

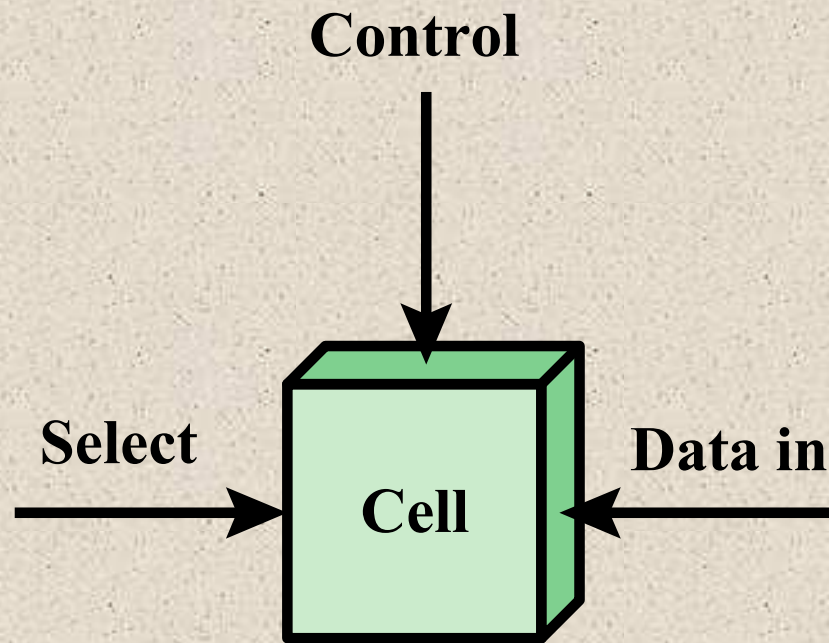
William Stallings Computer Organization and Architecture 10th Edition

© 2016 Pearson Education, Inc., Hoboken,
NJ. All rights reserved.

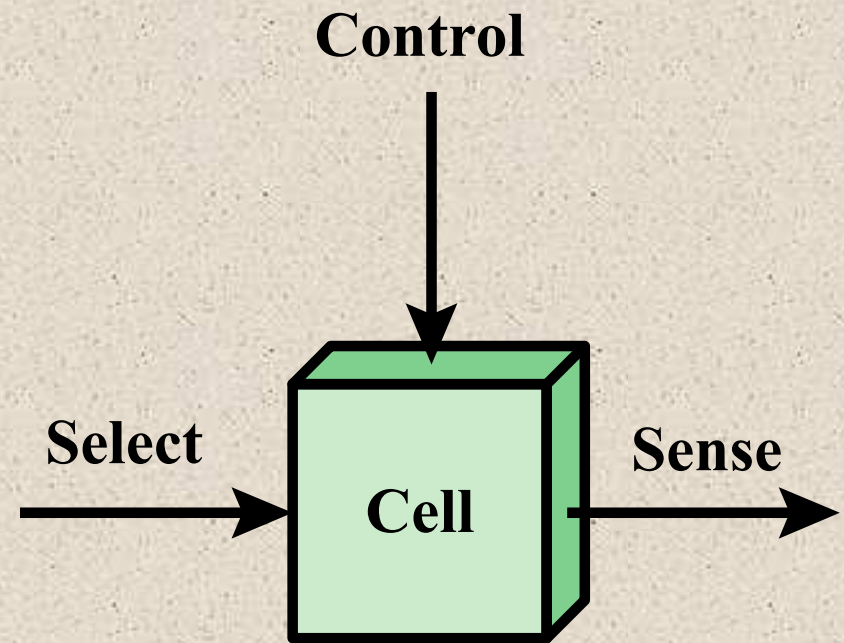


+ Chapter 5

Internal Memory



(a) Write



(b) Read

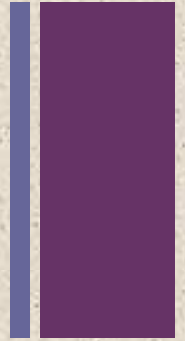
Figure 5.1 Memory Cell Operation

Memory Type	Category	Erasure	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)		UV light, chip-level	Electrically	
Electrically Erasable PROM (EEPROM)		Electrically, byte-level		
Flash memory		Electrically, block-level		

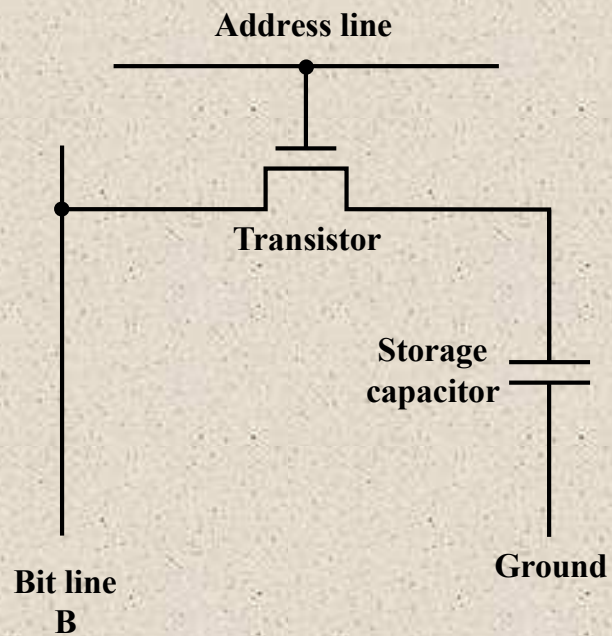
Table 5.1
Semiconductor Memory Types



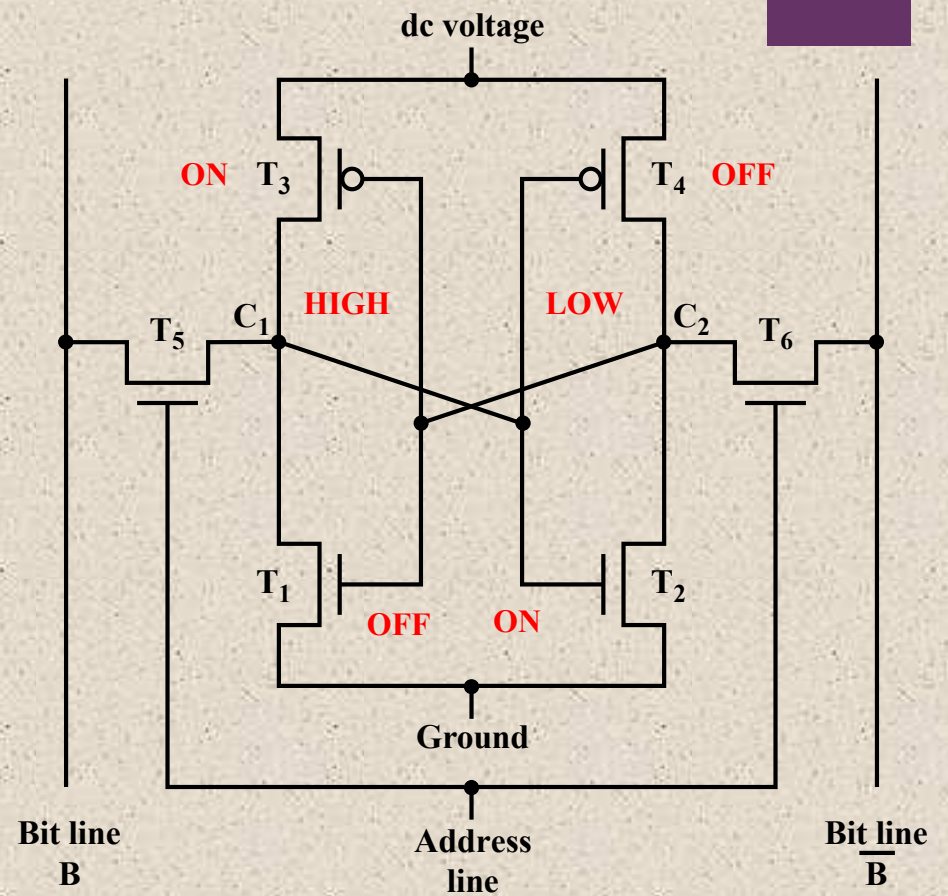
Dynamic RAM (DRAM)



- 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

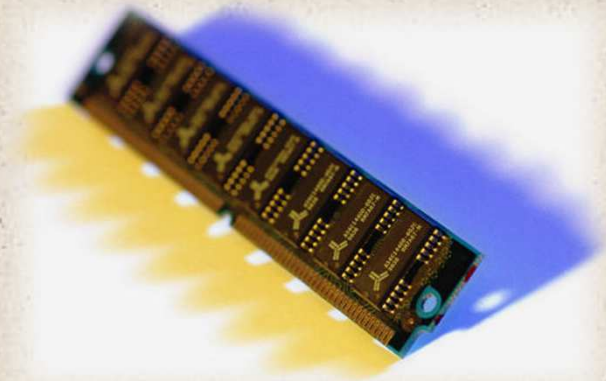
RED INDICATES STATE WHEN A BINARY "1"

Figure 5.2 Typical Memory Cell Structures



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

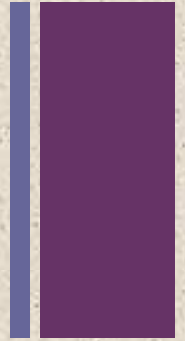
- Both volatile
 - Power must be continuously supplied to the memory to preserve the bit values
- Dynamic cell
 - Simpler to build, smaller
 - More dense (smaller cells = more cells per unit area)
 - Less expensive
 - Requires the supporting refresh circuitry
 - Tend to be favored for large memory requirements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

SRAM

DRAM



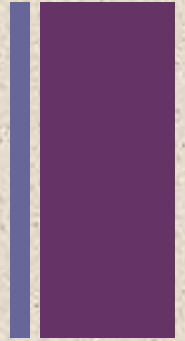
Read Only Memory (ROM)



- 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

Erase 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

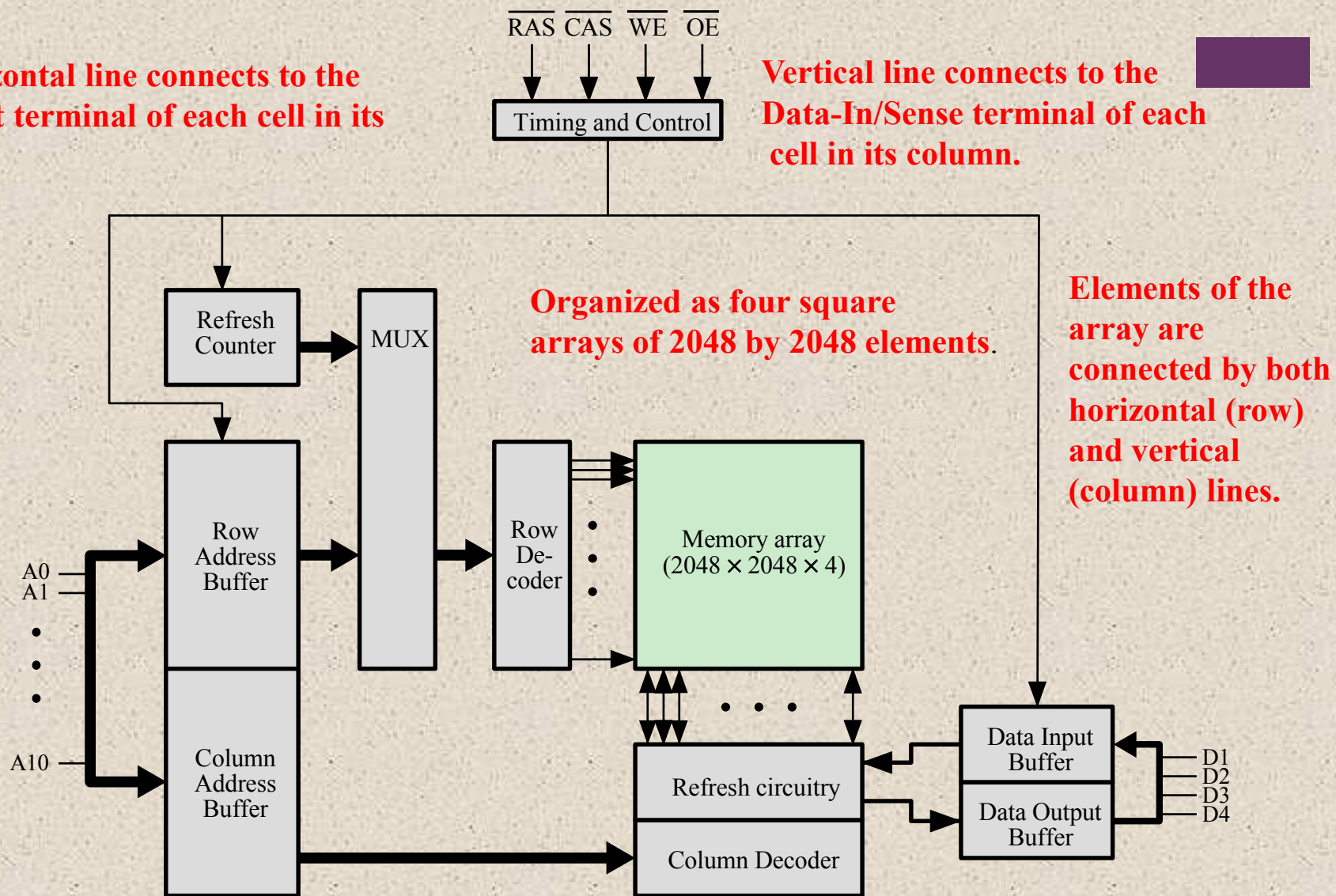
Intermediate between EPROM and EEPROM in both cost and functionality

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”

Horizontal line connects to the Select terminal of each cell in its row

Vertical line connects to the Data-In/Sense terminal of each cell in its column.



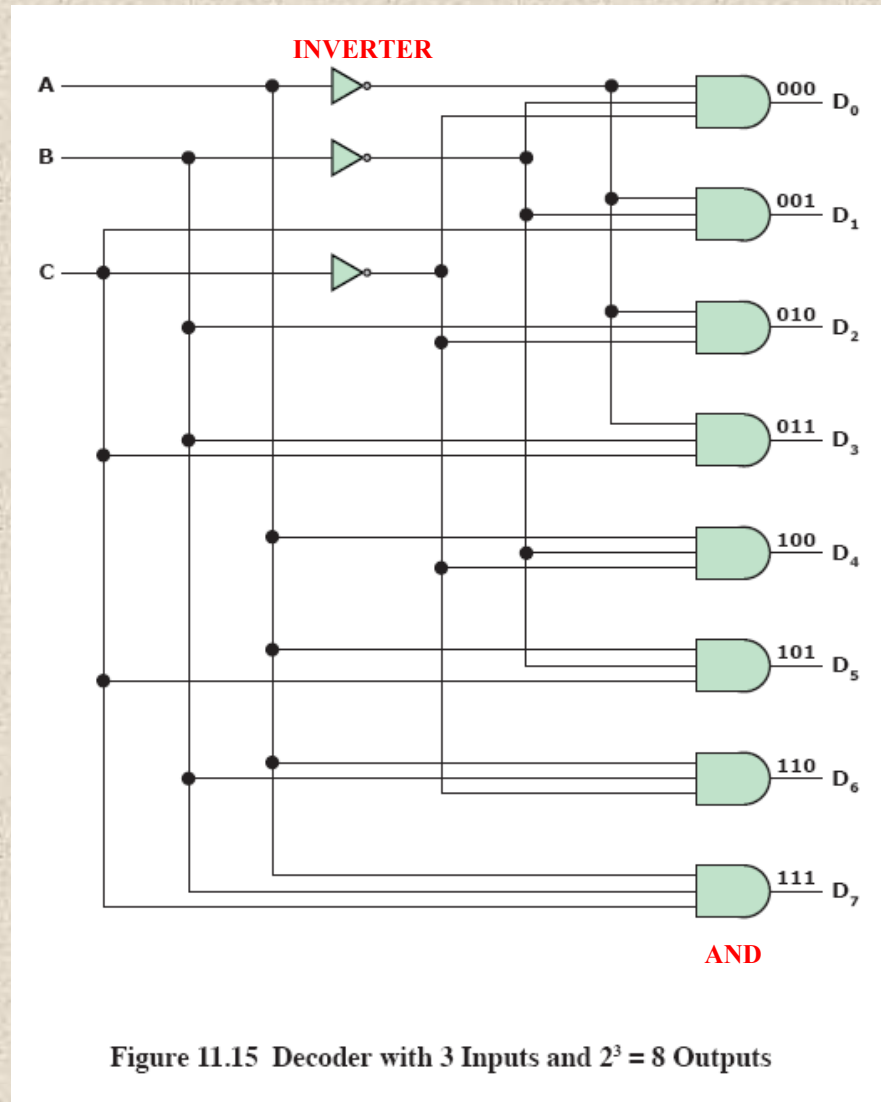
Organized as four square arrays of 2048 by 2048 elements.

Elements of the array are connected by both horizontal (row) and vertical (column) lines.

Refresh involves stepping through each row, reading the cells with RAS and then writing them right back.

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

3-to-8 Decoder



4-Bit Multiplexer

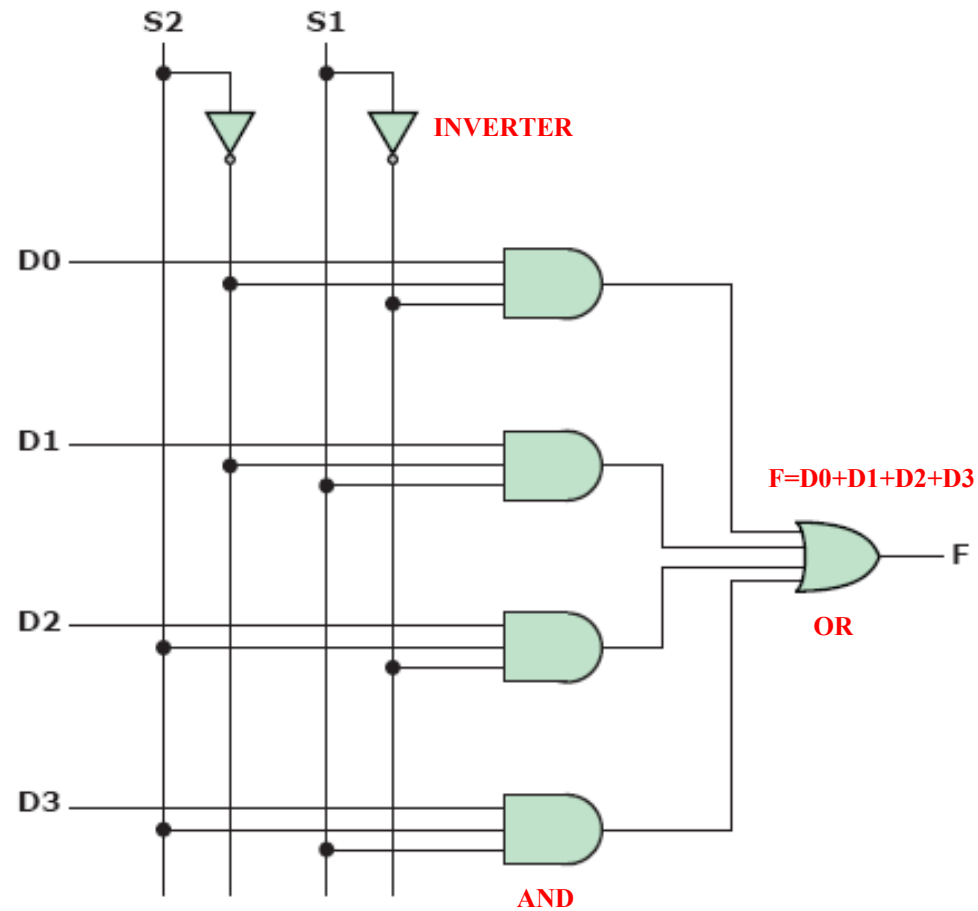
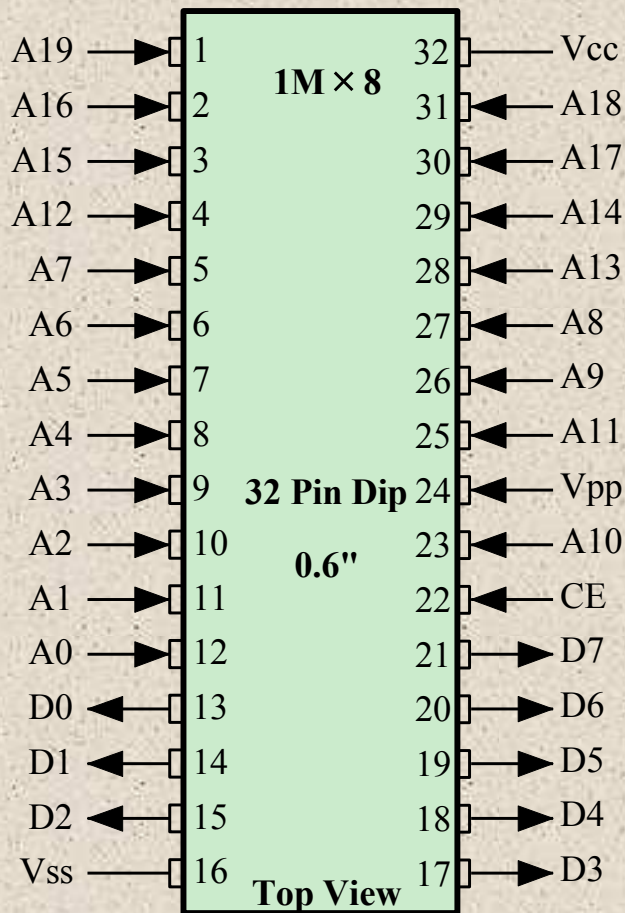
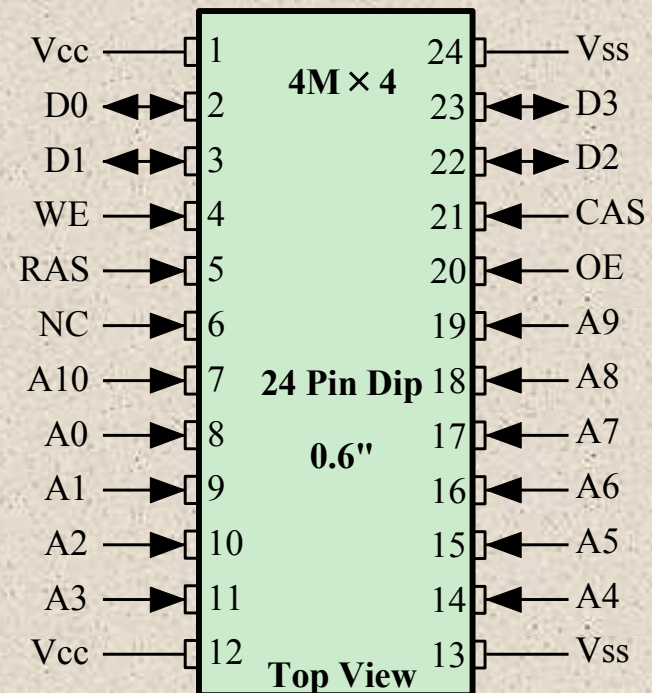


Figure 11.13 Multiplexer Implementation



(a) 8 Mbit EPROM



(b) 16 Mbit DRAM

Figure 5.4 Typical Memory Package Pins and Signals

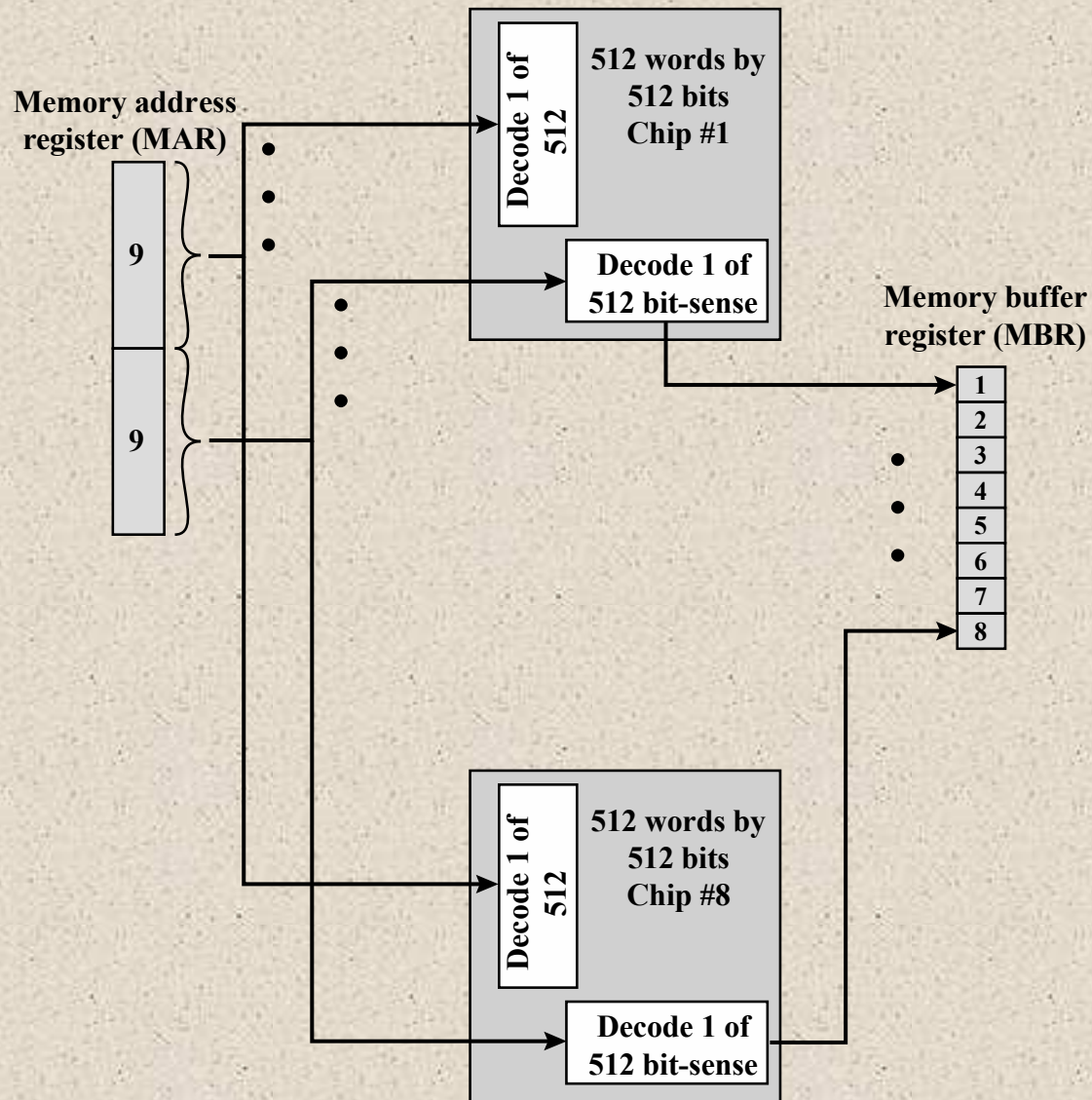


Figure 5.5 256-KByte Memory Organization

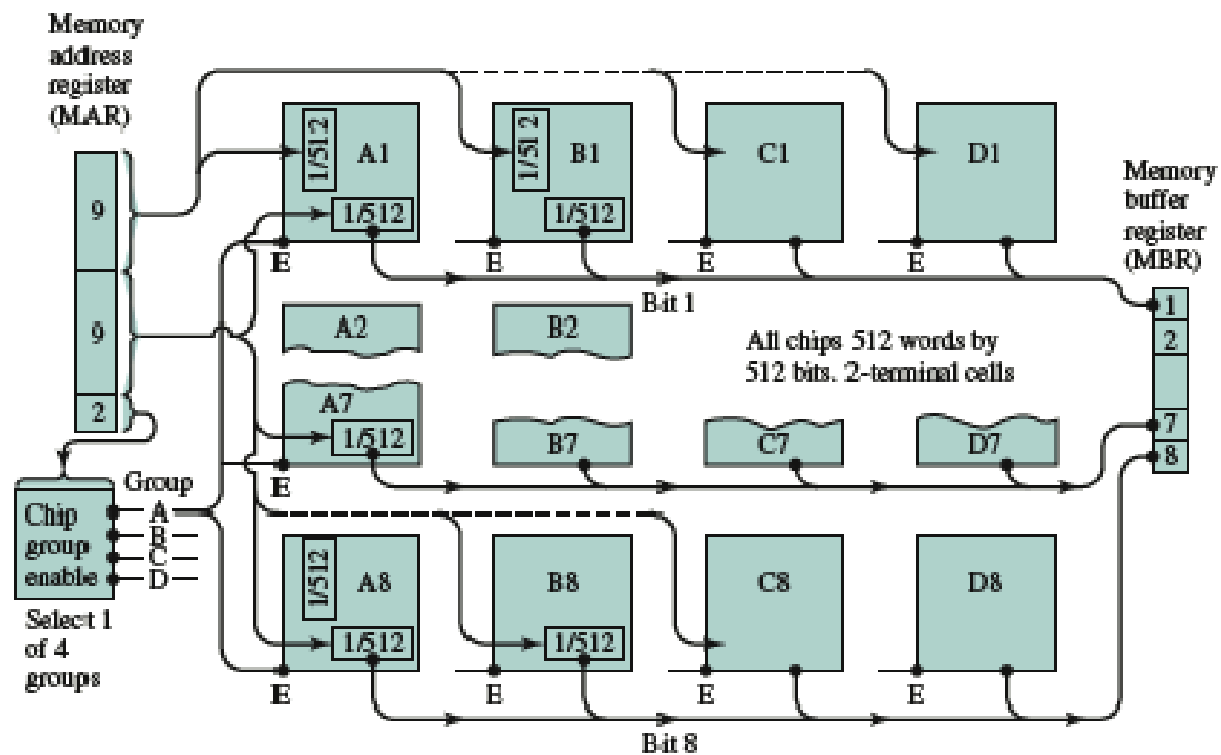
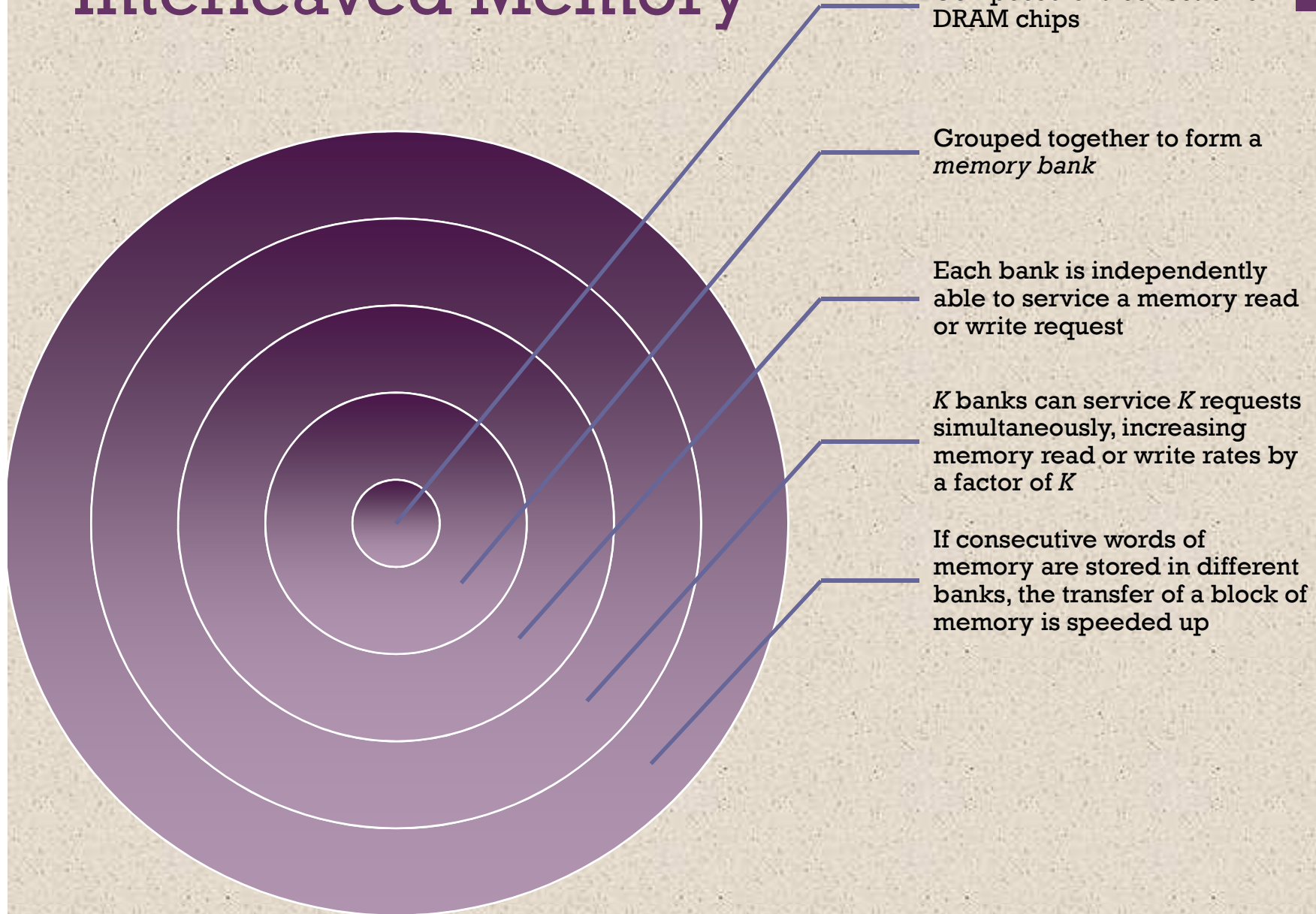


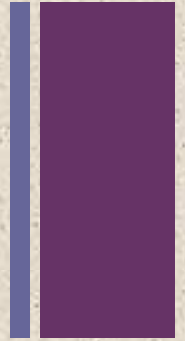
Figure 5.6 1-Mbyte Memory Organization

Interleaved Memory





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

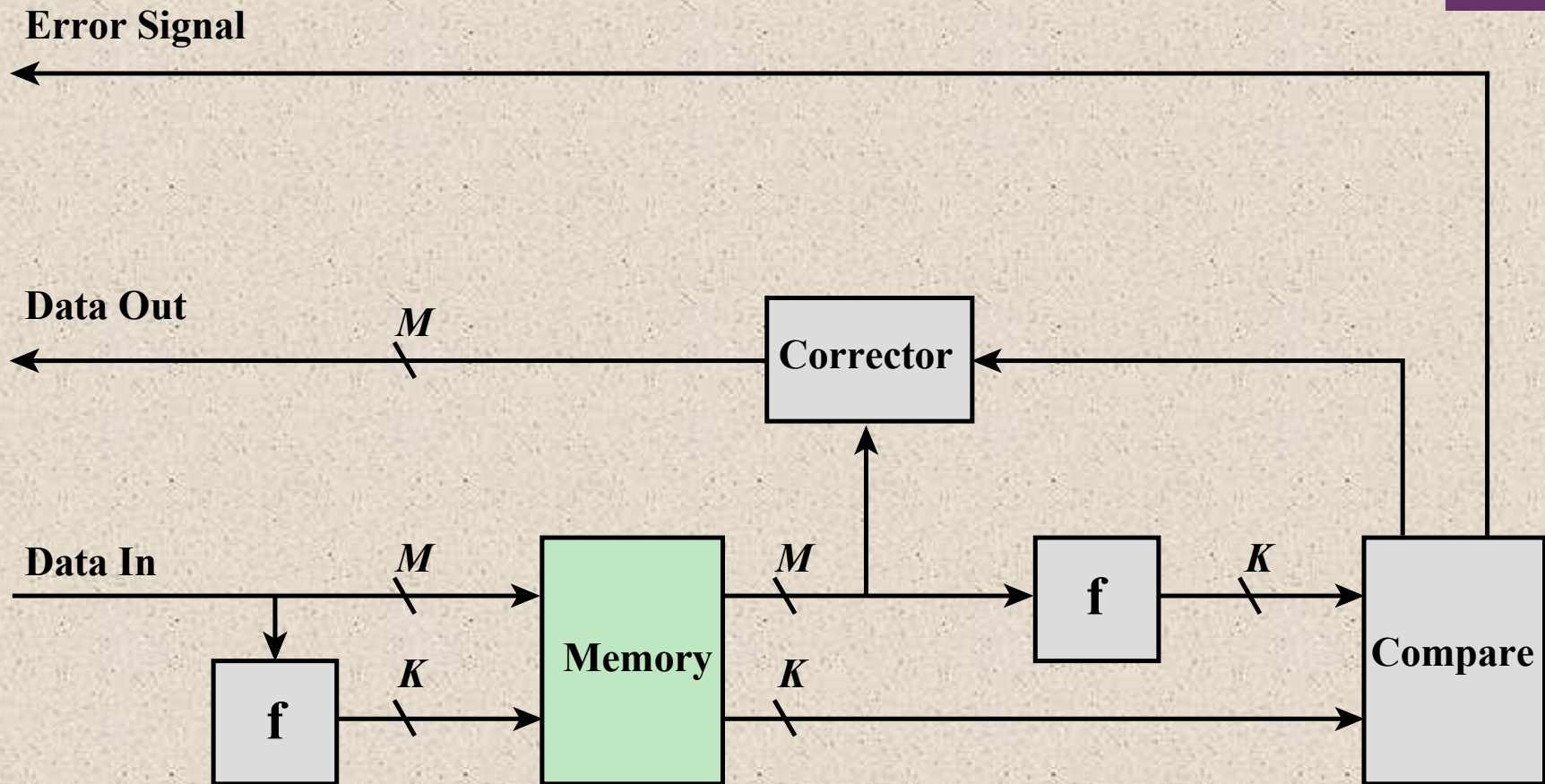


Figure 5.7 Error-Correcting Code Function

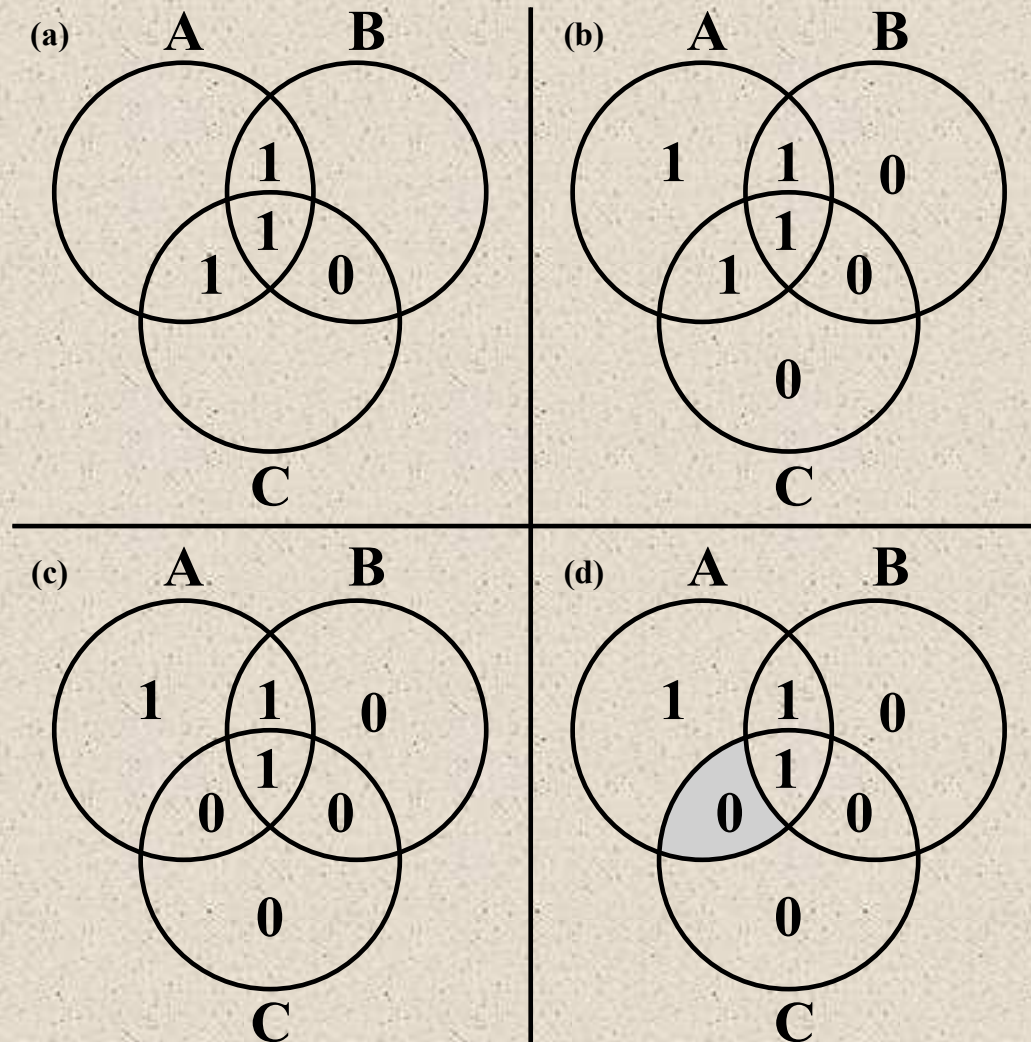


Figure 5.8 Hamming Error-Correcting Code



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



Bit Position	12	11	10	9	8	7	6	5	4	3	2	1
Position Number	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001
Data Bit	D8	D7	D6	D5		D4	D3	D2		D1		
Check Bit					C8				C4		C2	C1

Figure 5.9 Layout of Data Bits and Check Bits

Hamming Codes - SEC

$$C1 = D1 \oplus D2 \oplus D4 \oplus D5 \oplus D7$$

$$C2 = D1 \oplus D3 \oplus D4 \oplus D6 \oplus D7$$

$$C4 = D2 \oplus D3 \oplus D4 \oplus D8$$

$$C8 = D5 \oplus D6 \oplus D7 \oplus D8$$

Bit position	12	11	10	9	8	7	6	5	4	3	2	1
Position number	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001
Data bit	D8	D7	D6	D5		D4	D3	D2		D1		
Check bit					C8				C4		C2	C1
Word stored as	0	0	1	1	0	1	0	0	1	1	1	1
Word fetched as	0	0	1	1	0	1	1	0	1	1	1	1
Position Number	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001
Check Bit					0				0		0	1

Figure 5.10 Check Bit Calculation

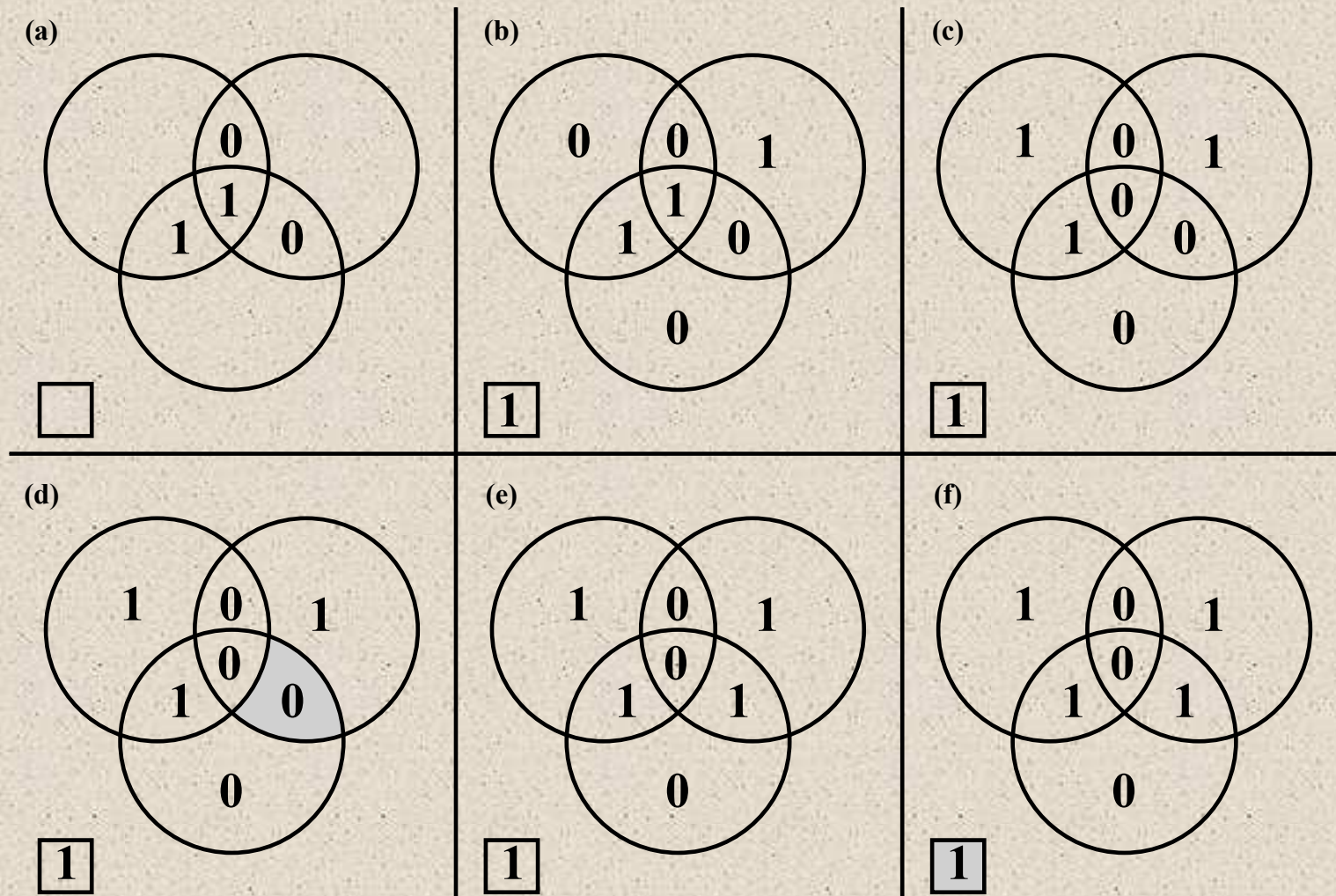


Figure 5.11 Hamming SEC-DED Code

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

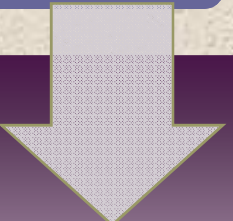
SDRAM

DDR-DRAM

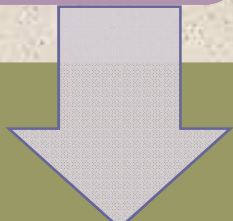
RDRAM

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

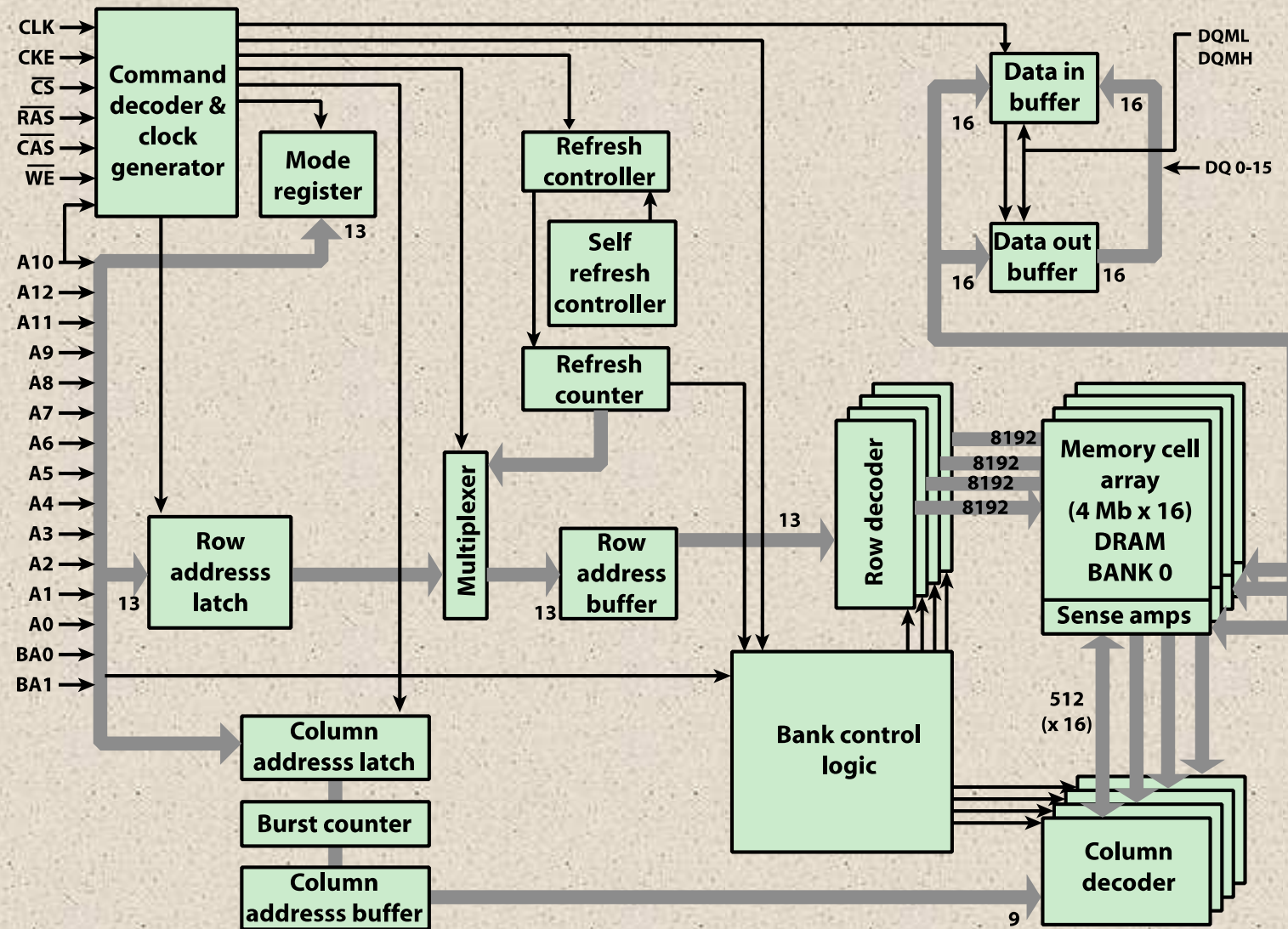


Figure 5.12 256-Mb Synchronous Dynamic RAM (SDRAM)

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
Assignment
s

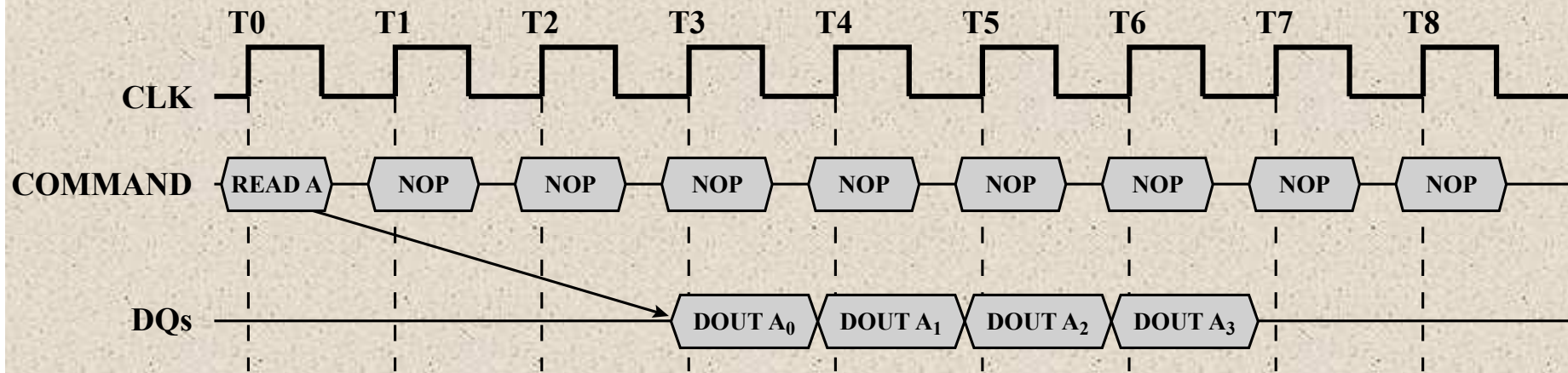
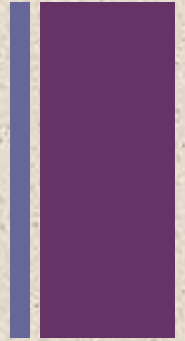


Figure 5.13 SDRAM Read Timing (Burst Length = 4, $\overline{\text{CAS}}$ latency = 2)



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

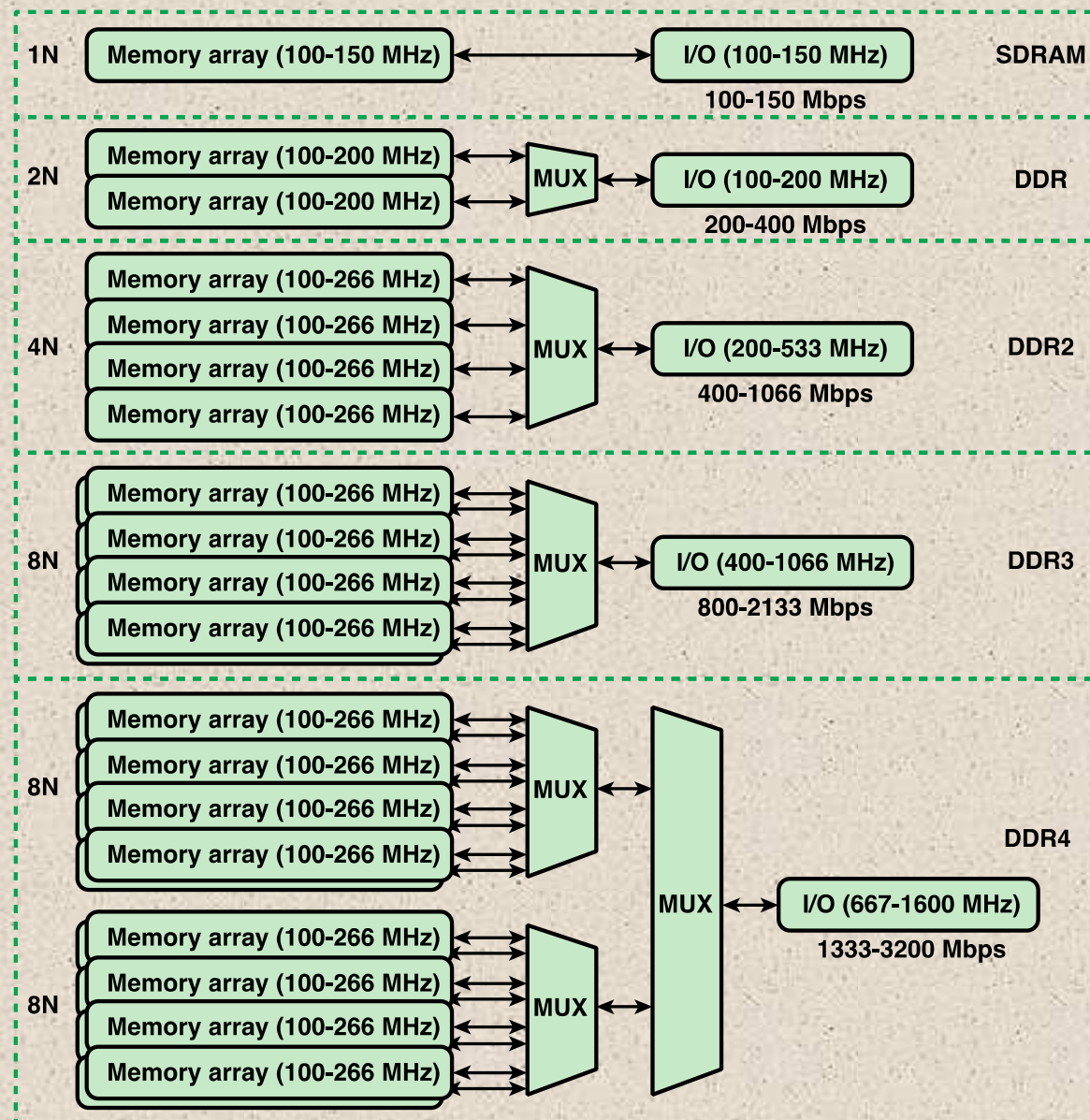


Figure 5.14 DDR Generations

SDR SDRAM (Single Data Rate synchronous DRAM)

This type of SDRAM is slower than the DDR variants, because only one word of data is transmitted per clock cycle (single data rate). But this type is also faster than its predecessors **EDO-RAM** and **FPM-RAM** which took typically 2 or 3 clocks to transfer one word of data.

DDR(1) SDRAM

While the access latency of DRAM is fundamentally limited by the DRAM array, DRAM has very high potential bandwidth because each internal read is actually a row of many thousands of bits. To make more of this bandwidth available to users, a double data rate interface was developed. This uses the same commands, accepted once per cycle, but reads or writes two words of data per clock cycle. The DDR interface accomplishes this by reading and writing data on both the rising and falling edges of the clock signal. In addition, some minor changes to the SDR interface timing were made in hindsight, and the supply voltage was reduced from 3.3 to 2.5 V. As a result, DDR SDRAM is not backwards compatible with SDR SDRAM.

DDR SDRAM (sometimes called *DDR1* for greater clarity) doubles the minimum read or write unit; every access refers to at least two consecutive words.

Typical DDR SDRAM clock rates are 133, 166 and 200 MHz (7.5, 6, and 5 ns/cycle), generally described as DDR-266, DDR-333 and DDR-400 (3.75, 3, and 2.5 ns per beat). Corresponding 184-pin DIMMs are known as PC-2100, PC-2700 and PC-3200. Performance up to DDR-550 (PC-4400) is available for a price.

DDR2 SDRAM

DDR2 SDRAM is very similar to DDR SDRAM, but doubles the minimum read or write unit again, to 4 consecutive words. The bus protocol was also simplified to allow higher performance operation. (In particular, the "burst terminate" command is deleted.) This allows the bus rate of the SDRAM to be doubled without increasing the clock rate of internal RAM operations; instead, internal operations are performed in units 4 times as wide as SDRAM. Also, an extra bank address pin (BA2) was added to allow 8 banks on large RAM chips.

Typical DDR2 SDRAM clock rates are 200, 266, 333 or 400 MHz (periods of 5, 3.75, 3 and 2.5 ns), generally described as DDR2-400, DDR2-533, DDR2-667 and DDR2-800 (periods of 2.5, 1.875, 1.5 and 1.25 ns). Corresponding 240-pin DIMMS are known as PC2-3200 through PC2-6400. DDR2 SDRAM is now available at a clock rate of 533 MHz generally described as DDR2-1066 and the corresponding DIMMs are known as PC2-8500 (also named PC2-8600 depending on the manufacturer). Performance up to DDR2-1250 (PC2-10000) is available for a price.

Note that because internal operations are at 1/2 the clock rate, DDR2-400 memory (internal clock rate 100 MHz) has somewhat higher latency than DDR-400 (internal clock rate 200 MHz).

DDR3 SDRAM

DDR3 continues the trend, doubling the minimum read or write unit to 8 consecutive words. This allows another doubling of bandwidth and external bus rate without having to change the clock rate of internal operations, just the width. To maintain 800–1600 M transfers/s (both edges of a 400–800 MHz clock), the internal RAM array has to perform 100–200 M fetches per second.

Again, with every doubling, the downside is the increased **latency**. As with all DDR SDRAM generations, commands are still restricted to one clock edge and command latencies are given in terms of clock cycles, which are half the speed of the usually quoted transfer rate (a **CAS latency** of 8 with DDR3-800 is $8/(400 \text{ MHz}) = 20 \text{ ns}$, exactly the same latency of CAS2 on **PC100** SDR SDRAM).

DDR3 memory chips are being made commercially,^[4] and computer systems using them were available from the second half of 2007,^[5] with significant usage from 2008 onwards.^[6] Initial clock rates were 400 and 533 MHz, which are described as DDR3-800 and DDR3-1066 (PC3-6400 and PC3-8500 modules), but 667 and 800 MHz, described as DDR3-1333 and DDR3-1600 (PC3-10600 and PC3-12800 modules) are now common.^[7] Performance up to DDR3-2800 (PC3 22400 modules) are available for a price.^[8]

DDR4 SDRAM

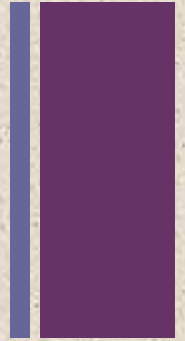
DDR4 SDRAM is the successor to **DDR3 SDRAM**. It was revealed at the **Intel Developer Forum** in **San Francisco** in 2008, and was due to be released to market during 2011. The timing has varied considerably during its development - it was originally expected to be released in 2012,^[9] and later (during 2010) expected to be released in 2015,^[10] before samples were announced in early 2011 and manufacturers began to announce that commercial production and release to market was anticipated in 2012. DDR4 is expected to reach mass market adoption around 2015, which is comparable with the approximately 5 years taken for DDR3 to achieve mass market transition over DDR2.

The new chips are expected to run at 1.2 V or less,^{[11][12]} versus the 1.5 V of DDR3 chips, and have in excess of 2 billion **data transfers** per second. They are expected to be introduced at frequency rates of 2133 MHz, estimated to rise to a potential 4266 MHz^[13] and lowered voltage of 1.05 V^[14] by 2013.

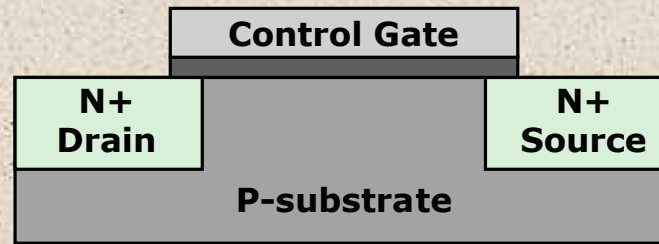
DDR4 will *not* double the internal prefetch width again, but will use the same 8n prefetch as DDR3.^[15] Thus, it will be necessary to interleave reads from several banks to keep the data bus busy.

In February 2009, **Samsung** validated 40 nm DRAM chips, considered a "significant step" towards DDR4 development^[16] since as of 2009, current DRAM chips were only beginning to migrate to a 50 nm process.^[17] In January 2011, **Samsung** announced the completion and release for testing of a 30 nm 2 GB DDR4 DRAM module. It has a maximum bandwidth of 2.13 Gbit/s at 1.2 V, uses pseudo open drain technology and draws 40% less power than an equivalent DDR3 module.^{[18][19]}

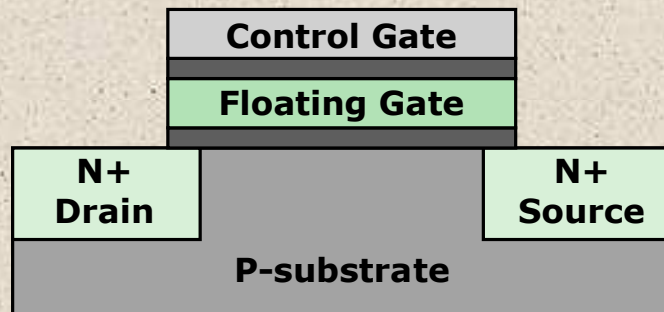
+ Flash Memory



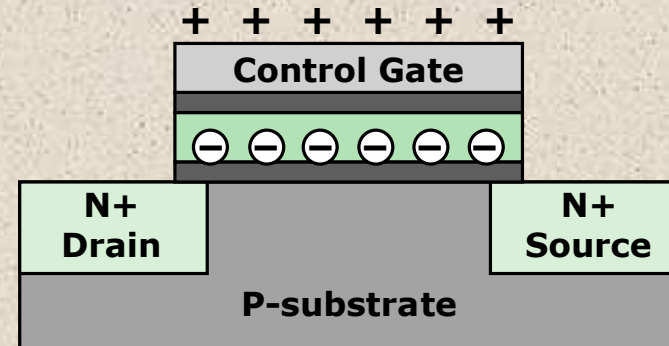
- Used both for internal memory and external memory applications
- First introduced in the mid-1980's
- Is intermediate between EPROM and EEPROM in both cost and functionality
- Uses an electrical erasing technology like EEPROM
- It is possible to erase just blocks of memory rather than an entire chip
- 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

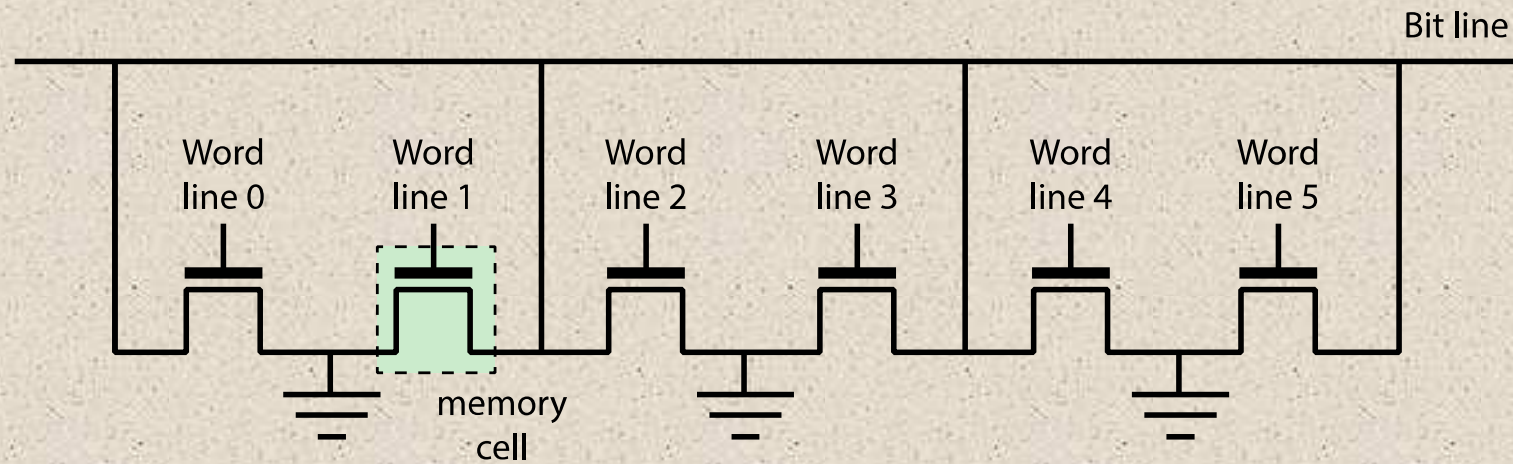


(b) Flash memory cell in one state

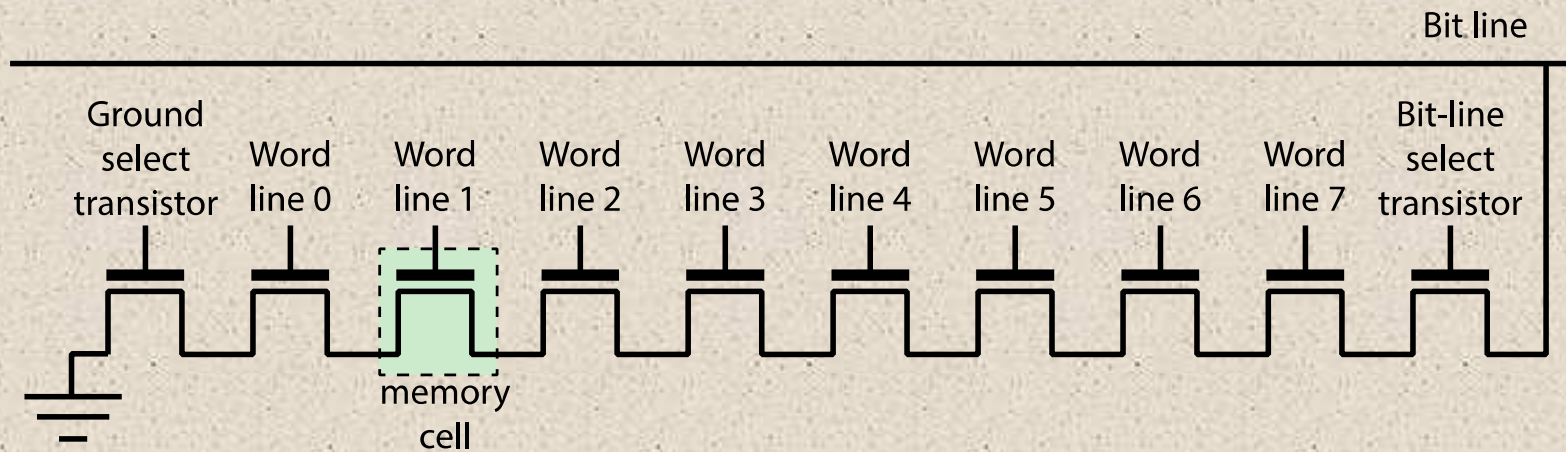


(c) Flash memory cell in zero state

Figure 5.15 Flash Memory Operation

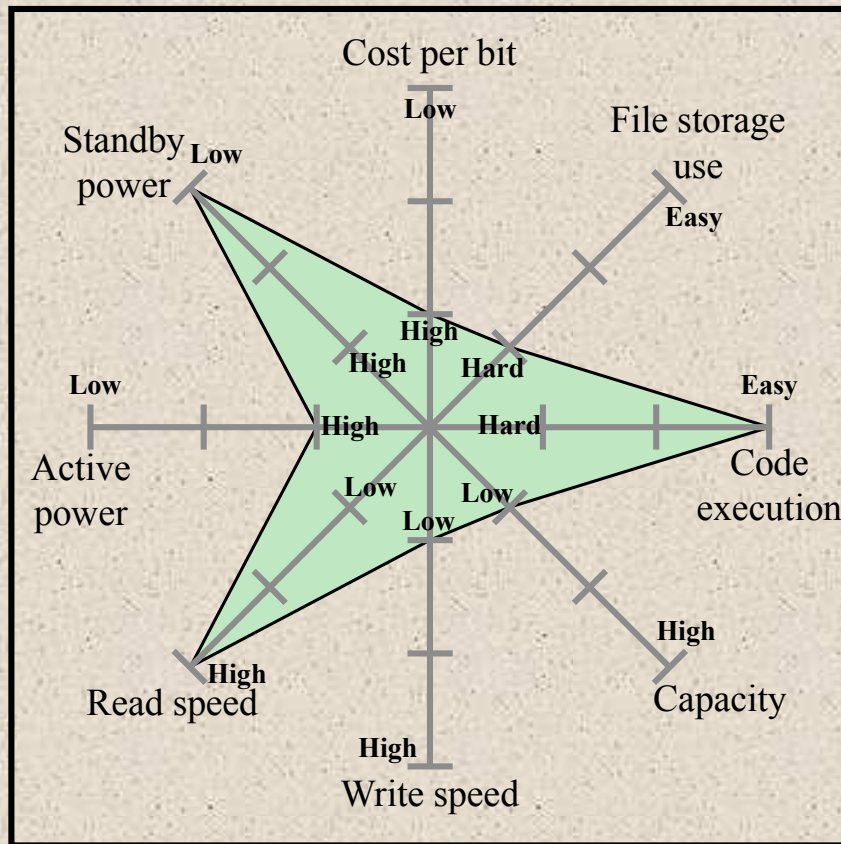


(a) NOR flash structure

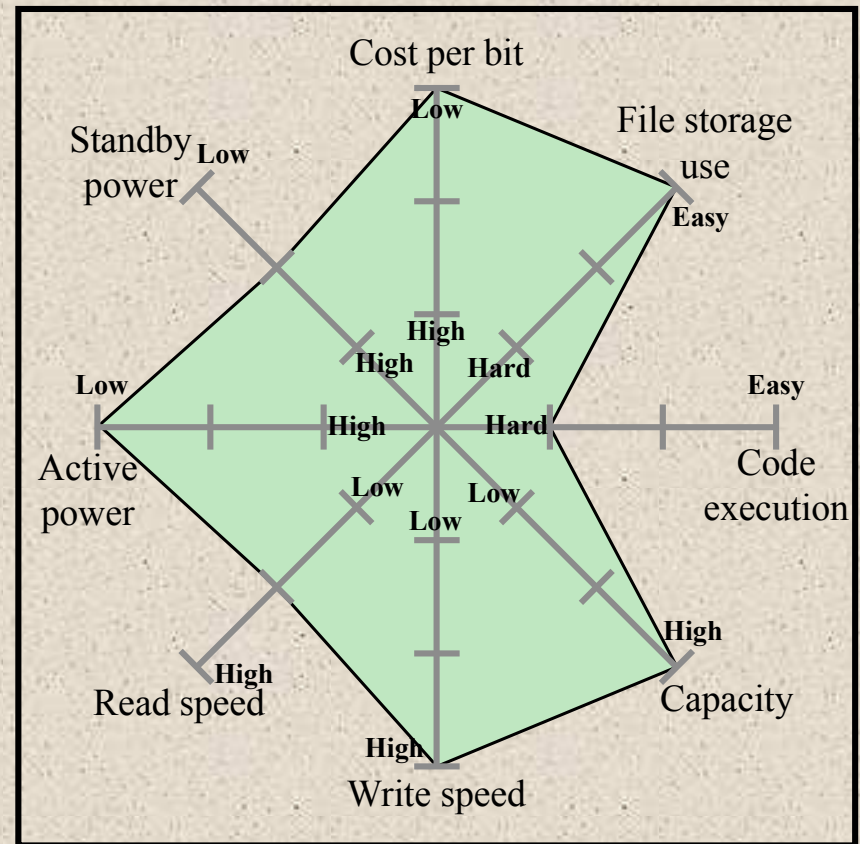


(b) NAND flash structure

Figure 5.16 Flash Memory Structures



(a) NOR



(b) NAND

Figure 5.17 Kiviat Graphs for Flash Memory

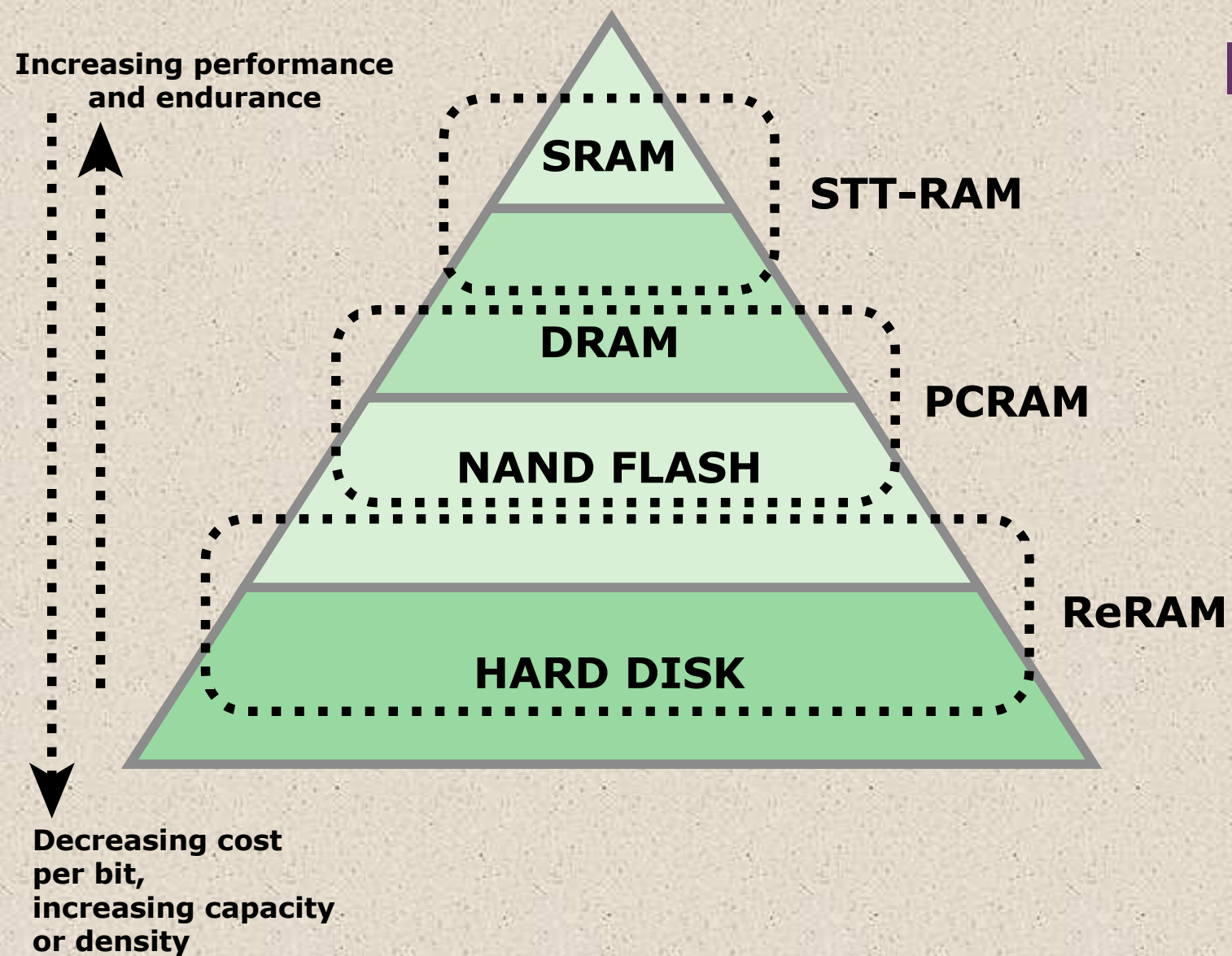
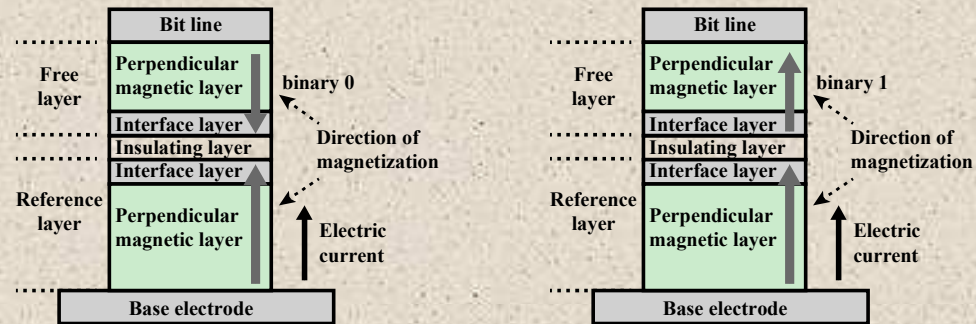
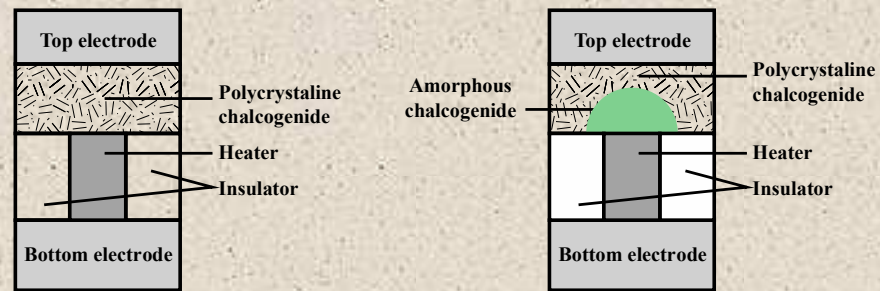


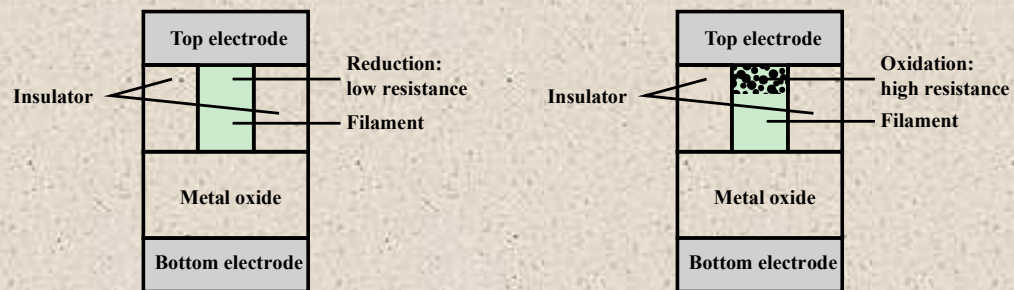
Figure 5.18 Nonvolatile RAM within the Memory Hierarchy



(a) STT-RAM



(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