PART5: Internal Memory

5.1 Semiconductor Main Memory

In earlier computers, the most common form of random access storage for computer main memory employed an array of doughnut-shaped ferromagnetic loops referred to as cores. Hence, main memory was often referred to as core, a term that persists to this day.Today, the use of semiconductor chips for main memory is almost universal. Key aspects of this technology are explored in this section.

Organization

The basic element of a semiconductor memory is the memory cell. Although a variety of electronic technologies are used, all semiconductor memory cells share certain properties:

- They exhibit two stable (or semistable) states, which can be used to represent binary 1 and 0.
- They are capable of being written into (at least once), to set the state.
- They are capable of being read to sense the state.

Figure 5.1 depicts the operation of a memory cell. Most commonly, the cell has three functional terminals capable of carrying an electrical signal. The select terminal, as the name suggests, selects a memory cell for a read or write operation. The control terminal indicates read or write. For writing, the other terminal provides an electrical signal that sets the state of the cell to 1 or 0. For reading, that terminal is used for output of the cell's state.

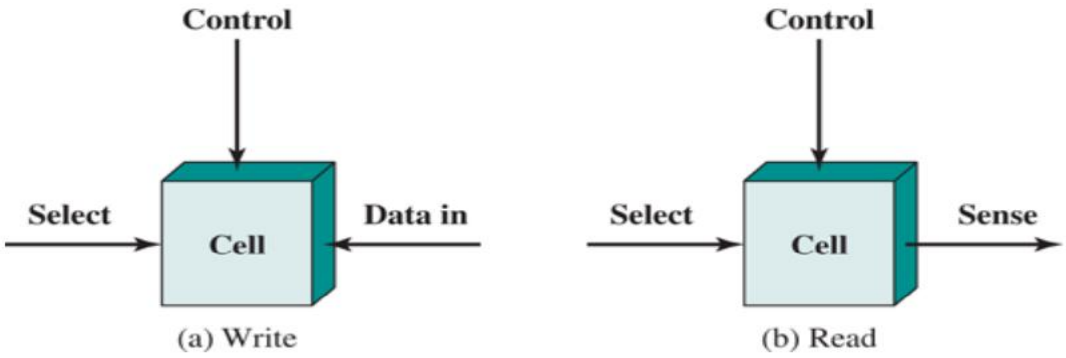


Figure 5.1 Memory Cell Operation

5.1.1 DRAM and SRAM

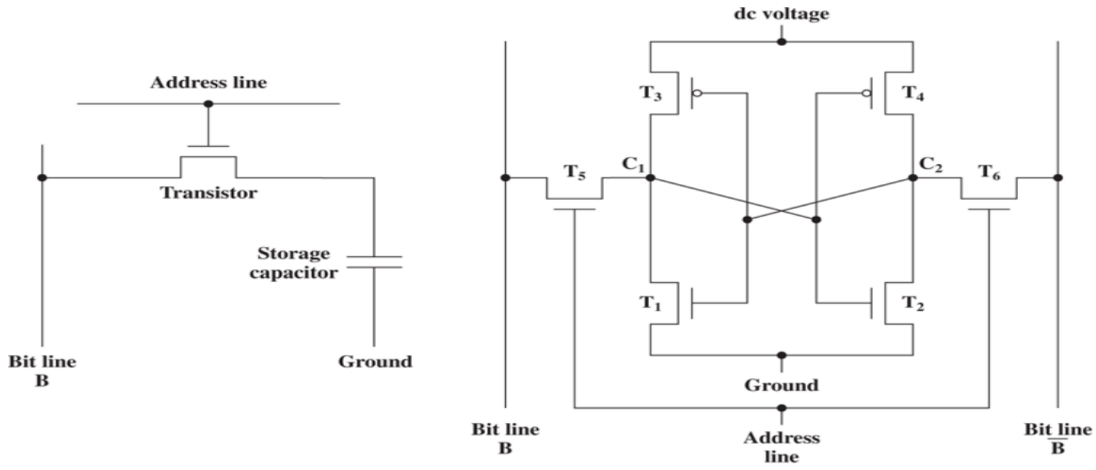
All of the memory types that we will explore in this chapter are random access. That is, individual words of memory are directly accessed through wired-in addressing logic.

DYNAMIC RAM : A dynamic RAM (DRAM) is made with cells that store data as charge on capacitors. The presence or absence of charge in a capacitor is interpreted as a binary 1 or 0. Because capacitors have a natural tendency to discharge, dynamic RAMs require periodic charge refreshing to maintain data storage. The term dynamic refers to this tendency of the stored charge to leak away, even with power continuously applied.See figure 5.2a

- For the write operation, a voltage signal is applied to the bit line; a high voltage represents 1, and a low voltage represents 0. A signal is then applied to the address line, allowing a charge to be transferred to the capacitor.
- For the read operation, when the address line is selected, the transistor turns on and the charge stored on the capacitor is fed out onto a bit line and to a sense amplifier. The sense amplifier compares the capacitor voltage to a reference value and determines if the cell contains a logic 1 or a logic 0. The readout from the cell discharges the capacitor, which must be restored to complete the operation.

STATIC RAM In contrast, a static RAM (SRAM) is a digital device that uses the same logic elements used in the processor. In a SRAM, binary values are stored using traditional flip-flop logic-gate configurations (see Chapter 12 for a description of flip-flops). A static RAM will hold its data as long as power is supplied to it.
Figure 5.2b is a typical SRAM structure for an individual cell. Four transistors (T1 , T2 , T3 , T4) are cross connected in an arrangement that produces a stable logic state.

- In logic state 1, point C1 is high and point C2 is low; in this state,T1 and T4 are off and T2 and T3 are on.
- In logic state 0, C1 point is low and C2 point is high; in this state, T1 and T4 are on and T2 and T3 are off. Both states are stable as long as the direct current (dc) voltage is applied. Unlike the DRAM, no refresh is needed to retain data



(a) Dynamic RAM (DRAM) cell (b) Static RAM (SRAM) cell

Figure 5.2 Typical Memory Cell Structures

5.1.2 Types of ROM

A read-only memory (ROM) contains a permanent pattern of data that cannot be changed. A ROM is nonvolatile; that is, no power source is required to maintain the bit values in memory. While it is possible to read a ROM, it is not possible to write new data into it. A ROM is created like any other integrated circuit chip,and can classify to:

PROM: When only a small number of ROMs with a particular memory content is needed, a less expensive alternative is the programmable ROM (PROM) . PROM is nonvolatile and may be written into only once. For the PROM, the writing process is performed electrically and may be performed by a supplier or customer at a time later than the original chip fabrication. PROMs provide flexibility and convenience.

EPROM :The optically erasable programmable read-only memory (EPROM) is read and written electrically, as with PROM. However, before a write operation, all the storage cells must be erased to the same initial state by exposure of the packaged chip to ultraviolet radiation. Erasure is performed by shining an intense ultraviolet light through a window that is designed into the memory chip. For comparable amounts of storage, the EPROM is more expensive than PROM, but it has the advantage of the multiple update capability.

EEPROM :A more attractive form of read-mostly memory is electrically erasable programmable read-only memory (EEPROM). This is a read-mostly memory that can be written into at any time without erasing prior contents; only the byte or bytes addressed are updated. EEPROM is more expensive than EPROM and also is less dense, supporting fewer bits per chip.

5.1.3 Chip Logic

Figure 5.3 shows a typical organization of a 16-Mbit DRAM. In this case, 4 bits are read or written at a time. Logically:

- the memory array is organized as four square arrays of 2048 by 2048 elements.
- Address lines :A total of $\log_2 W$ lines are needed. In our example,
  1. 11 address lines are needed to select one of 2048 rows. These 11 lines are fed into a row decoder, which has 11 lines of input and 2048 lines for output. The logic of the decoder activates a single one of the 2048 outputs depending on the bit pattern on the 11 input lines(2^{11}=2048) .
  2. An additional 11 address lines select one of 2048 columns
- Four data lines are used for the input and output of 4 bits to and from a data buffer.
  1. On input (write), the bit driver of each bit line is activated for a 1 or 0 according to the value of the corresponding data line.
  2. On output (read), the value of each bit line is passed through a sense amplifier and presented to the data lines. The row line selects which row of cells is used for reading or writing
- only 11 address lines (A0–A10), half the number you would expect for a array. This is done to save on the number of pins. The 22 required address lines are passed through select logic external to the chip and multiplexed onto the 11 address lines.
  1. First, 11 address signals are passed to the chip to define the row address of the array,
  2. and then the other 11 address signals are presented for the column address.
  3. These signals are accompanied by row address select(RAS) and column address select(CAS) signals to provide timing to the chip
  4. The write enable(WE) and output enable(OE) pins determine whether a write or read operation is performed.

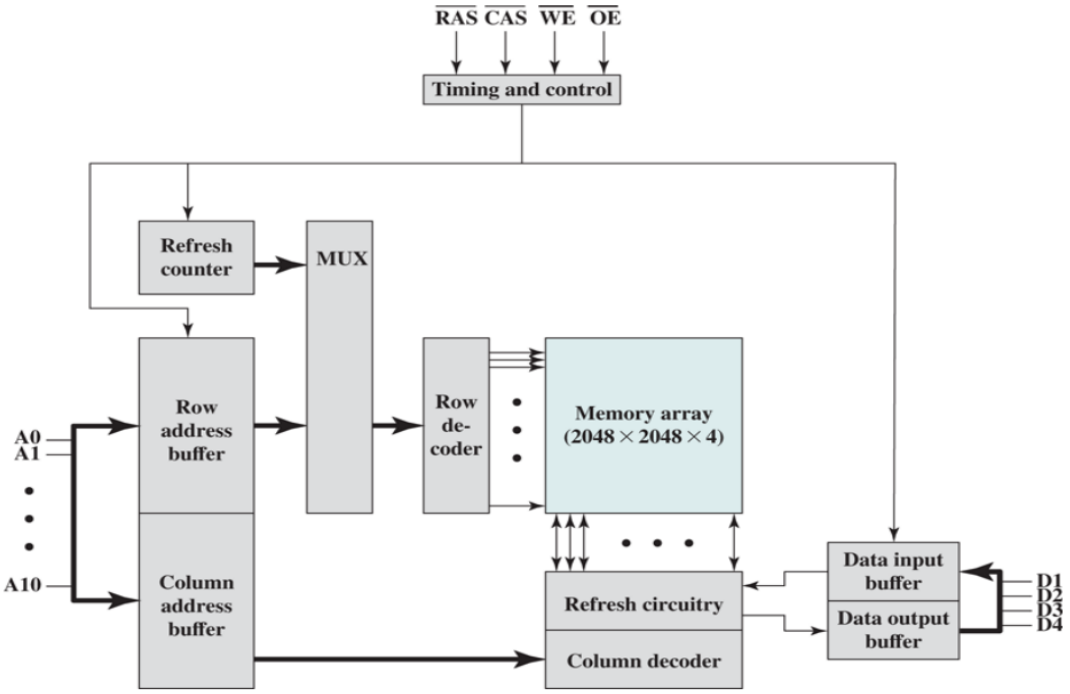


Figure 5.3 Typical 16-Mbit DRAM ($4M \times 4$)

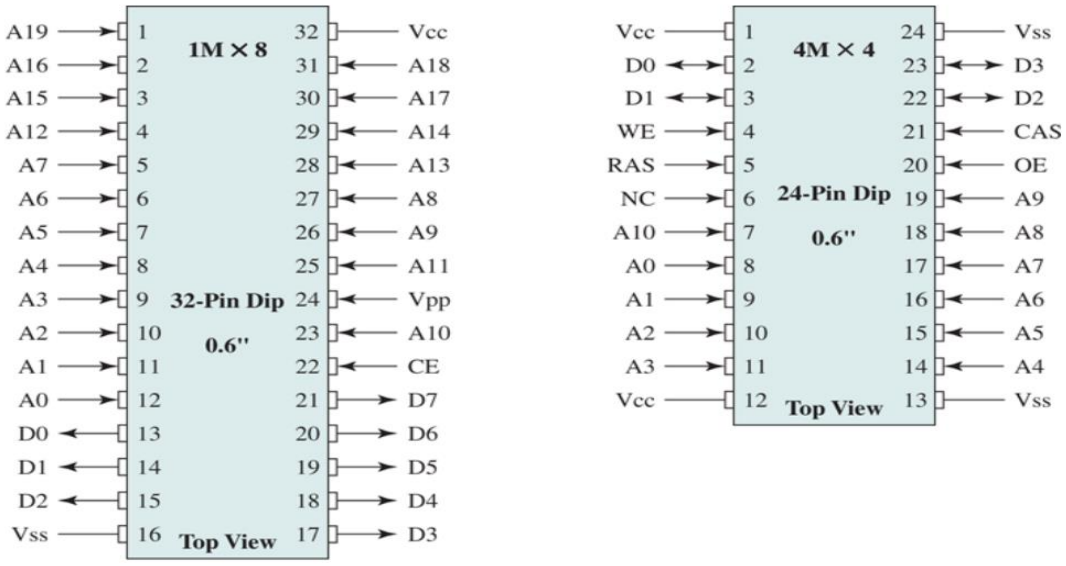
5.1.4 Chip Packaging

Figure 5.4a shows an example EPROM package, which is an 8-Mbit chip organized. The package includes 32 pins. The pins support the following signal lines:

- The address of the word being accessed. For 1M words, a total of 2^{20} pins are needed (A0–A19).
- The data to be read out, consisting of 8 lines (D0–D7).
- The power supply to the chip (V_{CC}) .
- A ground pin . A chip enable (CE) pin. Because there may be more than one memory chip, each of which is connected to the same address bus, the CE pin is used to indicate whether or not the address is valid for this chip..
- A program voltage(V_{pp})that is supplied during programming (write operations).

A typical DRAM pin configuration is shown in Figure 5.4b , for a 16-Mbit chip organized as 4M x 4. Because a RAM can be updated, the data pins are input/output.

- The write enable (WE) and output enable (OE) pins indicate whether this is a write or read operation.
- Because the DRAM is accessed by row and column, and the address is multiplexed, only 11 address pins are needed to specify the 4M row/column combinations: $2^{11} \times 2^{11} = 2^{22} = 4M$
- The row address select (RAS) and column address select (CAS) pins.
- Finally, the no connect (NC) pin is provided so that there are an even number of pins.



(a) 8-Mbit EPROM (b) 16-Mbit DRAM

Figure 5.4 Typical Memory Package Pins and Signals

5.2 Advanced DRAM Organization

In recent years, a number of enhancements to the basic DRAM architecture have been explored, and some of these are now on the market. The schemes that currently dominate the market are SDRAM, DDR-DRAM, and RDRAM**.**

5.2.1 SDRAM(Synchronous DRAM)

Synchronous dynamic random access memory (SDRAM) is dynamic random access memory (DRAM) with an interface synchronous with the system bus carrying data between the CPU and the memory controller hub. SDRAM has a rapidly responding synchronous interface, which is in sync with the system bus. SDRAM waits for the clock signal before it responds to control inputs.

- The speed of SDRAM is rated in MHz rather than in nanoseconds (ns).
- This makes it easier to compare the bus speed and the RAM chip speed.
- You can convert the RAM clock speed to nanoseconds by dividing the chip speed into 1 billion ns (which is one second). For example, an 83 MHz RAM would be equivalent to 12 ns.
- The SDRAM performs best when it is transferring large blocks of data sequentially, such as for applications like word processing, spreadsheets, and multimedia.

Table 5.1 DRAM Pin Assignments

A0 to A13	Address inputs
CLK	Clock input
CKE	Clock enable
$\overline{CS}$	Chip select
$\overline{RAS}$	Row address strobe
$\overline{CAS}$	Column address strobe
$\overline{WE}$	Write enable
DQ0 to DQ7	Data input/output
DQM	Data mask

In source-synchronous SDR interfaces, one edge of the clock, typically the rising edge, transfers the data.As shown in Figure 5.5,For the SDRAM operation, the burst length is 4 and the latency is 2. The burst read command is initiated by having and low while holding and high at the rising edge of the clock. The address inputs determine the starting column address for the burst, and the mode register sets the type of burst (sequential or interleave) and the burst length (1, 2, 4, 8, full page). The delay from the start of the command to when the data from the first cell appears on the outputs is equal to the value of the latency that is set in the mode register.

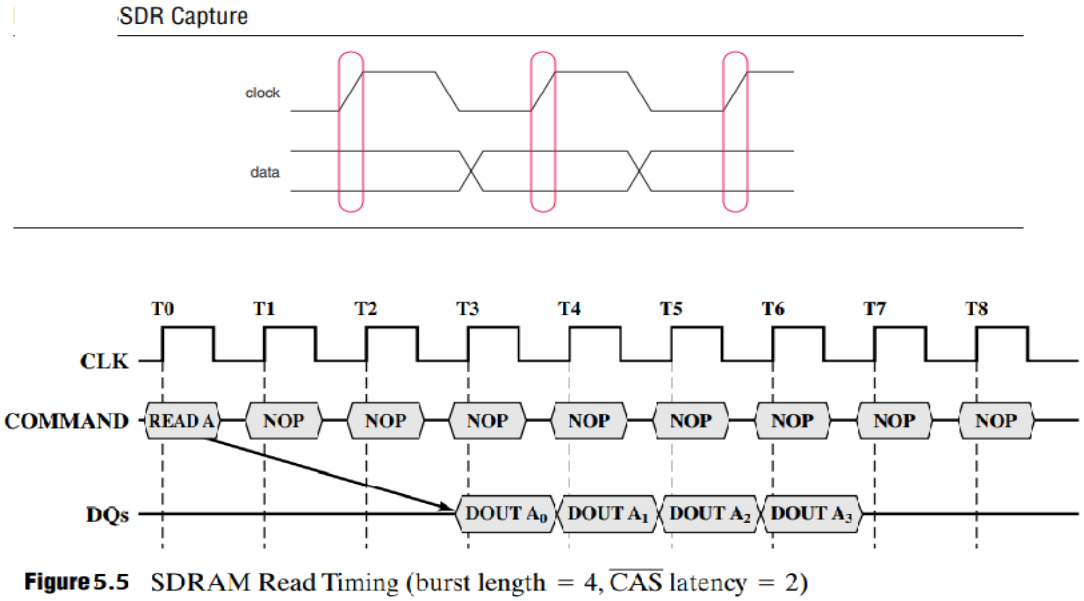


Figure 5.5 SDRAM Read Timing (burst length = 4, $\overline{\text{CAS}}$ latency = 2)

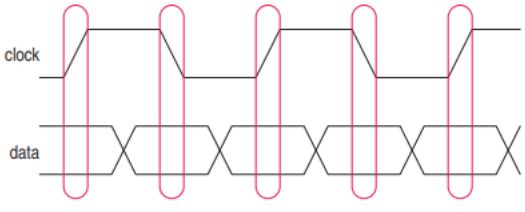
5.2.2 DDR SDRAM

SDRAM is limited by the fact that it can only send data to the processor once per bus clock cycle.A new version of SDRAM, referred to as double-data-rate SDRAM can send data twice per clock cycle, once on the rising edge of the clock pulse and once on the falling edge**.** DDR achieves higher data rates in three ways.

1. First, the data transfer is synchronized to both the rising and falling edge of the clock, rather than just the rising edge.
2. Second, DDR uses higher clock rate on the bus to increase the transfer rate.
3. Third, a buffering scheme is used, as explained subsequently.

In source-synchronous DDR interfaces, data is transferred on both edges of the clock, as shown below in Figure 5.6

Figure 5. DDR Capture



here are three significant characteristics differentiating SDRAM and DDR:

1. The main difference is the amount of data transmitted with each cycle, not the speed.
2. SDRAM sends signals once per clock cycle. DDR transfers data twice per clock cycle. (Both SDRAM and DDR use the same frequencies.)
3. SDRAM uses one edge of the clock. DDR uses both edges of the clock.

DDR SDRAM, also retroactively called DDR1 SDRAM, has been superseded by DDR2 SDRAM, DDR3 SDRAM, DDR4 SDRAM and DDR5 SDRAM. None of its successors are forward or backward compatible with DDR1 SDRAM, meaning DDR2, DDR3, DDR4 and DDR5 memory modules will not work in DDR1-equipped motherboards, and vice versa. JEDEC has thus far defined four generations of the DDR technology (Table 5.2).

Table 5.2 DDR Characteristics

	DDR1	DDR2	DDR3	DDR4
Prefetch buffer (bits)	2	4	8	8
Voltage level (V)	2.5	1.8	1.5	1.2
Front side bus data rates (Mbps)	200—400	400—1066	800—2133	2133—4266

In a prefetch buffer architecture, when a memory access occurs to a row the buffer grabs a set of adjacent data words on the row and reads them out ("bursts" them) in rapid-fire sequence on the IO pins, without the need for individual column address requests. This assumes the CPU wants adjacent datawords in memory, which in practice is very often the case. For instance, when a 64 bit CPU accesses a 16-bit-wide DRAM chip, it will need 4 adjacent 16 bit datawords to make up the full 64 bits. A 4n prefetch buffer would accomplish this exactly *("n" refers to the IO width of the memory chip; it is multiplied by the burst depth "4" to give the size in bits of the full burst sequence). An 8n prefetch buffer on a 8 bit wide DRAM would also accomplish a 64 bit transfer*..

NOTE: With data being transferred 64 bits at a time,
DDR SDRAM gives a max transfer rate (in bytes/s) = (memory bus clock rate=100MHz) × 2 (for dual rate) × 64 (number of bits transferred) / 8 (number of bits/byte)= 1600 MB/s.

5.3 Flash Memory

Another form of semiconductor memory is flash memory. Flash memory is used both for internal memory and external memory applications**.**
flash memory is intermediate between EPROM and EEPROM in both cost and functionality. Like EEPROM, flash memory uses an electrical erasing technology. An entire flash memory can be erased in one or a few seconds, which is much faster than EPROM. In addition, it is possible to erase just blocks of memory, rather than an entire chip. Flash memory gets its name because the microchip is organized so that a section of memory cells are erased in a single action or “flash.”

Figure 5.7 illustrates the basic operation of a flash memory. For comparison:

- Figure 5.7a depicts the operation of a transistor. Transistors exploit the properties of semiconductors so that a small voltage applied to the gate can be used to control the flow of a large current between the source and the drain.
- (Figure 5.7b),In a flash memory cell, a second gate—called a floating gate, is added to the transistor. Initially, the floating gate does not interfere with the operation of the transistor. In this state, the cell is deemed to represent binary 1.
- Applying a large voltage across the floating gate causes electrons to enter tunnel through it and become trapped on the floating gate, where they remain even if the power is disconnected (Figure 5.7c). In this state, the cell is deemed to represent binary 0.

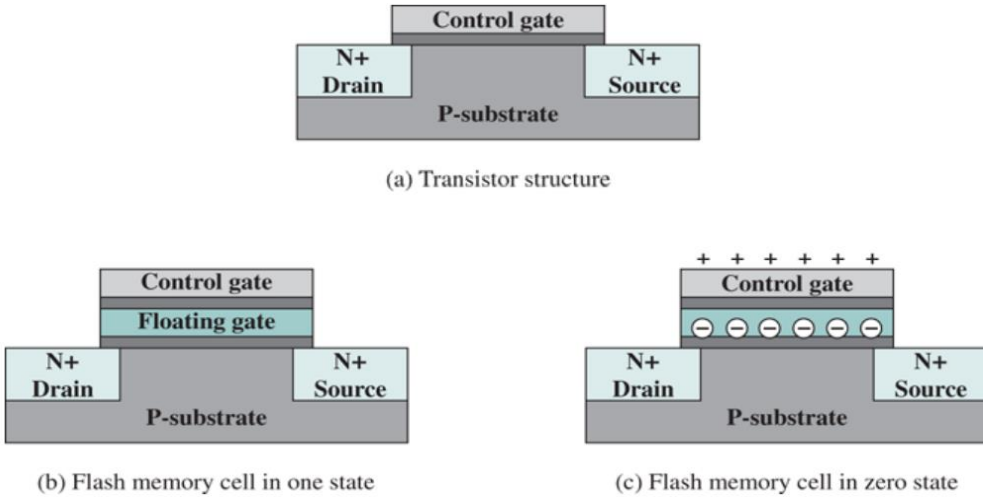
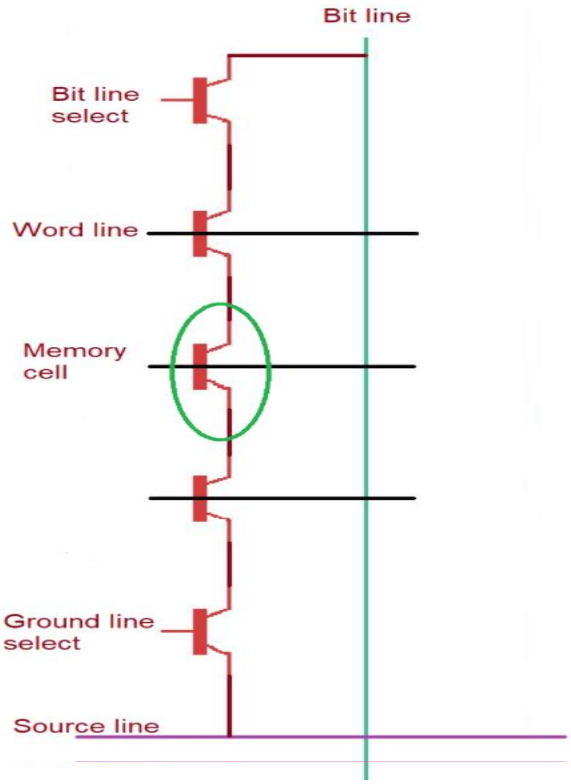


Figure 5.7 Flash Memory Operation

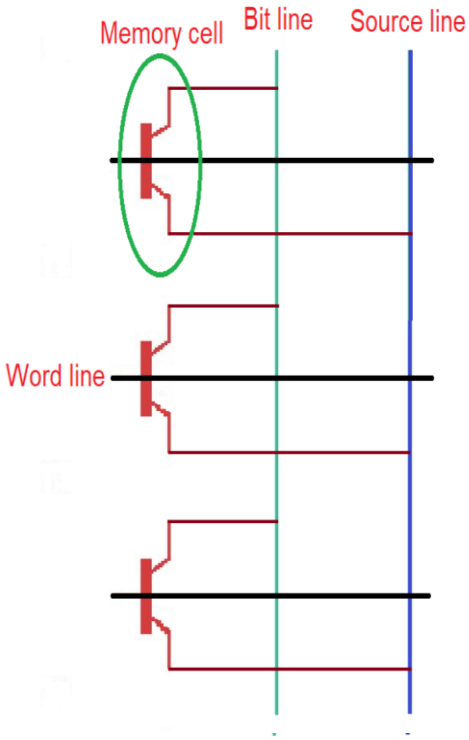
NOR and NAND Flash Memory

Flash memory architecture based on floating gate technology ,and There are two distinctive types of flash memory, designated as NOR and NAND (Figure 5.8).

- In NOR flash memory, the basic unit of access is a bit, referred to as a memory cell, every memory cell is connected to the floating gate. Cells in NOR flash are connected in parallel to the bit lines so that each cell can be read/write/erased individually. If any memory cell of the device is turned on by the corresponding word line, the bit line goes low. This is similar in function to a NOR logic gate**.** NOR memory is used for storing code and execution
- NAND flash memory is organized in transistor arrays with 16 or 32 transistors in series. The bit line goes low only if all the transistors in the corresponding word lines are turned on(several memory cells are connected in parallel.). This is similar in function to a NAND logic gate. NAND memory is used for data storage

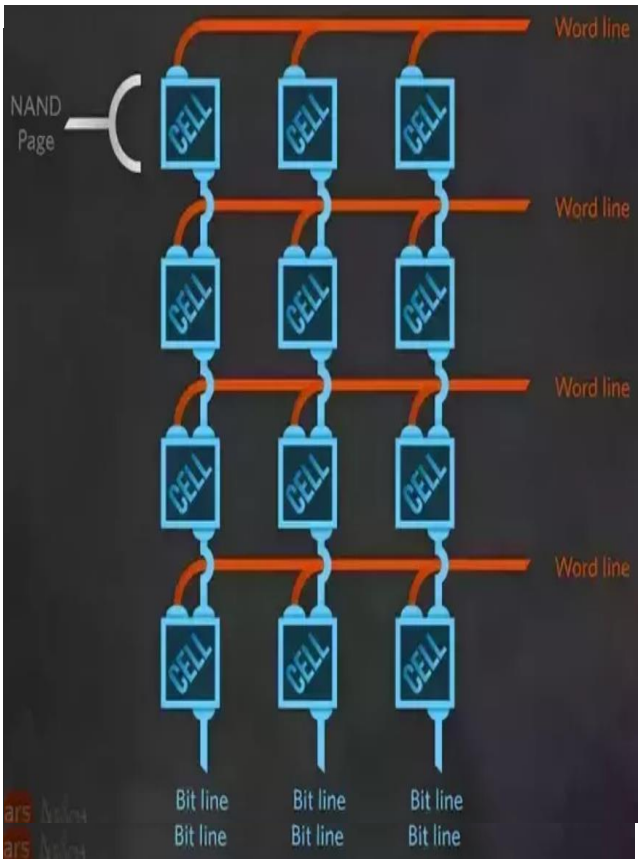


(b)NAND structure



(a)NOR structure

NAND



NOR

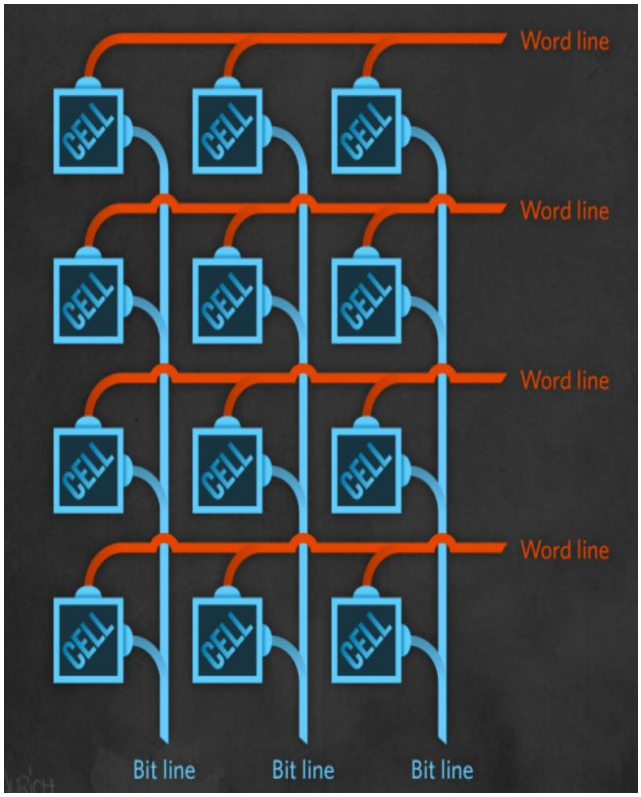


Figure 5.9 Flash Memory Structure