3D NAND Based on Conventional Flash Memory Architecture Research and Application Prospects

Botao Li [1, *], Ziqi Lu [2], Yuwei Xu [3]

[1] College of Microelectronics, Shanghai University, Shanghai, 201800, China

[2] Hangzhou No. 4 Middle School International Department, Hangzhou, 310000, China

[3] College of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin, 300382, China

* Corresponding Author Email: wulitao@shu.edu.cn

Abstract. With the rapid development of computer technology as a whole, the performance requirements of modern computer systems for storage technology are becoming higher and higher. This paper discusses the response to the problems that have arisen with the traditional storage in today's era, that is, the construction principle of the new storage three-dimensional memory (3D NAND) and the future development. Traditional storage in order to enhance the storage space of the plane flash memory, can only be components for shrinking, when shrinking to a certain extent when the flash memory will occur uncontrollable hidden danger, so the invention of 3D NAND. By vertically stacking many layers of memory cells, the storage density per unit area is greatly increased, with high-speed reading and writing, large capacity, low power consumption and other characteristics. Nowadays, 3D NAND is widely used in enterprises, and the industry is hopeful about the long-term prospects of the growing global demand for data generation and storage and the introduction of new technologies. The development of 3D NAND technology provides higher storage density and lower cost for NAND flash memory to meet the growing demand for data storage. This paper provides a systematic summary of existing research on 3D NAND in the hope that it will better enhance the overall technology improvement in the storage industry.

Keywords: 3D NAND; flash Memory; memory facilities; application.

1. Introduction

With the rapid development of cloud computing, big data and other technologies, the explosive growth of data poses a huge challenge to the capacity, performance and reliability of storage systems. Over the past decade, traditional memory has shown an extremely rapid replacement trend, after decades of rapid development, three-dimensional memory to solve the invisible pitfalls of traditional flash memory, with many advantages, this paper will focus on the exploration of three-dimensional memory based on traditional flash memory 3D NAND.

The current research in the field of 3D NAND mainly focuses on how to increase the number of layers of the chip to achieve higher storage functions, as well as its design principle, characteristics in some cases, there is still an obvious gap between the research level of 3D NAND at home and abroad. With the increasing global demand for data processing and storage, the NAND flash memory market is also valued and is about to usher in a new development opportunity. The development of 3D NAND technology provides higher storage density and lower cost for NAND flash memory, which meets the growing demand for data storage. Combined with the latest situation of 3D NAND research at home and abroad, this paper makes a supplement in the systematic review of 3D NAND.

This paper firstly introduces the memory and traditional flash memory as well as a series of problems, points out the limitations of planar memory and introduces three-dimensional memory, explains its structure and its principle, and finally introduces its application in modern life and its future development in view of the advantages of 3D NAND.

2. Characteristics of Storage and Traditional Flash Memory

Memory is made up of numerous storage cells that are ordered in the order of their cell numbers. Each cell is made up of a number of binary bits that indicate the value stored in the memory cell; this structure is quite similar to that of an array; hence memory is commonly expressed by an array in VHDL [1]. Memory is a type of memory that is used to store programs and various types of data. The main memory operates by saving or retrieving various sorts of information based on the address of the storage unit, which is referred to collectively as the access memory. The storage body is the carrier of the storage unit that the main memory puts together; each unit in the storage body can store a string of binary code which represents the message. the total number of bits of the information is known as the word length of a storage unit. The address of the storage unit and the information stored in it are one-to-one correspondence, there is only one unit address, which is fixed, while the information stored in it is replaceable [2].

Internal and external memory are two types of memory in a computer. The computer regularly uses internal memory during program execution and has direct access to it during each instruction cycle. External memory necessitates that the computer read data from an external storage device, such as a tape or disk. This is comparable to a student taking class notes. The information is preserved in "internal memory" if the learner knows the substance of the notes without reading them. If the learner has to refer to his or her notes, the knowledge is in "external memory" [3].

Internal memory comes in a variety of forms. During computing, random-access memory (RAM) is employed as a high-speed temporary memory space. In RAM, data may be stored, read, and replaced with fresh data. When the machine is turned on, RAM is accessible. It includes information related to the current problem that the computer is working on. The majority of RAM is "unstable," which implies that data is lost when the machine is shut off. The read-only memory (ROM) is reliable. It stores the instruction set that the computer need when needed. The information in ROM is "hard-wired" (i.e., it is a physical part of the electronics) and cannot be modified by the computer (thus the phrase "read-only"). Variable ROM, also known as programmable read-only memory (PROM), can be changed by exposing it to an external electrical or optical source (such as a laser) [3]. A digital imaging device's internal memory must be large enough to contain at least one digital image. 1/4 megabyte is required for a 512*512*8-bit picture. As a result, an imaging device that processes a large number of such photos requires several megabytes of memory. Fig. 1 is the outline structure of in-memory memory.

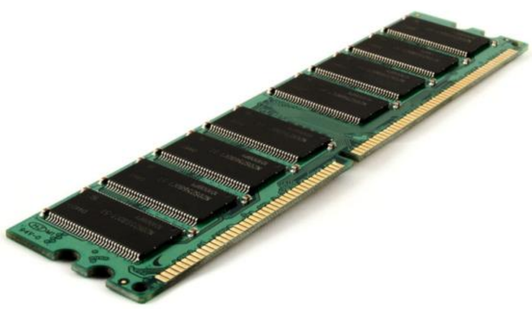

Fig. 1 Outline structure of in-memory memory.
https://blog.csdn.net/weixin_35353583/article/details/118846147

External memory is used to store images that are not acquired during real-time imaging jobs and are physically separated from the computer. Images that have been diagnosed are typically kept for several years for legal purposes. These images are referred to as "archives" (for example, magnetic tapes), and they must be reinstalled on the computer in order to recover the information. Hard disk drive images are physically placed on the computer and may be retrieved in milliseconds. Magnetic domains are used to store individual bits in magnetic memory.

Flash memory (Flash Memory) is a long-life non-volatile (in the event of a power outage, the stored data information can still be maintained) memory, data deletion is not a single byte as a unit, but a fixed block as a unit (Note: NOR Flash for bytes), the block size is normally bytes. Flash memory is a type of electronically erasable read-only memory (EEPROM). EEPROM may be wiped and rewritten at the byte level, whereas flash memory cannot. Most flash memory chips, on the other hand, need block erasure and can only be erased and rewritten in fixed blocks.

3. Problems with Traditional Flash Memory

Conventional flash memories are on a two-dimensional plane, and larger storage space is achieved by reducing the device size. However, conventional flash memories can cause uncontrollable pitfalls after the device size has been reduced to a certain level, which greatly hinders the continued development of conventional flash memories. These pitfalls, also known as reliability issues, can cause NAND flash memory to generate read and write errors. A related concept is the "bathtub hazard rate curve", which grows over time in three parts: initial failure rate, mid- term steady state, and late wear-out rate.

In order to increase storage size and efficiency, 2D NAND flash memories are getting smaller and smaller, and as the cell size gets smaller and smaller, the intersection of the mid-term steady state and the late wear-out efficiency of the "Bathtub Hazard Rate Curve" shifts to the left, which means that the device's stabilization time gets smaller, which means that its durability, data retention characteristics, read disturbances, and so on, all change, read interference, etc. have changed [4, 5].

In conventional flash memory, there are some memory cells that are inherently unstable and these bad memory cells will become more and more as the flash memory continues to be used, hence the introduction of the ECC error correction code, which corrects errors that occur because of bits that have been inverted. However, once the bits in error are out of the error correction range, the data is lost, and such flash memory blocks are discarded and no longer used. In addition to this flash memory also has a problem called read-write interference, from the principle of flash memory reading, when reading a flash page, the flash memory block in the unselected flash page are added positive voltage makes the MOS tube conduction, but frequently in the MOS tube added positive voltage will make the electrons are sucked into the floating gate, resulting in a slight write, resulting in the inversion of the bit. This results in a relatively short lifetime for conventional flash memory.

In order to correct these errors and improve flash memory life, some scholars have researched ways to extend flash memory life by wearing out balancing algorithms, reducing write amplification, and utilizing better error correction algorithms. And to address the problem that planar flash memory devices become unstable when their size is reduced to a certain degree, 3D NAND, a three-dimensional stereoscopic flash memory, has been designed.

4. Structure and Principle of 3D NAND Section Headings

4.1. Structure

Three-Dimensional (3D) uses stacking technology, which is more complex than flat NAND flash memory, and 3D NAND flash memory can significantly increase the storage density of flash memory by vertically stacking multi-layer flash memory units in physical space. The storage unit structure and manufacturing process are improved by using space compression technology. According to the characteristics of the unit structure, 3D stacking can be divided into Gate Stack (GS) architecture and Channel Stack (CS) architecture. In addition to the original multi-layer planar NAND flash memory, manufacturers have developed new stacking architectures in the era of 3D NAND flash memory, as shown in Fig. 2.

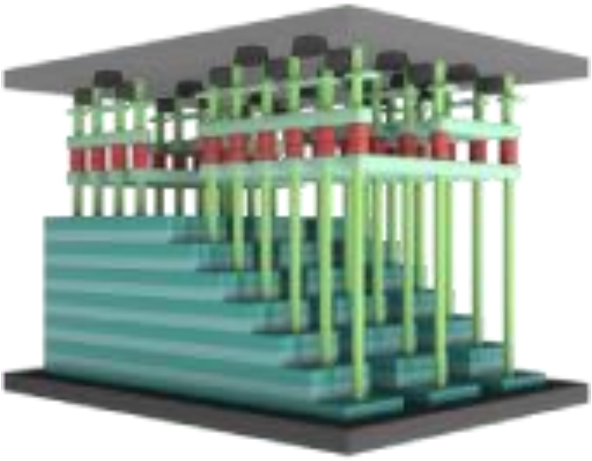

Fig. 2 3D NAND Flash (Xtacking Architecture).
https://www.ymtc.com/cn/technicalintroduction.html

The organizational structure of a NAND flash chip, from largest to smallest, is Chip > Wafer (DIE or LUN) > Plane > Block. The block is an independent unit of the above organisational structure, and is also the object of erase operations, and the stacking of 3D NAND flash memory is also manufactured on a block basis Fig. 3. A block contains several Word Lines (WL) as shown in the Fig. 3. A Word Line is a physical page, which consists of a number of storage units, and all the storage cells in a Word Line constitute a logical Page, which is usually the smallest unit of a flash read operation. The control gates of all memory cells in a word line are connected in parallel, while the source and source poles of flash memory cells with the same displacement in different word lines within a block are connected in series through a Bit Line.

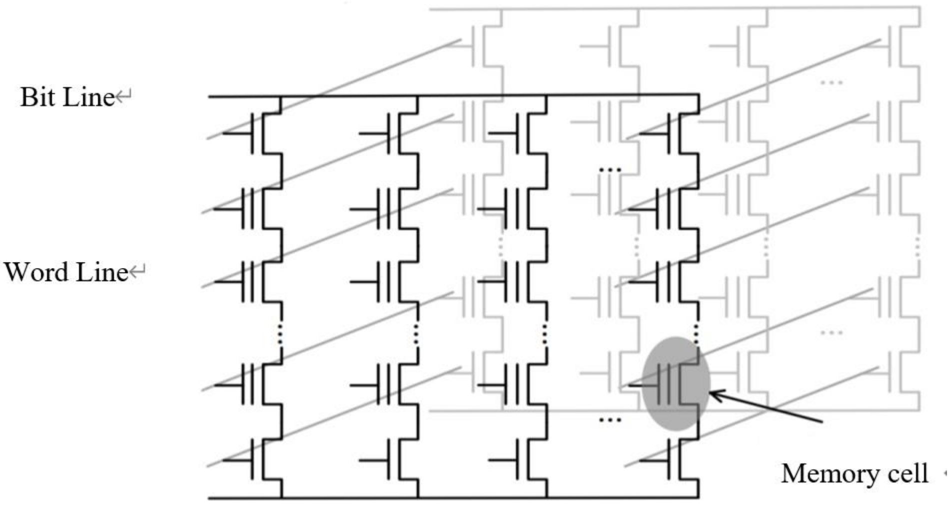

Fig. 3 3D NAND Flash Block Structure. (Photo credit: Original)

4.2. Principle

Solid State Drives use NAND-type Flash particles as the storage medium, and the data reading or writing process is carried out by the control IC (master chip), in which the flash memory storage cell (Cell) is a transistor similar to the structure of MOSFET, which is the smallest structure of the NAND flash memory to store the data, through the preservation of the electrons and quantisation of the electrons (voltage) to store and represent the data, and the flash memory storage cell is currently mainly of the Floating Grid (FG) Fig. 4.

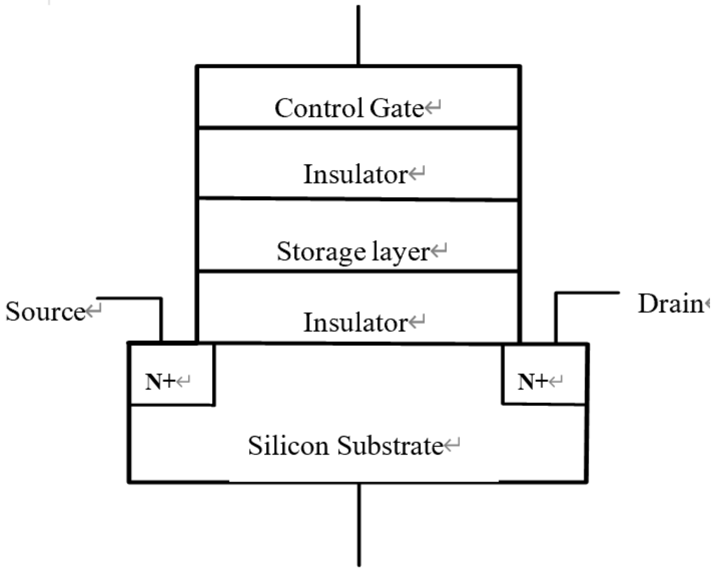

Fig. 4 Floating Grid Memory Cells. (Photo credit:Original)

It has three basic operations: erase, write, and read, and must follow the rule of erase before reading. The erase operation applies a high voltage to the substrate, sucking out all the electrons held within the floating gate or memory layer. When a write operation is performed, the control gate (cg) gradually applies a voltage to ensure that as many electrons as the expected amount of data to be written enter the floating gate or storage layer. If an erase operation is not performed prior to programming, there is no guarantee that the sum of the number of electrons written at a time and the number of electrons originally stored within the floating gate or storage layer matches the expectation of the data being written. Reading data requires a read reference voltage to be applied between the source and drain of the flash memory cell to create a potential difference between the source and drain. For SLC memory cells, if there are electrons held within the floating gate or Storage layer, electron tunnelling occurs under the potential difference to form a current, which is conducted between the source and drain. If there are no electrons in the floating gate or the storage layer, it cannot conduct, and by determining the conductive state, it can be assigned a value, which indicates that the storage bit is 0 or 1 and stored in binary form.

5. Applications of 3D NAND

Since its development, 3D NAND has been used for a variety of purposes. First, 3D NAND is widely used in enterprise solid state drives (SSDs) for data centers and servers. These devices require large amounts of storage space and high-speed data access, and 3D NAND can fulfill these needs. Second, hyper-converged infrastructure (HCI): 3D NAND can provide flash-speed mass storage, which reduces the cost of HCI and allows users to increase performance and scale at a reasonable price [6]. Finally, 3D NAND can provide high-capacity environments: 3D NAND SSDs of 15TB and above are already on the market, and 2U devices are being developed that can provide up to 6PB of storage.

In addition to this, 3D NAND offers significant improvements in all aspects of performance, storage density, scalability, and reliability, making it ideal for the development of next-generation smartphone applications such as holographic displays, artificial intelligence, drone simulation, high-end gaming, thermal imaging, real-time translation, 360-degree video, etc., all of which 3D NAND flash plays a large role in [7]. This means devices with higher data capacity, faster speeds, higher reliability, and greater energy efficiency to be able to transfer files faster, store more photos, analyze data in real time, use powerful apps without delays, and perform many other tasks.

And in recent years, when 3D NAND technology has been maturing, several Chinese semiconductor companies have also made amazing developments in related industries. 2023, Changjiang Storage announced that it could mass produce 232-layer 3D NAND flash memory, which, through mass production, could drive the rapid development of the high-speed solution market [8]. Fig 5. is the Changjiang Storage Xtacling.

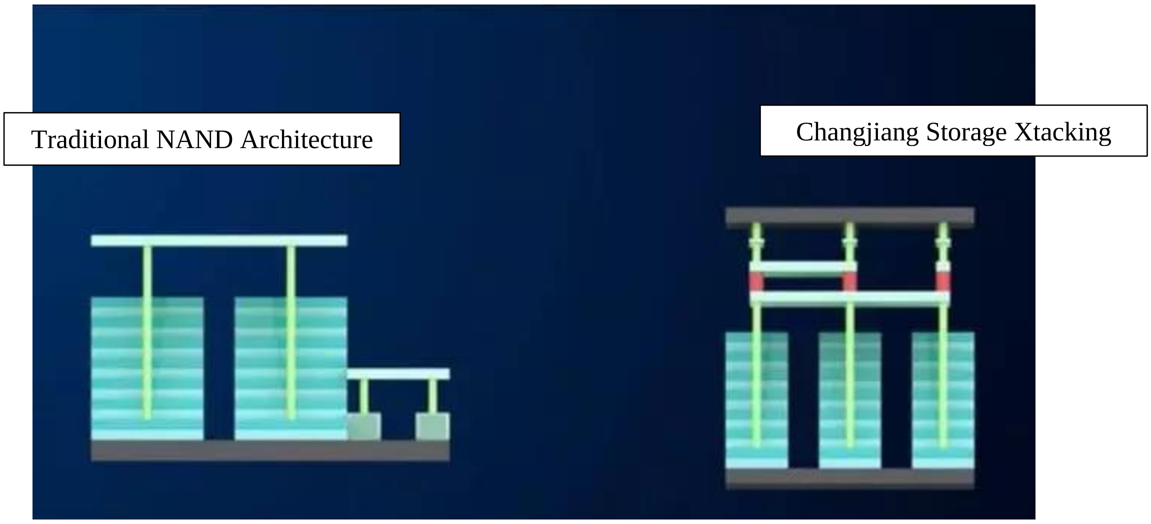

Fig. 5 Changjiang Storage Xtacling. https://m.163.com/dy/article/HP2FD83O0553445F.html

6. Future Prospects

Unlike the expansion practices in 2D NAND technology, the direct way to reduce cost and increase chip density in 3D NAND is to increase the number of layers, so wanting to continue to raise the upper limit of 3D NAND in the future is closely related to increasing the number of layers in it. 2023 In June, Tokyo Electron announced that it had developed a through-hole etching technology for memory chips that can be used to make more than 400-layer stacked 3D NAND flash chips, which is the highest number of stacked layers to date.

While 3D NAND is already incredibly advanced, however, several semiconductor companies today have begun working on 4D NAND, which further enhances the peripheral circuitry under the CTF NAND arrays, thus saving more space on the chip and further reducing production costs [9]. In June 2023, South Korean semiconductor company SK Hynix announced that it had begun mass production of 238-layer 4D NAND flash memory and is in the process of product validation. It is the highest number of stacked layers of mass-produced NAND flash memory [10].

7. Conclusion

With the continuous development of data storage, data storage for large capacity, high performance, high reliability needs more and more, as a data storage carrier, NAND flash memory is also more and more attention. This paper finds that there are a series of differences between flash memory and traditional memory, such as different erasure methods and different units for each erasure. In order to enhance the storage space of planar flash memory to shrink the components, when shrinking to a certain degree flash memory will occur uncontrollable hidden danger, so invented the three-dimensional memory. 3D NAND flash solid state hard disk thanks to the three-dimensional stacking, multi-order storage unit, and space compression and other technologies, has a high-speed read-write, high-capacity, low-power consumption, etc., and gradually replace the traditional equipment to become the mainstream of the storage device. In 3D NAND memory, the storage density per unit area is greatly increased by vertically stacking many layers of memory cells. To further increase the storage density, multiple tiers can be vertically stacked, in which many vertically stacked memory cells exist in each tier. In order to efficiently read, write, and erase in a 3D NAND flash memory having a plurality of tiers, each tier may be processed as a separate block of memory, i.e., each tier may be erased independently of the other tiers. Finally, 3D NAND is widely used in enterprise SSDs, hyper-converged infrastructures, etc., which also still has a lot of room for development and application prospects. And with the further development of the technology, the 3D NAND flash memory capacity will still keep growing tens of times in the next decade, and its structure will become more and more complex. In this paper, we make a systematic summary of the existing research on 3D NAND, which is meaningful for improving storage technology.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Wan G.C., Su L.F., Luo S.Q., Chen Y., System-on-Chip Design Methodology and Practice. Tongji University Press, Shanghai. 2016.

[2] Fang Z.J., Introduction to Computers 3rd edition. China Railway Press, Beijing. 2017.

[3] William R.E.R., Medical imaging physics. Zhongnan University Press, Changsha. 2014.

[4] Yang W.J., Characterization of NAND flash memory. Shandong University, Jinan. 2019.

[5] Geun H.L., Sungmin H., Junsu Y., Hyungjin K., Architecture and Process Integration Overview of 3D NAND Flash Technologies. Applied Sciences, 2021, 11(15): 6703.

[6] 3D NAND for flagship smartphones. Electronics Today, 2018, 292(04): 63.

[7] Lin Y.S. Research on the Reliability of 3D NAND under High and Low Temperature. Huazhong University of Science and Technology, 2021.

[8] Mathematical Sciences Excellence, Layer 232 Breakthrough! CK Storage Breaks Records and Pushes China's Flash Memory Technology Forward, 2023, https://baijiahao.baidu.com/s?id=1770453081541641160&wfr=spider&for=pc

[9] Jia M.H., Reliability research and system optimization of 3D NAND flash memory. Shandong University, 2021.

[10] PC Online, SK Hynix mass-produces 238-layer 4D NAND flash memory with 50% rate increase, 2023, https://diy.pconline.com.cn/1623/16237492.html