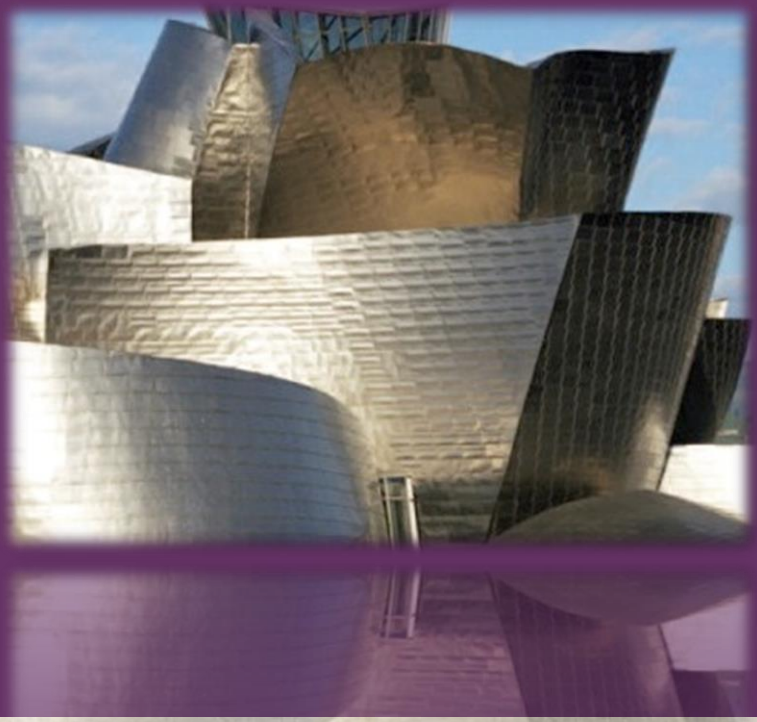
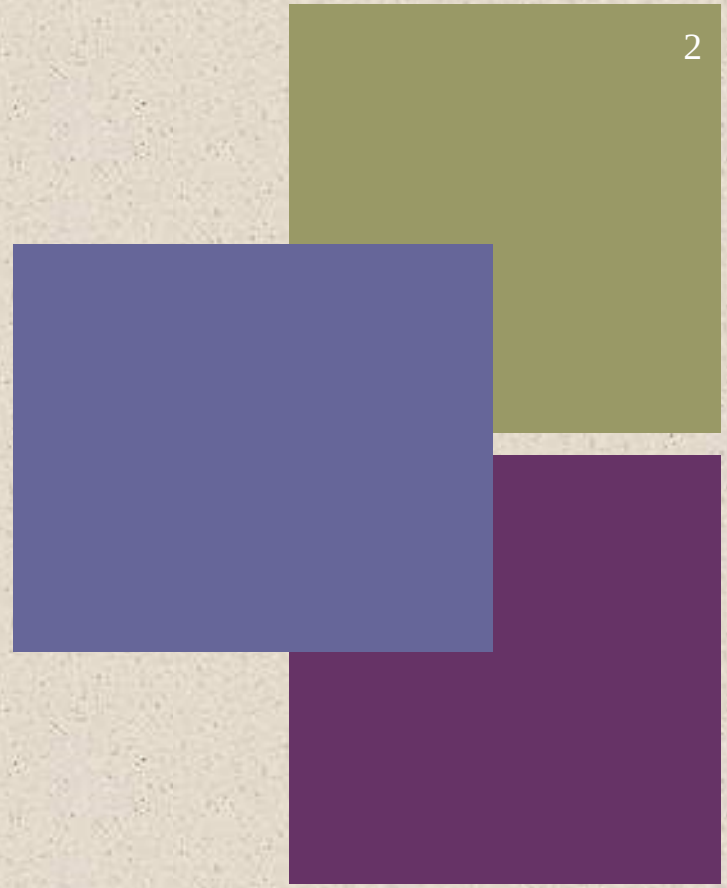


William Stallings Computer Organization and Architecture 9th Edition



Chapter 5

Internal Memory

Earlier, the most common form of random-access storage for computer main memory employed an array of doughnut-shaped ferromagnetic loops referred to as *cores.*

The advent of, and advantages of, microelectronics has long since outdone the magnetic core memory.

Today, the use of semiconductor chips for main memory is almost universal.

Memory Cell Operation

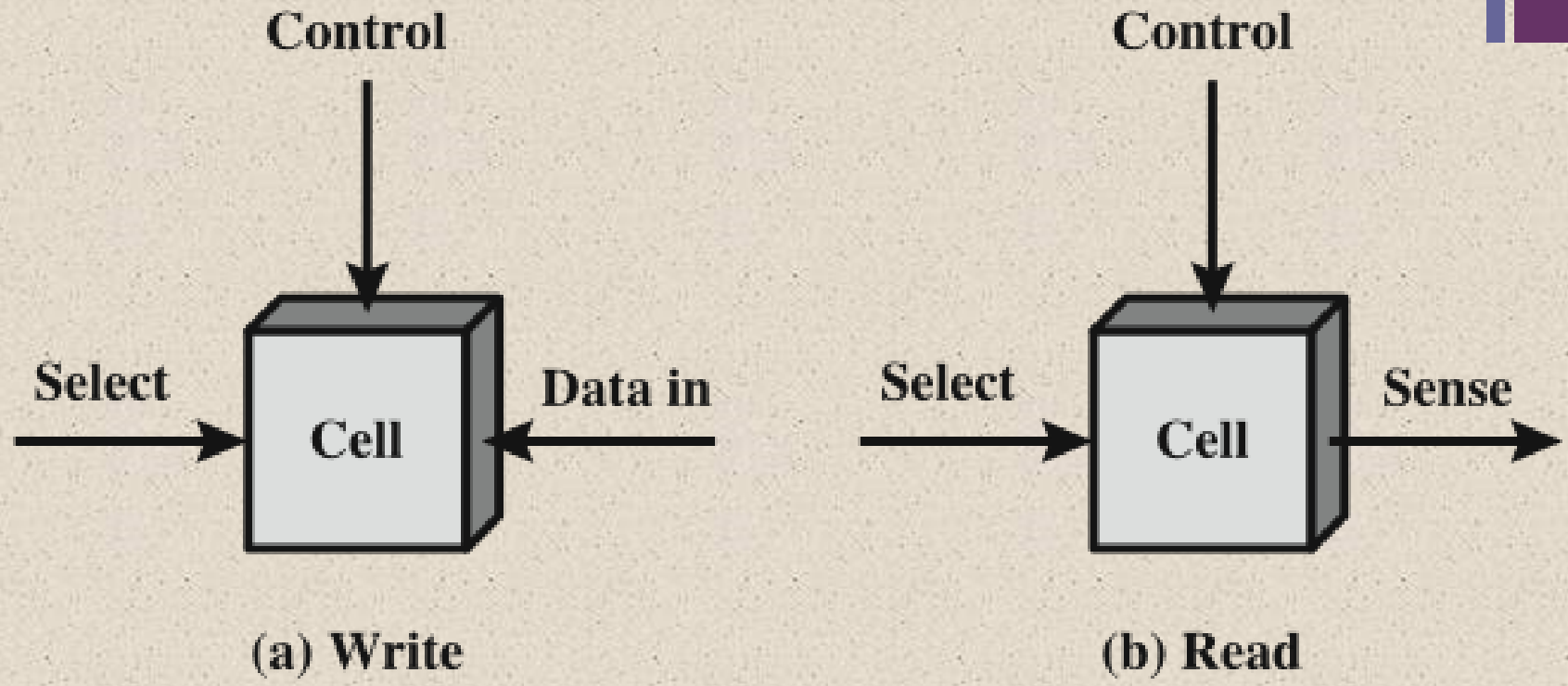


Figure 5.1 Memory Cell Operation

The basic element of a **semiconductor memory** is the memory cell. Few properties:

- They exhibit two stable (or semi-stable) 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.

Mostly, the cell has three functional terminals capable of carrying an electrical signal.

The select terminal, 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.

Semiconductor Memory Types

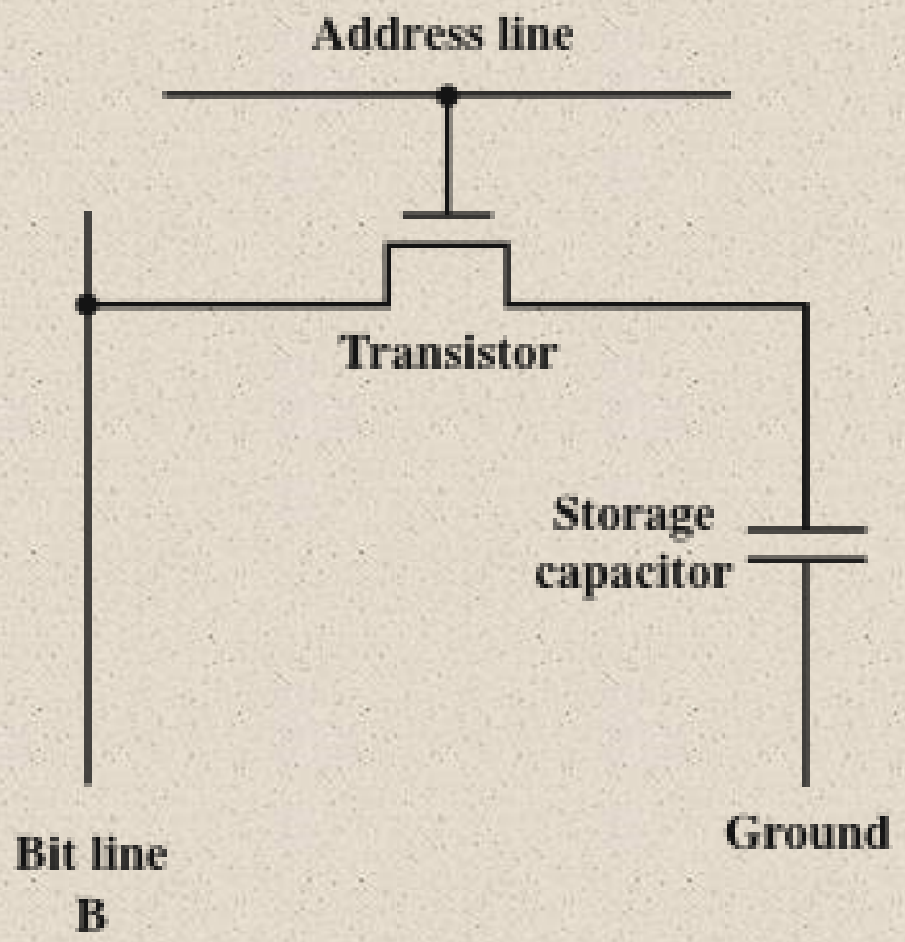
Memory Type	Category	Erasure	Write Mechanism	Volatility
Random-access memory (RAM)	Read-write memory	Electrically, byte-level	Electrically	Volatile
Read-only memory (ROM)	Read-only memory	Not possible	Masks	Nonvolatile
Programmable ROM (PROM)				
Erasable PROM (EPROM)		UV light, chip-level		
Electrically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

Dynamic RAM Structure



(a) Dynamic RAM (DRAM) cell

Figure 5.2a

Typical Memory Cell Structures

Address line is activated when the bit value fro this cell is to be read or written.

Transistor acts as a switch that is closed (allowing current to flow) if a voltage is applied to the address line and open (no current flows) if no voltage is present on the address line.

For the write operation, a voltage signal is applied to the bit line; (high voltage: 1 and low voltage: 0)

A signal is then applied to the address line, allowing a charge to be transferred to the capacitor.

For read operation, when the address line is selected, transistor turns on and charge stored on the capacitor is fed out onto a bit line and to a sense amplifier.

Sense amplifier compares the capacitor voltage to a reference value and determines if the cell contains a logic 1/0.

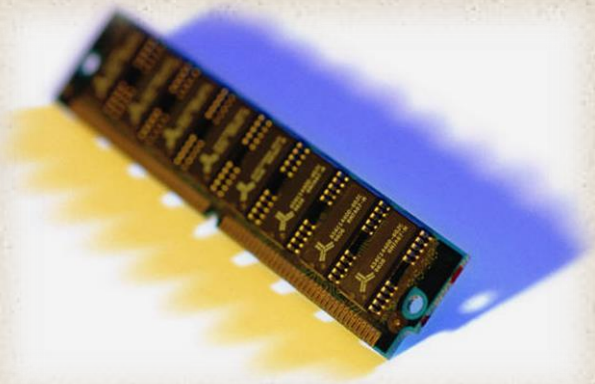
The readout from the cell discharges the capacitor, which must be restored to complete the operation.

Although DRAM cell is used to store a single bit, it is essentially an analog device.

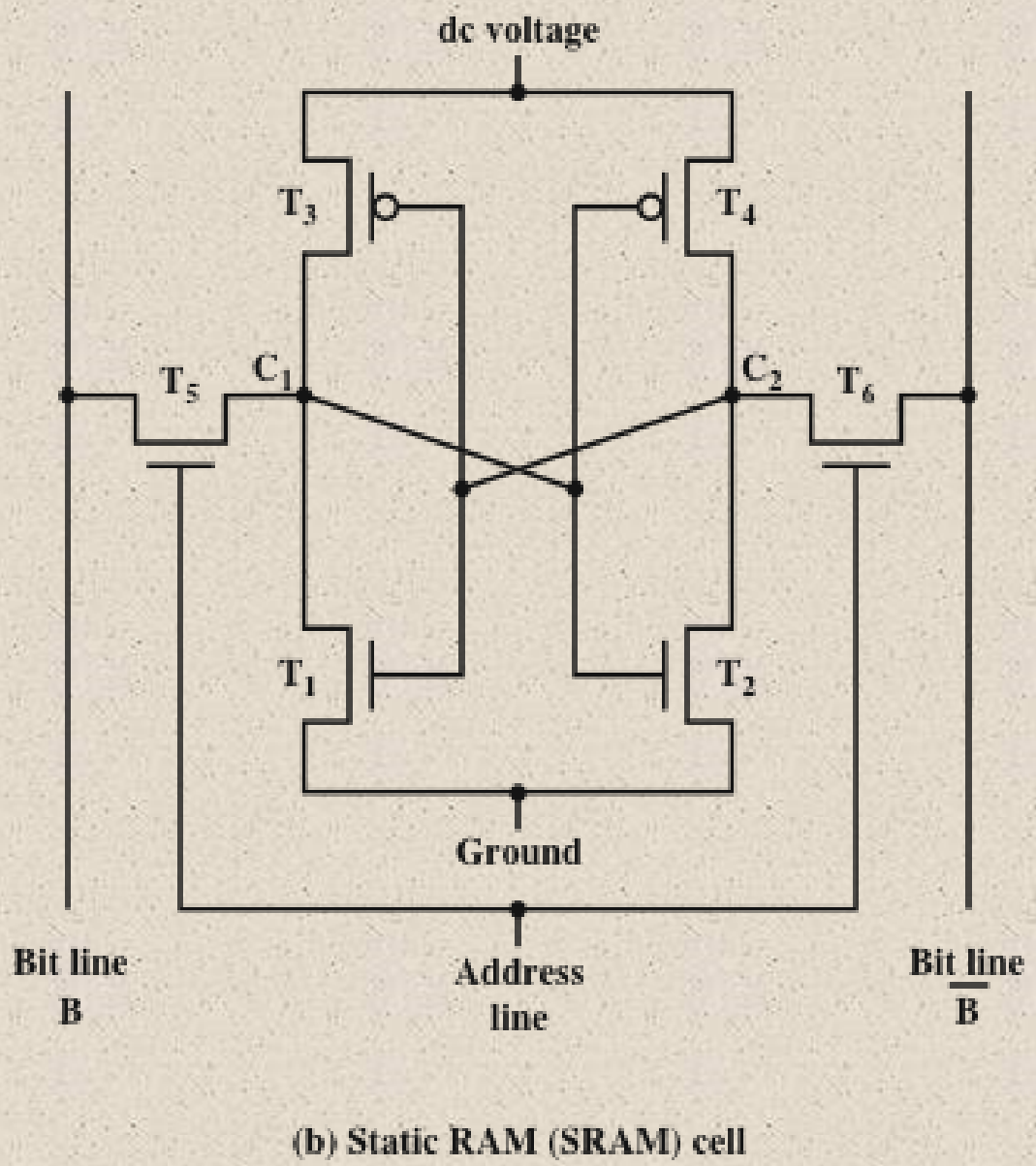
Capacitor can store any charge value within a range; a threshold value determines whether the charge is 1/0.

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



Static RAM Structure



(b) Static RAM (SRAM) cell

Figure 5.2b

Typical Memory Cell Structures

CMOS inverters: T_3 and T_4 (store the data)
Access transistors: T_1 and T_2

Logic state 1:
C_1 is high and point C_2 is low;
T_1 and T_4 are off and T_2 and T_3 are on.

Logic state 0:
C_1 is low and point C_2 is high;
T_1 and T_4 are on and T_2 and T_3 are off.

States are stable as the DC voltage is applied.
No refresh is needed to retain data.

Address line is used to open or close a switch. It controls two transistors (T_5 and T_6).

When a signal is applied to this line, the two transistors are switched on, allowing a read or write operation.

For a write operation, desired bit value is applied to line B, while its complement is applied to line B'.

This forces the four transistors (T_1, T_2, T_3, T_4) into the proper state. For a read operation, the bit value is read from line B.

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”

Typical 16 Mb DRAM (4M x 4)

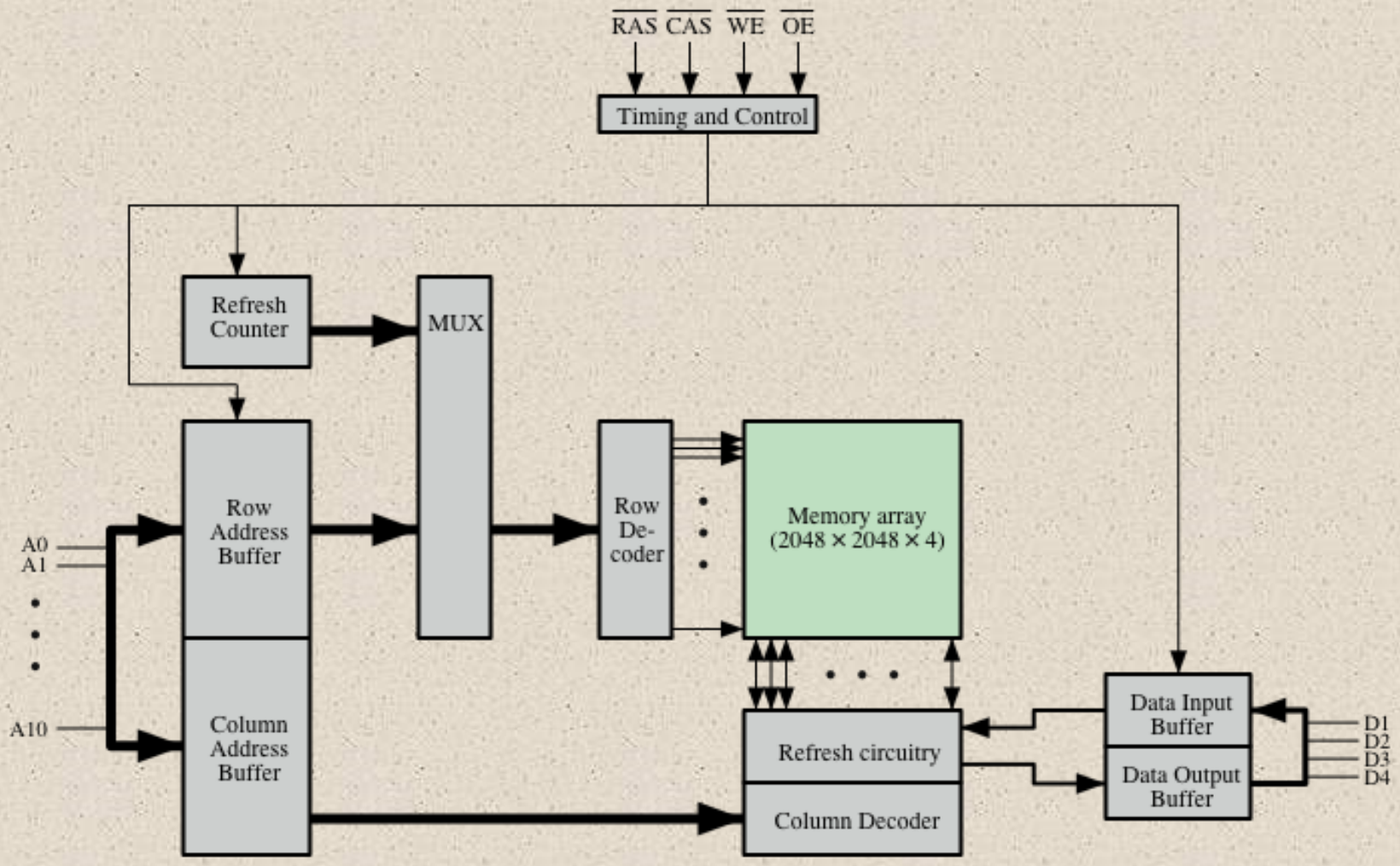


Figure 5.3 Typical 16 Megabit DRAM (4M × 4)

Here, 4 bits are read/written at a time.

Logically, memory array is organized as four square arrays of 2048x2048 elements. Various physical arrangements are possible.

In any case, the elements of the array are connected by both horizontal (row) and vertical (column) lines.

Each horizontal line connects to the Select terminal of each cell in its row; each vertical line connects to the Data-In/Sense terminal of each cell in its column.

+ Address lines supply the address of the word to be selected.

A total of $\log_2 W$ lines are needed. Here, 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 ($2^{11} = 2048$).

An additional 11 address lines select one of 2048 columns of 4 bits per column. 4 data lines are used for the input and output of 4 bits to and from a data buffer.

On input (write), bit driver of each bit line is activated for a 1/0 according to the value of the data line.

On output (read), 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.

As only 4 bits are read/written to this DRAM, there must be multiple DRAMs connected to the memory controller to read/write a word of data to the bus.

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.

Multiplexed addressing plus the use of square arrays result in a quadrupling of memory size with each new generation of memory chips.

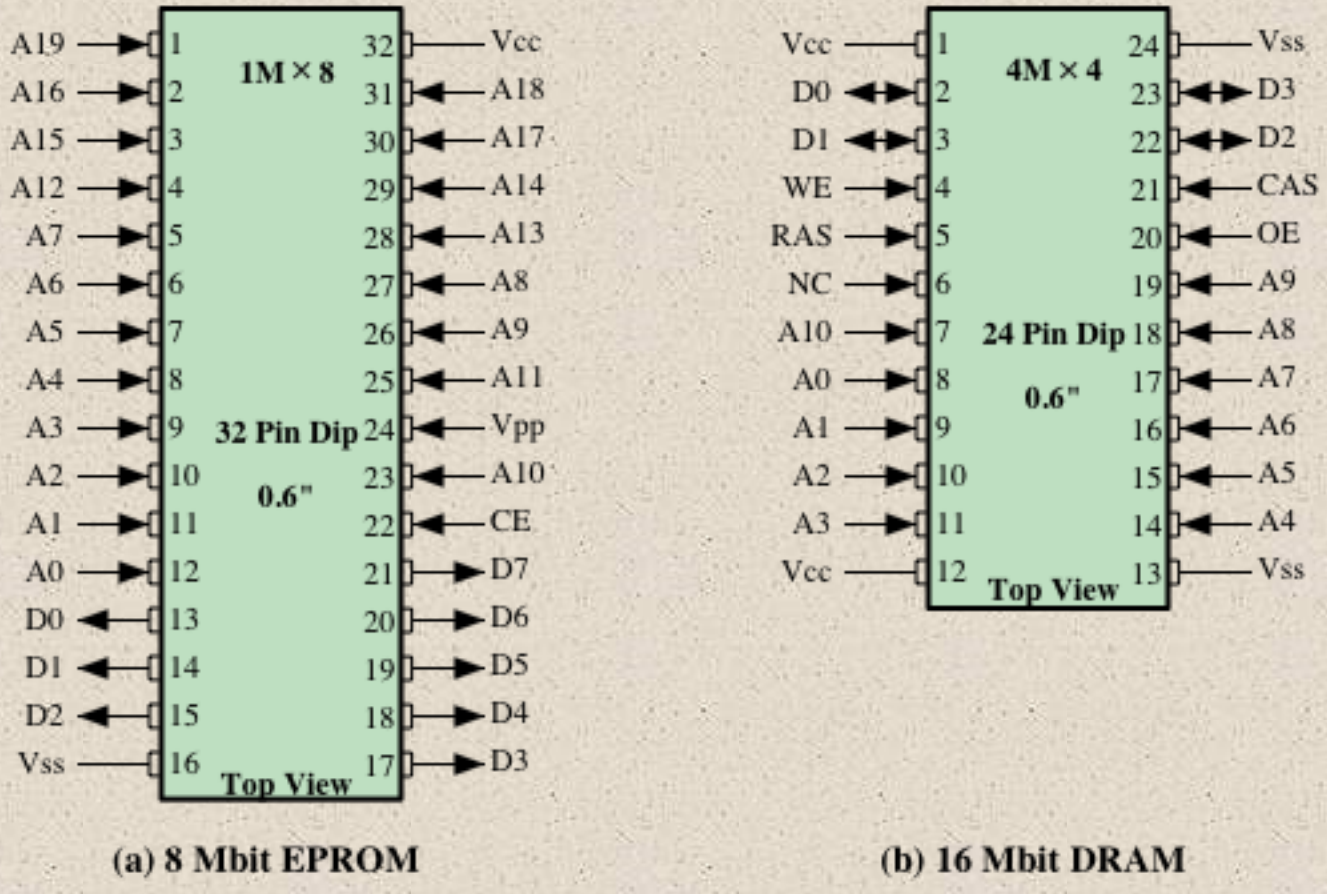
One more pin devoted to addressing doubles the number of rows and columns, and so the size of the chip memory grows by a factor of 4.

All DRAMs require a refresh operation. A simple technique for refreshing is, in effect, to disable the DRAM chip while all data cells are refreshed.

The refresh counter steps through all of the row values. For each row, the output lines from the refresh counter are supplied to the row decoder and the RAS line is activated.

The data are read out and written back into the same location. This causes each cell in the row to be refreshed.

Chip Packaging



(a) 8 Mbit EPROM

(b) 16 Mbit DRAM

Figure 5.4 Typical Memory Package Pins and Signals

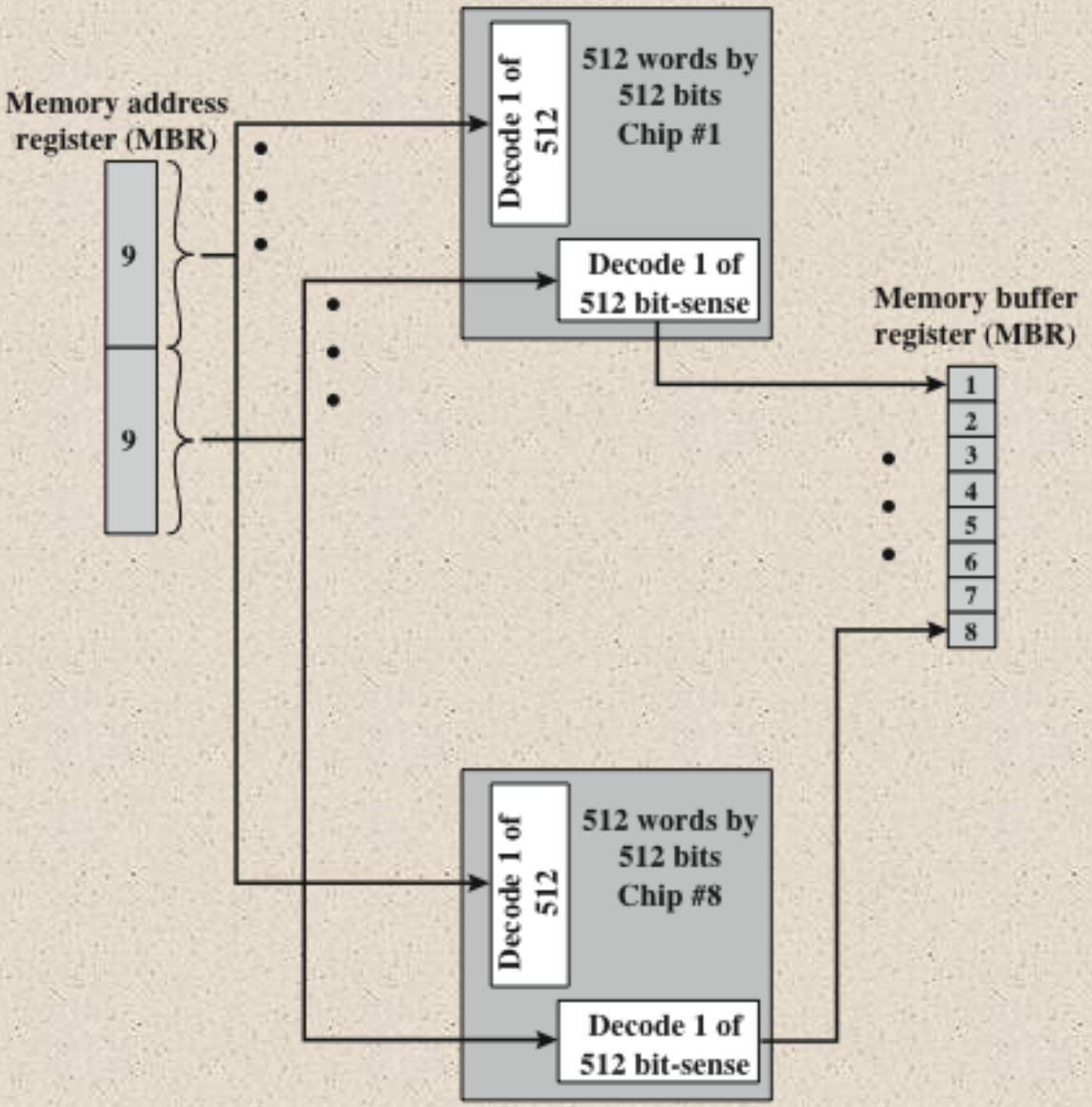


Figure 5.5 256-KByte Memory Organization

Figure 5.5

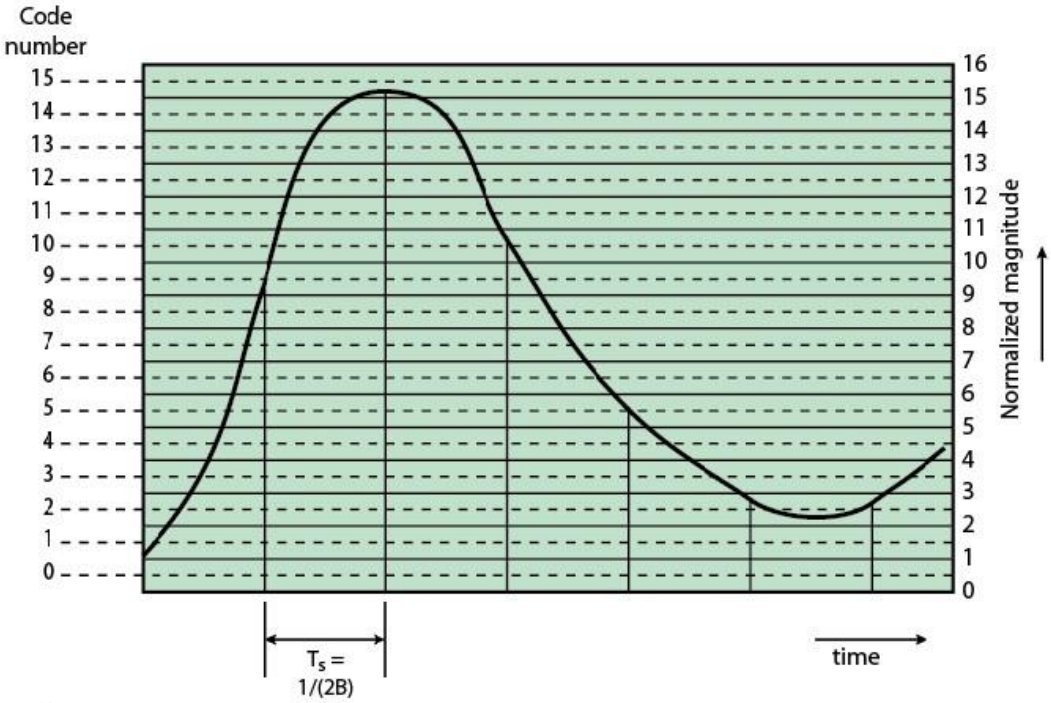
256-KByte Memory Organization

+ 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.

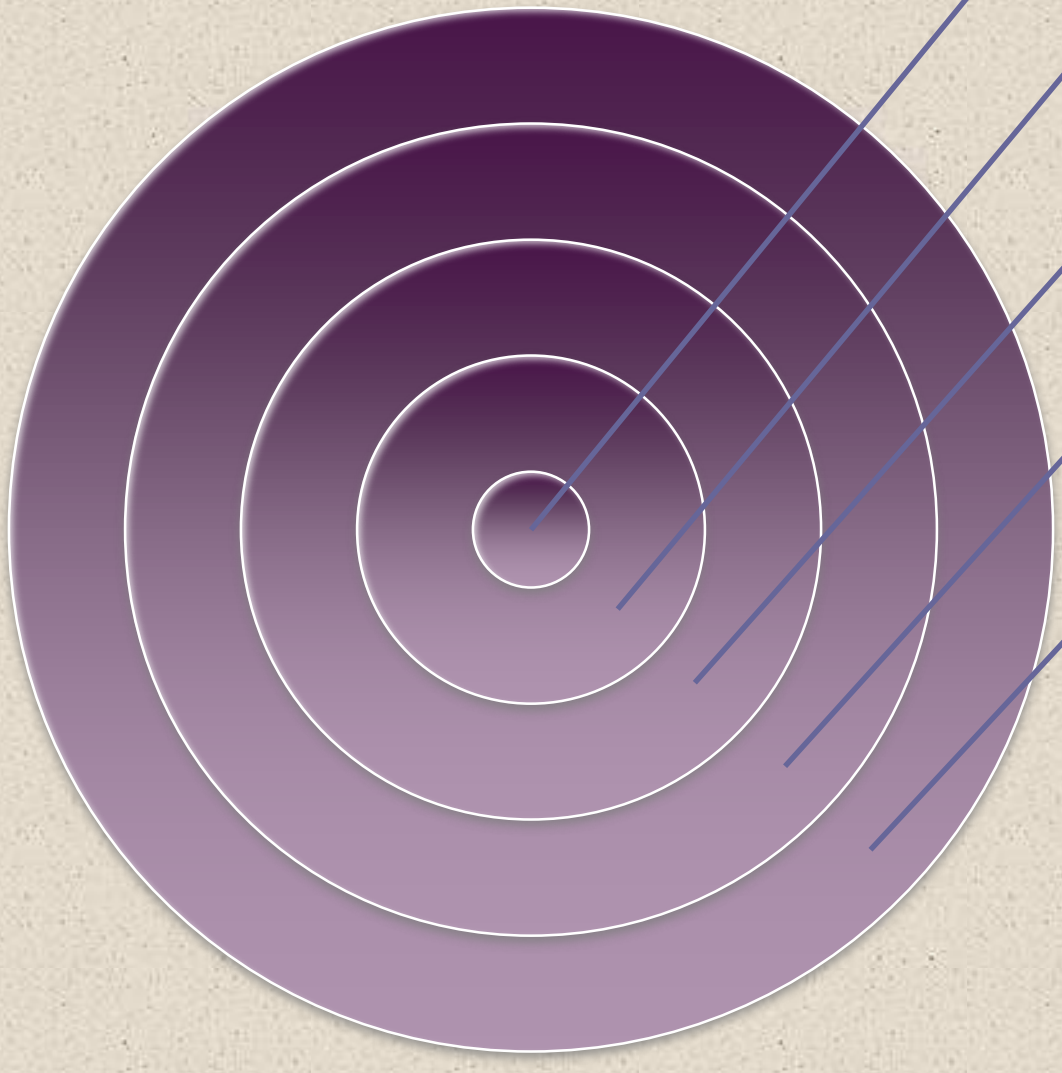
1MByte Module Organization



analog value	1.1	9.2	15.2	10.8	5.6	2.8	2.7
quantized code number	1	9	15	10	5	2	2
PCM code	0001	1001	1111	1010	0101	0010	0010

Figure 5.6 Pulse-Code Modulation Example

Interleaved Memory



Composed of a collection of DRAM chips

Grouped together to form a *memory bank*

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

Error Correcting Code Function

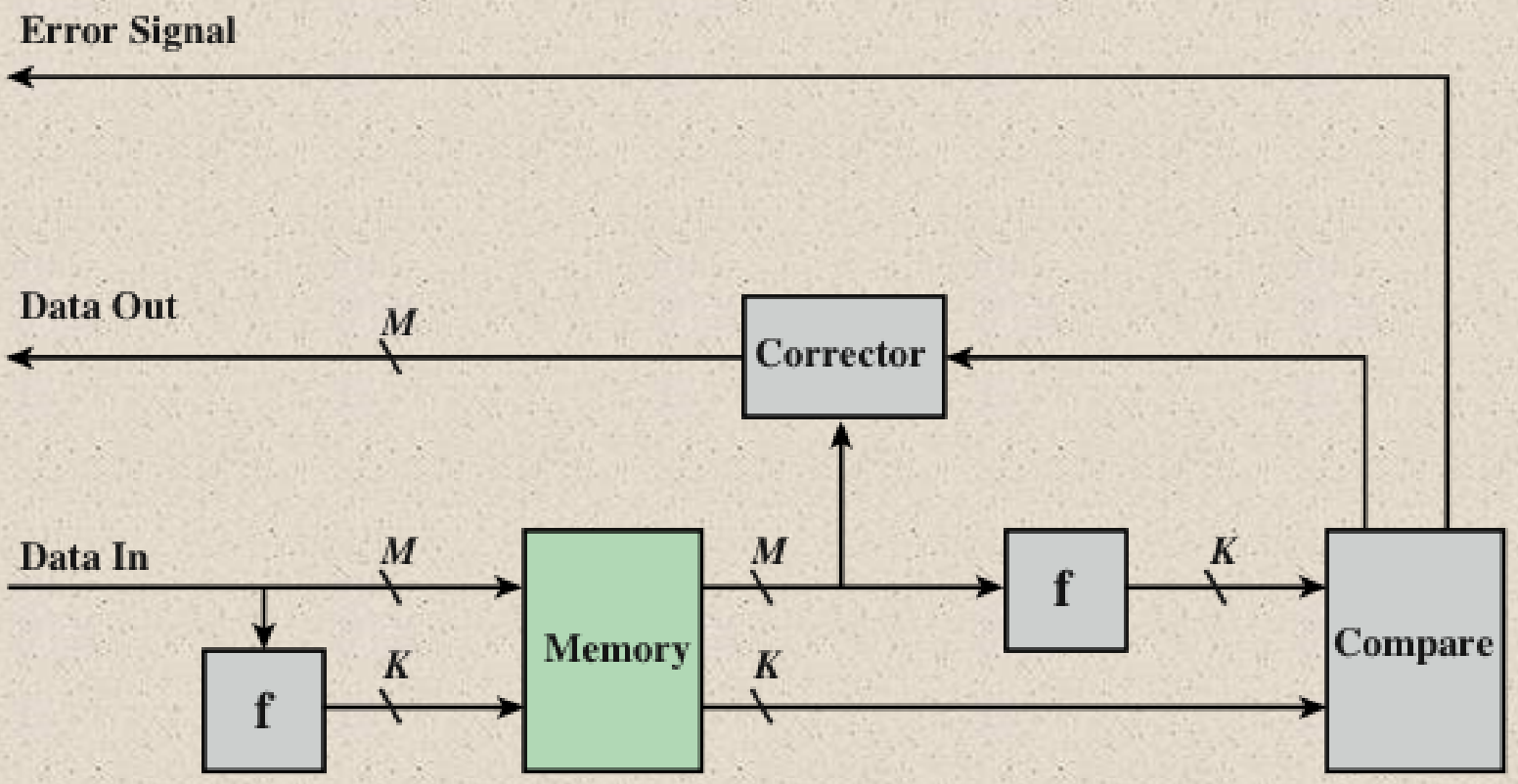


Figure 5.7 Error-Correcting Code Function

Hamming Error Correcting Code

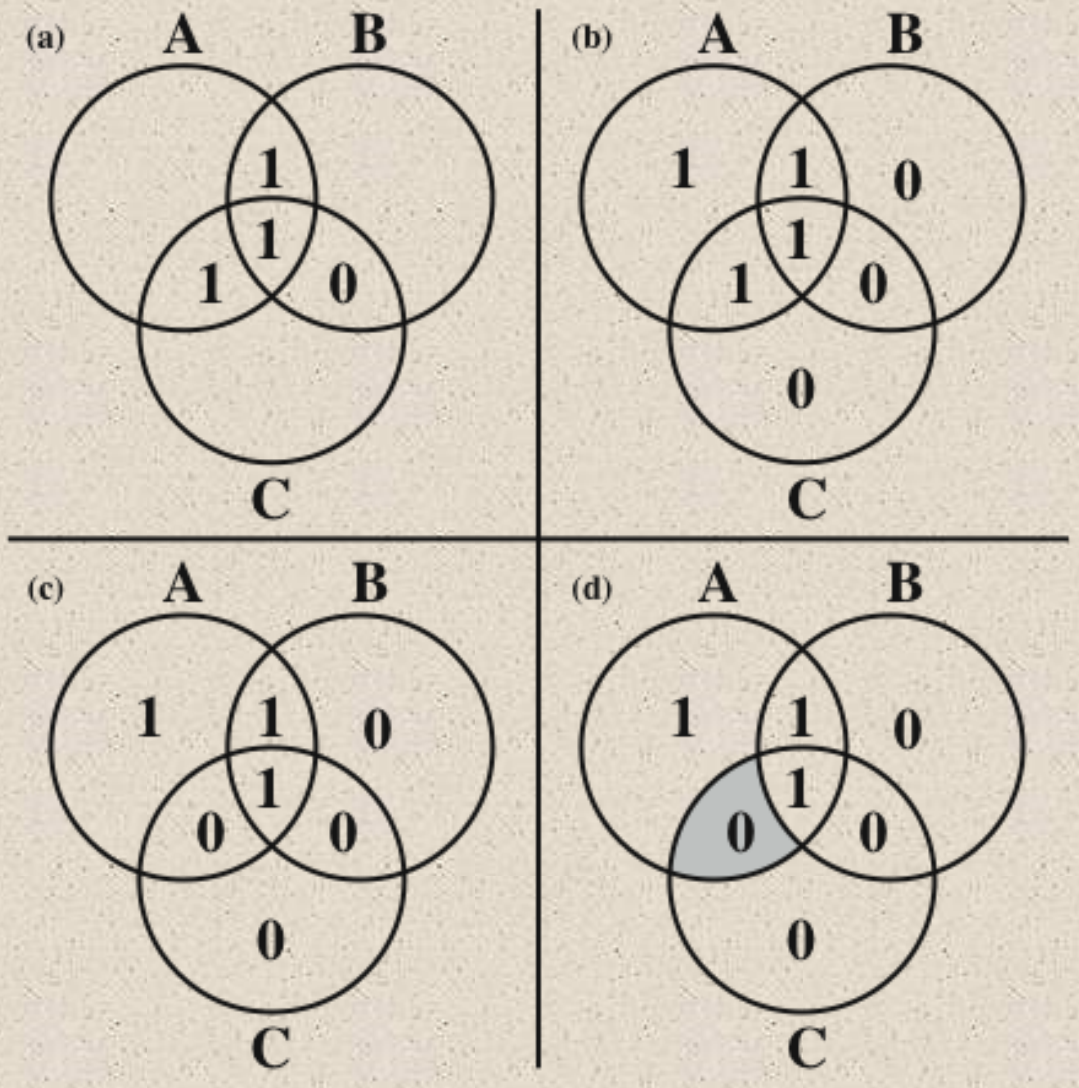


Figure 5.8 Hamming Error-Correcting Code

Table 5.3

Performance Comparison

DRAM Alternatives

	Clock Frequency (MHz)	Transfer Rate (GB/s)	Access Time (ns)	Pin Count
SDRAM	166	1.3	18	168
DDR	200	3.2	12.5	184
RDRAM	600	4.8	12	162

Table 5.3 Performance Comparison of Some DRAM Alternatives

Layout of Data Bits and Check Bits

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

Check Bit Calculation

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

Hamming SEC-DED Code

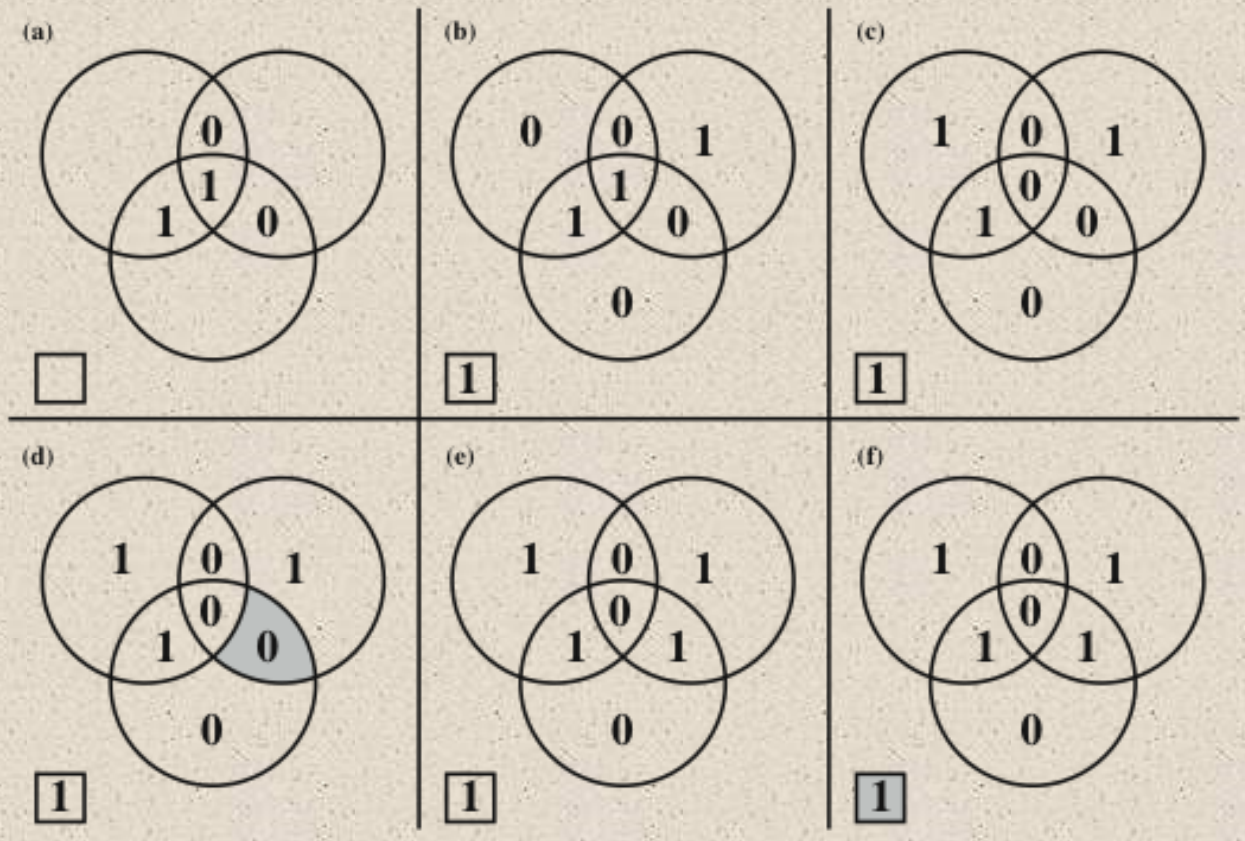


Figure 5.11 Hamming SEC-DED Code

Advanced DRAM Organization

SDRAM

DDR-DRAM

RDRAM

- 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:

	Clock Frequency (MHz)	Transfer Rate (GB/s)	Access Time (ns)	Pin Count
SDRAM	166	1.3	18	168
DDR	200	3.2	12.5	184
RDRAM	600	4.8	12	162

Table 5.3 Performance Comparison of Some DRAM Alternatives

+ Solution: Incorporate one or more levels of high-speed SRAM cache between the DRAM main memory and the processor.

But SRAM is much costlier than DRAM, and expanding cache size beyond a certain point yields diminishing returns.

The schemes that currently dominate the market are SDRAM, DDR-DRAM, and RDRAM. CDRAM has also received considerable attention.

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

Synchronous DRAM (SDRAM)

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

S A
D M
R

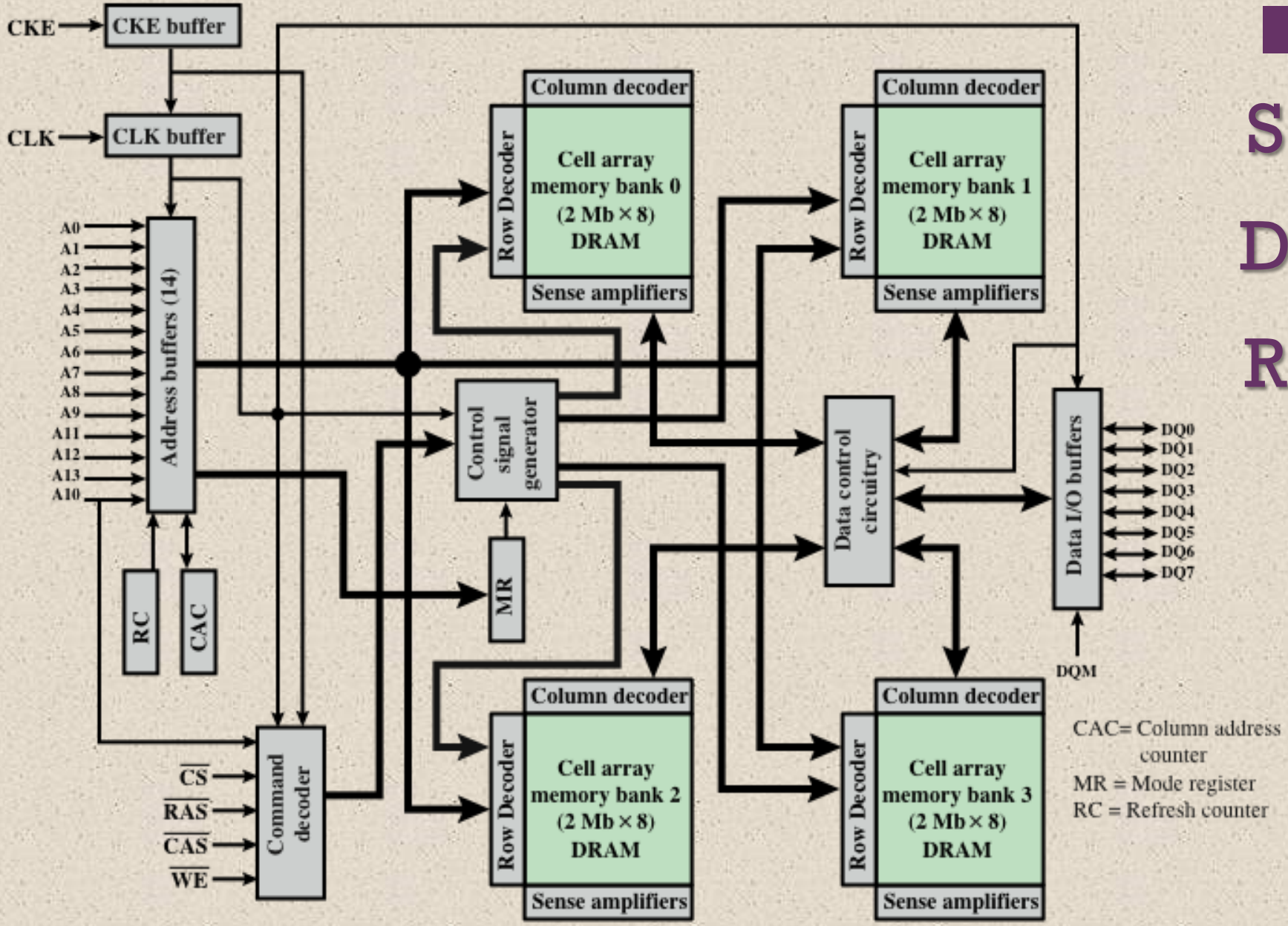


Figure 5.12 Synchronous Dynamic RAM (SDRAM)

SDRAM 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

Table 5.4 SDRAM Pin Assignments

SDRAM employs a burst mode to eliminate the address setup time and row and column line pre-charge time after the first access.

In burst mode, a series of data bits can be clocked out rapidly after the first bit has been accessed.

This mode is useful when all the bits to be accessed are in sequence and in the same row of the array as the initial access.

In addition, the SDRAM has a multiple-bank internal architecture that improves opportunities for on-chip parallelism.

+ Mode Register and associated control logic is a key feature differentiating SDRAMs from other DRAMs. It provides a mechanism to customize the SDRAM to suit specific system needs.

MR specifies the burst length, which is the number of separate units of data synchronously fed onto the bus. It also allows to adjust the latency between receipt of a read request and the beginning of data transfer.

SDRAM performs best when it is transferring large blocks of data serially, such as for applications like word processing, spreadsheets, and multimedia.

SDRAM Read Timing

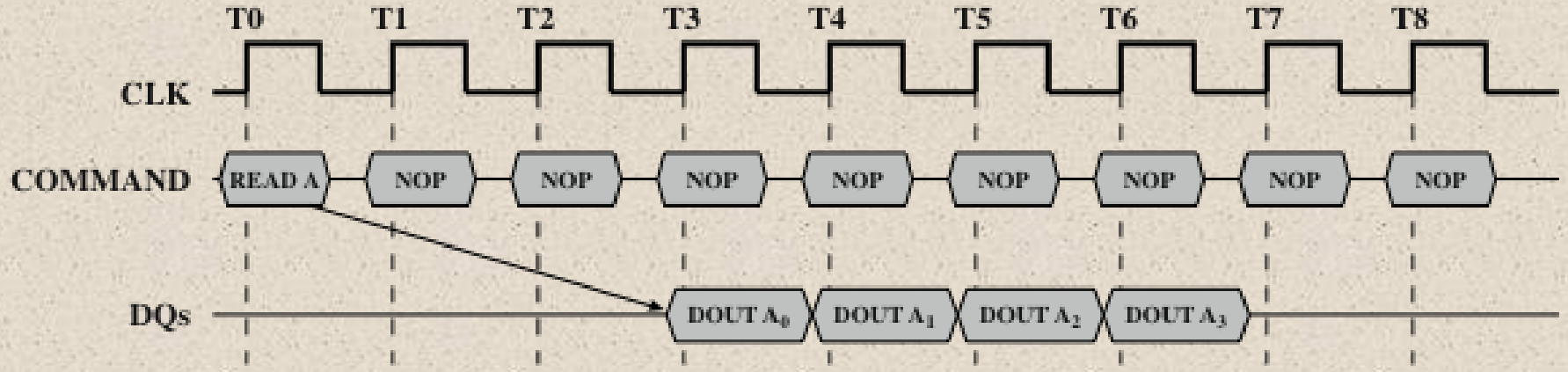


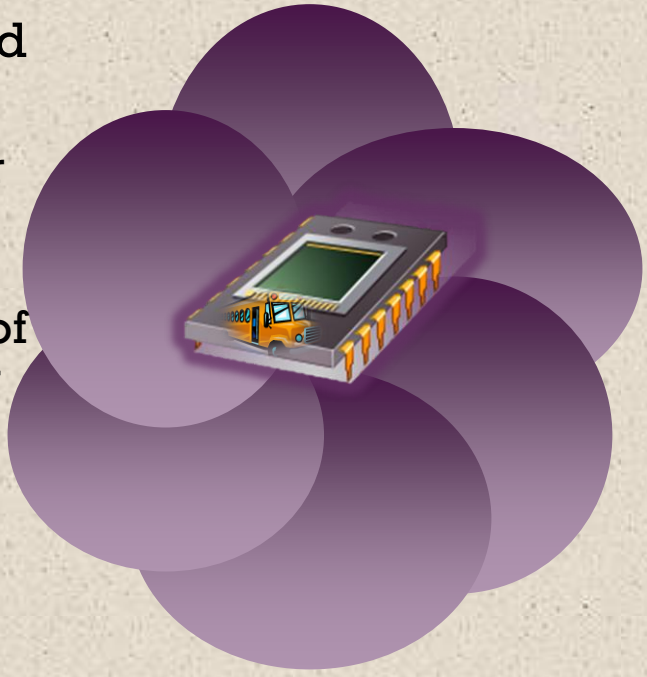
Figure 5.13 SDRAM Read Timing (Burst Length = 4, $\overline{\text{CAS}}$ latency = 2)

RDRAM

Developed by Rambus

Bus delivers address and control information using an asynchronous block-oriented protocol

- Gets a memory request over the high-speed bus
  - Request contains the desired address, the type of operation, and the number of bytes in the operation



Adopted by Intel for its Pentium and Itanium processors

Has become the main competitor to SDRAM

Bus can address up to 320 RDRAM chips and is rated at 1.6 GBps

Chips are vertical packages with all pins on one side

- Exchanges data with the processor over 28 wires no more than 12 centimeters long

RDRAM Structure

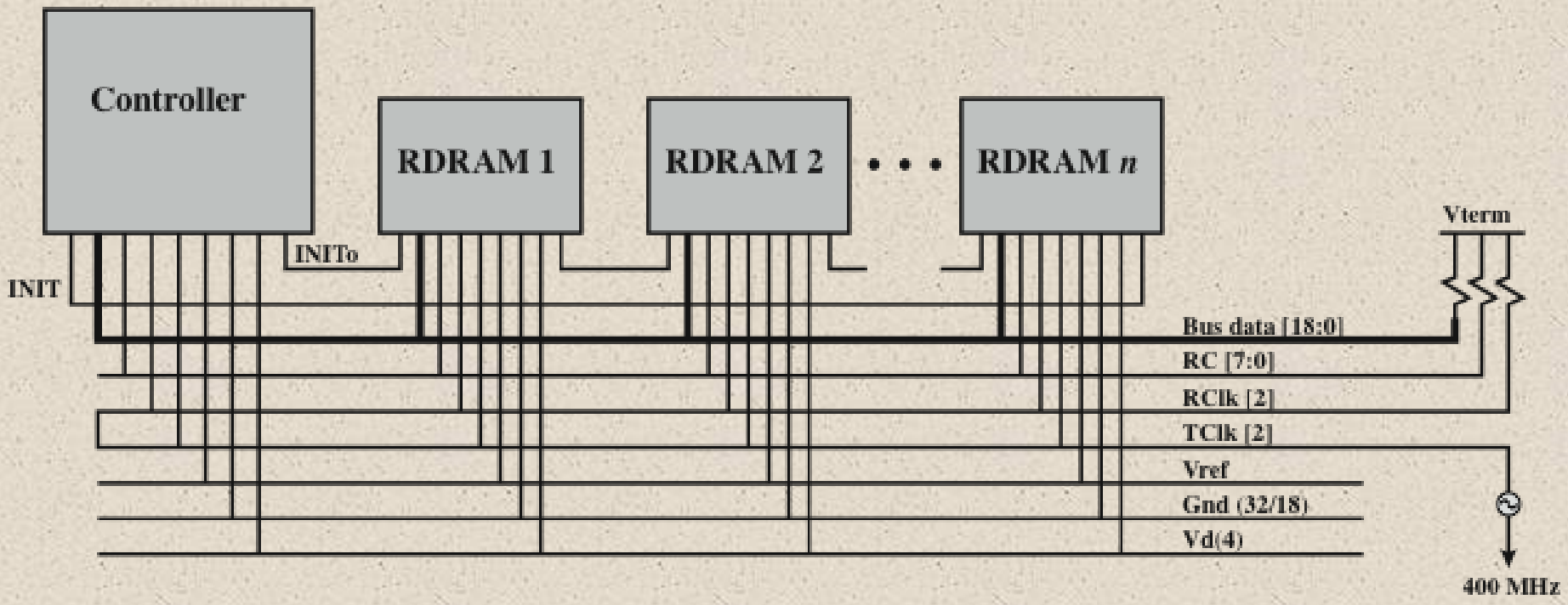


Figure 5.14 RDRAM Structure

The configuration consists of a controller and a number of RDRAM modules connected via a common bus.

The controller is at one end of the configuration, and the far end of the bus is a parallel termination of the bus lines.

The bus includes 18 data lines (16 actual data, two parity) cycling at twice the clock rate; that is, 1 bit is sent at the leading and following edge of each clock signal.

This results in a signal rate on each data line of 800 Mbps.

There is a separate set of 8 lines (RC) used for address and control signals, and a clock signal that starts at the far end from the controller propagates to the controller end and then loops back.

A RDRAM module sends data to the controller synchronously to the clock to master, and the controller sends data to an RDRAM synchronously with the clock signal in the opposite direction.

The remaining bus lines include a reference voltage, ground, and power source.

Double Data Rate SDRAM (DDR SDRAM)

- SDRAM can only send data once per bus clock cycle
- 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
- Developed by the JEDEC Solid State Technology Association (Electronic Industries Alliance's semiconductor-engineering-standardization body)

DDR SDRAM Read Timing

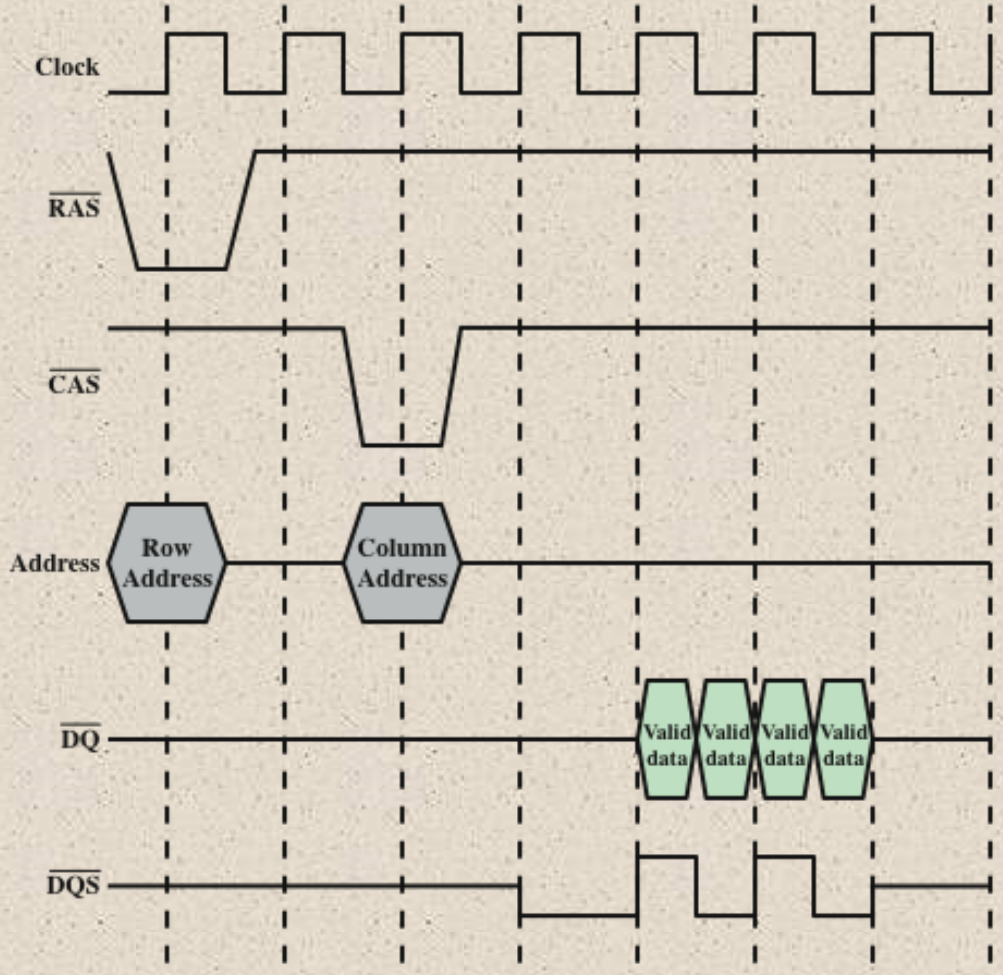


Figure 5.15 DDR SDRAM Read Timing

Data transfer is synchronized to both the rising and falling edge of the clock.

It is also synchronized to a bidirectional data strobe (DQS) signal that is provided by the memory controller during a read and by the DRAM during a write.

DDR2 increases the data transfer rate by increasing the operational frequency of the RAM chip and by increasing the prefetch buffer from 2 bits to 4 bits per chip.

The prefetch buffer is a memory cache located on the RAM chip.

The buffer enables the RAM chip to preposition bits to be placed on the data bus as rapidly as possible.

DDR3, introduced in 2007, increases the prefetch buffer size to 8 bits.

Theoretical transfer rate:
DDR: 200 to 600 MHz;
DDR2: 400 to 1066 MHz; and
DDR3: 800 to 1600 MHz.

In practice, somewhat smaller rates are achieved.

Cache DRAM (CDRAM)

- Developed by Mitsubishi
- Integrates a small SRAM cache onto a generic DRAM chip
- SRAM on the CDRAM can be used in two ways:
  - It can be used as a true cache consisting of a number of 64-bit lines
    - Cache mode of the CDRAM is effective for ordinary random access to memory
  - Can also be used as a buffer to support the serial access of a block of data

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
  - Hard failure
  - Soft error
- Hamming code
- Advanced DRAM organization
  - Synchronous DRAM
  - Rambus DRAM
  - DDR SDRAM
  - Cache DRAM