



**William Stallings
Computer Organization
and Architecture
10th Edition**

+ Chapter 5

Internal Memory



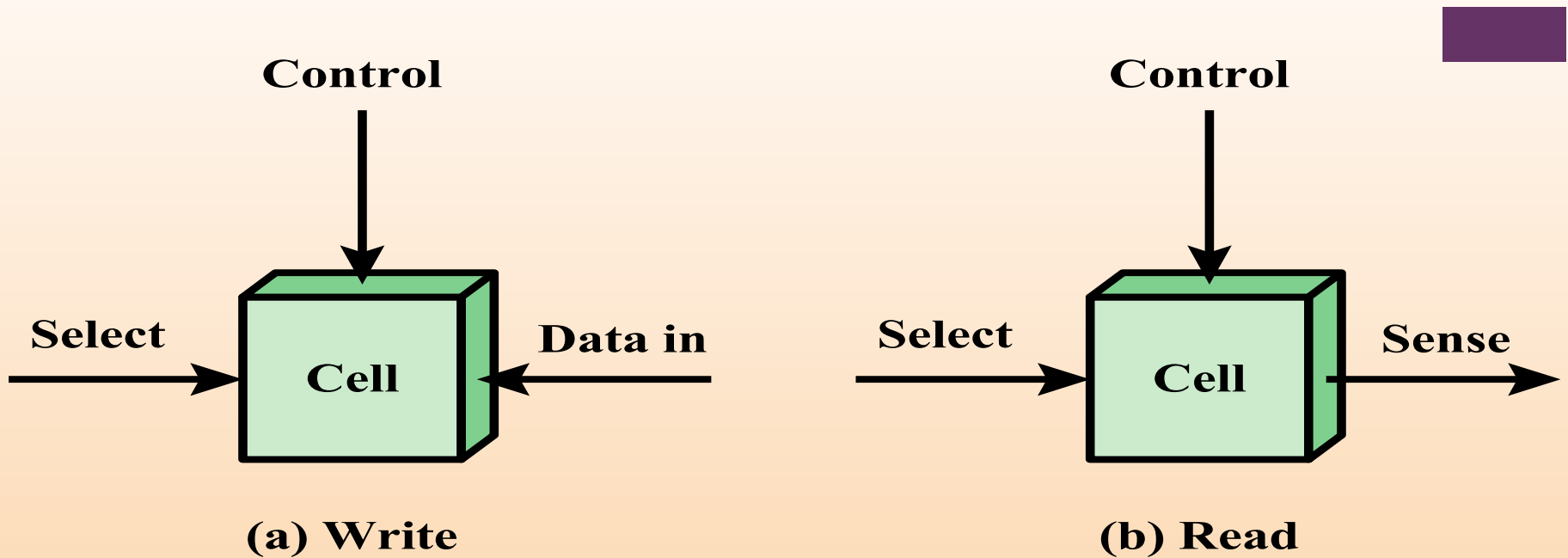


Figure 5.1 Memory Cell Operation

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.

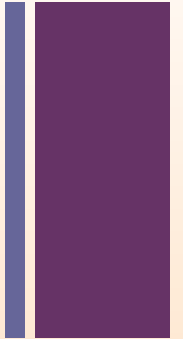


Memory Type	Category	Erase	Write Mechanism	Volatility
Random-access memory (RAM)	Read-write memory	Electrically, byte-level	Electrically	Volatile
Read-only memory (ROM)	Read-only memory	Not possible	Masks	Nonvolatile
Programmable ROM (PROM)				
Erasable PROM (EPROM)	Read-mostly memory	UV light, chip-level	Electrically	
Electrically Erasable PROM (EEPROM)		Electrically, byte-level		
Flash memory		Electrically, block-level		

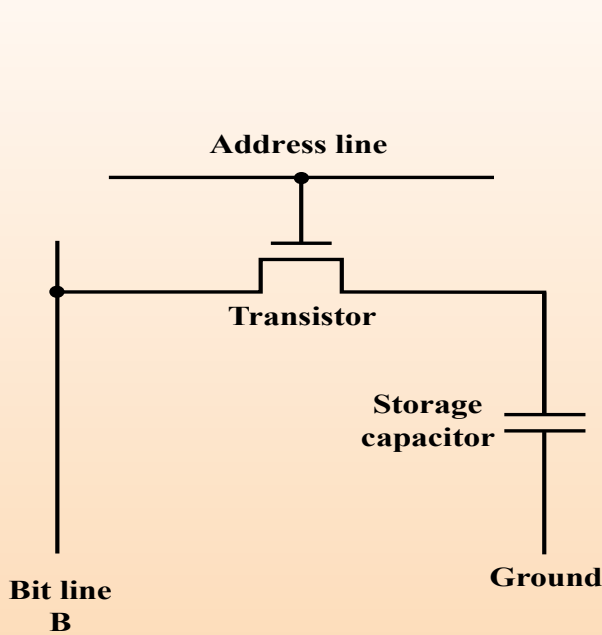
Table 5.1
Semiconductor Memory Types



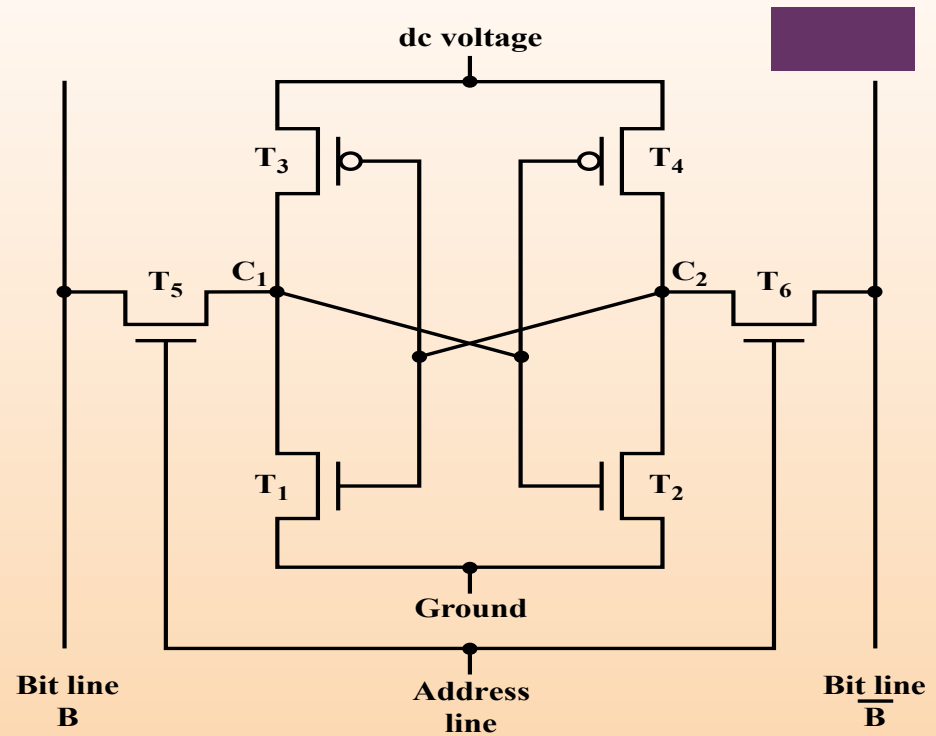
Dynamic RAM (DRAM)



- RAM technology is divided into two technologies:
 - Dynamic RAM (DRAM)
 - Static RAM (SRAM)
- DRAM
 - Made with cells that **store data as charge** on capacitors
 - Presence or absence of charge in a capacitor is interpreted as a binary 1 or 0
 - Requires **periodic charge refreshing** to maintain data storage
 - The term *dynamic* refers to tendency of the stored charge to leak away, even with power continuously applied



(a) Dynamic RAM (DRAM) cell



(b) Static RAM (SRAM) cell

Figure 5.2 Typical Memory Cell Structures

For the DRAM 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.

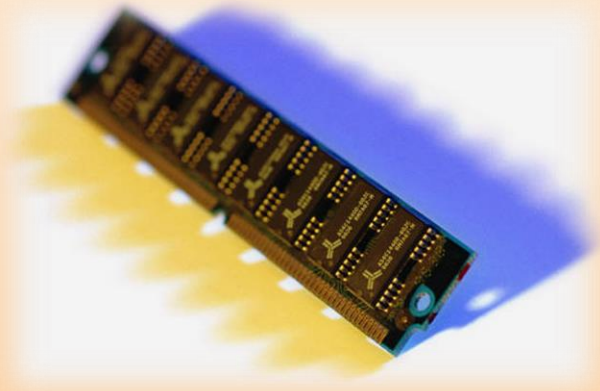
Figure 5.2b is a typical SRAM structure for an individual cell. Four transistors (T₁, T₂, T₃, T₄) are cross connected in an arrangement that produces a stable logic state.

As in the DRAM, the SRAM address line is used to open or close a switch. The address line controls two transistors (T₅ and T₆). When a signal is applied to this line, the two transistors are switched on, allowing a read or write operation.



Static RAM (SRAM)

- Digital device that uses the same logic elements used in the processor
- Binary values are stored using traditional flip-flop logic gate configurations
- Will hold its data as long as power is supplied to it



SRAM versus DRAM

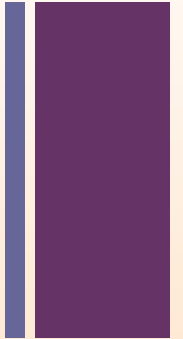
- Both volatile
 - Power must be continuously supplied to the memory to preserve the bit values
- Dynamic cell
 - Simpler to build, smaller
 - More dense (smaller cells = more cells per unit area)
 - Less expensive
 - Requires the supporting refresh circuitry
 - Tend to be favored for large memory requirements
 - Used for main memory
- Static
 - Faster
 - Used for cache memory (both on and off chip)

SRAM

DRAM



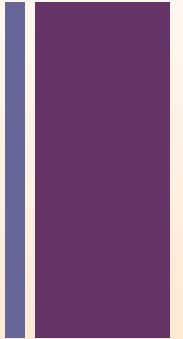
Read Only Memory (ROM)



- Contains a permanent pattern of data that cannot be changed or added to
- No power source is required to maintain the bit values in memory
- Data or program is permanently in main memory and never needs to be loaded from a secondary storage device
- Data is actually wired into the chip as part of the fabrication process
 - Disadvantages of this:
 - No room for error, if one bit is wrong the whole batch of ROMs must be thrown out
 - Data insertion step includes a relatively large fixed cost



Programmable ROM (PROM)



- Less expensive alternative
- **Nonvolatile and may be written into only once**
- Writing process is performed electrically and may be performed by supplier or customer at a time later than the original chip fabrication
- Special equipment is required for the writing process
- Provides flexibility and convenience
- Attractive for high volume production runs

Read-Mostly Memory



EPROM

Erasable programmable read-only memory

Erasure process can be performed repeatedly

More expensive than PROM but it has the advantage of the multiple update capability

EEPROM

Electrically erasable programmable read-only memory

Can be written into at any time without erasing prior contents

Combines the advantage of non-volatility with the flexibility of being updatable in place

More expensive than EPROM

Flash Memory

Intermediate between EPROM and EEPROM in both cost and functionality

Uses an electrical erasing technology, does not provide byte-level erasure

Microchip is organized so that a section of memory cells are erased in a single action or "flash"

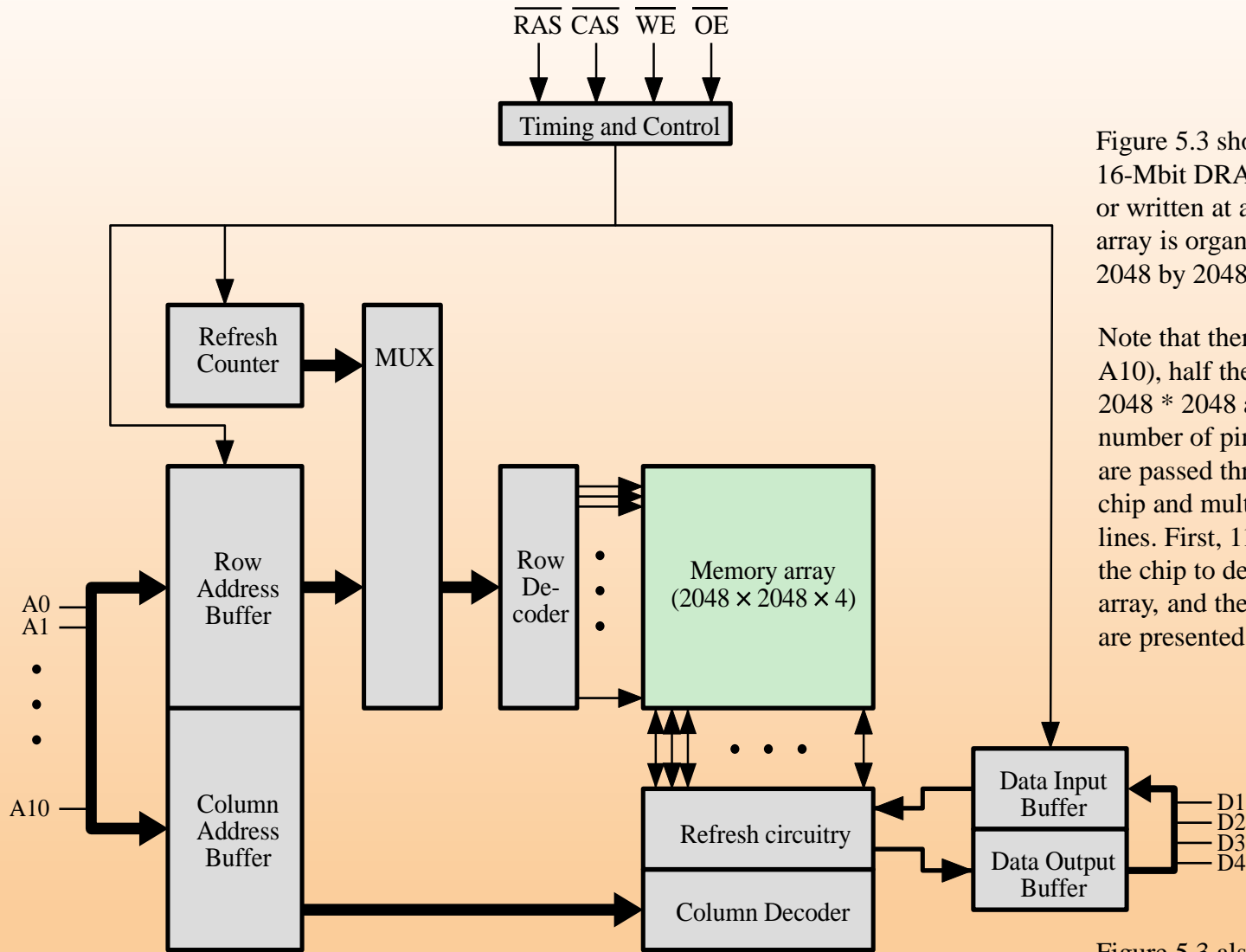
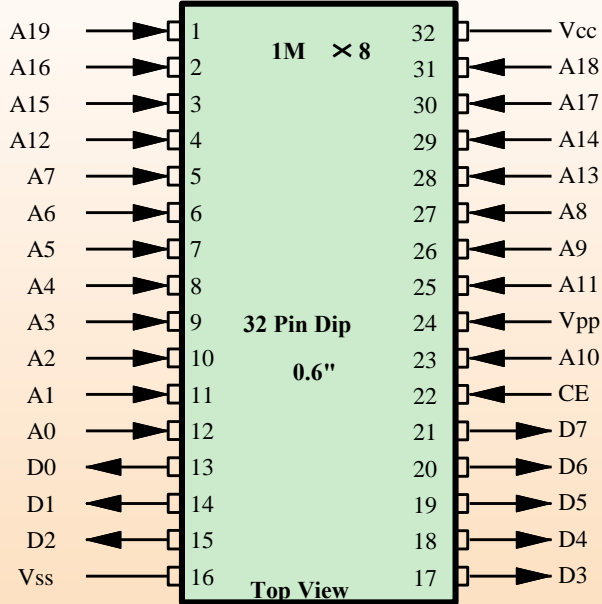


Figure 5.3 shows a typical organization of a 16-Mbit DRAM. In this case, 4 bits are read or written at a time. Logically, the memory array is organized as four square arrays of 2048 by 2048 elements.

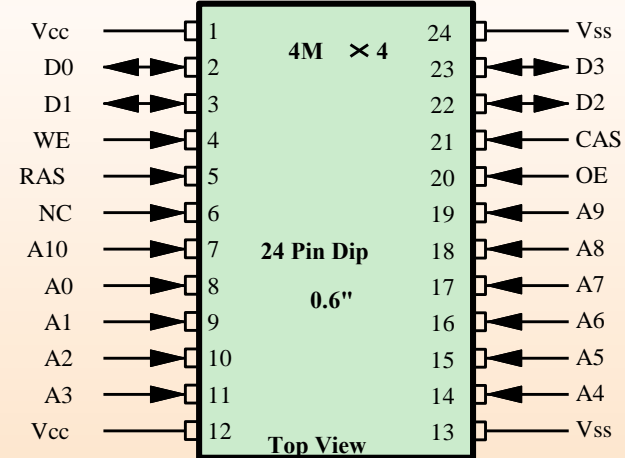
Note that there are only 11 address lines (A0–A10), half the number you would expect for a 2048 * 2048 array. This is done to save on the number of pins. The 22 required address lines are passed through select logic external to the chip and multiplexed onto the 11 address lines. First, 11 address signals are passed to the chip to define the row address of the array, and then the other 11 address signals are presented for the column address.

Figure 5.3 also indicates the inclusion of refresh circuitry. All DRAMs require a refresh operation

Figure 5.3 Typical 16 Megabit DRAM (4M × 4)



(a) 8 Mbit EPROM



(b) 16 Mbit DRAM

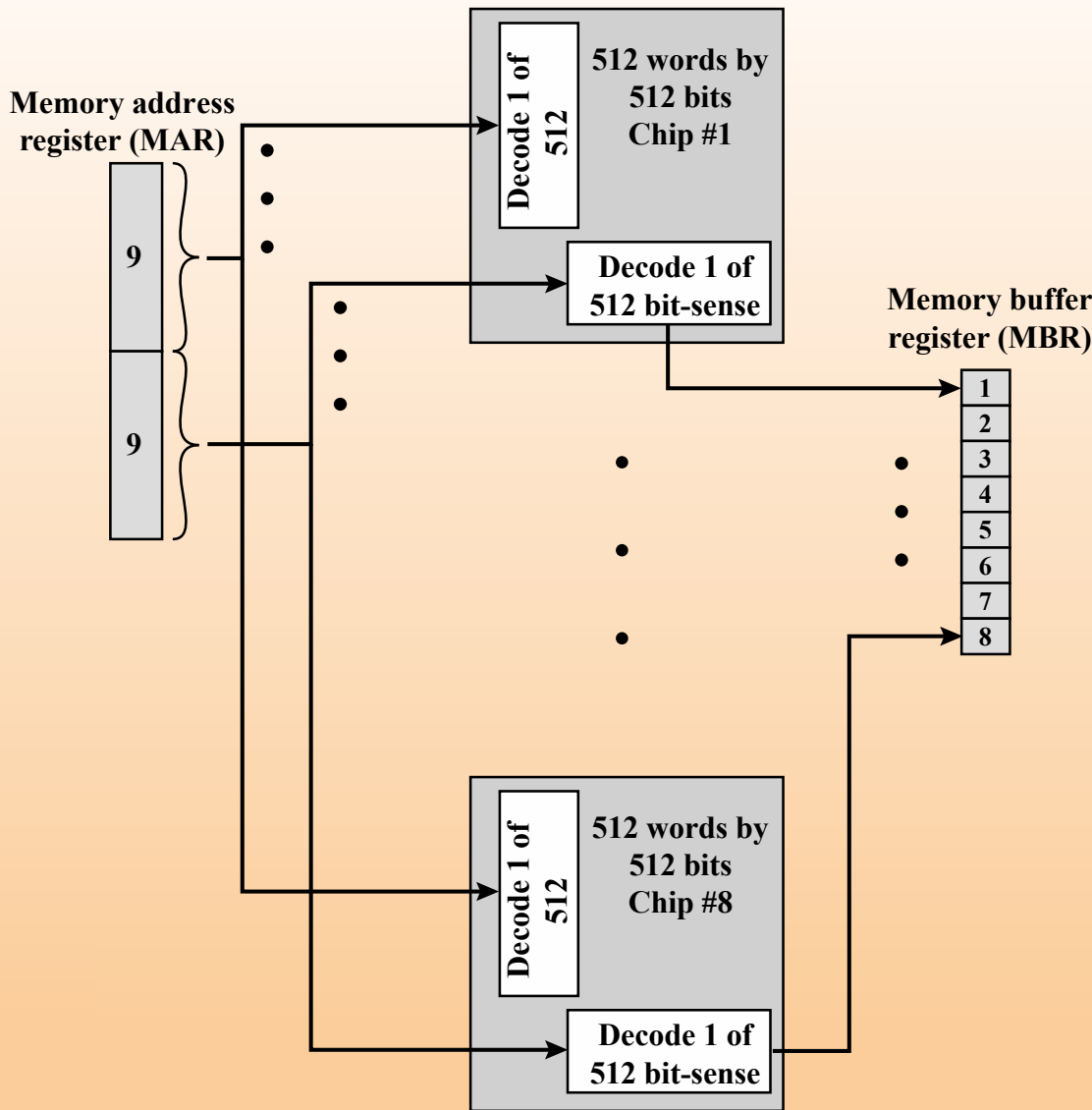
Figure 5.4 Typical Memory Package Pins and Signals

Figure 5.4a shows an example EPROM package, which is an 8-Mbit chip organized as 1M * 8.

- The address of the word being accessed. For 1M words, a total of 20 ($2^{20} = 1M$) 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 (V_{ss}).
- 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 * 4. 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 ($2^{11} * 2^{11} = 2^{22} = 4M$). The functions of the row address select (RAS) and column address select (CAS) pins were discussed previously. Finally, the no connect (NC) pin is provided so that there are an even number of pins.



If a RAM chip contains only 1 bit per word, then clearly we will need at least a number of chips equal to the number of bits per word. As an example,

Figure 5.5 shows how a memory module consisting of 256K 8-bit words could be organized. For 256K words, an 18-bit address is needed and is supplied to the module from some external source (e.g., the address lines of a bus to which the module is attached).

The address is presented to 8 256K * 1-bit chips, each of which provides the input/output of 1 bit.

Figure 5.5 256-KByte Memory Organization

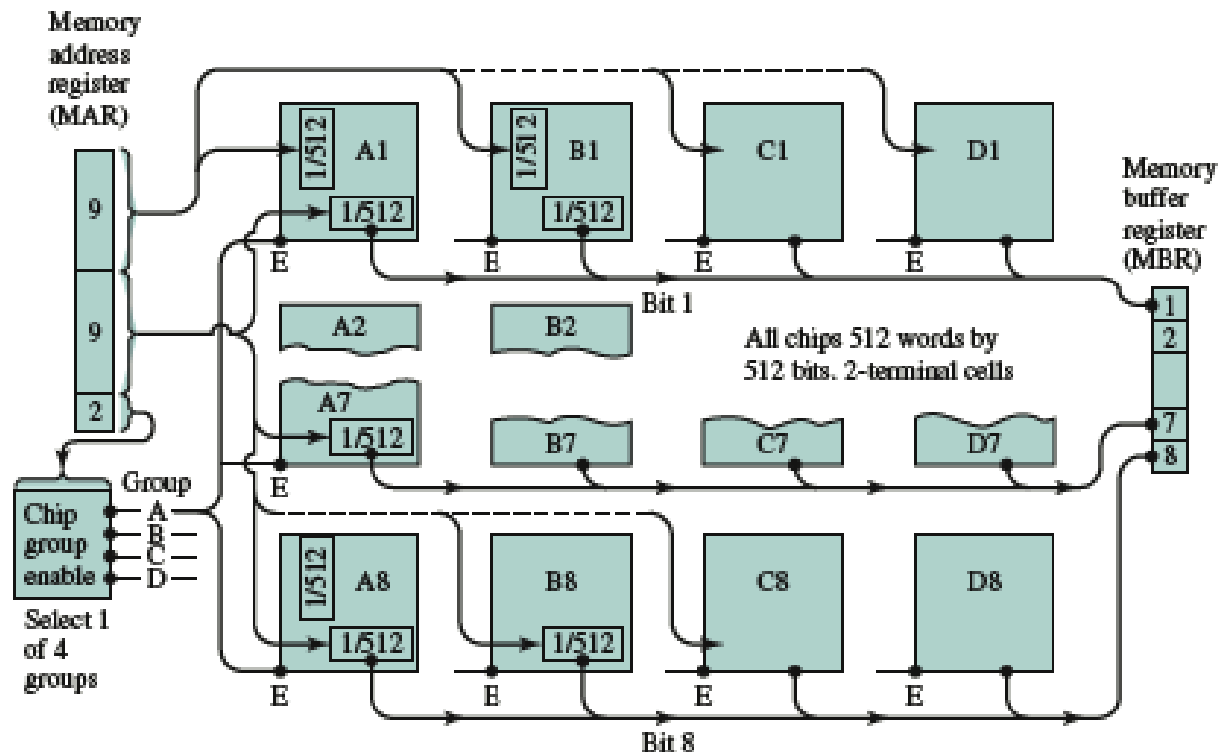
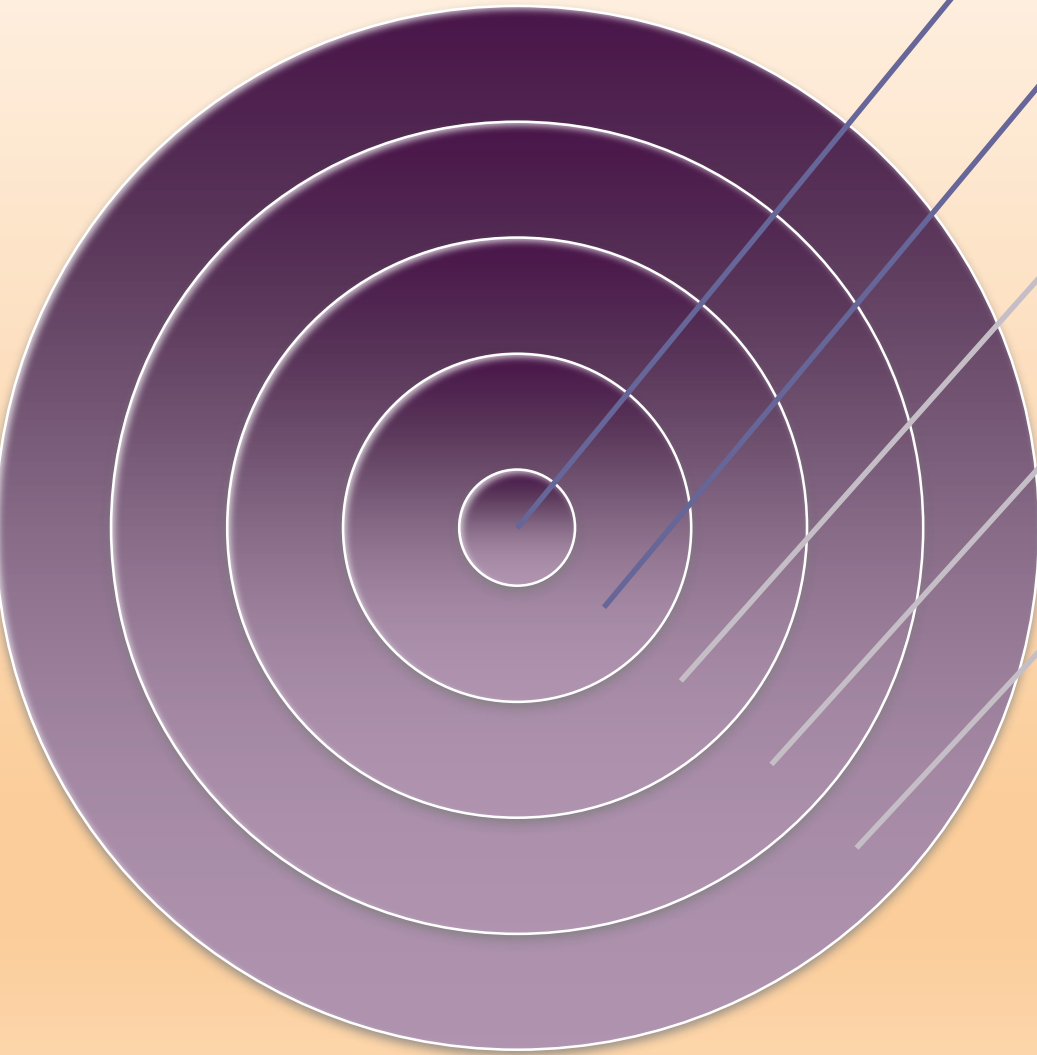


Figure 5.6 1-Mbyte Memory Organization

This organization works as long as the size of memory equals the number of bits per chip. In the case in which larger memory is required, an array of chips is needed. Figure 5.6 shows the possible organization of a memory consisting of 1M word by 8 bits per word. In this case, we have four columns of chips, each column containing 256K words arranged as in Figure 5.5. For 1M word, 20 address lines are needed. The 18 least significant bits are routed to all 32 modules. The high-order 2 bits are input to a group select logic module that sends a chip enable signal to one of the four columns of modules.

Interleaved Memory



Composed of a collection of DRAM chips

Grouped together to form a *memory bank*. It is possible to organize the memory banks in a way known as *interleaved memory*.

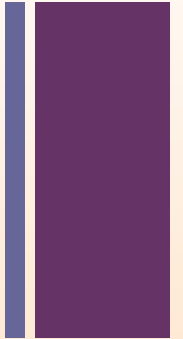
Each bank is independently able to service a memory read or write request

K banks can service K requests simultaneously, increasing memory read or write rates by a factor of K

If consecutive words of memory are stored in different banks, the transfer of a block of memory is speeded up



Error Correction



■ Hard Failure

- Permanent physical defect
- Memory cell or cells affected cannot reliably store data but become stuck at 0 or 1 or switch erratically between 0 and 1
- Can be caused by:
 - Harsh environmental abuse
 - Manufacturing defects
 - Wear

■ Soft Error

- Random, non-destructive event that alters the contents of one or more memory cells
- No permanent damage to memory
- Can be caused by:
 - Power supply problems
 - Alpha particles

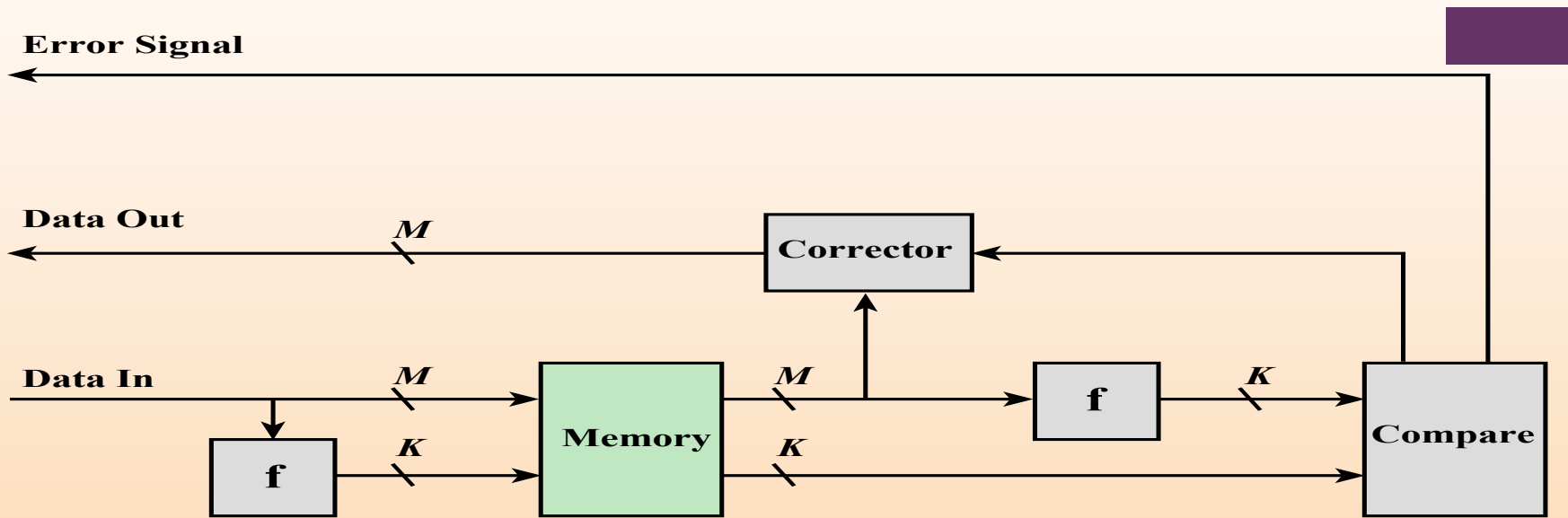


Figure 5.7 Error-Correcting Code Function

Figure 5.7 illustrates in general terms how the process is carried out. When data are to be written into memory, a calculation, depicted as a function f , is performed on the data to produce a code. Both the code and the data are stored. Thus, if an M -bit word of data is to be stored and the code is of length K bits, then the actual size of the stored word is $M + K$ bits.

When the previously stored word is read out, the code is used to detect and possibly correct errors. A new set of K code bits is generated from the M data bits and compared with the fetched code bits. The comparison yields one of three results:

- No errors are detected. The fetched data bits are sent out.
- An error is detected, and it is possible to correct the error. The data bits plus **error correction** bits are fed into a corrector, which produces a corrected set of M bits to be sent out.
- An error is detected, but it is not possible to correct it. This condition is reported.

Codes that operate in this fashion are referred to as **error-correcting codes**. A code is characterized by the number of bit errors in a word that it can correct and detect.

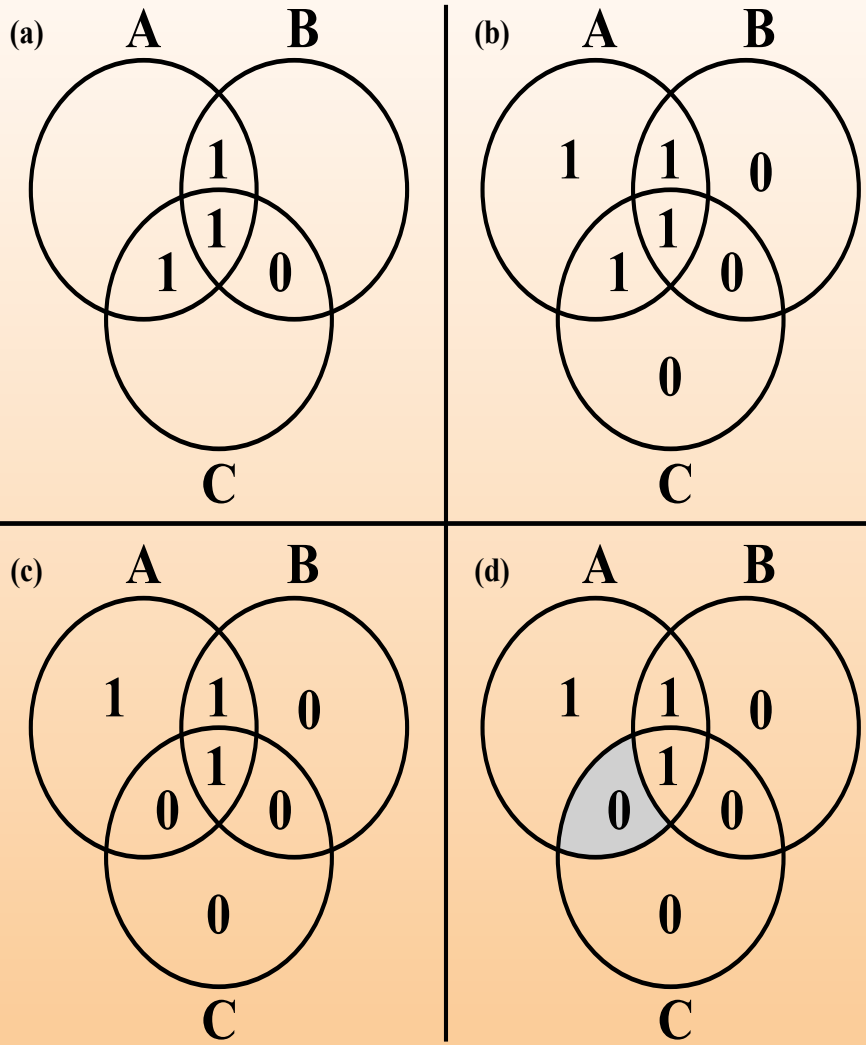


Figure 5.8 Hamming Error-Correcting Code

The simplest of the error-correcting codes is the **Hamming code** devised by Richard Hamming at Bell Laboratories. Figure 5.8 uses Venn diagrams to illustrate the use of this code on 4-bit words ($M = 4$).
 -With three intersecting circles, there are seven compartments.


-We assign the 4 data bits to the inner compartments (Figure 5.8a).

-The remaining compartments are filled with what are called *parity bits*.

-Each parity bit is chosen so that the total number of 1s in its circle is even (Figure 5.8b). Thus, because circle A includes three data 1s, the parity bit in that circle is set to 1.

-Now, if an error changes one of the data bits (Figure 5.8c), it is easily found. By checking the parity bits, discrepancies are found in circle A and circle C but not in circle B.

Only one of the seven compartments is in A and C but not B. The error can therefore be corrected by changing that bit.



To clarify the concepts involved, we will develop a code that can detect and correct single-bit errors in 8-bit words.

To start, let us determine how long the code must be. Referring to Figure 5.7, the comparison logic receives as input two K -bit values.

- A bit-by-bit comparison is done by taking the exclusive-OR of the two inputs. The result is called the *syndrome word*.

- Thus, each bit of the **syndrome** is 0 or 1 according to if there is or is not a match in that bit position for the two inputs.

- The syndrome word is therefore K bits wide and has a range between 0 and 2^K-1 .

- The value 0 indicates that no error was detected, leaving 2^K-1 values to indicate, if there is an error, which bit was in error.

- Now, because an error could occur on any of the M data bits or K check bits, we must have $2^K-1 \geq M + K$

This inequality gives the number of bits needed to correct a single bit error in a word containing M data bits. For example, for a word of 8 data bits ($M = 8$), we have

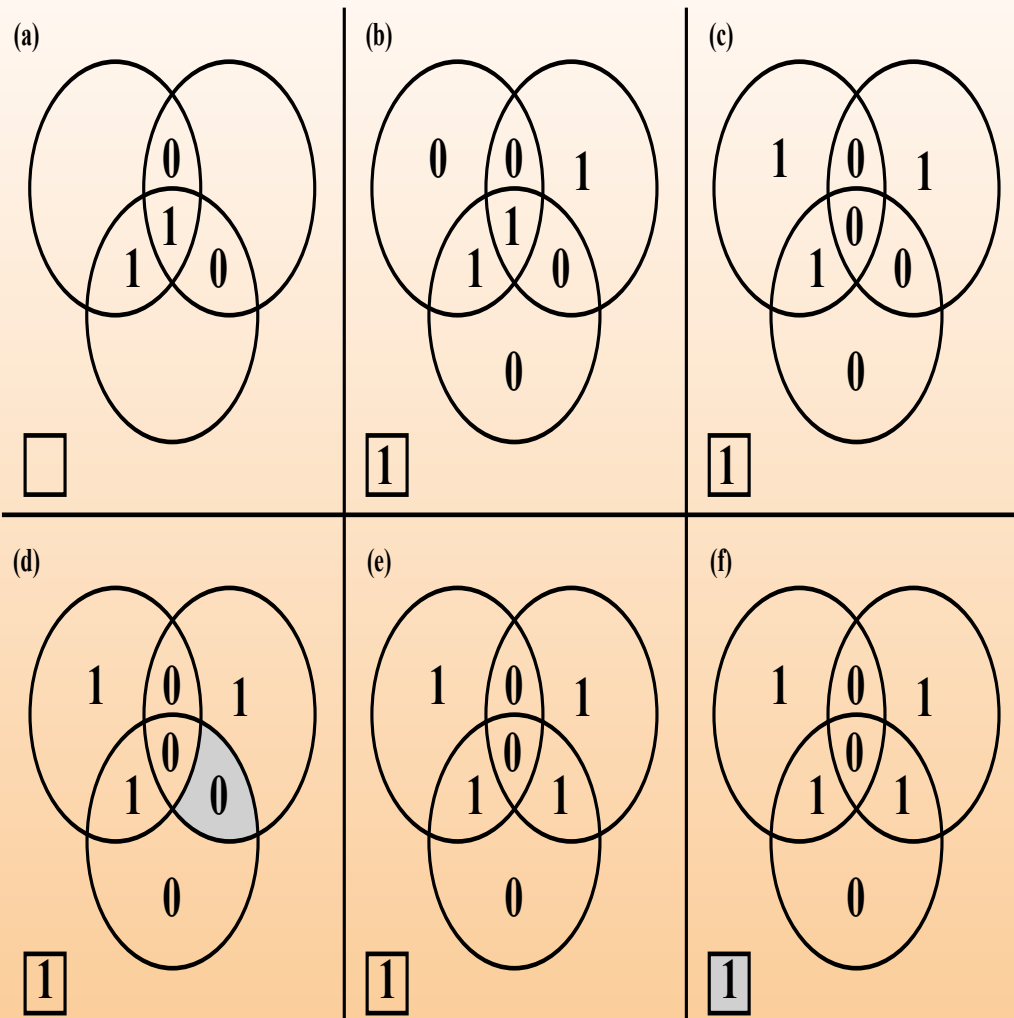
- $K = 3: 2^3-1 < 8 + 3$

- $K = 4: 2^4-1 > 8 + 4$



Data Bits	Single-Error Correction		Single-Error Correction/ Double-Error Detection	
	Check Bits	% Increase	Check Bits	% Increase
8	4	50	5	62.5
16	5	31.25	6	37.5
32	6	18.75	7	21.875
64	7	10.94	8	12.5
128	8	6.25	9	7.03
256	9	3.52	10	3.91

Table 5.2
Increase in Word Length with Error Correction



More commonly, semiconductor memory is equipped with a **single-error-correcting, double-error-detecting (SEC-DED) code**. As Table 5.2 shows, such codes require one additional bit compared with SEC codes.

Figure 5.11 illustrates how such a code works, again with a 4-bit data word. The sequence shows that if two errors occur (Figure 5.11c), the checking procedure goes astray (d) and worsens the problem by creating a third error (e). To overcome the problem, **an eighth bit is added that is set so that the total number of 1s in the diagram is even**. The extra parity bit catches the error (f).

Figure 5.11 Hamming SEC-DED Code

Advanced DRAM Organization

- One of the most critical system bottlenecks when using high-performance processors is the interface to main internal memory
- The traditional DRAM chip is constrained both by its internal architecture and by its interface to the processor's memory bus
- A number of enhancements to the basic DRAM architecture have been explored

+

- The schemes that currently dominate the market are SDRAM and DDR-DRAM

SDRAM

DDR-DRAM

RDRAM

Synchronous DRAM (SDRAM)



One of the most widely used forms of DRAM

Exchanges data with the processor synchronized to an external clock signal and running at the full speed of the processor/memory bus without imposing wait states

With synchronous access the DRAM moves data in and out under control of the system clock

- The processor or other master issues the instruction and address information which is latched by the DRAM
- The DRAM then responds after a set number of clock cycles
- Meanwhile the master can safely do other tasks while the SDRAM is processing

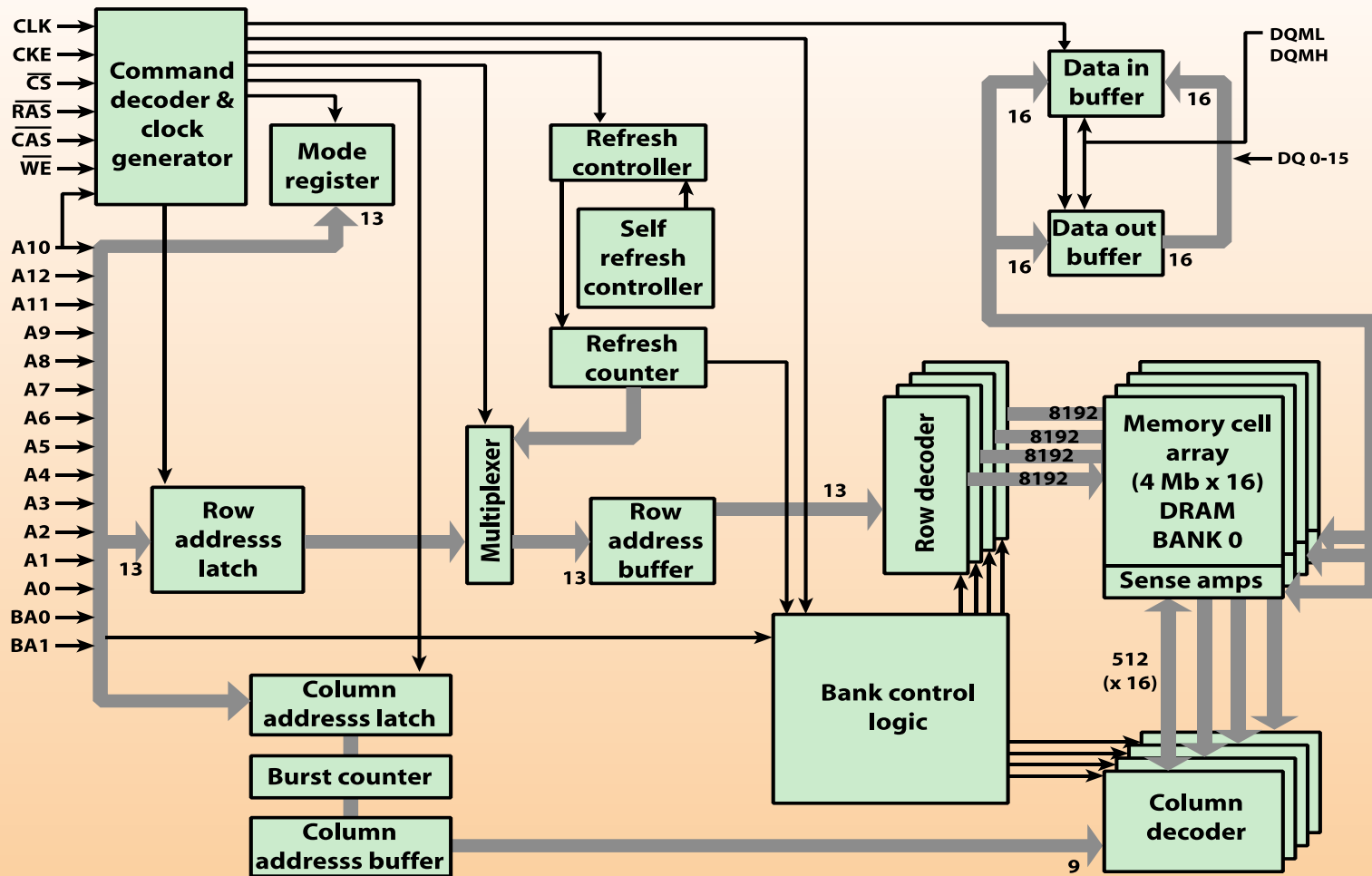


Figure 5.12 256-Mb Synchronous Dynamic RAM (SDRAM)

The internal logic of a typical 256-Mb SDRAM typical of SDRAM organization.

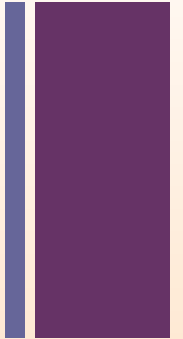
A0 to A12	Address inputs
BA0, BA1	Bank address lines
CLK	Clock input
CKE	Clock enable
$\overline{\text{CS}}$	Chip select
$\overline{\text{RAS}}$	Row address strobe
$\overline{\text{CAS}}$	Column address strobe
$\overline{\text{WE}}$	Write enable
DQ0 to DQ15	Data input/output
DQM	Data mask

Table 5.3

**SDRAM
Pin
Assignments**



Double Data Rate SDRAM (DDR SDRAM)



- Developed by the JEDEC Solid State Technology Association (Electronic Industries Alliance's semiconductor-engineering-standardization body)
- Numerous companies make DDR chips, which are widely used in desktop computers and servers
- DDR achieves higher data rates in **three ways**:
 - First, **the data transfer is synchronized to both the rising and falling edge of the clock**, rather than just the rising edge
 - Second, DDR **uses higher clock rate** on the bus to increase the transfer rate
 - Third, **a buffering scheme** is used



	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

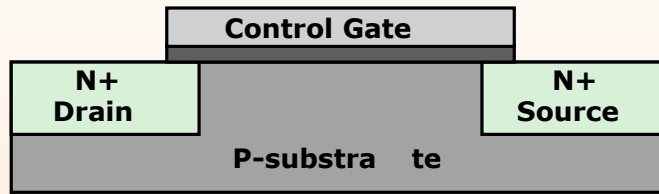
Table 5.4
DDR Characteristics

JDEC has thus far defined four generations of the DDR technology (Table 5.4).

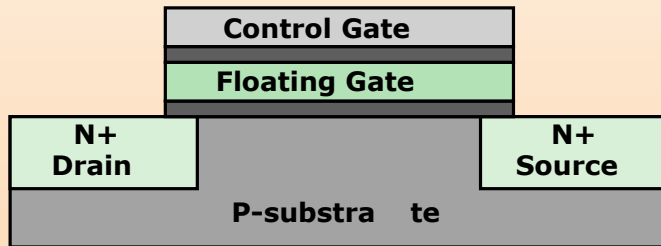
- The initial DDR version makes use of a 2-bit prefetch buffer. The prefetch buffer is a memory cache located on the SDRAM chip. It enables the SDRAM chip to preposition bits to be placed on the data bus as rapidly as possible.
- The DDR I/O bus uses the same clock rate as the memory chip, but because it can handle two bits per cycle, it achieves a data rate that is double the clock rate.
- The 2-bit prefetch buffer enables the SDRAM chip to keep up with the I/O bus.

+ Flash Memory

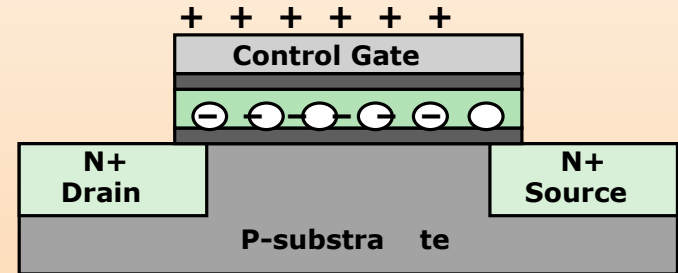
- Used both for internal memory and external memory applications
- First introduced in the mid-1980's
- Is intermediate between EPROM and EEPROM in both cost and functionality
- Uses an electrical erasing technology like EEPROM
- It is possible to **erase just blocks of memory** rather than an entire chip
- Gets its name because the microchip is organized so that a section of memory cells are erased in a single action
- Does not provide byte-level erasure
- Uses only one transistor per bit so it achieves the high density of EPROM



(a) Transistor structure



(b) Flash memory cell in one state



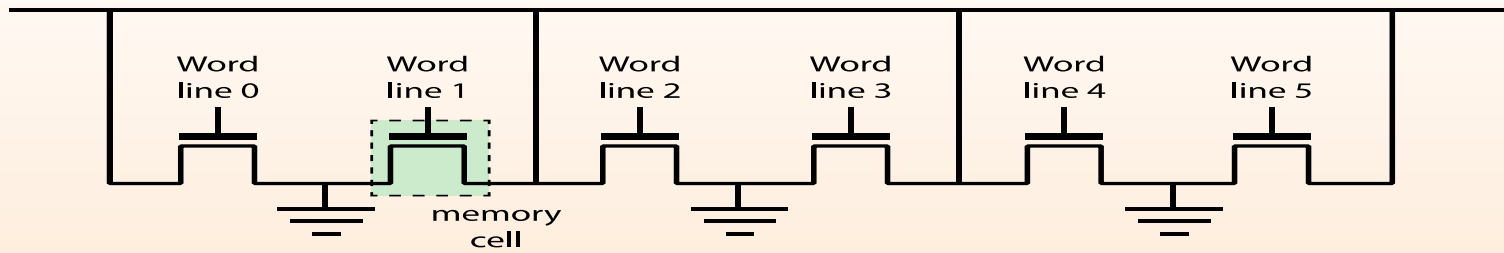
(c) Flash memory cell in zero state

Figure 5.15 Flash Memory Operation

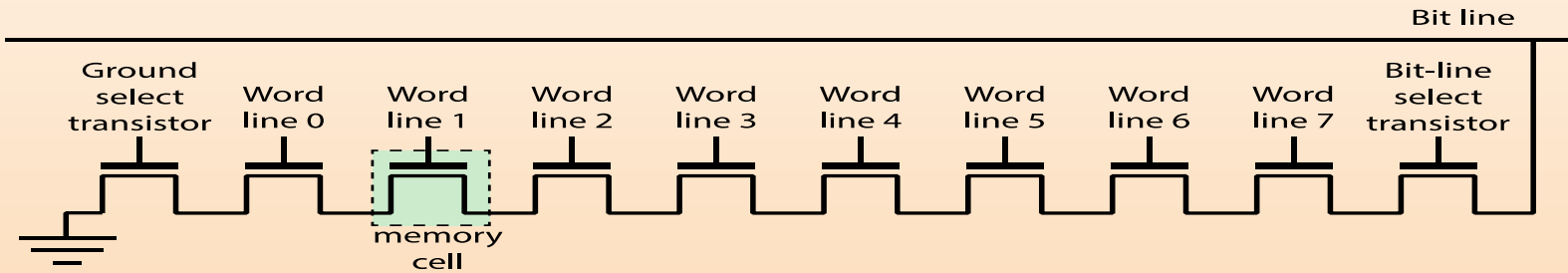
-Figure 5.15 illustrates the basic operation of a flash memory. For comparison, Figure 5.15a depicts the operation of a transistor. Transistors exploit the properties of semiconductors so that a small voltage applied to the gate can be used to control the flow of a large current between the source and the drain.

-In a flash memory cell, a second gate—called a floating gate, because it is insulated by a thin oxide layer—is added to the transistor. Initially, the floating gate does not interfere with the operation of the transistor (Figure 5.15b). In this state, the cell is deemed to represent binary 1. Applying a large voltage across the oxide layer causes electrons to tunnel through it and become trapped on the floating gate, where they remain even if the power is disconnected (Figure 5.15c). In this state, the cell is deemed to represent binary 0. The state of the cell can be read by using external circuitry to test whether the transistor is working or not. Applying a large voltage in the opposite direction removes the electrons from the floating gate, returning to a state of binary 1.

-An important characteristic of flash memory is that it is persistent memory, which means that it retains data when there is no power applied to the memory. Thus, it is useful for secondary (external) storage, and as an alternative to random access memory in computers.



(a) NOR flash structure



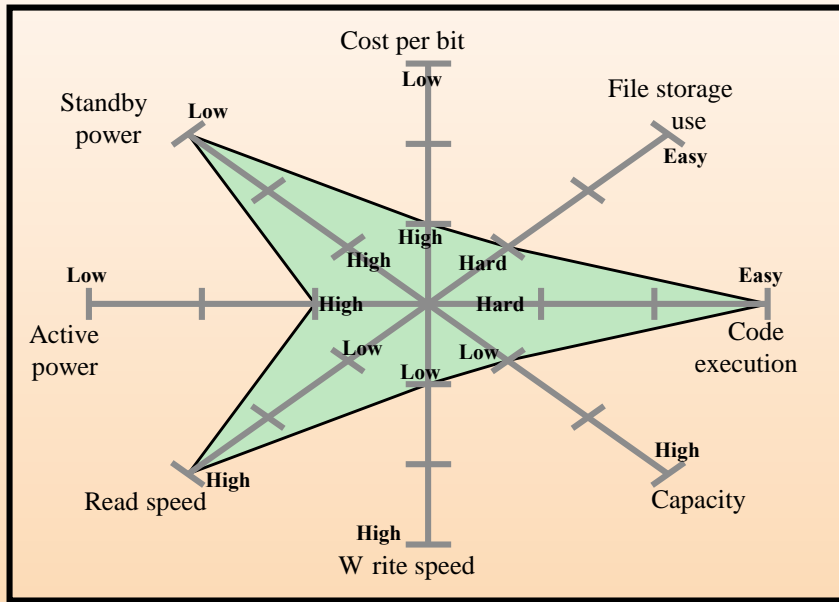
(b) NAND flash structure

Figure 5.16 Flash Memory Structures

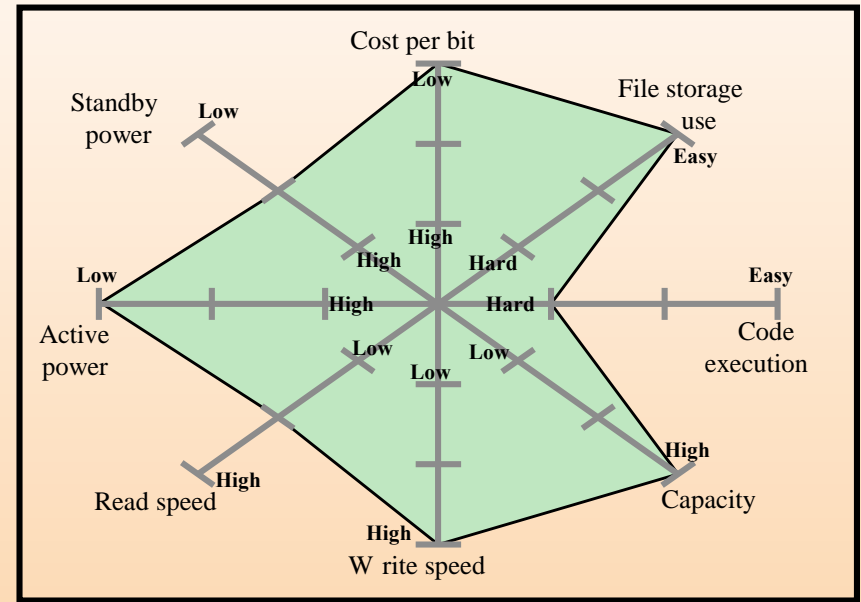
There are two distinctive types of flash memory, designated as NOR and NAND (Figure 5.16).

-In NOR flash memory, the basic unit of access is a bit, referred to as a memory cell. 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.

-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. This is similar in function to a NAND logic gate.



(a) NOR



(b) NAND

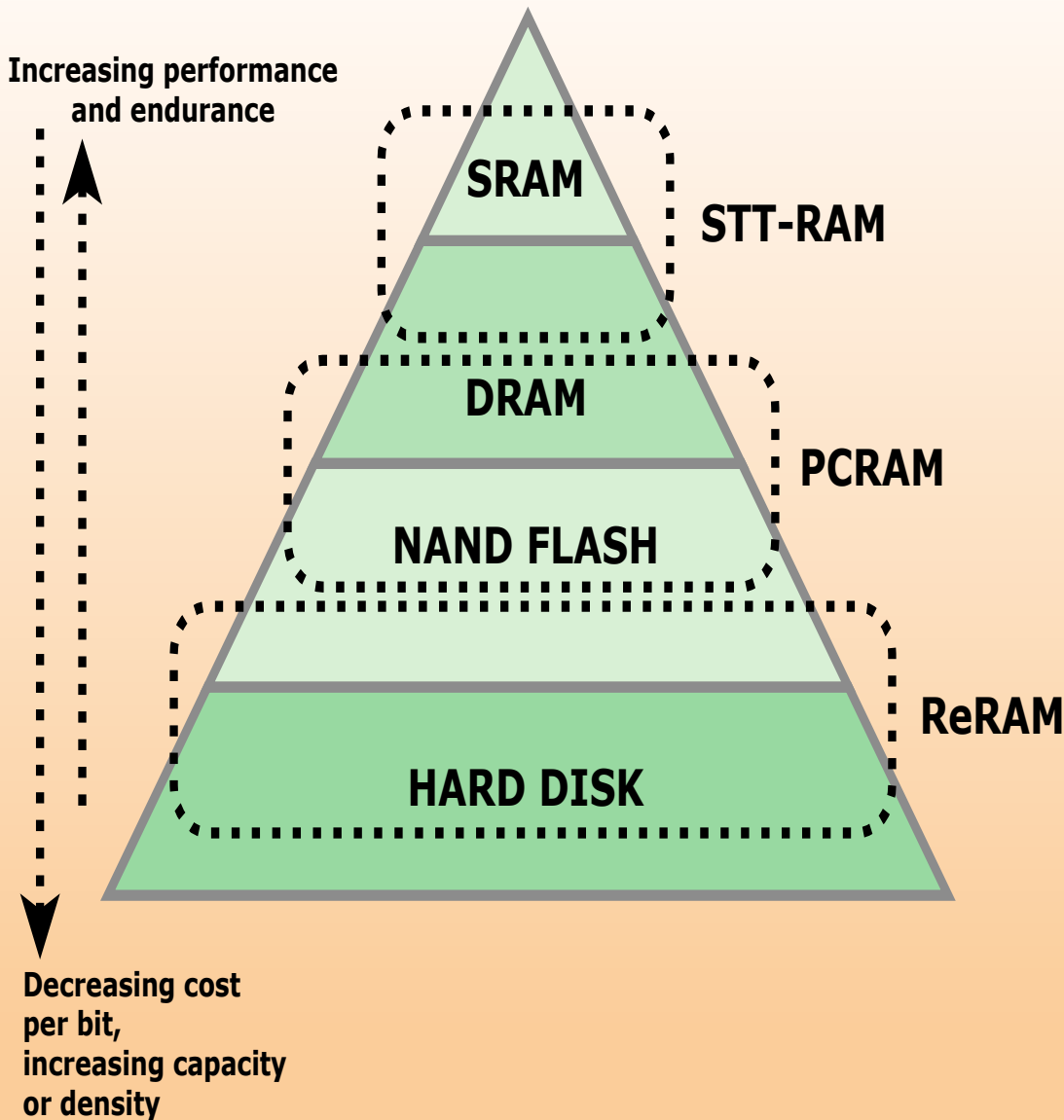
Figure 5.17 Kiviat Graphs for Flash Memory

-Although the specific quantitative values of various characteristics of NOR and NAND are changing year by year, the relative differences between the two types has remained stable. These differences are usefully illustrated by the Kiviat graphs shown in Figure 5.17.

-NOR flash memory provides high-speed random access. It can read and write data to specific locations, and can reference and retrieve a single byte. NAND reads and writes in small blocks. NAND provides higher bit density than NOR and greater write speed. NAND flash does not provide a random-access external address bus so the data must be read on a blockwise basis (also known as page access), where each block holds hundreds to thousands of bits

-For internal memory in embedded systems, NOR flash memory has traditionally been preferred. NAND memory has made some inroads, but NOR remains the dominant technology for internal memory. It is ideally suited for microcontrollers where the amount of program code is relatively small and a certain amount of application data does not vary. For example, the flash memory in Figure 1.16 is NOR memory.

-NAND memory is better suited for external memory, such as USB flash drives, memory cards (in digital cameras, MP3 players, etc.), and in what are known as solid-state disks (SSDs). We discuss SSDs in Chapter 6.



The traditional memory hierarchy has consisted of three levels (Figure 5.18):

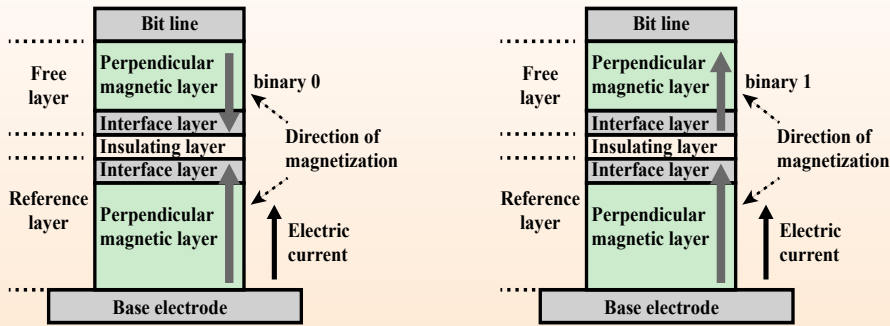
- Static RAM (SRAM): SRAM provides rapid access time, but is the most expensive and the least dense (bit density). SRAM is suitable for cache memory.

- Dynamic RAM (DRAM): Cheaper, denser, and slower than SRAM, DRAM has traditionally been the choice off-chip main memory.

- Hard disk: A magnetic disk provides very high bit density and very low cost per bit, with relatively slow access times. It is the traditional choice for external storage as part of the memory hierarchy.

Into this mix, as we have seen, **as been added flash memory. Flash memory has the advantage over traditional memory that it is nonvolatile.** NOR flash is best suited to storing programs and static application data in embedded systems, while NAND flash has characteristics intermediate between DRAM and hard disks.

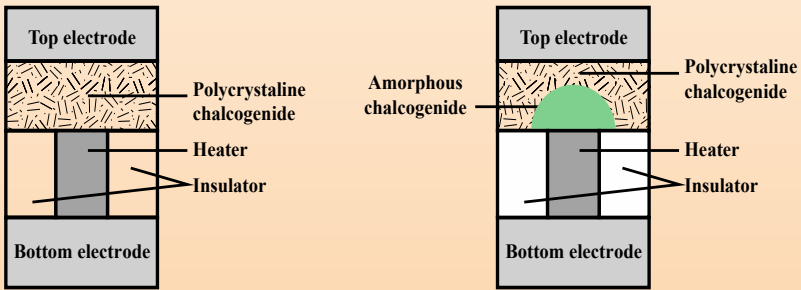
Figure 5.18 Nonvolatile RAM within the Memory Hierarchy



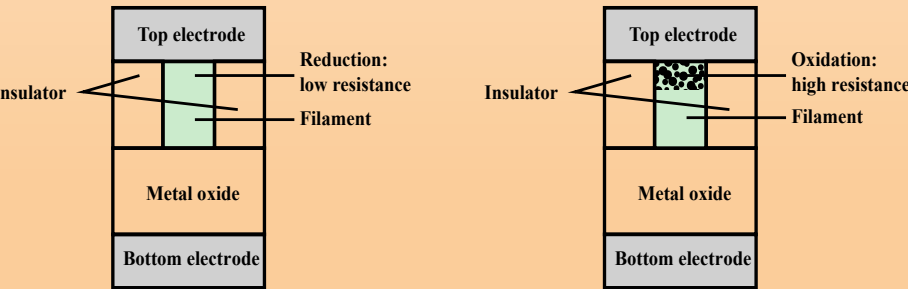
(a) STT-RAM

Recently, there have been breakthroughs in developing new forms of nonvolatile semiconductor memory that continue scaling beyond flash memory. The most promising technologies are spin-transfer torque RAM (STT-RAM), phase change RAM (PCRAM), and resistive RAM (ReRAM) ([ITRS14], [GOER12]).

All of these are in volume production. However, because NAND Flash and to some extent NOR Flash are still dominating the applications, these emerging memories have not yet fulfilled their original promise to become dominating mainstream high-density nonvolatile memory. This is likely to change in the next few years.



(b) PCRAM



(c) ReRAM

Figure 5.19 Nonvolatile RAM Technologies

+ Summary

Chapter 5

Internal Memory

- Semiconductor main memory
 - Organization
 - DRAM and SRAM
 - Types of ROM
 - Chip logic
 - Chip packaging
 - Module organization
 - Interleaved memory
- Error correction
- DDR DRAM
 - Synchronous DRAM
 - DDR SDRAM
- Flash memory
 - Operation
 - NOR and NAND flash memory
- Newer nonvolatile solid-state memory technologies