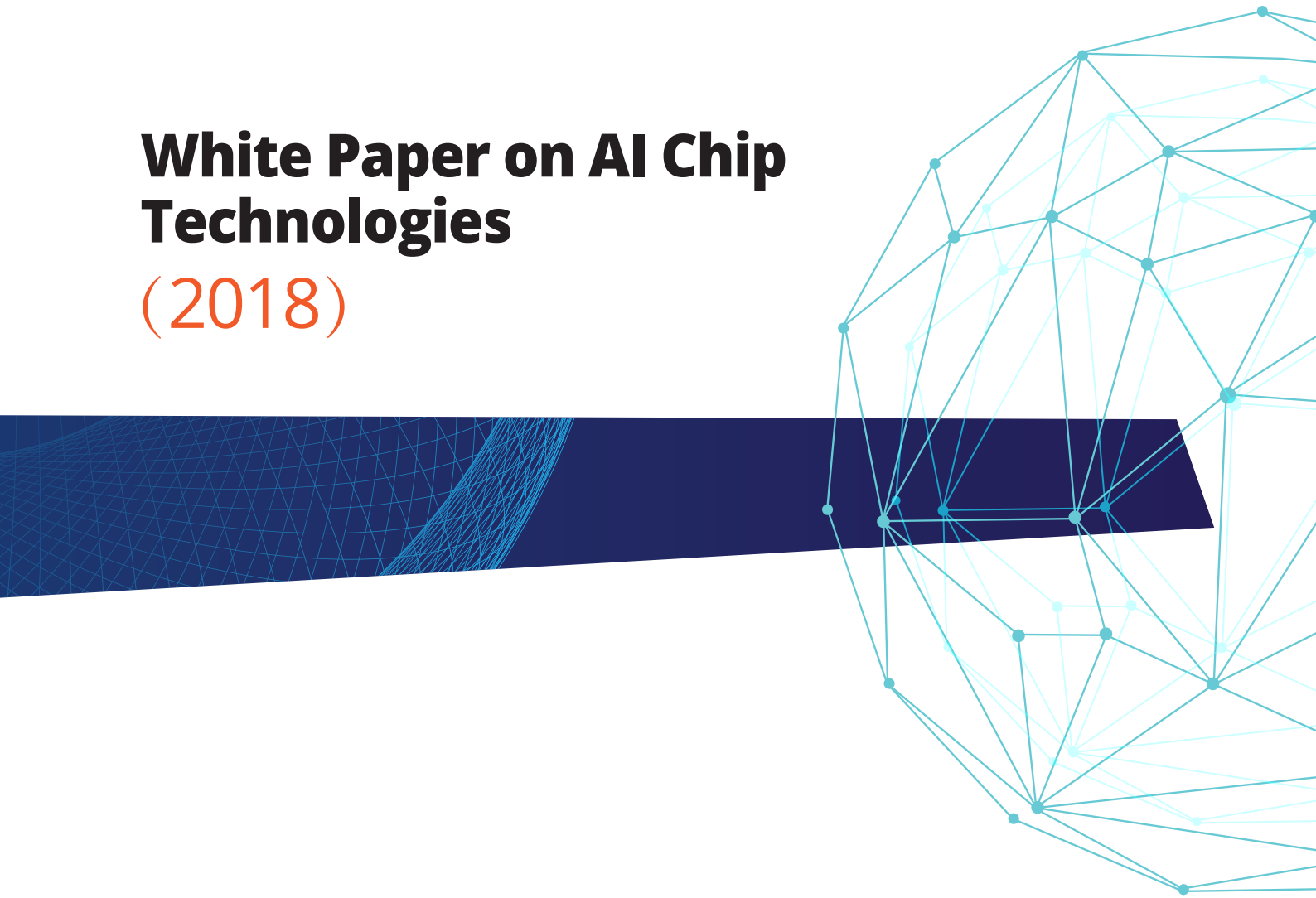
# White Paper on AI Chip Technologies

(2018)



清華大學 Tsinghua University

北京未来芯片技术高精尖创新中心
BEIJING INNOVATION CENTER FOR FUTURE CHIPS

# CONTENTS

Beijing Innovation Center for Future Chips (ICFC)

# White Paper on AI Chip Technologies (2018)


## Editor-in-Chief

| | | |
|---|---|---|
| **Zheng You** | Academician of CAE | Tsinghua University |
| **Shaojun Wei** | IEEE Fellow | Tsinghua University |

## Associate Editor-in-Chief

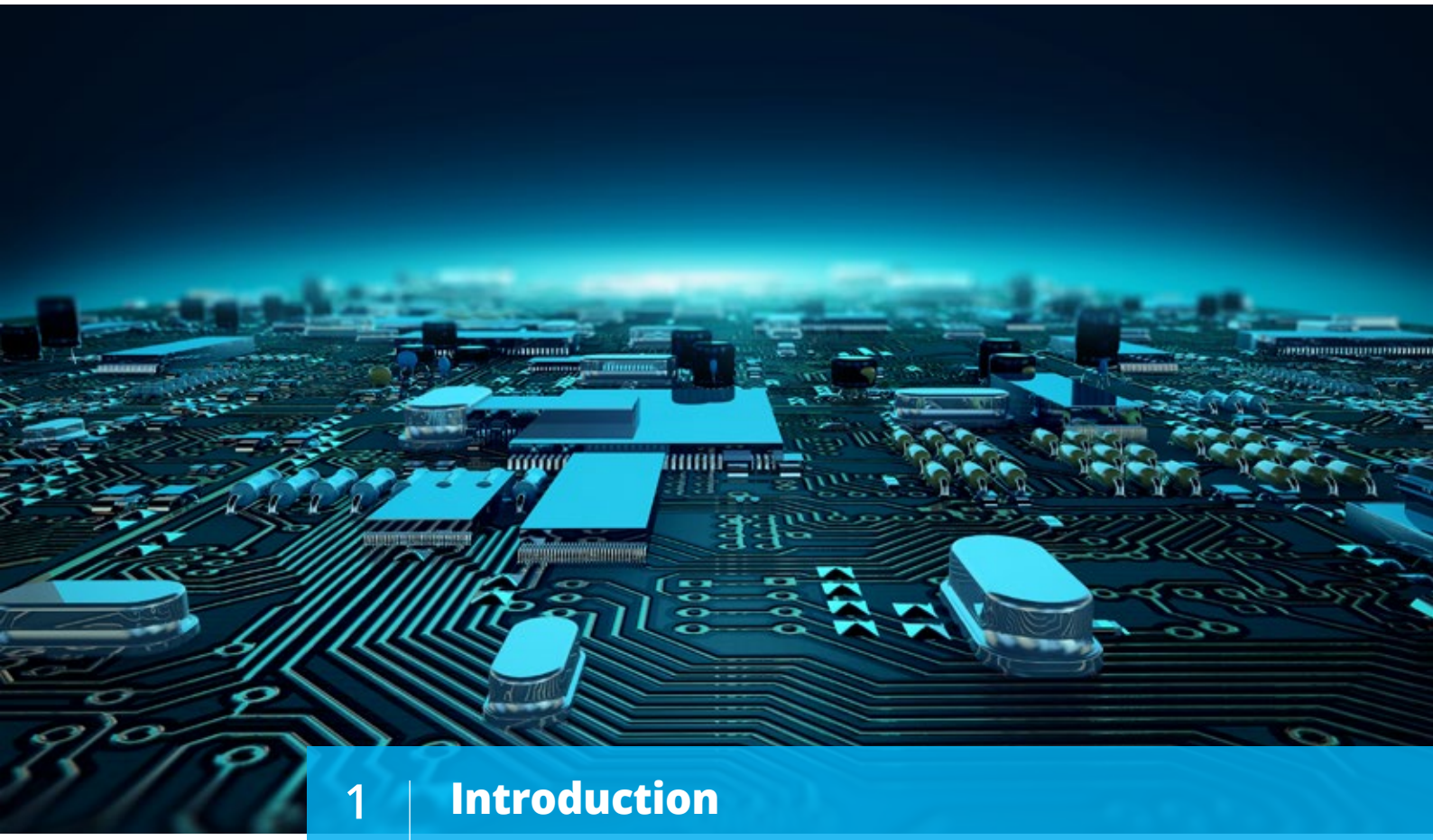
| | | |
|---|---|---|
| **Huaqiang Wu** | | Tsinghua University |
| **Ning Deng** | | Tsinghua University |

## Writing Committee (Listed Alphabetically)

| | | |
|---|---|---|
| **Meng-Fan Chang** | IEEE Fellow | Tsing Hua University (Hsinchu) |
| **An Chen** | | Semiconductor Research Corporation |
| **Yiran Chen** | IEEE Fellow | Duke University |
| **K.-T. Tim Cheng** | IEEE Fellow | Hong Kong University of Science and Technology |
| **X. Sharon Hu** | IEEE Fellow | University of Notre Dame |
| **MeiKei Ieong** | IEEE Fellow | Hong Kong Applied Science and Technology Research Institute |
| **Yongpan Liu** | | Tsinghua University |
| **Jan Van der Spiegel** | IEEE Fellow | University of Pennsylvania |
| **Shan Tang** | | Synopsys, Inc. |
| **Ling Wang** | | Tsinghua University |
| **Yu Wang** | | Tsinghua University |
| **H.-S. Philip Wong** | IEEE Fellow | Stanford University |
| **Zhenzhi Wu** | | Tsinghua University |
| **Yuan Xie** | IEEE Fellow | University of California, Santa Barbara |
| **Joshua Yang** | | University of Massachusetts |
| **Shouyi Yin** | | Tsinghua University |
| **Jing Zhu** | | Beijing Semiconductor Industry Association |

# About ICFC

Beijing Innovation Center for Future Chips (ICFC) has been established in 2015, with collaboration of Tsinghua University and the Beijing Municipal Education Commission. ICFC was built on Tsinghua University's existing exquisite technology and talent advantages. ICFC's main research fields include, low dimensional quantum materials and devices, brain-inspired computing chips, reconfigurable computing chips, emerging memories, smart sensing microsystem and optoelectronics and flexible electronics. The ICFC center director is the vice president of Tsinghua University, Dr. Zheng You, Academician of the Chinese Academy of Engineering. ICFC has grown into a solid team of about 200 and developed the world-class nanofabrication facilities. Look forward, we aim to develop original innovative chips and disruptive intelligent microsystems to facilitate the development of chip industry.



# 1 Introduction

## 1.1 Background

Artificial Intelligence (AI) is a scientific technology that studies and develops theories, methods, technologies, and application systems for simulating, extending, and expanding human intelligence. The fundamental nature of AI is to simulate the human thinking process. Over half a century has passed since the concept of AI was formally proposed in 1956. Over the course, people have been relentlessly pursuing scientific discoveries and technological development in related fields of research in the hope to better understand the essence of intelligence. Like any other emerging discipline that once went through an embryo phase, the early development of AI was beset with difficulties. It had been questioned, and underwent many ups and downs. In recent years, stimulus such as the rise of big data, the innovation of theoretical algorithms, the improvement of computing capabilities, and the evolution of network facilities brought revolutionary progresses to the AI industry which has accumulated knowledge for over 50 years. Research and application innovations surrounding AI has entered a new stage of development with unprecedented vigor.

Nowadays, AI has evolved into a general technology for a wide range of purposes and has been applied in all aspects of economy and society in an accelerated manner. It is already widely applied in sectors, including medical service, finance, security, education, transportation, and logistics, inducing the emergence of new categories of commercial activities, business models, and game-changing product applications. It facilitates the intelligentization of product production, information consumption, and product and services, as well as conducive to the high value-added transformation development.

Clearly, AI is acting as a driving force to economic and social development at the forefront of the technological revolution and industrial transformation. Whether it is the realization of algorithms, the acquisition and a massive database, or the computing capability, the secret behind the rapid development of the AI industry lies in the one and only physical basis, that is, the chips. Therefore, it is no exaggeration to say " No chip, no AI" given the irreplaceable role of AI chip as the cornerstone for AI development and its strategic significance.

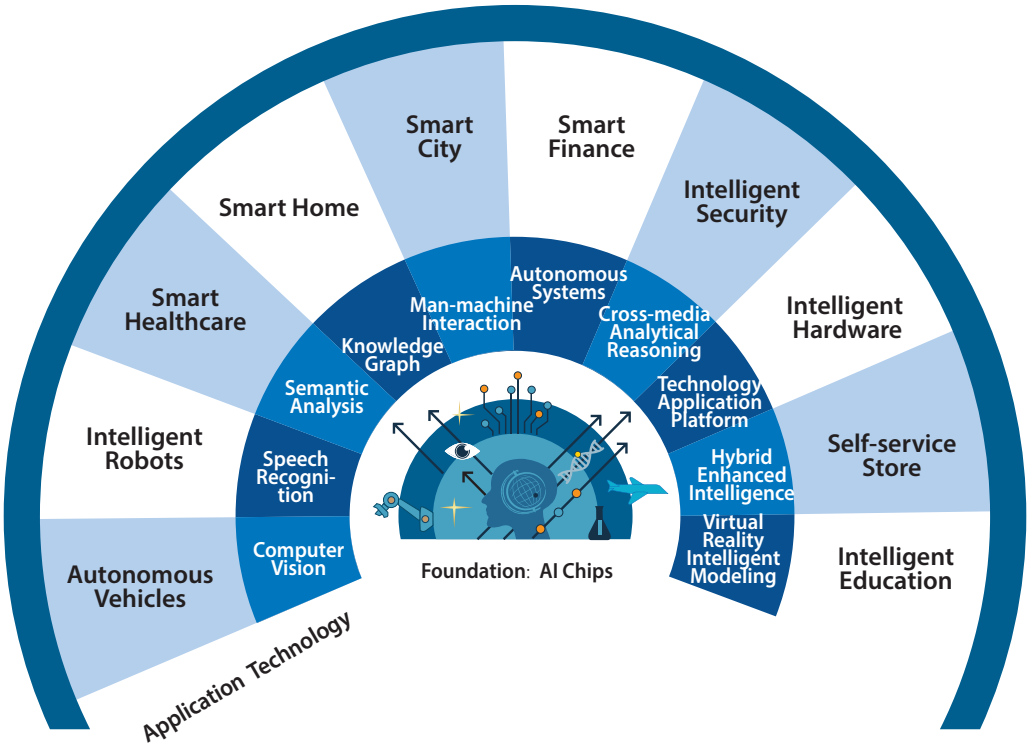


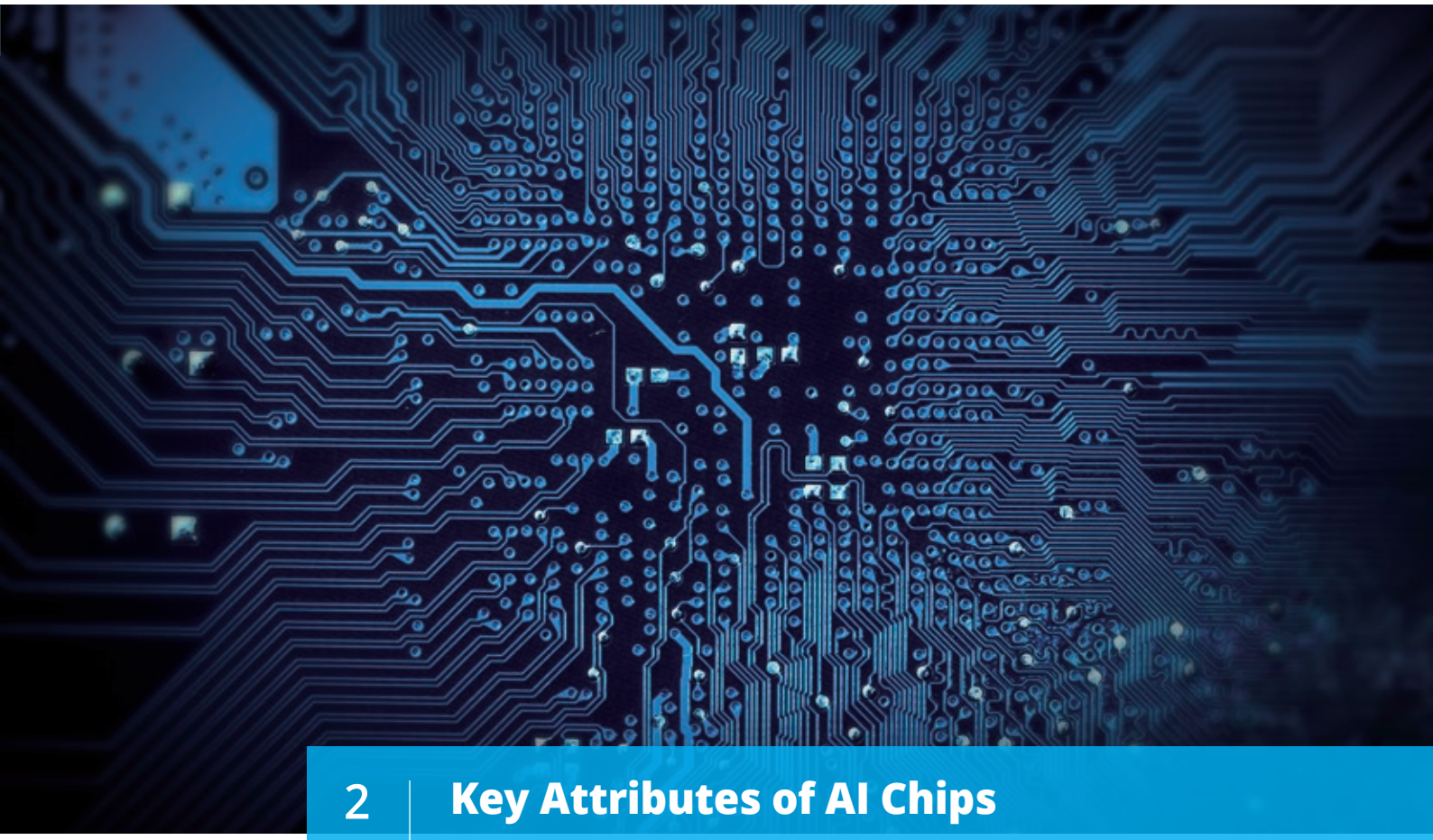

**Figure1-1** AI Industrial Structure and Technology Stack

In order to shed light on the present condition of AI chip development and to further clarify the future trajectory, critical phases, and application prospect against the backdrop of the latest technology trends, Beijing Innovation Center for Future Chips (ICFC), invites the top researchers around the world to jointly accomplish the "White Paper on AI Chip Technologies" in accordance with the latest practices of academia and industry. The work aims to inform readers about the competing technologies and the development trends of AI chips, and thereby gives momentum to the enhancement of the core AI technologies and other supporting capabilities, as well as making a contribution to an insightful understanding of the strategic opportunity driven by the AI industry development.

## 1.2 Contents

The work encompasses nine parts with each covering a specific area. Chapter 1 illustrates the strategic implications of the AI chip industry development and introduces the overall content of the work. Chapter 2 summarizes the technical background of AI chips and proposes the key features of AI chips and the relevant hardware platforms that satisfy various scenarios. Chapter 3 introduces the development of AI chips in a variety of settings such as cloud and edge devices in recent years. The chapter also summarizes the issues facing cloud and edge devices and lists the approaches for simultaneously deploying both cloud and edge devices to support AI applications. Chapter 4 analyzes the technological trends of AI chips in consideration of the architectural challenges faced by AI chips as the CMOS process feature size increasingly approaches the limit. Chapter 5 discusses cloud and edge AI chip architecture innovations based on the current CMOS technology integration. Chapter 6 introduces the storage technologies that are critical to AI chips, including the improvements in traditional storage technologies and the emerging non-volatile storage (NVM)-based memory solutions. Chapter 7 discusses the leading researches in processes, devices, circuits, and memory storage technologies, and the new trends of in-memory computing and biological neural networks based on these works. Chapter 8 introduces neuromorphic computation techniques and algorithms, models and key technical features, and analyzes the opportunities and challenges facing the technology. Chapter 9 focuses on discussing issues related to benchmarking tests and technology roadmaps for AI chips. Chapter 10 provides an outlook of the future of AI chips.

Against the backdrop of the AI boom, this work hopes to share the innovative achievements in the field of AI chips to the academia and industry actors worldwide, and calls for joint efforts to facilitate a rapid growth of the AI chip industry in part. Also, it hopes to provide a nuanced comprehension about the status, opportunities, and needs of AI chip industry. At the same time, it shares the prospect of AI chip development and points out that the AI technology is still at a nascent stage, a wide range of challenges are expected to arise in both technical and commercial spheres. It is necessary to refrain from impetuousness and short-term opportunist behaviors at a time when the industry is booming rapidly. To forge a steady, sustainable development of AI chip industry, it is necessary to be explorative and confident while remaining a calm and realistic attitude.



# 2 Key Attributes of AI Chips

## 2.1 Technology Overview

At present, there is no strict and widely accepted standard for the definition of AI chips. A broader view is that all chips for AI applications can be called AI chips. Nowadays, some chips based on traditional computing architectures are combined with various hardware and software acceleration schemes, which have achieved great success in some AI application scenarios. However, due to the diversity of requirements, it is difficult to have any single design and method that can be well applied to all kinds of situations. Therefore, many novel designs and methods for AI applications have emerged in academia and industry, covering all levels of semiconductor from materials, devices, circuits to architectures.

The AI chips discussed in this white paper mainly include three types: the first one is universal chips that can support AI applications efficiently through hardware and software optimization, such as GPU; the second one is chips that focus on accelerating machine learning (especially neural networks and deep learning), which is the most popular form of AI chips at present. The third one is the neuromorphic computing chips inspired by biological brain.

AI technology is multifaceted, which runs through the application, algorithm mechanism, chip, tool chain, device, process and material technology levels. These levels are closely linked to form the AI technology chain, as shown in Figure 2-1. The AI chip itself is in the middle of the whole chain, providing efficient support for applications and algorithms upwards and raising demand for devices and circuits, processes and materials downwards. On the one hand, the rapid development of applications and algorithms, especially deep learning and convolutional neural networks, has put forward requirements of 2-3 orders magnitude of performance


**Figure 2-1** Overview of AI Chip Related Technologies

optimization for AI chips, which has led to the upsurge of AI chip research and development in recent years. On the other hand, the rapid development of new materials, processes and devices, such as 3D stacked memory and process evolution, also provides the feasibility of significantly improving performance and reducing power consumption for AI chips. Generally speaking, these two kinds of power jointly promote the rapid development of AI chip technology in recent years.

## 2.2 New computational paradigms

Although AI computing is associated with traditional computation, it also has new computational characteristics, including:

1. The processing content is often unstructured data, such as video, image and voice, which is difficult to obtain satisfactory results through pre-programming. Therefore, it is necessary to train the models by means of sample training, fitting and environment interaction, and then use the trained models to process the data.

2. Processing usually requires a large amount of computation. The basic calculation is mainly linear algebraic operations, such as tensor processing, while the control process is relatively simple. For such operations, massively parallel computing hardware is more suitable than traditional universal processors.

3. Parameters are large, requiring huge storage capacity, high bandwidth, low latency memory access capacity, and rich and flexible connections between computing units and memory devices. Data localization is prevalent and suitable for data reuse and near memory computation.

## 2.3 Training vs inference

AI systems often involve training and inference for which the required computing resources are quite different. For artificial neural networks, training aims to minimize the error as a function of the parameters (i.e., weights) in a given neural network structure, and can be achieved either offline or online and with supervision or without supervision. Inference, on the other hand, is usually done online with straightforward evaluation of the network.

While having a large number of parallel functional units, high memory bandwidth, and low latency operations, are generally desirable for AI chips, training and inference, however, due to their own respective objectives, have notable differences in terms of needs for computing resources.

Training - First, computation accuracy is extremely important since it directly impacts the quality of the resulting neural network. To achieve expected accuracy, training hardware must support sufficient representation accuracy, e.g., floating-point or fixed-point with a longer word-length is often needed. In addition, error computation is often more computationally demanding and hardware for approximate error computation can be very helpful. Secondly, the training process, especially offline training, must deal with tremendous amount of data (as high as $10^{15}$ to $10^{18}$ bytes). Thus memory support including the amount of memory, memory bandwidth, and memory management approaches, is critically important. Third, since training requires both updating (writing) the weights and using (reading) the parameters in the neural network, more sophisticated synchronization techniques are needed in order to achieve high computational efficiency. Last but not least, the frequent writing of weight parameters also demands fast write time (particularly for online training), which for some memory technologies can be a significant challenge and must be taken into consideration in the hardware design.

Inference - Speed, energy and hardware cost are important considerations, and can be traded off with inference accuracy. Hardware designed specifically for inference needs to carefully explore the design space defined by these attributes. Fixed-point representation with relatively short word-length is often sufficient for inference hardware. Efficient access (reading) of the large number of weights is another key issue. Co-design of the network structure and hardware architecture is indispensable in achieving the most efficient inference hardware systems.

Although most machine learning methods can be clearly divided into training and inference processes, there are still some areas, such as Reinforcement Learning and On-line Learning, which are in the process of continuous evolving and model improvement. Therefore, in future AI applications, training (learning) and inference will be intertwined in more scenarios.

## 2.4 Being able to handle big data

AI's recent progress has heavily relied on the availability of big data. Its future development will be relying on even bigger data. Meeting the heavy data crunching requirements of machine learning with high power efficiency are among the most important consideration for AI chips. Even worse, the increasing gap between the computing of the processing units and the memory has created the "memory wall" problem in which the

memory sub-system becomes the bottleneck of the AI chips, while many AI workloads are data-intensive and mandate very high bandwidth and heavy data movement between the computing logic and the memory. To close the gap between the computing units and the memory, two possible approaches could be explored: (1) Memory rich processing units which increase the capacity of the on-chip memory and brings it closer to the computing units, so that the data movement between computing units and memory can be dramatically reduced. (2) Compute-capable memory which tries to move the computation inside (or closer) to memory. This approach has also been referred to as processing-in-memory (PIM) or near-data computing (NDC). For example, dot-product operations could be implemented inside or closer to the memory so that part of the neural computation can be done without moving the data to the processing units.

## 2.5 Precision for data representation

An interesting trend of emerging AI chips is that low-precision designs become increasingly prevailing. For some applications, reduced precision designs not only accelerate the inference (and possibly training too) of machine learning algorithms, but also might even conform better to the neuromorphic computing flow. It has been demonstrated recently that, for certain parts of the learning algorithms and networks, using the lowest possible precision, e.g. binary data, might be sufficient for achieving desirable accuracy while saving enormous amount of memory and energy consumption. Exploring architectures and designs of AI systems which can adapt and configure dynamically by examining data precision with respect to data context and sensitivity to rounding errors for application accuracy would be a necessary strategy for design space exploration of AI design optimization.

## 2.6 High configurability

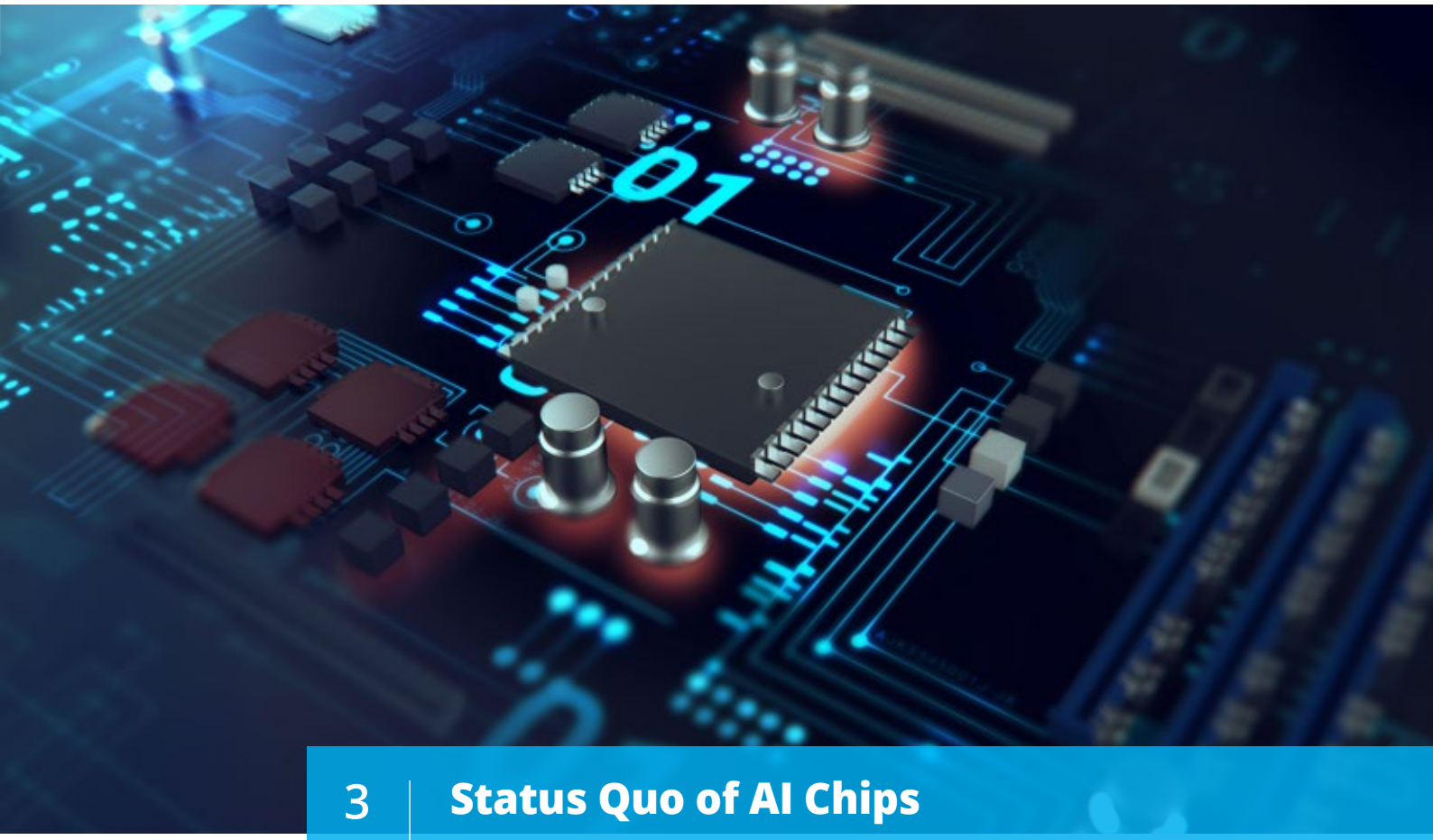
A specialized design targeting a specific application, algorithm or situation is not ideal due to the accelerating evolution of the field, continuing emergence of new algorithms and applications, and high cost of designing a new chip. Domain-specific, rather than application-specific, would be the right guiding principle for designing AI-chips, so that a new chip can be used more broadly across multiple applications and can be reconfigured to accommodate a broader range of AI algorithms, architectures and tasks.

## 2.7 Software toolchain

Like CUDA for GPU and GCC for CPU, AI chips also need supports from a software toolchain to efficiently translate different machine learning tasks and neural networks into executable binary codes, running on the AI chips. The proper assignment and scheduling of process elements, memory access and tasks will be the key factors to be considered in the toolchain.

There are plenty of code optimizing opportunities when a toolchain maps models and neural networks onto AI chips. Neural network pruning, weight compression and dynamic quantization are among the possible areas for optimization.

Finally, there are several platforms for AI algorithm development, such as TensorFlow, Caffee, etc. Building an integrated flow, which can seamless combine the AI model development and training, hardware-independent and -dependent code optimizations, and automatic instruction translation to AI chips, will be a key requirement for successful deployment of AI chips.



# 3 Status Quo of AI Chips

Since 2015, the R&D of AI chips has gradually become a hotspot in academia and industry. So far, there are many chips and hardware systems specifically designed for AI applications in the cloud and edge. At the same time, according to the target application ("training" or "inference"), we can divide the area of AI chips into four quadrants, as shown in Figure 3-1. Among them, the application of inference is mainly in edge / embedded devices, and the demand for training is not very clear. Some high-performance edge devices are also used for training, but in terms of hardware itself, they are more like cloud devices. Both future edge and embedded devices may need some learning ability to support online learning. The other quadrants have their own requirements and constraints, and there are also specific chips and hardware systems.

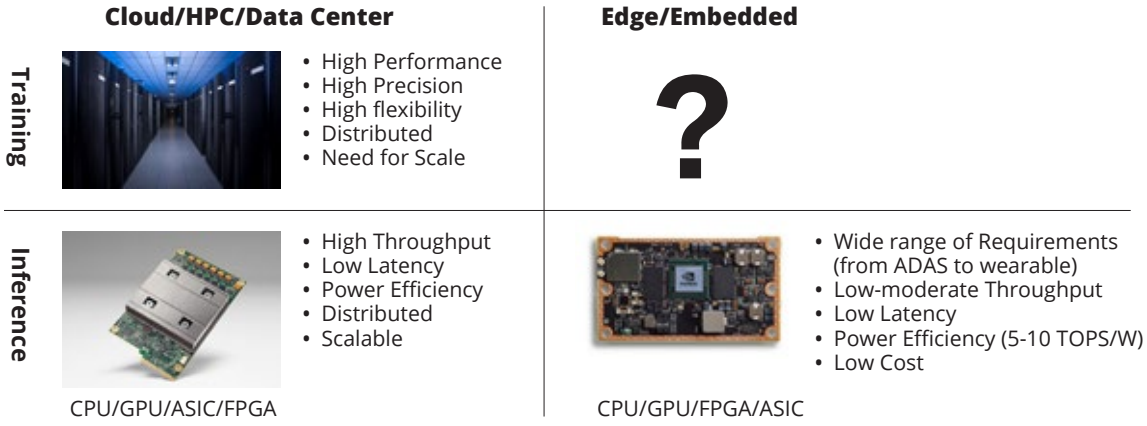



**Figure 3-1** AI Hardware Target Domains

## 3.1 Cloud AI computing

In the cloud, GPU, especially NVIDIAs' series GPU chip, have been widely used to do 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 main stream for machine learning for many of the largest web and social media companies. NVIDIAs' Tesla V100 is specially designed for Deep learning which incorporates Tensor Cores with GPU cores, could provide 120 TFLOPS (120 trillion floating-point instructions per second) processing power. Moreover, NVIDIAs' GPU also has an excellent software development environment, which is one of the most widely used platforms in the field of AI training.

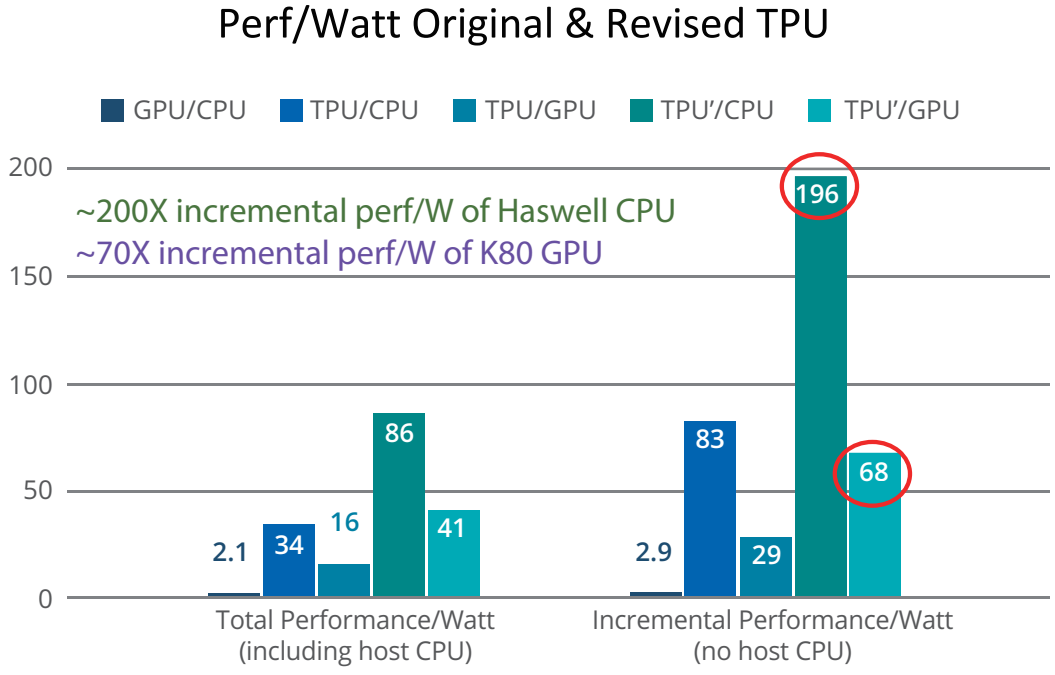



**Figure3-2** From Cliff Young of Google in Hot Chips 2017

Further 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, translation, and even AlphaGo match. Performance comparison of TPU, GPU, and CPU are shown in Figure3-2. Google announced the second version of their TPU (TPUv2) for neural net training and inference with 180 teraflops of computation, 64GB of HBM memory, and 2400 GB/s memory bandwidth.

Both chip giants and startups are attaching great importance to the market of cloud training and inference. More companies are releasing specialized AI chips. Intel announced its Nervana Neural Processor (Lake Crest family) which was also architected to optimize neural network computing with 32 GB HBM2, 1 TB/s Bandwidth, and 8 Tb/s Access Speeds. Startups such as Graphcore, Cerebras, Wave Computing, Cambrian and BitContinent also joined the competition.

Even higher parallelism can be achieved by using FPGAs that include millions of parallel system logic cells. It has become popular for data centers (e.g.: Microsoft Brainwave and Baidu) due to its excellent inference performance at low batch sizes. The FPGA can provide Inference-optimized numerical precision to exploit sparsity and deep compression for larger and faster models. The FPGAs were used in many and evolving machine learning algorithms

such as CNNs, LSTMs, MLPs, reinforcement learning, feature extraction, decision trees, etc. To support ever increasing sizes of dataset, all state-of-the-arts CPUs and FPGAs incorporated high-band-width-memory (HBM) and high-speed interconnects enabled by heterogeneous integration.

Major manufacturers of FPGA, such as Xilinx and Intel, have released specialized hardware (supporting higher storage bandwidth) and software tools for AI applications. Major cloud service providers, such as Amazon, Microsoft and Aliyun, have introduced specific cloud-based FPGA solutions to support AI applications. Some start-ups, such as DeePhi Tech are also providing software developing tools to support FPGA.

## 3.2 Edge AI computing

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. Another example is the face recognition task with millions of HD cameras in large cities. If the task is handed over to the cloud, the camera data transmission will overwhelm the communication network.

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 can not meet the need of local inference. Therefore, industry must empower more AI chips at edge to be applied in different AI application scenarios.

Many companies are developing specialized AI chips and IP for smart phones, drones, industry and service robots, intelligent cameras and speakers, and all kinds of IoT devices. Among these companies are Apple, Huawei, Qualcomm, ARM, CEVA, Cambrion, Horizon Robotics, BitMain. Unlimiter Hear, Synopsys etc.

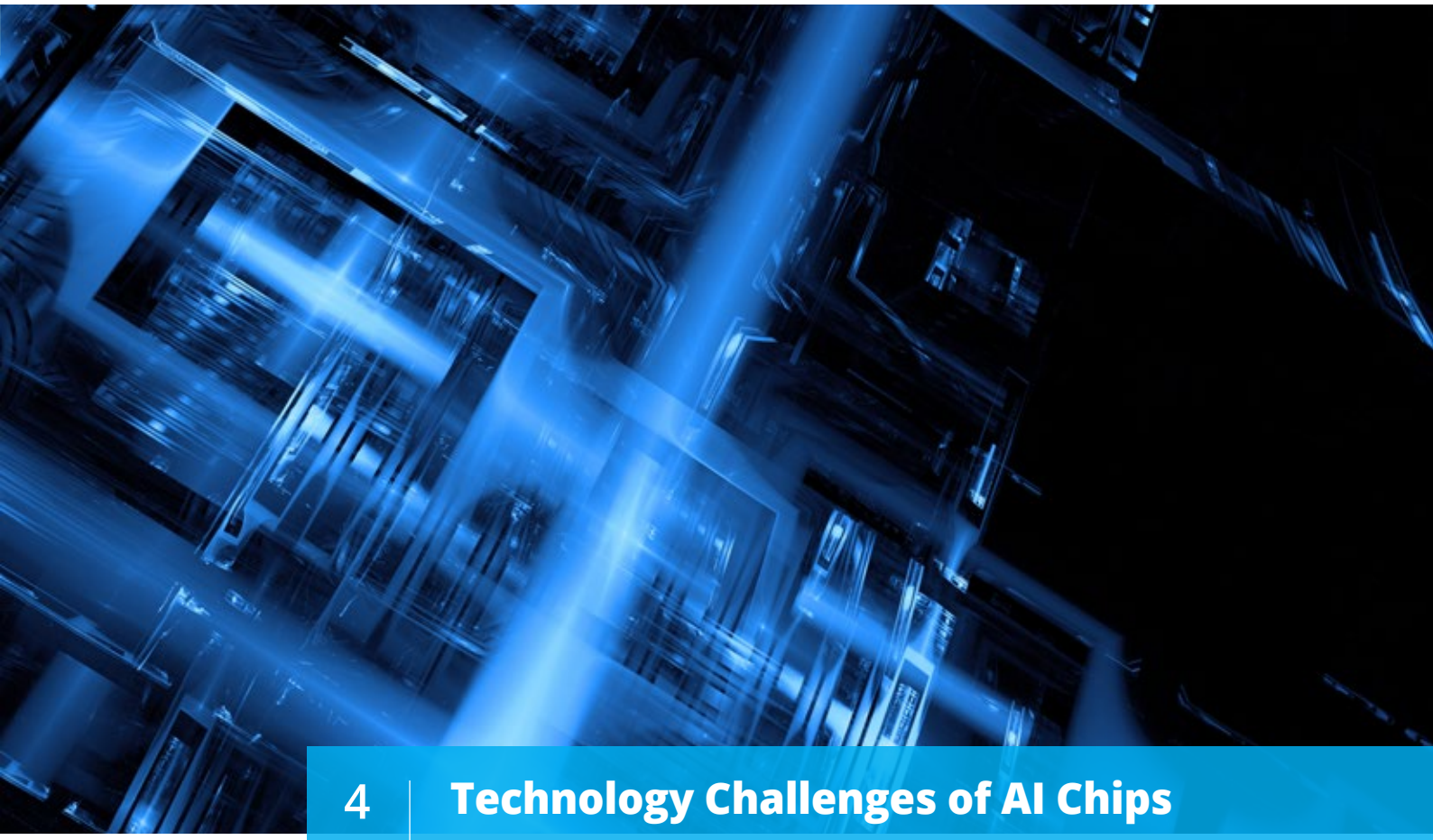
MobileEye SOCs and NVIDIA Drive PX family provides the processing power to support

semi and fully autonomous driving, having the bandwidth/throughput to stream and process the full set of surround cameras, radars and LiDARs.

## 3.3 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, cost and privacy issues.

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 an interesting direction to be explored.



# 4 Technology Challenges of AI Chips

When we discuss a new technology trend, we need to know the causes behind it first. Many large technological innovations are generated by both the market demand and the technological bottlenecks. The AI chip and the various technologies behind it are no exception. First of all, the huge demand, is reflected not only in the demand of AI applications, but also in the new computing paradigm required by AI, especially deep learning (which has been introduced earlier). The demands require more efficient hardware to processe AI computing, and we also encounter some bottlenecks in the current technical framework, especially the von Neumann bottlenecks, CMOS process and device bottlenecks. Before introducing the technological innovation and future development trend of various AI chips in detail, this section first briefly discusses these two issues.

## 4.1 Von Neumann bottleneck

Artificial intelligence (AI) chips have demonstrated their superior capabilities on a broad range of recognition and classification tasks and thus become one of the key enablers for the Internet of Things (IoT) systems and big data processing. One of critical keys to boost the performance and energy efficiency of AI chips is rooted in data accessing though the memory hierarchy, as shown in Figure 4-1. In conventional von-Neumann architecture, the data is serially fetched from the storage and driven to the working memory, which results in considerable latency and energy overhead in AI chips. The common method is to use hierarchical storage technology such as Cache to alleviate the speed difference between computing and storage.

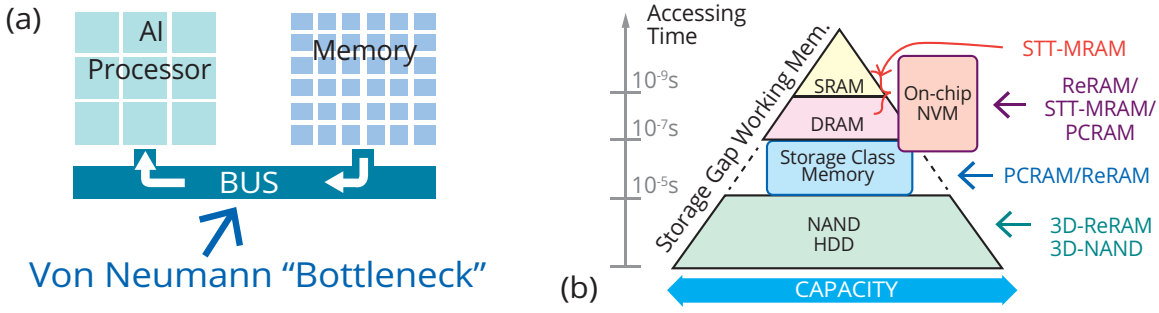



**Figure 4-1** (a) Von Neumann "bottleneck" in AI chips and (b) the memory hierarchy

| Metrics | AlexNet | VGG 16 | GoogLeNet V1 | ResNet 50 |
|---|---|---|---|---|
| Top-5 error | 16.4 | 7.4 | 6.7 | 5.3 |
| Of CONV Layers | 5 | 13 | 57 | 53 |
| Weights | 2.3M | 14.7M | 6.0M | 23.5M |
| MACs | 666M | 15.3G | 1.43G | 3.86G |
| Of FC Layers | 3 | 3 | 1 | 1 |
| Weights | 58.6M | 124M | 1M | 2M |
| MACs | 58.6M | 124M | 1M | 2M |
| Total Weights | 61M | 138M | 7M | 25.5M |
| Total MACs | 724M | 15.5G | 1.43G | 3.9G |

**Chart 4-2** Summary of Popular DNNs （source：[Vivienne17]）

However, the amount of data that need to be stored and processed in AI chips is much larger than that used in common applications before. Chart 4-2 lists the main parameters of some common deep neural networks, in which VGG16 network requires 138M total weights and 15.5G total MACs for one inference process. This makes the Von Neumann bottleneck more and more serious in AI applications. It is no exaggeration to say that most of the hardware architecture innovations for AI, especially for accelerating neural network processing, are struggling with this problem. Generally speaking, there are two basic ideas to find a way out of the dilemma at the architecture level: 1) reduce the number for accessing memory, such as reducing the storage need of neural networks (number of parameters, data accuracy, intermediate results), compressing data and supplanting storage with computing etc.; 2) reduce the cost of accessing memory and try to close the "distance" between storage devices and computing units, and even directly operate in storage devices.

## 4.2 Bottlenecks of CMOS process and device

Current computers are capable of petascale (1015 FLOPS) performance. These systems are playing a critical role in advanced research (biology, climate analysis, genomics, brain, genome, materials development, …), commercial and non-commercial applications. In many aspects, computing fuels advances of our modern society. Recent developments in AI and machine learning will require even more powerful computing systems, in excess of exascale (1018) computations per seconds.

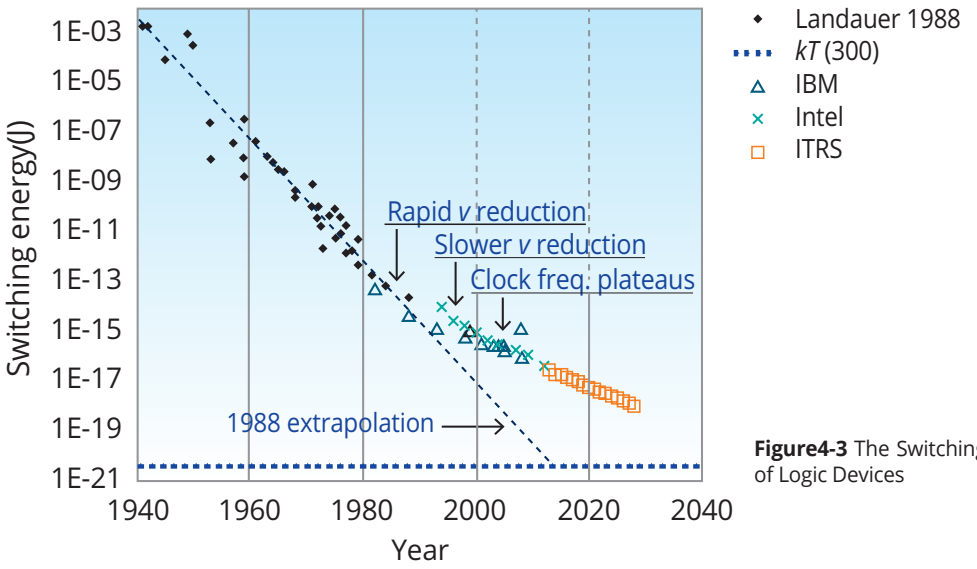


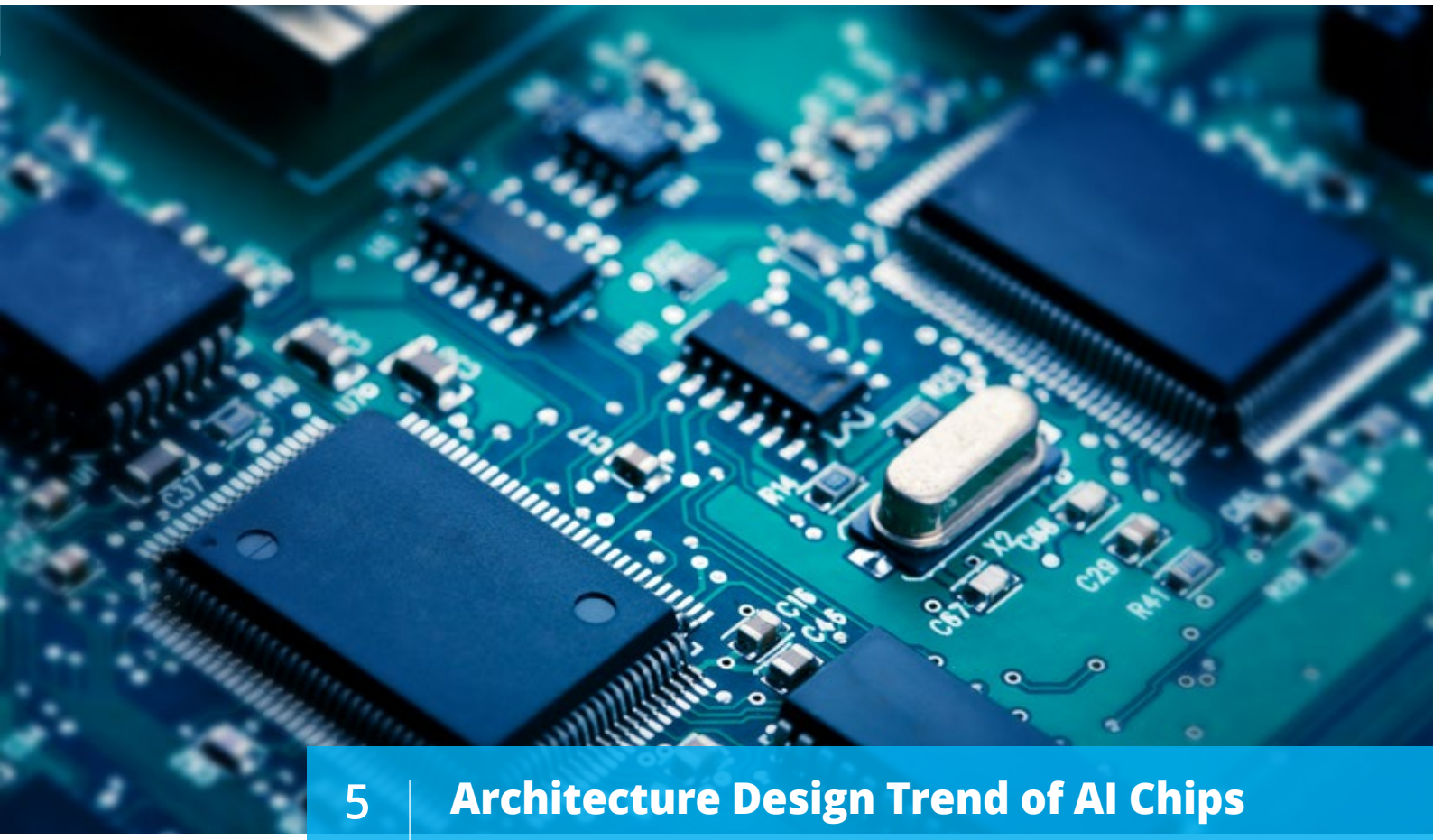

**Figure4-3** The Switching Energy of Logic Devices

The dominating technology to build these systems is CMOS that has benefited significantly from aggregative dimensional scaling of silicon CMOS. Over the past 30 years, Moore's law has predicted this progress very well. Major companies are ready for 10 nm node CMOS production in 2018, and technology definition for 7 nm and 5 nm nodes have already been finalized. However, continued scaling is not going to be feasible for several fundamental physical and economic reasons [Theis 16]. The lateral dimensions of CMOS devices are approaching a few nanometers with layer thicknesses of a few atomic layers. This causes significant leakage current what reduces the effectiveness of CMOS scaling. In addition, the energy consumption of these nanoscale transistors gets prohibitively high preventing the realization of densely packaged systems.

The Internet of Things (IoT), social media and security devices also generate a large amount of data, which requires a large amount of memory to store, exchange and process. DRAM scaling has largely saturated and advances in DRAM performance and memory capacity is mainly provided through packaging technologies that stack multiple DRAM chips and connect them by through silicon-vias (TSVs) to a logic chip that serves as the memory controller. Increased data bandwidth is achieved by having a wide data bus. NAND flash is the dominant data storage technology. 3D NAND with up to 64 layers and 256 Gb capacity is positioned to enter the market in 2018. Both DRAM and NAND flash are stand-alone chips that is off-chip from the computing cores. Therefore, the cost of data exchange with computing core (including time and energy consumption) is very large. For on-chip memory, the only solution today is SRAM, which as MB capacities. For on-chip memory, the only solution today is SRAM, which as MB capacities. Even if one populates half of a 1 cm2 chip with the smallest SRAM cell, there can only be ~128 MB SRAM on-chip. So there are very good reasons [Aly15] to develop future chip technologies that provide massive on-chip memory and explore future smart chip architectures that capitalize on the available on-chip memory that is not available in today’s semiconductor technologies.

A new, disruptive approach will also be needed both in terms of computing architecture and devices operation. The brain can serve as a computational model of how to deal with large amounts of (often fuzzy) information while being extremely dense, error resilient, and power efficient. E.g. the energy dissipation in neurons and synapses in the brain are several orders of magnitude lower than the most advanced CMOS devices. In addition, the brain can deal with a class of problems such as pattern recognition and feature extraction that is hard, if not impossible, for conventional architectures to achieve in real-time.

Ideally, one needs devices and materials that have the advantages of biological systems without any of its downsides such as slow component speed. Non-volatile memories that can store analog values are a recent development that shows promise as a possible implementation for brain-like synapses. It has processing and memory capabilities that can solve some of the fundamental limitations of conventional computing architectures. More importantly, these novel non-volatile memories can be fabricated at low temperatures, making them suitable to be integrated with the computing logic in 3D, potentially resulting in computing systems that have large amounts of memory connected with computing logic in a fine-grain fashion.



# 5 Architecture Design Trend of AI Chips

## 5.1 Cloud training and inference: big Storage, high Performance and scalability

As we mentioned before, although training and inference differ in data accuracy, architecture flexibility and real-time requirements, they have similar requirements in processing power (throughput), scalability, and power efficiency. Therefore, the development of dedicated chips and technological innovations for cloud training and inference are basically addressing these needs.

NVIDIA's V100 GPU and Google's Cloud TPU, which include four chips[Google], are the benchmarks for commercial AI chips in the cloud. In terms of processing power of deep learning computing, V100 achieves 120 TFLOPS, while Cloud TPU achieves 180 TFLOPS. It is worth mentioning that this processing capability is provided by the operation units designed specifically for deep learning. For storage and accessibility, V100 has 16 GB HBM2 memory to support 900 GB/s bandwidth, while Cloud TPU single chip has 16 GB HBM memory to support 600 GB/s bandwidth. In addition, they both could support the scalability of multi-chip. V100 supports NVIDIA's NvLink interconnection mode, and can be extended to 8-chip systems. Cloud TPU also supports high-speed inter-chip interconnection interface and board-level interconnection interface, which is very suitable for deployment in cloud and data center. Cloud TPU cabinets (Figure5-1), including 64 TPU2, provide 11.5 PFLOPS processing power and 4 TB HBM memory for training tasks of machine learning. At the same time, these computing resources can also be flexibly allocated and scaled to effectively support different application requirements.

**Figure 5-1** Google Cloud TPU Pod (Hot Chips 2017)

From the design practice of NVIDIA and Google, we can see several characteristics and trends of technology development in in architecture for cloud-based AI chips:

1. Storage requirements (capacity and access speed) are getting higher and higher. On the one hand, because of the requirement of handling large amounts of data, larger capacity memory is needed. On the other hand, the main factor limiting the increase of computing power is the speed of accessing memory. As a result, cloud AI chips will have more and more on-chip memory (such as Graphcore's 300 MB SRAM) and off-chip memory (HBM2 and other new types of advanced packaging) that can provide high bandwidth.

2. The processing power is pushed to PetaFLOPS (100 million times per second) and supports flexible scalability and deployment. For cloud AI chips, the processing power of single chip may reach PetaFLOPS level. Achieving this goal depends not only on the progress of CMOS technology, but also on the architecture innovation. For example, in the first generation of Google TPU, Systolic Array architecture is used, while in NIVIDIA's V100GPU, tensor kernels are included to handle matrix operations. In order to extend the GPU to a larger system, NVIDIA developed the NVSwitch switch chip, which can provide high bandwidth interconnection for multiple GPUs. In the newly released DGX-2 system, 16 V100 GPUs are connected together to provide processing capability of 2PFPLOS, which can realize the parallel training of large-scale neural networks. In addition, we see some more ‘extreme’ architecture designs. For example, wafer-level integration technology, which uses the whole wafer to make a "super chip"; or using clock-free circuit in the computing unit to achieve higher speed and lower power consumption. there is also another way to achieve stronger computing and storage capacity through the interconnection of multi-chip and multi-board, rather than simply pursuing the processing capacity of a single chip. In the future, we should see more and more products providing scalable and configurable processing capabilities in the form of systems (or cloud services) rather than single chips. The flexibility of this powerful processing power, is also reflected in the training and inference task deployment, such as more hardware resources are allocated to inference tasks during the day time, while these resources are allocated to training tasks at night.

3 FPGA and ASIC specific for inferring requirements. With the outbreak of AI applications, there will be more and more demands for inference computing, and a well-trained algorithm will be reused continuously. Compared with training, inference has its own particularity, which emphasizes throughput, energy efficiency and real-time. In the future, there will probably be ASIC chips in the cloud specially for inference computing (the first generation TPU of Google is a good example), providing better energy efficiency and achieving lower latency. FPGA also has unique advantages in this direction, which is indicated by Microsoft's BrainWave architecture.

## 5.2 Edge device: pushing efficiency to the extreme

Compared with cloud applications, the application requirements and scenario constraints of edge devices are much more complex, and special architecture design may be needed for different situations. In spite of the complexity of demand, the current edge devices mainly perform "inference". Under this goal, the most important thing for AI chips is to improve the efficiency of inference. At present, TOPs/W, an important index to measure the efficiency of AI chip performance, has become the focus of many technological innovation competitions. At the ISSCC2018 meeting, the single bit energy efficiency is said to reach an amazing level of 772 TOPs/W[Bankman18].

Among all the methods that can improve the inference efficiency and the inference accuracy in a permissible range, reducing the quantization bit accuracy of the inference is the most effective one. It can not only greatly reduce the accuracy of the operation unit, but also reduce the storage capacity requirements and reading and writing operations of the memory. However, lower bit accuracy also means lower inference accuracy, which is unacceptable in some applications. Therefore, the design trend of the basic arithmetic unit is to support variable bit precision. For example, the BitMAC of [Lee18] can support the weight accuracy from 1 bit to 16 bits.

In addition to reducing accuracy, reducing the computational complexity by combining some data structure transformations could also improve the efficiency of basic computing units (MACs). For example, fast Fourier transform (FFT) is used to reduce the multiplication in matrix operations (see [Vivienne 17]), and table lookup is also used to simplify the implementation of MAC.

For neural networks using modified linear unit (ReLU) as activation function, there are many cases where the activation value is zero, and after pruning the neural network, there will be many zero weights. Based on this sparsity feature, on the one hand, we can use special hardware architecture, such as the SCNN accelerator proposed in [Parashar17], to improve the efficiency of MAC, on the other hand, we can compress the weights and activation values [Vivienne17].

Another important direction is to reduce the access to memory, which is also the basic way to alleviate von Neumann's "bottleneck" problem. Another way is to close the distance between computing and storage, that is, the concept of "near data computing", such as putting neural network operations in sensors or in memory. For the former, there has been a lot of work trying to put the calculation in the analog part of the sensor, so as to avoid the cost of analog-to-digital conversion (ADC) and data transfer. Another trend is to preliminarily process the sensor data to reduce the amount of data storage and move. For example, a simple neural network is used to locate the target object based on the data obtained from the image sensor, and then only the object is stored and transmitted to the complex neural network for object recognition. The latter, that is, in store computing, will be discussed in detail later in the new memory technology.

Moreover, various low power design methods can be applied to AI chips of edge devices to further reduce the overall power consumption. For example, when the weight or the value of the intermediate result is zero, Clock-gating is applied to MAC. The frequency adjustment of dynamic voltage proposed by [Bert17] adds the consideration of inference accuracy to the traditional chip dynamic power adjustment technology. Furthermore, asynchronous design (or clock-free design) is currently used in some computing units to reduce power consumption, which is also a direction worth exploring.

In the future, more and more edge devices will need to have a certain "learning" ability to train, optimize and update models locally based on the collected new data. This will also put forward some new requirements for edge devices and the entire AI implementation system.

Finally, the AI chips in edge devices are integral parts of the whole SoC systems, and ultimately the efficiency of hardware should be reflected through the complete chip functions. In this case, we need to consider the optimization of the architecture from the perspective of the whole system. Therefore, the product AI chips are often presented as a heterogeneous system. Special AI accelerators and other components such as CPU, GPU, ISP, DSP work together to achieve the best efficiency.

## 5.3 Software-defined Chips

In AI computing, the chip is the basic component that supports the computing function, and software is the core to realize AI. The software here is the AI algorithms needed to fulfil the AI tasks of different objectives. For complex AI tasks, even a variety of different AI algorithms need to be combined. Even the same type of AI algorithm will have different parameters because of the different requirements of the specific task, such as accuracy, performance and efficiency. Therefore, AI chips must have the important feature: they could dynamically change their functions in real-time to meet the ever-changing computing needs of the software, that is, software-defined chips.

Universal processors such as CPU and GPU lack specific design of computing and memory unit for AI algorithm, which has low energy efficiency. ASIC is a single-function processor and has difficulty in adapting to flexible and diverse AI tasks. Field Programmable Gate Array (FPGA) can be reconstructed into different circuit structures by programming, but the time cost of reconfiguration is too large, and too much redundant logic leads to its high power consumption. The above traditional chips can hardly achieve the "software defined chip" feature required by AI chips.

Reconfigurable computing technology allows hardware architecture and functions to change with software. It has the features of flexibility, high performance and low power consumption of application-specific integrated circuit. It is also regarded as the core of software defined chip and a breakthrough of next generation IC technology. The AI chip (named Thinker [Shouyi17, Shouyi18]) designed by Tsinghua University's Microelectronics Institute adopts reconfigurable computing architecture, which can support many AI algorithms such as convolutional neural network(CNN), fully connected neural network (CNN) and recursive neural network(RNN). The Thinker chip achieves "Software Defined Chip" through three levels of reconfigurable computing technology, with the highest energy efficiency of 5.09 TOPS/W:

1. Computing array reconfiguration: the computing array of Thinker chips is constituted by multiple interconnected parallel computing units. Each computing unit can perform functional reconstruction according to the different basic operators needed by the algorithm. Moreover, in complex AI tasks, computing resource requirements of many AI algorithms are different, and Thinker chips could support on-demand resource partitioning of computing arrays to improve resource utilization and energy efficiency.

2. Memory bandwidth reconfiguration: The on-chip memory bandwidth of Thinker chips can be reconstructed according to the difference of AI algorithms. The data distribution in the memory will be adjusted with the change of bandwidth to improve data reusability and computing parallelism, thus to improve computing throughput and energy efficiency.

3. Bit-width reconfiguration: 16 bit-width is enough to meet the accuracy requirements of most applications, even 8 bit-width is enough for some low-precision scenarios. In order to meet the various precision requirements of AI algorithms, the computing unit of Thinker chip supports high and low (16/8 bits) bit width reconstruction. Computing accuracy is improved in high bit mode, and computing unit throughput is improved in low bit mode, thus improving the performance.

As an important technology to realize software-defined chip, reconfigurable computing technology is very suitable for AI chip design. By using reconfigurable computing technology, the level of software define is not limited to function. The accuracy, performance and energy efficiency of the algorithms can also be incorporated into the scope of software define. Reconfigurable computing technology realizes hardware and software co-design by virtue of its real-time dynamic configuration characteristics, which brings high flexibility to AI chips and broadens its application scope.

| Technology | TSMC 65nm LP |
|---|---|
| Supply | 0.67V~1.2V |
| Area | 4.4mm*4.4mm |
| SRAM | 348KB |
| Frequency | 10~200MHz |
| Peak performance | 409.6GOPS |
| Power | 4mW~447mW |
| Energy efficiency | 1.6TOSP/W~5.09TOPS/W |

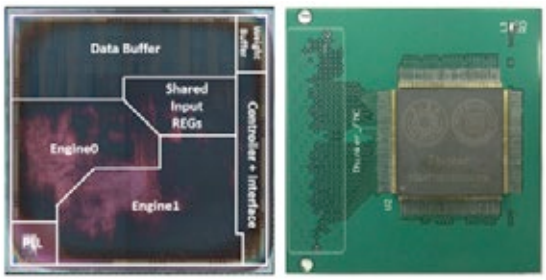
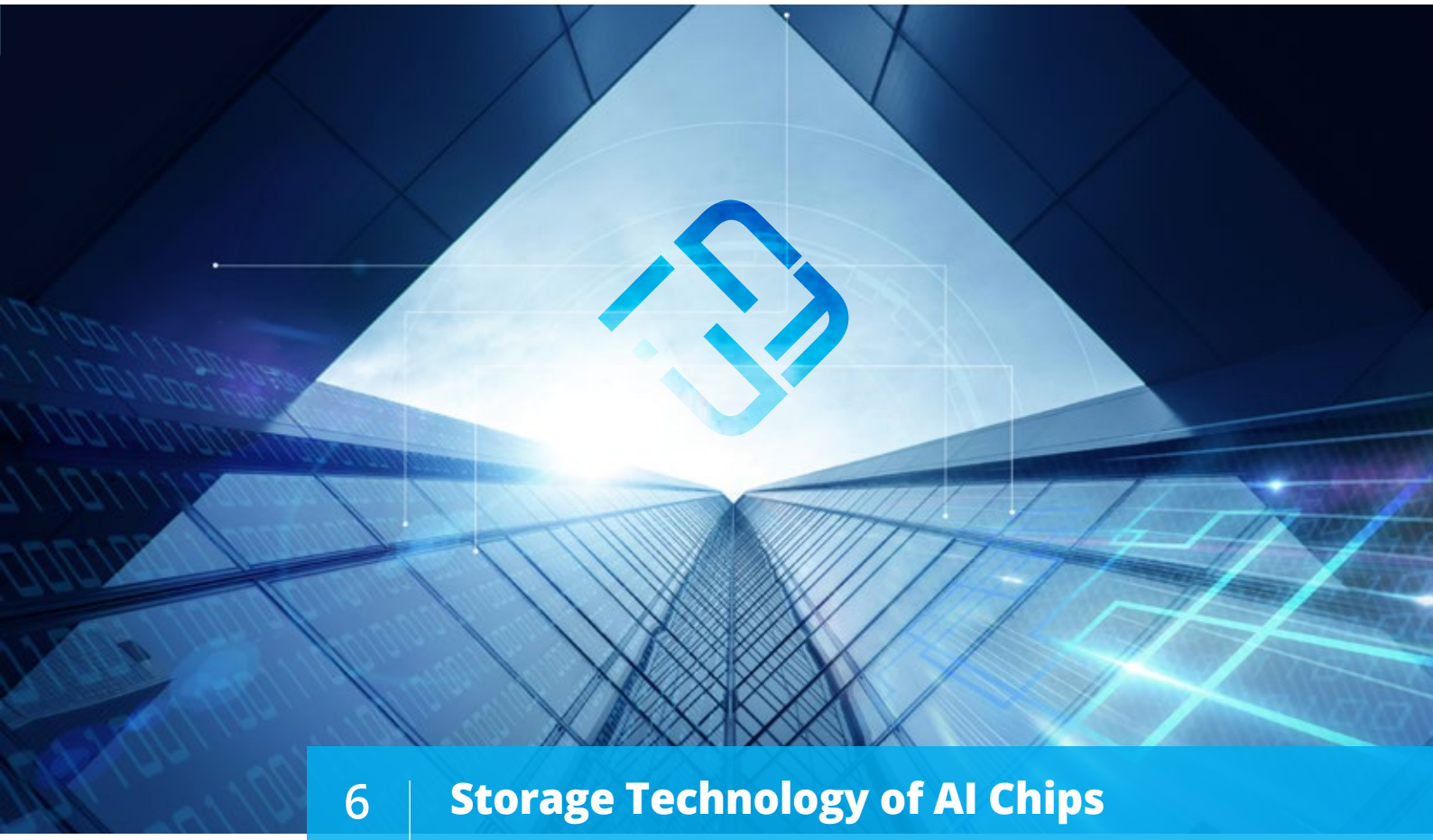



**Figure 5-2** Thinker Chip of Tsinghua University



# 6 Storage Technology of AI Chips

Artificial intelligence (AI) chips have demonstrated their superior capabilities on a broad range of recognition and classification tasks and thus become one of the key enablers for the Internet of Things (IoT) systems and big data processing. One of critical keys to boost the performance and energy efficiency of AI chips is rooted in data accessing though the memory hierarchy. In conventional von-Neumann architecture, the data is serially fetched from the storage and driven to the working memory, which results in considerable latency and energy overhead in AI chips. Comprehensive innovations from devices to architectures are expected to empower the AI chips. In the near-term, AI-friendly memories are urgently desired by current AI chips based on digital neural networks and accelerators (GPU, FPGA, and ASIC). In the mid-term, neural network based on in-memory-computing can provide an effective solution to circumvent the von Neumann Bottleneck. Lastly, memristor based neuromorphic computing that can mimic human brains have demonstrated promises as one of candidates for long-term solutions to AI chips.

## 6.1 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. Figure 4 shows the rapid growth of bandwidth and capacity of current major memory technologies. 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.

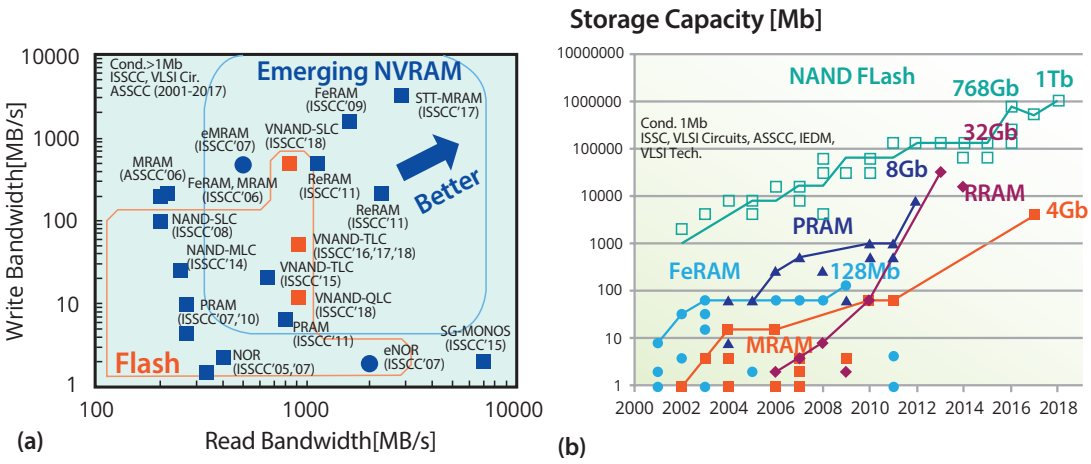



**Figure 6-1** (a) Bandwidth and (b) Storage Capacity of Flash and Emerging Memory [ISSCC Trend]

## 6.2 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 6-2 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.

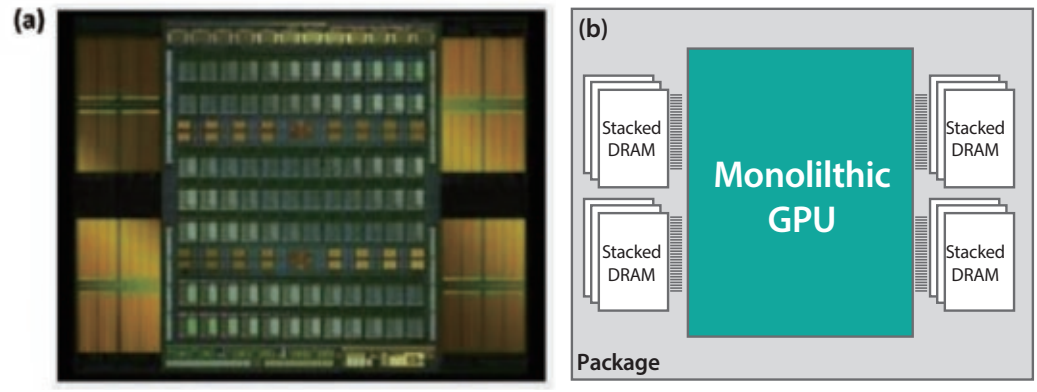



**Figure 6-2** (a) Die Photo and (b) Conceptual View of NVIDIA's GPU with High Bandwidth Memory (HBM) for Data Center Applications[NVIDIA]

## 6.3 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

| Device Metric | SRAM | DRAM | NAND | NOR | PCM | STT-MRAM | ReRAM |
|---|---|---|---|---|---|---|---|
| Write Energy | low | low | high | high | medium | medium | medium |
| Write Latency | ~1ns | ~5 ns | > 100μs | 10μs~1ms | 100~150ns | 2~200ns | 10~100ns |
| Read Latency | ~1ns | 20~80ns | 5~200μs | ~50ns | ~50ns | 1.3~25 ns | 3~200ns |
| Program Window | Good | Good | Good | Good | Variable | Small | Variable |
| Endurance | Unlimited | Unlimited | $10^4$-$10^5$ | $10^4$-$10^5$ | $10^8$-$10^9$ | ~$10^{15}$ | $10^5$-$10^{10}$ |
| Cell Size | ~100 $F^2$ | ~7 $F^2$ | ~4 $F^2$ | ~10 $F^2$ | ~4$F^2$ | ~12$F^2$ | ~4-6$F^2$ |

**Chart 6-3** Device Metrics of Current Major and Emerging Memories

## 6.4 Emerging memory

Emerging NVMs can significantly contribute to AI-friendly memory both for commodity and embedded applications. For commodity memory, emerging NVMs can serve as storage class memory (SCM) to bridge the accessing time gap between the working memory and the storage 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. For embedded applications, on-chip memory based on emerging NVMs can also provide better access speed and low power over conventional NVM, which is especially attractive for AI chips on IoT edge devices that work with very limited power availability.

# 7 Emerging Computing Technologies

Emerging computing technologies have been proposed and studied to mitigate or avoid the bottleneck of von Neumann architecture in the current computing technologies. The major new computing technologies include near-memory computing, in-memory computing and neuromorphic computing. While the mature CMOS devices have been utilized to implement these new computing paradigms, emerging devices are expected to further significantly improve the performance and reduce the circuit complexity in the future.

## 7.1 Near-Memory computing

In addition to placing the logic circuits or processing units (PUs) in proximity of memory and connect them with wide buses to minimize the delay and power caused by data transfer and increase the bandwidth, near memory computing further evolves by placing memory layers on top of logic layers to enable high performance parallel computing. Emerging NVM can also be suitable for this approach because it can be integrated with logic devices though CMOS back-end-of-line (BEOL) process.

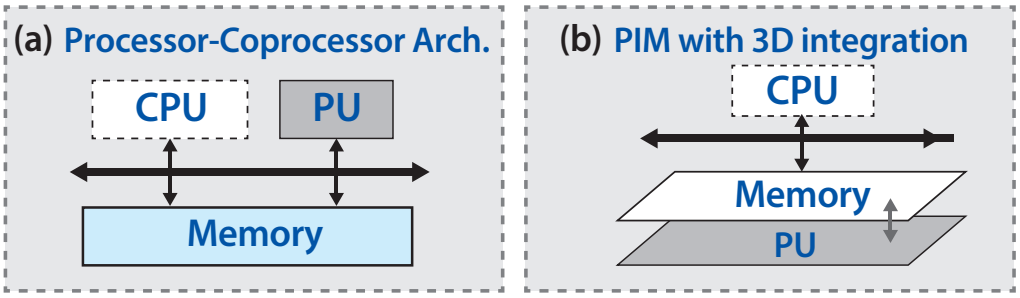



**Figure 7-1** (a) Traditional Von Neumann Architecture and (b) Near-memory Computing [Chi16]

## 7.2 In-memory computing

In-Memory-Computing signifies a fundamentally different approach compared to a conventional von Neumann architecture, because it directly carries out the computation inside the memory without requiring data transferring. Recent advances in this area have demonstrated the capability of logic operation as well as neural network processing [Chen17&18]. Figure 7 comparatively shows the conceptual views of AI chips based on von Neumann and in-memory computing architecture. The power and latency of the one leveraging the in memory computing macro can be significantly improved.

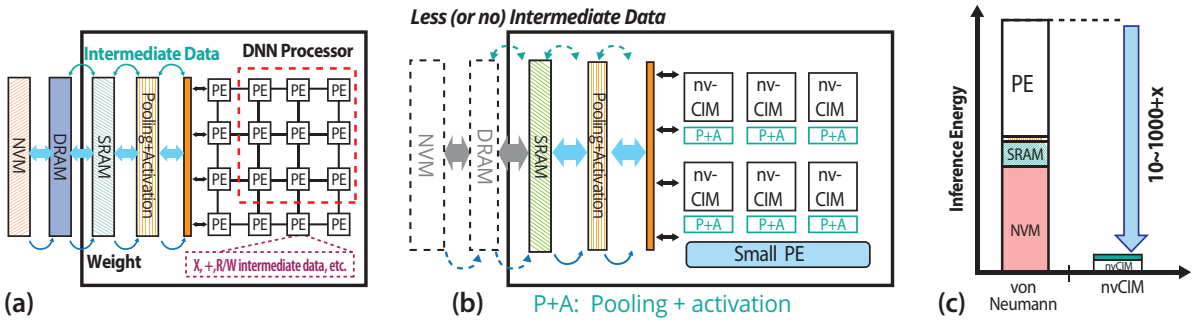



**Figure 7-2** AI Chips Based on (a) Von Neumann, (b) In-memory-computing, and (c) Comparison on Their Power Consumption [Chen18]

## 7.3 Artificial neural networks based on emerging non-volatile memory devices

Computing with artificial neural networks based on emerging non-volatile memory devices has attracted significant attention recently [Yang13]. This group of devices includes novel memories such as ferroelectric memory (FeRAM), magnetic tunneling junction memory (MRAM), phase change memory (PCM), and resistive switching memory (RRAM). These devices can be used to build extremely low standby power memory arrays. More importantly, they are candidates for analog in-memory-computing where memory elements not only store data but also actively participate in signal processing. These devices are generally assembled in crossbar formations with input/output signals across the row/column electrodes. An example RRAM crossbar is shown in Figure 7-3. With the weight matrix being represented as conductance, the crossbar array intrinsically implements vector-matrix multiplication, which is of great significance for a variety of AI based applications. Grey-scale face classification is experimentally demonstrated using the integrated 1024-cell array in Figure 8 with parallel online training. The energy consumption within the analogue synapses for each iteration is 1,000x (20x) lower compared to an implementation using Intel Xeon Phi processor with off-chip memory (with hypothetical on-chip digital resistive random access memory). The accuracy on test sets is close to the result using a central processing unit. [Yao17] Accurate Analogue signal and image processing has also been experimentally demonstrated with 5-8 bits output precision using 128 x 64 ReRAM crossbar array [Li18]. Compared to the CMOS approach, analog in-memory-computing offers high-throughput with parallel signal processing and low energy operation [Li18]. As the state of memory elements could be mapped to synapses, the crossbar array provides a physical embodiment of a fully connected hardware neural networks [Prezioso15] [Hu18].

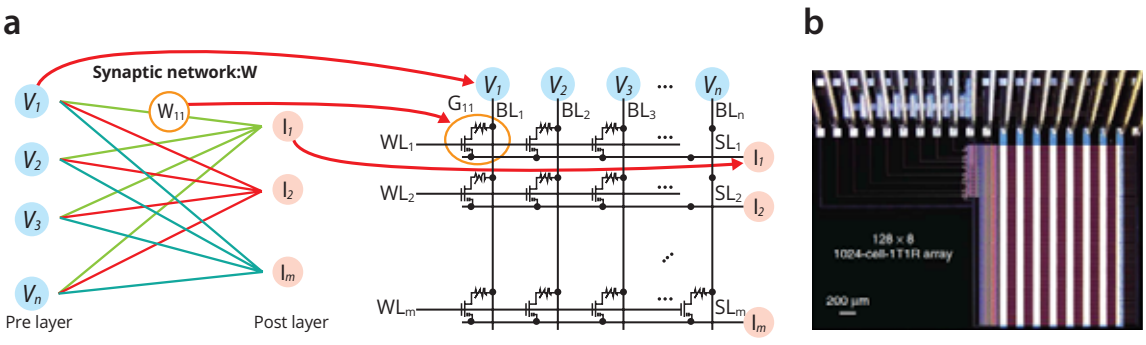



**Figure 7-3** Example of an ReRAM crossbar array used for face classification. (a) Mapping of a one-layer neural network on the 1T1R array, 'T' represents transistor, 'R' represents ReRAM. (b) The micrograph of a fabricated 1024-cell-1T1R array using fully CMOS compatible fabrication process. [Yao17]

## 7.4 Bio-inspired neural networks

The nature of the above arterial neural networks is essentially parallel in-memory computing. A more bio-inspired approach is to more faithfully mimic how the brain process information by using, for instance, spiking neural networks. Hardware implementations of bio-mimic spiking neural networks using CMOS devices have been demonstrated by IBM TrueNorth and the recently announced Intel Loihi. The former consists of 106 spiking neurons with 2.56×108 SRAM synapses [Merolla14] while the latter possesses 1.3×105 neurons and 1.3×108 synapses [Intel17]. The spiking neural network approach requires artificial synapses and neurons share similar dynamics with that of biological counterparts. However, CMOS devices were not created for this purpose and it takes multiple transistors to simulate a synapse or neuron. Therefore, new physics based compact devices with intrinsic similarity to biological synapses and neurons are needed to replicate the biological neural network behaviors in novel computing paradigms. Artificial synapse with diffusive dynamics that is critical for synaptic functions has been realized with a simple two terminal memristor [Wang17]. More recently, artificial neuron with leaky integration and fire function has also been achieved with a single memristor [Wang18]. The 1st integrated neural network with artificial synapses and neurons based on memristors is shown in Figure 7-4. Pattern classification with unsupervised learning has been experimentally demonstrated with this fully memristive neural network [Wang18]. Although neuromorphic computing is still in its early stage of technological maturity, it represents a promising direction for AI chips in the long-term.

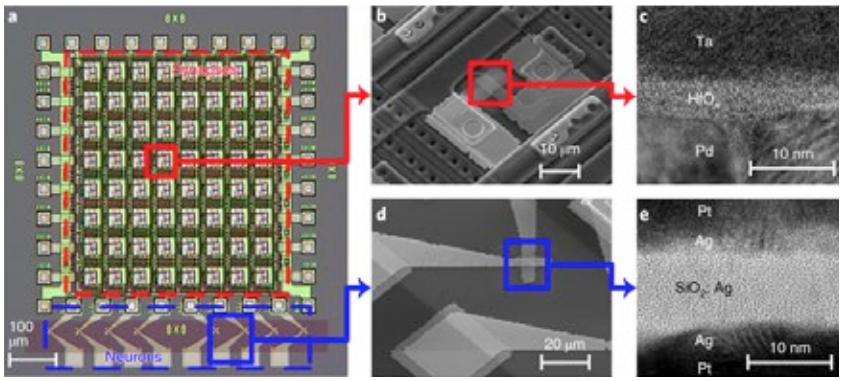



**Figure 7-4** The 1st fully integrated memristive neural network for pattern classification with unsupervised learning. a, Optical micrograph of the integrated memristive neural network, consisting of an 8×8 memristive synapse crossbar interfacing with 8 diffusive memristor artificial neurons. b, Scanning electron micrographs of a single artificial synapse. c, Cross-sectional transmission electron microscopy image of a synapse. d, Scanning electron micrograph of a single artificial neuron. e, High-resolution transmission electron micrograph of the cross-section of a neuron. [Wang18].

In addition to the two terminal devices, emerging three terminal transistors can also be used to build neural network for computing. For example, integrating a ferroelectric capacitor on to the gate of a transistor leads to the so called FeFET. The extent of polarization of the ferroelectric capacitor is correlated to the transconductance of the channel. FeFETs could offer fast programming speed, low power consumption, and smooth symmetric analog programming. With the aforementioned crossbar array structure, the FeFETs could naturally implement the vector-matrix multiplication. (See Figure 7-5). The analog states of FeFETs in crossbar array are able to represent the weights of synapses in a fully connected neural network [Jerry17]. In addition, the lattice polarization dynamics of the ferrorelectric layer also leads to temporal learning rules of SNN such as spiking-timing-dependent plasticity (STDP) [Boyn17].

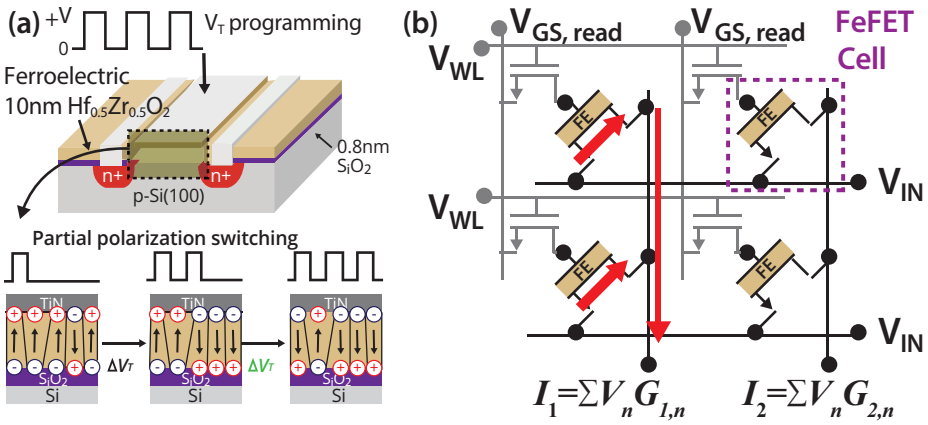



**Figure 7-5** (a) Schematic of a FeFET Showing Analog Programming Capability. (b) Structure of a FeFET Pseudo-crossbar Array for Analog In-memory-computing. [Jerry17].

pseudo-crossbar array for analog in-memory-computing. [Jerry17]

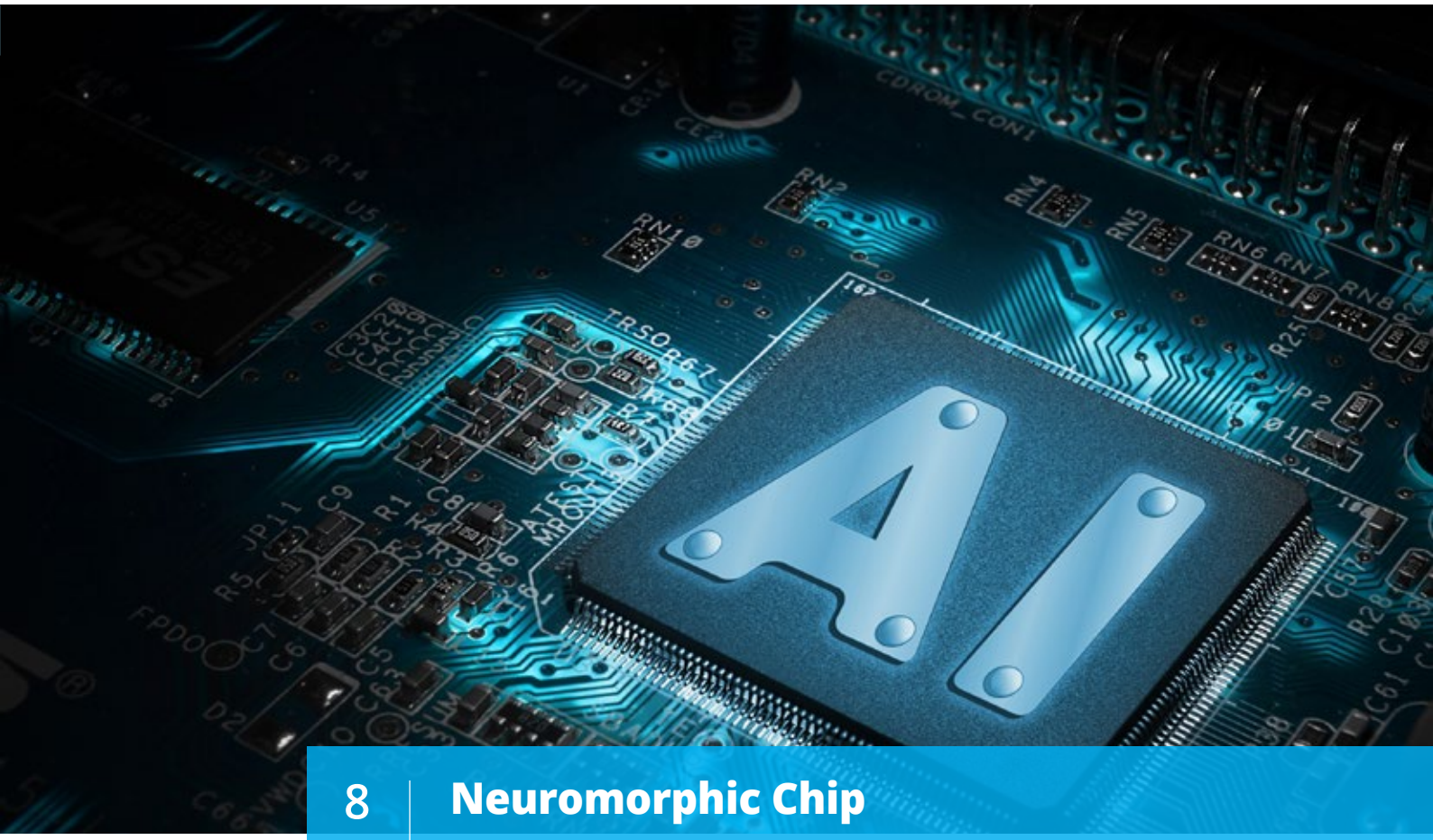
Emerging computing can be implemented using the current CMOS devices, but emerging memory technology is one of key enablers for the thriving of novel computing and AI technologies. While AI-friendly memory is urgently expected in near-term to mitigate the von Neumann bottleneck, near- and in-memory computing, and memristor-based neuromorphic computing are of importance for beyond von Neumann computing in the mid- and long-term to continually sustain the remarkable progress of AI technology.

## 7.5 Impact on Circuit Design

Analog in-memory-computing could potentially be faster and more energy efficient than digital multiply and accumulate unit. However, the analog circuit operation also comes with new challenges to the design of peripheral circuits.

Unlike the digital approach, analog weight representation requires high precision programming of memory elements because the error of each matrix element accumulates in the summation process and impacts the output. In addition, the analog programing process may be relatively stochastic for certain emerging memories. Therefore, achieving high precision analog state programming may need multiple cycles, which could be time consuming and energy inefficient for applications that need frequent reprogramming. Optimization of the programming circuit and algorithms is thus crucial to the computation efficiency and power advantage for those applications.

Furthermore, the analog in-memory-computing scheme also impacts signal conversion circuits. To interface with conventional digital signal inputs and pass results back to digital systems, fast and energy efficient signal conversion circuits (both digital-to-analog and analog-to-digital) are needed. For vector-matrix multiplication based on Ohm's law and Kirchhoff's law, the input generally takes the form of voltage signals while the output is received as current signals. Precise measurement of current over a wide range may need to be addressed.




# 8 Neuromorphic Chip

Neuromorphic chip uses electronic technology to simulate the revealed rules of operation of a biological brain, thereby constructing an electronic chip similar to a biological brain, namely a "bionic computer." It is similar to the meaning of neuromorphic engineering. Neuromorphic Engineering was proposed by Carver Mead, a professor at Caltech in the late 1980s, to use Very-large-scale integration (VLSI) with analog circuits to simulate the structure of a biological nervous system. Neuromorphic calculations have also been used in recent years referring the models of neural systems implemented using analog, digital, digital-analog hybrid VLSI, and software systems. Inspired by the results of brain structure research, the developed neuromorphic chip features low power consumption, low latency, high-speed processing, and joint space-time representation.

## 8.1 Algorithm Model of Neuromorphic Chip

In a broad sense, the algorithm models for neuromorphic computation can be broadly divided into Artificial Neural Networks (ANN), Spiking Neural Networks (SNN), and other extended models with specific data processing capabilities. Among them, ANN is the main model used in machine learning, especially deep learning, and it is the main content discussed in other parts of this article. Therefore, in this section we mainly discusses the SNN algorithm. This type of algorithm has several characteristics. First, the output of its neurons is a time-space-encoded pulses. Second, the timing domain information is expressed through the membrane potential value, ie, the membrane potential records the historically received and issued pulse energy. Therefore, multiple neurons can realize the expression ability of space-time two-dimensional space.

There are many simulation algorithms for dynamic behavior of neurons, which are generally expressed by differential dynamic equations, and have good bionic capabilities, but are not conducive to hardware implementation. Therefore, a method based on a simplified algorithm, such as Leaky Integrate and Fire (LIF) model, has received much attention. The principle is that the pulses connected to all axons of this neuron are weighted and summed according to the synaptic strength to obtain the integrated potential of the neuron, which is then added and updated with the previous membrane potential to obtain a new membrane potential. The pulse is issued if the membrane potential exceeds the set threshold, otherwise it is not issued. As shown in Figure 8-1. It can be seen that the LIF algorithm has the feature of expressing the temporal and spatial information jointly.

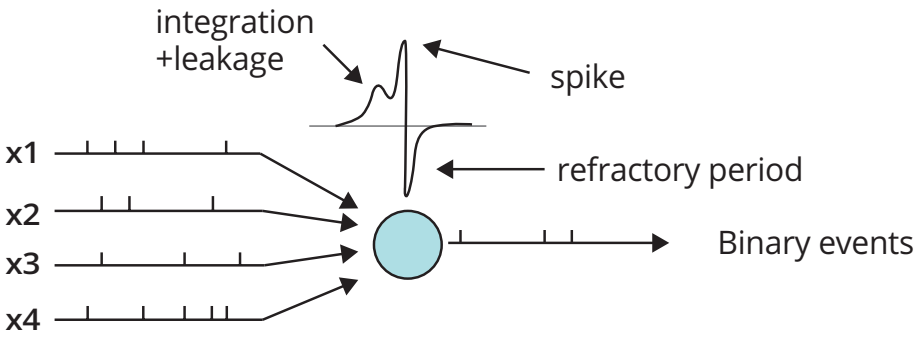



**Figure 8-1** A simple Demonstration of Leaky Integrate-and-Fire Algorithm

A commonly used SNN learning algorithm is spike-timing-dependent plasticity (STDP). STDP is a more effective training algorithm that has been validated in the biological brain. As a locally trained, non-back-propagation algorithm, it does not guarantee obtaining high-performance networks. In addition, because the training of the SNN network is difficult, in practical applications, back propagation algorithms are also used for training. SNN can work with low accuracy of weights, whereas a large number of multiplication and accumulation operations, hence ANN networks with low precision weights and low precision activation values are more suitable for porting to SNNs. A rougher comparison is that for a feedforward network topology such as MLP or CNN, the LIF model-based SNN network is similar to the ANN network which using binary activation neurons and ReLU as activation function. The membrain potentials in SNN have similar timing domain representation capability with the recurrent connections in ANN. Hence SNN and ANN have a certain degree of equivalence. The comparison between SNN and ANN is shown in Figure 8-2.

| Category | ANN<br>ANN (Artificial Neural Network, Deep Learning) | SNN<br>SNN (Spiking Neural Network Algorithm) |
|---|---|---|
| Neuronal Activations | Multi-level (fixed or floating point) | timing domain coded spikes (binary values) |
| Timing expression | Recurrent connections in RNN and other networks | Membrane potential and Recurrent connections |
| Spatial expression | usually a more regular interconnected neuronal array. The processing of images usually adopts sliding windows in convolution operation | non-regularly interconnected neurons. Generally, there is no sliding window process (requiring parallel expanding of convolutions). |
| Activation function | usually nonlinear activation | No activation function |
| Reasoning | Convolution, pooling, multilayer perceptron model (MLP), etc. | Leaky Integrate and Fire model (LIF), etc. |
| Training | Back-propagation is more popular | STDP, Hebb's law, back-propagation |
| Normalization method | Batch normalization, etc. | Winner takes all |
| Represent negative neuronal values | Negative activation value | inhibitory neurons |
| Typical Sensor | Digital Camera, Microphone | DVS Camera |
| Theoretical sources | Mathematical derivation | Brain enlightenment |
| Common point | integration process, MLP topology | |

**Figure 8-2** Comparison between SNN and ANN

## 8.2 Neuromorphic chip characteristics

### 8.2.1 Scalable, highly parallel neural network interconnection

Inspired by the interconnected structure of the biological brain, the neuromorphic chip can realize the interconnection between arbitrary neurons. That is, under a given-scale biomimetic neural network, any neuron can transmit information to any other neurons. Such a powerful fine-grained interconnect capability is currently unavailable in other neural network/deep learning chips.

In order to realize the complex interconnection, the neuromorphic chip is designed in a hiarachical way. It includes array cores which have crossbars, a network on chip (NoC), and a high-interconnection link outside the chip. The simplest scheme is the crossbar, as shown in Figure 8-1. There are connection points on the crossbar representing the synaptic connections, and the strength of the connections is represented by multi-valued weights. However, its scale is limited. To expand the scale, high-speed, shared physical links can be used, such as 2D Mesh networks inside chips and high speed (SerDes interfaces/fiber) interfaces that interconnect chips. These interconnection schemes are shared by multiple neurons in a time-sharing manner, which unlike biological brains where the transmission links are independent of each other. Therefore, the data to be transmitted needs to carry the target address information, and packets are transmitted on the shared link. Neuron-level transmissions have the arbitrariness of addresses, so each neuron destination address is different. The transmitted data packets use neuron-level packets, typically several tens of bits, and are therefore suitable for storage-forwarding methods, such as 2D-mesh network and high fanin/fanout bus. The neurons and axons in such a transmission network may not have absolute addresses. By introducing relative address, the network can be in arbitrarily large size. If the path is far away, it can be relayed using a neuron. It is noteworthy that ordinary computers using memory for data exchange can also achieve the same effect, but it is for serial and low speed exchanging. On the contrary this on-chip network can support hundreds of thousands of parallel transmission.

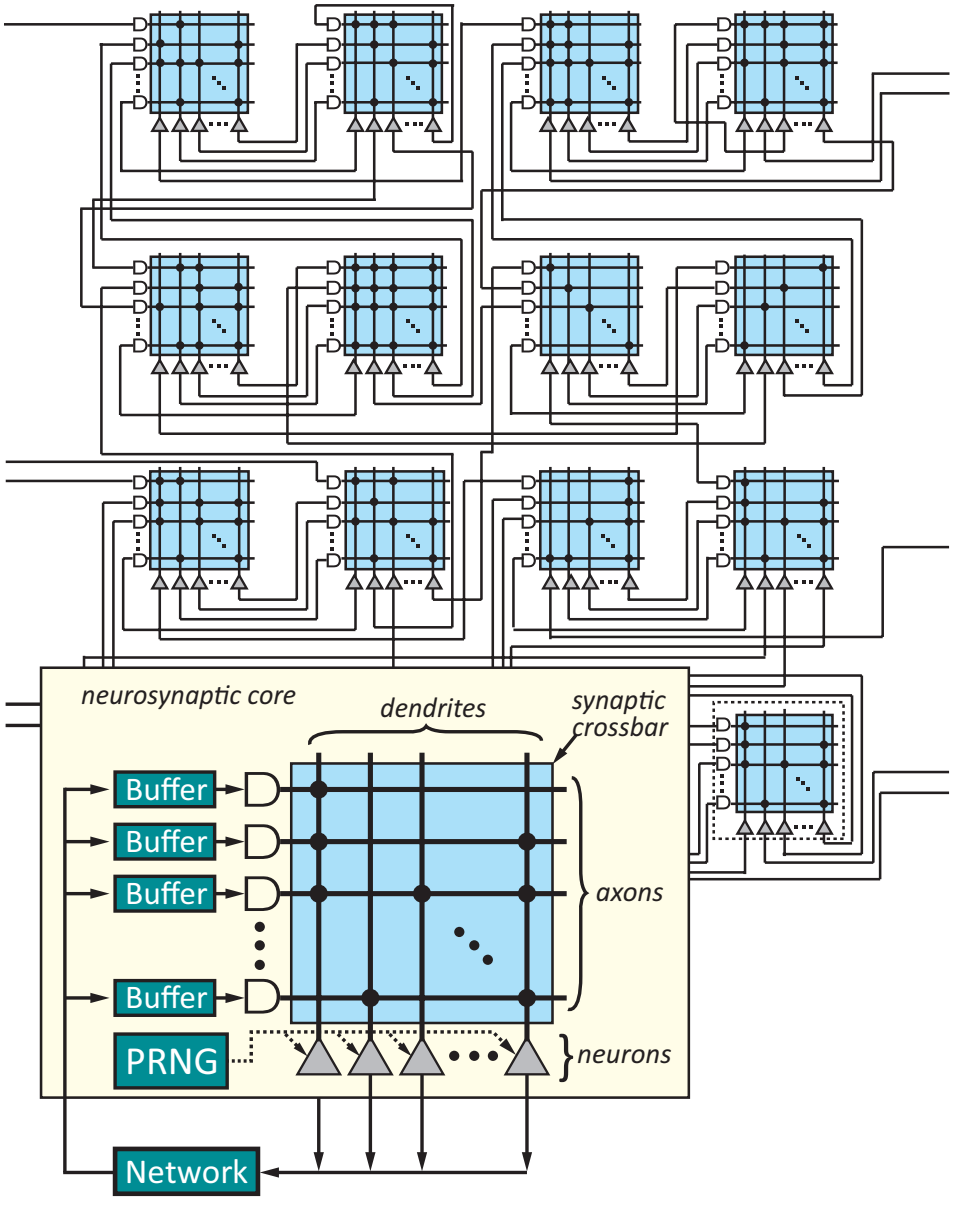



**Figure 8-3** Architecture of Neuromorphic Chip

### 8.2.2 Many-core architecture

Since the neurons in the biological brain are clustered, there are more intra-cluster interconnections and fewer connects among clusters, so it is feasible to use a clustering structure in the neuromorphic chip, that is, a group of neurons are included in the cluster, and interconnected with the incoming signals (output axons of other neurons) using the synaptic cross-bar, and inter-cluster networks such as 2D Mesh are used. The cluster becomes the functional base unit of the chip, sometimes called Core, and this structure is also known as many-core. This unit mainly includes input and output routes, synaptic cross-bar (actually implemented by memory), integral operations, and fire operations. The input and output data can be considered as the storage area that

interacts with the outside world. Synaptic storage is private within the cluster and cannot be accessed by other clusters. These clusters are equal in status. There is no master-slave relationship. The entire system is decentralized and scalable. Overall, synapse storage may occupy half or more of the memory, and is scattered across the clusters, and close to the computing unit. Therefore, this structure has near-memory or near-memory calculation characteristics. The theory stems from the computational-storage integration of biological brains. The advantage is that it can solve the“storage wall”in traditional von Neumann computers.

### 8.2.3 Event-driven

The event-driven in neuromorphic chip is also inspired by the brain: when biological neurons receive the pulse, the corresponding membrane potential changes. Without an input pulse, there is no need for integration and a change in membrane potential (excluding there is a potential leak). So whether the circuit works depends on whether there is a packet (event) trigger. Digital circuits are often post-selected, like arithmetic and logic units (ALUs) in the processor. All possible branches are first evaluated and the desired branch is selected. Event-driven is a predecessor, a module has many inputs, when the input receives information (such as pulse packets), the module starts the calculation. Therefore, whether or not the route transmits data and whether the neuron module operates depends on whether there is an event exist. This greatly reduces the dynamic power consumption of large-scale chips.

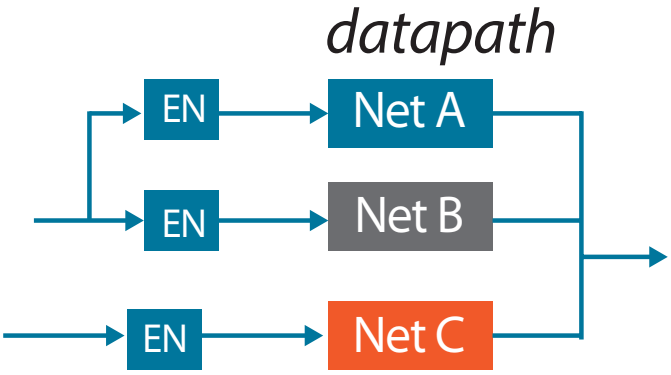



**Figure 8-4** Event Driven Architecture

### 8.2.4 Dataflow processing

There are many core viewpoints of data flow. The main point of reference in this paper is that the operation and transmission are represented by a directed graph. In this directed graph, nodes represent operations and edges represent transmission relations. The data flow diagram is like having a variety of branches, incorporating a complex process pipeline, allowing data to be output through this system. Whether or not each node operates depends on whether the result of its previous node's operation is already in place, and whether the input line of the succeeding node is capable of accepting new data. If it is satisfied, the operation is performed. It is a faster, non-centric, dynamic, many-core computing approach with the friendliness of many-core computing. The insufficiency of data flow calculation is that the control logic is simple and it is not easy to express a structure such as cyclic recursion. However, in the category of neural network it seems to be appropriate. Therefore, in Tensorflow and other neural network frameworks, the data flow is used to express the connection relationship of neural networks by default.

## 8.3 Opportunities and challenges

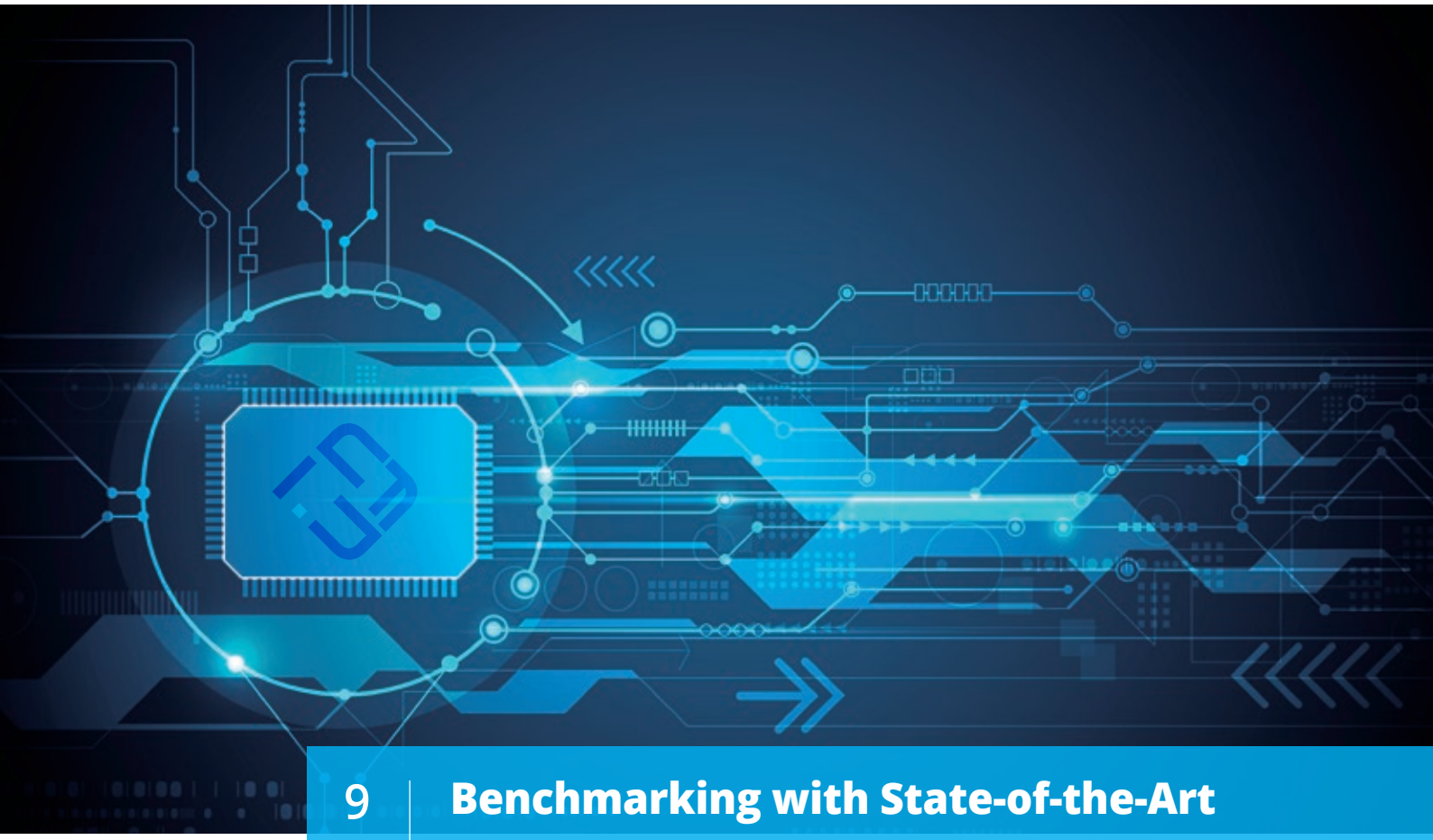
At present, the methods for designing neuromorphic microarrays are mainly divided into neuromorphic computational circuits based on traditional CMOS technology and neuromorphic computational circuits based on novel nanodevices. The development of traditional CMOS technology is relatively mature, as mentioned in Chapter 7. IBM TrueNorth chip is a representative of asynchronous-synchronous hybrid (no global clock) digital circuit, Tsinghua University's Tianjic is a pure synchronous digital circuit SNN-ANN mixed-mode chip; the ROLLS chip of the Swiss Federal Institute of Technology in Zurich and BrainScaleS of the University of Heidelberg are representative works of analog integrated circuits. The most important direction for neuromorphic computation circuits based on new nanodevices is the use of neuromorphic devices such as memristors (see Chapter 7 for details).

Neuromorphic chips have excellent applications in smart cities, real-time information processing for automatic driving, and face depth recognition. Such as IBM TrueNorth can be used to detect pedestrians, vehicles and other objects in the image, with the extreamly low power consumption (65mW). It can also be used for tasks such as voice and image identification, and it is not inferior to the CNN accelerator chip. In addition, online learning ability is also a highlight of the neuromorphic chip. Researchers have shown that Intel Loihi chips increase learning speed by 1 million times compared with other typical SNN networks in solving MNIST digital identification problems [Davies18].

In the traditional CMOS process, the physical structure of the neuromorphic chip is relatively mature, but there are still many challenges for the simulation of large-scale neural networks (such as systems larger than 1% of the human brain), including: (1) Heat dissipation. This will result in the inability of the single chip scale to increase further. The density of on-chip memory and integral computing units are still insufficient, resulting in an inability to increase the number of integrated synapses and neurons under the constraint of the heat dissipation. (2) Due to its array of many-core characteristics, the problems of interconnecting and synchronizing are prominent on multi-scales including on-chip, across-chip, cross-board, and multi-machine; (3) In order to increase the density, most of the ASIC chips can only simulate a single neuromorphic algorithm, and lack of flexibility and the ability to imitate real biological neurons.

For a neuromorphic chip based on a memristor cross array, in the foreseeable future, it is necessary to constantly optimize the parameters of memristor array to address the specific application requirements and energy efficiency goals. The optimiztions include increasing the size of the crossbar, improving synaptic connection, controlling the leakage current.

In addition to the above-mentioned physical structure problems, neuromorphic chips also face enormous challenges in terms of software, especially algorithms. A few years ago, when deep learning and CNN technology were not effectively developed, ANNs also faced similar problems. However, with the continuous advancement of deep convolutional neural networks based on back propagation algorithms, this problem has been greatly alleviated. At present, most of the SNN-related algorithms are in the research stage. The main reason is that people have not fully explored the operating mechanism of biological brains, and the enlightenment is limited. What is gratifying is that in recent years, brain mapping technology has developed rapidly, and it has become a reality to draw detailed animal-brain neuron-level static connection structures. The dynamic detection of brain mechanisms, such as the interpretation of low-level visual cortex in animals, has also made significant progress. It is believed that the experimental achievements from these brain researches will greatly assist in the breakthrough of neuromorphic computing.



# 9 Benchmarking with State-of-the-Art and Roadmap

With so many research teams in academia and industry working on developing chips for AI applications, we will undoubtedly see an ever-increasing number of AI chips being introduced. Two overarching efforts are indispensable in this AI chip development frenzy: (i) objectively evaluating and comparing different chips (i.e., benchmarking), and (ii) reliably projecting the grow paths of AI chips (i.e., roadmapping).

Benchmarking aims to provide a uniform methodology to evaluate and compare different chips for AI applications. A range of architectures and technologies (e.g., the examples discussed in the previous sections) has been adopted to implement various forms of neuromorphic computing and machine learning accelerations. CMOS technologies, circuits and architectures can be optimized in customized chips or specialized accelerators to execute mathematical calculation and approximation of neural networks (e.g., fast matrix operation), which can significantly surpass general-purpose CPUs in energy and computational efficiency. However, such systems may still be orders of magnitude poorer than biological brains in terms of efficiency and capabilities. On the other hand, direct implementations of neural networks may require materials and devices that possess native behaviors mimicking the

function of neurons and synapses. Numerous devices have been explored for these functions with different levels of success. For example, synaptic behaviors have been demonstrated in metal-insulator-transition devices, phase change memories, oxide resistive switches based on filament formation or oxygen migration, spin-transfer-torque devices, ferroelectric tunnel junctions, etc. It is important to clearly define a set of performance requirements and quantitative parameters, in order to benchmark these material, device, and architecture options to guide research directions.

Progress will typically start from smaller systems with modest functionalities and scale up to larger systems capable of solving complex problems. A well-defined roadmap based on commonalities in technology, design, or application will not only provide a measure of progress but also help to identify research gaps and critical challenges. Unlike CMOS benchmarking and roadmap where technology option (i.e., field-effect-transistor) and commonality (i.e., transistor feature size) are clearly defined and agreed upon, the large variety of applications, algorithms, architectures, circuits, devices for AI presents a challenge to identify the common basis for benchmarking and roadmapping.

Based on existing work in the general chip design area, it has been widely accepted that it is unlikely to find the universally "best" device, architecture, or algorithm. For example, CMOS devices seem to be hard to beat by emerging devices for von Neumann architectures [Nikonov15, Nikonov13], while some emerging devices, e.g., tunnel FETs and spintronic devices, may perform better in non-Boolean architectures, e.g., cellular neural network [Pan16]. What makes benchmarking for AI chips even more challenging is that, besides the need to include energy, performance, and accuracy associated with the computation itself, one must also consider performance and energy overheads due to other operations such as input, output, memory accesses etc. This is particularly difficult for non-von Neumann hardware since a true, "apples-to-apples" comparison must identify both the state-of-the-art von Neumann platform (CPUs, GPUs) and the associated algorithm. Moreover, the performance of the aforementioned platforms will change as technology scales and new technology innovations are introduced into future generations of chips. Last but not least, new algorithms (such as neural network structures and computing models) are being actively investigated and introduced by theoreticians and application experts. Benchmarking and roadmapping efforts must take all these factors into consideration.

We are not aware of any published comprehensive benchmarking work related to neuromorphic computing chips. The evaluation of AI chips in the industry mainly depends on running some common neural networks, such as Deepbench proposed by Baidu[Baidu]. Several research programs being funded by U.S. NSF, DARPA and SRC have recognized its importance. For example, researchers in the EXCEL Center (funded by U.S. NSF and SRC) are actively investigating benchmarking methodologies for non von Neumann hardwares, e.g., for tasks targeting at the MNIST dataset [EXCEL]. A benchmarking workshop will be organized in 2018 by X. Sharon Hu (a lead EXCEL researcher) in collaboration with colleagues from Tsinghua University, Beihang University, Hong Kong University of Science and Technology, etc.

**To overcome the challenges associated with hardware benchmarking for AI applications, we need to address the following issues.**

- Collect a set of architectural-level functional units
- Identify both quantitative and qualitative Figures of Merits (FoMs)
- Develop uniform methodologies for measuring FoMs

**Materials and devices for neuromorphic computing need to demonstrate the following properties:**

- Multistate behaviors: a physical property with different values depending on past history
- Low energy: switching from one state to another with low dissipation
- Nonvolatility: properties maintained without refreshing
- Threshold behavior: drastic change of a property after repetitive stimulus
- Fault tolerance

Whether a chip built with some specific device technology, circuit style and architecture are superior or not depends strongly on the specific applications and algorithms/models. To fairly benchmark the large variety of devices, it's necessary to identify the applications, the suitable algorithms and models, and the circuit designs that should be specific enough to define the device requirements without limiting the device options. The benchmarking of beyond-CMOS devices conducted by Nanoelectronics Research Initiative (NRI) uses inverter, NAND gate, and adder as standard Boolean logic circuit blocks to compare all the devices [Nikonov15, Nikonov13]. The benchmarking of neuromorphic chips would also need to define common functional units with quantifiable parameters. Many neural networks use convolution, rectilinear, and pooling functions, which can be suitable functional units. As new algorithms and computation models are developed, additional function units may be introduced. At the architectural level, operations/second/watt and throughput can be two complementary measurement of the system performance. How to capture other neuromorphic computing models such as spiking neural networks should also be investigated as they often carry computational forms different from other scalar neural networks.

Some quantitative parameters for neuromorphic devices have been reported and evaluated, including modulation precision (e.g., resistance levels) and range (e.g., on/off ratio), linearity, asymmetry, variability, etc. They have all been shown to be critical for neural network performance. Some device parameters commonly used for Boolean logic are also important for neuromorphic computing, including size, speed, operation voltage/current, power/energy, endurance, retention, yield, etc. Tradeoffs among these parameters need to be carefully evaluated. Accuracy is a key consideration for AI applications, and should also be included as a FoM. Different accuracy metrics should be considered in order to determine the most appropriate ones to use.

Developing a uniform benchmarking methodology for AI chips may leverage the knowledge and experience gained from the benchmarking effort sponsored by NRI. Though the work mainly focuses on the Boolean logic units (NAND gate and ALU), it lays the foundation for more advanced architectural level benchmarks. For example, in [Perricone17], the authors extend the work in [Nikonov15] to benchmark several emerging devices at the architectural level for multi-core processors executing parallel workloads. The proposed analytical framework projects performance and energy of multi-core processors based on emerging devices. Extending this framework to handle non-von Neumann architectures will require new performance and energy models based on the appropriate functional units. Benchmarking and roadmapping of neuromorphic computing have to move beyond the device and circuit levels (i.e., adders, multiply-accumulate units) to quantify how collective dynamics improves the energy/performance FoM of a computational primitive (e.g., convex optimization) and application-level tasks (e.g., image analysis). The work can best be accomplished through joint effort of algorithm researchers, architects, circuit designers and device experts.

# 10 Looking Ahead

At this moment, AI chip is still in its infancy stage and abundant uncertainties linger. The only sure thing is that it is the fundamental of development of AI technology and great impetus for semiconductor industry. Just as discussed in this white paper, in the global scope, giant companies and start-ups in this field are investing heavily to innovate AI chips at multiple hierarchies from architecture, circuits, device, to manufacture and materials. All the endeavors will probably lift the semiconductor techniques to an unprecedented level and benefit the overall advancement of science and technology. And we must realize the challenges ahead and make greater efforts to achieve the goal.

Firstly, the AI is at a preliminary stage and there are considerable uncertainties facing the AI chip industry. Nowadays, research surrounding AI chips has made significant progress in the area of machine learning based on neural network which is considered to be superior than human intelligence in solving some computing-intensive issues. However, it is still at an embryo stage when it comes to solving cognitive problems and has a long way to go before achieving general-purpose intelligence (Artificial General Intelligence, AGI). The ideal computing capability and energy efficiency of AGI should be at least several orders of magnitude higher than that of AI Chips today. The trickier fact is that; we do not even know whether the mainstream technology we pursuit today is the correct path to AGI. No one could give a definite answer to the question: will there be an ultimate algorithm for AGI? Chips are

intertwined with algorithm, if there would be a uniform algorithm emerging, a corresponding type of AI chips could be shaped too. For a long term, different applications still need different supporting algorithms including traditional algorithms. With development of underlying chip technology, AI algorithms will get better chance to flourish. In addition, some novel devices and materials would not only make new algorithms and models (like brain-inspired computing) more mature and efficient, but also facilitate us to explore AGI. During this process, AI self would be probably rendered to guide developing new chip technology, resulting in sustainable and mutual promoting scenario. In another words, the uncertainties of AI lend a huge performance stage for various technology innovations, until converged, we could anticipate brilliant shows on the stage.

Secondly, as the IoT industry continue to grow rapidly, many more diversified needs and application scenarios will emerge. Driven by user needs, innovation in AI chip technologies will facilitate a more closely connected R&D process and commercializing process and eventually forms an open, mutually beneficial industry ecosystem. Brought by the integration of CMOS technology and emerging information technologies and the emergence of open source software and hardware, we can anticipate an unprecedented era where innovations are achieved synergistically.

In conclusion, as the destination undefined, the exploration of AI chip is definitely strenuous and exciting, and requires efforts both from policy makers and other players involved in this race. With policy, capital, market and technology together forming an environment friendly to entrepreneurs, we have confidence in the future of AI chip and could envisage the bonus that brought by the development.

## References


**[Vivienne17]** Sze, Vivienne., et al. "Efficient processing of deep neural networks: A tutorial and survey." Proceedings of the IEEE 2017,105(12): 2295-2329. DOI: 10.1109/JPROC.2017.2761740

**[Theis16]** T.N. Theis., et al. "The End of Moore's Law: A New Beginning for Information Technology," Invited paper, IEEE Computing in Science & Engineering, 2017,19 (2): 41-50. DOI: 10.1109/MCSE.2017.29

**[Aly15]** M. M. Sabry Aly., et al. "Energy-Efficient Abundant-Data Computing: The N3XT 1,000X," IEEE Computer, 2015, 48(12): 24 – 33. DOI: 10.1109/MC.2015.376

**[Google]** 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

**[Bankman18]** 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

**[Lee18]** 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.2018.8310262

**[Parashar17]** A. Parashar., et al. "SCNN: An accelerator for compressed-sparse convolutional neural networks." 2017 ACM/IEEE ISCA, DOI: 10.1145/3079856.3080254

**[Bert17]** M, Bert., et al. "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltageaccuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI." Solid-State Circuits Conference (ISSCC), 2017 IEEE International. IEEE, 2017, 246 – 247. DOI: 10.1109/ISSCC.2017.7870353

**[Shouyi17]** S. Y. Yin., et al. "A 1.06-to-5.09 TOPS/W reconfigurable hybrid-neural-network processor for deep learning applications." VLSI Circuits, 2017 Symposium on. IEEE, 2017, C26 - C27. DOI: 10.23919/VLSIC.2017.8008534

**[Shouyi18]** S. Y. Yin., et al. "A High Energy Efficient Reconfigurable Hybrid Neural Network Processor for Deep Learning Applications." IEEE Journal of Solid-State Circuits 2018 53(4): 968-982. DOI: 10.1109/JSSC.2017.2778281

**[ISSCCTrends2017]** http://isscc.org/wp-content/uploads/2017/05/ISSCC2017_TechTrends.pdf

**[Lee14]** D. U. Lee., et al., "A 1.2V 8Gb 8-Channel 128GB/s High-Bandwidth Memory (HBM) stacked with effective Microbump I/O Test Methods Using 29nm Process and TSV", ISSSCC, 2014, 432–433. DOI: 10.1109/ISSCC.2014.6757501

**[Jeddeloh12]** J. Jeddeloh., et al., "Hybrid Memory Cube: New DRAM Architecture Increases Density and Performance", VLSI, 2012, 87 – 88. DOI: 10.1109/VLSIT.2012.6242474

**[Nvidia]** http://www.nvidia.com/object/what-is-gpu-computing.html

**[Chi16]** P. Chi., et al., "PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory," in Proc. ISCA, 2016, 27-39. DOI: 10.1109/ISCA.2016.13

**[Chen17]** Y. H. Chen., et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE Journal of Solid-State Circuits 2017,52 (1):127-138. DOI: 10.1109/JSSC.2016.2616357

**[Chen18]** W. -H. Chen., et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processor", ISSCC, 2018, 494-496 DOI: 10.1109/ISSCC.2018.8310400:

**[Yang13]** J. J. Yang., et al., "Memristive devices for computing", Nature Nanotechnology, 2013,8, 13-24. DOI: https://doi.org/10.1038/nnano.2012.240

**[Yao17]** Yao, Peng., et al. "Face classification using electronic synapses." Nature communications 2017,8 15199.

**[Li18]** C. Li., et al. "Analogue signal and image processing with large memristor crossbars." Nature Electronics 2018,1: 52-59. DOI: 10.1038/s41928-017-0002-z

**[Prezioso15]** M. F. Prezioso., et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors." Nature 2015,521 (7550):61-4. DOI: 10.1038/nature14441

**[Hu18]** M. Hu, et al., "Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine." Advanced Materirals 2018, 30(9): 1705914. DOI: 10.1002/adma.201705914. (2018)

**[Merolla14]** P. A. Merolla., et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface." Science 2014,345 (6197):668-673. DOI: 10.1126/science.1254642

**[Intel17]** https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-acceleratеartificial-intelligence/

**[Wang17]** Z. Wang., et al., "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing." Nature Materials 2017,16 (1):101-108. DOI: 10.1038/nmat4756.

**[Wang18]** Z. Wang., et al, "Fully memristive neural networks for pattern classification with unsupervised learning", Nature Electronics 2018,1, 137–145. DOI: https://doi.org/10.1038/s41928-018-0023-2

**[Jerry17]** M. Jerry., et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training." IEEE International Electron Devices Meeting (IEDM), San Francisco, USA. (2017) DOI: 10.1109/IEDM.2017.8268338

**[Boyn17]** Boyn, S., et al. "Learning through ferroelectric domain dynamics in solid-state synapses." Nature Communications 2017,8:14736. DOI: 10.1038/ncomms14736 (2017).

**[Davies18]** M. Davies., et al., "Loihi: a Neuromorphic Manycore Processor with On-Chip Learnin", IEEE Micro, 2018, 38(1): 82 – 99. DOI: 10.1109/MM.2018.112130359

**[Nikonov15]** D. Nikonov., et al. “Benchmarking of beyond-cmos exploratory devices for logic integrated circuits,” Exploratory Solid-State Computational Devices and Circuits, IEEE Journal on, 2015, 1, 3–11. DOI: 10.1109/JXCDC.2015.2418033

**[Nikonov13]** D. Nikonov., et al. “Overview of beyond-cmos devices and a uniform methodology for their benchmarking,” Proceedings of the IEEE, 2013, 101(12): 2498–2533. DOI: 10.1109/JPROC.2013.2252317

**[Pan16]** C. Pan., et al. "Non-Boolean Computing Benchmarking for Beyond-CMOS Devices Based on Cellular Neural Network," in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2016, 2, 36-43. DOI: 10.1109/JXCDC.2016.2633251

**[Baidu]** “DeepBench",https://github.com/baidu-research/DeepBench

**[EXCEL]** https://collectivecomputing.nd.edu/about/

**[Perricone17]** R. Perricone., et al. “Can Beyond CMOS Devices Illuminate Dark Silicon,” Conference on Design , 2016:13-18.

White Paper on AI Chip Technologies

Address: Building A, 3# Heqing Road, Haidian District, Beijing
Tel: 010-62799552
Web: http://www.icfc.tsinghua.edu.cn
Wechat: THU-ICFC