ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

AI Chips: New Semiconductor Era

Sorna Mugi Viswanathan

B.E. Student, Department of Electronics and Communication Engineering, KGiSL Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT: Today, semiconductors are important technology enablers that power many of the cutting-edge digital devices. The global semiconductor industries are assigned to maintain its robust growth due to arriving technologies such as autonomous driving, artificial intelligence (AI), 5G and Internet of Things in the following decade. Many budding divisions especially in the automotive sector and AI will provide huge opportunities for semiconductor companies. AI semiconductor has seen a sprint not just at the application level but also at the semiconductor chip level, commonly known as AI Chips. As the term suggests, AI chips refers to a recent generation of microprocessors which are particularly designed to process artificial intelligence tasks faster, using less power. AI chips could play a crucial function in economic growth moving forward because they will surely feature in cars which are becoming deliberately autonomous, smart homes where electronic devices are becoming more intelligent, robotics and many other technologies. This paper reviews about the competing technologies and the development trends of AI chips.

KEY WORDS: Semiconductors, cutting-edge digital devices, artificial intelligence (AI), automotive sector, AI chips, smart homes and robotics

I.INTRODUCTION

Artificial intelligence (AI) chips are comprehensive silicon chips which integrate AI technology and are used for machine learning. AI helps to eliminate or minimize the risk to human life in many industry shafts. The need for more productive systems to solve mathematical and computational problems is becoming critical, owing to the increase the volume of data. Thus, on developing AI chips and applications, a large number of the key players in the IT industry have dedicated themself. Moreover, the arrival of quantum computing and increased implementation of AI chips in robotics steer the growth of the global artificial intelligence chip market. In addition, the arrival of autonomous robotics (robots that develop and control themselves independently) is anticipated to provide potential growth opportunities for the market.

Till recent years, most of the computations of AI are almost been done distantly in data centres or on firm core appliances or on telecom edge processors (not internally on devices). This is because AI computations are requiring hundreds of varying types of chips to execute and are significantly processor-intensive. It is fundamentally incredible to integrate AI computations in anything smaller than a footlocker because of its size; cost and power drain of the hardware.

Presently, all those have been changed by AI chips. These AI chips are completely small, fairly inexpensive, use less power and generate very less heat. These parameters are making AI chips possible to integrate into handheld devices such as smartphones and even into non-consumer devices such as robots. Therefore, AI chips can deliver the data with high speed, security and privacy by allowing the above devices to execute processor-intensive AI computations locally thereby reducing or eliminating the necessity to send large amount of data to a remote location.

II. TECHNOLOGY OVERVIEW

At present, there is no fixed and widely accepted standard for the definition of AI chips. A wider view is that all chips for AI applications can be called as AI chips. Nowadays, some chips have achieved great success in some AI application scenarios by combining traditional computing 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. In Fig.1, the top down flow is driven by the application requirements and the bottom up flow is driven by the theoretical innovations. Here, the AI chip is in the middle of the whole chain and providing efficient support for applications and algorithms upwards and raising demand for devices and circuits, processes and materials downwards. On the other hand, the rapid development of new materials, processes and devices such as 3D stacked memory and process evolutions also provides the feasibility of significantly improving performance and reducing power consumption for AI chips. Generally speaking, in recent years, the rapid development of AI chip technology is jointly promoted by these two kinds of power.

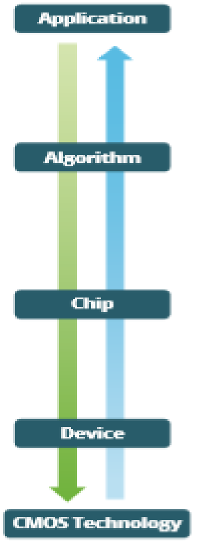

Fig.1. AI chip Technology chain

III. ROLE OF AI CHIP IN THE LAYERS OF AI

The artificial intelligence framework can be broadly characterized into three layers.

- Infrastructure
- Technology
- Application

The infrastructure layer consists of the core AI chips and the big data which will boost the sensing and cognitive computational performance of the technology layer. The application level is at the apex, supplying services such as virtual assistance, smart security, smart robotics and autonomous driving. AI chips form the heart of the AI technology chain and are central to the processing of AI algorithms, particularly for deep neural networks (DNN).

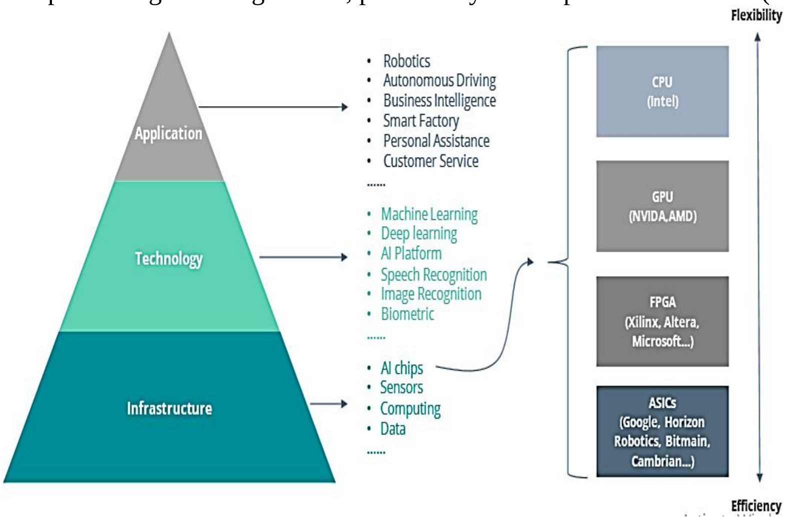

Fig.2. Role of AI chip in the layers of AI

There are numerous types of AI chips available for acceleration, including

A. Graphical Processing Unit (GPU),

B. Field Programmable Gate Arrays (FPGA), and
C. Application Specific Integrated Circuits (ASIC).

A. Graphical Processing Unit (GPU)
GPUs used to process graphic intensive tasks such as games, are built with parallelism. GPUs have very high performance suitable for deep learning AI algorithms that require a lot of parallelism. This makes GPUs as a great option for AI hardware. GPUs are now widely used in cloud and data centres for AI training. They are also used in automotive and security sectors. The GPU is currently the most widely used, most flexible AI chip available in the market.

B. Field Programmable Gate Array (FPGA)
FPGAs are programmable arrays suitable for clients and they will reprogram based on their own requirements. FPGAs are modelled by a faster development cycle when compared with ASIC and low power requirements compared to GPUs. But, the cost of FPGA is relatively high due to its flexibility. 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 technology disuse of the ASIC approach and to optimize the custom chips for their applications.

C. Application Specific Integrated Circuits (ASIC)
On the other hand, ASIC chips integrated with AI algorithm have exclusive design for AI applications. There are different types of ASIC-based AI chips have including TPU, NPU, VPU and BPU, etc. 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 die size and lower power consumption. But, the development cycle of ASIC chip is longer and less flexible which has found as a main reason for its slow commercial adoption.

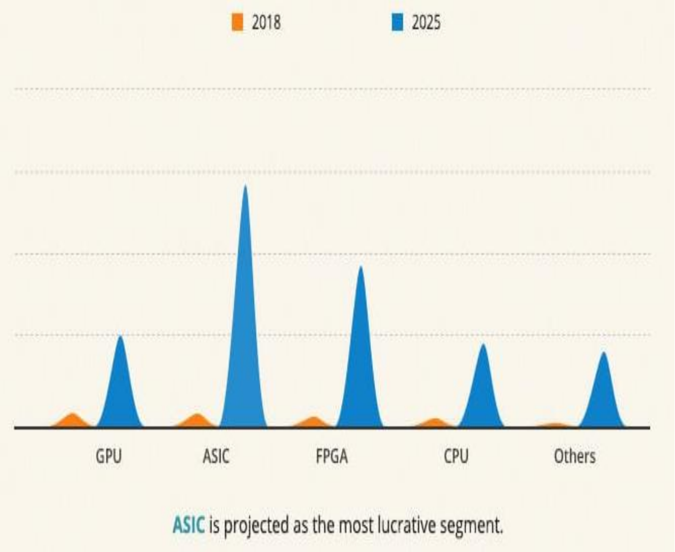

Fig.3. Growth of AI chips since 2018 to 2025

IV. CATEGORIES OF AI CHIP

The AI chip market can be split into two categories based on deployment methods, they are

1. Cloud-based AI chips,
2. Network edge based AI chips.

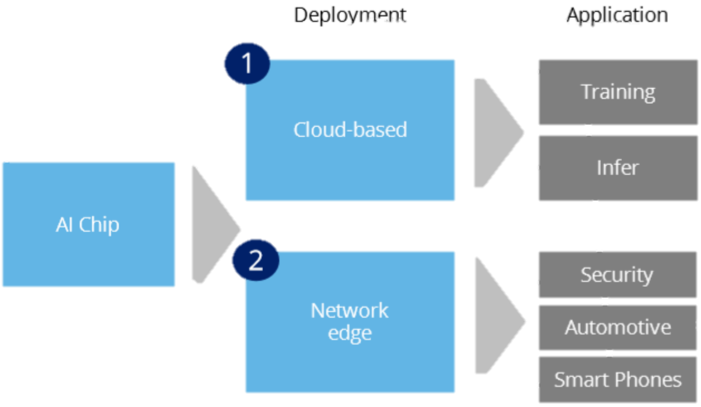

Fig.3. Categories of AI chip

1. Cloud based AI

The biggest market for AI chips is the cloud and their acceptance in data centres remains to increase, becoming more efficient, reducing operational cost and improving infrastructure management. Considering the product type, GPUs have become the conventional option for AI chips and have the largest share, resulting for over 30% of the market. In the cloud, GPU, especially NVidias' series GPU chip, have been extensively used for classification and to train deep neural networks. The GPU with thousands of computational cores can achieve 10-100x application throughput compared to CPUs alone. GPU accelerators are still conventional one for machine learning for many of the largest web and social media companies. Further, the performance can be achieved by specialized AI chips. The best-known example is the Google Tensor Processing Unit (TPUv1) currently used for all kinds of AI inference in the cloud, such as search queries and translation.

Training and Infer

In the cloud, AI utilizes big data as a foundation to "train" neural network models and these newly trained models are obtained using training datasets. A newly trained model is then servised with new capability to "infer" from new data sets to reach a conclusion. The training phase requires 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 advanced parallel 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.

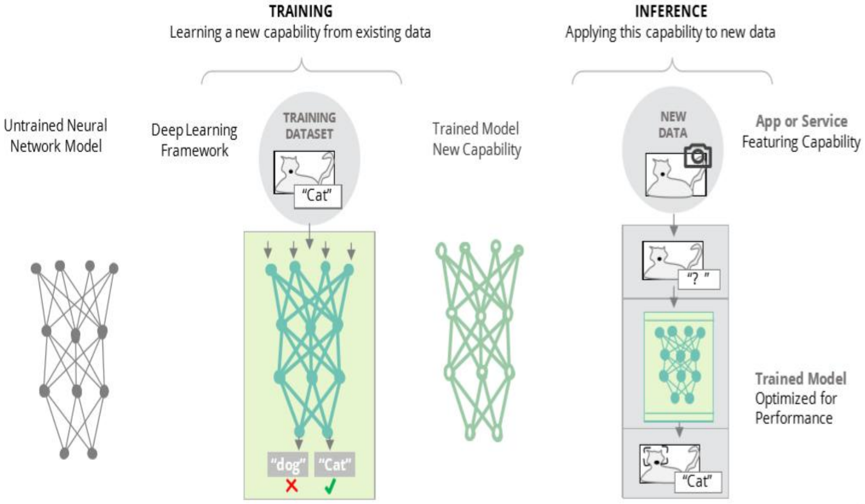
TRAINING
Learning a new capability from existing data
INFERENCE
Applying this capability to new data
Untrained Neural Network Model
Deep Learning Framework
TRAINING DATASET
"Cat"
Trained Model New Capability
NEW DATA
App or Service Featuring Capability
"dog"
"Cat"
"? "
Trained Model Optimized for Performance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

Fig.4. Phases of deep learning

2.Network edge based AI chip

AI chip deployment is not limited to the cloud, but can also be seen in a wide variety 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 actually 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 scenarios 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. Therefore, industry must commission more AI chips at the edge so that it can be applied at different AI application scenarios. The AI inference chip market is expected to grow at a CAGR of 40% and reach 2 billion USD by 2022.

Collaboration between cloud and edge

In summary, the cloud AI processing mainly emphasizes the peak performance, memory bandwidth and costs, where the requirement of accuracy, parallelism and data volume is quite high. GPU, FPGA and specialized AI chips will be promising candidates for in the cloud servers. On the other hand, the edge AI processing mainly focuses on the energy efficiency, response time, and cost and privacy issue. The present collaborative pattern for cloud and edge devices is to train neural network on the cloud and use edge devices for inference. With the increasing capability of edge devices, more and more computing workloads are executed on the edge devices. The collaborative training and inference among cloud and edge devices will be a curious direction to be discovered.

V. STORAGE TECHNOLOGY OF AI CHIPS

One of critical keys to boost the performance and energy efficiency of AI chips is rooted in data accessing though the memory hierarchy. Below are the memories which are used to empower the AI chips.

A) AI friendly memory
B) Commodity memory
C) On-Chip (Embedded) Memory

A) AI friendly memory

Considering the requirements of parallel accessing of large amount of data, AI and big data processing require memory with high bandwidth and large storage capacity. Considering increasing difficulties faced by conventional nonvolatile memory (NVM) in continual scaling, emerging NVMs can play a vital role in memory technologies for AI chips because of their relatively large bandwidth and rapidly increasing capacity.

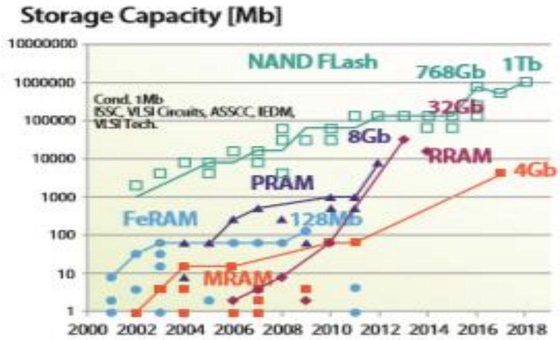

Fig.5. Storage capacity over years

B) Commodity memory

DRAM and NAND Flash memory are commonly used as off-chip memory with relatively large capacity because of their dense cell structure. Recently, 3D integration has been demonstrated to be an effective strategy to increase the bandwidth and capacity of commodity memory, which can be done by either stacking multiple dies using through silicon via (TSV) technology or monolithic fabrication from bottom to top. Representative works of DRAM in this direction include high bandwidth memory (HBM) [Lee14] and hybrid memory cube (HMC) [Jeddeloh12]. Figure 6 shows the NVIDIA's GPU product integrated with HBM for AI applications [NVIDIA]. As for NAND Flash, 3D NAND is being intensively studied. Recently, 96-layers 3D vertical NAND has been developed by Samsung.

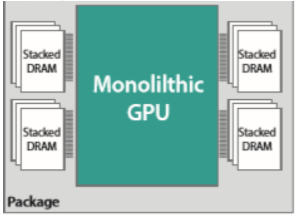

Fig.6. Conceptual View of NVIDIA's GPU with High Bandwidth Memory (HBM)

C) On-Chip (Embedded) Memory

Because of its capacity to interface the logic and memory circuits and its full compatibility with logic devices, SRAM is an indispensable on-chip memory and its performance and density continually benefit from relentless CMOS scaling. However, its volatility has necessitated the use of on- or off-chip NVMs. Although NOR Flash is widely used as on-chip NVM nowadays, it limits the system performance because of its relatively low access time and large write energy.

TABLE I

DEVICE METRICS OF DIFFERENT MEMORIES

Device Metric	SRAM	DRAM	NAND	NOR
Write Energy	low	low	high	high
Write Latency	~1ns	~5 ns	> 100µs	10µs~1ms
Read Latency	~1ns	20~80ns	5~200µs	~50ns
Program Window	Good	Good	Good	Good
Endurance	Unlimited	Unlimited	104-105	104-105
Cell Size	~100 F2	~7 F2	~4 F2	~10 F2

Emerging Memory

Emerging NVMs can significantly contribute to AI-friendly memory both for commodity and embedded applications. Arriving NVMs can serve as storage class memory (SCM) to bridge the accessing time gap between the working memory and the storage for commodity memory because of their appropriate speed. PCM and RERAM are main candidates for SCM because they can be integrated with high density. Besides, STT-MRAM is considered as a possible replacement for DRAM due to its high endurance and fast speed. Based on arriving NVMs, on-chip memory can also help good access speed and low power over conventional NVM which is generally attractive for AI chips on IOT edge devices for embedded applications. These devices work with very limited power accessibility.

VI. ADVANTAGES OF AI CHIPS

Problems that AI chips can help notice include

i. Data security and privacy
ii. Low connectivity
iii. Too big data
iv. Power constraints
v. Low latency requirements

i. Data security and privacy

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. This is extremely important because the risk is becoming even more crucial to address as time goes. Some devices, such as smart speakers, are started to be used in settings such as hospitals where patient privacy is controlled even more strictly. By permitting large amount of data to be handled locally, edge AI chips can limit the risk of personal or enterprise data being blocked or misused. for example, machine learning processing with security cameras can limit privacy risks by examining the video to choose which divisions of the video are applicable and sending only those to the cloud. Machine learning chips can also identify a wider range of voice commands so that only fewer audio is required to be recognized in the cloud.

ii. Low connectivity

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 drones 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. Drones with embedded machine learning guard beaches to keep swimmers safe in New South Wales and in Australia. They are helpful in identifying swimmers if they are taken by riptides or it also notifies the swimmers in case the sharks and crocodiles enter before an attack. All these are done without an internet connection.

iii. Too big data

Huge amounts of data can be generated by IOT devices. Around the world, security cameras construct about 2,500 petabytes of data per day. 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. For instance, Cameras could be equipped with vision processing units (VPUs) and low-power SOC processors are the experts for reviewing or pre-processing digital images. With 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.

iv. Power constraints

Machine learning chips with low-power permit devices with small batteries to execute AI computations without excessive power drain. For example, ARM chips are being integrated with respiratory inhalers to analyse data such as inhalation lung capacity and the flow of medicine into the lungs. The analysis of AI is performed on the inhaler and the results are then transmitted to a smartphone app, aiding health care professionals to improve personalized care for asthma patients. 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 SOCs but are smaller, less sophisticated and with low power which usually draws only mill watts or even microwatts. A version of Tensor Flow Lite is being developed which can enable microcontrollers to analyse data and compresses the data off-chip into a few bytes.

v. Low latency requirements

Performing AI computations over wired or wireless medium at a remote data centre means a round-trip latency of at least 1–2 milliseconds in the best case, and tens or even hundreds of milliseconds in the worst case. An edge AI chip would reduce its on-device performance to nanoseconds which is crucial for users where the device must instantaneously collect, process and act upon the data virtually. 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.

VII. FUTURE SCOPE

To achieve its processing power artificial intelligence does not depend only on AI chips. In the advancement of AI, one of the important components is memory where parallel processing with high throughput plays multiple strains on data bandwidth in the memory systems. The demand for AI system memory will createa great opportunities for memory vendors in 2025. Furthermore, the execution of interconnects between subsystems and devices will become a bottleneck as AI systems scale up. Thus, for semiconductor vendors there exists lot many opportunities to construct high speed interconnects to attain the demands of large amount data flowing between systems. Today, AI chips can contain multiple processors to attain maximum parallelism which results in a very large die size. This creates a great challenge for thermal and high voltage power industries in which custom cooling solution is be needed. This provides great opportunities for packaging vendors to come up with products that have thinner form factor and less thermal dissipations for a more cost-effective solution.

VIII.CONCLUSION AND FUTURE WORK

At this moment, AI chip is still in its infancy stage. The only sure thing is that it is the fundamental one of AI technology development and it is greater stimulus for semiconductor industry. Nowadays, research around AI chips has made significant progress in the area of machine learning based on neural network which is considered to be superior to human intelligence in solving some computing-intensive issues. By the integration of CMOS technology, emerging information technologies and the emergence of open source software and hardware, we can anticipate an unsuspected era where innovations are achieved synergistically.

REFERENCES

[1]. Meng-Fan Chang, An Chen, Yiran Chen, K.-T. Tim Cheng, X. Sharon Hu, MeiKeiIeong, Yongpan Liu, Jan Van der Spiegel, Ling, H.-S. Philip Wong, ZhenzhiWu,YuanXie, Santa Barbara Joshua Yang, Shouyi Yin, Jing Zhu, "White Paper on AI Chip Technologies", Beijing Innovation Center for Future Chips (ICFC), 2018.

[2]. William Chou, Jennifer Shao, Roger Chung, Leo Chen, Andrew Chen, Lisa Zhou, "Semiconductors – the Next Wave", Deloitte Touche Tohmatsu Limited (DTTL), April-2019.

[3]. Paul Lee, Jeff Loucks, Duncan Stewart, David Jarvis, Chris Arkenberg, "TMT Predictions 2020: The canopy effect", Deloitte Insights, 2019.

[4]. Han Dong, Zhou Shengyuan, ZhiTian, Chen Yunji, Chen Tianshi, "A Survey of Artificial Intelligence Chip", Journal of Computer Research and Development, Volume 56, Issue 1, pp: 7-21, 2019, DOI: 10.7544/issn 1000 1239.2019.20180693

[5]. TY GARIBAY, "Artificial Intelligence Chips: Past, Present and Future", AUGUST 2018.

[6]. Saif M. Khan, Alexander Mann, "AI Chips: What They Are and Why They Matter", Center for Security and Emerging Technology, APRIL 2020

[7]. Rahul Kumar , SupradipBau, "Artificial Intelligence Chip Market by Chip Type (GPU, ASIC, FPGA, CPU, and others), Application (Natural Language Processing (NLP), Robotic, Computer Vision, Network Security, and Others), Technology (System-on-Chip, System-in-Package, Multi-chip Module, and Others), Processing Type (Edge and Cloud), and Industry Vertical (Media & Advertising, BFSI, IT & Telecom, Retail, Healthcare, Automotive & Transportation, and Others): Global Opportunity Analysis and Industry Forecast, 2019-2025", Allied Market Research, May 2019.