Storing the huge mounds of data being generated every second is a challenge that scientists and engineers have been grappling with for the past many years.
The information era we are living in today is generating huge mounds of data. This data is in the form of numbers, text, graphic, audio, video or any hybrid combination of these. Every photograph and video that we shoot on our mobile and share over social networking sites also generates data. CCTV footage from offices or shopping complexes, business documents, official documents – every second we are generating data.

Data has its own life cycle consisting of four phases: Generation, Storage, Updating and Deletion. The time span from data generation to data deletion is known as the data lifetime. The moment data is generated a storage device is required – the two are inseparable.

In the digital world, these storage devices are called memory devices and come in various shapes, sizes and capacities to suit different needs. Today we want a storage device that is small, fast, of high capacity and durability. The lifetime of the memory device must be higher than the lifetime of the data.

All data types such as audio, images, graphics and video are stored in the form of 0s and 1s as binary digits. This binary digit is called bit. A group of 8 bits is called byte, group of 1024 bytes is termed as Kilobyte (KB). Other higher units of storage are Megabyte (MB), Gigabyte (GB), Terabyte (TB) and Petabyte (PB).

Digital data can be best understood by 3V: Velocity, Volume and Variety. Velocity is the speed by which data is being generated. Social networking sites are generating data very fast. Volume is defined by memory space consumed by data. Variety is the format of data such as audio, video or text.

Now, before we take a look at the specific requirements for varied data storage, let’s take a brief look at the evolution of storage devices over the years.

**Semiconductor-based Storage Device**

The semiconductor storage device era can be divided into three parts: pre-semiconductor era, semiconductor era and finally, post-semiconductor era.

Initially, developers tried to store data as a dot in a line and worked on increasing the data density per unit length. Later, a successful attempt was made to store data in two-dimensional areas. This was achieved in the form of concentric circles. In this way, more data capacity and data density became possible. The size and weight of the device also reduced significantly.

The first major development was the introduction of a magnetic tape in 1951 invented by Remington Rand with a capacity of 224 KB only. Today, magnetic tapes are available in 180 TB capacities. Magnetic tape stores the data linearly and so data can be accessed only sequentially. To overcome the sequential access problem, a two-dimensional device named magnetic hard disk was developed in 1956. It had an initial capacity of 3.75 MB with random access and sequential access capability. The hard disk solved the sequential access problem but weight became a big issue, as the hard disk weighed even one tonne. With time, the weight issue was resolved to suit practical purposes. Designers of hard disk were also successful in increasing the capacity and decreasing the physical size. Currently, magnetic hard disks are available in 500 TB capacity.

Another major challenge with magnetic hard disk was portability issue due to its weight and size. This issue was tackled by the development of the floppy disk, presented to the world by IBM in 1971 with a capacity of 80 KB and size of 8 inch. Later, it became available in different size variants such as 5.25 inch and 3.25 inch. The highest known available capacity was 1.44 MB.

One of the problems with floppy disk was that it was unable to meet the growing demand for personal data storage capacity. There was a need for a device that was small in size but high in capacity for personal storage requirement. This led to the development of the Compact Disk (CD), introduced in the market in 1982. It was co-developed by Philips and Sony.

A standard CD can store 74 minutes of audio data equivalent to 783,216,000 bytes. Further, Digital Versatile Disc (DVD) was introduced with higher data density. A DVD has 4.7 GB of capacity, which is almost seven times greater than that of a standard CD.

However, the growth of data and access speed of data posed severe challenges for engineers. The storage devices were consuming more power and space. This problem was solved with the emergence of semiconductor-based memory devices. These devices supported data access speed in the order of nano-seconds, were 10^6 times faster than their predecessors, and also consumed less power and less space.

Over the years, we have seen that the development of
computer systems has been driving the need for different kinds of memory devices. In the ENVIC computer, a vacuum tube was used as the memory device. At that time, the standard size of vacuum tube was approximately two feet and it operated at 200 volts. It could store only 2560 bits. From size and power consumption perspective, it was hard to operate and was not portable. Until the transistor was developed as a memory device.

The size of the transistor was greatly reduced to less than two inches and it operated with only 16 volts. This set the platform for reducing the size of the memory device. This great achievement was recognised by the Nobel committee and the 1956 Nobel Prize in Physics was awarded for research on “semiconductors and their discovery of the transistor effect” jointly to John Bardeen, Walter Houser Brattain and William Bradford Shockley.

But engineers were still using discrete components – diode, resistor, capacitor and transistor – to develop memory devices. In order to establish connection among all components, manual soldering was required. And if more storage space was required it meant thousands of components had to be connected manually. The technique was error prone, less reliable and was taking up more space.

In 1958, Jack Kilby of Texas Instruments developed the first integrated circuit (IC) by fabricating all the components and the chip out of the same block (monolith) of semiconductor material instead of discrete components. The unified circuit realised two benefits: reliability and space-saving. This great work was also recognised by the Nobel committee and Kilby received the Nobel Prize in Physics “for his part in the invention of the integrated circuit”.

Engineers were still looking to improve semiconductor memory in three major perspectives: data density, information retention capability (with and without continuous power) and faster re-writing (re-usability) capability. Further development of semiconductor memory was around information retention capability. Information retention capability can be improved by any one of the two methods. In one method, data can be retained only with the presence of continuous power supply and in another method, data can be retained in the absence of continuous power supply.

The progress in data retention capability with continuous power supply was very rapid and this type of memory was termed as volatile memory (VM). The first product in this series was called Random Access Memory (RAM). Further, two variants of RAM were developed – Static RAM (SRAM) and Dynamic RAM (DRAM). The first successful 64-bit high-speed SRAM was introduced by Intel in 1969 and the world’s first 1K DRAM in 1972. Currently, 8GB RAM is easily available in the market.

The main problem with volatile memory was supply of continuous power. So, engineers came up with the non-volatile memory (NVM). The first successful development was by Sylvania that produced a 256-bit Read Only Memory (ROM) using bipolar TTL technology in 1965 for Honeywell. However, it could be programmed/written only once.

And thus was developed the Erasable Programmed ROM (EPROM), which had re-writing capability. For re-writing, the user needed to first clean/delete memory and then write again. This is known as erasing. A new variant of NVM, EEPROM with electrical signal based erasing technique, came into existence in 1977. The order of erasing time was greatly reduced to seconds from minutes.

There were three possibilities of erasing techniques: bit level, byte level and block level. Block consists of several bytes. Bit level erasing was slower than byte level erasing, which in turn was slower than block level erasing. EEPROM still allowed only byte-level reading, erasure and writing. This meant EEPROM memory had longer erase time which implied longer wait time for re-writing. The need was felt for a faster erase time and this prompted researchers to develop a new technique for erasing data at the block level.

**Semiconductor Memory Flash**

Flash memory is a variant of solid-state non-volatile memory which evolved from EEPROM. Dr. Fujio Masuoka conceived the concept of flash memory and developed the flash while working for Toshiba in 1984. The main advantage of flash memory is data erasing time was reduced to fraction of seconds (in the order of milliseconds) compared to EEPROM, just like a “Flash of Camera”.

Due to these unique fast erasing characteristics, flash memory became dominant in the non-volatile memory segment...
and its utility range was wide from USB to the smartphone, MP3 players, and digital camera. Even data centres for cloud deploy flash-based memory for high performance.

For memory management, the flash memory is divided into multiple blocks, which are further divided into several pages. One block contains multiple pages. One page contains multiple bytes. The smallest unit used by the memory cell to store data is called a bit. The exact size of block and page varies according to the vendor. However, size of one block is in the order of MBs and size of one page is in the order of KBs.

Three operations are performed on flash memory: read, write and erase. Since the erase operation is performed at the block level, the erase time is thousand times (in the order of milliseconds) quicker than its previous variants.

Flash memory has two variants: NAND and NOR. Both variants of flash have their own advantages and disadvantages. The NAND flash memory writing time is faster than that of NOR flash and the NOR flash reading time is faster than that of NAND flash.

NAND flash memory was designed for low cost per bit of stored data. At the same time, it provides very small cell size and high density. That is why it is suitable for storage of large data in USB or PDA. NOR flash memory has been used for keeping code storage for faster read and high random access.

**Diversified Need of Storage Devices**

Data generation speed is doubling every year, so is the requirement for memory space. That is why the storage capacity of storage devices such as pen drives, hard disks and tape drives is also doubling every year.

There are several questions one needs to ask before going in for a specific storage device. What should be the storage capacity of the storage device? How long do you want your data to be active? What technology of storage device would serve the purpose of your data? What is your budget? Where do you want to keep your data physically? These are common questions faced by data storage administrators.

From the engineering point of view, knowing the attributes of a storage device is essential, before purchasing it. The first attribute is memory capacity. How much memory space is required for the user?

The second attribute is data access method. Data can be accessed using sequential or random method. Some storage devices support only sequential access, while others support random access. Sequential access is accessing of data in a linear fashion. Access time for nearer data is short while access time for farther data is more. Random access is where accessing of data from any location takes the same time. Tape-based storage devices support sequential access and flash-based devices support random access.

The third attribute is performance. This attribute is very important. The performance of a storage device is measured in reading and writing speed of the device. This performance is expressed in input-output per second (IOPS) or throughput. IOPS of any storage device for read-and-write operations would be different. Tape drive is available with 1000 IOPS. Solid State Device (SSD) based disk have maximum IOPS in the range of 5000 IOPS. Higher IOPS for any device is desirable.

Cloud is another data storage solution. It is an Internet-based data storage solution where data is stored at a remote location. Clouds also deploy flash-based data storage devices at the back-end for faster access. In cloud-based storage the data is maintained, managed and backed up over a network (typically the Internet) by a cloud service provider. In this service model, users only pay for what they use. Major players in this market are Amazon Web Services, Google and Microsoft Azure.

The existing storage devices – magnetic or semiconductor – can store data only in two-dimensional surfaces. A future research direction is to explore storage of data in 3D space to fulfil the growing data storage demand. This offers the potential for higher capacity in a smaller physical space.

In a biological system, the DNA itself is a data-storage device present in the human system. The DNA has capacity to store enormous data in a very small 3D space. Scientists and engineers are exploring DNA-based bio-storage devices. However, this research is still at a nascent stage.

Vikas Tripathi has been working as a storage consultant in Tata Consultancy Services (TCS) for more than 10 years. He has provided solutions in NAS (Network Attached Storage) and SAN (Storage Area Network) area of storage. Address: Gomati Nagar-226010, Lucknow, Uttar Pradesh; Email: Vikki.tripathi@gmail.com