Proceedings of the 4th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/84/20240894

The impact of semiconductor technology on edge computing

Weiyi Luo

Information Engineering, Nanchang University, Nanchang, Jiangxi, 330031, China

withbyfan@gmail.com

Abstract. Semiconductor technology has profoundly shaped the landscape of edge computing by enhancing energy efficiency, processing power, and device miniaturization. Continued advancements in semiconductor tech promise further integration of AI, improved edge-to-cloud synchronization, and heightened security, fueling growth and innovation in edge computing. This synergy is evident as semiconductor tech enables versatile edge device deployment, while edge computing applications like real-time analytics and machine learning drive innovation, resulting in purpose-built chips for edge computing applications. Semiconductor tech and edge computing are symbiotic, with one driving the other's evolution, forming an inseparable partnership in the realm of computing. This paper describes the role of semiconductor technology in driving edge computing technology in terms of energy, performance, and miniaturization, and incorporates data that predicts future directions for semiconductor-based technologies in edge computing.

Keywords: silicon integrated circuit. technology development. edge computing.

1. Introduction

1.1. Evolution of semiconductor technology

The evolution of semiconductor technology can be traced back to the mid-20th century. In the early 1940s, scientists at Bell Labs began researching new materials such as silicon and germanium to explore their potential applications in electronics. Subsequently, in 1947, Bell Labs released the first point-contact transistor made from germanium semiconductors, which marked the birth of semiconductor devices. The invention and widespread use of the transistor became a milestone in semiconductor technology. It replaced the early vacuum tubes with lower power consumption, smaller size, and longer life. Transistors can not only detect signals, but also be used for rectification, amplification, switching, voltage stabilization, signal modulation, and many other functions, so they are widely used in electronic equipment.

Then, the development of semiconductor technology entered the integrated circuit era. In 1958, Jack Kilby invented the silicon integrated circuit, a technology that allows multiple transistors, resistors, capacitors, and other components to be integrated into the same silicon wafer, thus dramatically increasing the complexity and performance of circuits and reducing costs. This breakthrough technology laid a solid foundation for the development of computers and electronic devices.

Over time, semiconductor technology continued to evolve into the era of large-scale integrated circuits (LSI) and very large-scale integrated circuits (VLSI). These technologies made it possible to integrate thousands or even millions of transistors on a single chip, driving a rapid increase in computer performance and thus giving rise to the revolution in modern computer and communications technology.

At the beginning of the 21st century, semiconductor technology continued to move forward into the era of microelectronics and nanotechnology. Nanofabrication techniques allowed for the integration of more transistors on a chip, improving energy efficiency and performance. This period also saw the emergence of innovative applications such as new memories, sensors, and processors, such as smartphones, cloud computing, and the Internet of Things.

Semiconductor technology now plays a key role in supporting edge computing and IoT applications. Advances in modern semiconductor technology have made it possible to realize more powerful computing in small, low-power devices, driving the development of smart sensors, embedded systems, and IoT devices. With the rapid development of IoT technology, new applications such as virtual reality, drones, and remote surgery have emerged, which are usually deployed on end devices and require low latency, high bandwidth, and greater storage capacity. Traditional cloud computing requires data to be transmitted to a cloud data center for processing, but this raises issues such as transmission latency, network bandwidth, energy consumption, and data security. Therefore, edge computing has emerged to address these issues. Edge computing is a new computing paradigm that reduces computation latency and device energy consumption by rapidly processing data on edge nodes at the edge of the network while reducing the burden on cloud servers. Semiconductor technology plays a key role in supporting edge computing and IoT applications. Advances in modern semiconductor technology have made it possible to realize more powerful computing capabilities in small, low-power devices, driving the development of smart sensors, embedded systems, and IoT devices.

1.2. Edge computing

Key technologies for edge computing include computing offloading, resource management, service orchestration, mobility management, and edge caching. Among them, computation offloading is offloading computing tasks from end devices to network resource-rich edge servers to reduce latency and power consumption and improve service quality. Offloading decisions are critical to reduce latency, and energy consumption, or balance the two. Resource management involves efficient allocation of server resources, single or multi-node allocation depending on the nature of the task. Edge caching, on the other hand, is used to store resources, reduce the pressure on network bandwidth from data traffic generated by end devices, and improve response time [1].

The application of these key technologies is critical when deploying edge computing platforms on devices to perform tasks. Compute offloading optimizes task processing, resource management ensures efficient server utilization and edge caching reduces network bandwidth pressure [2]. These technologies work in tandem to make edge computing more efficient and provide support for task execution on end devices.

2. Semiconductor Advancements for Edge Computing

2.1. Advancement

Semiconductor technology is the core driver of technological advancements in edge computing. Edge computing devices are single-board computers, IoT gateways, edge servers, etc.; all of these devices collect, analyze, transfer, and send data through integrated circuits. The basis of IC manufacturing is semiconductor materials, which have special conductive properties that can be used to realize logic operations and storage functions by controlling the flow of current; therefore, semiconductor technology is the key to realizing integrated circuits, which includes several process steps, such as wafer preparation, photolithography, thin film deposition, ion implantation, and metal joining. These steps not only require highly accurate equipment and process control but also operate at the nanometer scale. Advances in semiconductor technology directly impact the performance and functionality of integrated circuits, and thus edge computing devices. In other words, semiconductor technology is the cornerstone of edge computing, and advances in semiconductor technology inevitably drive advances in edge computing. At the same time, edge computing, as the end-use industry of semiconductor technology, its development needs will also push back the progress of semiconductor technology.

The goal of edge computing is to provide computation, storage, and network bandwidth close to data inputs or users. It has emerged to compensate for some of the shortcomings of cloud computing. Although cloud computing provides efficient computing and storage resources, there are still some problems in certain scenarios. First, cloud computing usually requires data to be transferred from end devices/edge devices to cloud servers for processing, which may trigger latency problems, and it is not suitable for some applications with high real-time requirements, such as intelligent transportation systems or industrial automation control systems; second, cloud computing also has a large amount of data transmission and storage requirements, which may lead to network congestion and privacy and security issues. The emergence of edge computing solves these problems well with its proximity to the data source, which reduces the latency of data transmission and provides real-time responsiveness. To give a concrete example, smart home systems can utilize edge computing technology to achieve intelligent local decision-making and control; for example, when a person walks into a room, the edge device can immediately detect and trigger the relevant lighting and temperature adjustments without waiting for the data to be transmitted to a cloud server and then return the command. At the same time, handing over uncomplicated calculations to edge devices also reduces energy consumption [3].

2.1.1. Miniaturization. Advances in semiconductor technology have brought about the miniaturization of integrated circuits, that is, the same volume of the case of integrating more semiconductor materials, the chip’s computing power has been greatly improved, and high-performance semiconductor chips can provide faster computing speed to meet the real-time requirements; at the same time, more miniaturized chips also make the edge of the device is more lightweight, portable, and to provide greater flexibility. The miniaturization trend of semiconductors is also reflected in the rapid industrial development in recent years. ArF immersion exposure uses an ArF light source with a wavelength of 193 nm as light source, and a “liquid” such as pure water is passed through the wafer (the wafer is immersed in the liquid) prior to irradiating the wafer with light. (the wafer is immersed in the liquid). The light that passes through the mask passes through a reduced projection lens and is projected, making it possible to create patterns on the wafer that are finer than the mask. By using this technique, patterns can be formed with the processing accuracy of the 10nm generation, and until around 2019, ArF immersion exposure is state-of-the-art. However, ASML has built EUV exposures using a 193.13nm EUV light source, which is well below the 5nm wavelength of the ArF light source, allowing for the formation of even finer patterns, such as those of the 5nm generation.

The improvement of the chip’s computing power, the deployment of complex security protection calculations on the edge device has also become possible [4]. Edge computing requires real-time data processing and analysis locally, and semiconductor technology can optimize chip structure and design to achieve more efficient energy use, reduce the power consumption required for computing, and extend the use of the device. Finally, industry advances in semiconductor technology will also bring down the cost of chip production, and with it the cost of edge devices. Thus, as semiconductor technology advances, semiconductor chips become more powerful, more efficient, and more affordable. This makes it possible to deploy edge devices in a wider range of applications and environments.

2.1.2. Processing Power. Semiconductor technology plays a pivotal role in enhancing the capabilities of edge computing across various domains. For instance, in the realm of IoT applications, semiconductor chips serve as the backbone for processing data originating from a multitude of sensors, including cameras and temperature sensors, as observed in smart home setups. In the context of autonomous driving, semiconductor chips come into play for the processing of data stemming from cameras, radar systems, and other sensors, such as those employed in self-driving vehicles to discern objects and navigate obstacles.

Furthermore, semiconductor chips are instrumental in the domain of virtual reality (VR) and augmented reality (AR) applications, where they are responsible for rendering lifelike images and videos, contributing to the immersive experiences within VR headsets. It is worth emphasizing that semiconductor technology stands as a fundamental enabler of edge computing. As advancements in semiconductor technology persist, they pave the way for increasingly innovative and robust edge computing applications, propelling the field forward. These developments could yield many benefits. For example, researchers have shown that using cloudlets to offload computing tasks for wear-able cognitive-assistance systems improves response times by between 80 and 200 ms[4] and reduces energy consumption by 30 to 40 percent. Cloud technology reduces response times and power usage by95percent for tested applications, in part via edge computing

2.2. Limitations

As transistor sizes in integrated circuits continue to decrease, more transistors can be accommodated in a circuit. While this theoretically improves the performance of the circuit, it also introduces several problems. First, as transistors become smaller, the voltage that needs to be applied to them must also be reduced to avoid problems such as electrical breakdown. This leads to a reduction in voltage, which limits the performance of the transistors. In addition, as the distance between the transistors in a circuit decreases, the electric field effect becomes more pronounced, further reducing the performance of the transistors.

For small transistors, the voltage span is usually only between zero and one volt, while for large transistors, the voltage span can be much larger, even up to more than ten volts. However, in modern tiny transistors, reducing the voltage to even lower levels is virtually impossible. This is because there exists a fundamental voltage noise level of about 26 million electron volts, caused by thermal fluctuations at room temperature. Even when a voltage is applied to the circuit, it still fluctuates randomly up and down at that voltage, about plus or minus 20 million electron volts. Therefore, in very small transistors, it is impractical to reduce the voltage to a low enough level to reduce power consumption.

This phenomenon has implications for deploying edge computing platforms on devices to perform tasks. Edge computing platforms typically require small, low-power devices to process data and run applications, and therefore require small transistors. However, as mentioned earlier, small transistors face challenges in terms of voltage reduction, which can limit the performance and energy efficiency of edge devices. Therefore, in the field of edge computing, it has become critical to research and develop low-power, high-performance small transistors, and circuit designs to meet the demand for low latency, high bandwidth, and greater storage capacity. This also reflects the critical role of semiconductor technology in supporting edge computing and IoT applications, as advances in modern semiconductor technology have made it possible to realize more powerful computing capabilities in small, low-power devices.

3. Future Trends

Currently, edge computing suffers from major problems such as network latency, resource constraints, and security risks. Since edge devices are usually located far away from data processing centers such as cloud platforms, network latency leads to increased delay in data transmission, affecting real-time and user experience [5]. Due to cost and IC size, edge devices have limited computing and storage resources to meet the demands of large-scale or complex tasks, so the performance and functionality of current edge devices are very limited. Edge devices are often distributed and operate in uncontrolled environments, making them susceptible to physical attacks as well as at risk of data leakage [6].

Edge computing applications have also driven innovation in semiconductor technology. For example, the growing need for real-time analytics and machine learning in edge computing is driving the development of new semiconductor chips optimized for these applications [7].

3.1. Performance Enhancement

As edge computing continues to grow in popularity, semiconductor technology is likely to play an even more important role. Future trends in semiconductor technology that may have an impact on edge computing include.

(1) Development of new chip architectures optimized for edge computing applications; edge computing applications have unique requirements that cannot be met by traditional chip architectures. For example, edge computing applications often need to be able to process data in real-time with low latency and low power consumption. They also need to be able to operate in a variety of environments, including harsh conditions such as extreme temperatures and vibration; the development of new chip architectures optimized for edge computing applications will enable the development of more powerful and efficient edge devices. This will lead to new and innovative applications for edge computing, such as real-time video analytics, predictive maintenance, and self-driving cars.

(2) Development of new manufacturing processes to produce semiconductor chips at lower cost and higher performance; Semiconductor manufacturing is a complex and expensive process. However, new manufacturing processes are being developed to produce semiconductor chips at lower cost and higher performance. This makes it possible to develop more economical and powerful edge devices. An example of a new manufacturing process is FinFET. FinFET chips are more powerful and efficient than traditional planar chips. However, FinFET chips are also more expensive to manufacture. New manufacturing processes are being developed that could reduce the cost of manufacturing FinFET chips. Another example of a new manufacturing process is 3D stacking. 3D stacking allows multiple chips to be stacked together. This can improve the performance and functionality of edge devices without increasing their size or power consumption. The development of new manufacturing processes capable of producing semiconductor chips at lower cost and higher performance will enable the development of more affordable and powerful edge devices. This will lead to the adoption of edge computing in a wider range of applications.

(3) The development of new packaging technologies can improve the performance and reliability of edge devices; semiconductor packaging is the process of encapsulating semiconductor chips in a protective layer. Packaging is important to protect the chip from damage and to provide a way to connect the chip to other components. New packaging techniques can improve the performance and reliability of edge devices. For example, some new packaging techniques can reduce the thermal resistance between the chip and its surroundings. This can help improve the performance of the chip by allowing it to operate at higher temperatures. Other new packaging technologies can improve the reliability of the chip by protecting it from vibration and shock.

3.2. Artificial intelligence hardware

Future edge computing devices will rely more on superior chip performance, especially in the AI space. On the data centre side, with increasing demand for AI applications in cloud computing data centers, GPUs and ASICs will compete for market share and are expected to split the market equally by 2025, while GPUs are likely to be more customized to meet deep learning needs. On the inference side, while CPUs currently dominate, ASICs will gradually replace CPUs as deep learning applications become more popular. Overall, the hardware market will become more diverse and competitive to cater to the needs of different application areas [8].

Although memory is growing at a relatively low CAGR, mainly driven by increased efficiency in algorithm design (e.g., reduced bit precision) and relaxed capacity constraints, the short-term memory market is being driven by increased demand for high-bandwidth DRAMs in data centers, especially to support artificial intelligence, machine learning, and deep learning algorithms. However, the demand for AI memory for edge computing will also increase over time, and more DRAM may be required for connected cars. Currently, memory is typically optimized for CPUs, but developers are looking for new architectures to meet high-performance AI needs [9]. The high memory bandwidth requirements for AI applications, as deep neural networks need to rapidly transfer data to thousands of cores, will create a huge opportunity for the memory market.

In addition, AI applications generate a large amount of data every year. This increases the demand for storage and manufacturers will accordingly increase the production of storage gas pedals, whose pricing will be influenced by supply and demand. Unlike traditional storage solutions, AI hardware developers are looking at new options such as programmable switches to improve data synchronization and training speed. Programmable switches can accelerate re-synchronization when model parameters are updated and improve training speed. Another way to improve networks is to use high-speed interconnects, which can triple performance, albeit at a relatively high cost. These innovations provide new opportunities for semiconductor companies [10].

4. Conclusion

Overall, the semiconductor industry has had a profound impact on the edge computing field. Various technologies in edge computing are associated with the miniaturization of integrated circuits and performance enhancement of chips. Despite the limitations of semiconductor technology, the field of edge computing will continue to evolve shortly as chip computing power increases. AI applications in edge computing will also benefit from this.

References

[1] K. -C. Chang, K. -C. Chu, Y. -C. Lin, H. -C. Wang, T. -L. Hsu and F. -H. Chang, “Study of Fpga-Based Edge Computing in Semiconductor Manufacturing Safety Management Application,” 2020 15th International Microsystems, Packaging, Assembly and Circuits Technology Conference (IMPACT), Taipei, Taiwan, 2020, pp. 273-276, DOI: 10.1109/IMPACT50485.2020.9268579.

[2] W. Shi and S. Dustdar, “The Promise of Edge Computing,” in Computer, vol. 49, no. 5, pp. 78-81, May 2016, DOI: 10.1109/MC.2016.145.

[3] Z. Ke, Silicon integrated circuit technology and development trend, 2004, DOI: 10.16257/j. cnki.1681-1070.2004.04.002

[4] Z. G. Yu, Development trend and prospect of silicon integrated circuit, 2003, DOI: 10.16257/j. cnki.1681-1070.2003.02.001

[5] R. Deng, R. Lu, C. Lai, T. H. Luan and H. Liang, “Optimal Workload Allocation in Fog-Cloud Computing Toward Balanced Delay and Power Consumption,” in IEEE Internet of Things Journal, vol. 3, no. 6, pp. 1171-1181, Dec. 2016, DOI: 10.1109/JIOT.2016.2565516.

[6] Kuo-Chi Chang, Yuh-Chung Lin, Kai-Chun Chu. Mobile Edge Computing Technology and Local Shunt Design. The Frontiers of Society, Science, and Technology (2019) Vol.1Issue10: 135-140. https: //doi. org/10.25236/FSST.2019.011017.

[7] Cagatay S, Atay O, Cem E. Edge Cloud Sim: An environment for performance evaluation of edge computing systems[J]. Transactions on Emerging Telecommunications Technologies, 2018. https://doi.org/10.1002/ett.3493

[8] Batra, Gaurav, et al. “Artificial-intelligence hardware: New opportunities for semiconductor companies.” McKinsey & Company: Hong Kong, China (2018).

[9] VerWey, John. “The health and competitiveness of the US semiconductor manufacturing equipment industry.” Available at SSRN 3413951 (2019).

[10] Chang KC. et al. (2020) Study on Hazardous Scenario Analysis of High-Tech Facilities and Emergency Response Mechanism of Science and Technology Parks Based on IoT. In: Pan JS., Lin JW., Liang Y., Chu SC. (eds) Genetic and Evolutionary Computing. ICGEC2019. Advances in Intelligent Systems and Computing, vol1107. Springer, Singapore. https: //doi. org/10.1007/978-981-15-3308-2_22