William Stallings

Computer Organization and Architecture 8th Edition

Chapter 5

Internal Memory

Semiconductor Memory Types

<table>
<tr><th>Memory Type</th><th>Category</th><th>Erasure</th><th>Write Mechanism</th><th>Volatility</th></tr>
<tr><td>Random-access memory (RAM)</td><td>Read-write memory</td><td>Electrically, byte-level</td><td>Electrically</td><td>Volatile</td></tr>
<tr><td>Read-only memory (ROM)</td><td rowspan="2">Read-only memory</td><td rowspan="2">Not possible</td><td>Masks</td><td rowspan="5">Nonvolatile</td></tr>
<tr><td>Programmable ROM (PROM)</td><td rowspan="4">Electrically</td></tr>
<tr><td>Erasable PROM (EPROM)</td><td rowspan="3">Read-mostly memory</td><td>UV light, chip-level</td></tr>
<tr><td>Electrically Erasable PROM (EEPROM)</td><td>Electrically, byte-level</td></tr>
<tr><td>Flash memory</td><td>Electrically, block-level</td></tr>
</table>

Read Only Memory (ROM)

Characteristics

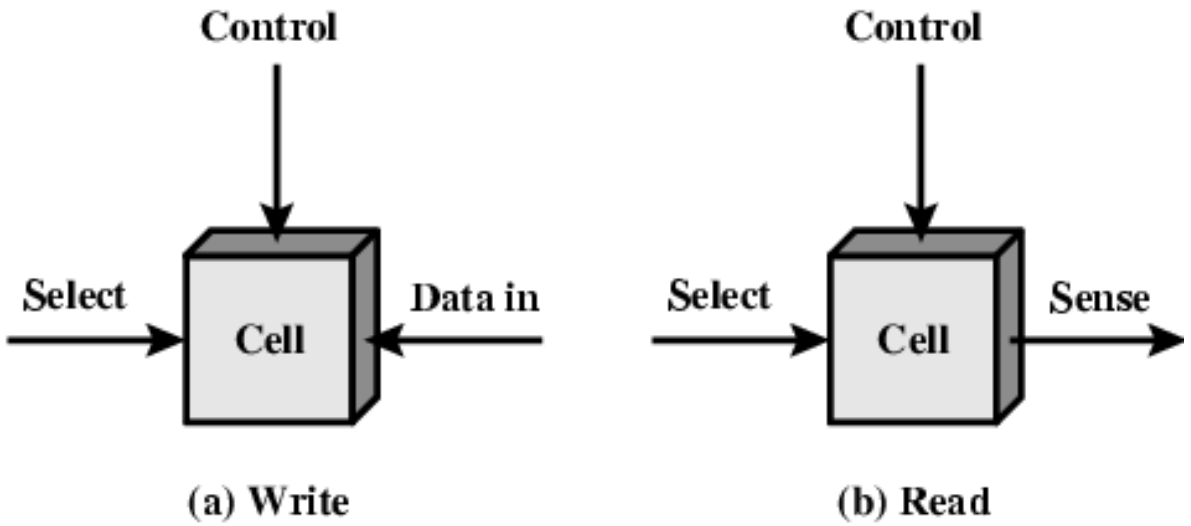
- Permanent storage
  - Nonvolatile
- Microprogramming
- Library subroutines
- Systems programs (BIOS)
- Function tables

Types of ROM

- Written during manufacture
  - Very expensive for small runs
- Programmable (once)
  - PROM
  - Needs special equipment to program
- Read “mostly”
  - Erasable Programmable (EPROM)
    - Erased by UV
  - Electrically Erasable (EEPROM)
    - Takes much longer to write than read
  - Flash memory
    - Erase whole memory electrically

Semiconductor Memory

- RAM
  - —Misnamed as all semiconductor memory is random access
  - —Read/Write
  - —Volatile
  - —Temporary storage
  - —Static or dynamic

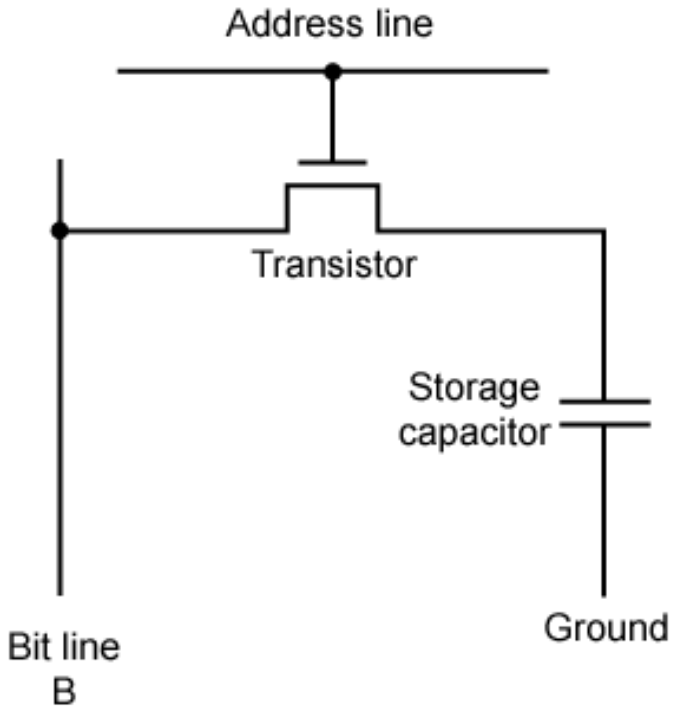


Dynamic RAM

- Bits stored as charge in capacitors
- Charges leak
- Need refreshing even when powered
- Simpler construction
- Smaller per bit
- Less expensive
- Need refresh circuits
- Slower
- Main memory
- Essentially analogue
  - Level of charge determines value

DRAM Operation

- Address line active when bit read or written
  - Transistor switch closed (current flows)
- Write
  - Voltage to bit line
    - High for 1 low for 0
  - Then signal address line
    - Transfers charge to capacitor



- Read
  - Address line selected
    - transistor turns on
  - Charge from capacitor fed via bit line to sense amplifier
    - Compares with reference value to determine 0 or 1
  - Capacitor charge must be restored

Static RAM

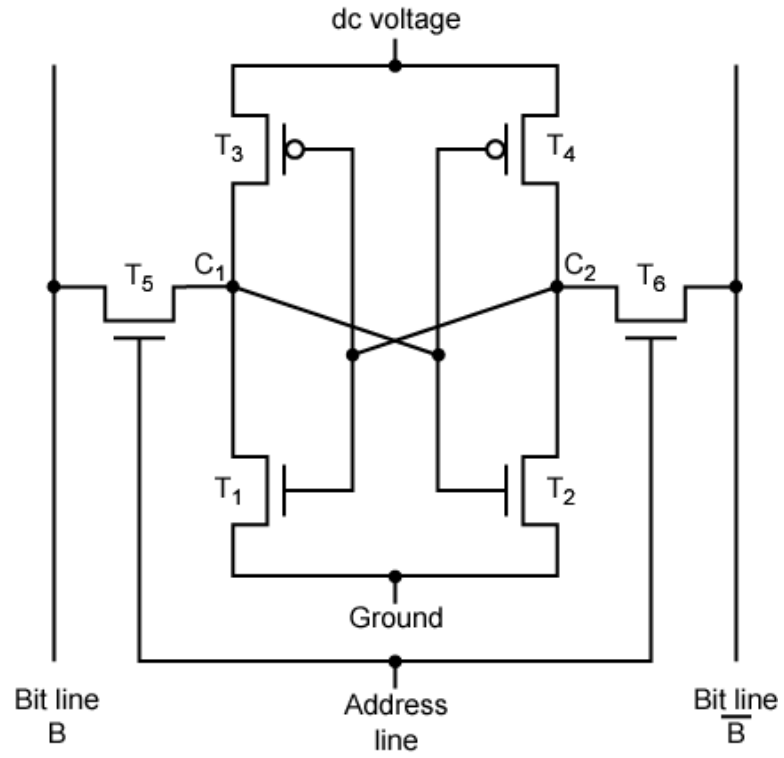
- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Does not need refresh circuits
- Faster
- Cache
- Digital
  - Uses flip-flops

Static RAM Structure & Operation

Four transistors (T1,T2,T3,T4) are cross connected that produces a stable logic state (LS).

In LS1, C1 is high & C2 is low; T1 and T4 are off and T2 and T3 are on.
In LS0, C1 is low and C2 is high; 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.

The AL controls T5 & T6. When a signal is applied to this line, the two transistors are switched on, allowing a read or write operation

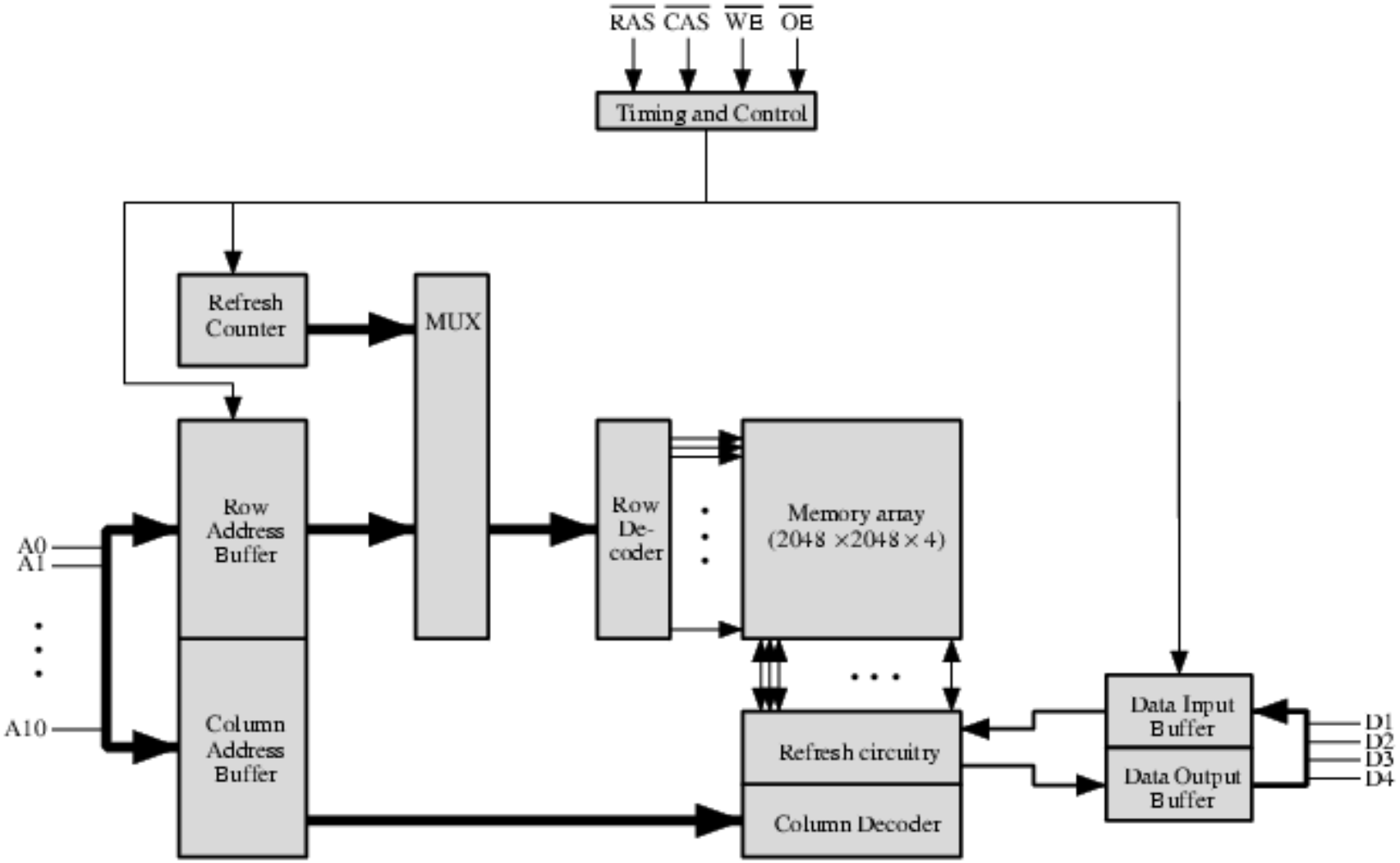


For a write operation, the desired bit value is applied to line B & $\overline{B}$
This forces T1, T2, T3, T4 into the proper state.
For a read operation, the bit value is read from line B.

SRAM v DRAM

- Both volatile
  - Power needed to preserve data
- Dynamic cell
  - Simpler to build, smaller
  - More dense
  - Less expensive
  - Needs refresh
  - Larger memory units
- Static
  - Faster
  - Cache

Typical 16 Mb DRAM (4M x 4)



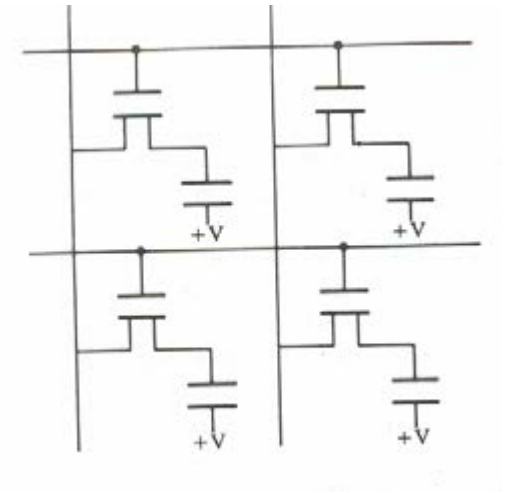
Operation of Typical 16 Mb DRAM (4M x 4)

- 11 address signals are passed to define the row address, accompanied by asserting RAS' (row address select)
- Then, 11 other address signals are passed to define the column address, accompanied by asserting CAS' (Column address select)
- If WE' (write enable) is asserted:
  -Bit driver of each bit line is activated; (4 lines, one for each bit)
  -Voltage signal is applied to bit line
  -A signal is applied to address line, allowing capacitor to charge

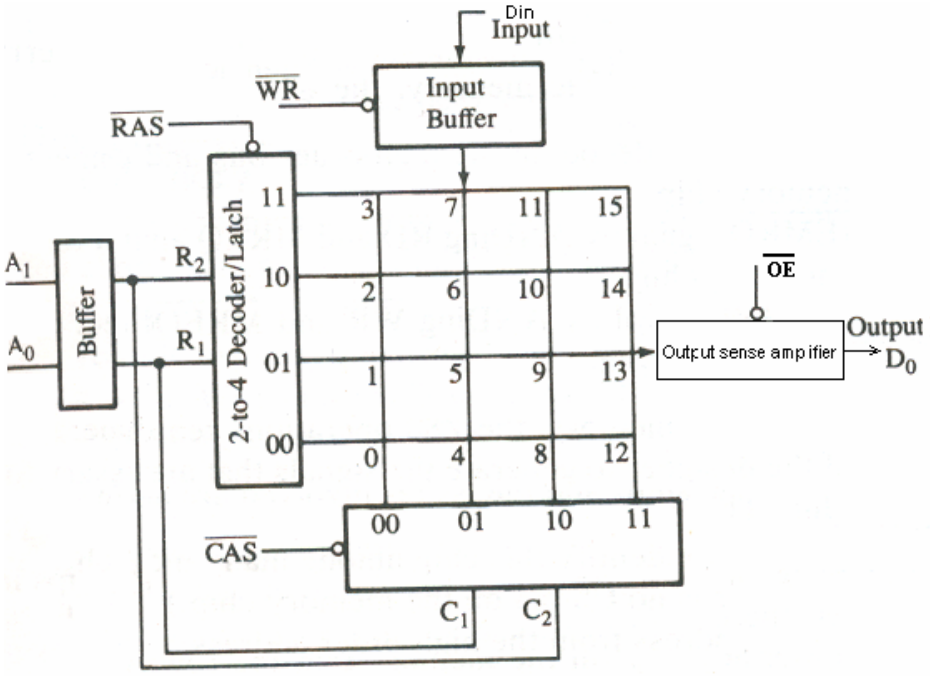
- If OE' (output enable)is asserted:
  -Address line is selected, transistor turns on
  -Charge stored in capacitor is fed out onto a bit line to a sense amplifier
  -Sense amplifier compares capacitor voltage to a reference value & determines if cell contains a logic 0 or 1
  -Value is presented to data line; (4 lines, one for each bit)

DRAM Organization

- To select a specific bit:
  - — Select an address, then assert RAS' (row address select
  - — Then, select another address, then assert CAS' (column address select)



- Then assert WE', or OE' (write enable and output enable)
- Reduces number of address pins
  - — Multiplex row address and column address
  - — 11 pins to address ($2^{11}=2048$)
  - — Adding one more pin doubles range of values so x4 capacity



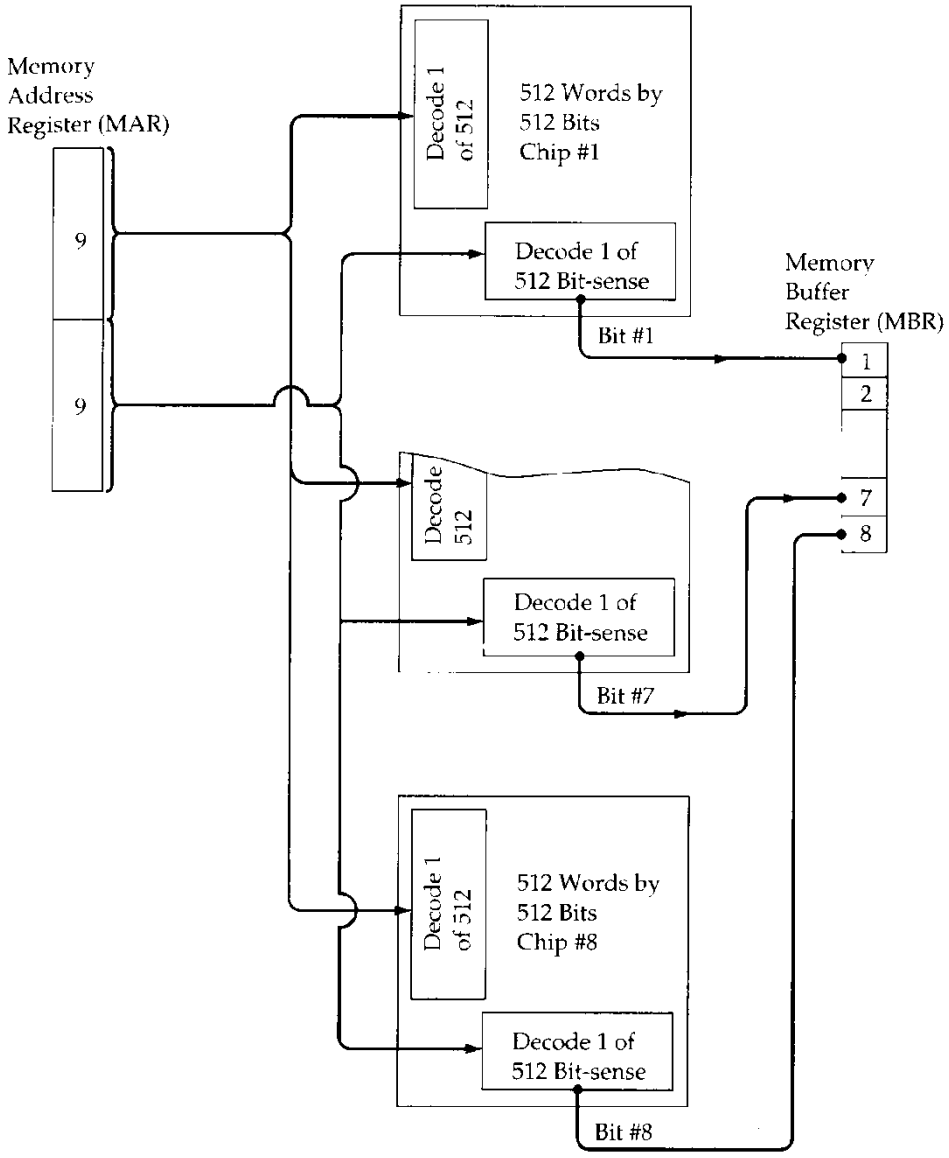
Refreshing

- Refresh circuit included on chip
- A memory cell is automatically refreshed by:
  - Accessing its row
  - Reading from it
  - Writing into it
- Usually, DRAM is refreshed by:
  - Placing a row address on address lines & asserting RAS'
  - Use a counter to go through all row addresses, while RAS' is activated
- Disable chip
- Count through rows
- Read & Write back
- Takes time
- Slows down apparent performance

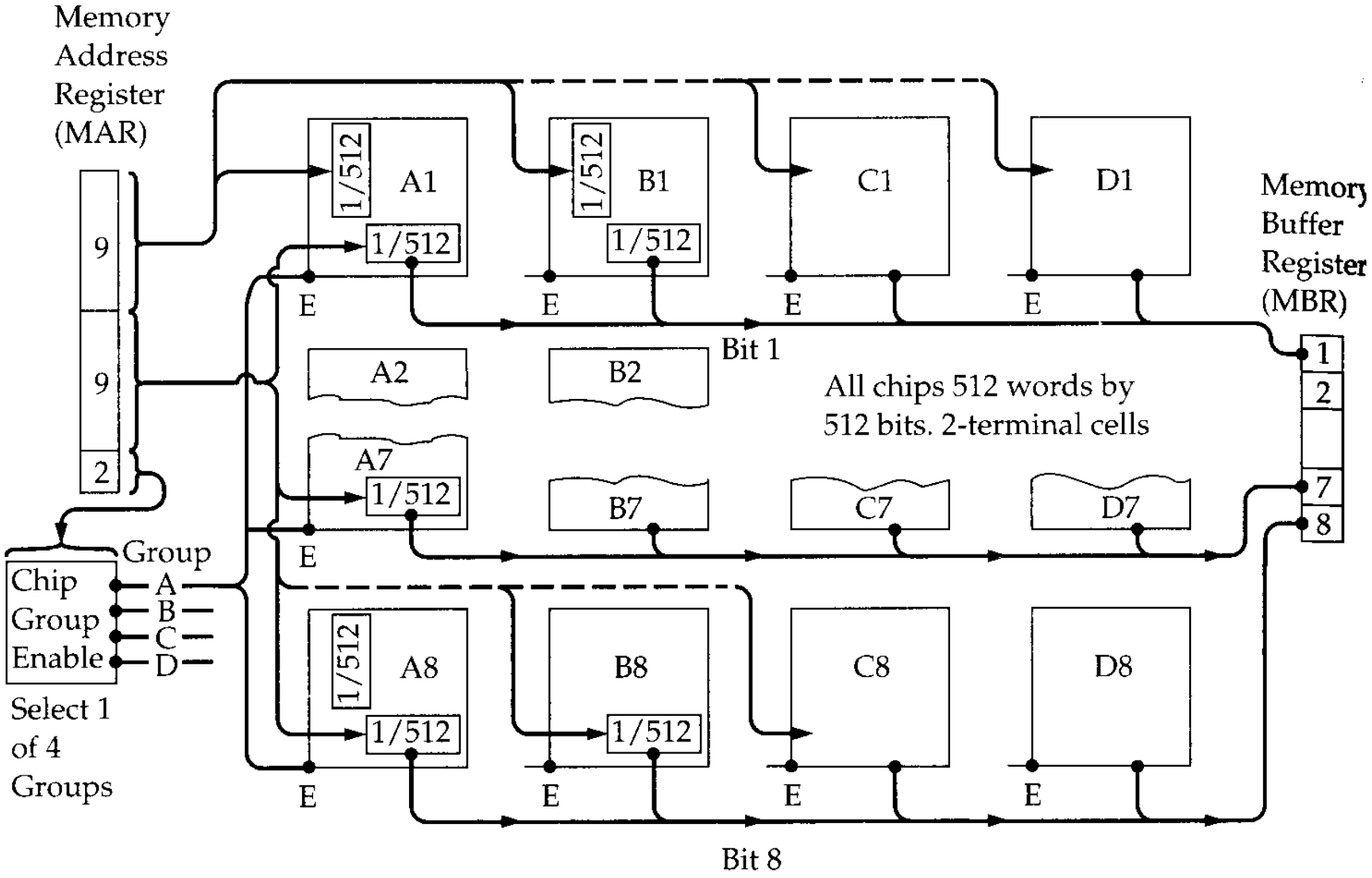
256kByte Module Organisation

In this memory module of 256k 8-bit words, 18 bit address is needed & is supplied to the module from some external source. The address is presented to 8 256Kx1 bit chips each provides I/O of 1 bit

This works as long as size of memory = # of bits per chip. When larger memory is required, an array of chips is needed



1MByte Module Organisation (8-bit word)



Interleaved Memory & S.C Memory Errors

- Collection of DRAM chips
- Grouped into memory bank
- Banks independently service read or write requests
- K banks can service k requests simultaneously
- Memory read or write rate increases by factor k

Semi-conductor memory system is subject to errors

- Hard Failure
  - Permanent defect caused by environment abuse
  - manufacturing defects.
- Soft Error
  - —Random, non-destructive
  - —No permanent damage to memory
  - —It is caused by power supply problems or alpha particles
- Detected using Hamming error correcting code

Advanced DRAM Organization

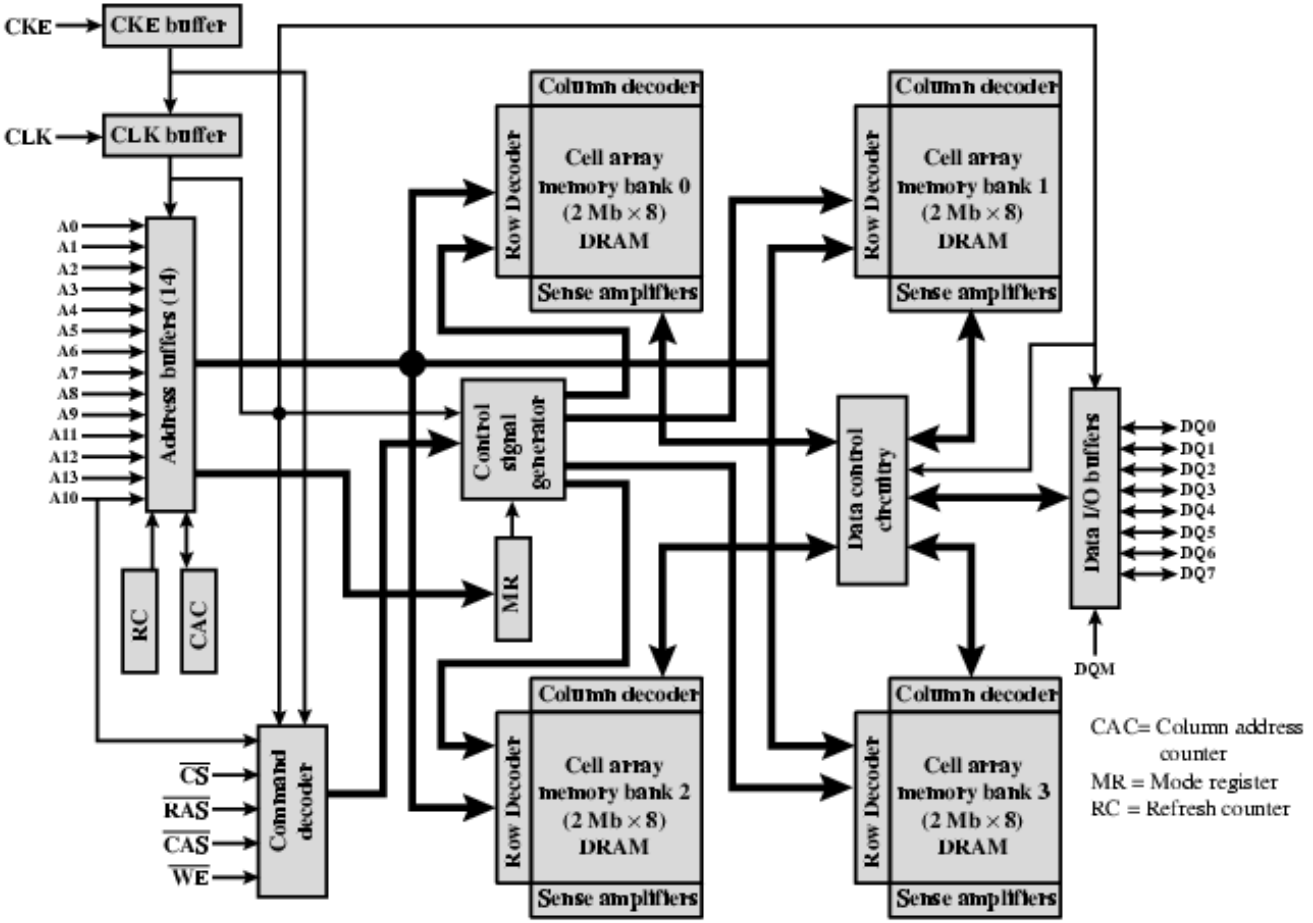
- Basic DRAM same since first RAM chips
- Enhanced DRAM
  - Contains small SRAM as well
  - SRAM holds last line read
- Cache DRAM
  - Larger SRAM component
  - Use as cache or serial buffer

Synchronous DRAM (SDRAM)

- Access is synchronized with an external clock
- Address is presented to RAM
- RAM finds data
  - CPU waits in conventional DRAM
  - amount of wait = access time)
- Since SDRAM moves data in time with system clock:
  - CPU knows when data will be ready
  - CPU does not have to wait, it can do something else
- Burst mode allows SDRAM to set up stream of data and fire it out in block
- Mode register specifies the burst length & Allows the adjusting of the latency between receipt of a read request & the beginning of data transfer
- DDR-SDRAM sends data twice/clock cycle (leading & trailing edge)

SDRAM (IBM 64 Mb)

The 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

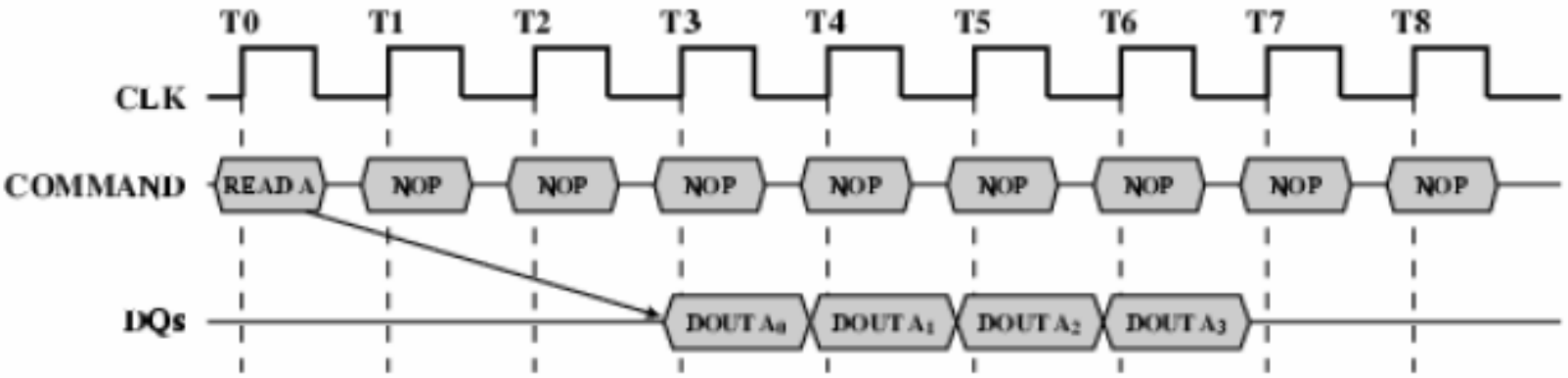


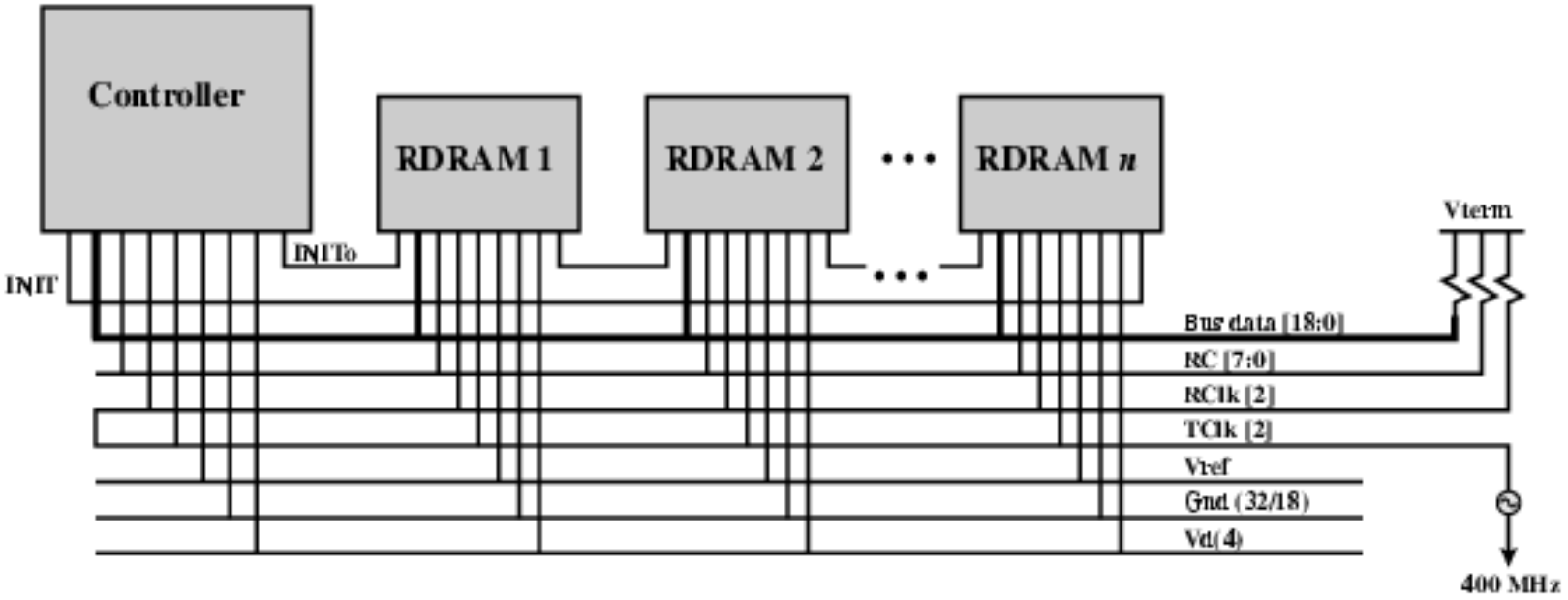
Figure 5.13. SDRAM read timing (burst length = 4 & $\overline{CAS}$ latency = 2

Burst read command is initiated by having $\overline{CS}$ & $\overline{CAS}$low while $\overline{RAS}$ & $\overline{WE}$ high at the rising edge of the clock. The address inputs determine the starting column address of the burst and the mode register sets of type of burst and the burst length. The delay from the start of the command and the data from the first cell appears on the outputs is equal to $\overline{CAS}$ latency.

RAMBUS

- Adopted by Intel for Pentium & Itanium
- Main competitor to SDRAM
- Vertical package – all pins on one side
- Data exchange over 28 wires < 12cm long
- Bus addresses up to 320 RDRAM chips at 1.6Gbps
- Asynchronous block oriented protocol
  - 480ns access time
  - Then 1.6 Gbps

RAMBUS Diagram



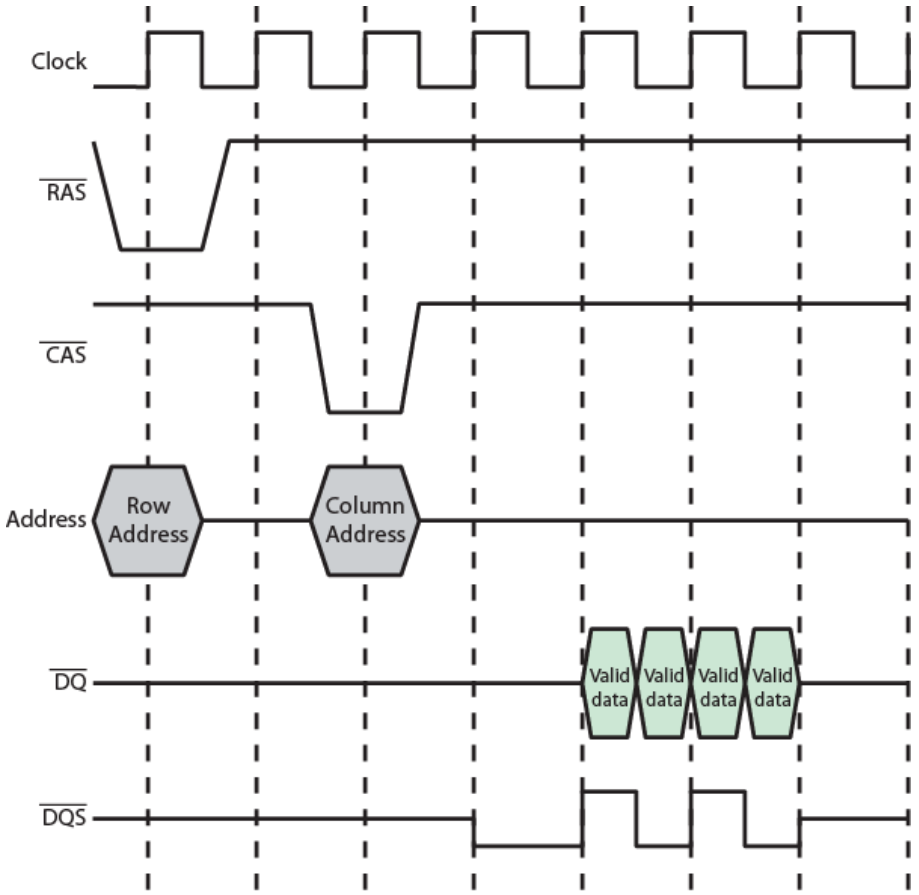
A controller connects a # of RDRAM module via a common bus. The bus includes 18 data lines (16 actual data & 2 parity) cycling at twice the clock rate; 1 bit is sent at the leading and following edge of each clock cycle

- This results in a signal rate on each data line of 800 MHz. there is a separate set of 8 lines (RC) used for address and control signals.
- There is also a clock signal that starts at the far end from the controller propagates to the controller end and then loops back.

DDR SDRAM Read Timing

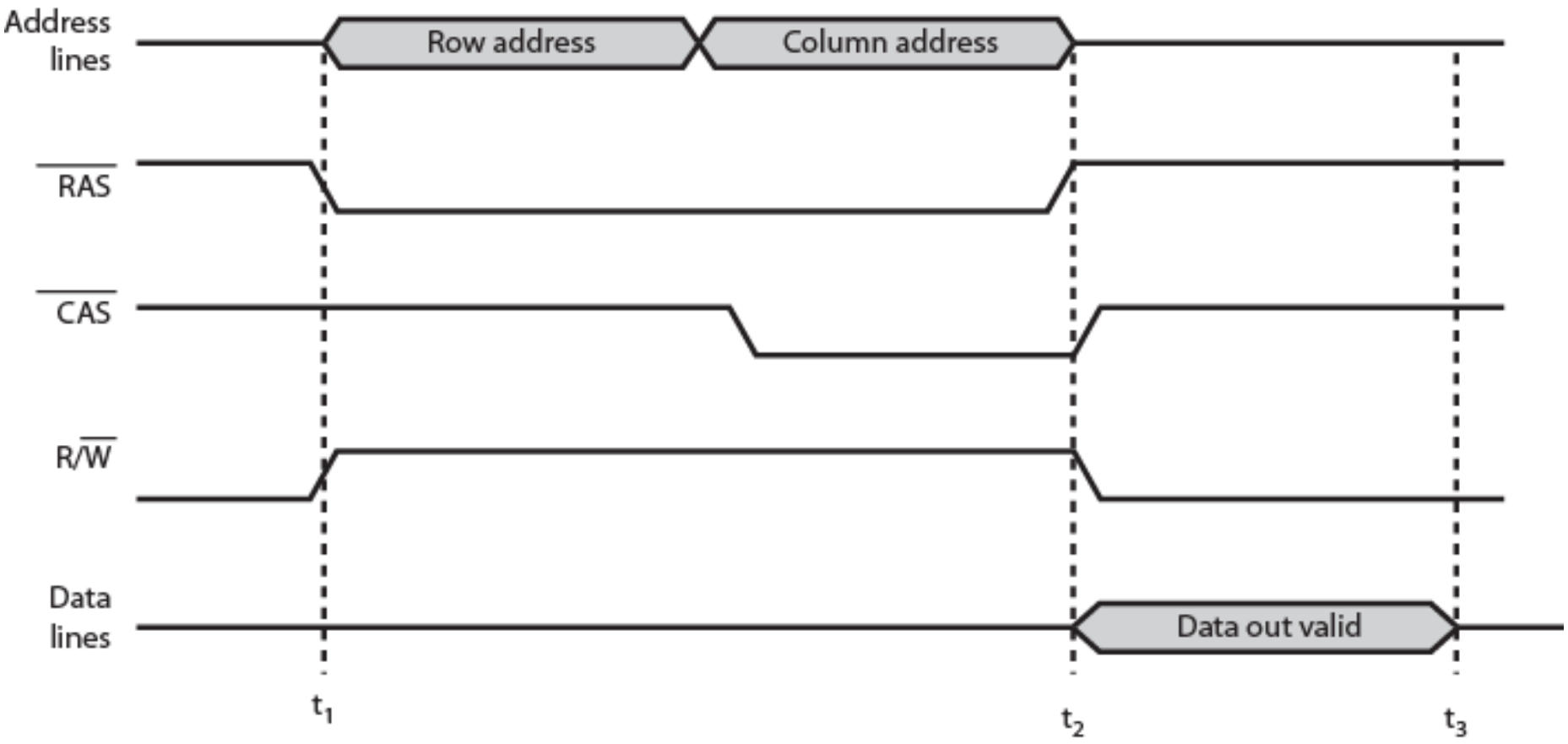
- SDRAM can only send data once per clock
- Double-data-rate SDRAM can send data twice per clock cycle - Rising edge and falling edge

Data transfer is synchronized to the rising & 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 & by DRAM during a write. In typical implementation DQS is ignored during the read.



RAS = row address select
CAS = column address select
DQ = data (in or out)
DQS = DQ select

Simplified DRAM Read Timing

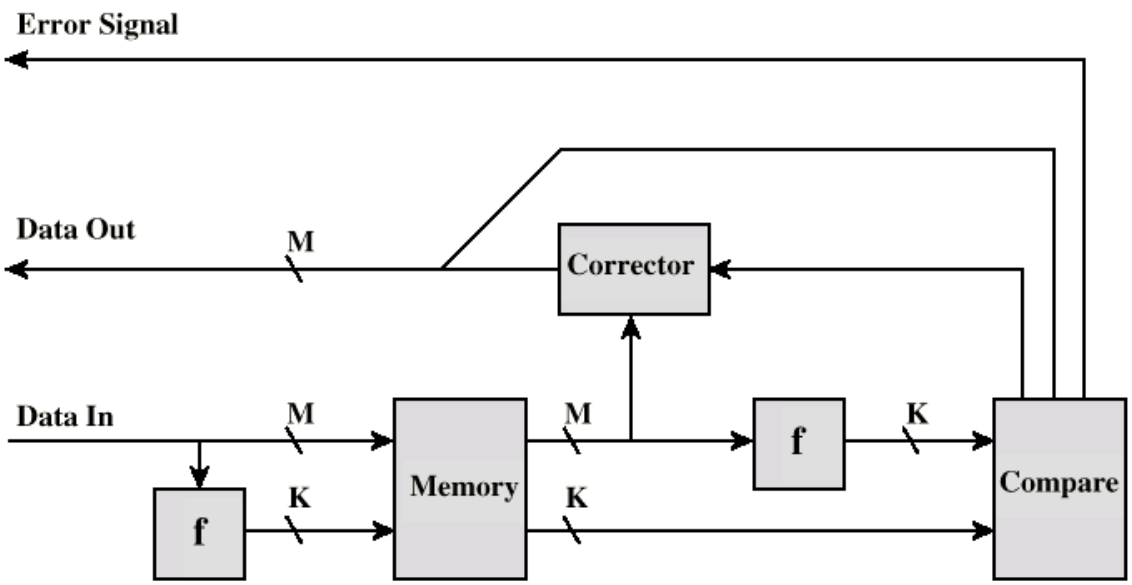


Cache DRAM

- Mitsubishi
- Integrates small SRAM cache (16 kb) onto generic DRAM chip
- Used as true cache
  - 64-bit lines
  - Effective for ordinary random access
- To support serial access of block of data
  - E.g. refresh bit-mapped screen
    - CDRAM can prefetch data from DRAM into SRAM buffer
    - Subsequent accesses solely to SRAM

Error Correcting Code Function

When data are to be read into memory, a calculation is performed on the data to produce a code of length k. Code and data are stored. If an m bit word is to be stored, then the size of stored word is M + K bits



Code is used to detect and possibly correct errors. A new set of k code is generated form k data bits and compared with fetched code bits. This yields to:

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

Organization in detail

A 16Mbit chip can be organised as:

- 1M of 16 bit words
- 16 of 1Mbit chip with bit 0 of each word in chip 0, bit 1 of each word in chip 1, and so on
- 4 of 2048 x 2048 bit arrays (4Mx4)
  - i.e. 4 parallel planes each containing 2048 cellsx2048 cells
  - Cells of the array are connected by horizontal/vertical lines
  - Horizontal lines connect to the select of each cell in the row (2048 horizontal lines)
    - Need a decoder with 11 inputs, whose outputs are those 2048 horizontal lines, i.e. to select one of the 2048 rows; a.k.a. row decoder
    - Need another decoder with another 11 inputs to select one of the 2048 columns; a.k.a. column decoder
    - This way, one of 4 million bits per plane gets selected; i.e. 4 bits in total are selected
    - Using the same lines for both decoders, only half the number of address lines is required => smaller size
- Vertical lines connect data-in/sense of each cell in the column
- Adding one more pin doubles range of values so x4 capacity

Dynamic RAM Structure & Operation

DRAM structure for a cell that stores 1 bit. Address line (AL) is activated when bit value from this cell is to be read or written. Transistor acts as a switch that is closed if voltage is applied to AL and open if no voltage is applied to AL

For write operation, a voltage signal is applied to the bit line. A signal is applied to AL allowing a charge to be transferred to the capacitor

For read operation, when AL is selected, transistor turns on and charge stored on capacitor is fed out onto a bit line and to a sense amplifier SA. SA compare the capacitor voltage t a reference value and determine if the cell contain a logic 1 or logic 0. readout from the cell discharges the capacitor, which must restored to complete the operation