MEGHE GROUP
VALUES REDEFINED

DTEL (Department for Technology Enhanced Learning)

The Centre for Technology enabled Teaching & Learning , N Y S S, India

Teaching Innovation - Entrepreneurial - Global

MEGHE GROUP
VALUES REDEFINED

DEPARTMENT OF COMPUTER TECHNOLOGY

IV-SEMESTER

COMPUTER ARCHITECTURE AND ORGANIZATION(CT 207)

Unit 5

Semiconductor Memory

UNIT 5:- SYLLABUS

1. Integer Division,
2. Floating point numbers and operations
3. Types of Memory
4. Basic Concepts Related to Semiconductor Memory
5. Internal Organization of Memory
6. Connection of Memory with Processor
7. RAM and ROM

UNIT 5:- SYLLABUS

8 Cache Memory

9 Locality of Reference

10 Performance considerations of Memory

UNIT 5 SPECIFIC OBJECTIVE / COURSE OUTCOME

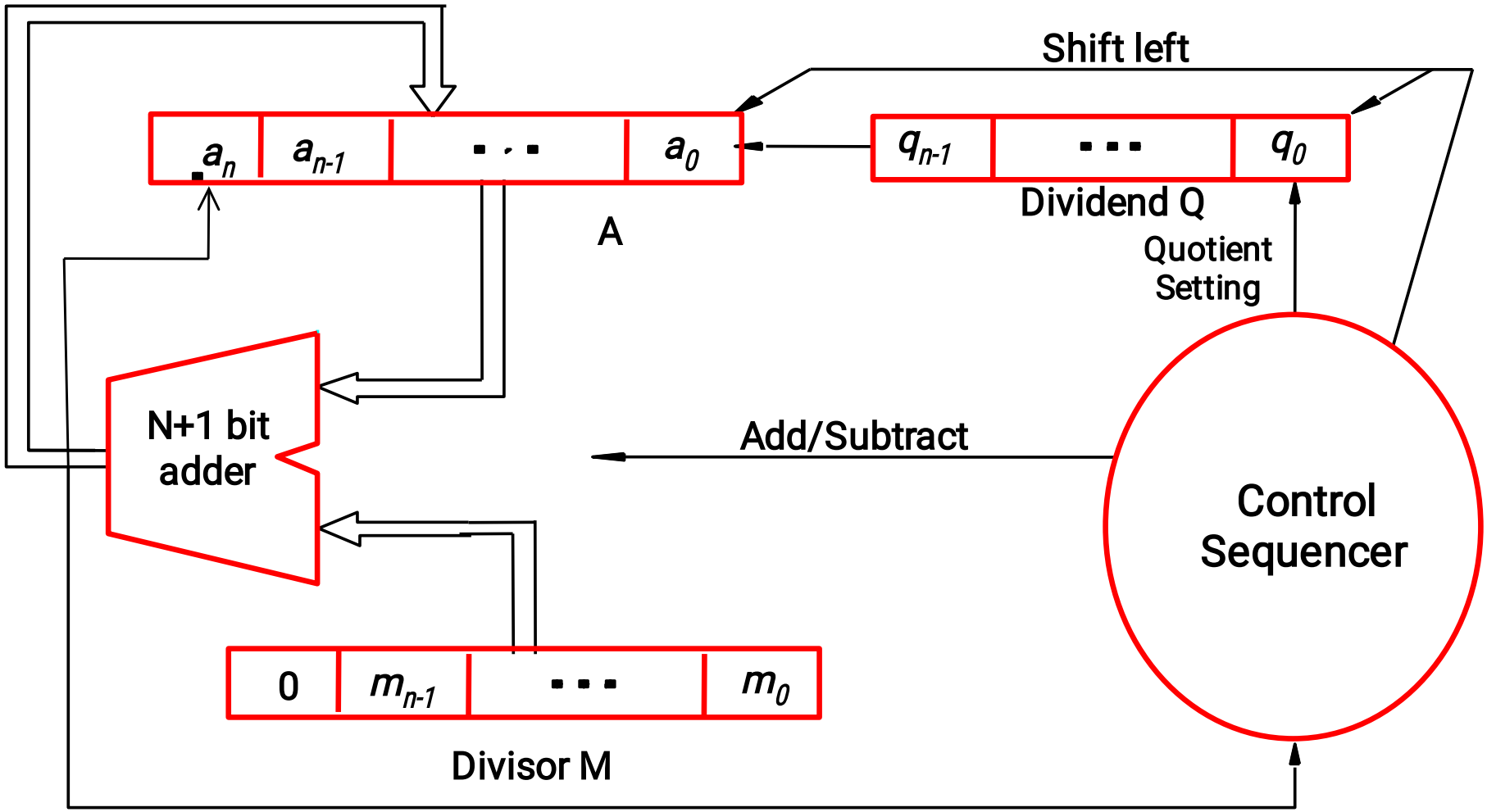
The student will be able to:

1. **Design** of Arithmetic unit to perform fixed point and floating point arithmetic operations
2. Compare various memory types i.e. RAM, ROM, cache, etc.
3. Understand design of Internal Organization of Memory

```
      21
  ____
13 ) 274
     26
    ---
     14
     13
    ---
      1
```

```
          10101
      ----------
1101 )100010010
       1101
      -------
        10000
         1101
        ------
         1110
         1101
        ------
            1
```

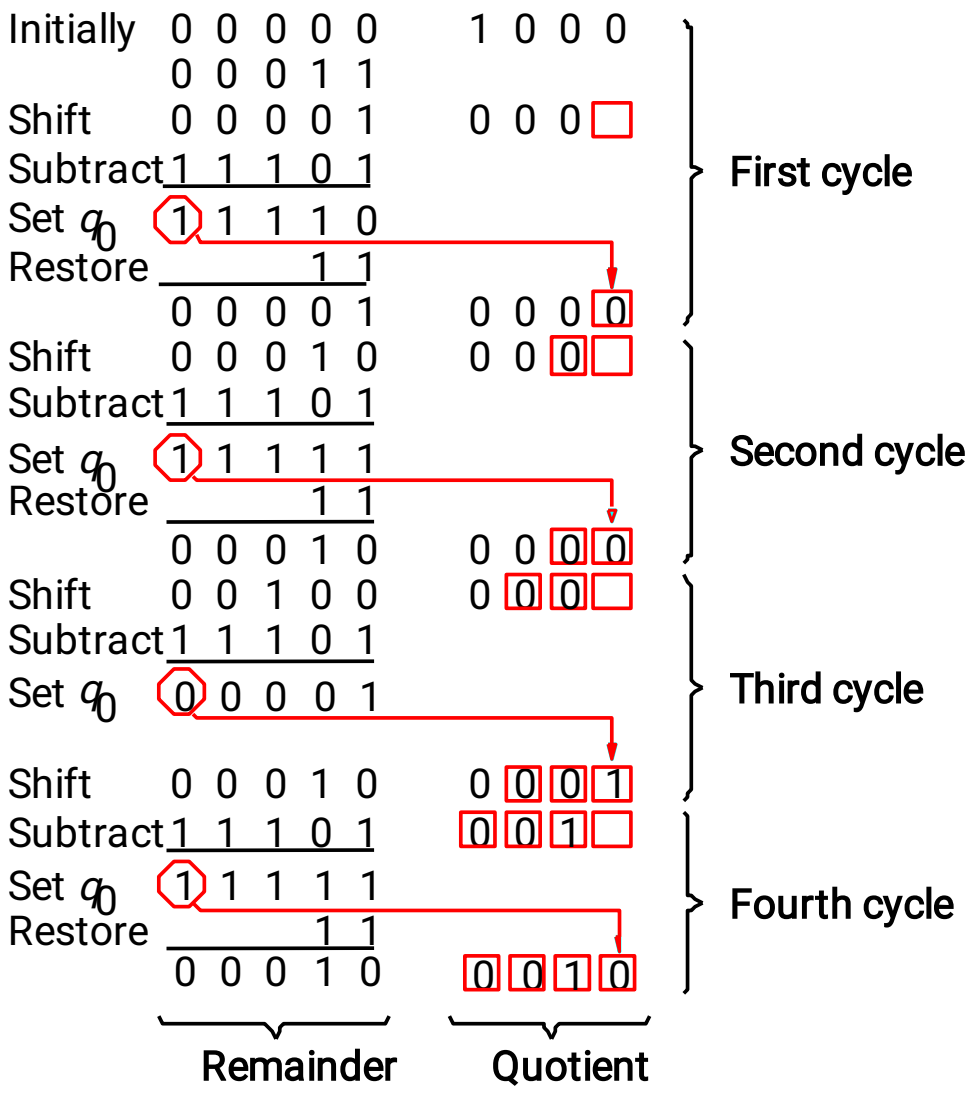
Longhand division examples.



Circuit arrangement for binary division

- Shift A and Q left one binary position
- Subtract M from A, and place the answer back in A
- If the sign of A is 1, set q_0 to 0 and add M back to A (restore A); otherwise, set q_0 to 1
- Repeat these steps *n* times

```
          1 0
      ------------
1 1 ) 1 0 0 0
        1 1
        ---
          1 0
```



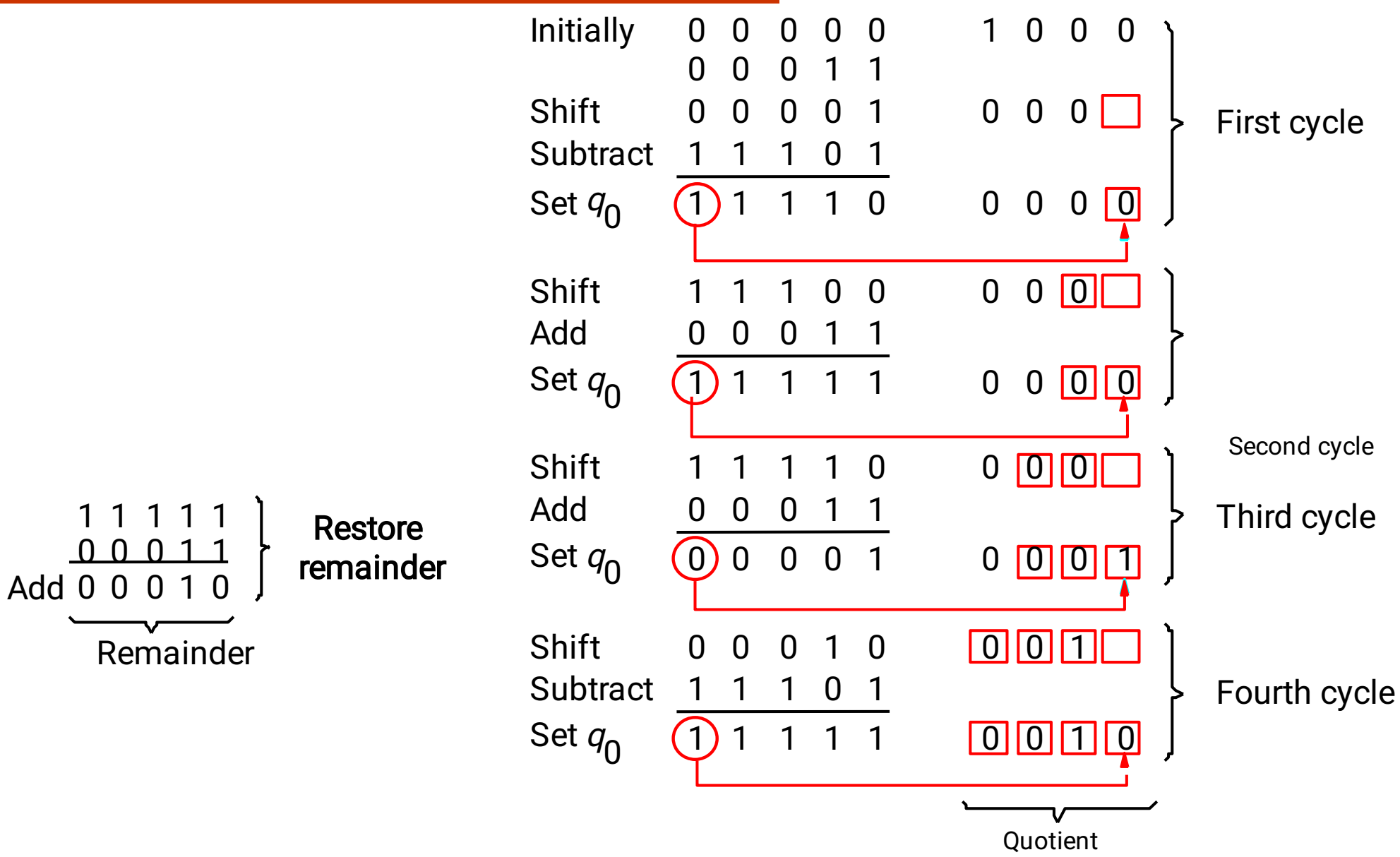
THANK YOU

- Avoid the need for restoring A after an unsuccessful subtraction.
- Any idea?
- Step 1: (Repeat *n* times)

➤ If the sign of A is 0, shift A and Q left one bit position and subtract M from A; otherwise, shift A and Q left and add M to A.

➤ Now, if the sign of A is 0, set q_0 to 1; otherwise, set q_0 to 0.

- Step2: If the sign of A is 1, add M to A



A nonrestoring-division example.

THANK YOU

IEEE Floating Point notation is the standard representation in use. There are two representations:

- Single precision.
- Double precision.

Both have an implied base of 2.

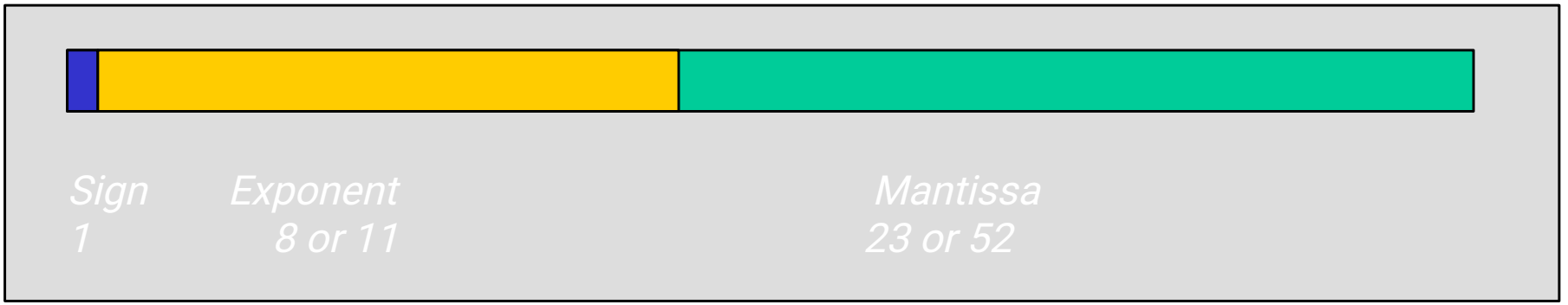
Single precision:

- 32 bits (23-bit mantissa, 8-bit exponent in excess-127 representation)

Double precision:

- 64 bits (52-bit mantissa, 11-bit exponent in excess-1023 representation)

Fractional mantissa, with an implied binary point at immediate left.


Sign
1
Exponent
8 or 11
Mantissa
23 or 52

- Floating point numbers have to be represented in a normalized form to maximize the use of available mantissa digits.
- If every number is normalized, then the MSB of the mantissa is always 1.
- The mantissa is a 1 and does not store this bit.
- So the real MSB of a number in the IEEE notation is either a 0 or a 1.
- The values of the numbers represented in the IEEE single precision notation are of the form:

$$(+,-)\ 1.M \times 2^{(E-127)}$$

-126 <= E <= 127 and -1022 <= E <= 1023
not
-127 <= E <= 128 and -1023 <= E <= 1024

- Choose the number with the smaller exponent.
- Shift its mantissa right until the exponents of both the numbers are equal.
- Add or subtract the mantissas.
- Determine the sign of the result.
- Normalize the result if necessary and truncate/ round to the number of mantissa bits.

- Add the exponents.
- Subtract the bias.
- Multiply the mantissas and determine the sign of the result.
- Normalize the result (if necessary).
- Truncate/round the mantissa of the result.

- Subtract the exponents
- Add the bias.
- Divide the mantissas and determine the sign of the result.
- Normalize the result if necessary.
- Truncate/round the mantissa of the result.

• While adding two floating point numbers with 24-bit mantissas, we shift the mantissa of the number with the smaller exponent to the right until the two exponents are equalized.

•This implies that mantissa bits may be lost during the right shift.

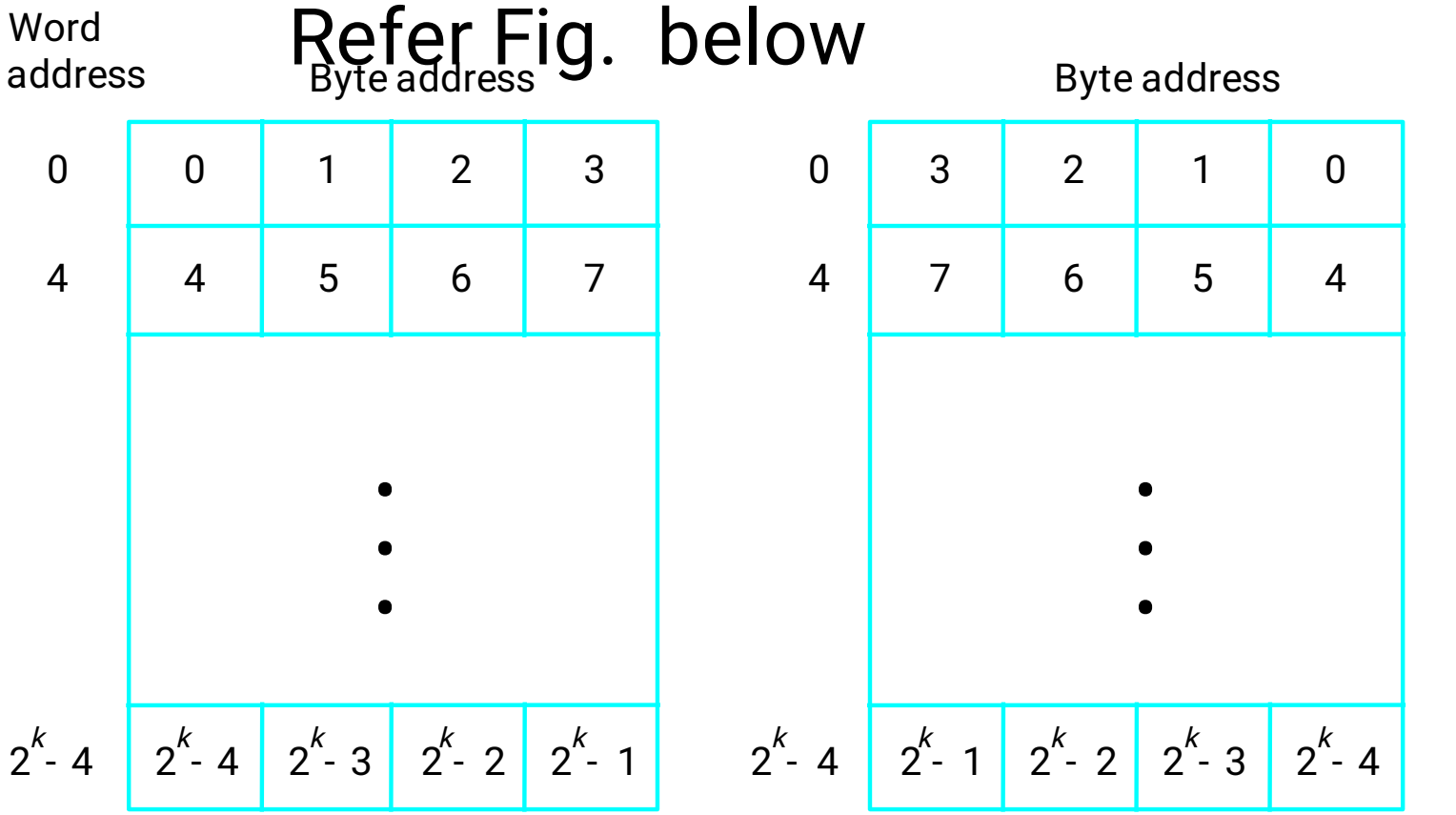
•To prevent this, floating point operations are implemented by keeping guard bits, that is, extra bits of precision at the least significant end of the mantissa.

THANK YOU

Basic Concepts

- The maximum size of the memory that can be used in any computer is determined by the addressing scheme.

  16-bit addresses = 2^{16} = 64K memory locations
- Most modern computers are byte addressable.
- Refer Fig. below



Basic Concepts

Traditional Architecture

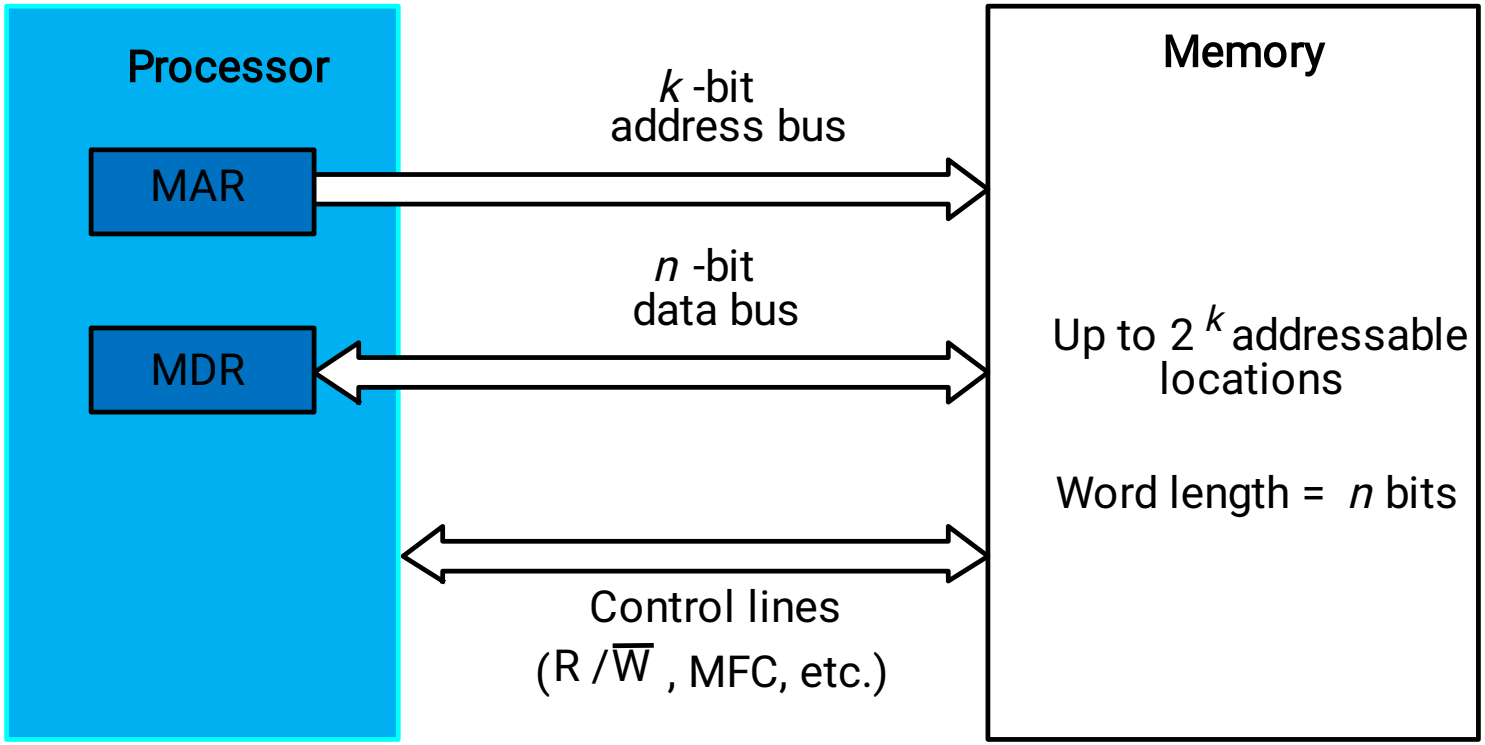
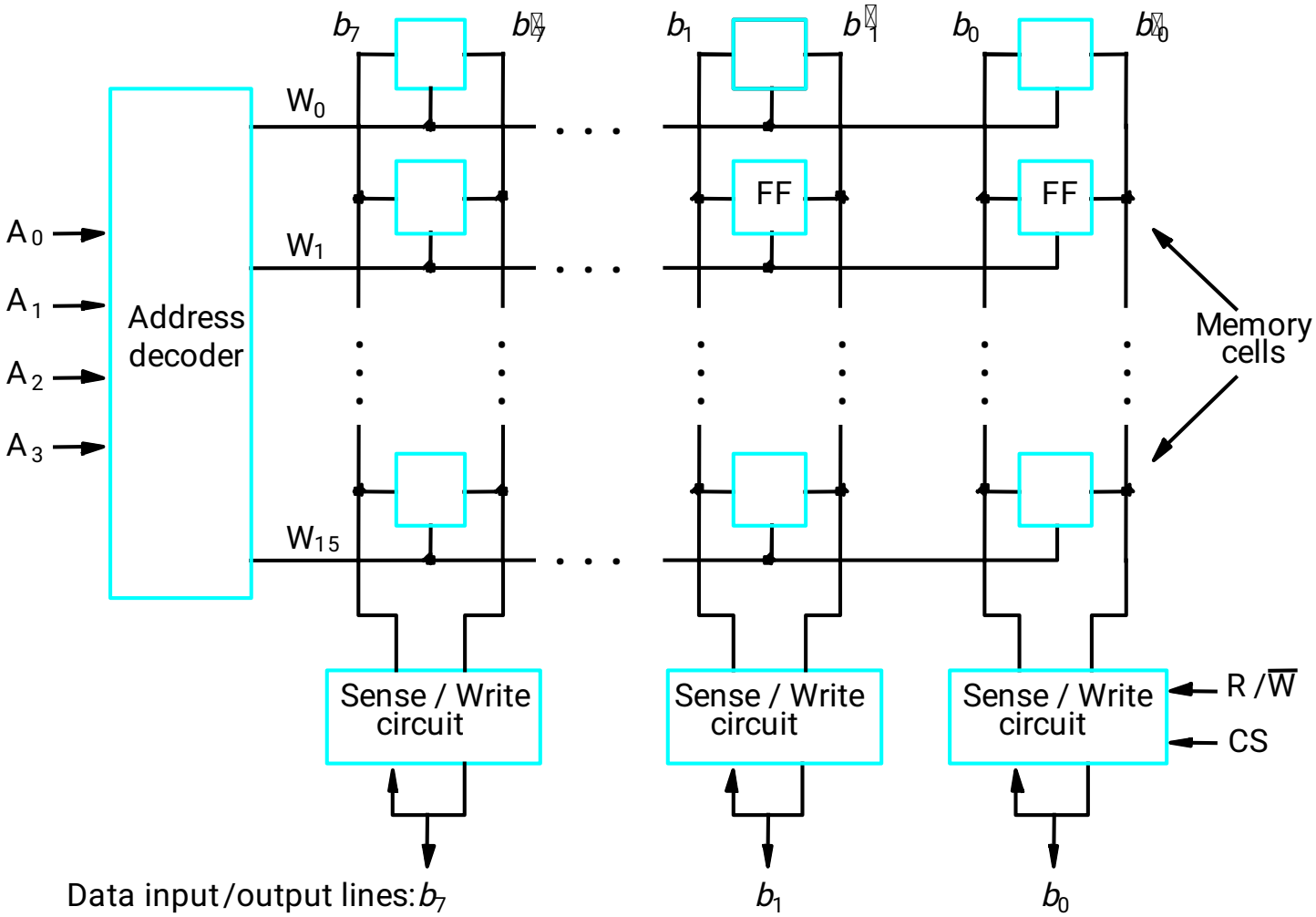


Figure 5.1. Connection of the memory to the processor.

Basic Concepts

- “Block transfer” – bulk data transfer
- *Memory access time*
- *Memory cycle time*
- RAM – any location can be accessed for a Read or Write operation in some fixed amount of time that is independent of the location’s address.
- Cache memory
- Virtual memory, memory management unit

Semiconductor RAM Memories



Internal Organization of Memory Chips (fig.5.2)

Figure 5.2. Organization of bit cells in a memory chip.

Internal Organization of Memory Chips (fig.5.2)

16 words of 8 bits each: 16x8 memory org.. It has 16 external connections: addr. 4, data 8, control: 2, power/ground: 2

1K memory cells: 128x8 memory, external connections: ? 19(7+8+2+2)

1Kx1:? 15 (10+1+2+2)

Figure 5.2. Organization of bit cells in a memory chip.

A Memory Chip

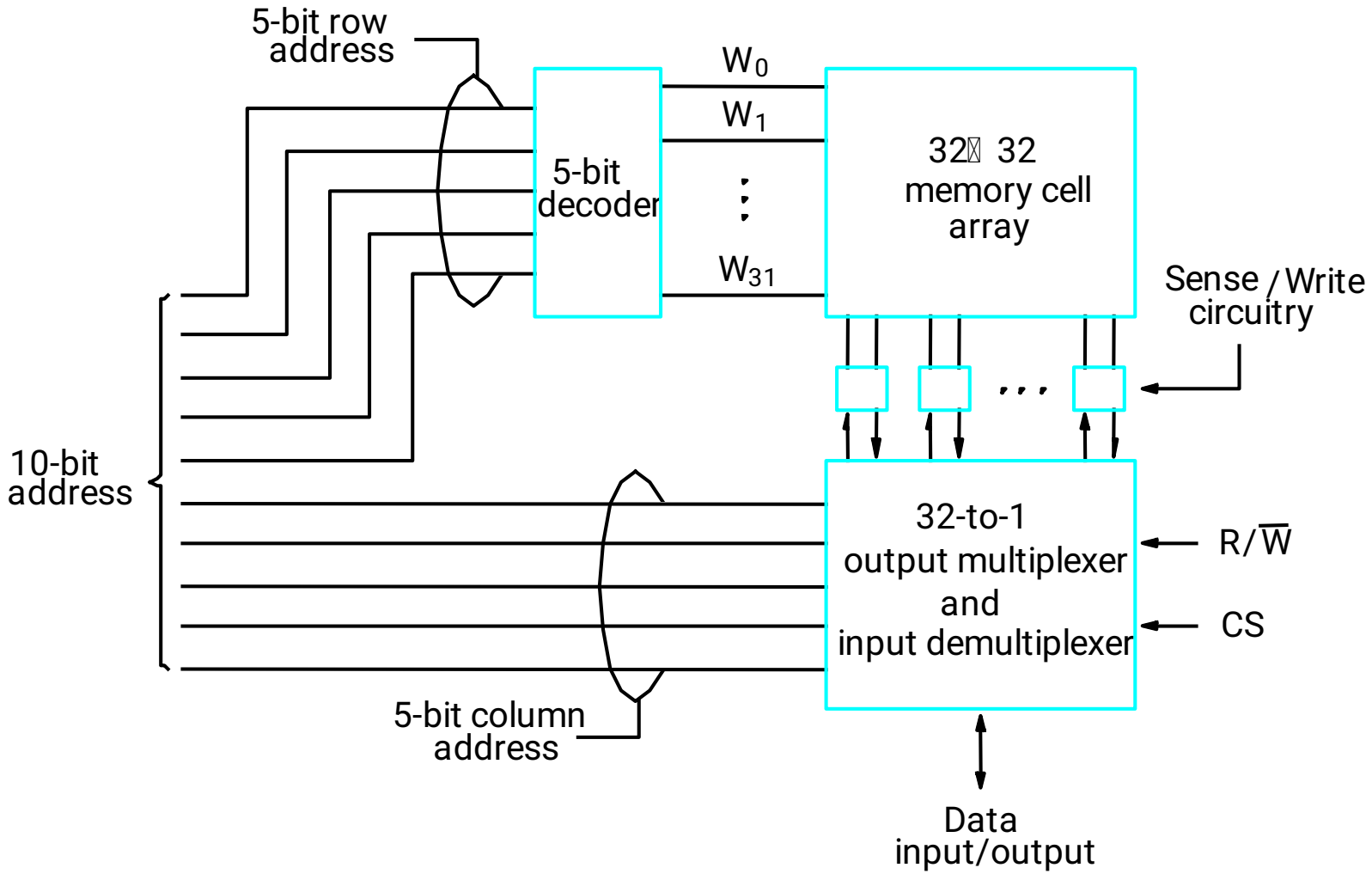


Figure 5.3. Organization of a 1K × 1 memory chip.

Static Memories

- The circuits are capable of retaining their state as long as power is applied.

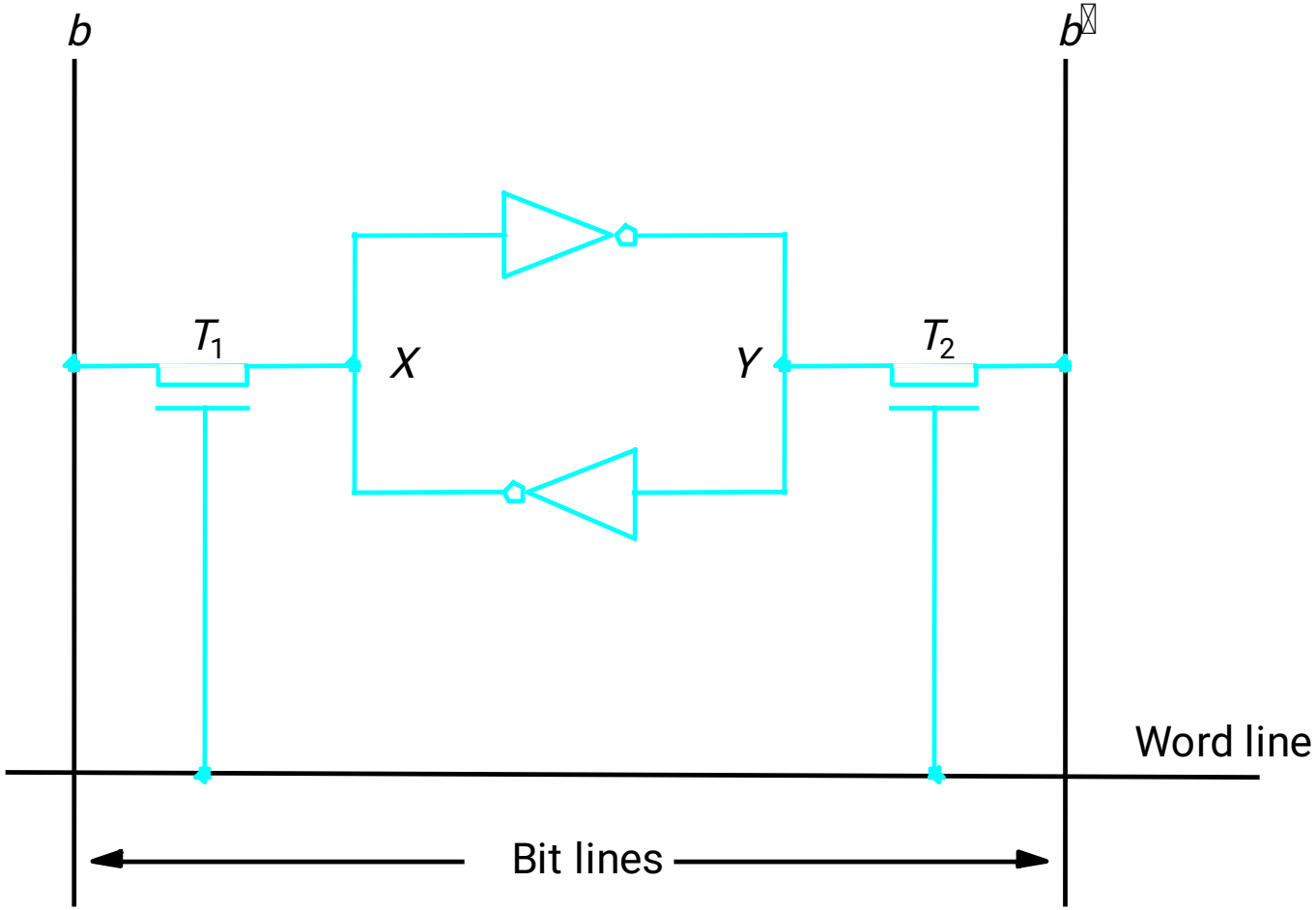


Figure 5.4. A Static RAM cell.

Static Memories

- CMOS cell: low power consumption

Semiconductor RAM Memories

Asynchronous DRAMs

- Static RAMs are fast, but they cost more area and are more expensive.
- Dynamic RAMs (DRAMs) are cheap and area efficient, but they can not retain their state indefinitely – need to be periodically refreshed.

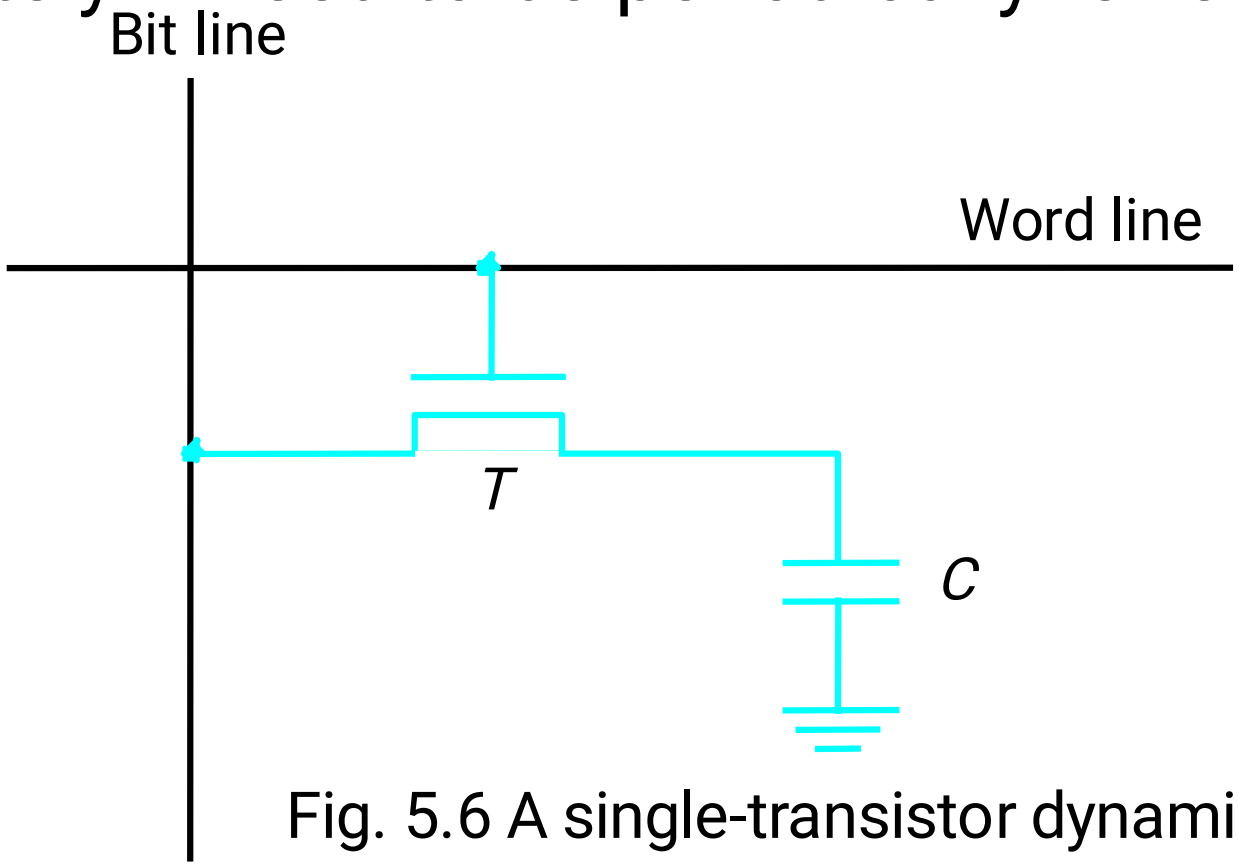


Fig. 5.6 A single-transistor dynamic memory cell

A Dynamic Memory Chip

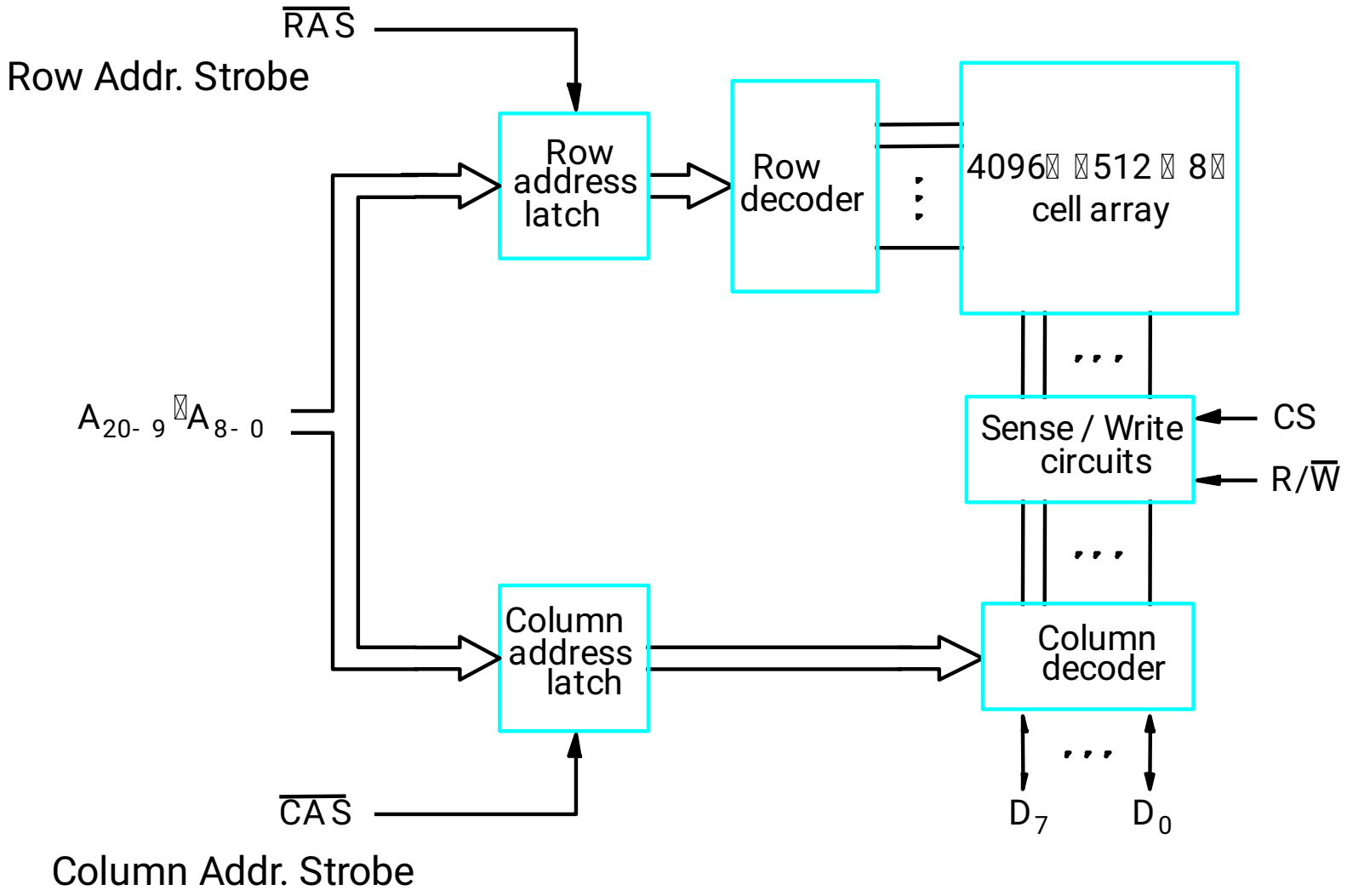


Figure 5.7. Internal organization of a 2M × 8 dynamic memory chip.

Fast Page Mode

- When the DRAM in last slide is accessed, the contents of all 4096 cells in the selected row are sensed, but only 8 bits are placed on the data lines $D_{7\text{-}0}$, as selected by $A_{8\text{-}0}$.
- Fast page mode – make it possible to access the other bytes in the same row without having to reselect the row.
- A latch is added at the output of the sense amplifier in each column.
- Good for bulk transfer.

Synchronous DRAMs

- The operations of SDRAM are controlled by a clock signal.

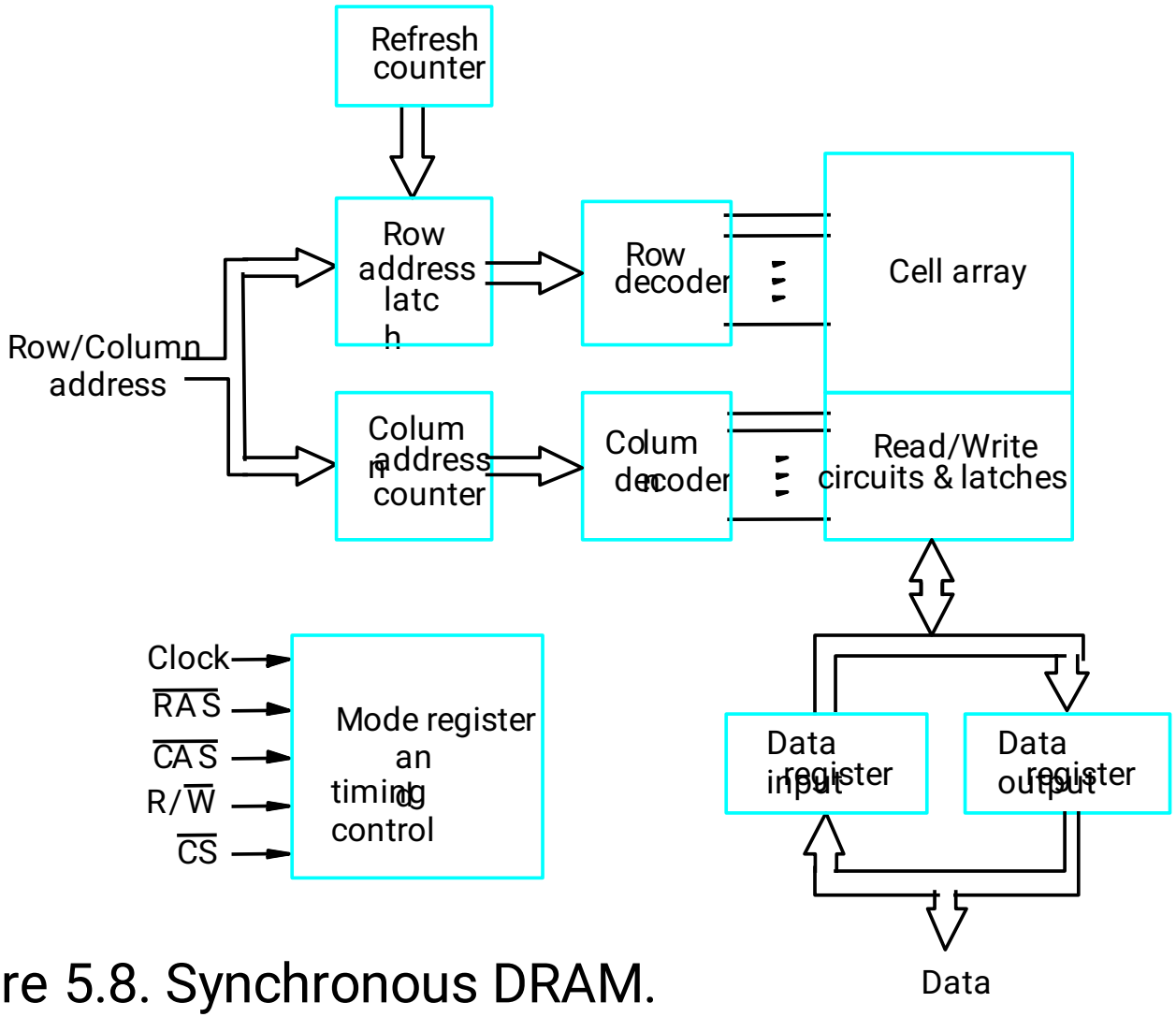


Figure 5.8. Synchronous DRAM.

Synchronous DRAMs

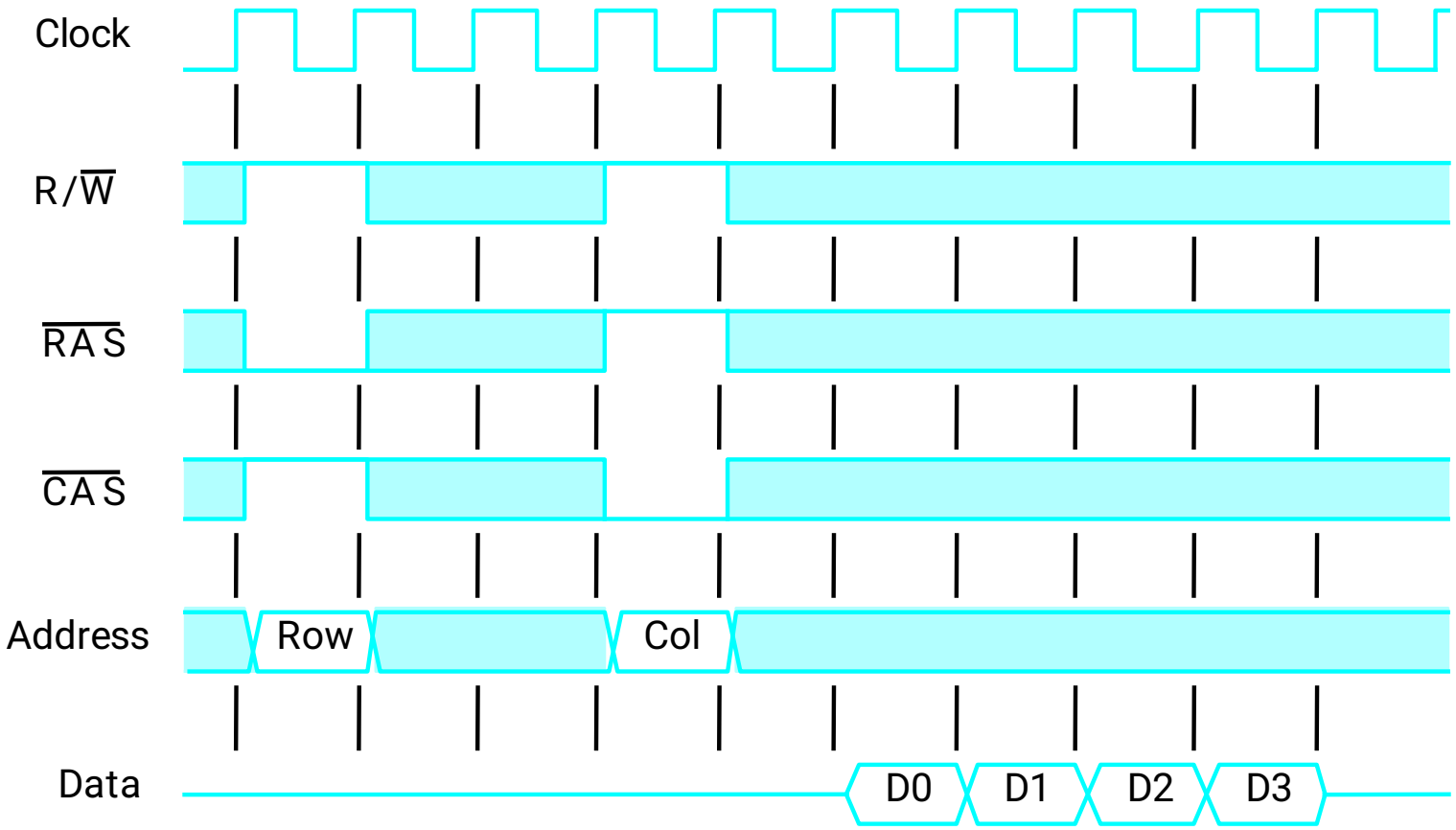


Figure 5.9. Burst read of length 4 in an SDRAM.

Synchronous DRAMs

- No CAS pulses is needed in burst operation.
- Refresh circuits are included (every 64ms).
- Clock frequency > 100 MHz
- Intel PC100 and PC133

Latency and Bandwidth

- The speed and efficiency of data transfers among memory, processor, and disk have a large impact on the performance of a computer system.
- Memory latency – the amount of time it takes to transfer a word of data to or from the memory.
- Memory bandwidth – the number of bits or bytes that can be transferred in one second. It is used to measure how much time is needed to transfer an entire block of data.
- Bandwidth is not determined solely by memory. It is the product of the rate at which data are transferred (and accessed) and the width of the data bus.

DDR SDRAM

- Double-Data-Rate SDRAM
- Standard SDRAM performs all actions on the rising edge of the clock signal.
- DDR SDRAM accesses the cell array in the same way, but transfers the data on both edges of the clock.
- The cell array is organized in two banks. Each can be accessed separately.
- DDR SDRAMs and standard SDRAMs are most efficiently used in applications where block transfers are prevalent.

Structures of Larger Memories

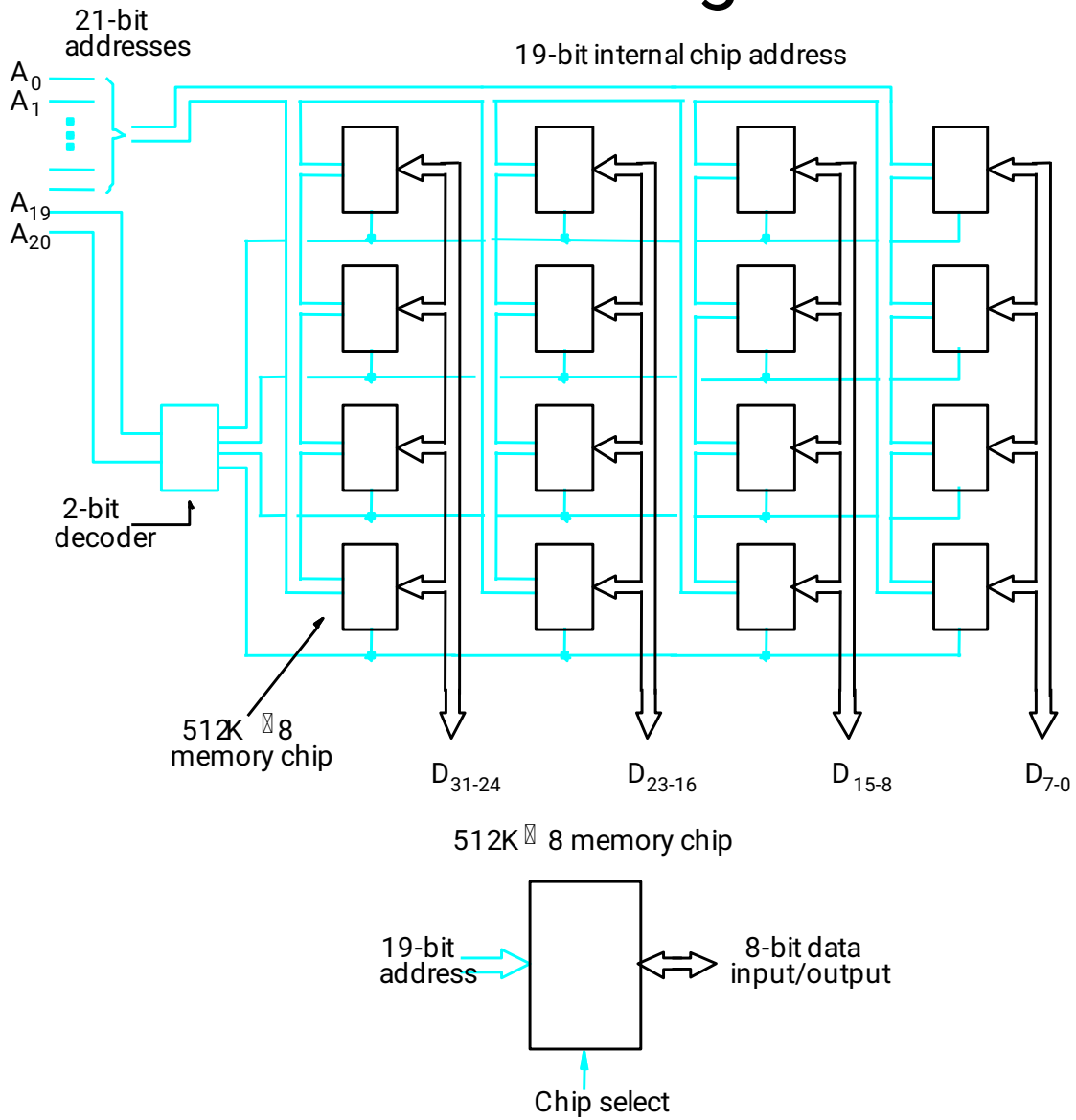


Figure 5.10. Organization of a 2M × 32 memory module using 512K × 8 static memory chips.

Semiconductor RAM Memories

Memory System Considerations

- The choice of a RAM chip for a given application depends on several factors:

  Cost, speed, power, size…

- SRAMs are faster, more expensive, smaller.

- DRAMs are slower, cheaper, larger.

Memory System Considerations

- Which one for cache and main memory, respectively?
- Refresh overhead – suppose a SDRAM whose cells are in 8K rows; 4 clock cycles are needed to access each row; then it takes 8192×4=32,768 cycles to refresh all rows; if the clock rate is 133 MHz, then it takes $32{,}768/(133\times10^{-6})=246\times10^{-6}$ seconds; suppose the typical refreshing period is 64 ms, then the refresh overhead is 0.246/64=0.0038<0.4% of the total time available for accessing the memory.

Memory Controller

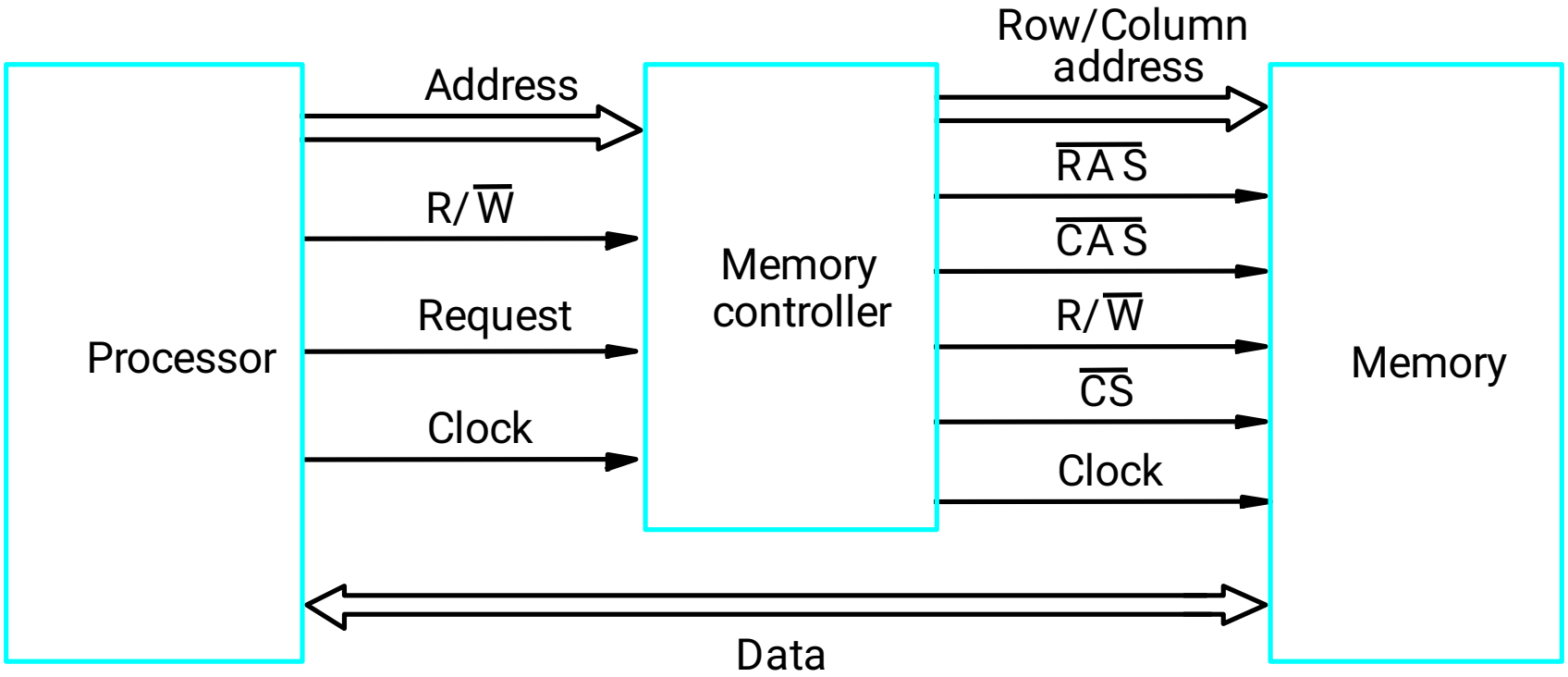


Figure 5.11. Use of a memory controller.

THANK YOU

Read-Only Memories

- Volatile / non-volatile memory
- ROM
- PROM: programmable ROM
- EPROM: erasable, reprogrammable ROM
- EEPROM: can be programmed and erased electrically

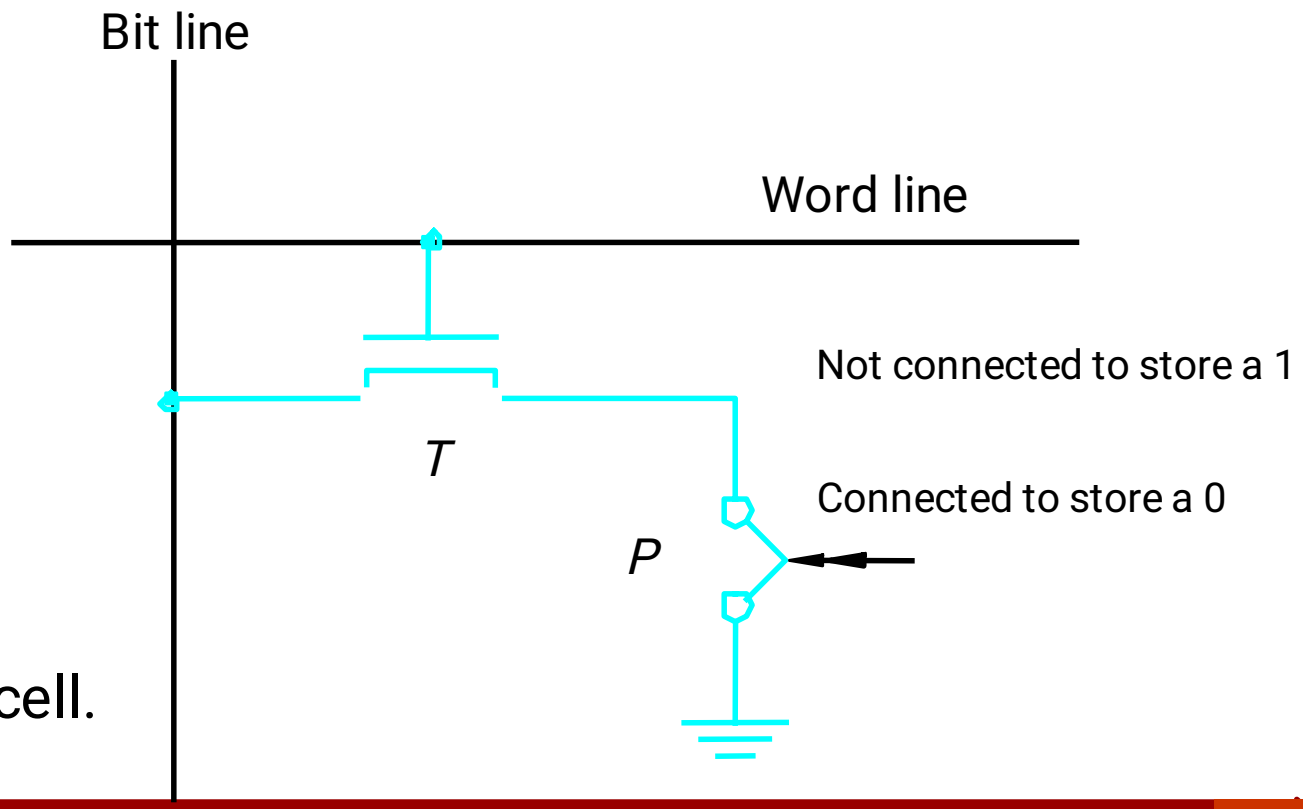


Figure 5.12. A ROM cell.

Flash Memory

- Similar to EEPROM
- Difference: only possible to write an entire block of cells instead of a single cell
- Low power
- Use in portable equipment
- Implementation of such modules
  - Flash cards
  - Flash drives

Read-Only Memories

Speed, Size, and Cost

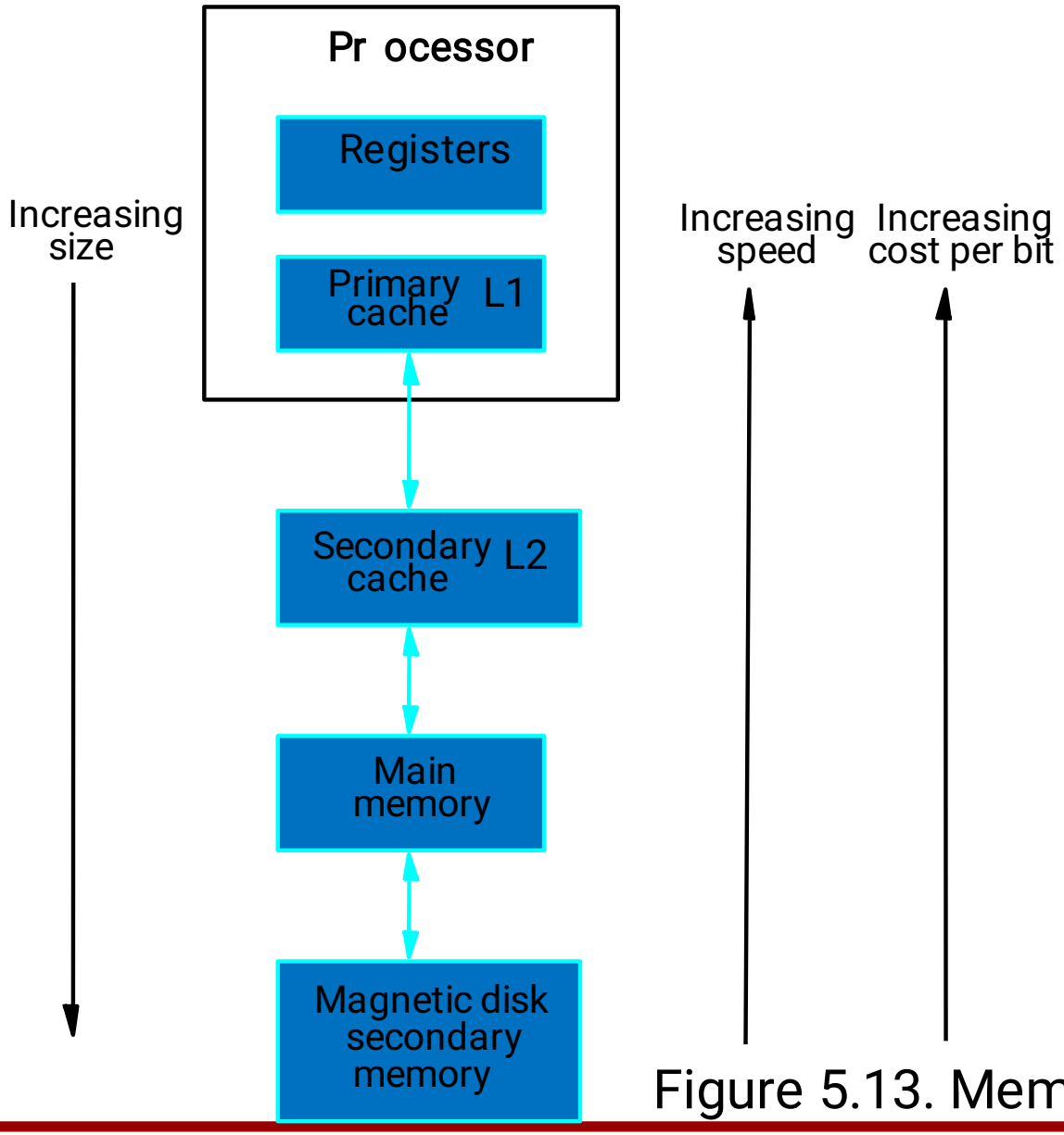


Figure 5.13. Memory hierarchy.

Cache

- What is cache?
- Why we need it?
- Locality of reference (very important)

  - temporal

  - spatial
- Cache block – *cache line*
  - *A set of contiguous address locations of some size*

Cache Memories

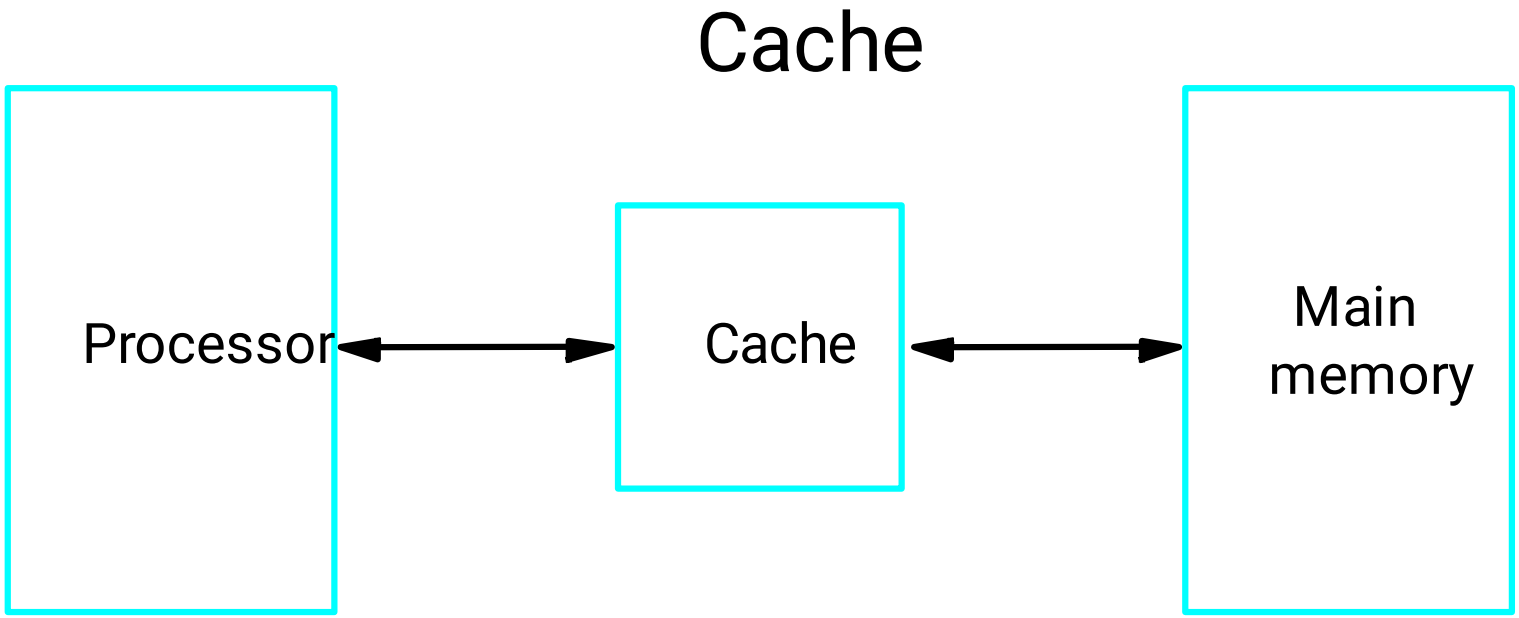


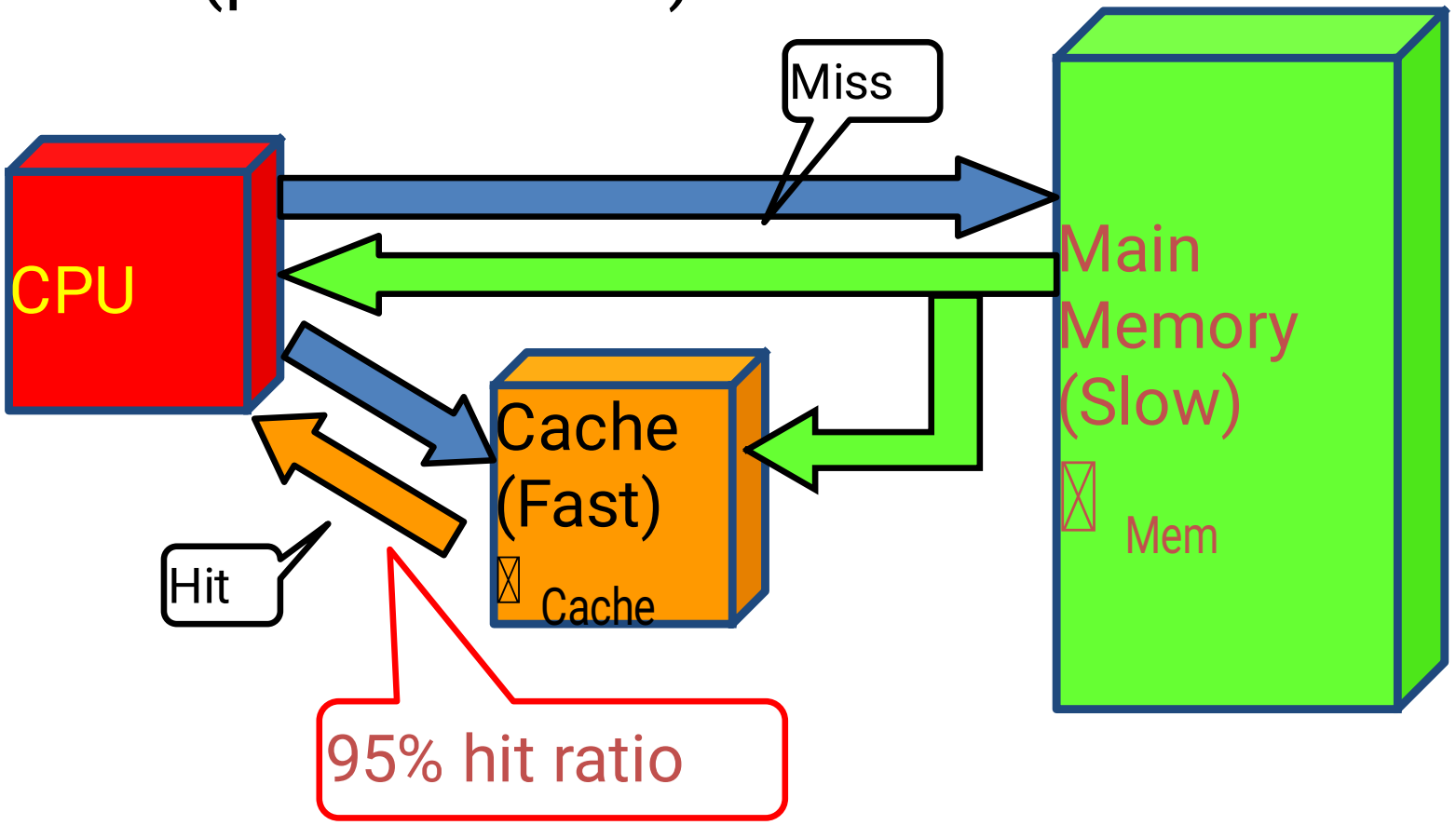
Figure 5.14. Use of a cache memory.

- Replacement algorithm
- Hit / miss
- Write-through / Write-back
- Load through

Cache Memories

Cache Memory (Ref. Internet Fig a)

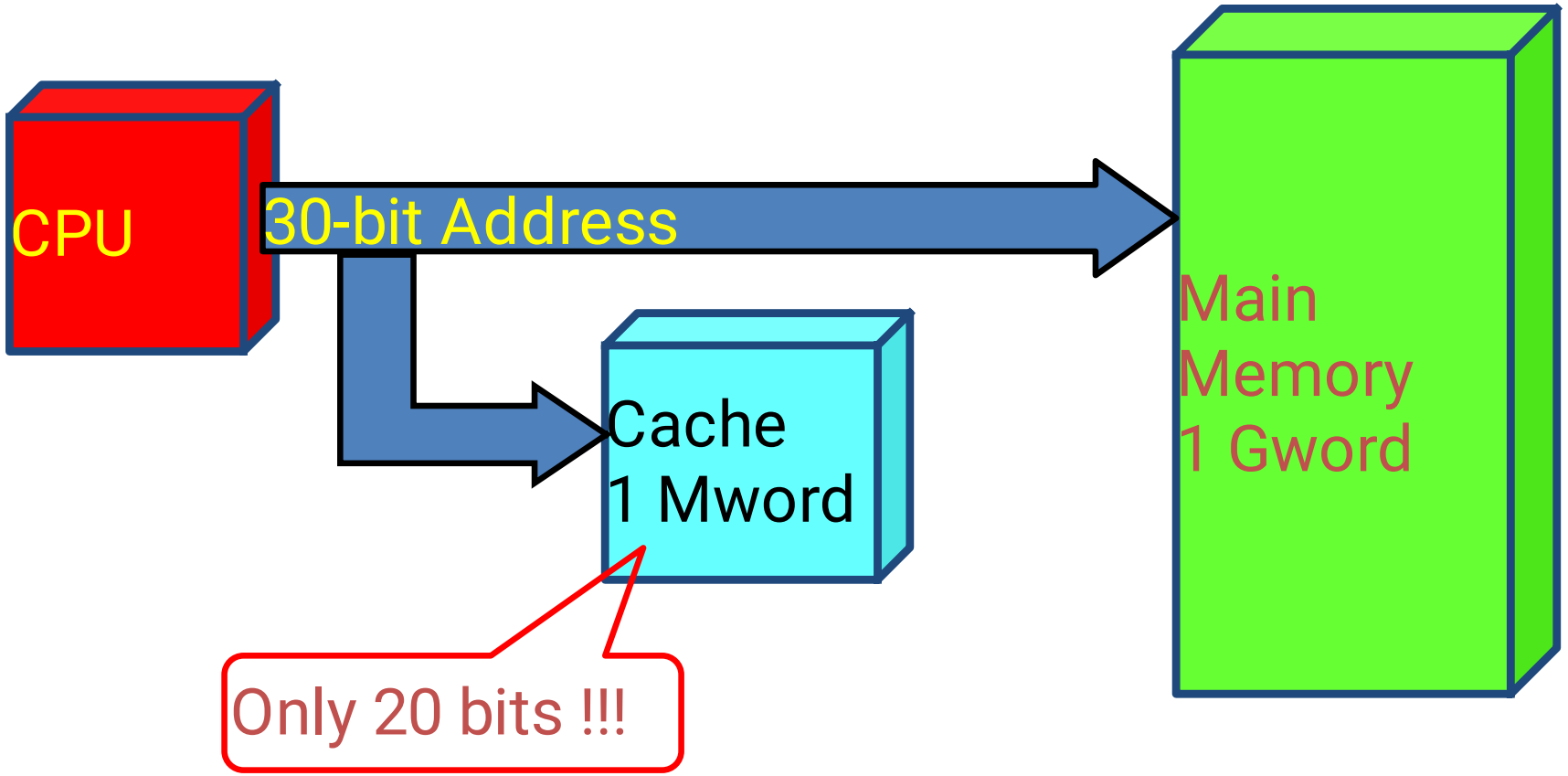
- High speed (towards CPU speed)
- Small size (power & cost)a



$T_{Access} = 0.95\ T_{Cache} + 0.05\ T_{Mem}$

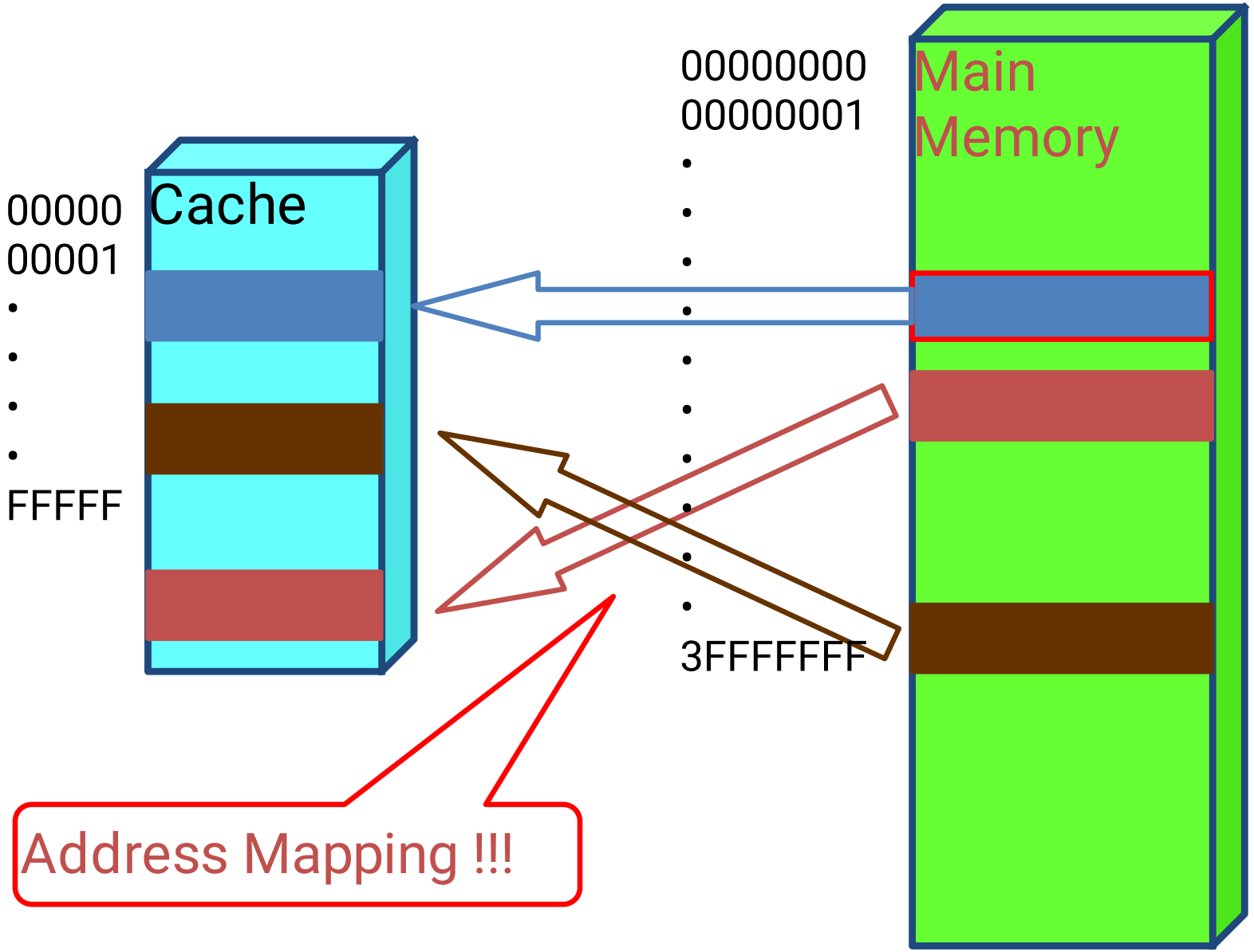
Cache Memories

Cache Memory (Ref. Internet Fig b)



Cache Memories

Cache Memory (Ref. Internet Fig c)



THANK YOU

Performance Considerations

Overview

- Two key factors: performance and cost
- Price/performance ratio
- Performance depends on how fast machine instructions can be brought into the processor for execution and how fast they can be executed.
- For memory hierarchy, it is beneficial if transfers to and from the faster units can be done at a rate equal to that of the faster unit.
- This is not possible if both the slow and the fast units are accessed in the same manner.
- However, it can be achieved when parallelism is used in the organizations of the slower unit.

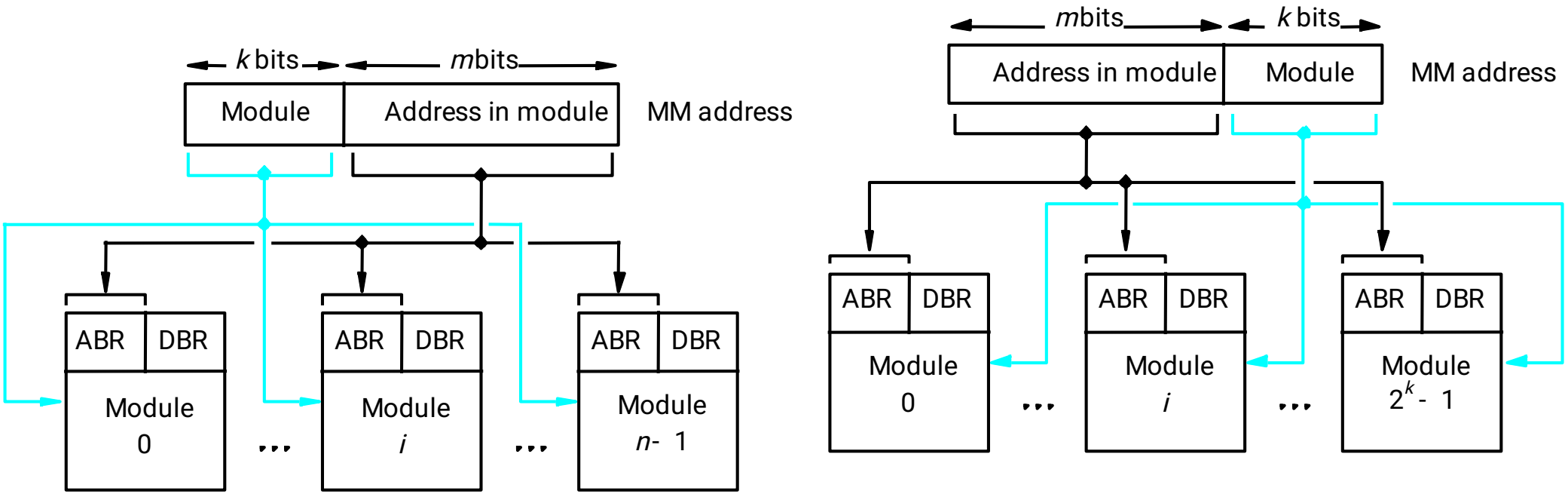
Interleaving

- If the main memory is structured as a collection of physically separated modules, each with its own ABR (Address buffer register) and DBR(Data buffer register), memory access operations may proceed in more than one module at the same time.

Figure 5.25. Addressing multiple-module memory systems.

Performance Considerations

Interleaving



(a) Consecutive words in a module

(b) Consecutive words in consecutive modules

Figure 5.25. Addressing multiple-module memory systems.

Hit Rate and Miss Penalty

- The success rate in accessing information at various levels of the memory hierarchy – hit rate / miss rate.
- Ideally, the entire memory hierarchy would appear to the processor as a single memory unit that has the access time of a cache on the processor chip and the size of a magnetic disk – depends on the hit rate (>0.9).
- A miss causes extra time needed to bring the desired information into the cache.
- Example 5.2, page 332.

Hit Rate and Miss Penalty

- $T_{ave}=hC+(1-h)M$
  - $T_{ave:}$ average access time experienced by the processor
  - h: hit rate
  - M: miss penalty, the time to access information in the main memory
  - C: the time to access information in the cache

Hit Rate and Miss Penalty

- Example:
  - Assume that 30 percent of the instructions in a typical program perform a read/write operation, which means that there are 130 memory accesses for every 100 instructions executed.
  - h=0.95 for instructions, h=0.9 for data
  - C=10 clock cycles, M=17 clock cycles, interleaved memory

  Time without cache 130x10

  Time with cache
  100(0.95x1+0.05x17)+30(0.9x1+0.1x17)
  - The computer with the cache performs five times better

How to Improve Hit Rate?

- Use larger cache – increased cost
- Increase the block size while keeping the total cache size constant.
- However, if the block size is too large, some items may not be referenced before the block is replaced – miss penalty increases.
- Load-through approach

Performance Considerations

Caches on the Processor Chip

- On chip vs. off chip
- Two separate caches for instructions and data, respectively
- Single cache for both
- Which one has better hit rate? -- Single cache
- What's the advantage of separating caches? – parallelism, better performance
- Level 1 and Level 2 caches

Performance Considerations

Caches on the Processor Chip

- L1 cache – faster and smaller. Access more than one word simultaneously and let the processor use them one at a time.
- L2 cache – slower and larger.
- How about the average access time?
- Average access time:

$$t_{ave} = h_1C_1 + (1-h_1)h_2C_2 + (1-h_1)(1-h_2)M$$

where h is the hit rate, C is the time to access information in cache, M is the time to access information in main memory.

Other Enhancements

- Write buffer – processor doesn't need to wait for the memory write to be completed
- Prefetching – prefetch the data into the cache before they are needed
- Lockup-Free cache – processor is able to access the cache while a miss is being serviced.

THANK YOU

References

1. Computer Organization, 5th Chapter, Carl Hamacher, TMH publication 5th Edition.
2. Online ppts by Carl Hamacher www.slideworld.com/**ppts**lides.../computer-organization-**carl-hamacher**