Lecture 02: Memory and I/O Modules

William Stallings Computer Organization and Architecture

Chapter 5

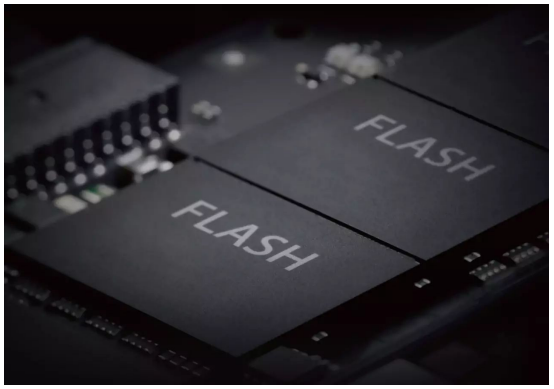
Internal Memory

Key Content

- Memory taxonomy & characteristics
  - Physical type, location, capacity, unit of transfer, access method, performance
- Memory hierarchy in computer system
  - Registers -> Cache -> Main memory -> Disk
- RAM organization (SRAM and DRAM)
  - Cell -> Array -> Chip
- Memory module extension
  - Bit & word extension

Physical Types of Memory

- Semiconductor
  - RAM & ROM
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD



Memory Characteristics Comparison

- Location
- Capacity
- Unit of transfer
- Access method
- Performance
- Physical type
- Organisation

Location

- CPU
  - registers
- Internal
  - Cache, main memory
- External
  - Disk, tape, DVD

Capacity

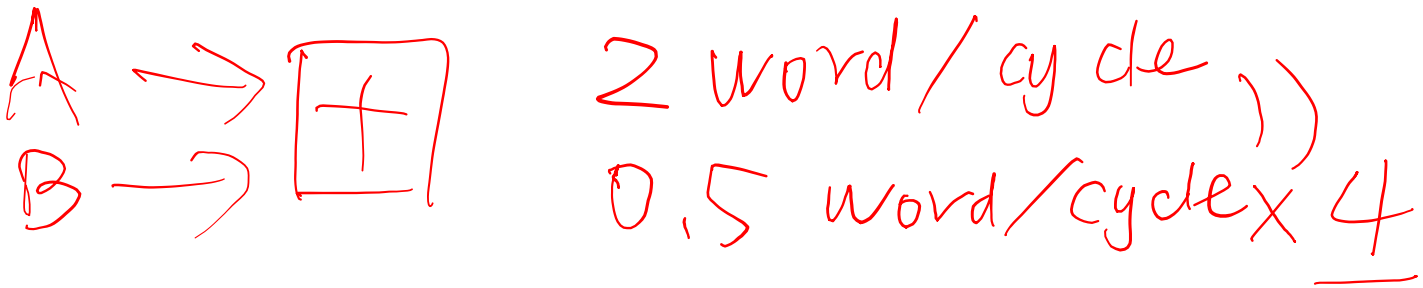
- Word size
  - The natural unit of organization
- Number of words
- E.g., 2M 8-bit, 16M 1-bit
  - Same capacity but different organization

2M x 8B = 2 MB

16M x 1b = 16Mb = 2 MB

Unit of Transfer

- Internal
  - Usually a word, governed by data bus width
- External
  - Usually a block which is much larger than a word
  - E.g., the CPU can calculate one addition every cycle, but a memory transfer takes two cycles. With the memory interface width to be 4 words, the CPU can be kept with 100% utilization.

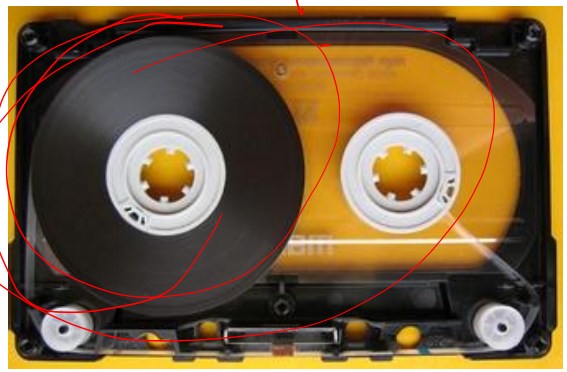


Addressable unit

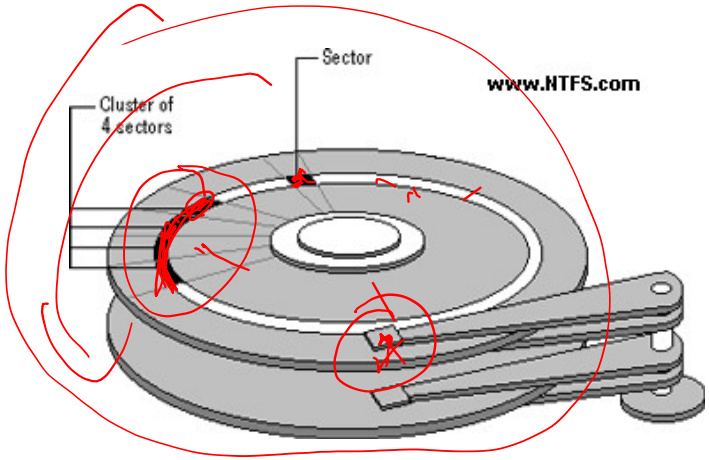
- Smallest location which can be uniquely addressed
- Normally a Byte for internal memory
- Cluster on disks

Access Methods (1)

- Sequential
  - Start at the beginning and read through in order
  - Access time depends on location of data and previous location
  - e.g. tape

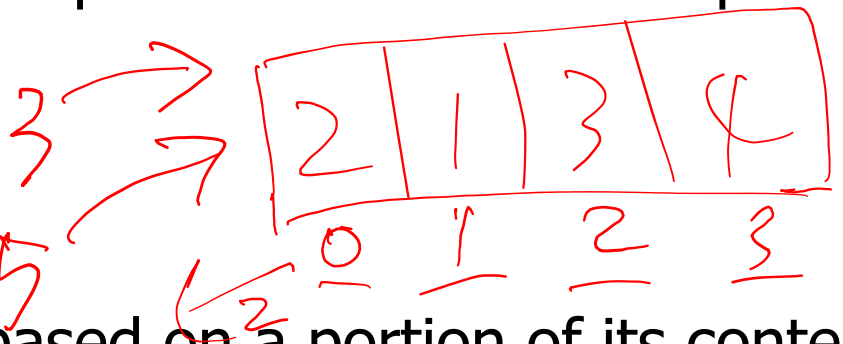


- Direct
  - Individual blocks have unique address
  - Access is by jumping to vicinity plus sequential search
  - Access time depends on location and previous location
  - e.g. disk



Access Methods (2)

- Random
  - Individual addresses identify locations exactly
  - Access time is independent of location or previous access
  - e.g. RAM, ROM
- Associative
  - Data is located based on a portion of its contents rather than its address
  - Access time is independent of location or previous access
  - e.g. cache



Performance

- Access time: time between presenting the address and getting the valid data
- Memory cycle time: time may be required for the memory to "recover" before next access
  - Cycle time is access + recovery
- Transfer rate: rate at which data can be moved
  - Unit: transfer per second (e.g., GT/s)
- Transfer bandwidth
  - Equals Transfer rate * Transfer unit size
  - Unit: byte per second (e.g., GB/s)

Quiz

- Determine the access time, transfer rate, and transfer bandwidth for the following memory
  - A memory transfer takes two cycles, and each transfer has 4 bytes
  - Clock frequency is 1 GHz

A. T. = 2 cycles = 2 ns

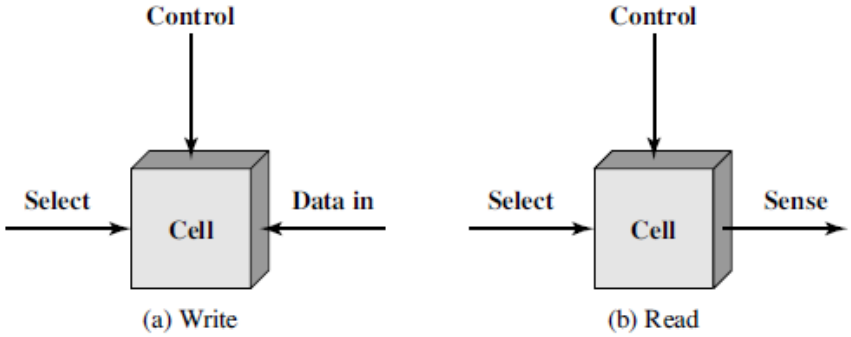
T. R. = 0.5 T/cycle = 0.5 GT/s

T. B. = 0.5 GT/s × 4 B

= 2 GB/s

Semiconductor Memory

- The basic element of a semiconductor memory is the memory cell
  - exhibit two stable states, representing binary 1 and 0
  - capable of being written into to set the state
  - capable of being read from to sense the state



- RAM (Random Access Memory)
  - Misnamed as all semiconductor memory is random access
- ROM (Read-Only Memory)

Memory Basics

- RAM: Random Access Memory
  - historically defined as memory array with individual bit access
  - Refers to memory with both Read and Write capabilities
- ROM: Read Only Memory
  - No capabilities for "online" memory Write operations
  - Write typically requires high voltages or erasing by UVlight

Memory Basics

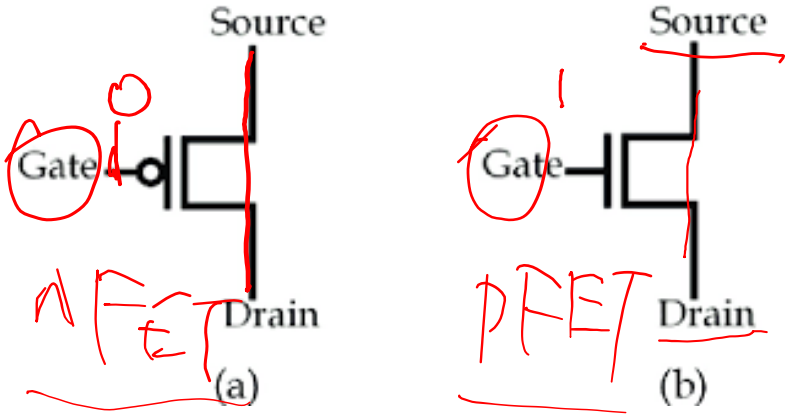
- Volatility of Memory 易失性
  - Volatile memory loses data over time or when power is removed RAM is volatile
  - non-volatile memory stores date even when power is removed
  - ROM is non-volatile
  - Static vs. Dynamic Memory
    - Static: holds data as long as power is applied (SRAM)
    - Dynamic: will lose data unless refreshed periodically (DRAM)

Random Access Memory (RAM)

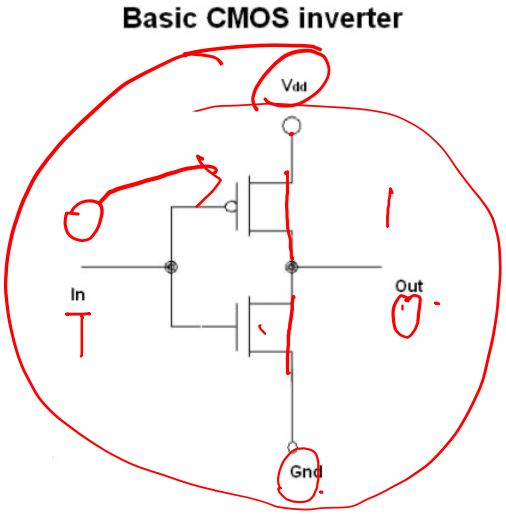
- Read/Write
- Volatile
  - must be provided with a constant power supply
- Temporary storage
  - When the power is gone, data are lost
- Static or dynamic
  - SRAM or DRAM

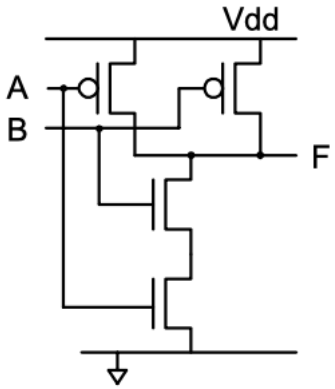
Recap on Digital Circuit

- Field effect transistor (FET)
  - pFET vs nFET



- CMOS gate:
  - Complementary metal–oxide–semiconductor





Quiz: what is this gate for?

Quiz: can you identify those gates?

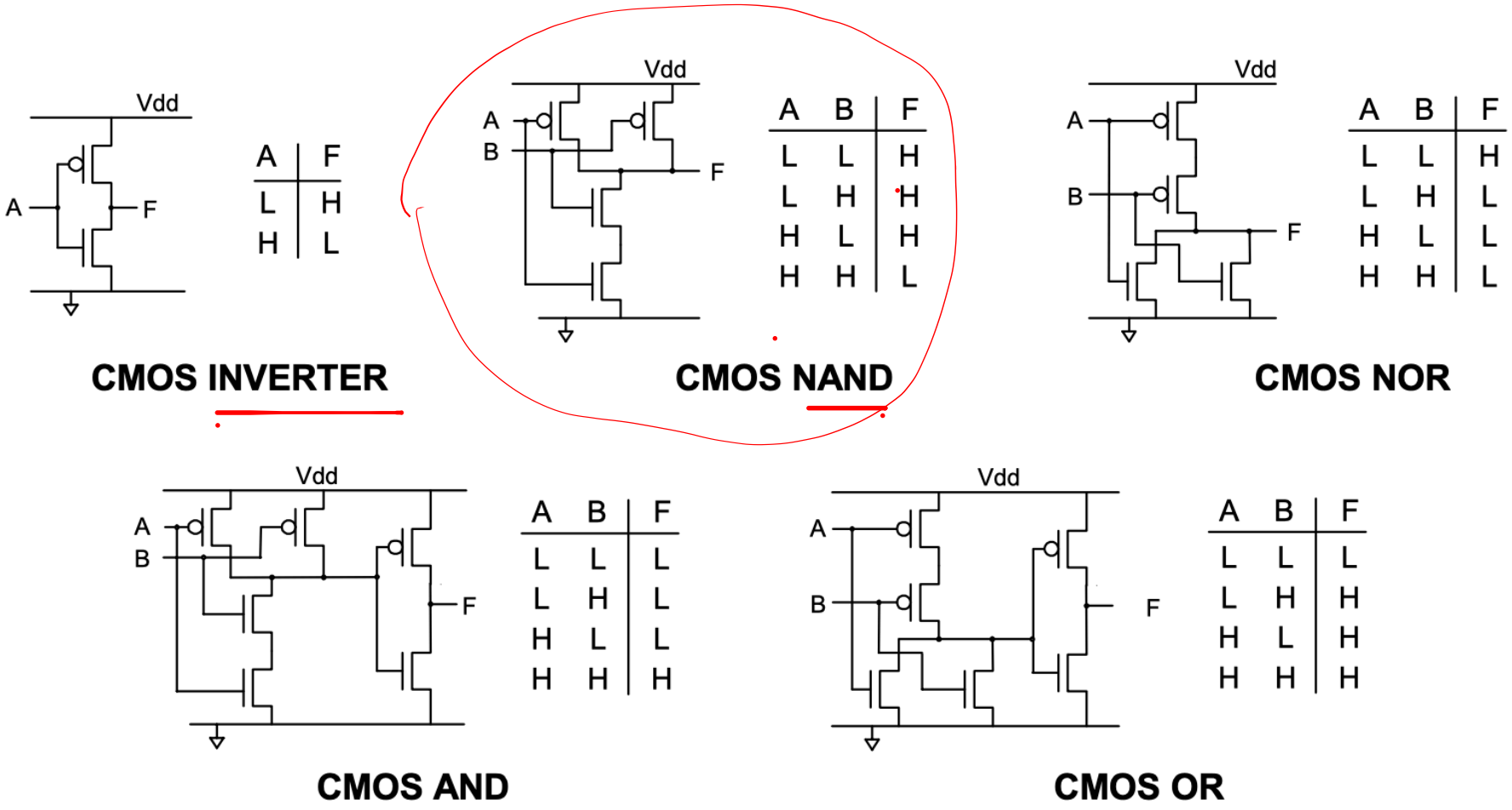
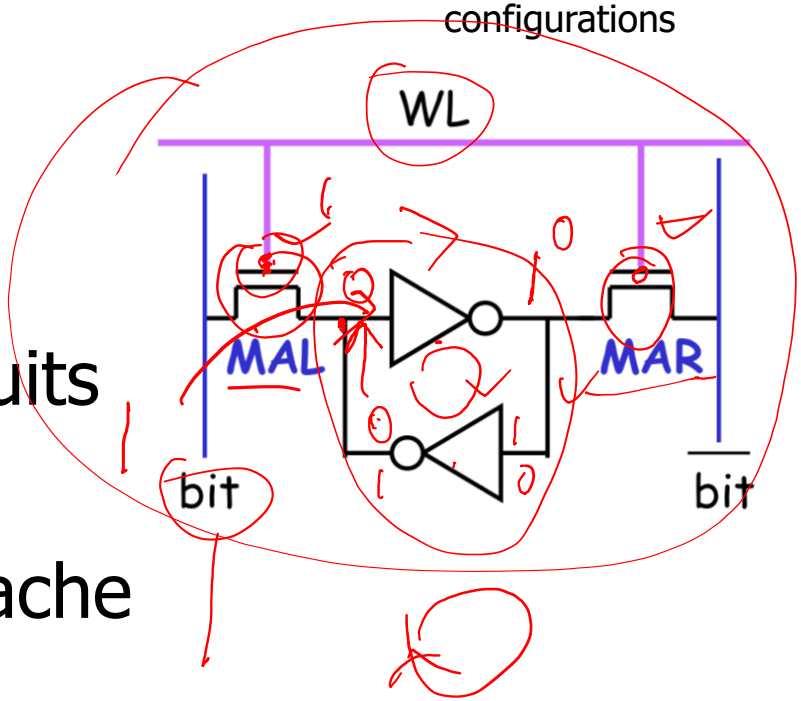


Figure 4. Basic CMOS gates and their truth tables.

Static RAM

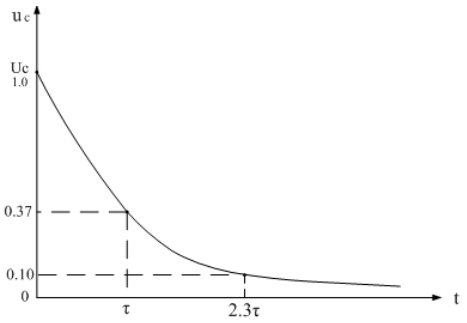
- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Does not need refresh circuits
- Faster
- For the use of register & cache

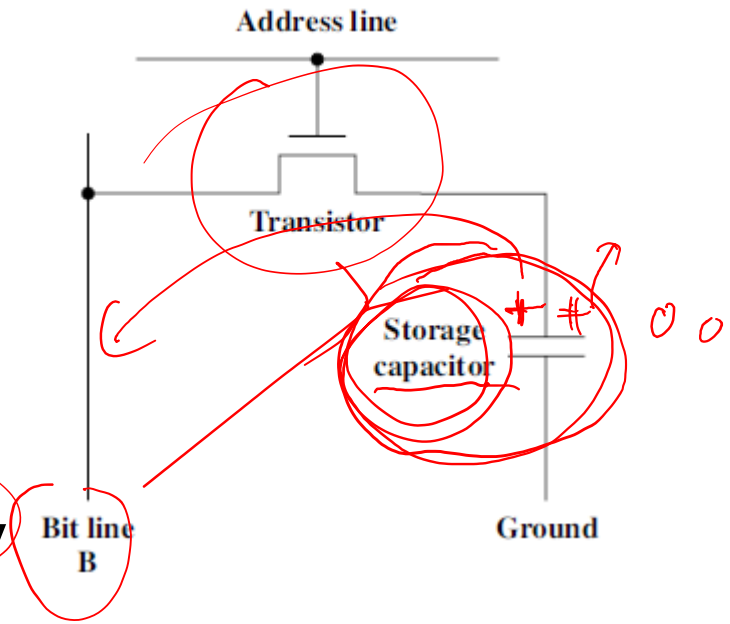
traditional **flip-flop** logic-gate configurations



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
- For the user of main memory



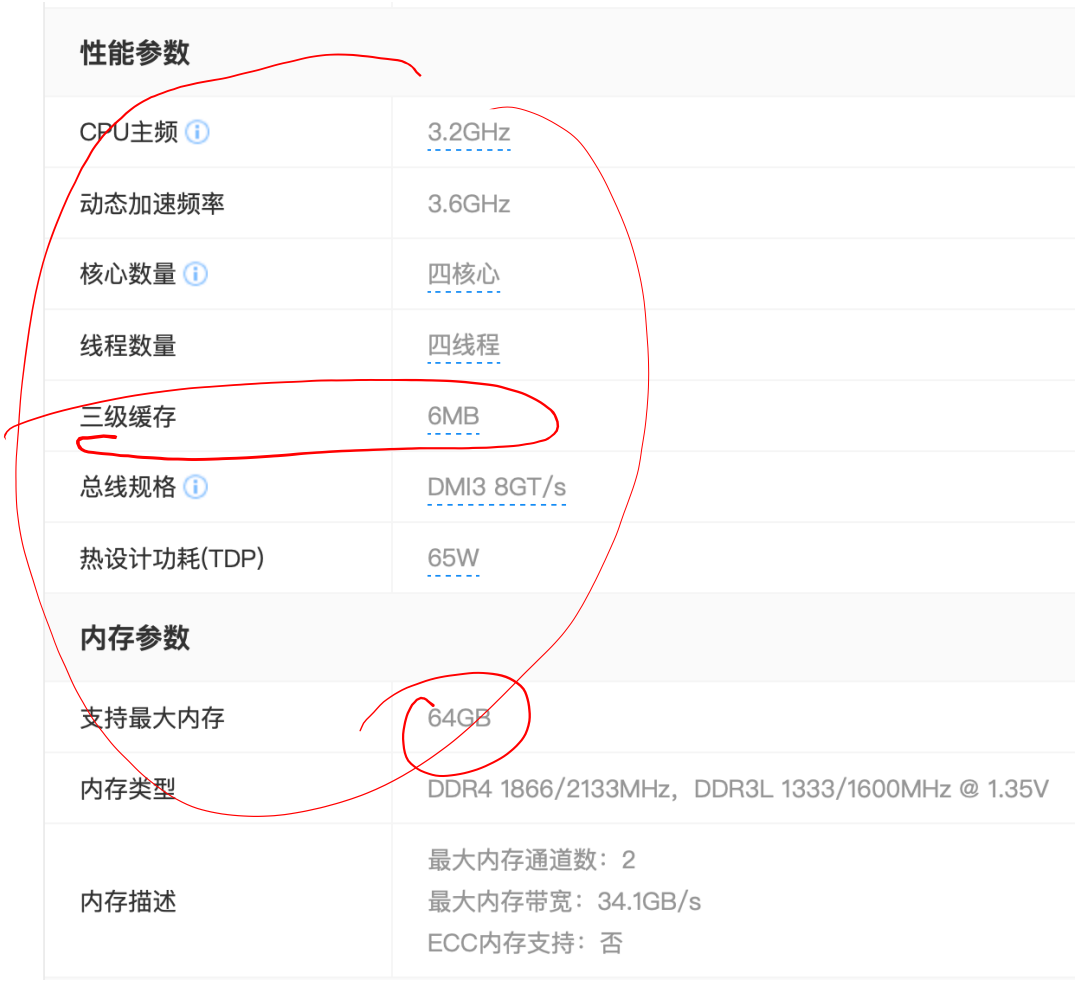


Key Content

- Memory taxonomy & characteristics
  - Physical type, location, capacity, unit of transfer, access method, performance
- Memory hierarchy in computer system
  - Registers -> Cache -> Main memory -> Disk
- RAM organization (SRAM and DRAM)
  - Cell -> Array -> Chip
- Memory module extension
  - Bit & word extension

Memory is Organized with Hierachy

- Registers
- L1 Cache
- L2 Cache
- Main memory
- Disk cache
- Disk
- Optical
- Tape

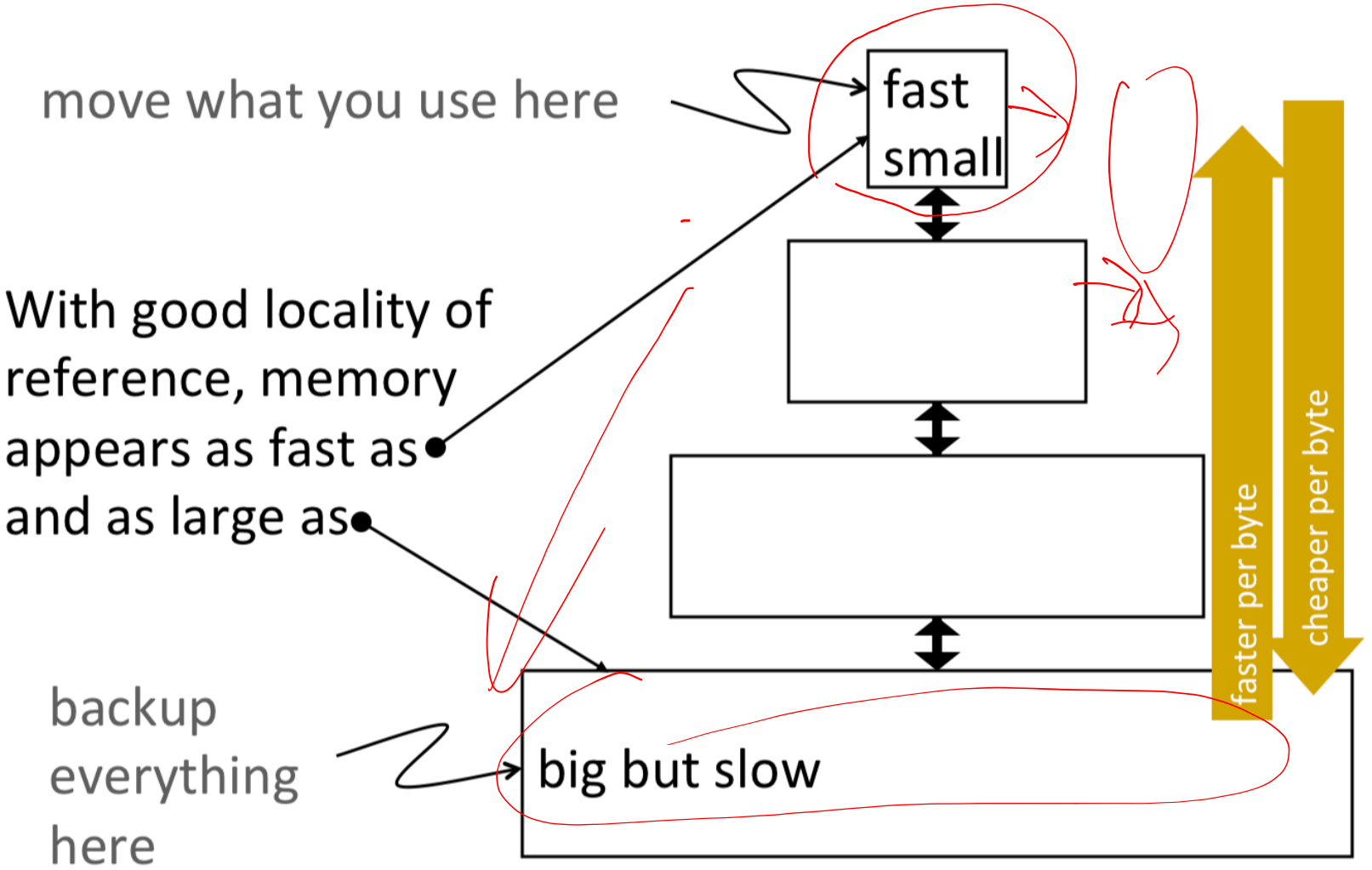


性能参数	
CPU主频	3.2GHz
动态加速频率	3.6GHz
核心数量	四核心
线程数量	四线程
三级缓存	6MB
总线规格	DMI3 8GT/s
热设计功耗(TDP)	65W
内存参数	
支持最大内存	64GB
内存类型	DDR4 1866/2133MHz，DDR3L 1333/1600MHz @ 1.35V
内存描述	最大内存通道数：2 最大内存带宽：34.1GB/s ECC内存支持：否

Why Memory Hierarchy

- Bigger is slower
  - SRAM, 512 Bytes, sub-nanosec
  - SRAM, KByte~MByte, ~nanosec
  - DRAM, Gigabyte, ~50 nanosec
  - Hard Disk, Terabyte, ~10 millisec
- Faster is more expensive (dollars and chip area)
  - SRAM, < 10$ per Megabyte
  - DRAM, < 1$ per Megabyte
  - Hard Disk < 1$ per Gigabyte
  - These sample values scale with time
- Other technologies have their place as well
  - Flash memory, PC-RAM, MRAM, RRAM (not mature yet)

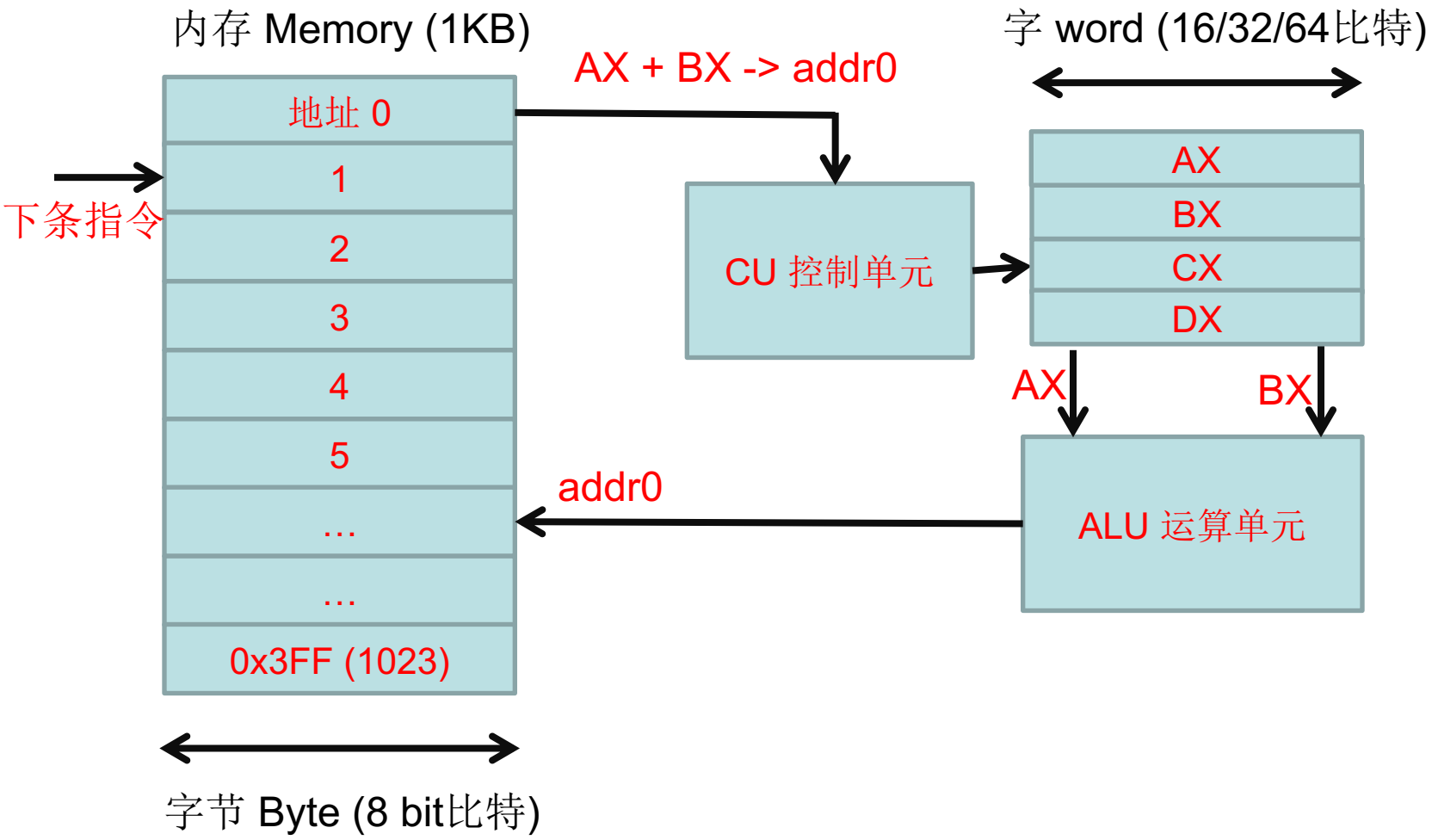
Idea Behind Memory Hierarchy



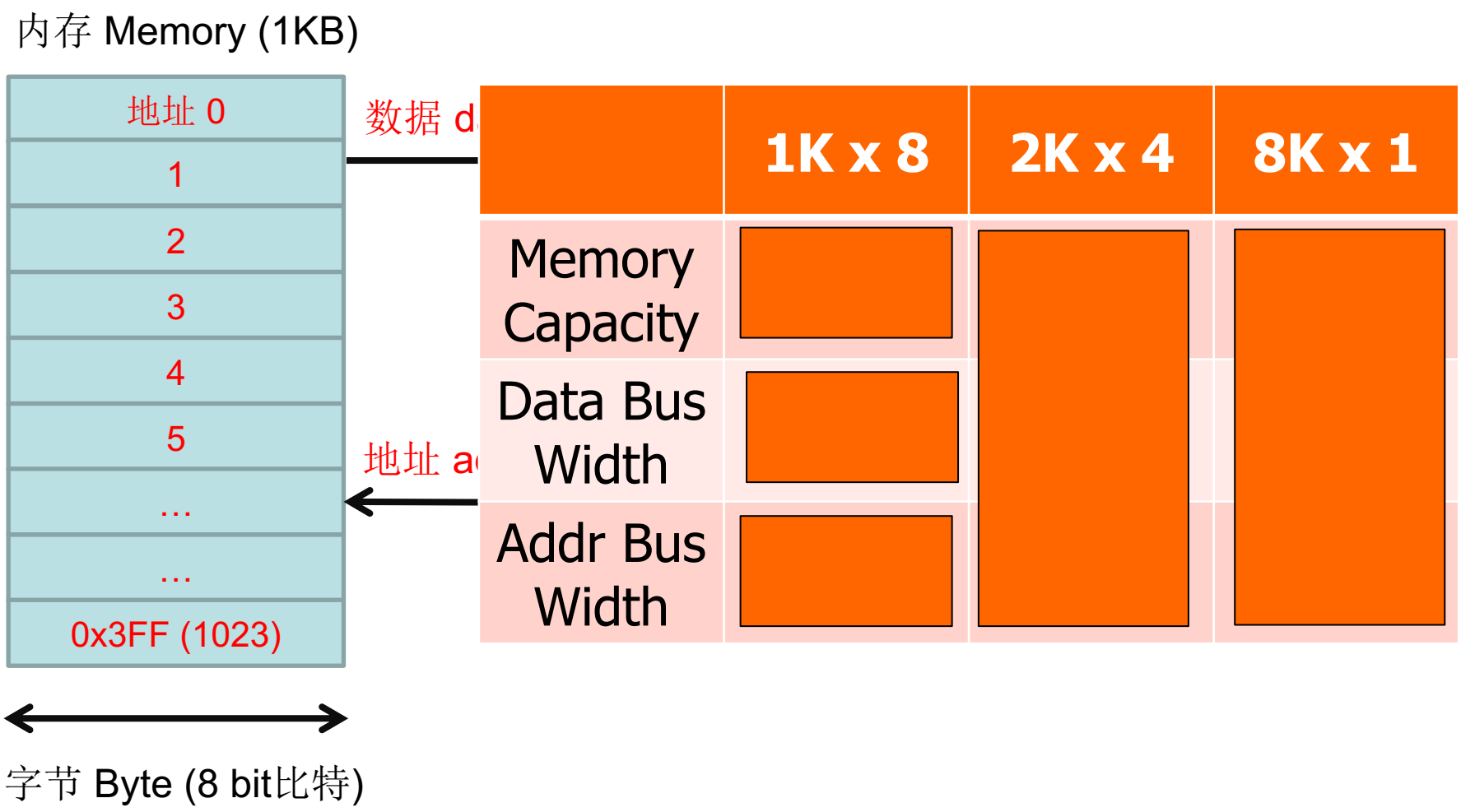
Key Content

- Memory taxonomy & characteristics
  - Physical type, location, capacity, unit of transfer, access method, performance
- Memory hierarchy in computer system
  - Registers -> Cache -> Main memory -> Disk
- RAM organization (SRAM and DRAM)
  - Cell -> Array -> Chip
- Memory module extension
  - Bit & word extension

Memory Interface with CPU



Memory Interface with CPU

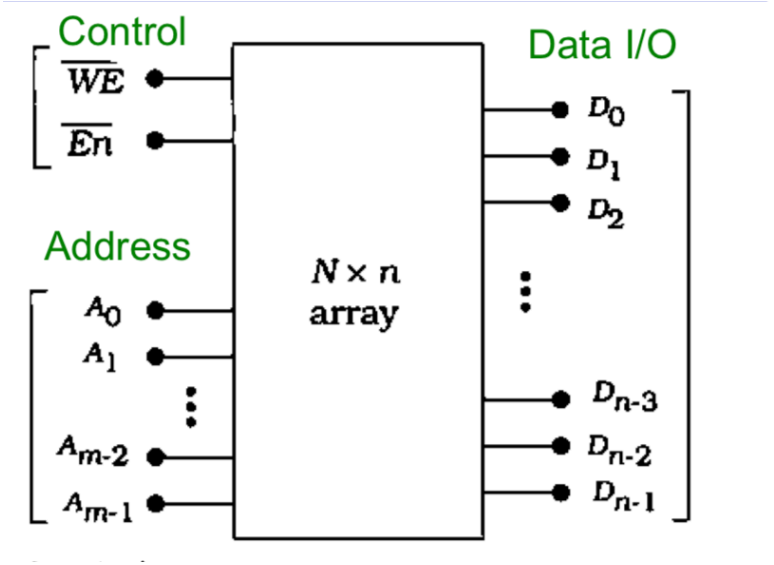


Memory Interface with CPU

	1K x 8	2K x 4	8K x 1
Memory Capacity	1 KB	1 KB	1 KB
Data Bus Width	8	4	1
Addr Bus Width	10	11	13

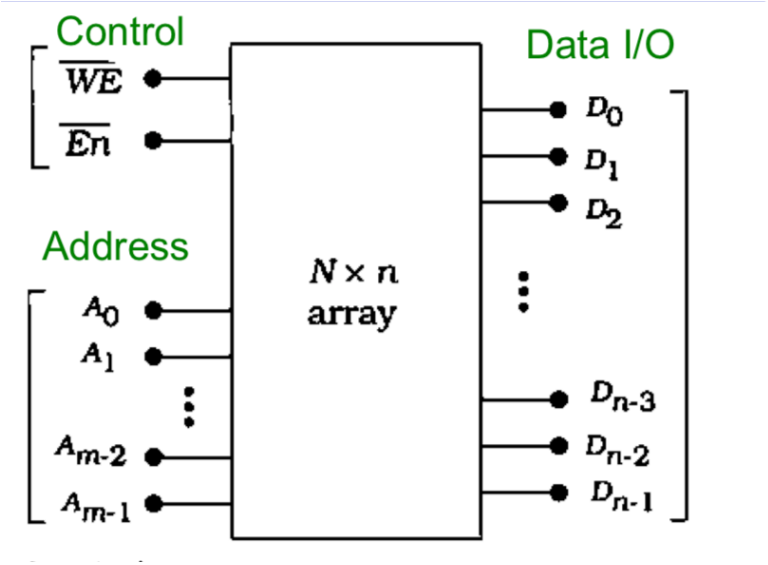
Memory Chip

- N x n memory chip
  - n = chip word width (note that this "word" is chip word not CPU word)
  - N = number of n-bit words
- Chip I/O
  - Data (in and out): Dn-1 – D0
  - Address: Am-1 – A0
  - Control
    - WE = write enable (assert low)
      - WE=1: read, WE=0: write
    - En block enable (assert low)



Memory Chip (Logical View)

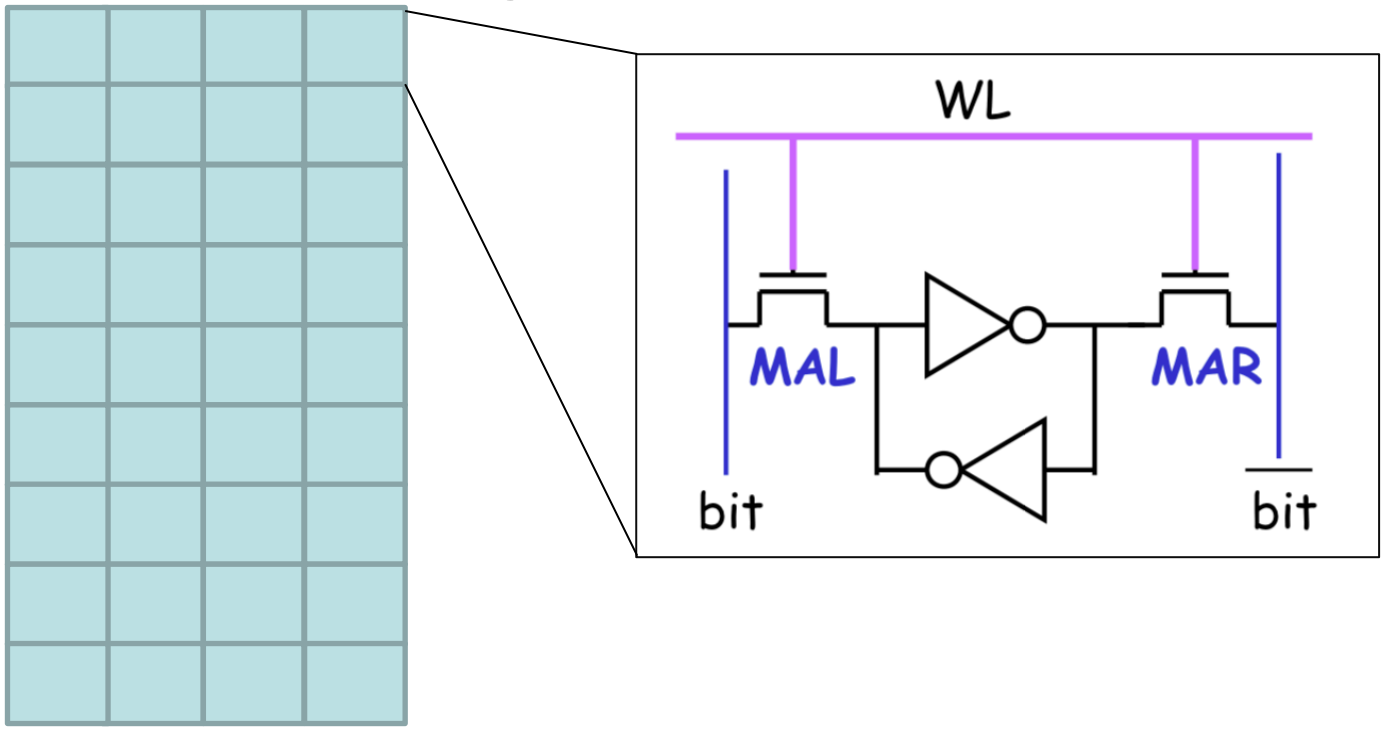
- Example: 4M x 4 RAM (16-Mbit RAM chip)
  - 2^{22} = 4 M (4x1024x1024) word
  - Each word has 4 bits
  - 22 address lines (reduced to 11 if multiplexed)
  - 4 data lines



Memory Chip (Physical View)

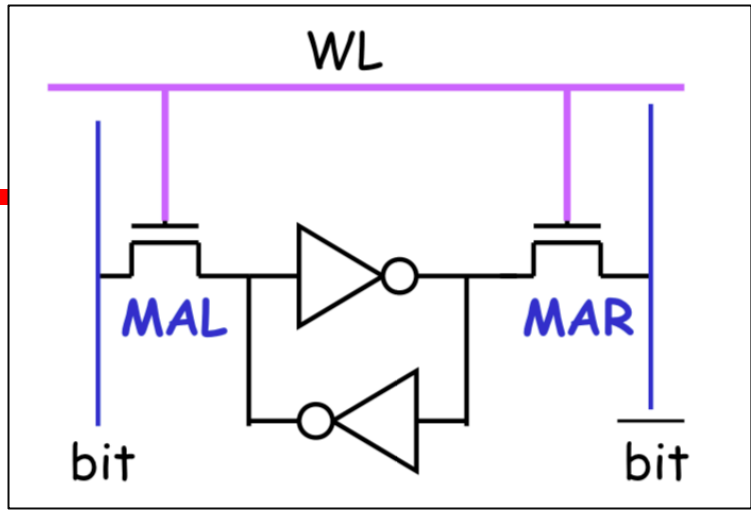
- Example: 4M x 4 RAM (16-Mbit RAM chip)
- A 16Mbit chip has 16M SRAM cells
- But how are they organized?

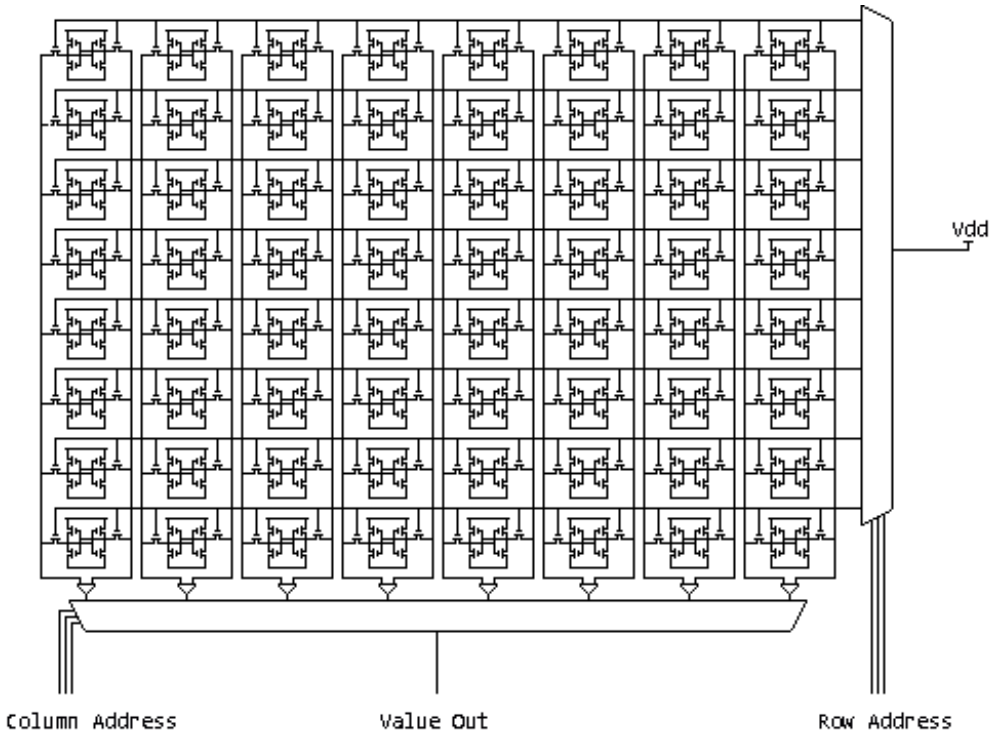
Is this 4M x 4 array a good organizat ion?



Memory Array

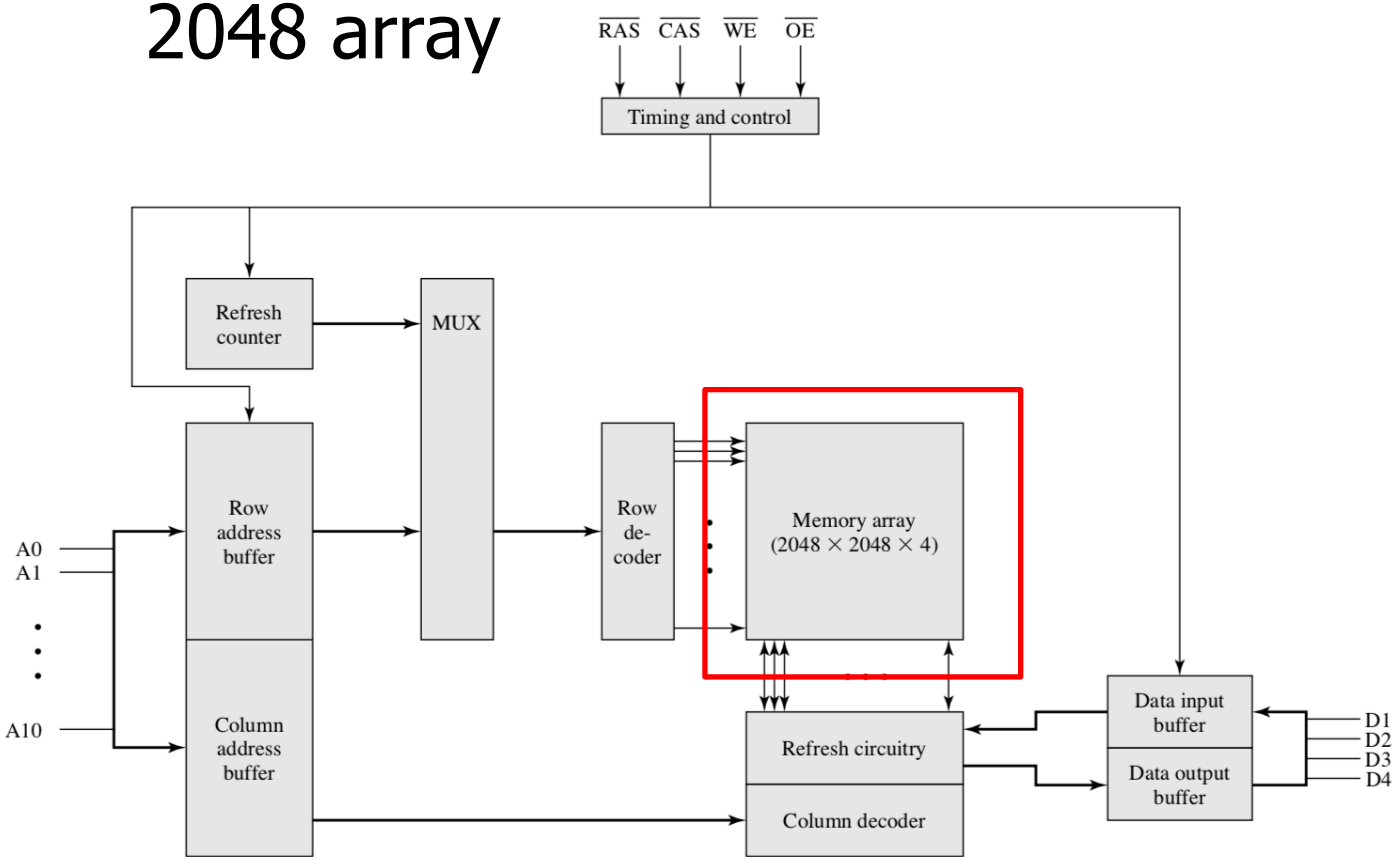
- N_R x N_C array of 1-bit cells
  - N_R = number of rows
  - N_C = number of columns

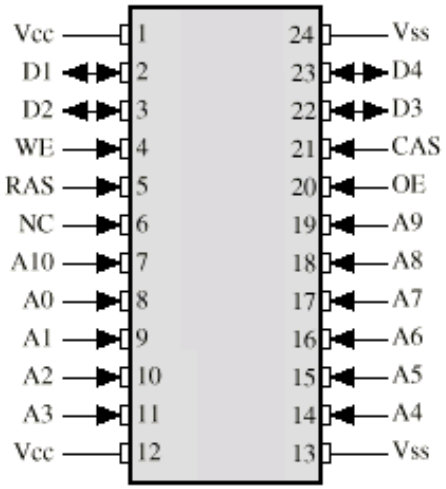




Typical 16 Mb RAM (4M x 4)

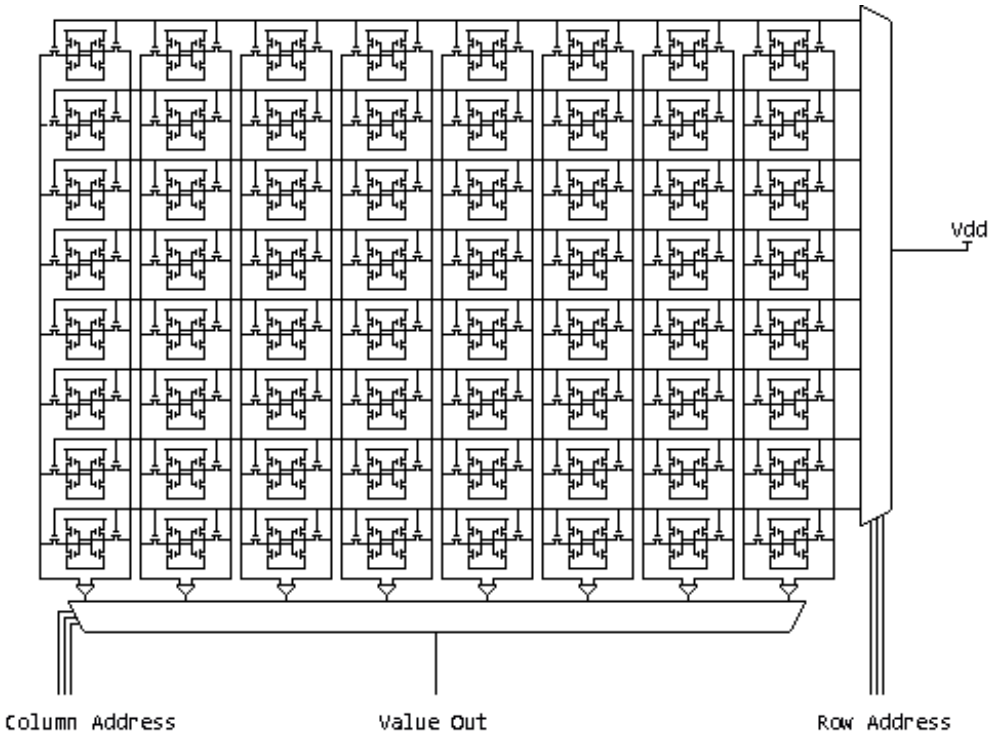
- The 4M x 4 RAM is organized as four 2048 x 2048 array

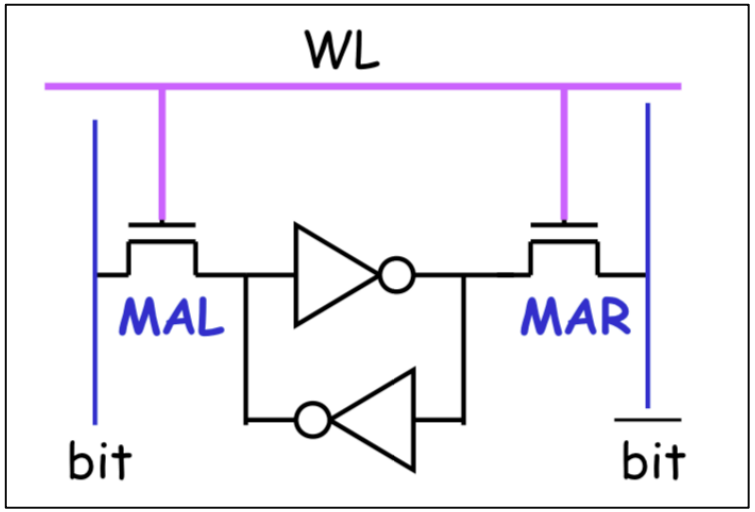




(b) 16 Mbit DRAM

How to Access Memory Array



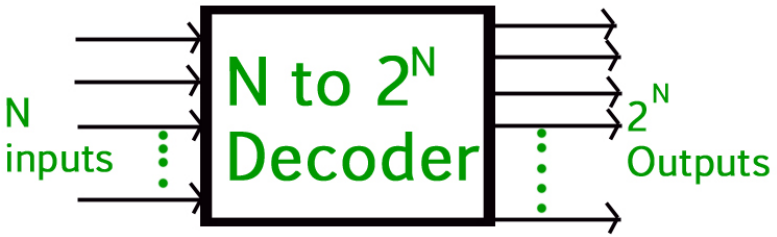


How to Access Memory Array

- Example of 4 x 1 array
- Example of 16 x 1 array

How to Access Memory Array

- Example of 4 x 1 array
- Example of 16 x 1 array

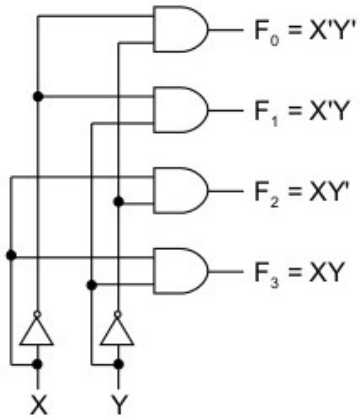


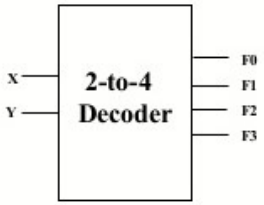
2-to-4 Binary Decoder

Truth Table:

X	Y	F_0	F_1	F_2	F_3
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	1

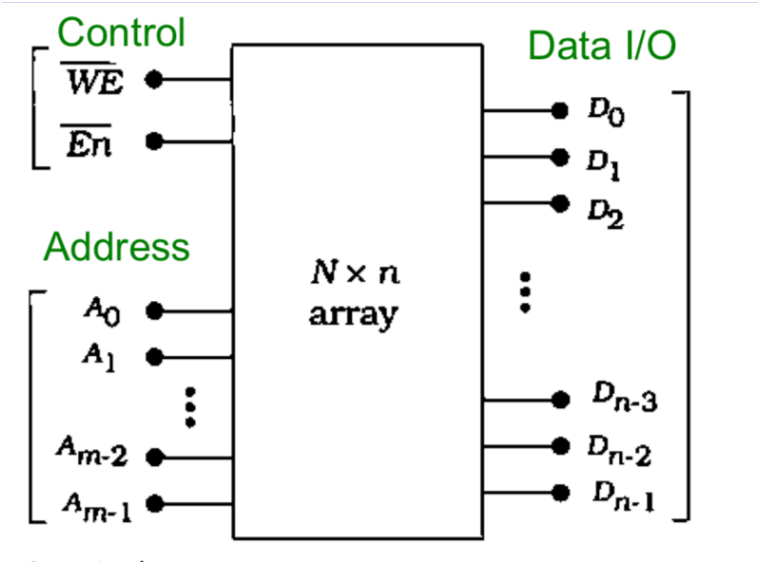
- From truth table, circuit for 2x4 decoder is:
- Note: Each output is a 2-variable minterm (X'Y', X'Y, XY' or XY)

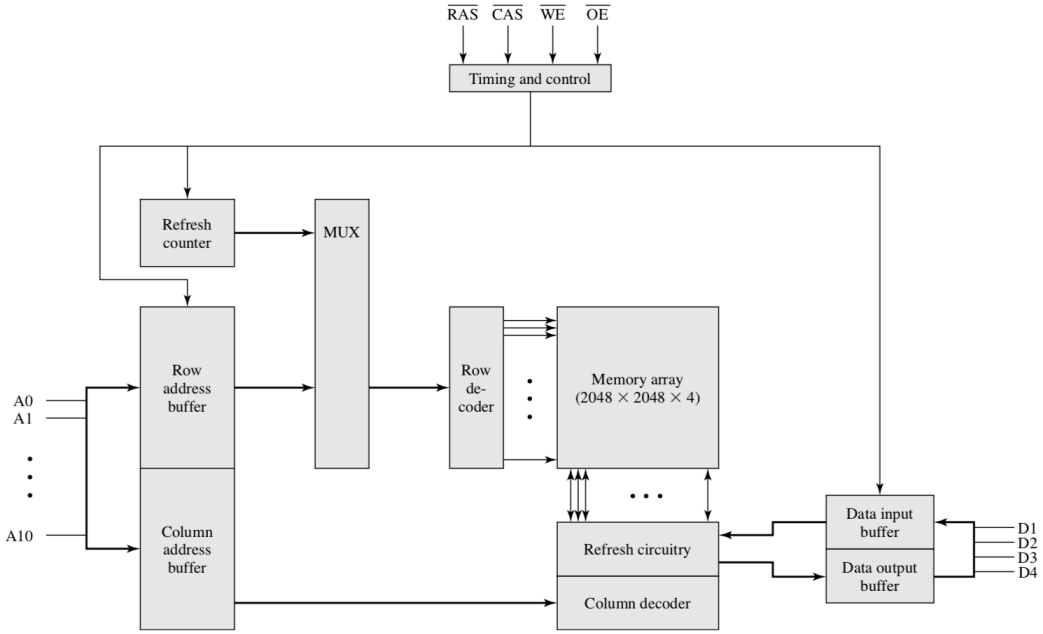




Key Takeaway in This Part

- Understand the logical interface for a memory chip
- Understand the internal organization of the chip



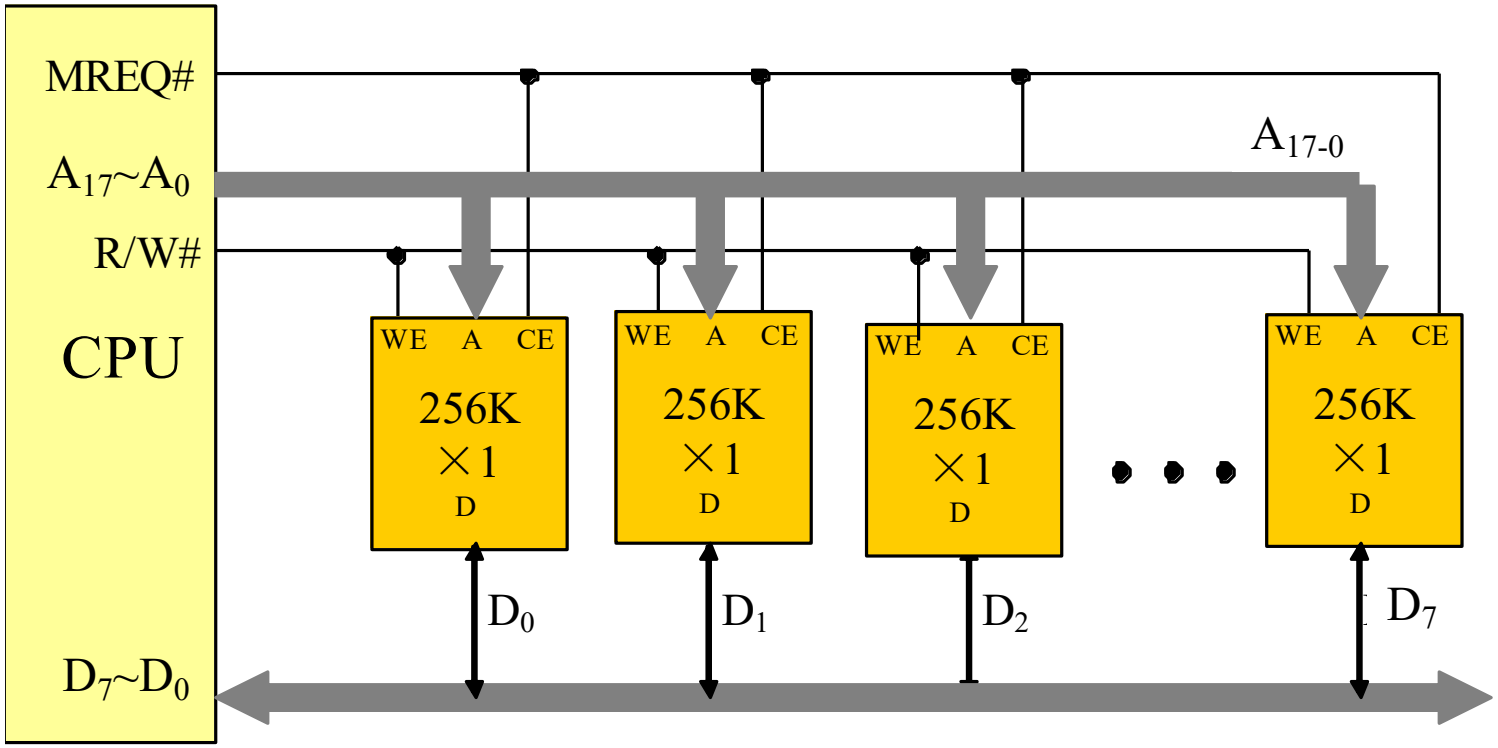


Key Content

- Memory taxonomy & characteristics
  - Physical type, location, capacity, unit of transfer, access method, performance
- RAM organization (SRAM and DRAM)
  - Cell -> Array -> Chip
- Memory hierarchy in computer system
  - Registers -> Cache -> Main memory -> Disk
- Memory module extension
  - Bit & word extension

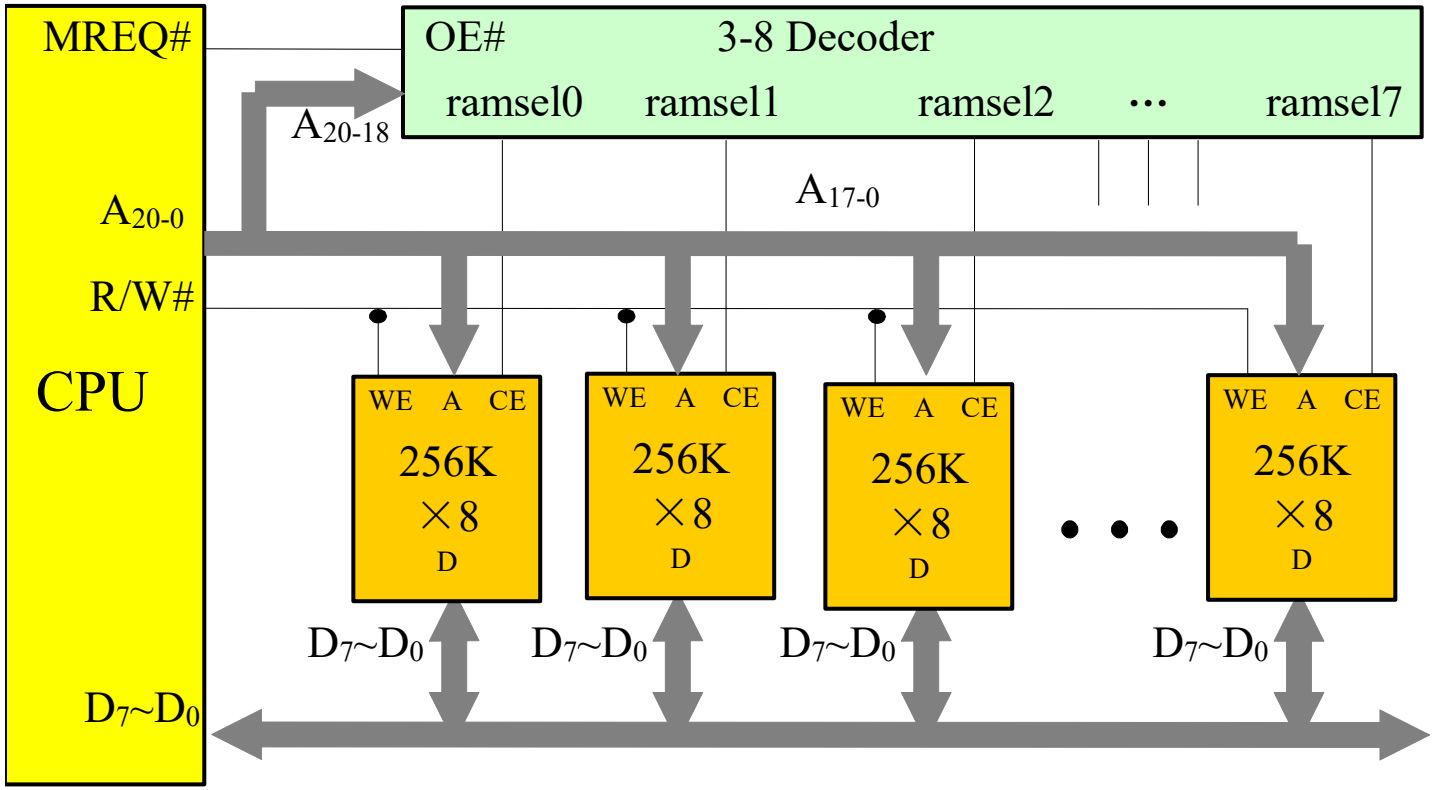
Bit Extension Example

You have 256K×1-bit RAM chips. How can you build a memory module of 256 KB (with word size of 8 bits) and how to connect this module with a computer system?



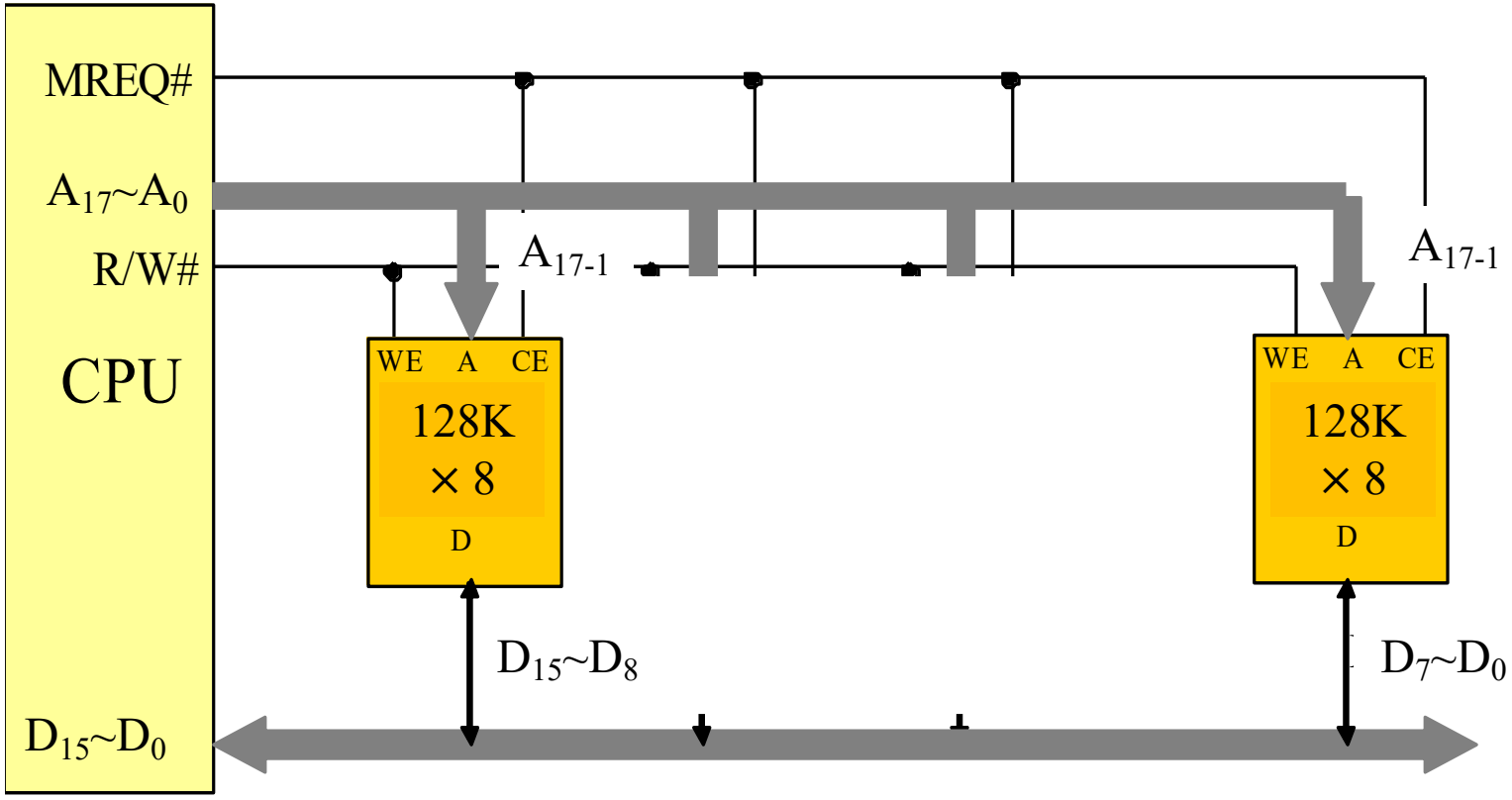
Word Extension Example

You have 256K×8-bit RAM chips. How can you build a memory module of 2M×8-bit and how to connect this module with a computer system?



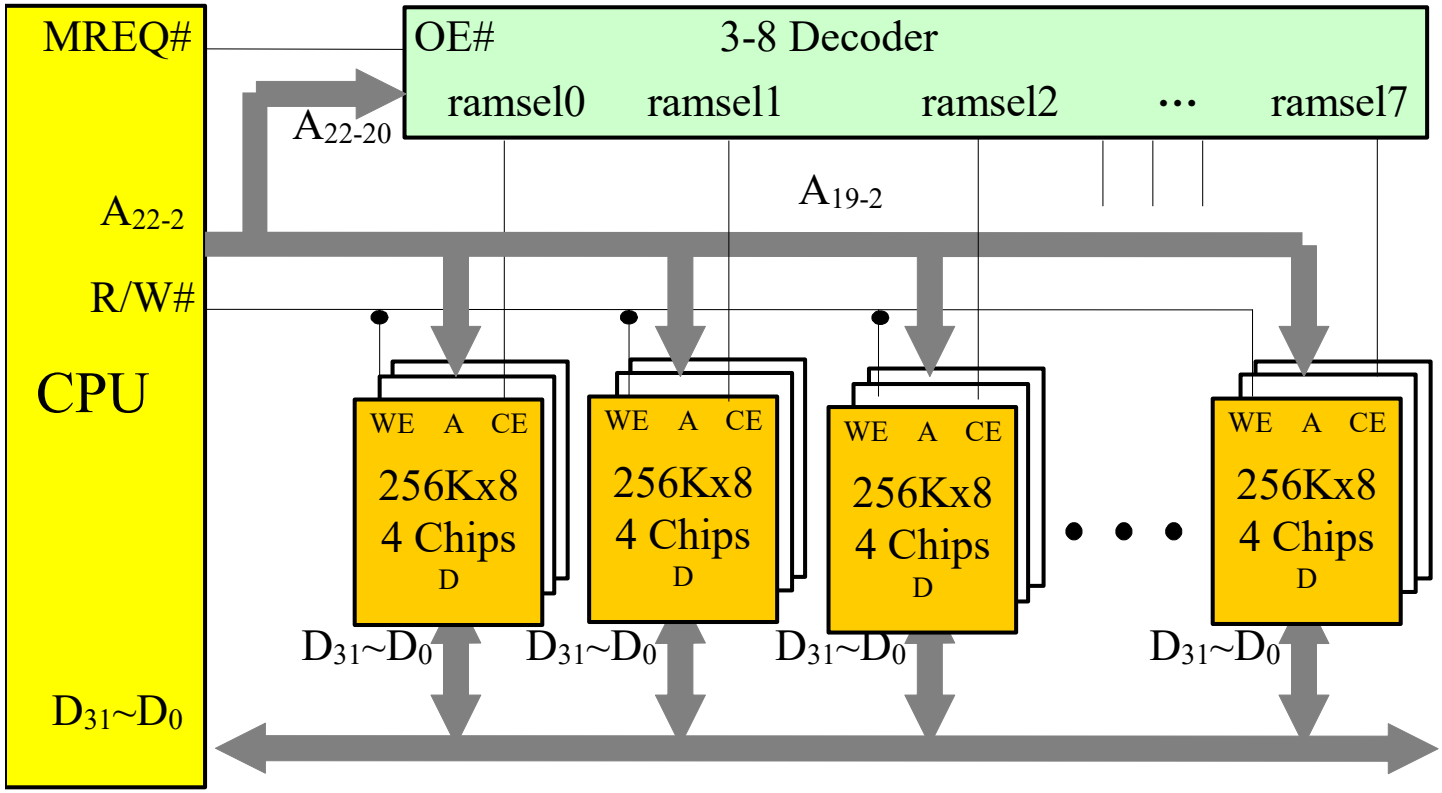
Quiz

You have 128K×8-bit RAM chips. How can you build a memory module of 128K×16-bit and how to connect this module with a computer system?

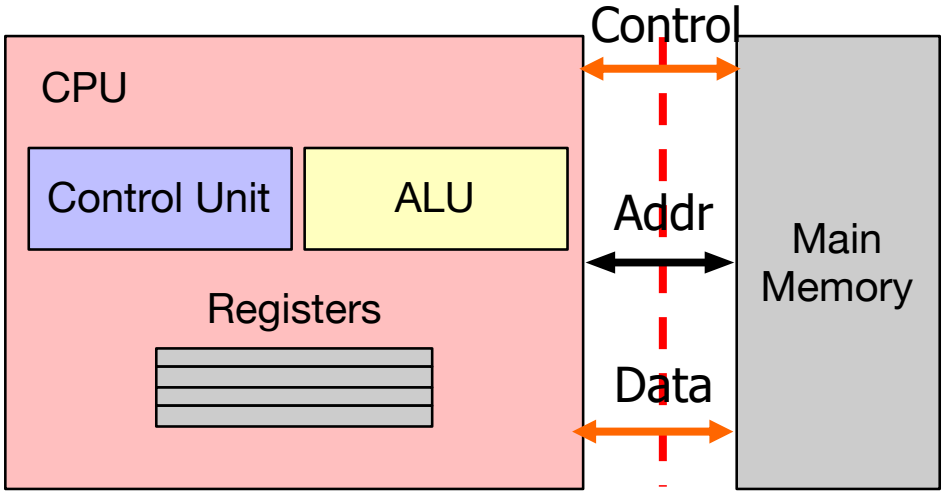


Word and Bit Extension Example

Now you have 256K×8-bit RAM chips. How can you build a memory module of 2M×32-bit and how to connect this module with a computer system if the addressable unit is byte?



Memory Module Extension Summary



- CPU Word
  - The size of instruction operand
    一条指令中数据的宽度
  - Add AX, BX, CX (Size of register AX/BX/CX)
- CPU addressable unit
  - The size of data represented by an address
  - Usually one byte (8 bits)

Interface
- Address bus width
- Data bus width

Memory chip N x n
- n: size of memory word
- N: number of words

Memory Module Extension Summary

- When to use word/bit extension
  - CPU wants the memory module of N_{CPU} x n_{CPU}
  - Memory chip has N_{mem} x n_{mem}
- Word extension: $N_{CPU} > N_{mem}$
  - Connect data bus of CPU and all memory chips directly
  - Requires extra decoder for address bus connection
- Bit extension : $n_{CPU} > n_{mem}$
  - Connect address bus of CPU and all memory chips directly
    - May discard some address lines when n_{CPU} > CPU addressable unit
  - Assemble memory chips' data buses to connect with CPU

Quiz

- For a 1MB memory module
  - If addressable unit equals 8 , What's the address range
  - If addressable unit equals 32 , What's the address range
- If we have 256K x 4bit memory chip
  - If data bus width = 8 , how do we extend?
  - If data bus width = 32 , how do we extend?

0x00000-0xFFFFF

0x00000-0x3FFFF

CPU wants 1M x 8:
(1M/256K)x(8/4)=4x2
4 times word extension and 2 times bit extension

CPU wants 256K x 32:
(256K/256K)x(32/4)=1x8
8 times bit extension (need to discard address A_0-A_1lines)

Read Only Memory (ROM)

- Permanent storage
- Nonvolatile
  - Needs no power
- Normal applications of ROM (Firmware)
  - 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 Ultraviolet radiation
  - Electrically Erasable (EEPROM)
    - Takes much longer to write than read
  - Flash memory
    - Erase whole memory electrically

Key Content Review

- Memory taxonomy & characteristics
  - Physical type, location, capacity, unit of transfer, access method, performance
- RAM organization (SRAM and DRAM)
  - Cell -> Array -> Chip
- Memory hierarchy in computer system
  - Register -> cache -> main memory -> disk
- Memory module extension
  - Bit & word extension

Assignment Two

- Now you have 128K×4-bit RAM chips. How can you build a memory module of 1M×64-bit and how to connect this module with a computer system if the addressable unit is byte?
  - Draw the address/data lines according to the example in the slides