
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2021 SCPE. Volume 22, Issues 3, pp. 365–385, DOI 10.12694:/scpe.v22i3.1904

REVIEW OF RESEARCH ON STORAGE DEVELOPMENT

HONGHONG ZHANG*AND GUOGUO ZHANG†

Abstract. The development of computer external storage has undergone the continuous change of perforated cassettes, tapes, floppy disks, hard disks, optical disks and flash disks. Internal memory has gone through the development of drum storage, Williams tube, mercury delay line, and magnetic core storage, until the emergence of semiconductor memory. Later RAM and ROM were born. RAM was divided into DRAM and SRAM. Due to its structure and cost advantages, DRAM has gradually developed into the widely used DDR series. At the same time, the low-power LPDDR series has also been advancing. At present, with the development of NVRAM technology, non-volatile random access memory with both internal and external storage functions is born. Dual-space storage based on NVRAM combines internal and external storage into one, and large capacity dual-space storage has become the development trend of storage.

Key words: external storage, internal memory, hard disk, DDR, dual-space storage

AMS subject classifications. 68M99

1. Introduction. The storage is an important part of the computer system, and its function is to store programs and data. The world's first computer "ENIAC" was born in the University of Pennsylvania in 1946. In 1945, von Neumann's "stored program" design idea played a key role in the invention of computers. In fact, the appearance of storage predates the birth of computers. The punch card invented by the Frenchman Joseph Jacquard in 1801 was the earliest storage [1]. From punch cards to mercury storage, to magnetic core storage, then to hard disks, optical disks, U disks, RAM and ROM, until now NVRAM, storage continues to advance with the development of computers. This article details the complete process of storage development. Based on the latest non-volatile random access memory, a large-capacity dual-space storage architecture is designed and constructed. The main contributions of this article are as follows:

1. Study the complete process of storage development;
2. The future development trend of storage is predicted;
3. Designed a large-capacity dual-space storage structure.

2. Related research. Guo Jianqing of the Technology Department of China Huajing Electronics Group Co., Ltd. once reviewed the technological level and development status of domestic and foreign storage in 1993, and predicted the future development trend of storage [2]. Increasing areal density is the development direction of hard disks, and there is still room for development of tape storage technology [3]. Jiang Xinghua from the Department of Audio-visual Education of Hebei Normal University introduced the development of semiconductor integrated circuit storage applications in computing in 1999 [4]. He Cang of CST in the United States and Huang Xinyu and Li Peng of Dongfang Integrated Electronic Commerce Network introduced the development trend of storage for different types of equipment in 2000 [5]. Compared with DRAM, Flash Memory is easier to shrink in size and easier to manufacture. Therefore, its cost reduction is relatively faster [6]. The use of storage is becoming diversified, with high speed, large capacity and low power consumption, and the demand for storage is greatly increasing [7]. Tang Haiyan pointed out in 2005 that low power consumption and high integration are the development trend of mobile phone storage [8].

Liu Ming from the Institute of Microelectronics of the Chinese Academy of Sciences discussed the development status and challenges of non-volatile memory in 2012, and introduced the research work of emulation,

*Shanghai University, Shanghai, 200444, China; Henan University of Animal Husbandry and Economy, Zhengzhou, 450044, China(zhhh7921@126.com).

†Shanghai University, Shanghai, 200444, China.‡

reliability, materials, devices and integration of emerging memories [9]. Magnetic random access memory (MRAM) is a potential replacement product for the memory of most digital products such as mobile phones, mobile devices, notebook computers, and PCs, and is the new star of storage [10]. Resistive random access memory is a special application of memristors in binary conditions. Because of its simple structure, high density, high speed, low power consumption, compatibility with CMOS technology, and three-dimensional integration capabilities, it has become one of the next generation non-volatile memory technologies with the most potential for development [11]. The ferroelectric random access memory (FeRAM) has the main characteristics of many erasing times, low operating voltage, low power consumption, fast reading and writing, etc. It also has radiation resistance. This intrinsic radiation resistance is suitable for national defense, military, aerospace, satellite communications and other fields [12]. The phase change memory and storage materials developed by Song Zhitang's team have reached the international advanced level in comprehensive performance indicators, and are said to herald the arrival of the next generation of storage [13].

Many simple physical devices can be used to store information, the best storage to use is the acoustic delay line and drum [14]. A small ring-shaped ferromagnetic core with appropriate "rectangular" characteristics can be used as a storage device, and the storage unit is selected at the intersection of two or three spatial coordinates, and then assembled into a two- or three-dimensional storage system [15]. Among many different types of memory such as punched cards or tapes, films, magnetic tapes and drums, the drum has an outstanding feature, which is suitable as an internal memory for machines that are not too fast [16] . The analog-to-digital conversion controller and MOS shift register buffer form a unit, although it is customized for a special computer, it can run on different systems [17]. The IBM 650 calculator uses a magnetic drum storage program for control, and punch cards for input and output, gaining the flexibility of computers required in the field of commercial and scientific computing [18]. A perfect memory system can immediately provide any data requested by the CPU. However, this kind of ideal memory cannot be actually realized because the three factors of memory capacity, speed and cost are directly opposed [19]. Hard disk drives are an important bottleneck of system performance. Intel Turbo memory solves these problems by adding a new layer, non-volatile disk cache based platform, to the storage hierarchy [20]. Researchers are exploring the use of several emerging storage technologies, such as embedded DRAM, spin transfer torque RAM, resistive RAM, phase change RAM, and domain wall memory [21]. Compared with traditional DRAM and flash memory, phase change memory (PCM) is becoming more and more popular in next-generation systems [22].

At present, with the continuous development of non-volatile random access memory (NVRAM) and its technology, various NVRAMs have emerged. Because NVRAM has both the random access of memory and the permanent preservation of external storage, Jin Yi et al. proposed a dual-space storage based on NVRAM technology [23], which combines the current internal memory and external storage into one and integrates them into a storage body on NVRAM.

3. The development of external storage. The earliest external storage was punched cards (Fig. 3.1), but the invention of punched cards was much earlier than the birth of computers. In the 18th century, French weavers invented tandem punched cards when printing patterns. Joseph Jacquard invented the Jacquard Loom in 1801, which used punched cards to automatically control the pattern of weaving. American statistician Herman Hollerith used punched cards to store population data, and then used a tabulating machine to perceive the cards, helping the United States quickly complete the census. The tabulation machine company founded by Herman Hollerith later became part of IBM [24]. The non-perforated and perforated punched cards are precisely the expression of binary information 0 and 1. Jacquard and Herman used punched cards for automatic control and data processing respectively, which can be regarded as the prototype of computer software. In 1928, IBM used punched cards of 12×80 array to store information in its tabulating machine system. In 1935, IBM launched a punch-card computer, the 601 model, which could calculate multiplication in one second. The world's first programming language, FORTRAN, used punched cards to represent programs (Fig. 3.2). Until 1970, punched cards were used still as storage media in computer equipment. For computer systems, punched cards seemed to be chaos beginning. It can be used as an input device or an output device for a computer. The calculation result is printed on the punched card so this is the attribute of the external storage for it. Furthermore, the program and data stored on the card can also be regarded as the "predecessor" of the internal memory.

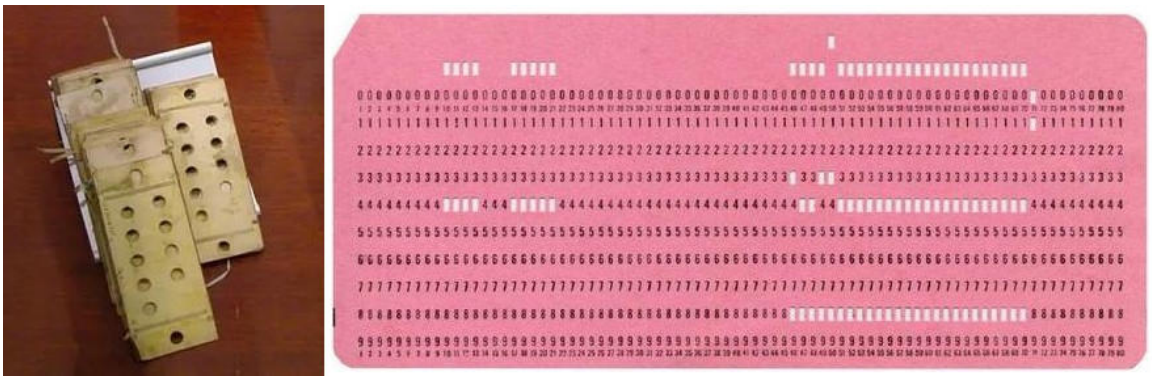

Fig. 3.1: Punched cards.

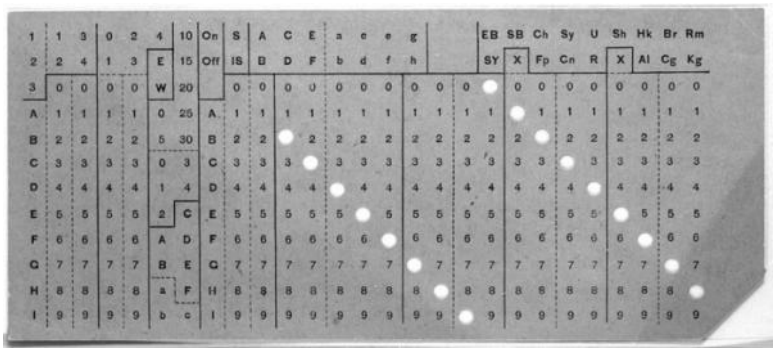

Fig. 3.2: Punched cards of FORTRAN program.

Due to the ever-increasing data and results to be processed, punched cards naturally evolved into punched paper tapes (Fig. 3.3). Punched paper tape was also born in the French textile industry. In the textile industry, it even predates punched cards. The aforementioned series of punched cards are improved on the basis of punched paper tape. In 1846, British physicist Alexander Bain invented a telegraph device[25] that used punched paper tape to send telegrams. In 1857, British scientist Charles Wheatstone used perforated paper tape for data storage and transmission. The principle of storing data in a punched paper tape is the same as that of a punched card. A punched paper tape row can be pierced 8 holes in a row to represent a character. Some systems also use parity check perforation method, using the 8th column whether to be perforated to ensure that the number of holes in a row is even (even parity).

Following punched cards and paper tapes, magnetic tapes had become the most important external storage. The birth of the tape started with the phonograph. In 1877, the American inventor Edison invented the phonograph, which can store sounds and restore them for playback. In 1898, the Danish scientist Valdemar Boll stored the sound on the magnetized piano strings. After repeated adjustments, the steel wire was replaced with a paper tape coated with metal powder. This was the embryonic form of the magnetic tape. In the 1930s, the German companies "Falben" and "Wireless Telecom" improved the metal paper tape, coated the plastic tape with iron oxide, and wrapped the plastic tape around the reel. Thus, the tape (Fig. 3.4) was officially born. In 1951, UNIVAC (UNIVersal Automatic Computer) used magnetic tape as a data storage device for the first time [26]. Since then, magnetic tape has developed rapidly, and later it has been widely used in the fields of recording and imaging. With the development of disk technology, tape has slowly withdrawn from the consumer field. Because the tape is very long, the access speed is very slow. However, tape also has many advantages. One is that it is cheap. with the same amount of data, the cost of tape storage is one-sixth that of

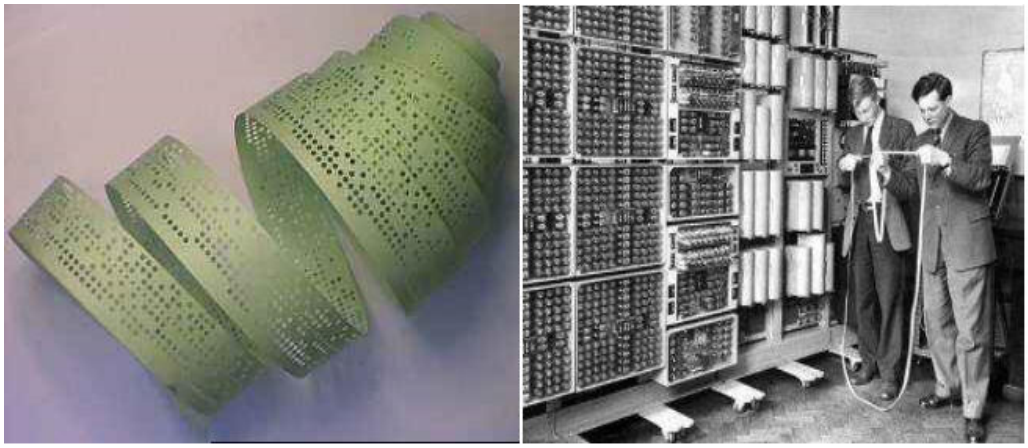

Fig. 3.3: Punched paper tapes.

Fig. 3.4: Tape.

hard disk storage. The other is security. The off-line characteristics of tape make it very defensive. Moreover, with the development of magnetic tape storage technology, the capacity of magnetic tape continues to increase, and the cost continues to decrease, so magnetic tape still has a strong vitality.

The operating instructions of the IBM System370 computer were stored in the semiconductor memory and were lost once shut down. Therefore, it was necessary to develop a cheap and portable device that could store and transmit the operating code of the computer. In 1967, the storage team of IBM's SanJose Lab started the road to their development of floppy disks (Fig. 3.5). Four years later, in 1971, Alan Shugart (who later left IBM to found Seagate) invented a plastic disk coated with oxide on the surface. This was the world's first floppy disk launched by IBM. The floppy disk [27] had a diameter of 8 inches and a capacity of 79.7KB. It was read-only. A readable-writable floppy disk appeared a year later. In fact, as early as 1952, Dr. Yoshiro Nakamura (Japan), the world's great inventor, who enjoyed the reputation of contemporary Edison, invented the floppy disk, but this was not recognized by IBM. In 1976, IBM introduced a 5.25-inch floppy disk (often called a 5-inch disk) with a capacity of 180KB (single-sided low-density) and 360KB (double-sided low-density). Later, the capacity was increased to 1.2MB (double-sided high-density). Although the volume was smaller than the 8-inch disc, it was still not convenient to carry, and the packaging was fragile and easy to break. In 1980, Sony first introduced 3.5-inch floppy disks (commonly called 3-inch disks) and floppy drives (Fig. 3.6), which were smaller in size and larger in capacity, reaching 1.44MB. It has 80 tracks, each track has 18 sectors, 512 bytes are stored in each

Fig. 3.5: Floppy disks of different types.

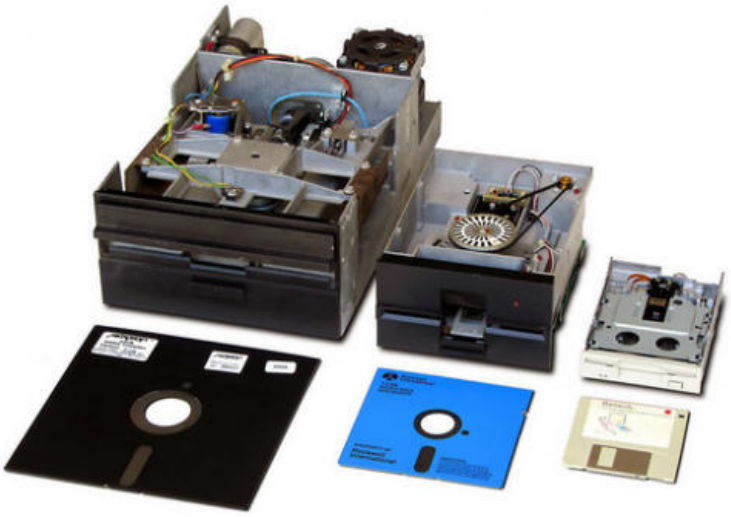

Fig. 3.6: Floppy drives of different types.

sector, and both sides can be stored. 80×18×2×512B=1440×1024B=1440KB 1.44MB. But at the time, 5-inch disks were still popular on the market. It wasn't until 1987, when IBM deployed 3.5-inch floppy drives on its personal computers, that 3-inch disks officially became popular. Because the previous computer programs were very small and the floppy disk was very cheap, it occupied the mobile storage monopoly for 20 years. Until the beginning of this century, many people still prepared floppy disks as system startup disks to install and restore operating systems. Moreover, many college and technical secondary school students used 3-inch disks to submit various computer assignments. In 2010, Sony discontinued production of floppy drives, and stopped selling floppy disks in March of the following year. Until then, the floppy disk had completely withdrawn from the storage market.

In 1956, IBM introduced 350 RAMAC [28] (Random Access Method of Accounting and Control) (Fig. 3.7), which was the world's first hard drive. The hard drive used 50 24-inch platters with a capacity of only 5MB. In 1968, IBM invented the Winchester technology, which sealed the magnetic head, seek mechanism and disk in a closed body to form a Head Disk Assembly (HAD) (Fig. 3.8). This enclosure is not a vacuum, it just isolates the dust from the external environment. This technology uses a lightly buoyant small head slider to implement contact start and stop, that is, the head does not touch the disk surface until it starts and stops. To prevent data from being destroyed, no data is stored in the start-stop area. Relying on ingenious aerodynamic design, when accessing data, the disk rotates at high speed, and the magnetic head flies at a height of 0.2

Fig. 3.7: The first hard drive.

Fig. 3.8: Hard disk.

to 0.5 m on the disk surface. Moreover, the surface of the disk is coated with lubricant, which can better protect the head and disk. This is the prototype of modern hard drives. In 1979, IBM introduced a thin film induction magnetic head, which reduced the weight of the magnetic head, thereby speeding up the access speed and increasing the storage density. In 1980, the 3380 hard drive manufactured by IBM had a total capacity of 2.52GB, but it was twice the size of a refrigerator and weighed 250KG. By the end of the 1980s, IBM invented the Magneto Resistive (MR) head, which was another major contribution to storage. Because magnetoresistive heads are very sensitive to data changes, the storage density of magnetic disks has been increased by tens of times compared with the past. In 1991, IBM applied magnetoresistive heads to its 3.5-inch hard drive, and the hard drive capacity exceeded 1GB, and the hard drive entered the GB era. In 1997, IBM introduced Giant Magneto Resistive (GMR) heads, which have higher sensitivity than MR heads, thus further increase the storage density. In 2003, IBM sold its hard-disk division to Hitachi of Japan, ending its glorious history in the disk field. Hitachi thus established Hitachi Global Storage Technologies (Hitachi GST, HGST for short).

The main part of a hard disk is a circular disk and a magnetic head. Each disk has a magnetic head on each side. Each disk has the same size and is divided into the same number of circular tracks, from the outside to the inside numbered from 0 , and the tracks with the same number form a cylinder. For example, the hard drive with the model number ”HGST HTS545050A7E680” is a Hitachi hard drive. The parameters

are: 969021 cylinder, 16 heads, 63 sectors/track, 512 bytes/sector. The number of sectors of the hard disk = head × cylinder × sector = 16 × 969021 × 63 = 976773168, this value is the number of LBA sectors. LBA is logical block addressing. In this mode, parameters such as head, cylinder (track), and sector are not actual physical parameters. In the old hard disk, the number of sectors in each track is the same, which will cause a great waste of outer track storage space. Later, in order to increase the capacity of the hard disk, the number of sectors set for each track is no longer the same, and the outer track will have much more sectors than the inner track. But in order to be compatible with the previous addressing mode, logical block addressing is adopted.

Because hard disks have the advantages of large storage capacity, convenient use, and high cost performance, they are still widely used until now. The parameters of the hard disk include capacity, rotation speed, average seek time, transmission rate, cache and so on. Rotation speed refers to the maximum number of revolutions of the disc in one minute, that is, the rotation speed of the spindle motor, and the unit is Revolutions Per minute (RPM). The common revolutions are 5400 or 7200 revolutions. The larger the revolutions, the faster the hard disk speed. Average Seek Time refers to the average time it takes for the head to move from the starting position to the specified track. The shorter the time, the faster the hard disk speed. Data Transfer Rate refers to the speed at which the hard disk reads and writes data, in megabytes per second (MB/s). The transmission rate is divided into internal transmission rate and external transmission rate. The internal transfer rate refers to the data transfer rate from the head to the hard disk cache, which mainly depends on the hard disk speed. The external transfer rate refers to the data transfer rate between the hard disk cache and the system bus, which mainly depends on the cache size and interface type. The cache is a memory chip on the hard disk controller [29], which has a fast access speed. It is a buffer between the internal storage of the hard disk and the external interface. A large cache can increase the external transmission rate of the hard disk. Interface types include ATA (Advanced Technology Attachment), SATA (Serial ATA), SCSI (Small Computer System Interface) and SAS (Serial Attached SCSI), etc. The ATA interface uses a traditional 40-core cable to connect to the motherboard. It is a parallel interface, which is different from the serial ATA later, also known as PATA (Parallel ATA). Usually known as IDE (Integrated Drive Electronics) hard disk, originally meant the hard disk drive that integrates the controller and the disk body, and also indicates the hard disk interface, commonly known as the PATA interface. SATA is a serial interface developed on the basis of ATA, which overcomes the crosstalk of the PATA interface, and the external transmission rate has reached 150MB/s. SATA also has stronger error correction capabilities and has basically replaced the traditional PATA interface. The SATA interface was later developed into SATA and SATA , and the transmission speed continued to increase, reaching 600MB/s. The SCSI interface is not an interface specifically produced for hard disks, but a high-speed data transmission interface for small computer systems. The SCSI interface is connected by a 50 or 68-core cable, which is a parallel interface, supports multi-tasking, hot-plugging, and has a low CPU occupancy rate. But it is not compatible with ATA and the price is high, so it is mainly used for advanced servers. SAS is a serial SCSI, which improves the transmission speed and is compatible with SATA.

RAID (Redundant Arrays of Independent Disks) is a disk group with huge capacity composed of multiple independent disks. This technology expands the disk capacity, improves data access speed by accessing data in blocks or accessing several disks at the same time, realizes redundant backup of data through mirroring and so on. The concept of RAID was proposed by David Patterson, Garth A. Gibson, and Randy Katz of the University of California at Berkeley in 1987 [30], but the original I was not independent, but inexpensive, which can be seen from their paper "A Case for Redundant Arrays of Inexpensive Disks (RAID)" published in June 1988. In fact, the essence of RAID technology can be traced back to the patent US4092732 in 1978. This is the patent "SYSTEM FOR RECOVERING DATA STORED IN FAILED MEMORY UNIT" invented by Norman Ken Ouchi of IBM. The patent proposes technologies such as disk mirroring and special parity check codes. RAID is divided into multiple levels according to the technology and structure used, including RAID 0, RAID 1, RAID 0+1, RAID 2, RAID 3, RAID 4, RAID 5, RAID 6, RAID 7, RAID 53. RAID 0 mainly realizes data striping access, at least two disks, but no redundancy; RAID 1 mainly realizes disk mirroring, at least two disks, but the utilization rate is 50%; RAID 0+1, as the name implies, is the two technologies are combined, divided into RAID 01 and RAID 10, according to the order of using RAID 0 and RAID 1. It is a highly reliable and efficient disk structure, but at least four disks are required and the cost is high; RAID 2 uses Hamming code calibration; RAID 3 adopts bit-wise parallel transmission technology with parity check code; RAID 4 adopts

Fig. 3.9: CD and DVD.

data block transmission technology with parity check code; RAID 5 uses distributed parity check; RAID 6 uses two kinds of distributed parity check; RAID 7 adopts optimized high-speed data transmission technology; RAID 53 is a combination level like RAID 0+1, however it is not a combination of RAID 5 and RAID 3, but RAID 0 and RAID 3 The combination.

CD (Compact Disc) is what we usually call optical discs (Fig. 3.9), which use optical storage media that are different from previous magnetic media, rely on laser principles to read and write storage devices. In 1972, Philips of the Netherlands successfully developed the use of laser beams to record and reproduce information. In 1978, the Laser Vision Disc (LD) was put on the market [31]. What LD stores is the analog signal of image and sound. In 1982, Philips and Sony jointly formulated the CD-DA (Compact Disc-Digital Audio) standard. CD-DA records the analog signal on the disc after PCM (Pulse Code Modulation) digital processing. The advantage of digital recording is that it is not sensitive to interference and noise. Although the thickness of a CD disc is only 1.2mm, it is a multilayer structure: substrate, recording layer, reflective layer, protective layer and printing layer. Since the laser wavelength is 780nm, the numerical aperture (NA) of the objective lens is 0.45, and the distance for the laser beam to converge to a point needs 1.2mm, so the thickness of the CD disc is 1.2mm, too thick or too thin will affect the data access . The substrate is a circular polycarbonate (Polycarbonate, high molecular polymer containing carbonate base, PC plastic for short)sheet with a diameter of 12cm and a hole in the middle, which is the appearance of an optical disc that is usually seen. Regarding the diameter of the disc, Philips recommends 11.5cm, which can record for 60 minutes, which is suitable for car audio systems on the European market. Sony believes that it should be set to 12cm, recording 74 minutes and 42 seconds. Because Norio Oga, who leads Sony, is a musician, he believes that ”it is incomplete for the recording of Beethoven’s Ninth Symphony to be difficult.” In the end, Sony’s claim was passed. The recording layer is the place where information is recorded. The substrate is coated with special organic dyes. When burning information, the laser burns the organic dyes into ”pits” one after another. Both pits and no pits represent the information ”0”, the edge of the pit Represents information ”1”. Since these pits cannot be recovered, this is a non-rewritable CD. For rewritable optical discs, the substrate is coated with carbonaceous material, which changes the polarity of the carbonaceous material during recording. Since the polarity of the carbon substance can be changed repeatedly, this forms a rewritable optical disc. The reflective layer is the third layer, which is used to reflect the laser beam of the optical drive. The reflected laser beam is used to read the information in the optical disc. The material of this layer is 99.99% pure silver, aluminum or copper. When light reaches this layer, it will be reflected back, making the disc look like a mirror. The protective layer is a light-curing acrylic material to protect the reflective layer and the dye data layer from damage. The printing layer is where the information and capacity of the optical disc are printed, that is the back of the optical disc, and it also protects the optical disc.

VCD is a video compact disc. The difference between it and CD is that it records film and television information, so it is commonly called video disc. There are different standards for the compression of data on optical discs. According to the color of each standard package, it mainly includes the following types: 1) Red Book, a CD-DA disc jointly developed by Philips and Sony for storing audio sound tracks Standard, contains only audio sector tracks. 2) Yellow Book, the CD-ROM data disc standard jointly developed by Philips and Sony, contains only data sectors. 3) Green Book, the CD-I (Compact Disc Interactive) standard developed in 1986.

4) Yellow Book Advanced, the CD-ROM/XA (CD-ROM eXtended Architecture) disc standard supplemented in 1989, can interlace data or audio and video storage. 5) The Orange Book has formulated a standard for rewritable discs. The disc is divided into data area, data area is divided into tracks, and the track is divided into sectors. 6) Blue Book stipulates the CD-Extra in extra mode, the first track is CD-DA music, and the second track is CD-ROM data. 7) White Book, which defines the VCD standard, and the video CD mentioned here is the white paper standard. VCD uses MPEG-1 compression method to compress images. MPEG is the Moving Picture Experts Group, an organization that specializes in formulating international standards in the multimedia field. It includes three parts: MPEG video, MPEG audio, and MPEG system (audio and video synchronization). The audio format of VCD uses 44.1KHz sampling frequency, 16 Bit sampling value and Stereo, before uncompressed, such audio format is CD sound quality. The CD restored by this compression method has the same sound quality as the original CD, and even professionals cannot hear the difference, so this is a lossless compression. MPEG-1 layer 1 and layer 2 are compression processing methods that are specially designed to deal with the audio format of VCD, and layer 3 is the MP3 music format that we know well. MP3 uses 192Kb/s Audio Bit Rate compression. Compared with the original music, there is some slight distortion after restoration, so MP3 is a lossy compression, which is compressed by as much as 10 times. VCD can be played directly on a VCD player, or on a computer equipped with an optical drive. But because VCD has strict format regulations, sometimes the discs burned by oneself cannot be played on the VCD player.

DVD (Fig. 3.9) is a digital versatile disc, with MPEG-2 as the standard, with a capacity of 4.7G, which can store 133 minutes of high-resolution full-motion movie and television programs, and the image and sound quality is beyond the reach of VCD. The surround channel adopts Dolby Digital technology, which is a new generation of home theater surround sound system released by Dolby Laboratories. The digital sound includes front left, center, front right, surround left, and surround right. The signal of 5 channels, plus a single subwoofer effect channel namely 0.1 channel, is called 5.1 channel together. The difference between DVD and VCD is as follows: 1) The color of VCD discs is aluminum white, or slightly light blue, light green, etc.; while DVD discs are very obvious dark purple, which is the color of alumina after treatment. After the data has been burned, the color of the inner circle becomes darker, making it easier to see if there is data on the DVD disc. 2) The capacity of CD/VCD discs is about 600-700MB, while that of DVD is much larger, reaching about 4.3GB. 3) CD/VCD discs use a near-infrared invisible laser with a wavelength of 780nm for reading and writing data, and DVD discs use a red laser with a wavelength of 650nm for reading and writing operations. 4) The image resolution of VCD is only 352×240 (NTSC: National Television Standards Committee in Japan and the United States) or 352×288 (PAL: Phase Alteration Line in China and Europe). The DVD with MPEG2 standard image compression technology has a resolution of up to 720×480. MPEG2 also has dynamic stream control technology, which can flexibly adjust the video reading rate. In 2002, Sony Group led the planning and development of Blu-ray Disc (BD). The English name of BD is Blu-ray Disc, not Blue-ray Disc. This is because the word Blue-ray Disc is too colloquial in Europe and the United States and cannot be approved for trademark application, so the trademark registration is completed by removing the letter e. Blu-ray discs are named because they use a blue laser beam with a wavelength of 405nm for reading and writing operations. Since the blue light has a shorter wavelength, the focal diameter after focusing is smaller, and the use of different reflectivity for multi-layer writing, so the Blu-ray disc has a larger capacity. A single-layer Blu-ray disc has a capacity of 25 or 27GB and can burn up to 4 hours of high-resolution video.

Hard disks and floppy disks are magnetic media storage, optical disks are optical storage, and semiconductor storage is now widely used. Flash memory is a long-life memory that does not lose data (non-volatile) when power is off. It was invented by Dr. Fujio Masuoka when he was working at Toshiba in 1984 [32]. It is named flash because the memory erasing process is reminiscent of a camera's flash. Flash memory technology uses charges stored on a piece of floating polysilicon (floating gate) placed on a gate oxide layer to achieve storage. In 1988, Intel and Toshiba introduced NOR-type and NAND-type flash memory respectively. The NOR type belongs to the internal memory described later, and the NAND type flash memory is described here. So far, NAND flash memory has various forms (Fig. 3.10), including SSD (Solid State Drive), U disk, CF (Compact Flash) card, SD (Secure Digital) card, TF (Trans-Flash) card, memory stick and so on. The naming of the U disk comes from the use of the USB (Universal Serial Bus) interface, which supports hot swap. From 1998 to 2000, many companies claimed that they were the first to invent USB flash drives, including China Netac

Fig. 3.10: Various forms of flash memory.

Fig. 3.11: Different interfaces of USB.

Technology (Netac), Israel M-Systems, Singapore Trek Company. However, it is China Netac [33] that has obtained the basic invention patent of the U disk. Netac has registered the trademark "USB flash drive", so many people also call the USB flash drive "U Disk". The USB 1.0 standard was proposed in 1996, and the speed was only 1.5 Mbps. Two years later, the USB 1.1 speed reached 12 Mbps. The speed of USB 2.0 introduced in 2000 reached 480Mbps, and the maximum speed of USB 3.0 used today can reach 5.0Gbps. The color of the plastic sheet of the U disk interface can help us identify its version, the black one is 2.0, and the blue one is 3.0. The USB interface has only 4 wires, 2 power wires and a pair of differential signal wires, so it is serial transmission. The package of the USB interface (Fig. 3.11) includes Type A (usually USB flash drive interface), Type B (square ladder type port), Mini B type (small square ladder type port), Micro USB type (smaller than Mini B type, the previous Android phone charging interface), Type-C (current Android phone charging interface, regardless of direction) type. The Type-C interface puts an end to the worries of people plugging in the wrong direction. According to the half probability that one billion people in the world make mistakes every day, and each error takes extra 2 seconds, which can save more than 277777 hours, a total of about 31.7 years. SD card was jointly launched by Panasonic, Toshiba and SanDisk in 1999. It has a size of 32mm x 24mm x 2.1mm and supports hot swap. It is widely used in portable devices such as digital cameras. Ordinary computers can use a USB card reader to access the SD card, and some notebooks come with an SD card interface. The memory stick is a memory card developed by Sony. Its early size was 50mm x 21.5mm x 2.8mm, which was similar to chewing gum. Unlike other flash memory cards, the interface standard of the memory stick is non-public. SSD is a solid state drive (Fig. 3.12). Unlike traditional hard drives that require high-speed rotating disks and moving heads, SSD uses microchips and has no moving parts, so it has strong shock resistance, low noise, short read and write and short delay times. And it has the same interface as traditional hard disks, it is easy to replace traditional hard disks, but the current price is still relatively high.

4. The development of internal memory. The earliest internal memory can be traced back to Magnetic Drum [34], which was invented in 1932 by Gustav Tauschek, an Austrian engineer from IBM. After 20 years, the drum memory technology was widely adopted, and it was the first computer memory to be widely used. The

Fig. 3.12: Solid state drive.

magnetic drum (Fig. 4.1) uses a magnetic material coated on the surface of a circular aluminum drum to store data. The drum rotates at high speed during operation. A set of fixed magnetic heads are arranged close to the surface of the cylinder to read and write data. The magnetic drum adopts saturation magnetic recording, from fixed magnetic head to floating magnetic head, and from magnetic glue to electroplated continuous magnetic medium, which laid the foundation for later disk storage. The drum was first used in IBM 650 as internal memory. The drum is 16 inches long, has 40 tracks, can rotate 12,500 revolutions per minute, and can store 10KB of data. The biggest disadvantage of the magnetic drum is that the storage capacity is too small. A large cylinder has only one surface layer for storage. More than ten years after the invention of the drum memory, some other types of memory have also been born. One of them is the Williams tube [35] in 1947, the full name is Williams-Kilburn tube, a kind of storage device composed of cathode ray tube (CRT) . The name of the Williams tube comes from the developers Freddie Williams and Tom Kilburn. Williams tubes store data as charged points on the surface of the cathode ray tube. Since the electron beam of the cathode ray tube can read and write the points on the electron tube in any order, it is random access, which is in line with the characteristics of our current internal memory. The capacity of Williams tubes is only a few hundred to about 1,000 bit. The Manchester Baby [36] computer (also known as Small-Scale Experimental Machine, SSEM for short) that appeared in 1948 successfully ran an electronic storage program for the first time, using a Williams tube. Although the Williams tube was not designed for SSEM, the use of SSEM had verified its reliability.

Another well-known early memory is the Mercury Delay Line [37], its principle is very simple, a stone is thrown into the water to form a wave, and the wave head can spread to a distant place after a period of time. In other words, the pulse is transmitted from one end of the mercury-filled tube to the other, assuming that it takes one second to propagate, the data will be stored in this second. Since no ready-made devices were available at that time, in order to find a suitable memory, researchers explored almost all physical phenomena-electricity, magnetism, light and sound. Finally, because the acoustic impedance of mercury is close to piezoelectric quartz crystal, when the signal is transmitted from the crystal to the medium and then back, the energy loss and echo are minimized. Therefore, although mercury has defects such as weight, cost and toxicity, mercury is still used to make memory. Moreover, in order to make the acoustic impedance as close as possible, the mercury must be kept at a constant temperature, that is, the mercury must be heated to 40 degrees Celsius above room temperature. In March 1951, the first universal automatic computer UNIVAC-1 designed by Mauchly and Eckert (the main designers of ENIAC) used a mercury delay line memory device [38]. The mercury delay line (Fig. 4.2) is a tube with a length of 150cm and a diameter of 10 mm, filled with mercury. There are converters at each end for electrical-acoustic conversion and acoustic-electrical conversion. In this way, the pulse signal enters from one end of the tube and is converted into ultrasonic waves. After 960ms, the ultrasonic waves reach the other end of the tube, and then are converted into electrical signals for output. The mercury delay line is the heaviest internal memory in history, and each mercury tank weighs more than one ton.

Magnetic-core memory was invented by Wang An of Ethnic Chinese in 1948. Magnetic-core memory is usually called core memory [39], so until today, people are used to calling memory as core because of this. A wire is inserted into the ferrite magnetic ring. When currents in different directions flow through the wire, the

Fig. 4.1: Magnetic drum memory.

Fig. 4.2: Mercury delay line memory.

magnetic ring can be magnetized in two different directions, and the information representing ”1” or ”0” is stored in the form of a magnetic field. Each core has wires in two directions that are perpendicular to each other, x and y, and there is also a readout line that traverses diagonally. As shown in Fig. 4.3, x and y are addressed in two different directions.When a certain current flows through the wire, the magnetic core will be magnetized or the direction of magnetization will be changed. The minimum threshold value of the current that can magnetize the magnetic core can be obtained through experiments in advance. When writing data, input a current slightly higher than 50% of the magnetization threshold of the magnetic ring on the x and y coordinate lines corresponding to the magnetic core to be written, so that only the magnetic core corresponding to the x and y coordinates will have current in both lines at the same time. And the threshold will be exceeded after superposition, and the magnetic core will be magnetized or change the direction of magnetization, thus writing one bit of data. At the same time, the current passing through all other magnetic cores is either 0 or 50% of the magnetization threshold, and cannot be magnetized, so no data is written. The process of reading data is more troublesome. A read current slightly greater than 50% of the magnetization threshold is sent in the x and y directions. The direction of the read current is known in advance, so that the core corresponding to the x and y addressing coordinates will have current exceeding the threshold. If its original magnetic field direction is the same as the direction of the magnetic field corresponding to the read current, magnetic state of the magnetic core will not change, and there will be no induced current on the readout line traversing diagonally. It can be seen that the data stored in the magnetic core and the read signal are the same. On the contrary, if its

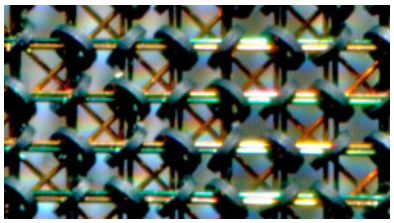

Fig. 4.3: Magnetic-core memory.

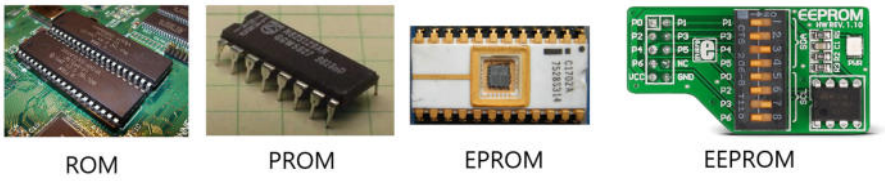

Fig. 4.4: ROM of Various types.

original magnetic field direction is opposite to the direction of the magnetic field corresponding to the read current, magnetic state of the magnetic core will be reversed [40], there will be a huge magnetic flux change, so there will be a large induced current on the readout line traversing diagonally. It can be seen that the data stored in the magnetic core is opposite to the read signal. However, the reading process is destructive. After reading the data, the magnetic core is written with the same read data. The way to recover the data is to write back the original data in the cache. Magnetic core memory is different from ordinary internal memory in two points: first, reading data is slower than writing data, because reading data also includes one-time write recovery; second, data is not lost when power is off, which is similar to the non-volatile random access memory that is being developed today. IBM purchased the magnetic core memory patent with a huge amount of money from Wang An in 1956, in order to solve the problem of data storage in large computers at that time. However, the magnetic core memory also has its shortcomings. The magnetic core is easily damaged, expensive, and slow in operation. In order to solve these problems, IBM has conducted research more than ten years. In 1961, IBM established the Thomas Watson Research Center in New York State, with semiconductors as the direction, and its semiconductor device supplier was Fairchild. Gordon Moore from Fairchild published an article in the "Electronics" magazine in 1965 and predicted: when the price remains the same, the number of transistors that can be accommodated on an integrated circuit will double about every 18 months, and the performance will also be increased doubled [41]. This prediction was later called "Moore's Law".

With the increase in storage density, smaller magnetic cores and wires are required, and the difficulty of manufacturing is also increasing. By the end of the 1960s, when the first generation of semiconductor memory chips came out, the magnetic core memory market began to suffer. Semiconductor memory can be divided into read only memory (ROM) and random access memory (RAM) according to functions. As the name suggests, ROM (Fig. 4.4) can only read the content of the memory. The content stored in it is written in one time by the manufacturer using mask technology and stored permanently. ROM is generally used to store fixed and dedicated programs and data. It is a non-volatile memory. Once the information is written, it will not be lost due to power failure. The content of read-only memory required by different users is different. For ease of use and mass production, programmable read-only memory [42] (PROM: Programmable ROM) has been further developed. PROM is generally programmable once, and each memory unit is all 1 or all 0 when leaving the factory. When the user uses it, use the programming method (electricity or light) to store the required data. For example, bipolar PROM has two structures: one is the fuse blown type, and the other is the PN junction breakdown type. They can only be rewritten once, and once the programming is completed, their content is permanent. In order to overcome the drawbacks of programming only once, EPROM (Erasable Programmable

Read Only Memory) was born. EPROM allows users to write content according to their needs, and can erase and rewrite the written content. It is a ROM that can be rewritten multiple times. Because it can be rewritten, the user can correct the written information and rewrite it after correcting the error. EPROM can use high voltage to program and write data, and the data can be cleared by exposing the circuit to ultraviolet light when erasing. To facilitate exposure, a transparent quartz window is reserved on the package shell, but usually, the window is sealed with an opaque sticker or tape to protect the data. In 1979, Intel introduced the first electrically erasable programmable read only memory [43] (EEPROM, also known as E2PROM), which solved the inconvenience of EPROM operation. The erasing of EEPROM is carried out by electronic signals, without the help of other equipment, and the minimum modification unit is byte, so there is no need to clear all the data and write it. EEPROM is a dual-voltage chip. When writing data, a certain programming voltage must be used, and the content can be easily rewritten by using the dedicated refresh program provided by the manufacturer. BIOS can prevent viruses by virtue of the dual-voltage feature of the EEPROM chip: when refreshing, turn the jumper switch to the "on" position, and apply the programming voltage to the chip; in normal use, turn the jumper switch to the "off" position, prevent viruses from illegally modifying the BIOS chip.

In 1964, the original Fairchild semiconductor company(known as the cradle of Silicon Valley talent) developed a 64-bit metal oxide semiconductor (MOS) static random access memory [44] (SRAM: Static RAM). In 1966, Dr. Robert H. Dennard of 34 years old from the IBM Thomas Watson Research Center, proposed the idea of using MOS transistors to make memory chips. The principle is to use the amount of charge stored in the capacitor to represent whether a binary bit is 1 or 0. Each bit only needs one transistor and one capacitor (1T/1C structure). In June 1968, IBM registered a patent for transistor DRAM (Dynamic RAM) (Patent No. 3387286). The dynamic here is relative to the static state. DRAM needs to be charged and refreshed every certain period of time, otherwise the data will be lost, but the SRAM can save the data without refreshing. SRAM does not need to cooperate with memory refresh, so it has high working efficiency and high speed; but its unit circuit is complicated, the integration level is low, and the price is high. Because IBM was undergoing an antitrust investigation by the US Department of Justice at the time, it delayed the commercialization of DRAM. In August 1968, Fairchild's general manager Bob Noyce and head of R&D department Gordon Moore resigned from Fairchild. They pulled 2.5 million US dollars from venture capitalist Arthur Locke and registered a company, which is now the famous Intel company. The word of "Intel" means wisdom and integrated circuits in English, and the trademark was bought from a hotel for $15,000. At that time, the company had only two employees, Noyce and Moore, and they recruited Andy Grove, a process development expert from Fairchild. Then in 1969, Fairchild's marketing director Jerry Sanders pulled 7 employees to form AMD (Advanced Micro Devices) company. Due to financing difficulties, Sanders pulled $1.55 million in investment from Intel's Noyce. Over the next half century, Intel and AMD have become a pair of intractable competitors. In 1969, Intel Corporation developed the first 256-bit SRAM [45], named the 1101 chip, which was officially launched in 1971. The 1101 chip was the world's first mass-produced MOS internal memory using silicon semiconductor technology. The 1101 chip was later improved into the DRAM chip 1103. 1103 was Intel's first star product, so that all companies at that time had to be compatible with this immature chip. For example, due to a design error in 1103, three kinds of voltages have appeared. For compatibility, other companies can only design 3 voltages. This small step by Intel can be said to be a big step in memory history. At that time, the memory was directly installed on the motherboard's DRAM socket in the form of a DIP (dual in-line package) chip, and the package interface was 30-pin SIPP (Single In-line Pin Package) interface. With a capacity of only 64KB to 256KB, the memory is not easy to expand, but it is sufficient for the processors and programs at the time. Until the emergence of 80286, software and hardware put forward greater demand for memory, so the memory module was born, and the package interface was upgraded to 30pin SIMM [46] (Single In-line Memory Modules). The pin definition is actually the same as SIPP. Golden fingers of both sides transmit the same signal. Although SIMM quickly replaced SIPP, the two forms of memory coexisted for a long time. The first-generation SIMM memory with only 30 pins, a single memory data bus has only 8 bits, a 16-bit data bus processor requires two, and a 32-bit data bus processor requires four, which is costly and frequently fails. The emergence of 72pin SIMM memory solves these problems. The bit width of a single memory is increased to 32 bits, and a 32-bit data bus processor only needs one memory. The early memory frequency is not synchronized with the CPU external frequency. It is asynchronous DRAM, including fast page mode dynamic memory

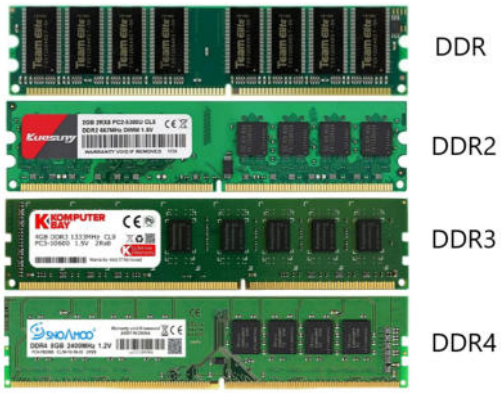

Fig. 4.5: DDRx.

(FPM DRAM: Fast Page Mode DRAM) and extended data output memory [47] (EDO DRAM: Extended data out DRAM). Operating voltage is all 5V. FPM DRAM is improved from the early Page Mode DRAM. When reading the same row of data, it can continuously transmit the column address without the need to transmit the row address. This is much more advanced than that of Page Mode DRAM, which must transmit the row address and column address every time it is accessed. The common capacity of 30pin FPM DRAM is 256KB, and the capacity of 72pin FPM DRAM ranges from 512KB to 2MB. EDO DRAM improves the access mode of FPM DRAM, without waiting for the completion of read and write operations, as long as the specified valid time expires, the next address can be output. EDO DRAM is 72pin, and the capacity of a single EDO memory ranges from 4MB to 16MB. A processor with a 64-bit data bus needs to use two EDO memories.

With the development of processors, memory technology has also undergone changes. The socket has been upgraded from the original SIMM to DIMM (Dual In-line Memory Module). The golden fingers on both sides transmit different data and enter the era of classic synchronous dynamic random access memory (SDRAM: Synchronous DRAM). Synchronization refers to the synchronization of the memory frequency with the CPU FSB, which greatly improves the efficiency of data transmission. The first is the single data rate synchronous dynamic random access memory SDR SDRAM [48] , which was born in 1993. Here, SDR is relative to the later double data rate DDR (Double Data Rate). The DDR X (Fig. 4.5) mentioned in the article is actually the abbreviation of DDR X SDRAM. The interface of the SDR SDRAM memory socket is 168Pin, and the number of pins on one side is 84. It uses a 64-bit data bus. A 64-bit data bus processor only needs a single memory. The frequency of the first generation SDR SDRAM is 66MHz, usually called PC66 memory [49]. Later, as the frequency of the processor increased, PC100 and PC133 memory appeared one after another. SDR SDRAM has a standard operating voltage of 3.3V and a capacity ranging from 16MB to 512MB. In order to further increase the transmission rate, an upgraded version of SDR SDRAM, double data rate synchronous dynamic random access memory (DDR SDRAM) appeared in 1996. SDR SDRAM only transmits data on the rising edge of the clock period, while DDR SDRAM transmits the signal at the rising edge and the falling edge of the clock cycle respective [50], making its data transmission speed twice that of SDR SDRAM, and it will not increase power consumption. So until now the technology used in memory is still DDR, but the version has been upgraded from DDR1 to DDR4, and DDR5 is also being developed. Corresponding to the synchronous dynamic random access memory (SDRAM), the synchronous static random access memory (SSRAM) is also developing synchronously. And the speed of SSRAM is faster, but due to its complex unit structure, the integration is not as high as SDRAM, which brings about a cost increase. This limits its popularity, mainly used in caches and some servers.

DDR SDRAM uses a 184pin DIMM slot. The fool-proof notch is changed from two for SDR SDRAM to one. The operating voltage is 2.5V. The initial DDR memory frequency is 200MHz, and then 266MHz, 333MHz and mainstream 400MHz are slowly born. The capacity ranges from 128MB to 1GB. DDR2/DDR II (Double Data Rate 2) SDRAM is a new generation memory technology standard developed by JEDEC (Joint Electron

Device Engineering Council) in 2003. The biggest difference between DDR2 and DDR is the use of 4bit data prefetching, that is to say, DDR2 memory can access data at 4 times the speed of the external bus per clock cycle. DDR2 uses a 240pin DIMM slot, the working voltage is reduced to 1.8V, which is more energy-efficient than DDR. The mainstream frequency is 800MHz, and the capacity ranges from 256MB to 4GB, mostly with a single 2GB capacity. In 2007, JEDEC developed the DDR3 memory standard. Compared with DDR2, DDR3 has made new specifications in many aspects. The DDR3 core voltage drops to 1.5V, and the data prefetching increases from 4bit to 8bit. The bandwidth of DDR3 at the same core frequency is twice that of DDR2. DDR3 also uses 240pin DIMM slots, the frequency ranges from 1066MHz to 2400Mhz, and the capacity ranges from 512MB to 8GB. In 2011, JEDEC developed the DDR4 memory standard. Compared with DDR3, DDR4 also made some new specifications. The DDR4 operating voltage is further reduced to 1.2V, and the data prefetching still maintains 8bit, because it is too difficult to increase the prefetching to 16bit. However, DDR4 has increased the number of banks, using Bank Group (BG) design, 4 banks as a BG group, you can use 2 to 4 BGs. If 2 BGs are used, the data of each operation is 16bit, and 4 BGs can reach 32bit operations. It can be said that the number of prefetch bits is increased in disguise. DDR4 uses a 288-pin DIMM slot, and the location of the fool-proof notch is also different from that of DDR3. In addition, the gold fingers of DDR4 have a slight curve—the middle is slightly protruding and the edges are shortened, while the previous memory gold fingers are straight, which can ensure that there is enough signal contact area between DDR4 and DIMM slot, so the signal is more stable. And it will be easier to remove the memory module.

LPDDR is called Low Power Double Data Rate SDRAM, which is a type of DDR SDRAM, also called mDDR [51] (Mobile DDR SDRAM). LPDDR is a low-power memory standard drafted by the JEDEC Solid State Technology Association. It is characterized by low power consumption and small size, specially used for mobile electronic products. The JEDEC released the second-generation low-power memory technology LPDDR2 standard in December 2010, and the LPDDR3 standard in May 2012. Its working voltage is 1.2V, which is lower than the 1.5V of DDR3. Compared with LPDDR3, the first-generation LPDDR4 standard released in 2014 has an operating voltage drop of 1.1V, which can be said to be the lowest power solution for mobile devices such as large-screen mobile phones and tablets. Currently, the faster LPDDR5 is also being developed and implemented.

The flash memory mentioned in the external memory above is divided into Nor(or not) Flash and Nand(and not) Flash, the difference is as follows. The first is the interface: address and data bus of Nor Flash are separated; but that of Nand Flash are shared. The second is the reading and writing unit: Nor Flash is read and written by byte; but Nand Flash is read and written by page. The third is the composition structure: the structure of Nor Flash is sectors and bytes; while Nand Flash is divided into blocks and pages, and only supports page-level writing. Even if only one byte is changed, the entire page must be rewritten. The fourth is the erasing unit: the erasing unit of Nor Flash is sector, which takes about 5 seconds, and the maximum number of erasing is 100,000; that of Nor Flash is block, which only requires 4 milliseconds, and the maximum erasing number is 1 million times. The fifth is the application scenario: Nor Flash is mainly used for program storage in the industrial field; Nand Flash is mainly used for large data storage in the consumer field. Nor Flash can be used as internal memory.

In recent years, with the continuous development of non-volatile random access memory (NVRAM Non-Volatile RAM) and its technology, various NVRAMs have emerged. Ferroelectric RAM (also known as FeRAM or FRAM) uses a layer of ferroelectric material to replace the original dielectric, making it also have the function of non-volatile memory [52]. Ferroelectric memory can be traced back to 1952. The master's thesis "Digital Information Storage and Data Exchange of Ferroelectrics" by MIT graduate student Dudley Allen Buck discussed "ferroelectric memory". But it was nearly 40 years later that the Jet Propulsion Laboratory of the National Aeronautics and Space Administration put this technology into practice in 1991. However, the research and development of new FeRAM chips is still the current research topic of major research centers. FeRAM can be divided into destructive readout and non-destructive readout according to the working mode. The destructive readout mode is to use the capacitance effect of the ferroelectric film, replace the conventional capacitor with the ferroelectric film capacitor, and use the polarization reversal of the ferroelectric film to realize the writing and reading of data. Data needs to be rewritten after destructive reading, so FeRAM is accompanied by a large number of rewriting operations in the information reading process. Non-destructive

readout replaces the gate silicon dioxide layer in the MOSFET with a ferroelectric thin film, and modulates the source-drain current through the gate polarization state. The stored information can be read according to the size of the source-drain current value. Because the polarization state of the gate does not have to be reversed, the readout method is non-destructive. However, this approach is still in the research stage.

Magnetic RAM (MRAM) stores data by magnetic field polarization instead of electric charge. The memory cell of MRAM includes a free magnetic layer, a tunnel gate layer and a fixed magnetic layer. The direction of the magnetic field of the fixed layer is unchanged, and the polarization direction of the free magnetic layer is variable. When the direction of the magnetic field of the free layer and the fixed layer are parallel, the memory cell exhibits low resistance; on the contrary, it exhibits high resistance. By detecting the resistance of the memory cell, It can be judged whether the data is 0 or 1. Since the magnetic properties of ferromagnetic materials will not disappear due to power failure, MRAM is non-volatile. Theoretically, ferromagnets are permanently effectual, so the number of writes is also unlimited. Currently commonly used is the current-driven spin transfer torque MRAM (STT-MRAM). This technology originated in 1996. Slonczewski and Berger theoretically predicted a purely electrical magnetic tunnel junction writing method called spin transfer torque [53]. STT-MRAM can achieve a good compromise in terms of speed, area, and power consumption, so it has a wider range of applications. In 2008, Siemens Industrial Automation Division adopted 4Mb MRAM for human-computer interaction interface for industrial control. In 2009, French Airbus decided to replace the original SRAM and FLASH with MRAM in the flight control computer of the A350 XWB aircraft. In 2013, Buffalo Memory announced the use of STT-MRAM instead of traditional DRAM as a high-speed cache in solid-state hard drives (SSDs). In 2018, IBM introduced 256Mb STT-MRAM chips into the latest generation of enterprise SSDs.

Resistive RAM (RRAM) consists of two metal electrodes sandwiching a thin dielectric layer, which serves as an ion transmission and storage medium [54]. The storage medium ion movement and local structural changes caused by external voltages, which in turn cause resistance changes, RRAM uses this resistance difference to store data. Different media materials have different storage effects. The materials selected for RRAM are mostly metal oxides. In addition, sulfides and organic dielectric materials have also received a certain degree of attention. At present, the most accepted is the conductive filament, because it does not depend on the area of the device, and the potential for miniaturization is huge. The memory matrix of RRAM can be divided into passive matrix and active matrix. The memory cell of the passive matrix is connected by a resistive element and a diode. The function of the diode is to prevent the memory cell information from being lost when the resistive element is in a low resistance state. The advantage of the passive matrix is simple design and good miniaturization; the disadvantage is that there is inevitably interference between adjacent units. The active matrix is just the opposite. The transistor controls the reading and writing of the resistive element. The advantage is to isolate the interference of adjacent cells; the disadvantage is that the design is complicated and the miniaturization is poor. In addition, some of the resistive random access memory materials also have multiple resistance states, making it possible for a single memory cell to store multiple bits of data, thereby increasing the storage density.

Phase-change RAM (PRAM or Phase-change memory: PCM) is made of Chalcogenide glass containing one or more chalcogenides [55]. Chalcogenide glass can change its state after heating and become crystalline or amorphous. Different states have different resistance values and can be used to store different values. In fact, as early as 1969, Charles Sie of Iowa State University discussed this memory technology in his doctoral thesis. Intel's co-founder Gordon Moore also published an article in 1970 describing phase change memory. However, due to the complexity of the technology, many research centers are still under development. At present, only Samsung, Micron, and Intel have launched phase change storage products in the world. In May 2017, Intel released Optane storage products based on 3D Xpoint, which is phase change memory. The Song Zhitang research team of the Shanghai Institute of Microsystems, Chinese Academy of Sciences, has been devoted to phase change memory research for 15 years. They have cooperated with SMIC(Semiconductor Manufacturing International Corporation) to achieve a major breakthrough in the research and development of phase change memory at the 130nm technology node. Industrialized sales have been realized through the phase-change memory chip for printers developed in collaboration with Zhuhai Apex Microelectronics Co., Ltd.

The development history of external storage and internal memory is detailed above, and the development

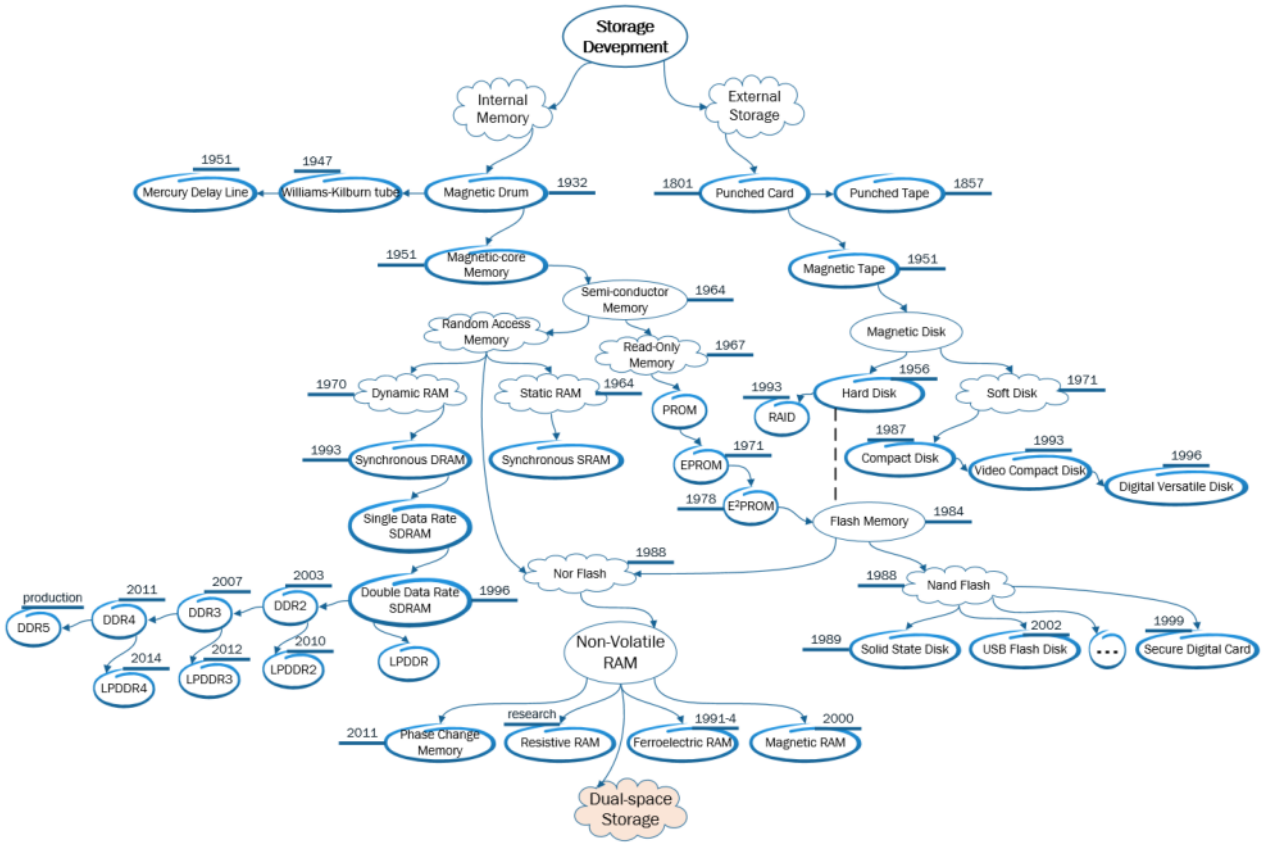

Fig. 4.6: Memory development history.

context is shown in Fig. 4.6.

5. The development trend of memory. It is generally believed that the reason why memory is divided into internal memory and external storage is due to the significant difference between the computing speed of the CPU and the access speed of the memory. In order to match the computing speed of the CPU as much as possible and improve the computing efficiency of the CPU, the memory is divided into internal memory and external storage, and the high-speed memory communicates directly with the CPU, and the low-speed external storage communicates with the memory. The memory speed is fast, but the capacity is small and the data is lost after power failure; the external memory capacity is large, and the data is permanently stored, but the speed is slow. Since computer memory has been divided into internal memory and external storage for a long time, most people think that the existence of internal and external memory is determined by the structure of the computer itself, and the computer should have internal and external memory. This can actually be called a misunderstanding of computer storage. The characteristics of computer memory are non-volatile and random access. Since no storage medium with these two characteristics has been found before, scientists cleverly used non-volatile storage and random access memory respectively.

At present, with the continuous development of non-volatile random access memory (NVRAM:Non-Volatile RAM) and its technology, various NVRAMs have emerged, and the theory of dual-space storage has also been born. This is precisely because NVRAM memory has both random access of internal memory and permanent preservation of external storage. The dual-space storage combines internal memory and external storage in the same storage body, which will change the situation that the computer memory has been divided into two for many years, eliminate the data copy between the internal and external storage, and greatly speed up the

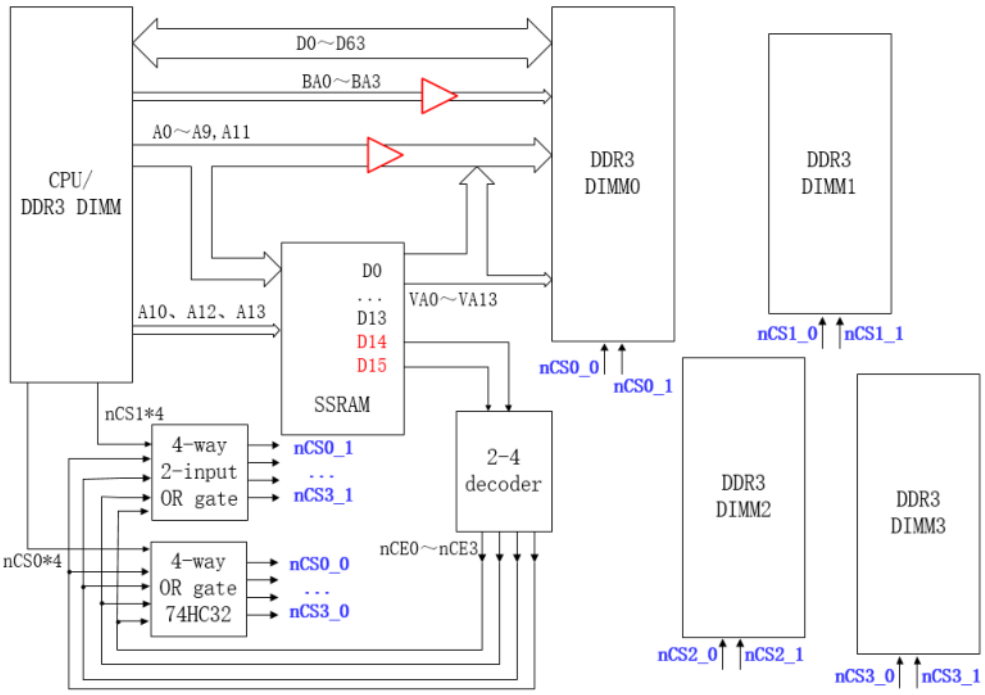

Fig. 5.1: Large-capacity dual-space storage architecture diagram.

operation of the computer. Our team has developed the structure and theory of dual-space storage, and has carried out memory shift experiments to verify it.

The dual-space storage uses NVRAM as the storage body. The basic idea is to use the random access of NVRAM to construct the word space of the dual-space storage to replace the current memory; use the permanent preservation of NVRAM to construct the block space of the dual-space storage to replace the current external storage. So the current internal and external memory is integrated into a storage body, which we call dual-space storage . That is to say, word space and block space do not refer to different spaces, but the space on the same storage body, but their division and usage methods are different. Among them, word space is randomly accessed by word or byte, so it is called word space; block space is accessed by block, so it is called block space. Based on the theory of dual-space storage, we propose a new computer architecture, see another article. Based on the theory of dual-space storage, we designed a large-capacity dual-space storage architecture, as shown in the Fig. 5.1. In this structure, 4 DDR3 DIMM slots are used, and 4 DDR3s are inserted to simulate large-capacity dual-space storage. The capacity of the memory space that the CPU needs to randomly access suddenly increases to the capacity of the external storage space, and the address bus of the CPU itself cannot be increased to the number required for random access to the external storage space. We use the memory space shift technology to solve this problem. The core is to construct a shift latch group to automatically map the memory space that can be directly accessed by the CPU itself to any section of 4 times the dual-space storage word space. The memory data buses from the CPU are directly connected to the corresponding data buses of the dual-space storage; A0∼A9 and A11 of the memory address buses from CPU are directly connected to the corresponding address buses of the dual-space storage, because these address buses are used to transfer column addresses in DDR3. The address buses A0∼A13 are connected to the shift latch group SSRAM (synchronous static random access memory), because these address buses are used to transfer row addresses. Data stored in SSRAM is the row address of the dual-space storage, which can be adjusted in actual operation as needed to carry out the memory shift, thereby realizing random access to the entire dual-space storage. The D0∼D13 output of SSRAM is connected to the row address bus of the dual space storage. D14 and D15 are outputted

as nCE0∼nCE3 through the decoder. They are ORed with nCS0 and nCS1 respectively to obtain nCSx_0 and nCSx_1 (x is 0, 1, 2, 3). Among them, nCS0 and nCS1 are chip select signals of a single DDR3. The nCS04 in the figure represents 4 nCS0 signals respectively ORed with nCE0∼nCE3, and the same is true for nCS14. So, nCSx_0 and nCSx_1 are chip select signals of one dual-space storage body.

The application of large-capacity dual-space storage will historically change the computer architecture and will also change people's traditional concept of computer storage. On the other hand, the current artificial intelligence, big data and cloud computing, etc., all of them require high-performance computing. Large-capacity dual-space storage will lay a solid foundation for high-speed storage. In summary, large-capacity dual-space storage will become the final trend of memory development.

Acknowledgments. We would like to thank Weimin Lian for supporting this research.

REFERENCES

[1] T. Koetsier, *On the prehistory of programmable machines: musical automata, looms, calculators*, Mechanism and Machine Theory, 36(2001), pp. 589–603.
[2] J. Guo, *Memory development status and future trends*, Semiconductor Technology, 1993, pp. 6–10.
[3] M DavidStone and J. Yang, *Mass storage development forecast*, Computer and Peripherals, 1997, pp. 80–81.
[4] X. Jiang, *Overview of the development of semiconductor integrated circuit memory*, Journal of Hebei Normal University, 1999, pp. 3–5.
[5] C. He, X. Huang, and P. Li, LU-*Storage industry development trend*, Electronic Products, 2000, pp. 33–34.
[6] Y. Su, *The world's flash memory is developing rapidly*, Information Times, 2000, pp. 19–21.
[7] Y. Shao and M. Zhao, *Aspects of the development of new memory*, Electronic Engineering & Product World, 2003, pp. 41–43.
[8] H. Tang, *Low power consumption and high integration: the development trend of mobile phone memory*, EDN China, 12(2005), pp. 112.
[9] M. Liu, *Development status and challenges of non-volatile memory*, in 2012 Academic Annual Meeting of China Vacuum Society, Lanzhou, Gansu, China, 20(2012), pp. 1.
[10] L. Yan, *The new star of memory: MRAM*, Software and Integrated Circuit, 2015, pp. 14–16.
[11] S. Liu and Q. liu, *Development status of resistive random access memory*, National Defense Technology, 37 (2016), pp. 4–8.
[12] L. Su, *Preparation and study of flexible or transparent Bi3.25La0.75Ti3O12 ferroelectric memories*, Master, Nanjing University of Science & Technology, 2018.
[13] X. Fei, *Phase change memory, heralding the arrival of the next generation of memory*, East China Science & Technology, 2019, pp. 62–63.
[14] J. W. Forrester, *Digital information storage in three dimensions using magnetic cores*, Journal of Applied Physics, 22(1951), pp. 44–48.
[15] W. N. Papian, *A coincident-current magnetic memory cell for the storage of digital information*: 40 (1952), pp. 475–478.
[16] W. L. V. d. poel, *Dead programmes for a magnetic drum automatic computer* , applied scientific research section b, 3(1954), pp. 190–198.
[17] C. W. Brandon, J. W. Cox, F. W. Keller, and D. A. Brody, *An ADC controller and MOS buffer for the laboratory-oriented computer*, computers and biomedical research, 5(1972), pp. 291–300.
[18] F. E. Hamilton and E. C. Kubie, *The IBM magnetic drum calculator type 650*, ieee annals of the history of computing, 8 (1986), pp. 14–19.
[19] D. Burger, *Memory systems*, acm computing surveys, 28 (1996), pp. 63–65.
[20] J. Matthews, S. Trika, D. Hensgen, R. Coulson, and K. Grimsrud, *Intel® Turbo Memory: Nonvolatile disk caches in the storage hierarchy of mainstream computer systems*, acm transactions on storage, 4 (2008).
[21] S. Mittal, J. S. Vetter, and D. Li, *A survey of architectural approaches for managing embedded DRAM and Non-Volatile on-chip caches*, ieee transactions on parallel and distributed systems, 26 (2015), pp. 1524–1537.
[22] S. M. Seyedzadeh, D. Kline, A. K. Jones, and R. Melhem, *Sustainable disturbance crosstalk mitigation in deeply scaled phase-change memory*, sustainable computing informatics and systems, 28 (2020).
[23] S. OuYang, J. Peng, Y. Jin, Y. Shen, X. Liu, Y. Han, et al., *Structure and theory of dual-space storage for ternary optical computer*, Scientia Sinica(Informationis), 46(2016), pp. 743–762.
[24] H. Zhao, *Introduction to Computer Science*, 3rd, Post & Telecom Press, 2014.
[25] J. Dai, *Research on modem in high-speed facsimile communication system*, Master, Dalian University of Techlology, 2006.
[26] J. Kang, *Research and implementation of read-only RAID and tape library system*, Doctor, Tsinghua University, 2009.
[27] Q. Huang, *The birth of computer disks, the prelude to the information age*,Xinmin Weekly, 2020, pp. 42–43.
[28] C. Liu and C. Xie, *The history, development and future of hard drives-Commemorating the 50th anniversary of the birth of hard drives*, China Mediatech, 2006, pp. 57–61.
[29] Y. Luo, *Research and realization of the implementation of mass storage device in embedded system*, Master, East China Normal University, 2004.
[30] L. Wang, *Disk array storage system and its choice*, Library Work in Colleges and Universities, 2003, pp. 33–34.
[31] K. Chen, *Laser Disc (LD) ushered in a new era of optical disc storage technology*, China Mediatech, 2006, pp. 54–58.

[32] Y. Liao, *Design and implementation of BCH error-correction algorithm based on NAND flash controller*, Master, Harbin Institute of Technology, 2014.
[33] S. Xu, *"Father of U disk" Deng Guoshun: continue the "revolution" (Part 1)*, Guangdong Science & Technology, 2008, pp. 86–91.
[34] S. An, *Ten thousand years of evolutionary history of information storage*, Super Science, 2014, pp. 11–15.
[35] J. Li, *The wonderful equipment in store history*, Computer News, pp. 014.
[36] D. He, *The English-Chinese translation practice report of science and technology academic biography*, Master, Heilongjiang University, 2014.
[37] R. D. Ryan, *A mercury delay line storage unit*, Journal of the British Institution of Radio Engineers, 15(1955).
[38] R. Su, *Chronology of major storage technology developments*, Radio Engineering, 1987, pp. 100–101.
[39] H. Chen, *Introduction to Computer*, 1st, Post & Telecom Press, 2014.
[40] J. Cai, *Principles of magnetoelectronics device application*, Progress in Physics, 2006, pp. 180–227.
[41] G. Gu, *Research on the innovation process of computer technology*, Master, Northeastern University, 2008.
[42] Q. Zhang, *Design and realization of MTM type anti-fuse PROM reading system*, Master, University of Electronic Science and Technology of China, 2017.
[43] H. Ning, *Research on embedding and application design of EEPROM in SoC*, Master, University of Electronic Science and Technology of China, 2009.
[44] R. Chen, *Research on memory failure and reliability test*, Master, Guangdong University of Technology, 2016.
[45] F. Tang and X. Song, *Intel's road to transformation*, Business Management Journal, 2007, pp. 83–87.
[46] S. Lan, *Overview of memory technology evolution*, Personal Computer, 23(2017), pp. 79–83.
[47] H. Wang and R. Song, *Also talk about EDO memory and FPM memory*, China Economy & Informatization, 1997, pp. 54–55.
[48] Y. Yang, *Design and application of SDRAM controller based on FPGA*, Master, Lanzhou University, 2007.
[49] J. Zhao, *The design of memory controller in HDTV chip*, Master, Fudan University, 2010.
[50] X. Wei, *Design and verification of DDR SDRAM controller*, Harbin Institute of Technology, 2009.
[51] X. Wu, *Application and performance analysis of LPDDR in intelligent terminal*, Electronic Design Engineering, 24(2016), pp. 164–166.
[52] C.-H. Chiu, C.-W. Huang, Y.-H. Hsieh, J.-Y. Chen, C.-F. Chang, Y.-H. Chu, et al., *In-situ TEM observation of Multilevel Storage Behavior in low power FeRAM device*, nano energy, 34 (2017), pp. 103–110.
[53] Z. Pajouhi, X. Fong, and K. Roy, *Device/Circuit/Architecture Co-Design of Reliable STT-MRAM*, elements, 2019.
[54] G. Charan, A. Mohanty, X. Du, G. Krishnan, R. V. Joshi, and Y. Cao, *Accurate Inference With Inaccurate RRAM Devices: A Joint Algorithm-Design Solution*, ieee journal on exploratory solid state computational devices and circuits, 6(2020), pp. 27–35.
[55] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, et al., *Phase change memory technology*, journal of vacuum science & technology b nanotechnology and microelectronics materials processing measurement and phenomena, 28(2010), pp. 223–262.

Edited by: Dana Petcu
Received: Jun 26, 2021
Accepted: Sep 30, 2021