Computer Organization and Architecture

Module 4

Prof. Indranil Sengupta

Dr. Sarani Bhattacharya

Department of Computer Science and Engineering

IIT Kharagpur

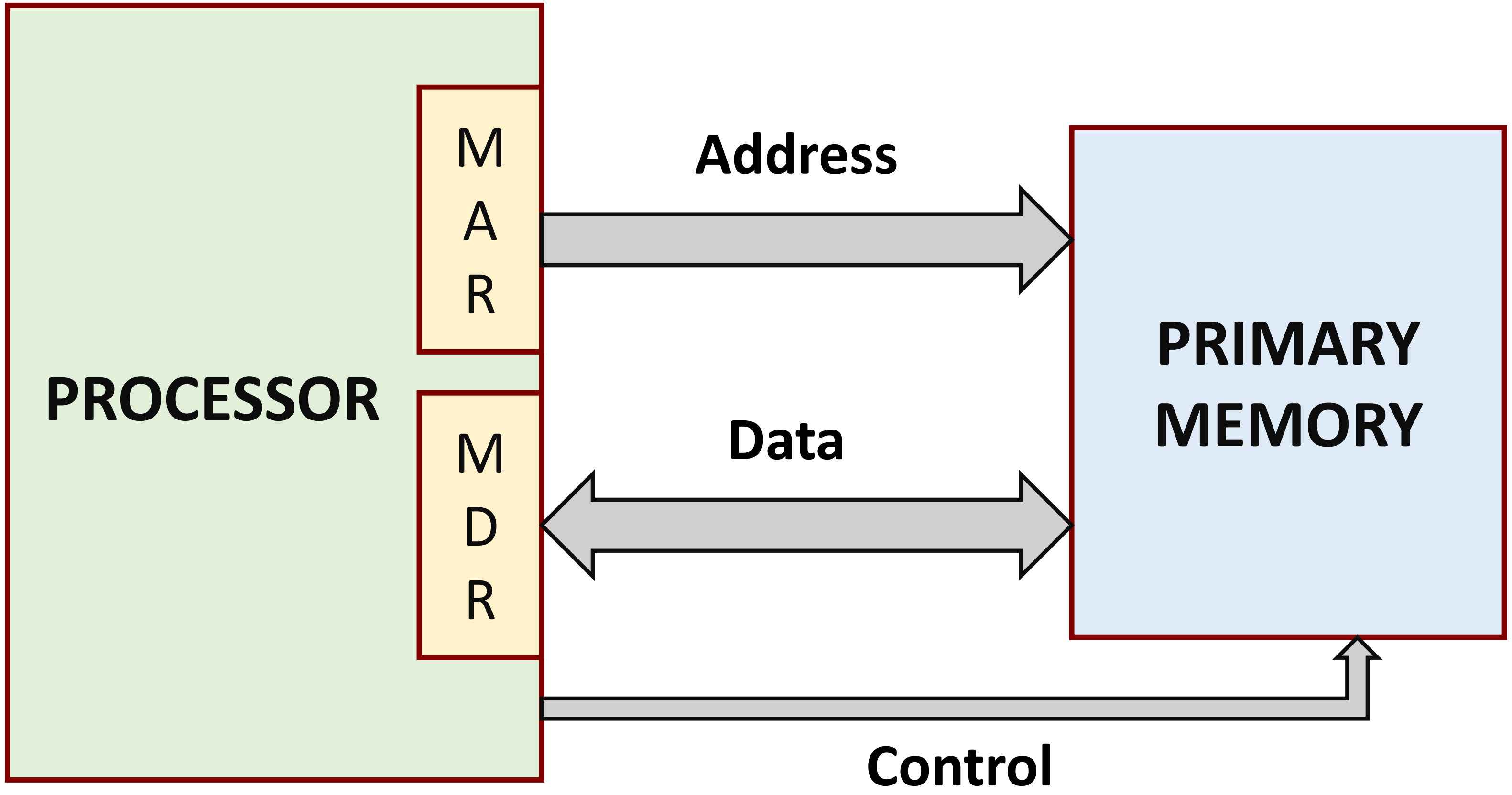
PROCESSOR MEMORY INTERACTION

Introduction

- Memory is one of the most important functional units of a computer.
    - Used to store both instructions and data.
    - Stores as bits (0’s and 1’s), usually organized in terms of bytes.
- How are the data stored in memory accessed?
    - Every memory location has a *unique address*.
    - A memory is said to be *byte addressable* if every byte of data has a unique address.
    - Some memory systems are *word addressable* also (every addressed location consists of multiple bytes, say, 32 bits or 4 bytes).

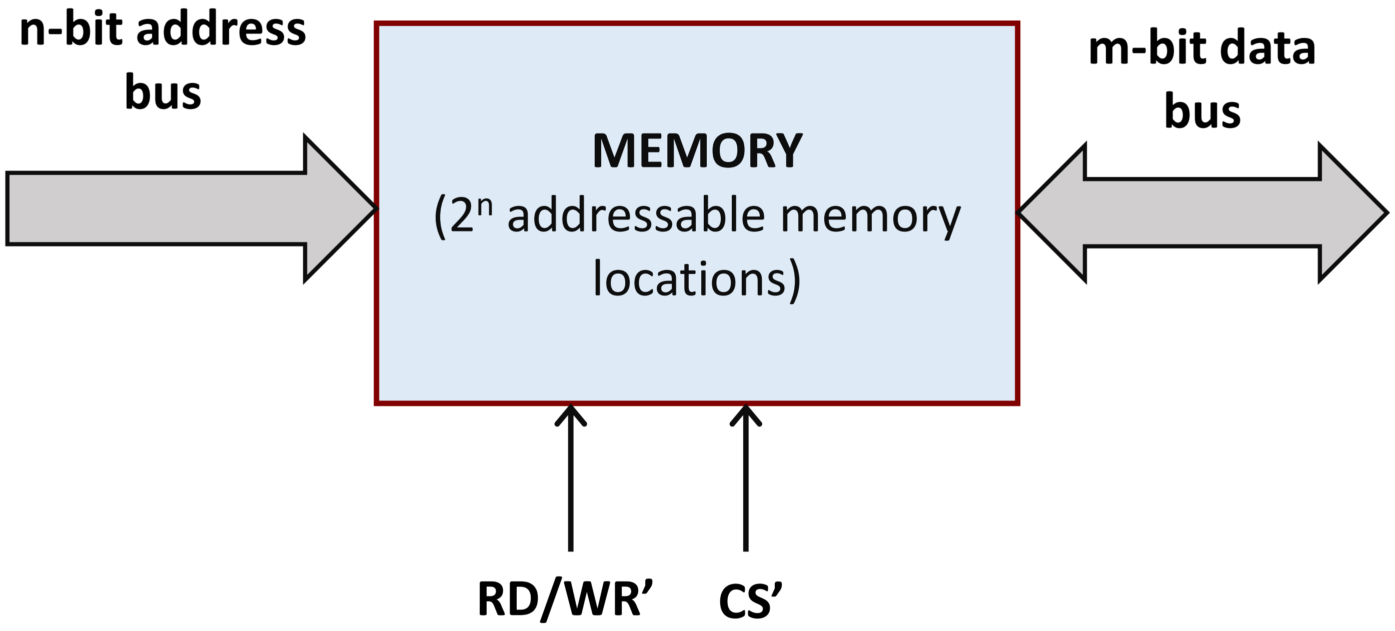
Connection between Processor and Memory

- Address bus provides the address of the memory location to be accessed.
- Data bus transfers the data read from memory, or data to be written into memory.
  - Bidirectional.
- Control bus provides various signals like READ, WRITE, etc.



An Example Memory Module

- *n address lines* :: The maximum number of memory locations that can be accessed is 2^n.
- *m data lines* :: The number of bits stored in every addressable location is *m*.
- The RD/WR’ control line selects the memory for reading or writing (1: read, 0: write).
- The chip select line (CS’) when active (=0) will enable the chip; otherwise, the data bus is in the ***high impedance state***.



The memory size is specified as 2^n x m

Classification of Memory Systems

a) Volatile versus Non-volatile:

- A *volatile memory* system is one where the stored data is lost when the power is switched off.
  - Examples: CMOS static memory, CMOS dynamic memory.
  - Dynamic memory in addition requires periodic refreshing.
- A *non-volatile memory* system is one where the stored data is retained even when the power is switched off.
  - Examples: Read-only memory, Magnetic disk, CDROM/DVD, Flash memory, Resistive memory.

b) Random-access versus Direct/Sequential access:

- A memory is said to be *random-access* when the read/write time is independent of the memory location being accessed.
  - Examples: CMOS memory (RAM and ROM).
- A memory is said to be *sequential access* when the stored data can only be accessed sequentially in a particular order.
  - Examples: Magnetic tape, Punched paper tape.
- A memory is said to be *direct* or *semi-random access* when part of the access is sequential and part is random.
  - Example: Magnetic disk.
  - We can directly go to a track after which access will be sequential.

c) Read-only versus Random-access:

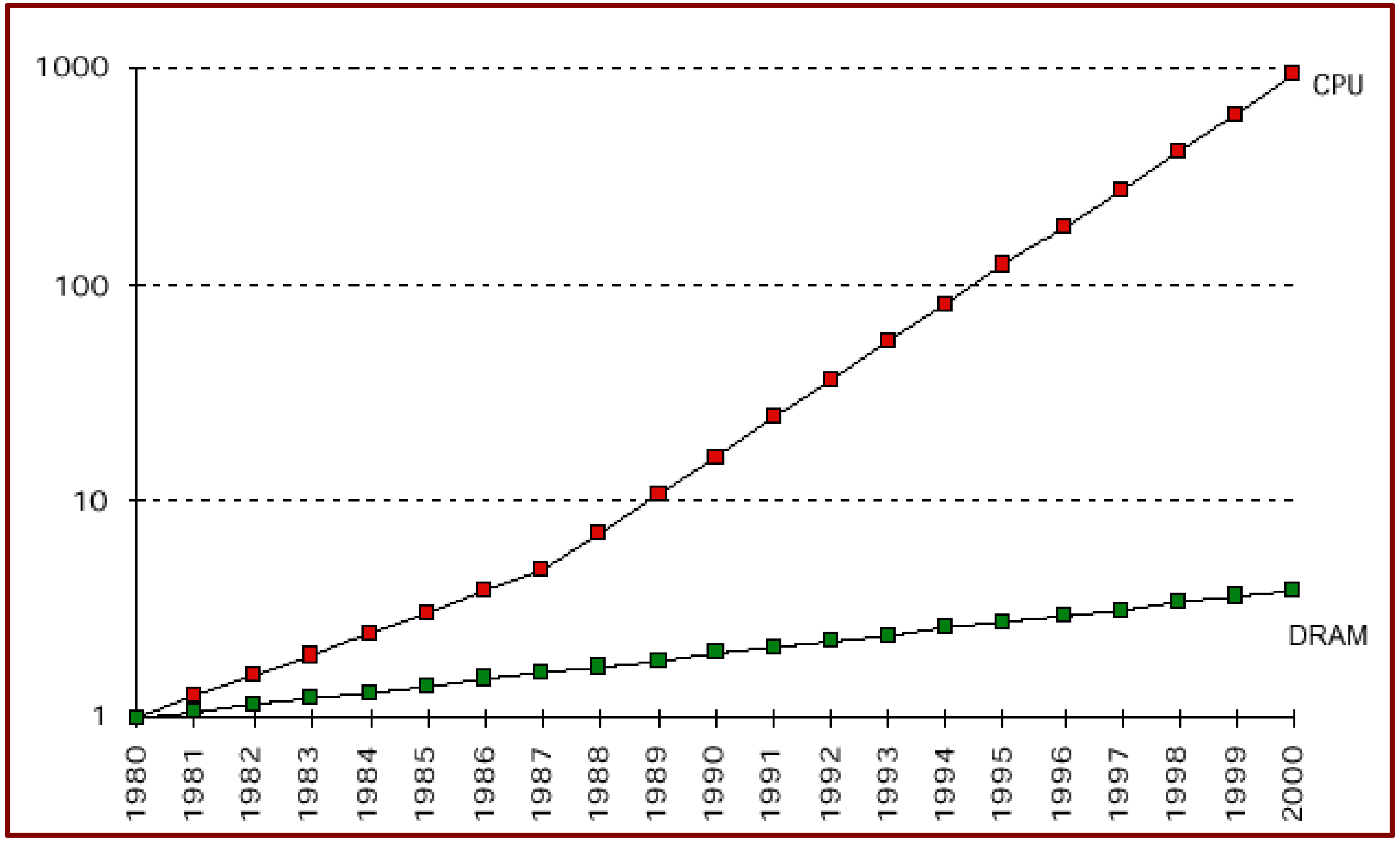
- *Read-only Memory* (ROM) is one where data once stored in permanent or semi-permanent.
  - Data written (programmed) during manufacture or in the laboratory.
  - Examples: ROM, PROM, EPROM, EEPROM.
- *Random Access Memory* (RAM) is one where data access time is the same independent of the location (address).
  - Can be read as well as written.
  - Used in main / cache memory systems.
  - Example: Static RAM (SRAM) → data once written are retained as long as power is on.
  - Example: Dynamic RAM (DRAM) → requires periodic refreshing even when power is on (data stored as charge on tiny capacitors).

Access Time, Latency and Bandwidth

- Terminologies used to measure speed of the memory system.
  a) **Memory Access Time**: Time between initiation of an operation (Read or Write) and completion of that operation.
  b) **Latency**: Initial delay from the initiation of an operation to the time the first data is available.
  c) **Bandwidth**: Maximum speed of data transfer in bytes per second.
- In modern memory organizations, every read request reads a block of words into some high-speed registers (*LATENCY*), from where data are supplied to the processor one by one (*ACCESS TIME*).

Design Issue of Memory System

- The most important issue is to bridge the processor-memory gap that has been widening with every passing year.
  - Advancements in memory technology are unable to cope with faster advancements in processor technology.

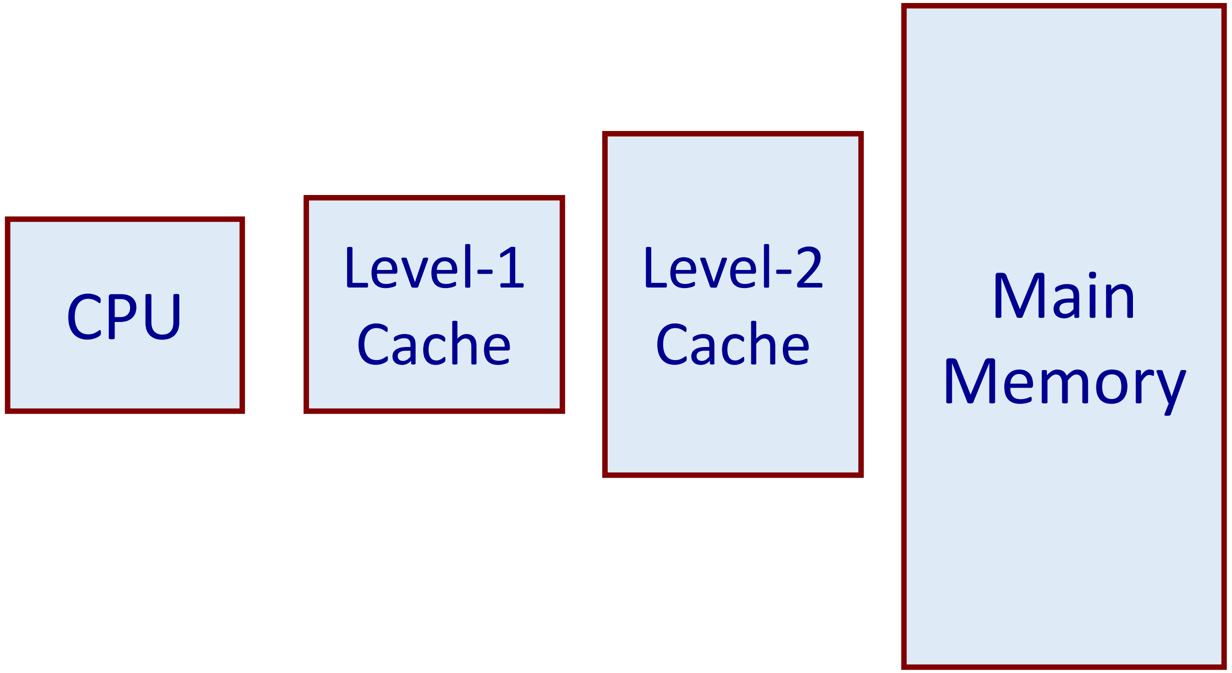


- Some important questions?
  - How to make the memory system work faster?
  - How to increase the data transfer rate between CPU and memory?
  - How to address the ever increasing storage needs of applications?

- Some possible solutions:
  - **Cache Memory**: to increase the effective speed of the memory system.
  - **Virtual Memory**: to increase the effective size of the memory system.

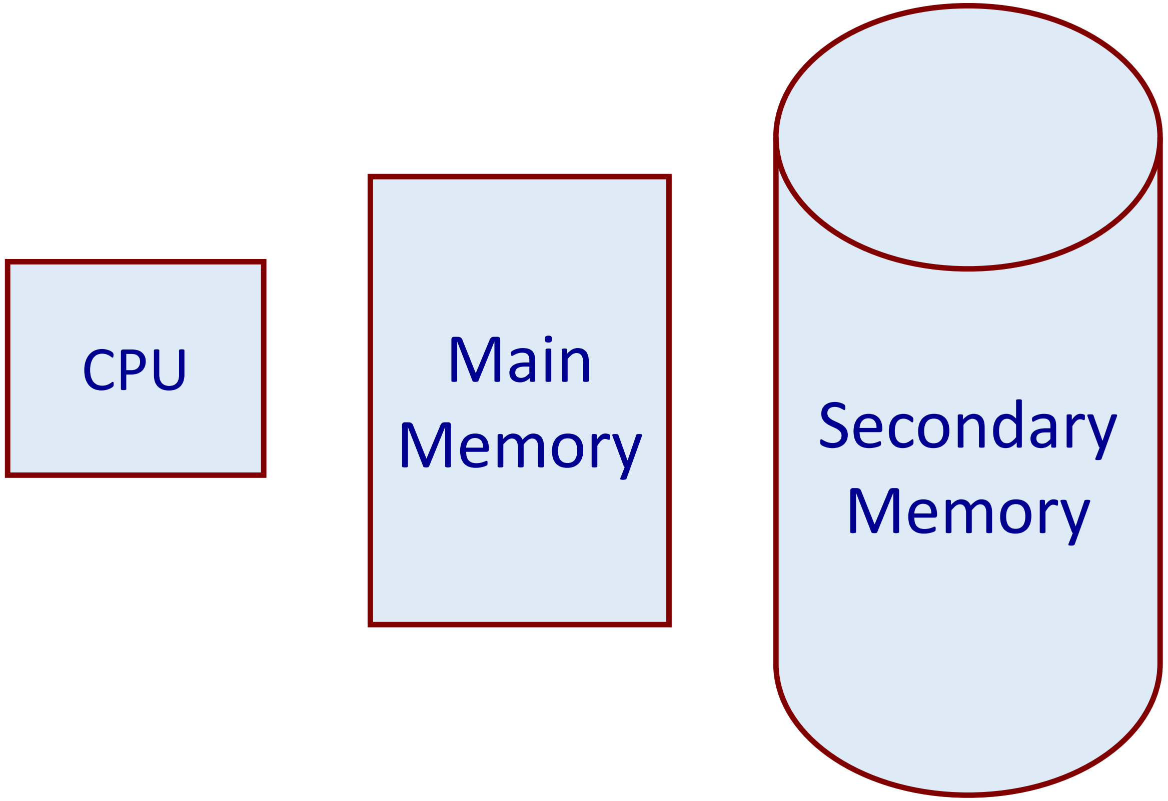
What is Cache Memory?

- A fast memory (possibly organized in several levels) that sits between processor and main memory.
- Faster than main memory and relatively small in capacity.
- Frequently accessed data and instructions are stored here.
- Cache memory makes use of the fast SRAM technology.



What is Virtual Memory?

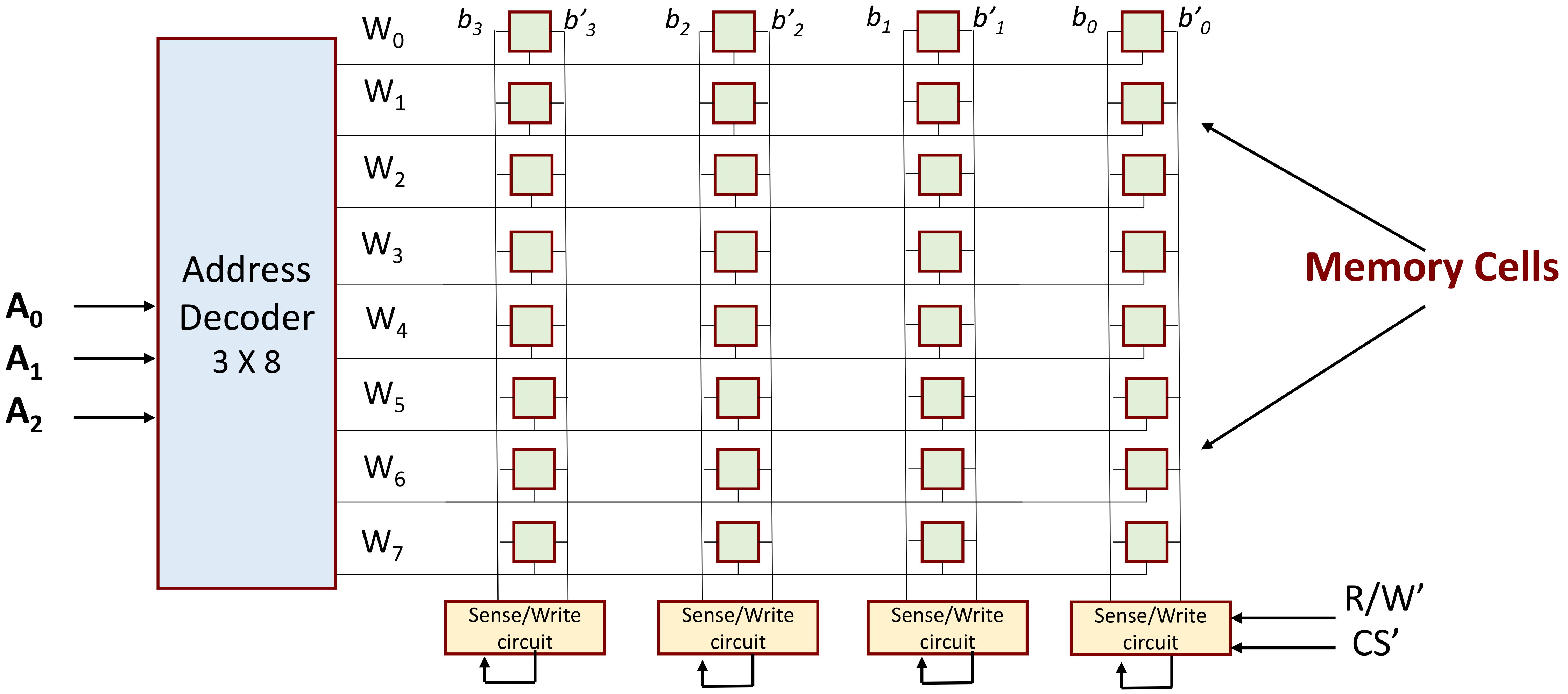
- Technique used by the operating system to provide an illusion of very large memory to the processor.
- Program and data are actually stored on secondary memory that is much larger.
- Transfer parts of program and data from secondary memory to main memory only when needed.



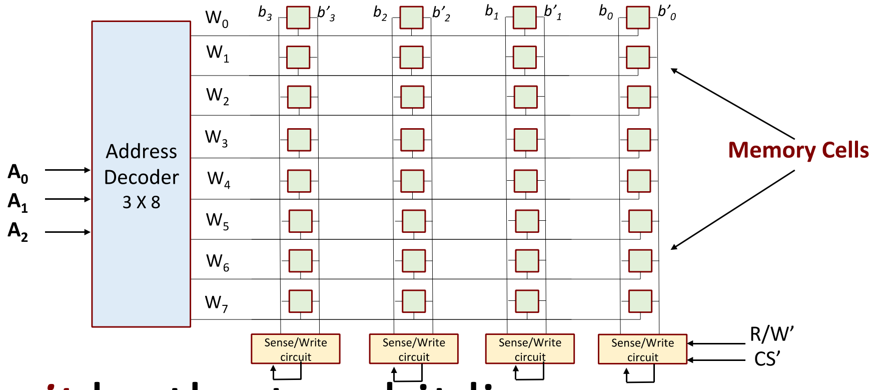
How does a Memory Chip look like?

- Memory cells are organized in the form of an array.
- Every memory cell holds one bit of data.
- Present-day VLSI technology allows one to pack billions of bits per chip.
- A memory module used in computers typically contains several such chips.

Organization of Cells in an 8x4 Memory Chip



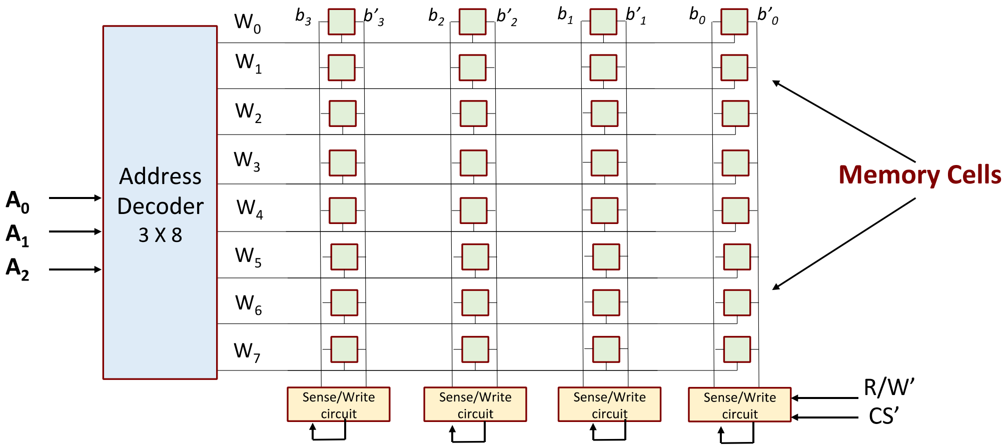
- A 32-bit memory chip organized as 8 x 4 is shown.
- Every row of the cell array constitutes a *memory word*.
- A 3 x 8 *decoder* is required to access any one of the 8 rows.
- The rows of the cells are connected to the *word lines*.
- Individual cells are connected to two *bit lines*.
    - Bit *b* and its complement *b'*.
    - Required for reading and writing.



- Cells in each column are connected to a *sense/write circuit* by the two bit lines.
- Other than address and data lines, there are two *control lines*: R/W' and CS' (Chip Select).
    - CS is required to select one single chip in a multi-chip memory system.

External Connection Requirements

- The 8 x 4 memory requires the following external connections:
  - Address decoder of size: 3 x 8
    - 3 external connections for address.
  - Data output : 4-bit
    - 4 external connections for data.
  - 2 external connections for R/W’ and CS’.
  - 2 external connections for power supply and
  - Total of 3 + 4 + 2 + 2 = 11.



What About a 256 X 16 Memory?

- Here the total number of external connections are estimated as follows:
    - Address decoder size: 8 x 256
        - 8 external connections for address.
    - Data output : 16-bit
        - 16 external connections for data.
    - 2 external connections for R/W’ and CS’.
    - 2 external connections for power supply and ground.
    - Total of 8 + 16 + 2 + 2 = 28.

STATIC AND DYNAMIC RAM

Introduction

- Broadly two types of semiconductor memory systems:
  a) Static Random Access Memory (SRAM)
  b) Dynamic Random Access Memory (DRAM)
     - Asynchronous DRAM
     - Synchronous DRAM
- Vary in terms of speed, density, volatility properties, and cost.
  - Present-day main memory systems are built using DRAM.
  - Cache memory systems are built using SRAM.

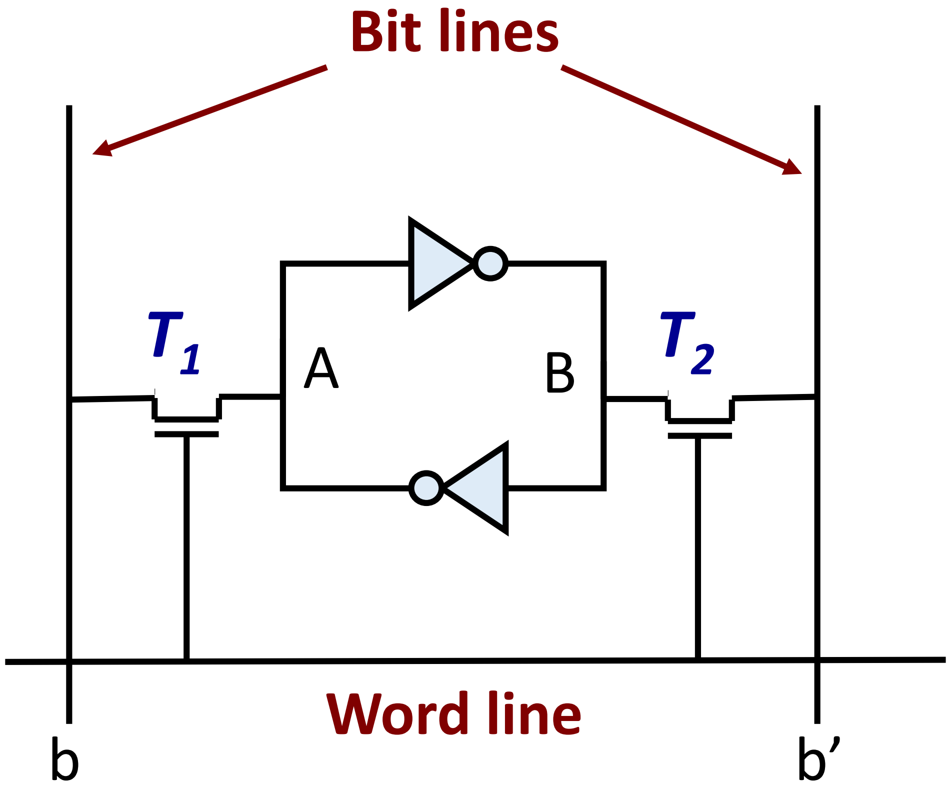
Static Random Access Memory (SRAM)

- SRAM consists of circuits which can store the data as long as power is applied.
- It is a type of semiconductor memory that uses bistable latching circuitry (flip-flop) to store each bit.
- SRAM memory arrays can be arranged in rows and columns of memory cells.
  - Called *word line* and *bit line*.

- SRAM technology:
  - Can be built using 4 or 6 MOS transistors.
  - Modern SRAM chips in the market uses 6-transistor implementations for CMOS compatibility.
  - Widely used in small-scale systems like microcontrollers and embedded systems.
  - Also used to implement cache memories in computer systems.
    - To be discussed later.

A 1-bit SRAM Cell

- Two inverters are cross connected to form a *latch*.
- The latch is connected to two bit lines with transistors *T1* and *T2*.
- Transistors behave like switches that can be opened (OFF) or closed (ON) under the control of the word line.
- To retain the state of the latch, the word line can be grounded which makes the transistors off.

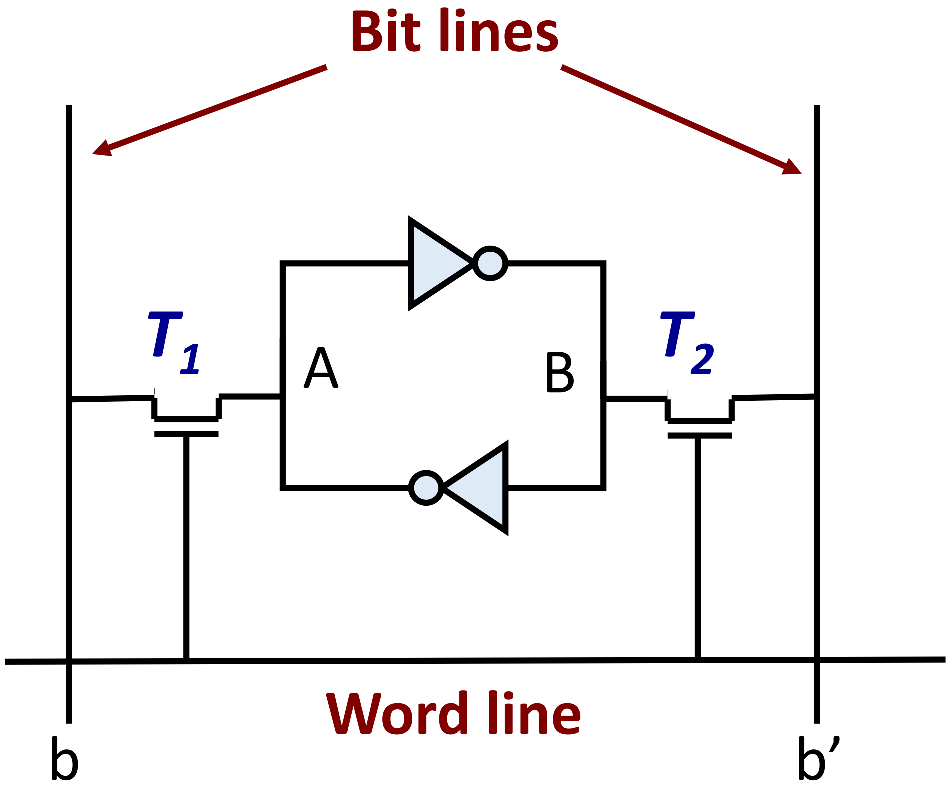


(a) READ Operation in SRAM

- To read the content of the cell, the word line is activated (= *1*) to make the transistors *T1* and *T2* on.

- The value stored in latch is available on bit line *b* and its complement on *b’*.

- Sense/write circuits connected to the bit lines monitor the states of *b* and *b’*.

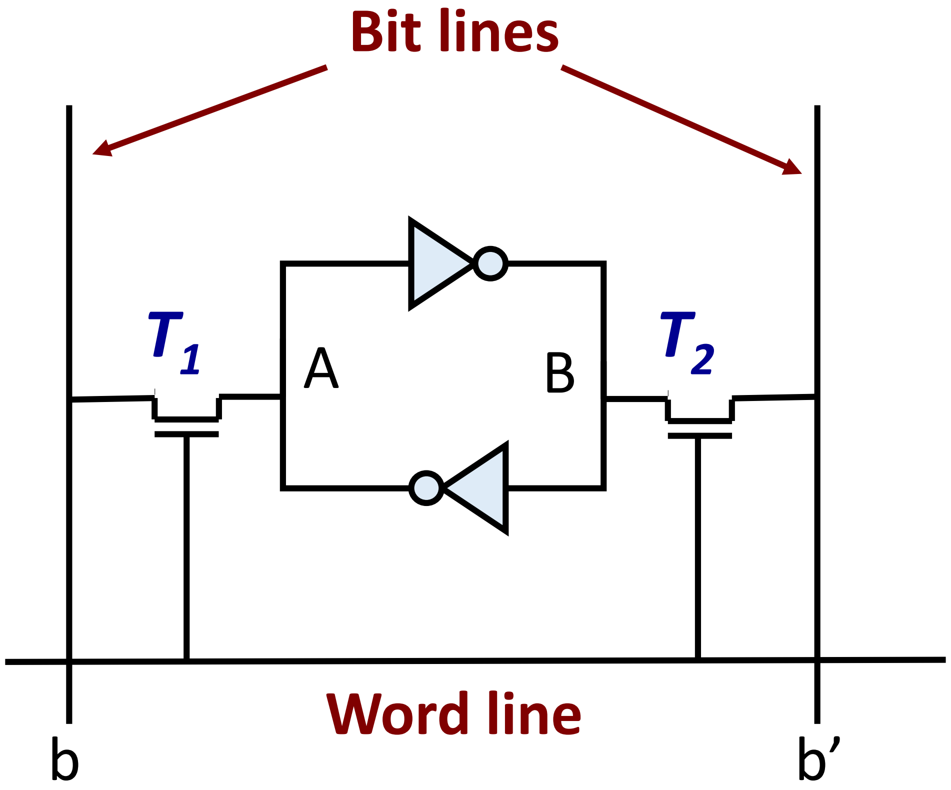


(b) WRITE Operation in SRAM

- **To write 1**: The bit line *b* is set with *1* and bit line *b'* is set with *0*. Then the word line is activated and the data is written to the latch.

- **To write 0**: The bit line *b* is set with *0* and bit line *b'* is set with *1*. Then the word line is activated and the data is written to the latch.

- The required signals (either *1* or *0*) are generated by the sense/write circuit.

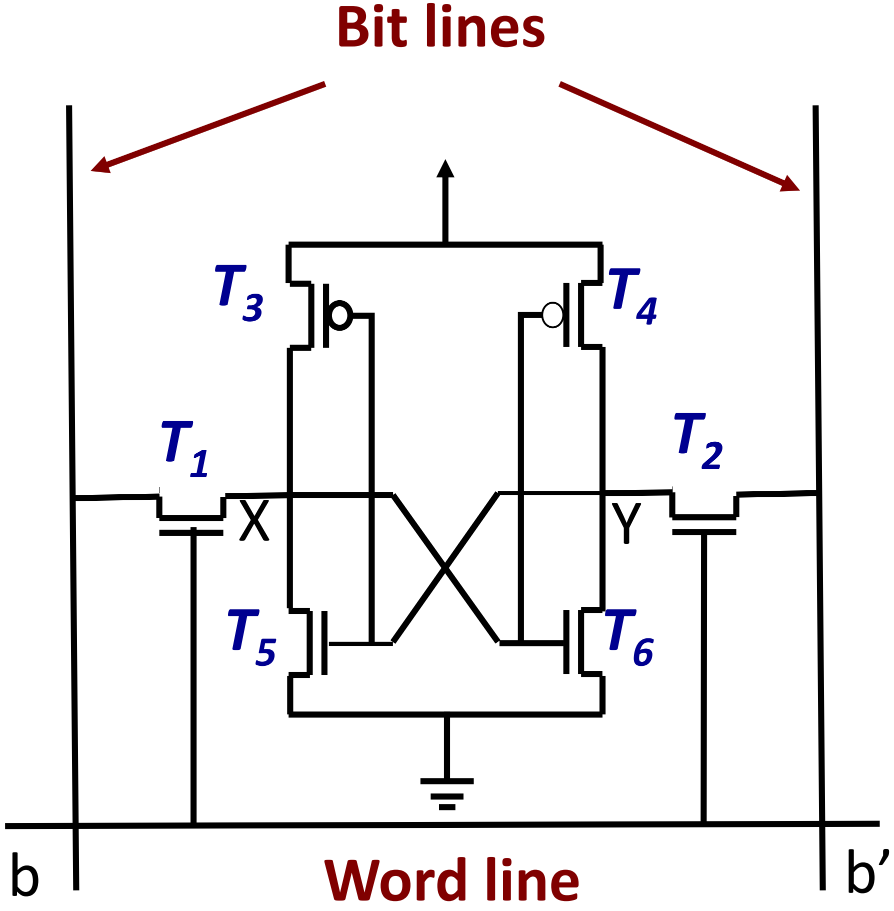


6-Transistor Static Memory cell

- 1-bit SRAM cell with 6-transistors are used in modern-day SRAM implementations.

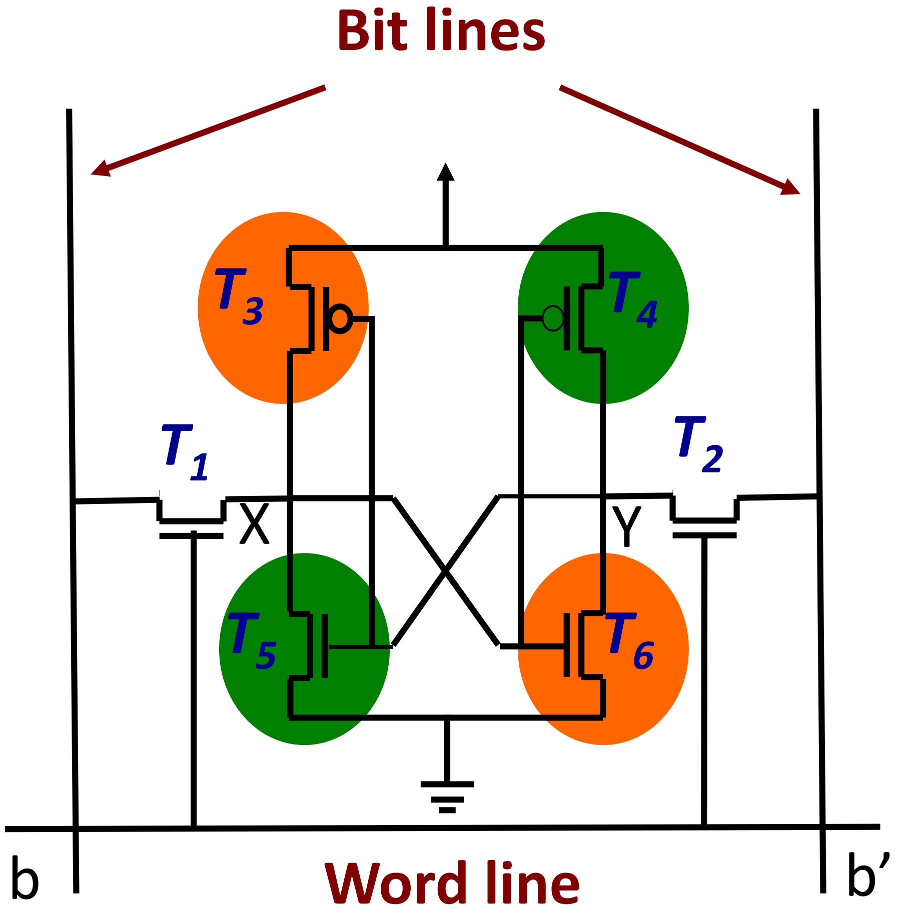
- Transistors (*T3* &*T5*) and (*T4* &*T6*) form the CMOS inverters in the latch.

- The data can be read or written in the same way as explained.



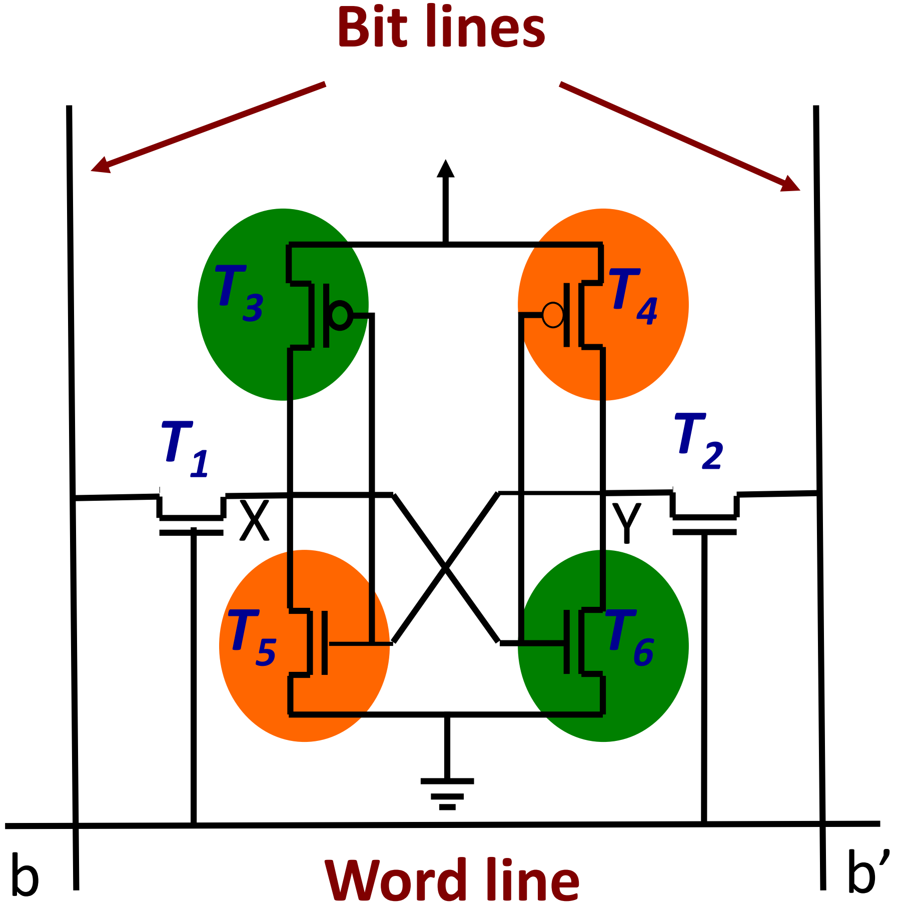
In State 0

- In state 0 the voltage at *X* is low and the voltage at *Y* is high.
- When the voltage at *X* is low, transistors (*T4* & *T5*) are on while (*T3* & *T6*) are off.
- When word line is activated, *T1* and *T2* are turned on and the bit lines *b* will have *0* and *b'* will have *1*.



In State 1

- In state 1 the voltage at *X* is high and the voltage at *Y* is low.
- When the voltage at *X* is high, transistors (*T3* & *T6*) are on while (*T4* & *T5*) are off.
- When word line is activated, *T1* and *T2* are turned on and the bit lines *b* will have *1* and *b'* will have *0*.

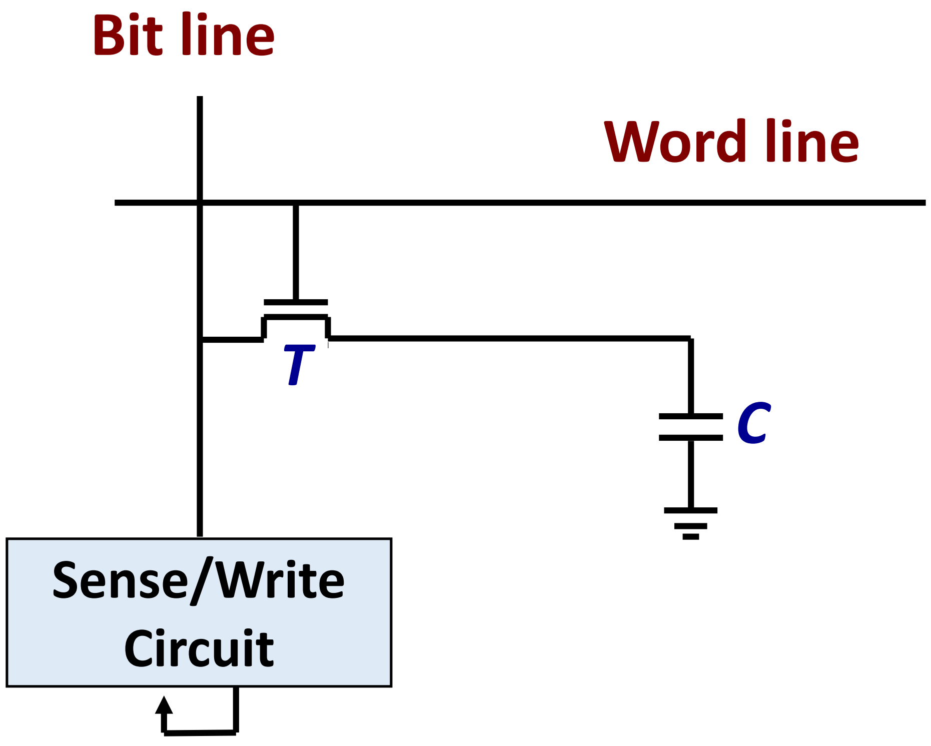


Features of SRAM

- Moderate / High power consumption.
  - Current flows in the cells only when the cell is accessed.
  - Because of latch operation, power consumption is higher than DRAM.
- Simplicity – refresh circuitry is not needed.
  - Volatile :: continuous power supply is required.
- Fast operation.
  - Access time is very fast; fast memories (cache) are built using SRAM.
- High cost.
  - 6 transistors per cell.
- Limited capacity.
  - Not economical to manufacture high-capacity SRAM chips.

Dynamic Random Access Memory (DRAM)

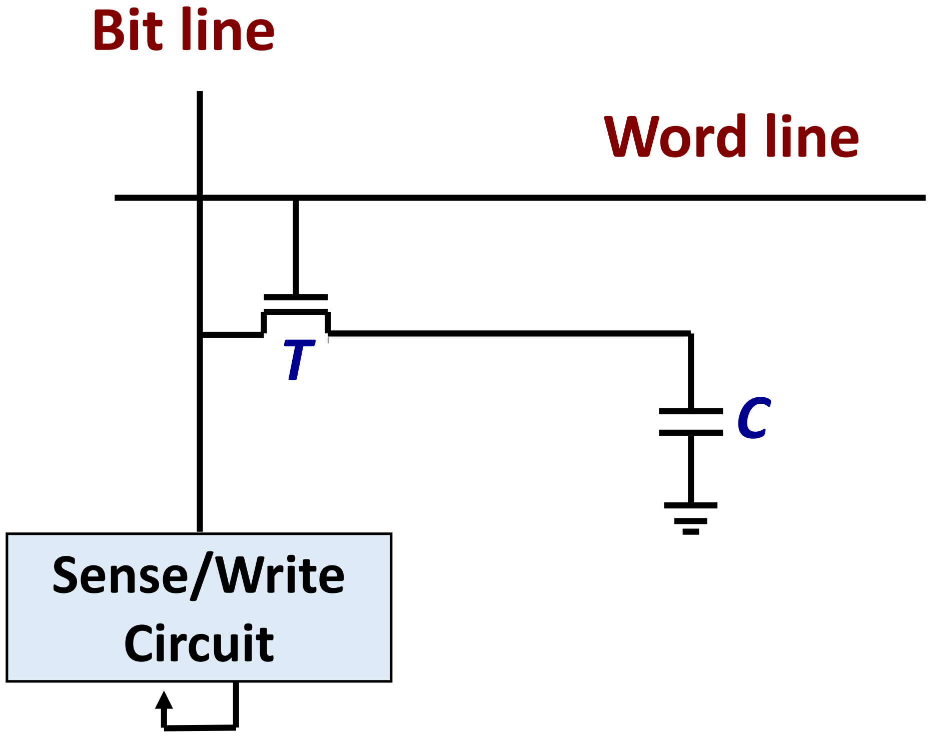
- Dynamic RAM do not retain its state even if power supply is on.
  - Data stored in the form of charge stored on a capacitor.
- Requires *periodic refresh*.
  - The charge stored cannot be retained over long time (due to leakage).
- Less expensive that SRAM.
  - Requires less hardware (one transistor and one capacitor per cell).
- Address lines are multiplexed.



1-transistor DRAM Cell

(a) READ Operation in DRAM

- The transistor of the particular cell is turned on by activating the word line.
- A sense amplifier connected to bit line senses the charge stored in the capacitor.
- If the charge is above threshold, the bit line is maintained at high voltage, which represents logic *1*.
- If the charge is below threshold, the bit line is grounded, which represent logic *0*.



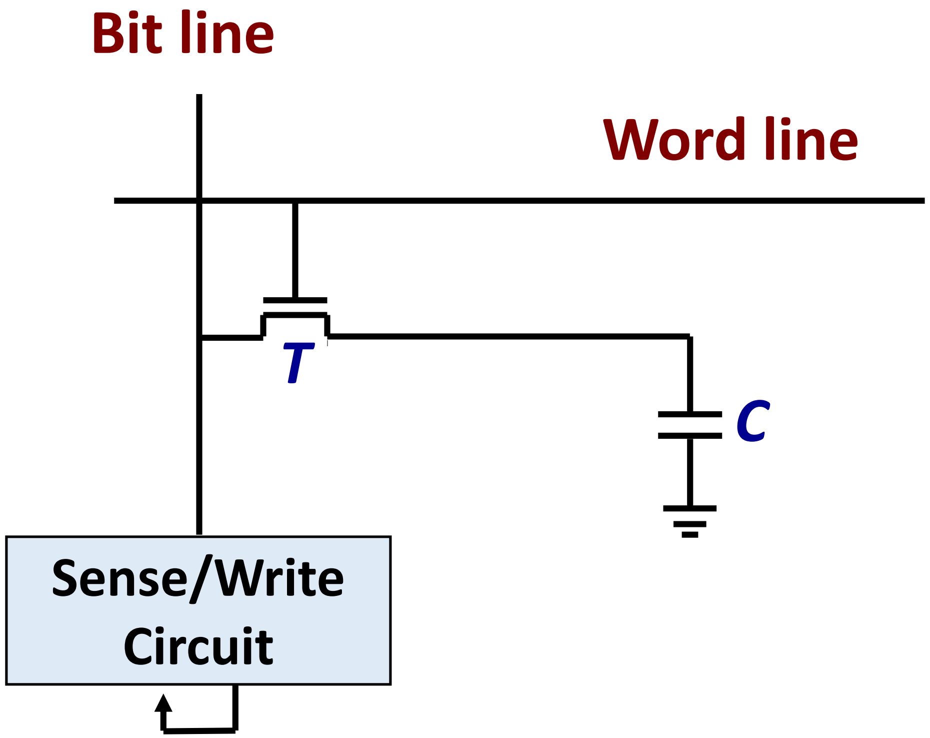
(b) WRITE Operation in DRAM

- The transistor of the particular cell is turned on by activating the word line.

- Depending on the value to be written (*0* or *1*), an appropriate voltage is applied to the bit line.

- The capacitor gets charged to the required voltage state.

- Refreshing of the capacitor requires periodic READ-WRITE cycles (every few msec).



Types of DRAM

a) Asynchronous DRAM (ADRAM)

- Timing of the memory device is handled asynchronously.
- A special memory controller circuit generates the signals asynchronously.
- DRAM chips produced between the early 1970s to mid-1990s used *asynchronous* DRAM.

b) Synchronous DRAM (SDRAM)

- Memory operations are synchronized by a clock.
- Concept of SDRAM came in the 1970s.
- Commercially made available only in 1993 by Samsung.
- By 2000 SDRAM replaced almost all types of DRAMs in the market.
- Performance of SDRAM is much higher compared to all other existing DRAM.

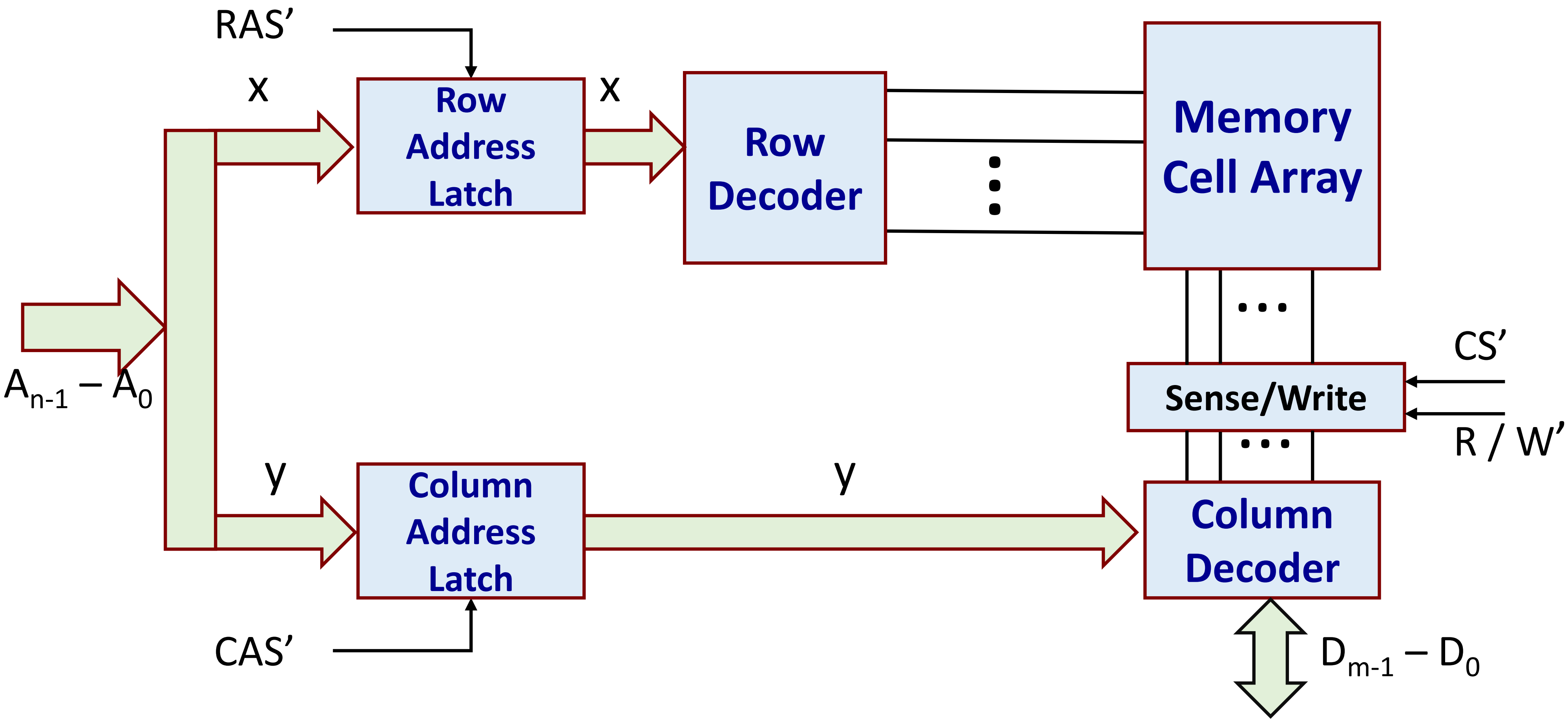
Asynchronous DRAM

Asynchronous DRAM

- The timing of the memory device is controlled asynchronously.
- The device connected to this memory is responsible for the delay.
- Address lines are divided into two parts and multiplexed.
    - Upper half of address:
        - Loaded into *Row Address Latch* using *Row Address Strobe* (RAS).
    - Lower half of address:
        - Loaded into *Column Address Latch* using *Column Address Strobe* (CAS).

Internal Organization of a DRAM Chip

- Cells are organized in the form of an array, in rows and columns.
- Cells of each row are divided into fixed number of columns, *m* bits each.
- *m* is 8, 16, 32 or 64.



- Suppose that the memory cell array is organized as $r \times c$.
  - r rows and c columns.
- An x-bit address is required to select a row r, where $x = log_2 r$.
- An y-bit address is required to select a column c, where $y = log_2 c$.
- Total address bits: *$n = x$ (high order) + y (low order)*

READ or WRITE Operation

- For a read operation, the *x*-bit row address is applied first.
  - It is loaded into *Row Address Latch* in response to the signal *RAS’*.
  - The read operation is performed in which all the cells of the selected row are read and refreshed.

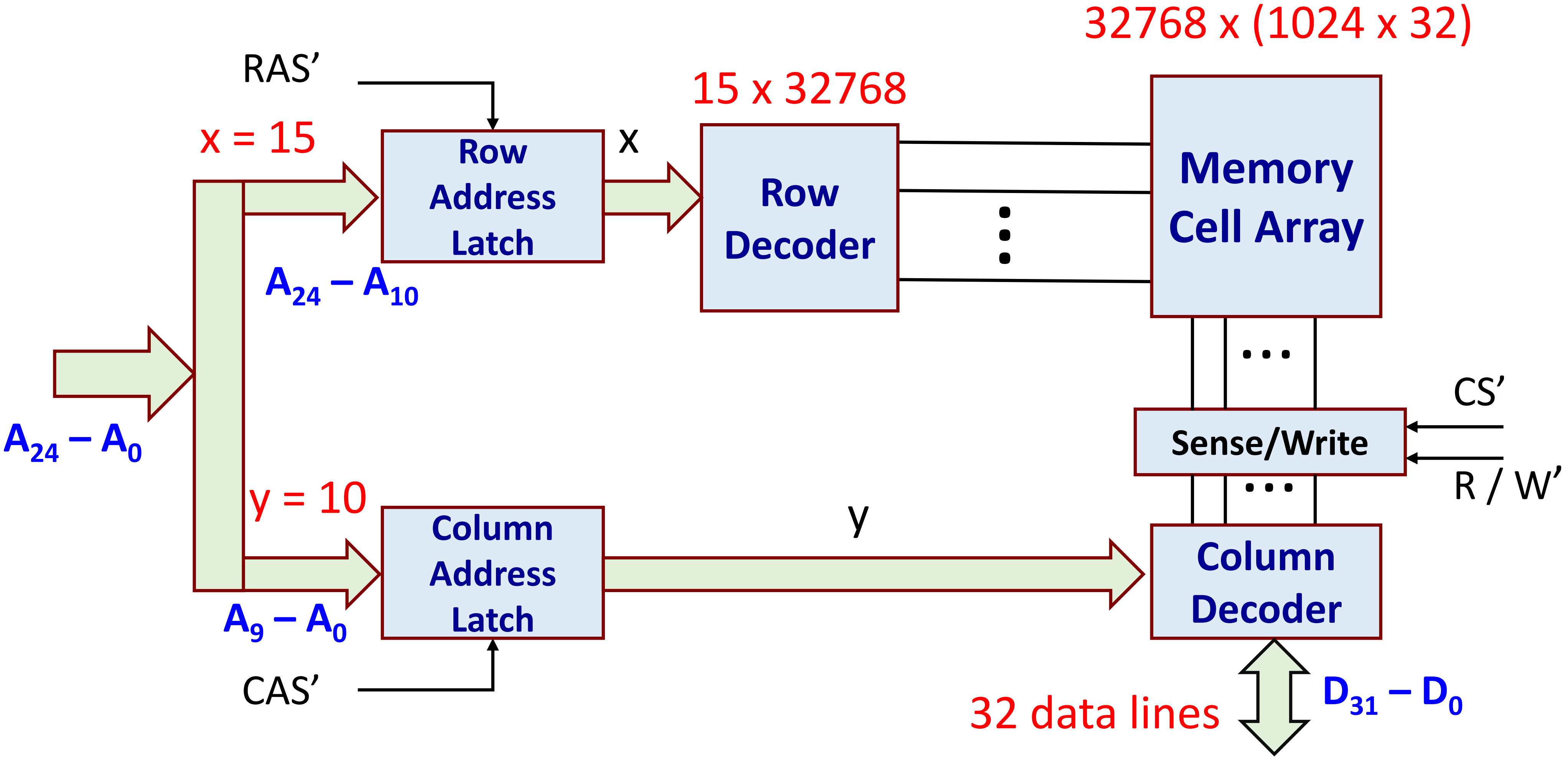
- After loading of row address, the column address is selected.
- In response to *CAS’* the column address is loaded into *Column Address Latch*.
- Then the column decoder selects a particular column from *c* columns and an appropriate group of *m* sense/write circuits are selected.

- For a READ operation, the output values of the selected circuits are transferred to data lines D_{m-1} to $D_{0.}$
- For a WRITE operation, the data available on the data lines D_{m-1} to D_0 is transferred to the selected circuits.
  - This information is stored in the selected cell.
- Both *RAS'* and *CAS'* are active low signals. That is they cause latching the addresses when they move from high to low.

- Each row of the cell array must be periodically refreshed to prevent data loss.
- Cost is low but access time is high compared to SRAM.
- Very high packing density (few billion cells per chip).
- Widely used in the main memory of modern computer systems.

An Example: 1 Gbit ADRAM Chip

- We assume that the 1 Gbit memory cells are organized as 32768 (2^{15}) rows and 32768 (2^{15}) columns.
- Let us assume that data bus is 32-bit long.
- So, the memory can be organized as (2^{15}) x (2^{10} x 2^{5}).
  - Total number of address lines is 25 bits.
- High order 15 bits of the address is used to select a row.
  - Requires a 15 x 32768 row-address decoder.
- Low order 10 bits of the address is used to select a column.
  - Requires a 10 x 1024 column decoder.


32768 x (1024 x 32)
RAS’
15 x 32768
x = 15
Row Address Latch
x
Row Decoder
Memory Cell Array
$A_{24} - A_{10}$
$A_{24} - A_{0}$
CS’
Sense/Write
R / W’
y = 10
y
Column Address Latch
Column Decoder
$A_{9} - A_{0}$
CAS’
32 data lines
$D_{31} - D_{0}$

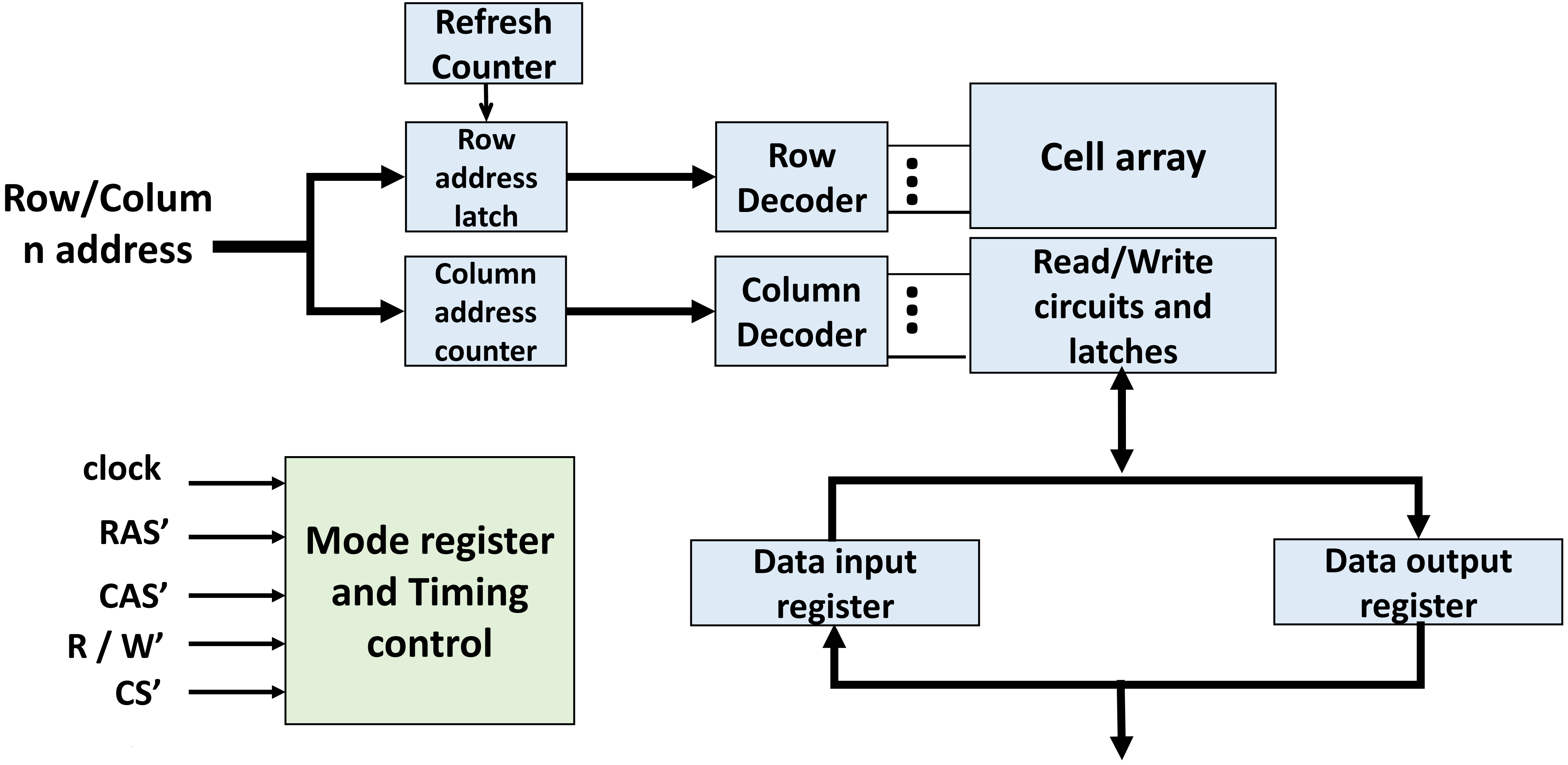
- **Operation:**
  - 15-bit row address is selected (i.e., x = 15).
  - With the help of RAS control signal the row address is latched. The 15 x 32768 Row Decoder selects a particular row.
  - Then the 10-bit column address is applied and with the help of CAS the address is latched. The 10 x 1024 column decoder selects a particular column.
  - A group of 32 bits are selected as the 32-bit word to be accessed.

Synchronous DRAM

Synchronous DRAM

- SDRAM is the commonly used name for various kinds of dynamic RAM that are synchronized with clock.
- The structure of this memory is same as asynchronous DRAM.
- The concept of SDRAM were known from 70’s but it is first developed by Samsung in the year 1993 (KM48SL2000).
    - By 2000 all kinds of DRAM were replaced by SDRAM.

Internal Organization of a SDRAM Chip



- In SDRAM address and data connections are buffered by registers.
- The output of individual sense amplifier is connected to a latch.
- Mode register is present which can be set to operate the memory chip in different modes.
- To select successive columns it is not required to provide externally generated pulses on CAS line.
- A column counter is used internally to generate the required signals.

READ and WRITE Operations

- For READ operation, the row address is applied first, and in response to the column address, the data present in the latches for the selected columns are transferred to the data output register.
  - Then the data is available on the data bus.
- For WRITE operation, the row address is applied first, and in response to the column address, the data present in the data bus is made available to the latches through data input register.
  - The data is then written to the particular cell.

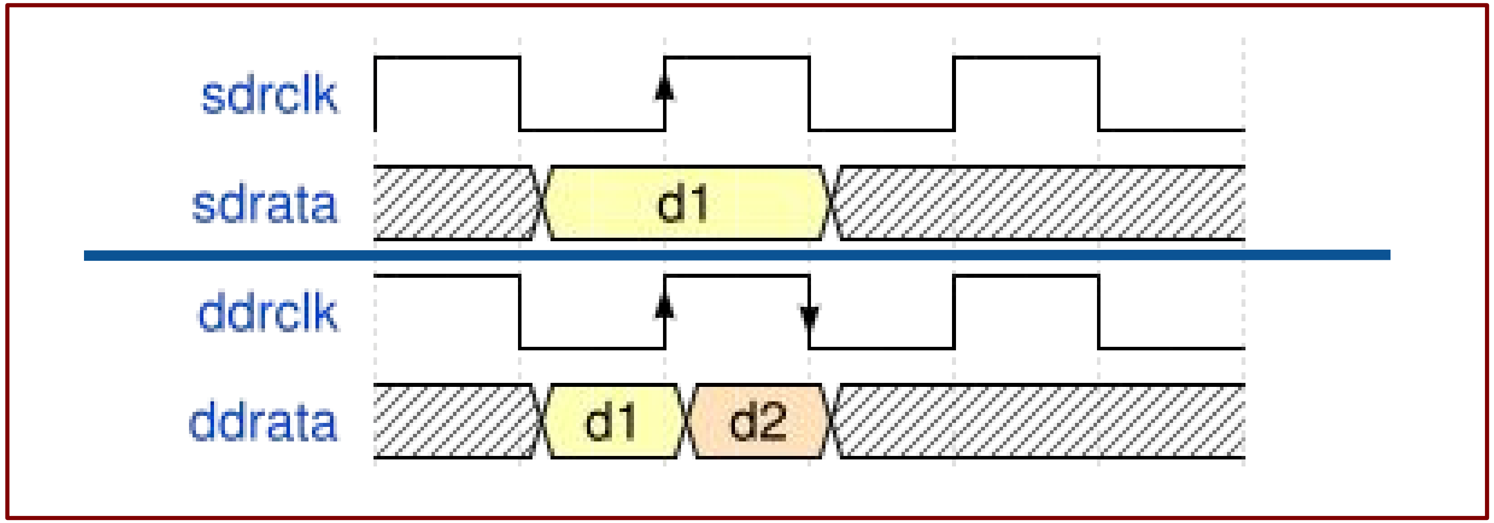
Types of SDRAM

- Single data rate SDRAM (called SDR) can accept one command and transfer one word of data per clock cycle.
  - Data transferred typically on the rising edge of the clock.

- Double data rate SDRAM (called DDR) transfers data on both the rising and falling edges of the clock.

- DDR SDRAM was launched in 2000.

- DDR2 (2003), DDR3 (2007), DDR4 (2014).



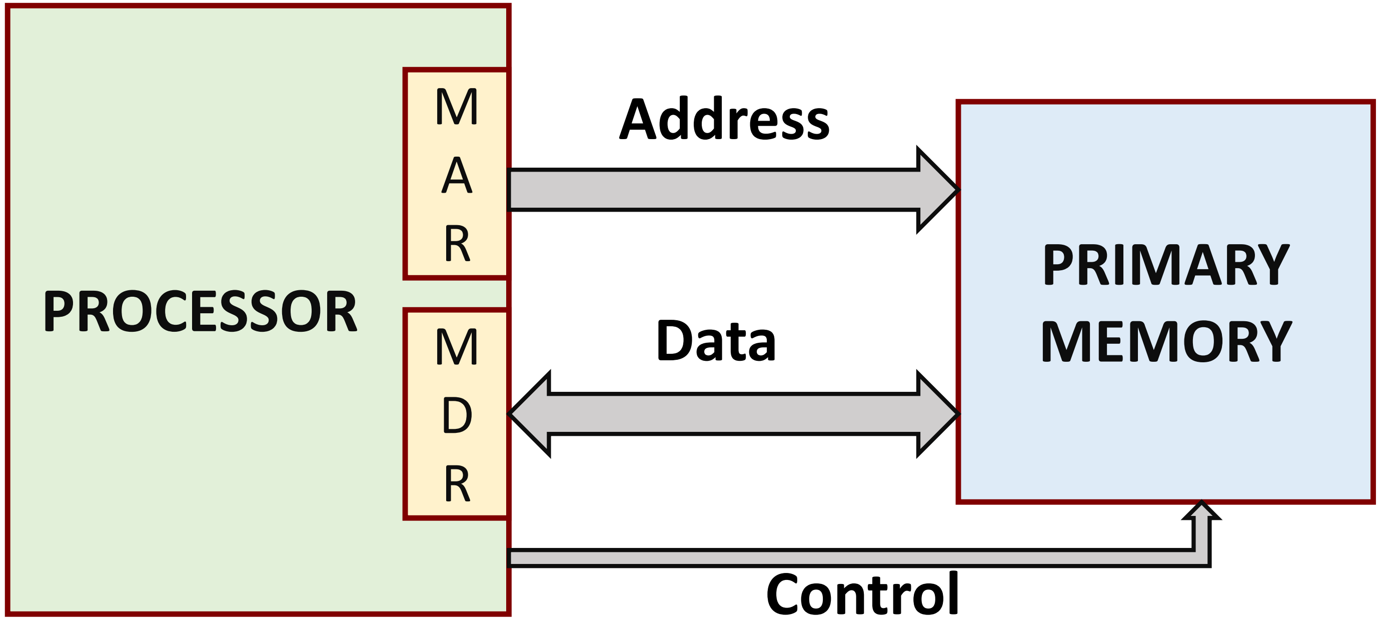
Speed of DDR Memories Across Generations

Year	Chip size	Type	Slowest DRAM	Fastest DRAM	CAS transfer time	Cycle time
2000	256 Mb	DDR1	65 ns	45 ns	7 ns	90 ns
2002	512 Mb	DDR1	60 ns	40 ns	5 ns	80 ns
2004	1 Gb	DDR2	55 ns	35 ns	5 ns	70 ns
2006	2 Gb	DDR2	50 ns	30 ns	2.5 ns	60 ns
2010	4 Gb	DDR3	36 ns	28 ns	1 ns	37 ns
2012	8 Gb	DDR3	30 ns	24 ns	0.5 ns	31 ns

Memory Interfacing and Addressing

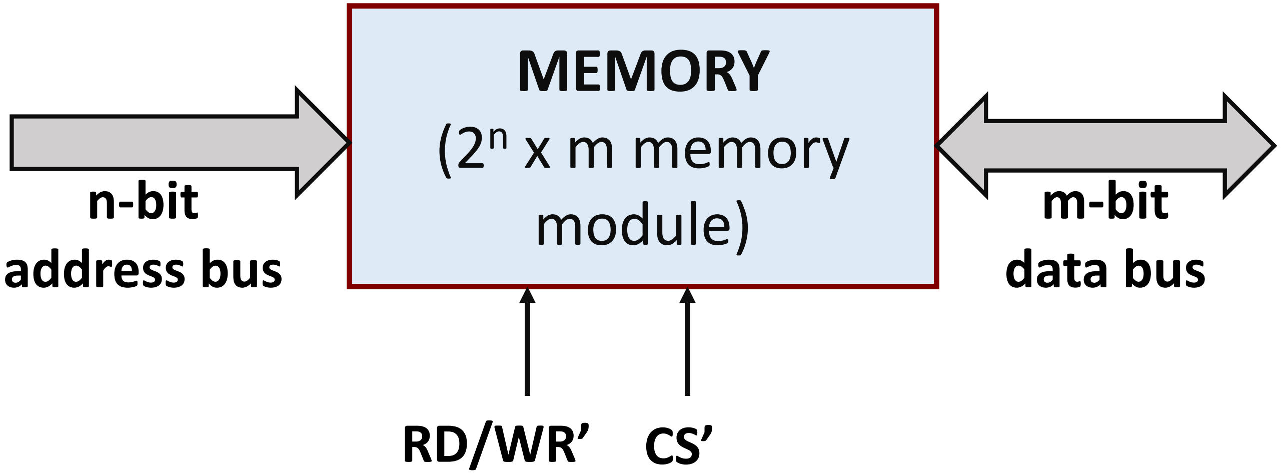
Memory Interfacing

- Basic problem:
  - Interfacing one of more memory modules to the processor.
  - We assume a single level memory at present (i.e. no cache memory).
- Questions to be answered:
  - How the processor address and data lines are connected to memory modules?
  - How are the addresses decoded?
  - How are the memory addresses distributed among the memory modules?
  - How to speed up data transfer rate between processor and memory?



The processor's view of memory

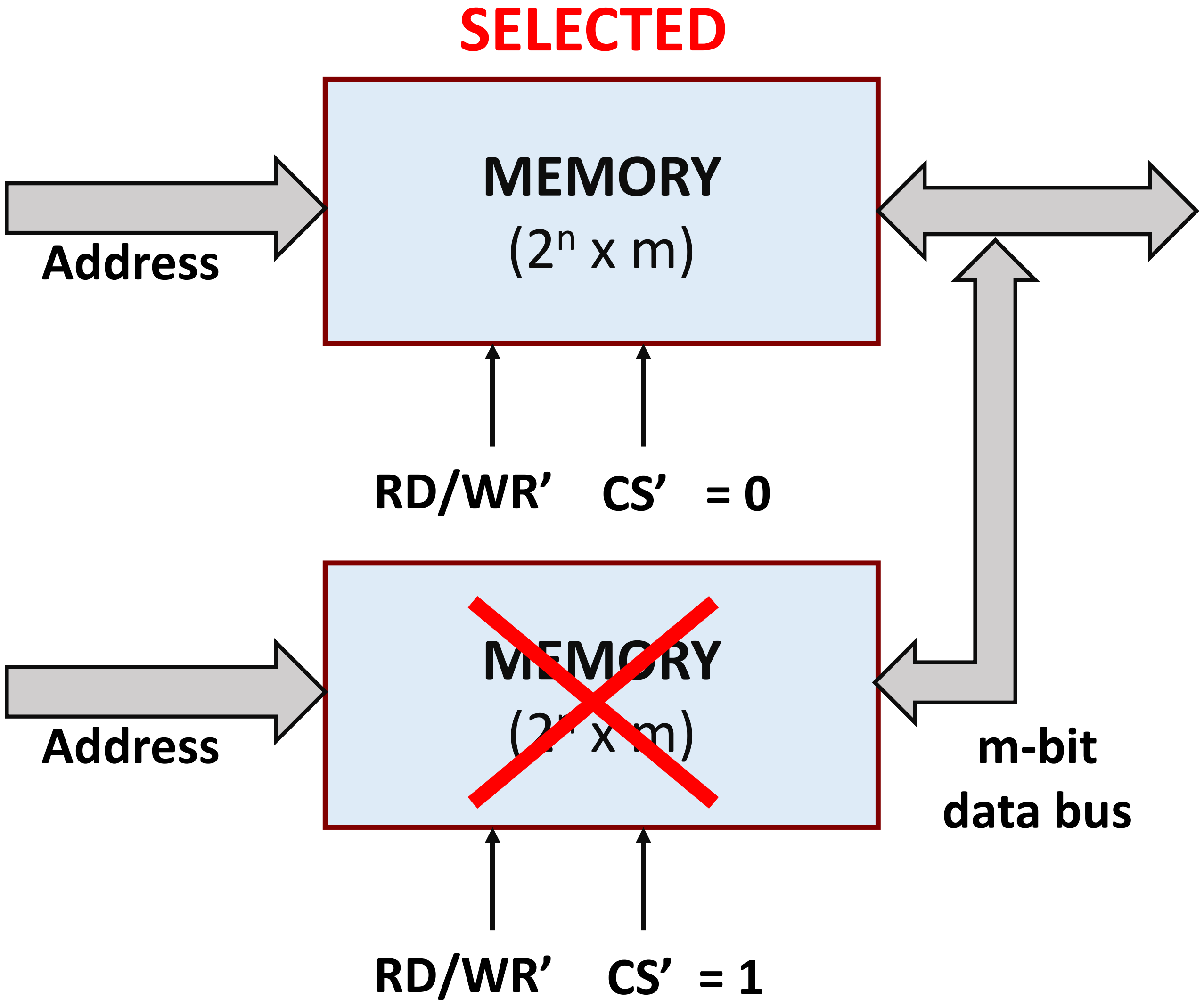
- Typical interface of a memory module.
- Real chip may contain more signal lines (e.g. DRAM).



A Note About the Memory Interface Signals

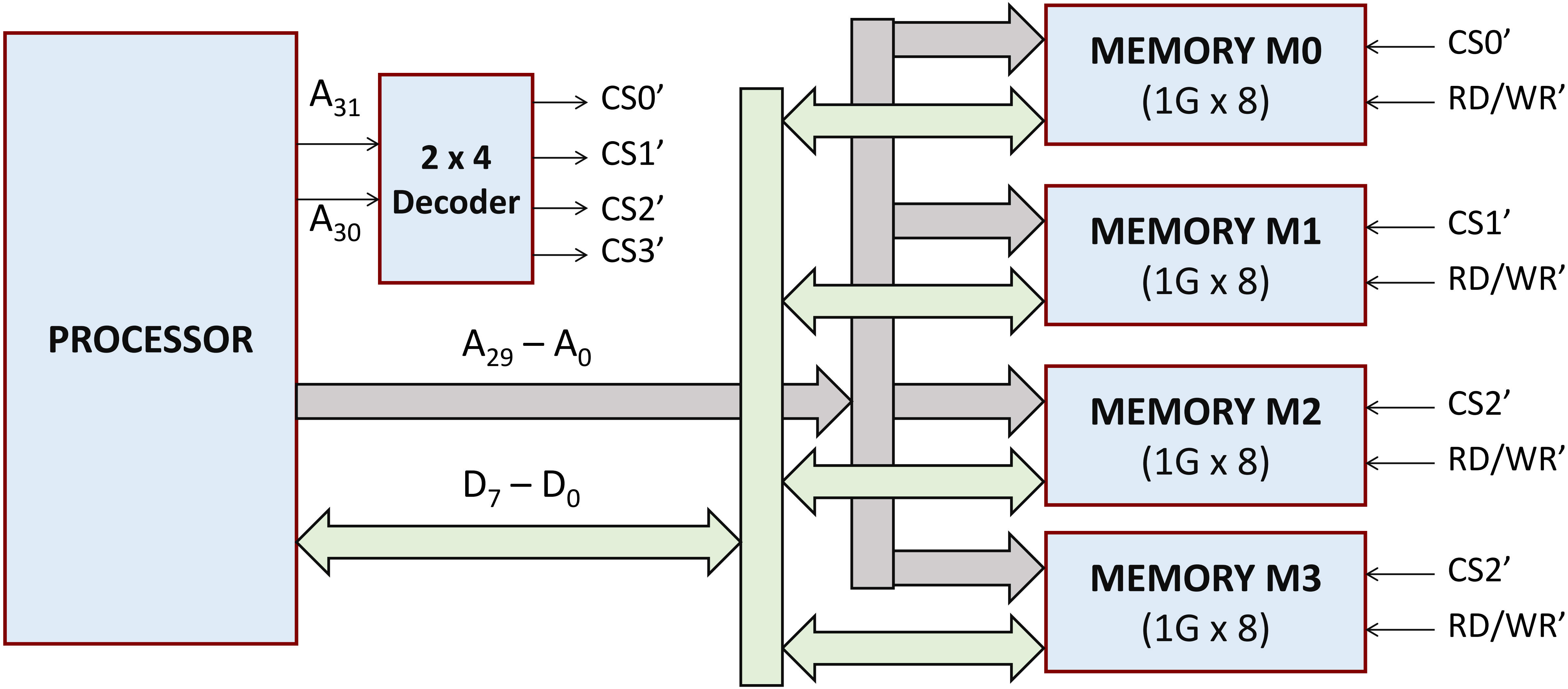
- The data signals of a memory module (RAM) are typically bidirectional.
  - Some memory chips may have separate data in and data out lines.
- For memory READ operation:
  - Address of memory location is applied to address lines.
  - RD/WR’ control signal is set to 1, and CS’ is set to 0.
  - Data is read out through the data lines after memory access time delay.
- For memory WRITE operation:
  - Address of memory location is applied to address lines, and the data to be written to data lines.
  - RD/WR’ control signal is set to 0, and CS’ is set to 0.

- Why is CS’ signal required?
  - To handle multiple memory modules interfacing problem.
  - We typically select only one out of several memory modules at a time.
- What happens when CS’ = 1?
  - When a memory module is *not selected*, the data lines are set to the high impedance state (i.e. electrically disconnected).
  - An example scenario is shown.



An Example Memory Interfacing Problem

- Consider a MIPS32 like processor with a 32-bit address.
  - Maximum memory that can be connected is 2^{32} = 4 Gbytes.
  - Assume that the processor data lines are 8 bits.
- Assume that memory chips (RAM) are available with *size 1 Gbyte*.
  - 30 address lines and 8 data lines.
  - Low-order 30 address lines (A_{29}-A_0) are connected to the memory modules.
- We want to interface *4 such chips* to the processor.
  - Total memory of 4 Gbytes.


PROCESSOR
A_{31}
A_{30}
2 x 4
Decoder
CS0'
CS1'
CS2'
CS3'
$A_{29} – A_0$
$D_7 – D_0$
MEMORY M0
(1G x 8)
CS0'
RD/WR'
MEMORY M1
(1G x 8)
CS1'
RD/WR'
MEMORY M2
(1G x 8)
CS2'
RD/WR'
MEMORY M3
(1G x 8)
CS2'
RD/WR'

- High order address lines (A_{31} and A_{30}) select one of the memory modules.

- **When is M0 selected?**
  - Address is: 0 0 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
  - Range of addresses is: 0x00000000 to 0x3FFFFFFF

- **When is M1 selected?**
  - Address is: 0 1 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
  - Range of addresses is: 0x40000000 to 0x7FFFFFFF

- **When is M2 selected?**
  - Address is: 1 0 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
  - Range of addresses is: 0x80000000 to 0xBFFFFFFF

- **When is M3 selected?**
  - Address is: 1 1 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
  - Range of addresses is: 0xC0000000 to 0xFFFFFFFF

- **An observation:**
  - Consecutive block of bytes are mapped to the same memory module.
  - For MIPS32, we have to access 32 bits (4 bytes) of data in parallel, which requires four sequential memory accesses here.
  - We shall look at an alternate memory organization later that would make this possible.
    - Called *memory interleaving*.

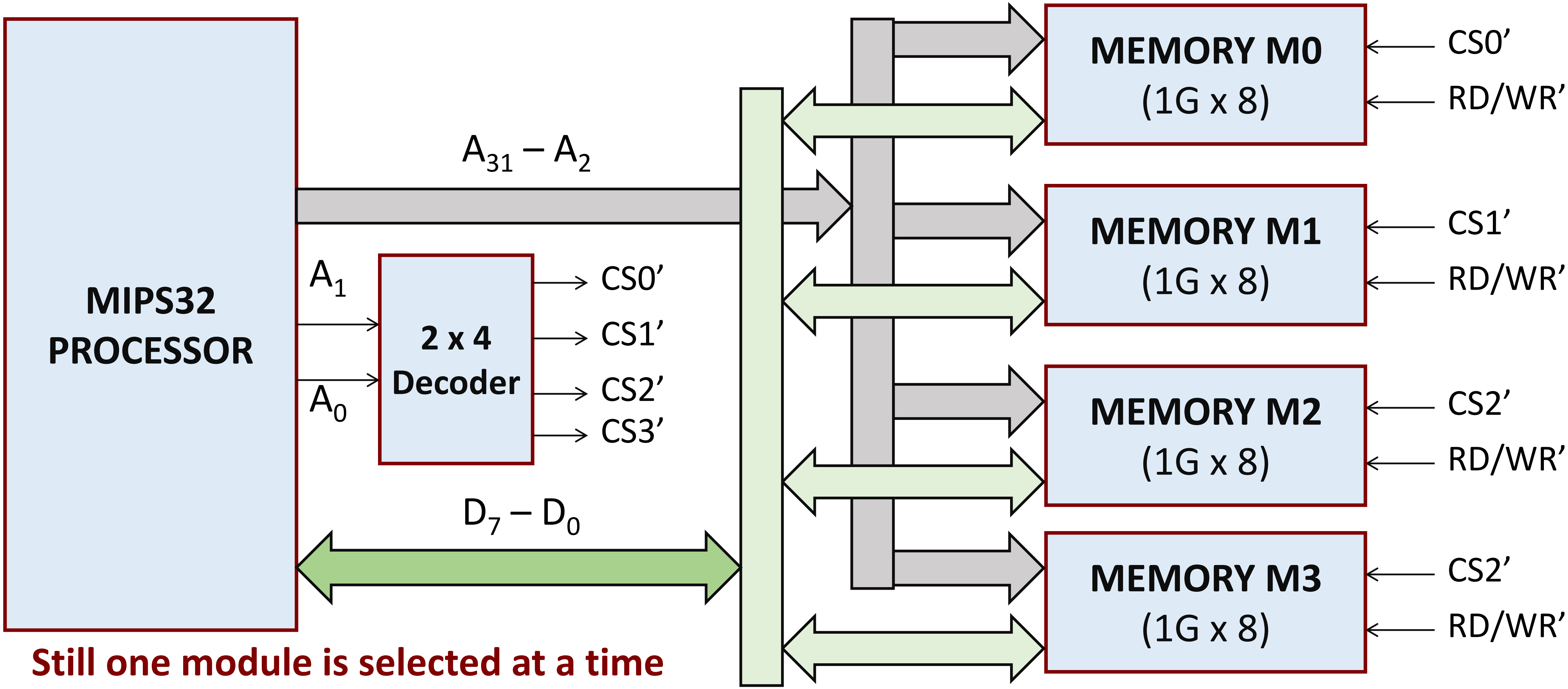
Improved Memory Interface for MIPS32

- We make small changes in the organization so that 32-bits of data can be fetched in a single memory access cycle.
  - Exploit the concept of memory interleaving.
- The main changes:
  - High order 30 address lines (A_{31}-A_2) are connected to memory modules.
  - Low order two address lines (A_1 and A_0) are used to select one of the modules.

- How are the addresses mapped to memory modules?
  - ***Module M0***: 0, 4, 8, 12, 16, 20, 24, …
  - ***Module M1***: 1, 5, 9, 13, 17, 21, 25, …
  - ***Module M2***: 2, 6, 10, 14, 18, 22, 26, …
  - ***Module M3***: 3, 7, 11, 15, 19, 23, 27, …
- Memory addresses are *interleaved* across memory modules.
- What we can gain from this mapping?
  - Consecutive addresses are mapped to consecutive modules.
  - Possible to access four consecutive words in the same cycle, if all four modules are enabled simultaneously.

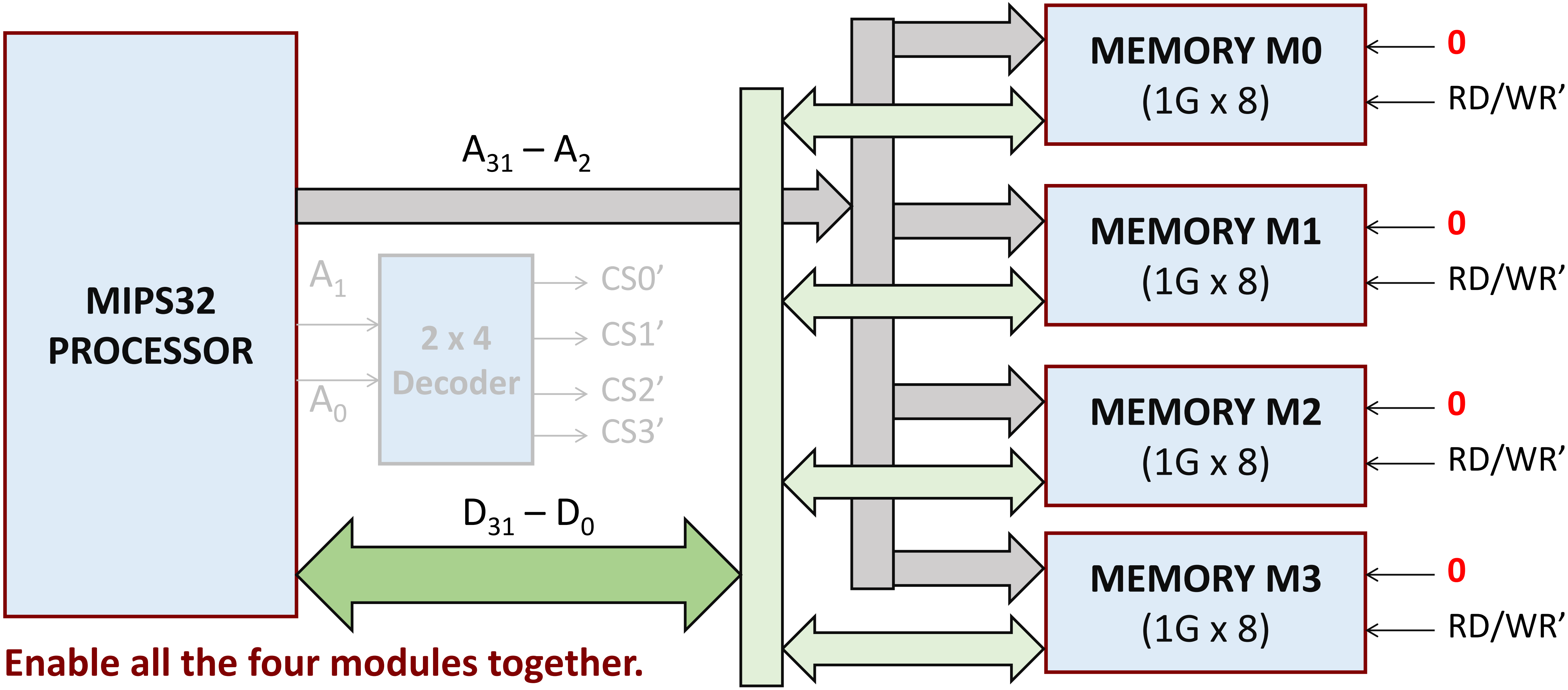
- Motivation for word alignment in MIPS32 data words.
  - 32-bit words start from a memory address that is divisible by 4.
    - Corresponding byte addresses are (0, 1, 2, 3), (4, 5, 6, 7), (8, 9, 10, 11), (12, 13, 14, 15), etc.
    - Possible to transfer all the four bytes in a *single memory cycle*.
  - What happens if a word is not aligned?
    - Say: (1, 2, 3, 4) or (2, 3, 4, 5) or (3, 4, 5, 6).
    - Two of the bytes will be mapped to the same memory mod
    - Hence the word cannot be transferred in a single memory cycle.

2 memory cycles required



Still one module is selected at a time

:: 8 bits data transfer per cycle.



Enable all the four modules together.
32-bit parallel data transfer.

Memory Latency and Bandwidth

- **Memory Latency:**
  - The delay from the issue of a memory read request to the first byte of data becoming available.
- **Memory Bandwidth:**
  - The maximum number of bytes that can be transferred between the processor and the memory system per unit time.

- **Example 1:**

  Consider a memory system that takes 20 ns to service the access of a single 32-bit word.

  - Latency L = 20 ns per 32-bit word.
  - Bandwidth BW = 32 / (20 x 10^{-9}) = 200 Mbytes per second.

- **Example 2:**
  - The memory system is modified to accept a new (still 20ns) request for a 32-bit word every 5 ns by overlapping requests.
    - Latency L = 20 ns per 32-bit word (*no change*).
    - Bandwidth BW = 32 / (5 x 10^{-9}) = 800 Mbits per second.