Heterogeneity in IT Infrastructures

Rena Oi

Abstract

To supply the massive computing performance required for artificial intelligence and the Internet of Things, new infrastructure is needed for both general use and specific purposes. Cloud services will rapidly enable such a flexible future infrastructure.

Keywords: graphics processing unit, field-programmable gate array, edge computing

1. Change in the leading processor

The evolution of technology is insatiable, demanding more and more calculation ability and processor power. In the field of artificial intelligence (AI), it is robust processor power that led to the breakthrough and commercialization of deep learning. In self-driving cars, it is processor power that enables real-time identification of objects surrounding the car from four directions to determine the next operation of the steering wheel. It is processor power that, based on the movement of the user, allows real-time generation of high-definition images on a virtual reality display. Finally, it is processor power that makes possible the more in-depth analysis of Internet of Things (IoT)-generated data for marketing purposes.

Processor manufacturers continue to compete in the development to meet the insatiable demand for advanced calculation ability. In the past, the calculation ability of the central processing unit (CPU), a multi-purpose processor, was continuously being improved as the standard. Now processors that excel in parallel execution of many homogeneous processors have become the leader in calculation. Because emerging technologies of recent years required large-scale parallel processing, the adoption of the graphics processing unit (GPU) has increased.

In the AI area, the GPU achieved an efficiency rate ten times higher than that of the CPU, making the GPU a de facto standard. In addition, the emergence of the GPU has coincided with the enhancement of software libraries, which utilize the GPU's characteristic features. As a result, the use of the GPU has been increasing from image processing, which was its original purpose, to advanced scientific computation. GPUs and processors with many cores that compete are used in devices ranked high in TOP500, which ranks the world's supercomputers. These parallel processors are also continuing to evolve, maintaining their leadership in speed for the foreseeable future.

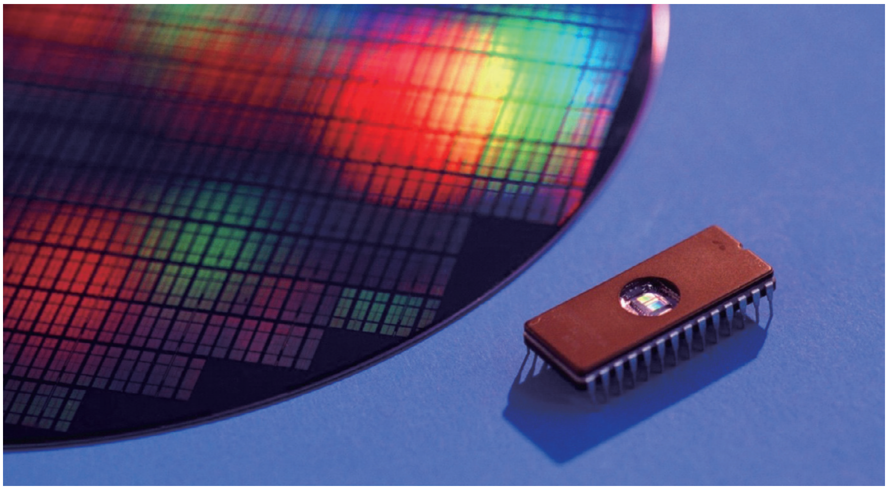
2. Diversification of processors

In some cases, even the power of the GPU is not adequate, so a processor called the field-programmable gate array (FPGA) is used. While the logic on regular processors is printed at the time of manufacture and cannot be changed, FPGA processors can be rewritten later and as many times as necessary. With FPGAs, it is also possible to manufacture only a few units of processors specialized in specific applications. In the financial industry, FPGAs are used to perform high-speed algorithmic trading of stocks and currency exchange, with the logic being changed as needed and tuned at the high-speed level of milliseconds to chase profits. In the competitive area of AI, dedicated logic is implemented on an increasing number of FPGAs in an attempt to achieve speed and efficiency that exceed those of GPUs. In addition, efforts are underway to install several thousand FPGAs at a time to improve the overall processing ability of datacenters.

A trend to use smartphone processors is also attracting attention. The processor of a smartphone uses multiple cores with different abilities to enable the phone to go to sleep or shut down in a short period of time, improving the efficiency of electric power usage. Low electric power usage and robust processing ability at an advanced level are spreading the use of smartphone processors to drones, connected cars, and other IoT devices. In the future, the use of smartphone processors is expected to further streamline the operation of datacenters.

Some companies design, manufacture, and use their own dedicated processors for specific purposes. One such company developed its own processor that specializes in deep learning and that is supporting the fast development of AI today. This specialized processor is currently installed in datacenters to streamline operations. It is a low-power-consuming processor based on the knowledge that even if the computation of floating-point arithmetic is less accurate than normal, it can be used in deep learning. The latest GPUs have also adopted this approach of lowering the computation accuracy of floating-point arithmetic to achieve higher speed and improved efficiency. Deep learning is a particularly competitive area, and it is driving increased competition in the development of specific-purpose processors.

The important characteristic of this new calculation ability is the development of software that utilizes the features of a processor. In particular, parallel processing would not exist without the support of software. The difficulty level is high to develop this type of software, which makes it challenging for the processor to reach its potential. Processor manufacturers are enhancing software parts called libraries to make it easier for software developers to harness the processor power.



3. Diversification of architecture

Like processors, system architectures are also diversifying. Edge computing is an architecture that uses huge numbers of IoT devices in the field, that is, the edge, to make distributed processing more flexible and dynamic. When immediate processing is necessary in the field, or accumulation and compression of collected data are required, the edge processes these tasks while coordinating with the center as needed. Application to automatic operation is expected soon, and development is ongoing along with the use of next-generation 5G (fifth-generation mobile communications) networks.

The blockchain, a distributed ledger technology, is also attracting attention as an infrastructure technology that supports Bitcoin virtual currency. The blockchain enables distributed data to be shared on the network while preventing data falsification. It is innovative in that it does not require the building of a robust and centralized database. Consequently, systematization of complex areas that have traditionally been impossible to do on a cost-effective basis is now possible. In addition, funds transfer and payment uses, financing, contract management, and the government's notarization service are undergoing proof-of-concept trials. Bitcoin has been operating ever since its release, and new implementations of high-quality blockchain technology are expected in the future.

4. Utilization of flexible options

To bring business ideas that constantly use new technologies to fruition, it is essential to master the use of a system infrastructure that includes diversified and complicated processors and architectures. However, this kind of system infrastructure does not need to be owned. Clouds are taking the place of diversified infrastructures at a rapid pace. For example, users of the GPU, which is required for machine learning by AI, can select from available menus on the cloud. Even the special FPGA hardware is on clouds.

Mechanisms that never existed before, such as the blockchain, are now provided by SaaS (software as a service). Whenever a new architecture appeared in the past, the barrier that faced an engineer was the preparation of the environment in which it operated. Now this barrier is considerably lower. New streamlining methods also exist, for example, server-less architectures, that only clouds can generate. As a result, it is now possible to launch a worldwide data processing service at low cost and without considering infrastructure preparation or performance design. It is also easier than before to migrate large amounts of data from an on-premise environment to a cloud and to retrieve it or migrate it to another cloud. Clouds enable the user to acquire any amount of data at any time, and to withdraw it at any time. In addition to being of great benefit to users, clouds bring flexible options in the use of system infrastructures to business.

5. Changes required of users

Users need to acquire more knowledge and know-how to be able to handle more diversified and complicated infrastructures. Processors today achieve high speeds by cooperating with software. However, once a performance problem occurs, it is difficult to repair it by simply scaling up or scaling out. Although a GPU's software library reduces the burden on developers, its performance may be compromised without tuning it in a way that considers the library features. Even though the FPGA is on clouds and is operable on a browser, it is essential to have the necessary knowledge to implement the logic unique to this hardware. Because clouds present flexible and diversifying options, it is not easy to review the entire system lifecycle and select suitable cloud services, along with an option to leave part of the system on premises. As a consequence, it is necessary to develop human resources who can actively adapt to changes and today's technological trends, which go beyond the framework of traditional software and hardware, and who can reexamine system infrastructure.

Rena Oi

Assistant Manager, Strategy Development Section, Research and Development Headquarters, NTT DATA Corporation.

She received a B.A. in sociology from Keio University, Tokyo, in 2008. After joining NTT DATA in 2008, she worked on the development of a customer contact system. She joined the NTT DATA Technology Foresight team in 2017.