Flash Memory and Its Manufacturing Technology for Sustainable World

KAZUNARI ISHIMARU (Fellow, IEEE), MAKOTO FUJIWARA (Member, IEEE), HIDENORI MIYAGAWA, AND YUTA AIBA
Institute of Memory Technology Research and Development, Kioxia Corporation, Yokkaichi 512-8550, Japan
CORRESPONDING AUTHOR: K. ISHIMARU (e-mail: kazu.ishimaru@kioxia.com)

ABSTRACT In today’s data-driven society, memory is an unavoidable component. IoT and 5G are accelerating data generation exponentially, and the demands of high-capacity memory are increasing. Flash memory is a crucial infrastructure component and inevitable for Data Center. The advantages of a Solid-State Drive over a Hard Disk Drive are power consumption and volume per giga-byte. Low power consumption of storage devices, which are components of the data center itself, is critical in reducing power consumption, which has become a primary social concern. This paper describes current efforts and future trends in storage devices and their manufacturing technologies for realizing a sustainable society. New device structure which increases storage capacity with significant reduction of both materials and process steps, environmentally friendly and efficient manufacturing technologies and future technologies under consideration for even higher capacity will be shown. These technological innovations will make it possible to realize a society that enriches people’s lives.

INDEX TERMS Flash memory, storage, sustainability.

I. INTRODUCTION
The amount of data generated in today’s information society is growing at an exponential rate, and high-speed networks like 5G are accelerating this trend. Within a few years, it is expected that the annual amount of data generated will exceed 175 ZB [1]. The data are not only created by people, but also by machines which become dominant. These enormous amounts of data are being actively used to generate new value and business. However, less than 10% of these data are currently stored [2]. People’s lives would be better enriched if more data could be stored and used as information and knowledge. For example, if we can store people’s memories and retrieve them when needed, we can use them to treat dementia, pass on the knowledge of experts, and do many other things. Thus, the requirement for huge storage is also increasing, and it is important to meet such demands while achieving a sustainable society. By lowering the cost of each bit of data and making it more widely available, a safe and convenient lifestyle with high-quality data can be achieved. In this paper, flash memory technologies for storing the explosive growth of data, manufacturing technologies that are friendly to the global environment, and what kind of future can be envisioned by these technologies are discussed.

II. MEMORY SURROUNDING US
We welcome the challenge of pursuing limitless possibilities of “Memory.” There are two types of “Memory.” One is a kind of “Record” or a “Media” to store them. This memory as a record is generally tangible, and culture, history, scenery, technology are stored in text, sound, images, video format. It can be digitized and utilized in a variety of ways. These memories are everywhere in our lives as shown in Fig. 1. Memory cards or USB sticks, for example, are used to store data on cell phones, digital cameras, tablets, game consoles, and personal computers in our home. Some are embedded in the system, and people do not realize it in daily life. People can enjoy reading books and watching movies through smartphones and tablets. The latest high-end game console equips with high speed and large capacity SSD (Solid-State Drive), which stores many game software with high-quality video and sounds. Online games and e-sports are a growing market, the demand for those SSD is increasing rapidly. In the office, PCs and servers use high-speed and large-capacity...
SSDs. Many applications are moving to cloud services these days, and Data Centers store large amount of data that people access via the network. Peoples’s work styles have shifted as a result of remote access, and work from home is now possible. From powertrains to entertainment systems, our cars rely heavily on semiconductors. Electric vehicles and self-driving cars will boost semiconductor device usage in coming years, and automobiles will serve as mobile edge servers. The transition from data-gathering devices to memory-gathering devices that inspire the future is underway.

The other memory is “Human Memory” in Japanese “Ki-Oh-Ku.” “Ki-Oh-Ku” is different from “Record”; it is a record that is accompanied by information from emotion and the five senses, such as “touch,” “taste,” “hearing,” “eyesight,” and “smell.” Some are already written in books or recorded in videos; however, they are intangible and need to be digitized for utilization as knowledge and wisdom. There is currently no commercial tool that can copy human memory to storage today, and researchers are working to enable it near in the future.

Both “Memory” is essential, and we can provide the value of “Memory” to society through products such as flash memory and SSDs, which will continue to develop a sustainable society. We can enrich people’s lives by uplifting the world with “Memory.” However, to do so, we must first create a lifestyle infrastructure that cannot be realized unless global environmental problems are solved. We also need to implement measures against climate change and ensure the effective use of resources in its business activities. Furthermore, with flash memory and SSD products, we are helping people achieve convenient and comfortable lifestyles where Internet access is available anytime and anywhere. In addition, we believe that once such lifestyles have been achieved, we can provide people with more creative lifestyles through “Memory” as shown in Fig. 2 [3], [4].

In addition to the physical space in which people live, technological advancements have created digital space. Environmental conservation, smart cities, the ability to provide people with creative experiences that transcend time and space, and other benefits will all be realized as digital technology advances. Storage products serve as the foundation for a digital society, and manufacturers continue to pursue further research and development. In addition, it is crucial to contribute to future society by undertaking digital technology research and development through “backcasting” from the perspective of various future social issues in the physical environment. Such technology to realize a sustainable society will be shown in the next section, using NAND flash memory as an example.

III. TARGET APPLICATION

The amount of data generated and accumulated by humans is steadily increasing, thanks to the widespread adoption of cloud computing in the IoT era and the increased use of AI technology. On a global scale, our research and technological advancements help to support the future information society. Some examples are shown in Fig. 3.

A. AUTOMOTIVE

The automobile industry is undergoing a once-in-a-century transformation. Automobiles that handle large amounts of data will serve as mobile edge servers as a result of technological advancements such as connected cars, autonomous vehicles, ride-sharing services, and vehicle electrification. Large-capacity flash memory that is mobile and able to function even in harsh environments is expected to play a role as infrastructure in support of these mobile edge servers.

B. ENTERTAINMENT

People can enjoy new forms of entertainment as a result of technological advancements. Expectations are higher than ever for high-capacity, high-speed, and low-cost flash memory to allow for enormous amounts of information to be processed at high speeds and enable more people to enjoy stress-free, high-quality interactions.

C. ROBOTICS

For future robots that can naturally communicate with people, it is necessary to utilize significant data to realize...
sophisticated interactions based on individuals’ actions, situations, contexts, emotions, and more. Flash memory supports smooth interactions between humans and robots by storing and processing large amounts of data acquired by new sensor technology at high speeds.

D. MEDICAL/HEALTH CARE

Every five years, the amount of data in life science research increases by a thousand times. It is expected to reach the Exabyte level in 10 years. Therefore, flash memory that can store huge amounts of data and systems that can analyze and process this data at high speeds are needed. We will continue to support the elucidation of new knowledge as well as appropriate diagnoses, preventive medicine, and drug discovery.

E. INDUSTRY

Digital twins, factories in the real world reproduced in virtual spaces, are achieved by sending factory equipment, machinery, and environmental information to a virtual space in near real-time using technology such as IoT. Advanced simulations that use large amounts of data in virtual spaces, as well as technology that allows for rapid feedback to the physical site, are supported by flash memory. This contributes to the creation of new value and significant productivity improvements.

IV. TECHNOLOGY FOR SUSTAINABLE STORAGE

A. MEMORY CELL TECHNOLOGY

As mentioned earlier, to store the explosive growth of data, the storage capacity must be increased to match the growth. Over the last 20 years, flash memory density per unit area (GB/cm²) has grown at a rate of about 40% per year, resulting in increased storage capacity. It will continue at least coming decade and will be more than one terabyte per cm² in 2030. The transition from 2D to 3D structure enabled this continuous growth of flash memory capacity as shown in Fig. 4. The number of memory layers in 3D flash memory, such as BiCS FLASH, has increased from 96 layers to 170 layers over the past three years [5]. If the number of layers continues to increase at this rate, it is expected to reach around 1000 layers by 2030.

On the other hand, an increase in stacking cell layers will lead to an increase in the number of manufacturing processes and CO₂ emissions associated with manufacturing. To realize a sustainable society, it is necessary to be environmentally friendly and improve performance. Multi-level cell technology [6] is one of several ways to increase memory capacity without increasing the number of layers as much as possible. One of the features that distinguish flash memory from other types of memory is multi-level cell technology. In the past, one memory cell stored 1bit information and was increased to 2bits, 3bits, and now it stores up to 4bits of information (Fig. 5). Multi-bit technology has the advantage of requiring no extra memory cell layers or process steps. However, as can be seen in the figure, as the number of bits increases, the width of the threshold voltage (Vth) distribution needs to be narrower, which makes the programming sequence complicated and degrade performance. The benefit of increasing ratio in storing capacity also decreases as several bit increase. The metal-assisted solid-phase crystallization process [7] is one of the promising technology to improve Vth variability as shown in Fig. 6 and expands multi-bit capability further.

Another approach is to multiply cells “physically,” by splitting memory cells. Fig. 7 shows the schematic image of Twin BiCS FLASH [8], which is made with the same technology of BiCS FLASH and just adds one etching process. The cell size of BiCS FLASH memory cells can be halved by dividing them and making them semicircular. As a result, doubling the capacity is realized. This technique could be used to multiply memory cells further, such as quadruple and octuple. However, there is a trade-off between capacity and stored charges, and less charge causes performance

FIGURE 4. FLASH memory density increases roughly 40%/year by shifting 2D to 3D structure.

FIGURE 5. Threshold voltage distribution of multi-level cell from SLC(1bit/cell) to QLC(4bit/cell).

FIGURE 6. Improvement of Vth variability by metal-assisted solid phase crystallization process [7] ©IEEE 2019.
FIGURE 7. Double the density by Twin BiCS FLASH technology which enables significant reduction of both material and process time.

FIGURE 8. A multi-tier structure is one of the alternative options to enable capacity BiCS FLASH from cost and sustainability points of view.

degradation that 2D structure faced. Further research and development are conducted in order to realize high-density memory with low environmental impact.

B. PROCESS TECHNOLOGY

The challenge for the BiCS FLASH process is how far we can stack memory layers by minimizing manufacturing costs. The etching speed of the memory hole becomes slower when the aspect ratio increases. The difference of memory hole diameter between top and bottom also becomes large, which causes variability in cell characteristics. New process gas and hardware are required to improve etching speed and improve memory hole depth profile, which leads to manufacturing cost increase and product lead time. From a sustainability point of view, it is desired to use the same process tools and materials as much as possible. The introduction of a multi-tier structure, as shown in Fig. 8 [9], is one solution for mitigating this issue. The memory hole aspect ratio is cut in half when the array stack height is divided in half, and the memory hole size difference between the top and bottom layers can be minimized. Reducing the hole aspect ratio leads to higher throughput and even repeating stack, etch, and fill sequence than the single-tier structure. To meet the demand for device miniaturization, shorter wavelengths and higher NAs that increase the lens diameter have been introduced into the optical lithography process. As wavelength reduction and NA heightening are approaching their physical limitations, new techniques emerge, such as multiple patterning that repeats optical lithography several times or EUVL (Extreme Ultra-Violet Lithography). However, process costs will inevitably increase due to the cost of process step increase and additional process tools. To overcome the lithography process cost increase, we are developing nanoimprint lithography that can miniaturize devices at a lower cost. Imprinting nanoscale patterns on a template onto a Si wafer is used in the nanoimprint technique. It does not require a lens optical system for reduction projection, unlike traditional lithography tools, shown in Fig. 9. The nanoimprint is a highly anticipated next-generation lithography method to realize advanced memory devices with reduced cost.

C. PACKAGING AND SYSTEM INTEGRATION TECHNOLOGY

In comparison to hard disk drives, SSDs have a smaller physical size, are lighter, and use less energy. Their lack of mechanically moving parts, which makes them more shock resistant, is suitable for mobile applications, a growing market. Another advantage is that it is easy to integrate with other LSIs and enables the addition of more functions. Multiple BiCS FLASH dies are contained in the current flash memory package. Wire bonding is used to connect each memory dies and stacking by shifting them so that wires do not interfere with each other. The bonding pads and wires increase parasitic capacitance and inductance, affecting and degrading read and write speeds. Each dies to require I/O circuits to drive signals to the bonding pad, which results in a chip size increase. Through Silicon Vias (TSV) technology enables to stack dies vertically without shifting them. It does not need a bonding pad that can eliminate the I/O circuit for each dies as shown in Fig. 10. Smaller die sizes reduce parasitic capacitance and inductance, improving both speed and power consumption [10]. TSV can stack different dies, it can integrate the memory controller and other system LSIs in the future, in addition to the BiCS FLASH, because it can stack different dies. Fig. 11 shows one example of a future storage system. Current high-end SSDs consist of multiple BiCS FLASH packages, a memory controller, and DRAMs. These dies can be integrated into one package as a one package SSD. Furthermore, those one package SSD will be directly attached to accelerators or even integrated with the accelerator, DRAM, XL-FLASH, or other storage-class memory, and realizes a tiny mobile edge server. There are many issues for realization, such as cost and thermal management. However, it is a viable future.
V. TECHNOLOGY FOR SUSTAINABLE MANUFACTURING

A. HIGH-EFFICIENCY MANUFACTURING

Every day, we collect over two billion pieces of data in real-time from production equipment and transportation systems at the BiCS FLASH manufacturing plant, and we use AI technology to conduct big-data analysis. By doing so, we analyze any failures and strive for increased productivity. Smart factories that utilize these types of AI technology go further than simple automation with “smart” functions that lead to a sustainable society. These include reducing the number of materials used in production by avoiding the production of defective products, reducing energy consumption by shortening turnaround times, and ensuring a stable supply of high-quality products (Fig. 12). Big data is a goldmine of useful information and contains a lot of buried wisdom. It is, however, difficult to locate manually. Therefore, AI technology is used to understand the big picture while also unraveling the complex relationships between individual pieces of data. In addition to helping us understand the present, this technology also allows us to predict the future, leading to more efficient development and production. Internal and external open innovation was used to develop these state-of-the-art AI technologies. We will continue to use AI technology in semiconductor manufacturing to maintain and improve high productivity as the industry advances in highly integrated products. State-of-the-art semiconductor manufacturing requires highly accurate defect inspection, even if the defects are very small. A new inspection technique was developed utilizing not only conventional image processing but also machine learning. A Die to Die (DD) comparative inspection requires image data for comparison and reference in addition to the inspection image, which increases the imaging time. The method of detecting the difference between SEM images and CAD data is faster than the DD method. However, since the actual pattern does not perfectly match the CAD data, there is a problem of over-detection failure when simply evaluating the difference shown in Fig. 13 (a). Using Generative Adversarial Network (GAN) to generate a virtual SEM image from CAD data, and actual SEM images were added as supplemental information for learning. This method can suppress excessive detection, and defect inspection can be performed with practical accuracy, enabling higher yields and higher quality of products [11].

B. WASTE RECYCLING

In wafer cleaning and etching processes, a large amount of hydrofluoric acid is used, resulting in the generation of high-concentration hydrofluoric acid waste liquid and low concentration hydrofluoric acid wastewater. Traditionally, the coagulation-sedimentation method was used to remove fluorine from wastewater. Due to the large amounts of impurities such as silica and the excessive moisture, this method was problematic in terms of recycling and reuse. Recycling technology was developed in collaboration with partner companies, which allowing high-concentration hydrofluoric acid waste liquid to be recycled as high-purity calcium fluoride (artificial fluorite) as shown in Fig. 14. We have reduced the amount of sludge containing hydrofluoric acid by roughly 30% through this method of recycling. In addition, manufacturers of fluorine-based products use recycled calcium fluoride (artificial fluorite) to produce fluorine-based materials such as hydrofluoric acid. Fluorite, an important natural source, is used less because of this recycling.

C. PFC REDUCTION

PFC (a greenhouse gas) used in the manufacturing process is used to form wiring and contact holes during P-CVD (plasma CVD) process, metal-CVD process, and plasma
etching process. The installation of abatement equipment is effective in reducing PFC gas emission because it breaks down PFC gases and emits them as gases with low global warming potential (Fig. 15). In addition, by introducing high-efficiency equipment and optimizing cleaning time, we are also working to reduce the amount of PFC gases used during the cleaning of reaction chambers.

VI. FUTURE TECHNOLOGY OPTION

One of the issues, as previously mentioned, is the server’s power consumption. These days, liquid cooling of the entire server system is considered. We can put a small edge server in a liquid nitrogen ambient if we can build one. Figs. 16 and 17 show preliminary results of the 77K operation of BiCS FLASH™. It shows tight Vt distribution and robust cycling characteristics compare to those of room-temperature operation. This enables a 6bit per cell (HLC) operation. These are just preliminary results. However, cryogenic sever is a considerable future system. This cryogenic server will help to improve performance and reduce the physical size of the future quantum computing system, which uses very low temperature in the mK range. It can also expand the application to another area such as aerospace, and we can build Data Center on the Moon or Mars.

Finally, future CMOS manufacturing is discussed. The planar structure was used for transistors and migrate to 3D structures such as FinFET. Gate-All-Around (GAA) or nano-sheet or nano-ribbon structure will be used. When looking at the cross-section of these transistors, we can find similarities between nano-sheet or nano-ribbon structure with BiCS FLASH (Fig. 18). We can use the BiCS FLASH process and integration scheme if the future CMOS has multiple nano-sheet or nano-ribbon structures because BiCS FLASH focuses on low manufacturing costs. There are several challenges. However, from a sustainability standpoint, it will be a viable option.

VII. CONCLUSION

Flash memory is an inevitable component of our society. It spreads around everywhere and is essential to reduce the cost of every bit of data and make it more widely available. Leveling up the world with every bit of data and giving value to a new way of life with “Bits” are desired. It is essential for the realization of a sustainable society to not only provide those but also to enable high-efficiency manufacturing. Twin BiCS FLASH and multi-tier technology enable high-density storage with minimizing the usage of resources with higher productivity. TSV technology realizes one package SSD, which enhances the performance and scales down the total storage system. The cryogenic operation can be a future option enabling not only further density enhancement but also expands SSD application to the aerospace. We can use the BiCS FLASH process and integration scheme for the future CMOS, such as multiple nano-sheet and nano-ribbon structures enabling low manufacturing costs. There are several challenges; however, from a sustainability standpoint, it will be a viable option.

ACKNOWLEDGMENT

The author would like to thank colleagues of Kioxia Corporation who are working on technology development and also partner companies to support our R&D activities.

REFERENCES

[1] K. Ishimaru, “Future of non-volatile memory -from storage to computing,” in Tech. Dig. Int. Electron Devices Meeting, 2019, pp. 12–17.
[2] [Online]. Available: https://www.statista.com/statistics/871512/ worldwide-data-created (accessed: Dec. 29, 2021).

[3] [Online]. Available: https://www.kioxia-holdings.com/en-jp/ sustainability/management/sdgs.html (accessed: Dec. 29, 2021).

[4] K. Ishimaru. "‘Memory’ for sustainable society," in Proc. 21th Int. Workshop Junction Technol. (IWJT), 2021, pp. 12–14.

[5] T. Higuchi et al., “30.4 a 1Tb 3b/Cell 3D-flash memory in a 170+ word-line-layer technology,” in Proc. IEEE Int. Solid State Circuits Conf., Feb. 2021, pp. 428–430.

[6] N. Shibata et al., “A 70nm 16Gb 16-level-cell NAND flash memory,” in IEEE Symp. VLSI Circuits Dig. Tech. Papaper, Jun. 2007, pp. 190–191.

[7] H. Miyagawa et al., “Metal-assisted solid-phase crystallization process for vertical monocrystalline Si channel in 3D flash memory,” in Tech. Dig. Int. Electron Devices Meeting, 2019, pp. 28.3.1–28.3.4.

[8] M. Fujiwara et al., “3D semicircular flash memory cell: Novel split-gate technology to boost bit density,” in Tech. Dig. Int. Electron Devices Meeting, 2019, pp. 642–645.

[9] S. Inaba, “3D flash memory for data-intensive applications,” in Proc. IEEE 10th Int. Memory Workshop, 2018, pp. 1–4.

[10] K. Matsudera and K. Kawasaki. “World’s first 16-die stacked NAND flash memory package fabricated using TSV technology,” Toshiba Rev., vol. 71, no. 6, pp. 20–23, 2016.

[11] Y. Konno et al., “A study on effective use of machine learning for inspection and metrology of SEM image,” in Proc. 37th NANO Test. Symp. (NANOTS2017), Nov. 2017, pp. 13–18.

[12] Y. Aiba et al., “Cryogenic operation of 3D flash memory for new applications and bit cost scaling with 6-bit per cell (HLC) and beyond,” in Proc. 5th IEEE Electron Devices Technol. Manuf. Conf., 2021, pp. 9–11.