Currently, there is no fixed and extensively accepted standard for the explanation of AI applications. some chips have achieved great success in some AI application scenarios by combining traditional computation architectures with various hardware and software acceleration schemes. AI technology is a multilayered technology which flows through the layers of application, algorithm mechanism, chip, tool chain device process and material technology levels. These layers are subsequently connected to form the AI technology chain. the top-down flow is driven by the application requirements and the bottom-up flow is driven by the hypothetical innovations. Here, the AI chip is in the middle of the whole chain and providing effective support for applications and algorithms upwards and raising demand for devices and circuits, progressions and materials downwards. On the other hand, new materials development, evolutions, and devices such as 3D stacked memory and process developments also provides the possibility of encouragingly educating performance and reducing power consumption for AI chips. These two types of power the rapid development of AI chip technology is jointly

ISBN: 978-93-5779-832-7

promoted in recent years.

Types of AI chips

GPUs have very high performance appropriate for deep learning AI algorithms that require a lot of parallelism. This makes GPUs as a great selection for AI hardware. GPU used to progression realistic comprehensive responsibilities such as games, are built with correspondence. GPUs are now broadly used in cloud and data centers for AI training, motorized and security segments. The GPU is the most broadly used and flexible AI chip available in the market.

FPGAs are programmable arrays appropriate for consumers and they will reprogram based on their own necessities. These are modelled by a faster development cycle when compared with ASIC and low power necessities compared to GPUs. But, the cost of FPGA is relatively high due to its elasticity. Between efficiency and flexibility, FPGAs can be viewed as a best deal, particularly when an AI algorithm has been integrated with it. This allows the chip dealers to avoid the cost and potential knowledge disuse of the ASIC approach and to optimize the convention chips for their applications.

ASIC chips combined with AI algorithm have limited design for AI applications. These are all intended at different, computer-intensive, rules-based workloads with high flexibility, efficiency and performance. Generally, when compared with GPUs and FPGAs, ASIC AI chips have higher efficiency, a smaller diesize and lower power consumption. But, the development cycle of ASIC chip is longer and less flexible which has found as a main reason forits slow commercial adoption.

Classification of AI chip

Training Data set

In the cloud, AI utilizes big data as a substance to network models and these newly accomplished models are obtained using training datasets. A newly trained model is then serviced with new capability to "infer" from new datasets to reach a conclusion. The training phase involves a tremendous amount of computational power because it requires the application of a huge data set to a neural network model. This demands high-end servers which have progressive parallel

ISBN: 978-93-5779-832-7

performance. This enables to process large, diverse, and highly parallel datasets and are therefore typically done in the cloud via hardware. On the other base, the inference phase can be handled either in the cloud or on devices(products) at the edge. In comparison with training chips, inference chips require more attentive consideration of power usage, latency, and cost.

Network edge-based AI

AI chip settlement is not incomplete to the cloud, but can also be seen in a wide variability of network edge devices such as smartphones, autonomous vehicles, and security cameras. Most AI chips at the edge are inference chips and they are becoming increasingly specialized. For some applications, machine learning models that have been trained in the cloud must be inferred at the edge due to various reasons such as latency, bandwidth, and privacy concerns. Power and costs are additional constraints for AI at the edge. For autonomous driving, the inference should be implemented at edge instead of in cloud, in case of network delay. Edge devices cover a large scope, and their application scenarios are also varied. For example, the automatic driving may need a very strong computing device, while wearable devices must achieve certain intelligence under the strict constraints of power consumption and cost. In the future, a lot of edge devices in AI application mainly perform inference computing, which requires the edge devices have sufficient inference computing ability. However, the computing power of edge AI chips cannot meet the need of local inference.

Applications of AI chips

i. Data safety and confidentiality
ii. Low connectivity
iii. Large data
iv. Power limitations
v. Low potential necessities

Data safety and confidentiality

This is enormously important because the hazard is becoming even more critical to address as time goes. Some devices, such as smart speakers, are started to be used in hospitals where

ISBN: 978-93-5779-832-7

patient privacy is controlled by favorable large amount of data to be handled edge AI chips can limit the risk of personal or enterprise data being blocked or misused. Machine learning chips can also identify a wider range of voice so that only fewer audio is required to be recognized in the cloud. Collecting, storing, and transmitting data to the cloud necessarily appears for an organization to cyber security and privacy threats, even when companies are attentive about data protection.

Low connectivity

These are helpful in recognizing swimmers if they are taken by rip current or it also notifies the swimmers in case the sharks and crocodiles enter before an attack. All these are done without an internet connection. Any device must be connected to internet for the data to be accessed in the cloud. However, in some cases, connecting the device is unrealistic, for example consider buzzes in which maintaining connectivity with a drone is very difficult depending on where they operate and both the connection and uploading data to the cloud can limit the battery life.

Large data

Huge amounts of data can be generated by IOT device Transmitting all these data to the cloud for storage and analysis is bit costly and very complex. Integrating machine learning processors enable the sensors or cameras to solve this issue. Embedded edge AI chips, a device can review data in real time, transmit only the data which are relevant for further analysis in the cloud. Therefore, it reduces the cost of storage and bandwidth.

Power limitations

Machine learning chips with low-power permit devices with small batteries, in addition to the low-power edge AI chips currently available, the companies are working to develop deep learning on devices as small as microcontroller units which are same as, less sophisticated and with low power which usually draws only mill watts or even microwatts. ARM chips are being integrated with respiratory inhalers to analyze data such as inhalation lung capacity and the flow of medicine into the lungs. The analysis of AI is performed on

ISBN: 978-93-5779-832-7

the inhaler and the results are then transmitted to a smartphone app, aiding health care professionals to improve personalized care for asthma patients.

Low potential necessities

Consider autonomous vehicles for example, it must collect and process huge amounts of data from computer vision systems to recognize objects as well as from the sensors which controls all the functions of the vehicle. Then they must immediately convert these data into decisions like when to turn, apply brake or accelerate in order to provide safely. To execute this, autonomous vehicles must process all the data they collect in the vehicle itself and autonomous vehicles of today use a variety of chips for this purpose which includes standard GPUs as well as edge AI chips.

Conclusion

Now a days, AI chip is still in its early stages stage. The only sure thing is that it is the essential one of AI technology development and it is better incentive for semiconductor industry. Nowadays, research around AI chips has completed important progress in the area of machine learning based on neural network which is superior to human intelligence in solving some computing-intensive problems. By the integration of CMOS technology, emerging information technologies and the emergence of open-source software and hardware, we can anticipate an un suspected era where innovations are achieved synergistically.

Reference

1.N. P. Joupp., et al. "In-datacenter performance analysis of a tensor processing unit". In Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA) 2017, 1-12. DOI: 10.1145/3079856.3080246

2. D. Bankman., et. al. "An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor with All Memory on Chip in 28nm CMOS," Solid-State Circuits Conference (ISSCC), 2018 IEEE International. IEEE, 2018, 222 – 224. DOI: 10.1109/ISSCC.2018.8310264

3. J, Lee., et al. "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision" Solid-State Circuits Conference (ISSCC), 2018 IEEE

International. IEEE, 2018, 218 – 220. DOI: 10.1109/ISSCC.2018.831026
4. Kriegbaum, Jeff (13 September 2004). "FPGA's vs. ASIC's". EE Times.
5. William Chou, Jennifer Shao, Roger Chung, Leo Chen, Andrew Chen, Lisa Zhou, "Semiconductors– the Next Wave", Deloitte ToucheTohmatsuLimited(DTTL),April-2019.
6.PaulLee,JeffLoucks,DuncanStewart,DavidJarvis,ChrisArkenberg,"TMTPredictions2020:Thecanopyeffect",DeloitteInsights,2019.
7. Han Dong, Zhou Shengyuan, ZhiTian, Chen Yunji, Chen Tianshi, "A Survey of Artificial Intelligence Chip", Journal of Computer Research andDevelopment, Volume56, Issue1, pp:7-21,2019,DOI:10.7544/issn10001239.2019.20180693
8.TYGARIBAY,"ArtificialIntelligenceChips:Past,PresentandFuture",AUGUST2018.
9. Saif M. Khan, Alexander Mann, "AI Chips: What They Are and Why They Matter", Center for Security and Emerging Technology, APRIL 2020
10. Rahul Kumar ,SupradipBau, "Artificial Intelligence Chip Market by Chip Type (GPU, ASIC, FPGA, CPU, and others), Application (NaturalLanguage Processing (NLP), Robotic, Computer Vision, Network Security, and Others), Technology (System-on-Chip, System-in-Package, Multi-chip Module, and Others),
11. Processing Type (Edge and Cloud), and Industry Vertical (Media & Advertising, BFSI, IT & Telecom, Retail, Healthcare,Automotive&Transportation,andOthers):GlobalOpportunityAnalysisandIndustryForecast,2019-2025",AlliedMarketResearch,May2019.