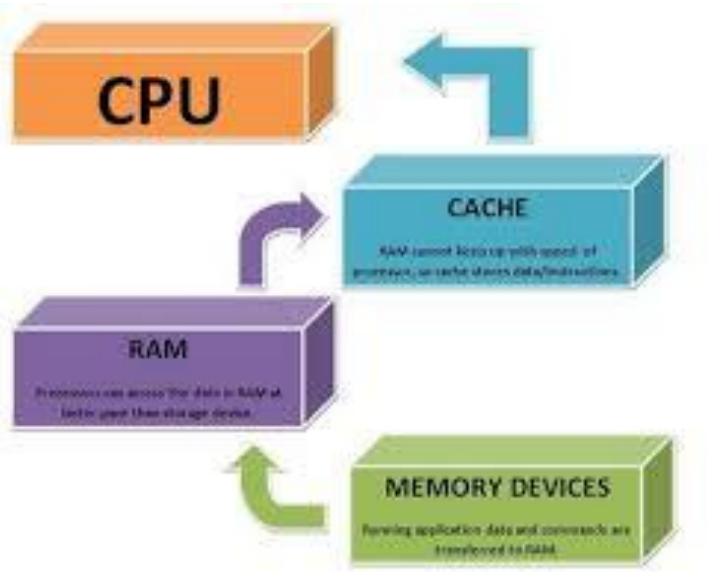
ساختار و زبان کامپیوتر

فصل نهم

سازمان حافظه



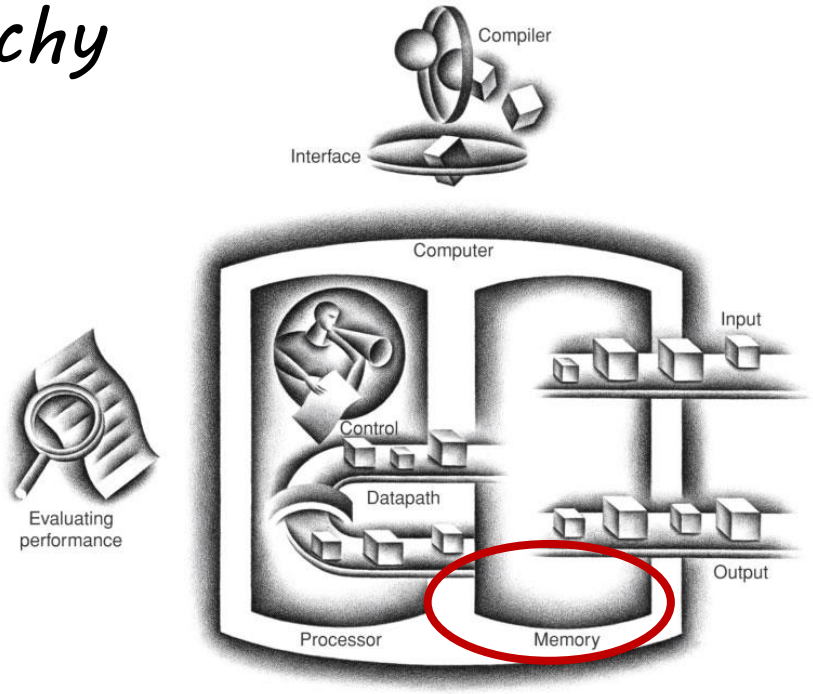
Copyright Notice

Parts (text & figures) of this lecture are adopted from :

- D. Patterson & J. Hennessey, "Computer Organization & Design, The Hardware/Software Interface", 5th Ed., MK publishing, 2014
- W. Stallings, "Computer Organization and Architecture, Designing for Performance", 10th Ed., Pearson
- Morris Mano and Michael Ciletti, "Digital design: with an introduction to the Verilog HDL", 5th Edition, Pearson, 2013

Outlines

- Characteristics of Memory Systems
- Internal Memory Classification
- Memory Hierarchy



Memory

- A collection of cells capable of storing binary information
- Contains electronic circuits for storing and retrieving the information
- Used in many different parts of a computer, providing temporary or permanent storage for substantial amounts of binary information

Key Characteristics

Location
- Internal (e.g., processor registers, cache, main memory)
- External (e.g., optical disks, magnetic disks, tapes)

Capacity
- Number of words
- Number of bytes

Unit of Transfer
- Word
- Block

Access Method
- Sequential
- Direct
- Random
- Associative

Performance
- Access time
- Cycle time
- Transfer rate

Physical Type
- Semiconductor
- Magnetic
- Optical
- Magneto-optical

Physical Characteristics
- Volatile/nonvolatile
- Erasable/nonerasable

Organization
- Memory modules

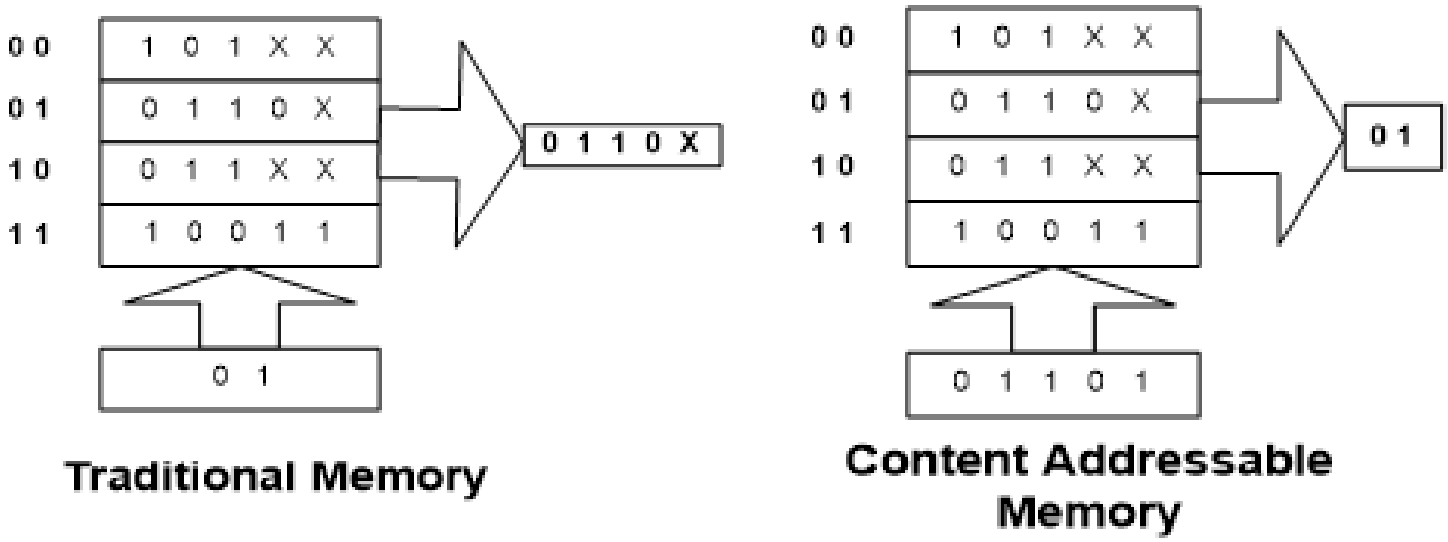
Internal vs. External Memory

- Internal
  - Semiconductor memories
  - Register, Cache, Main Memory
- External
  - Magnetic/ Optical/ Semiconductor
  - Hard disks, Optical disks, SSD

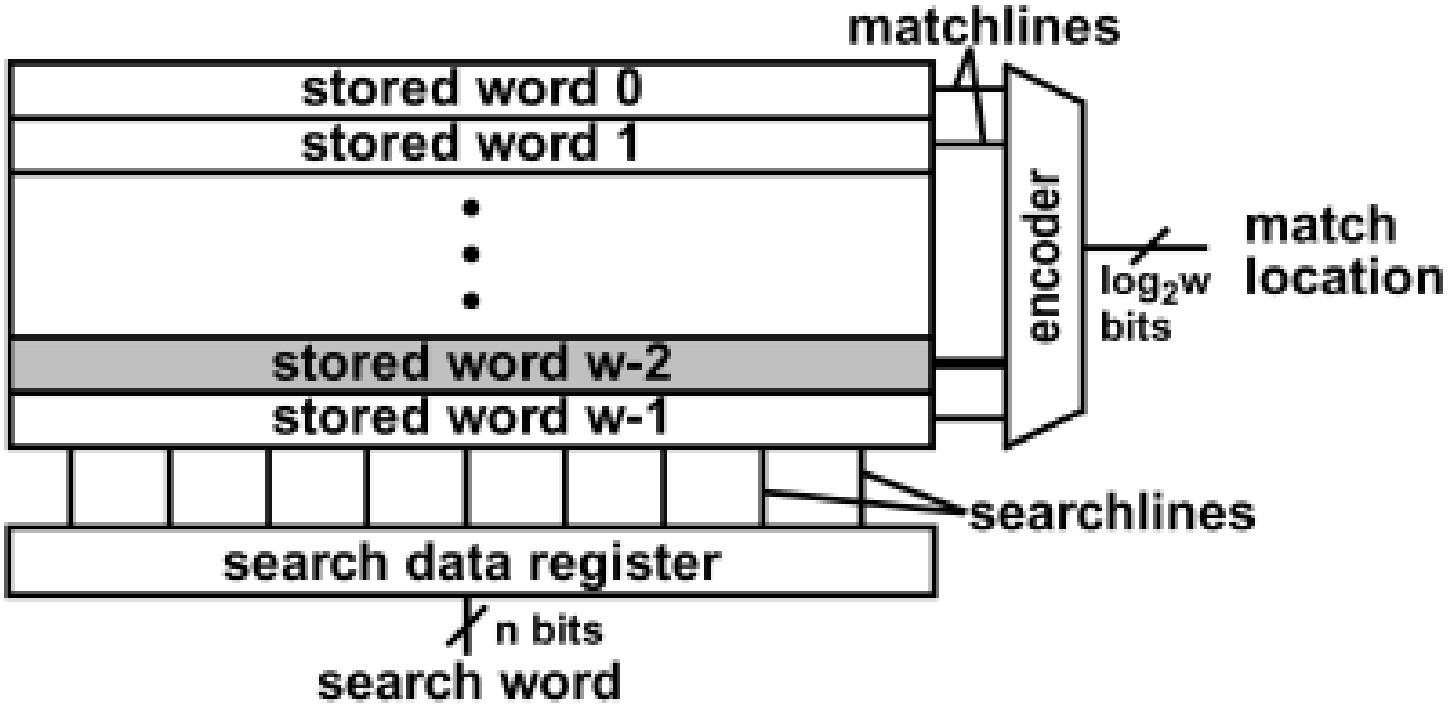
Semiconductor Memory

- Content Addressable Memory (CAM)
- Sequential Access Memory (SAM)
- Random Access Memory (RAM)

Content Addressable Memory (CAM)



CAM: Closer View



K. Pagiamtzis, A. Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey," *IEEE J. of Solid-state circuits. March 2006*

Sequential Access Memory (SAM)

- Shift Registers
  - SISO (Serial-in-Serial-out)
  - SIPO (Serial-in-Parallel-out)
  - PISO (Parallel-in-Serial-out)
- Queues
  - FIFO (First-in-First-out)
  - LIFO (Last-in-First-out)

Random Access 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 (RWM)</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 (ROM)</td><td rowspan="2">Not possible</td><td>Masks</td><td rowspan="6">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 (RMM)</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>

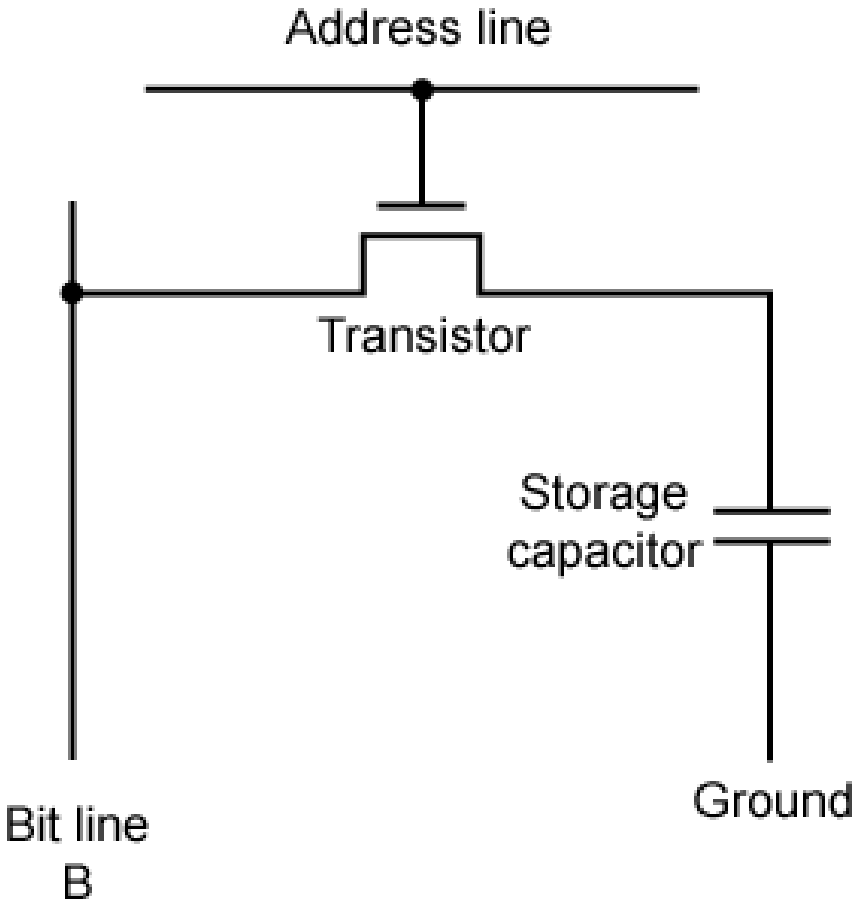
Semiconductor Memory

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

Dynamic RAM (DRAM)

- 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

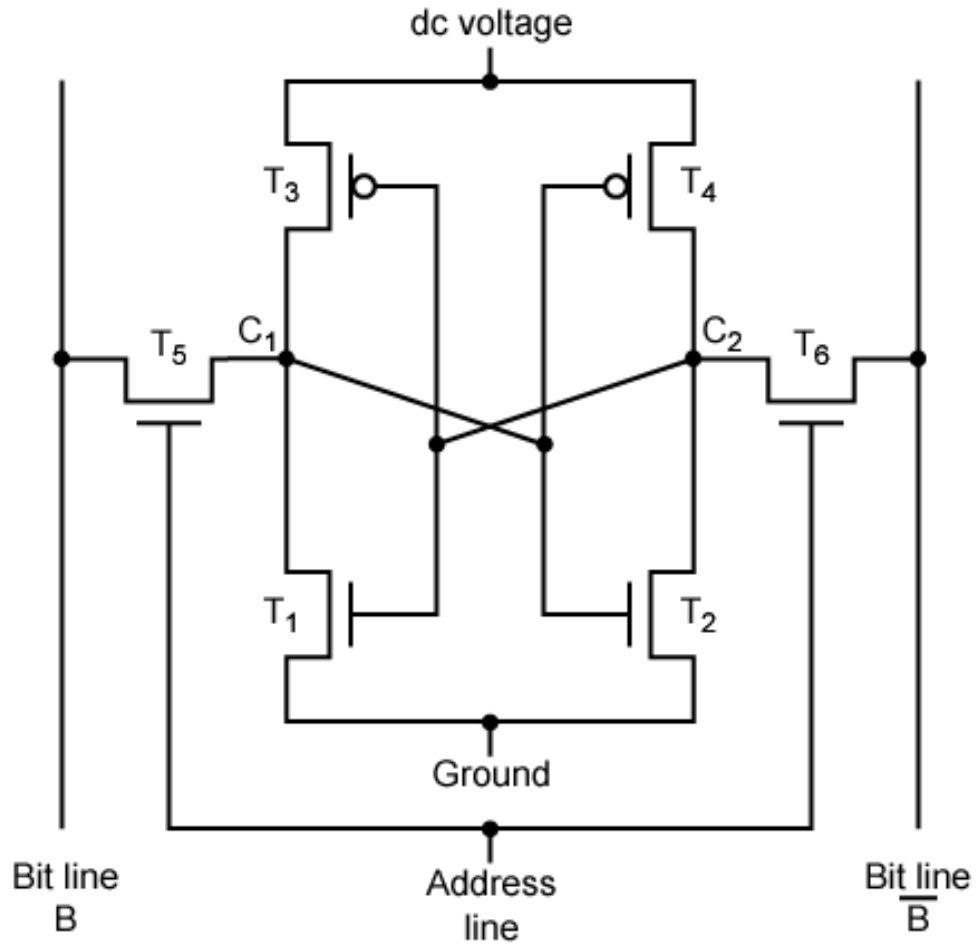
Dynamic RAM Structure



Static RAM (SRAM)

- Bits stored as on/off switches
- No charges to leak
- No refreshing needed when powered
- More complex construction
- Larger per bit
- More expensive
- Faster
- Cache
- Digital
  - Uses flip-flops

Static RAM Structure



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

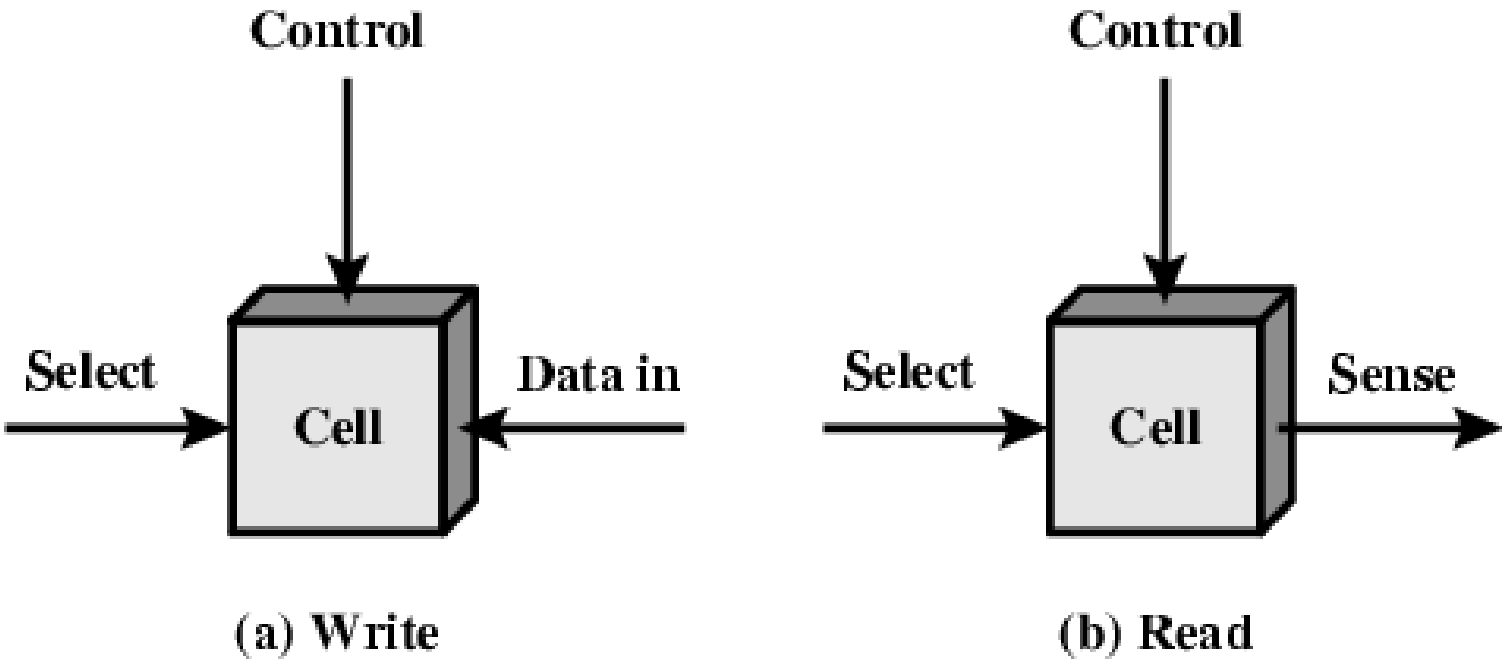
Read Only Memory (ROM)

- Permanent storage
  - Nonvolatile
- Written during manufacture
  - Very expensive for small runs
- Programmable (once)
  - PROM
  - Needs special equipment to program

Read Mostly Memory (RMM)

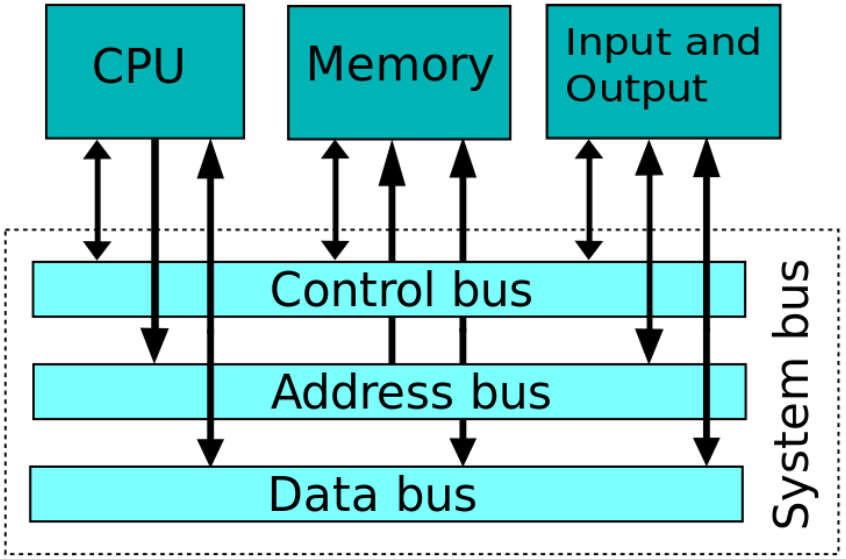
- Read "mostly"
  - Erasable Programmable (EPROM)
    - Erased by UV
  - Electrically Erasable (EEPROM)
    - Takes much longer to write than read
  - Flash memory
    - Erase blocks of memory electrically

Memory Cell Operation



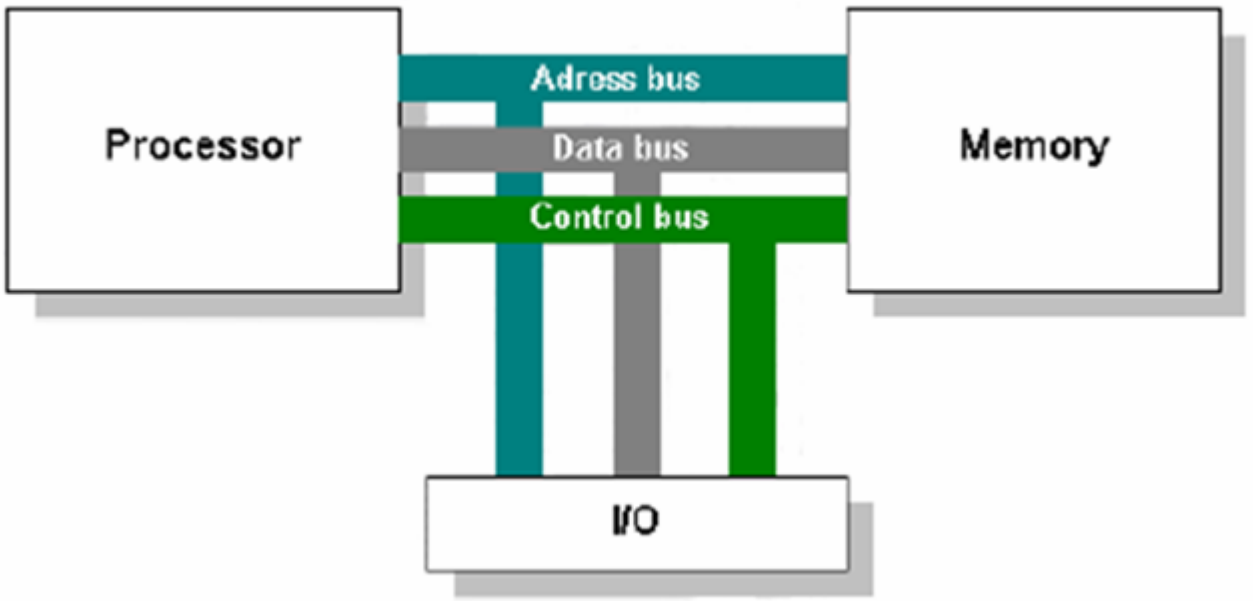
Communication with CPU

Memory, CPU & I/O devices communicate via **BUS**



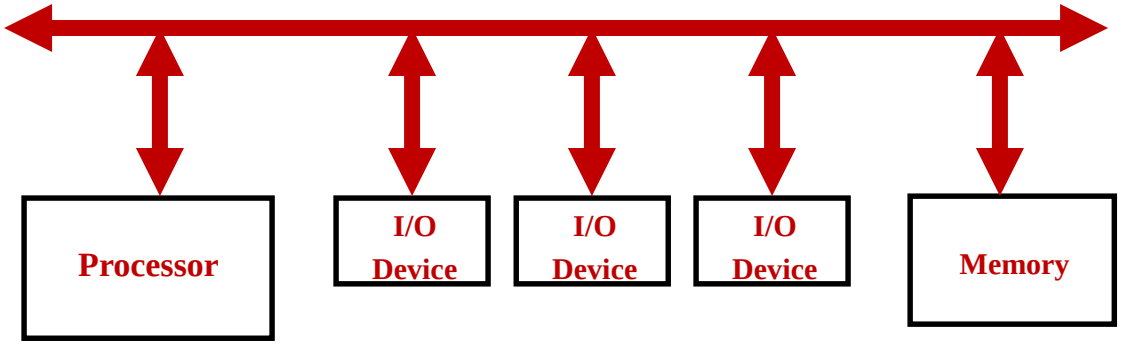
What is a Bus?

- Single set of wires used to connect multiple subsystems
- Shared communication link with multiple drivers



Disadvantage of Buses

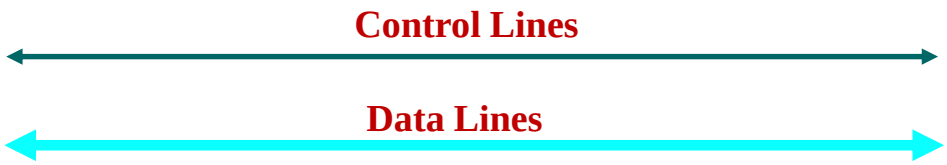
- Creates a communication bottleneck
  - Bus bandwidth limits maximum I/O throughput
- Maximum bus speed is largely limited by:
  - Length of bus
  - Number of devices on bus
  - Slowest device on bus



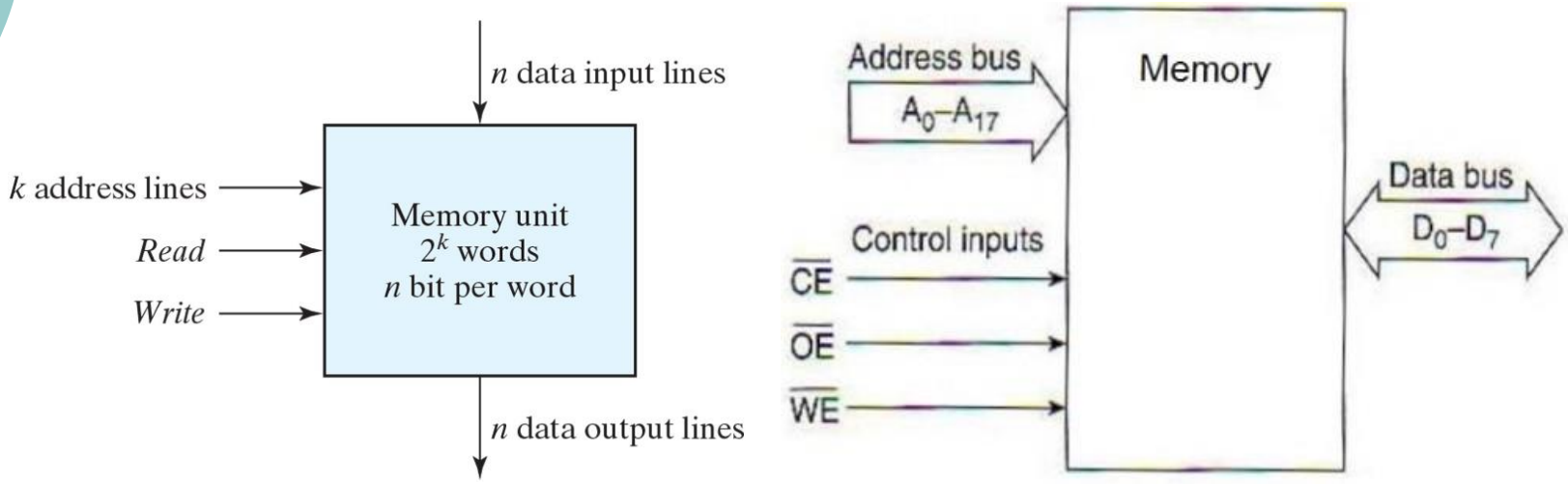
23

General Organization of Buses

- Control Lines
  - Signal requests and acknowledgments
  - Indicate what type of information is on data lines
- Data Lines
  - Carry information between source and destination
  - Data and Addresses
  - Complex commands



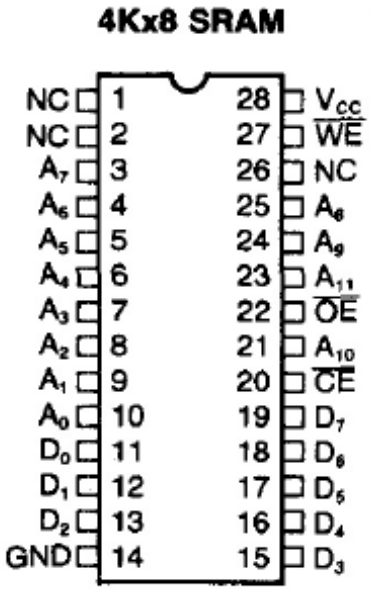
Memory Unit Access

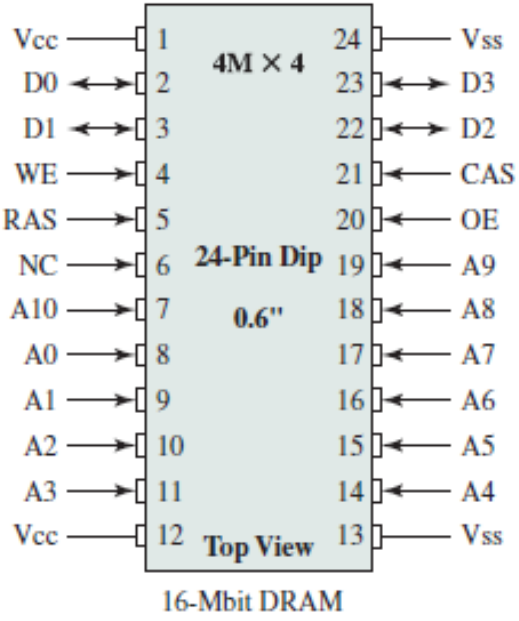


Memory Contents

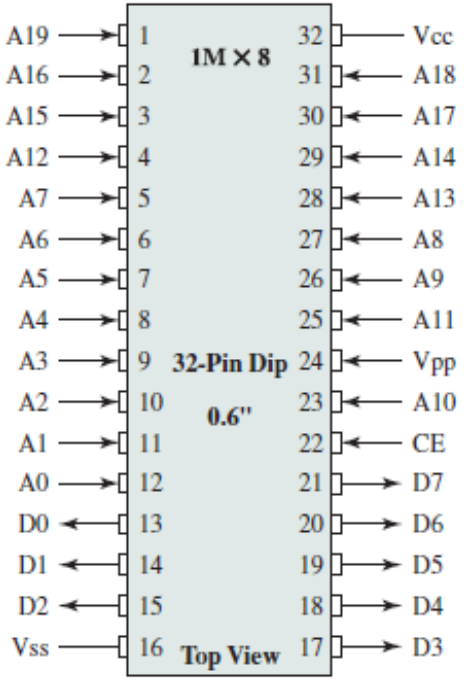
Memory address		Memory content
Binary	Decimal	
0000000000	0	1011010101011101
0000000001	1	1010101110001001
0000000010	2	0000110101000110
⋮	⋮	⋮
1111111101	1021	1001110100010100
1111111110	1022	0000110100011110
1111111111	1023	1101111000100101

Memory Chips



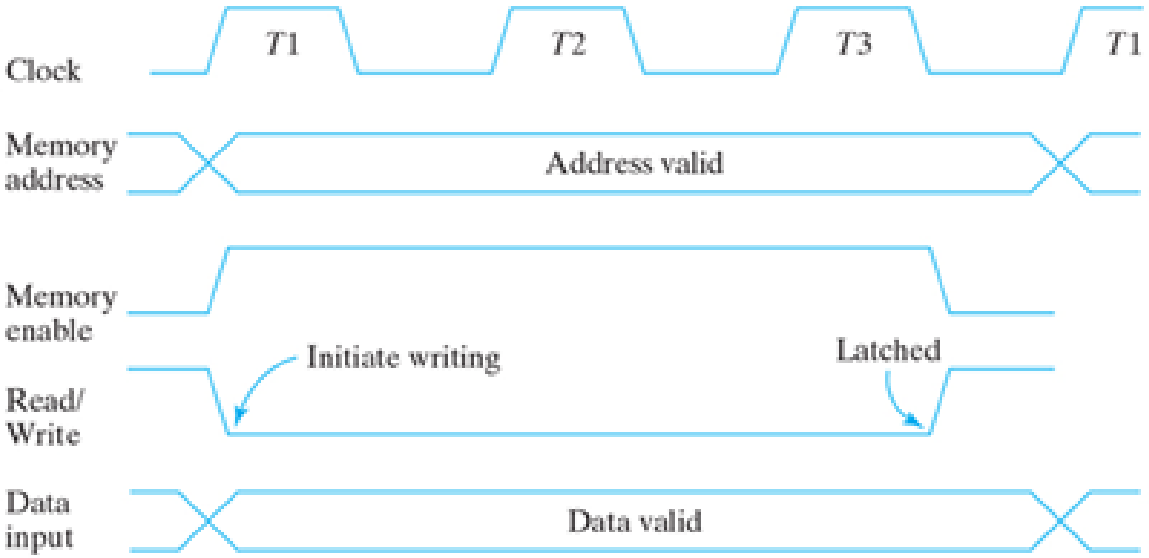


16-Mbit DRAM



8-Mbit EPROM

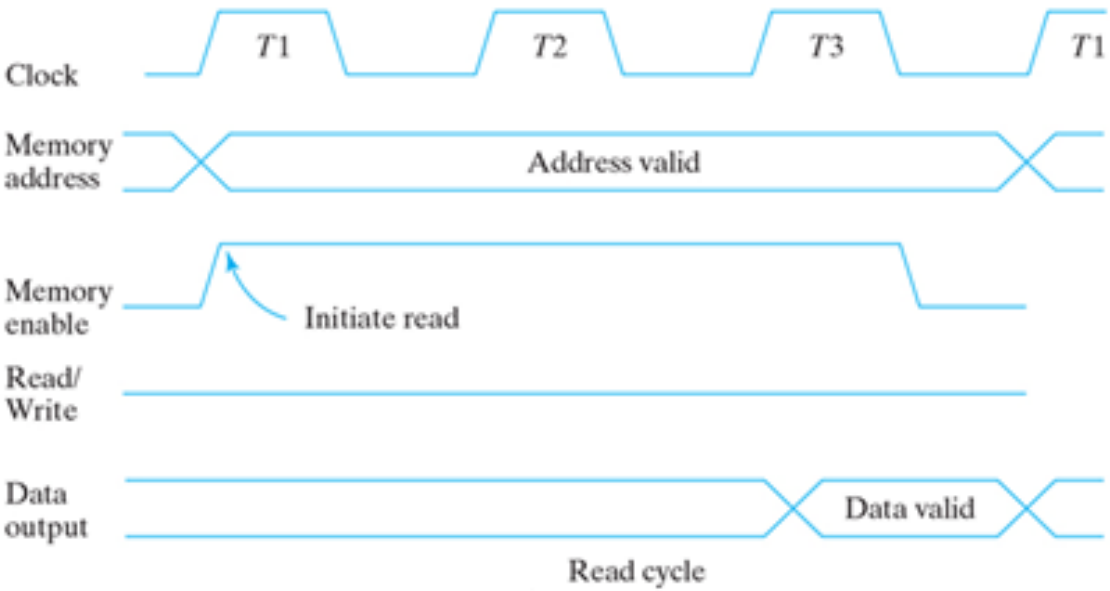
Memory Write Cycle



(a) Write cycle

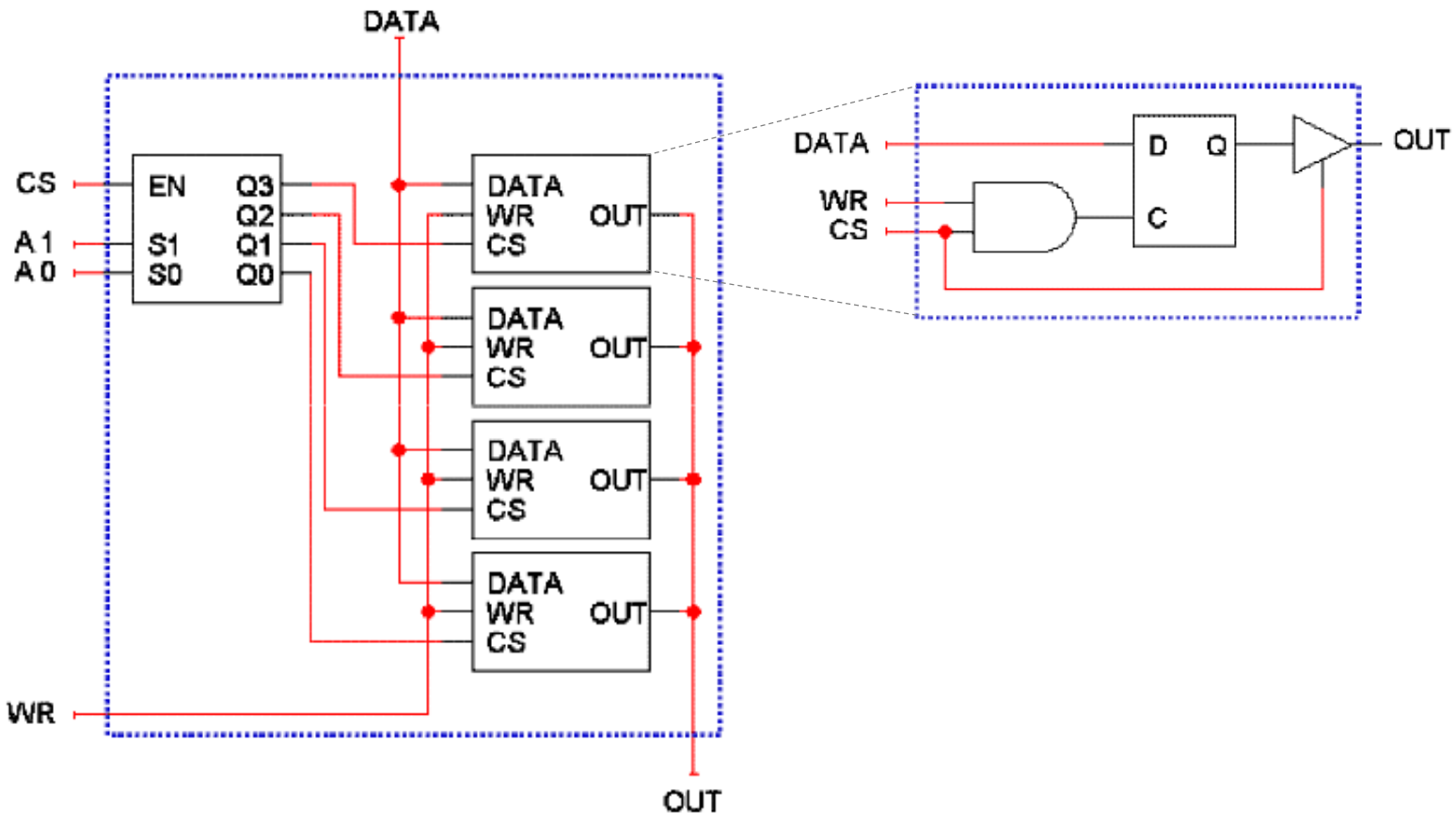
Memory Enable	Read/Write	Memory Operation
0	X	None
1	0	Write to selected word
1	1	Read from selected word

Memory Read Cycle

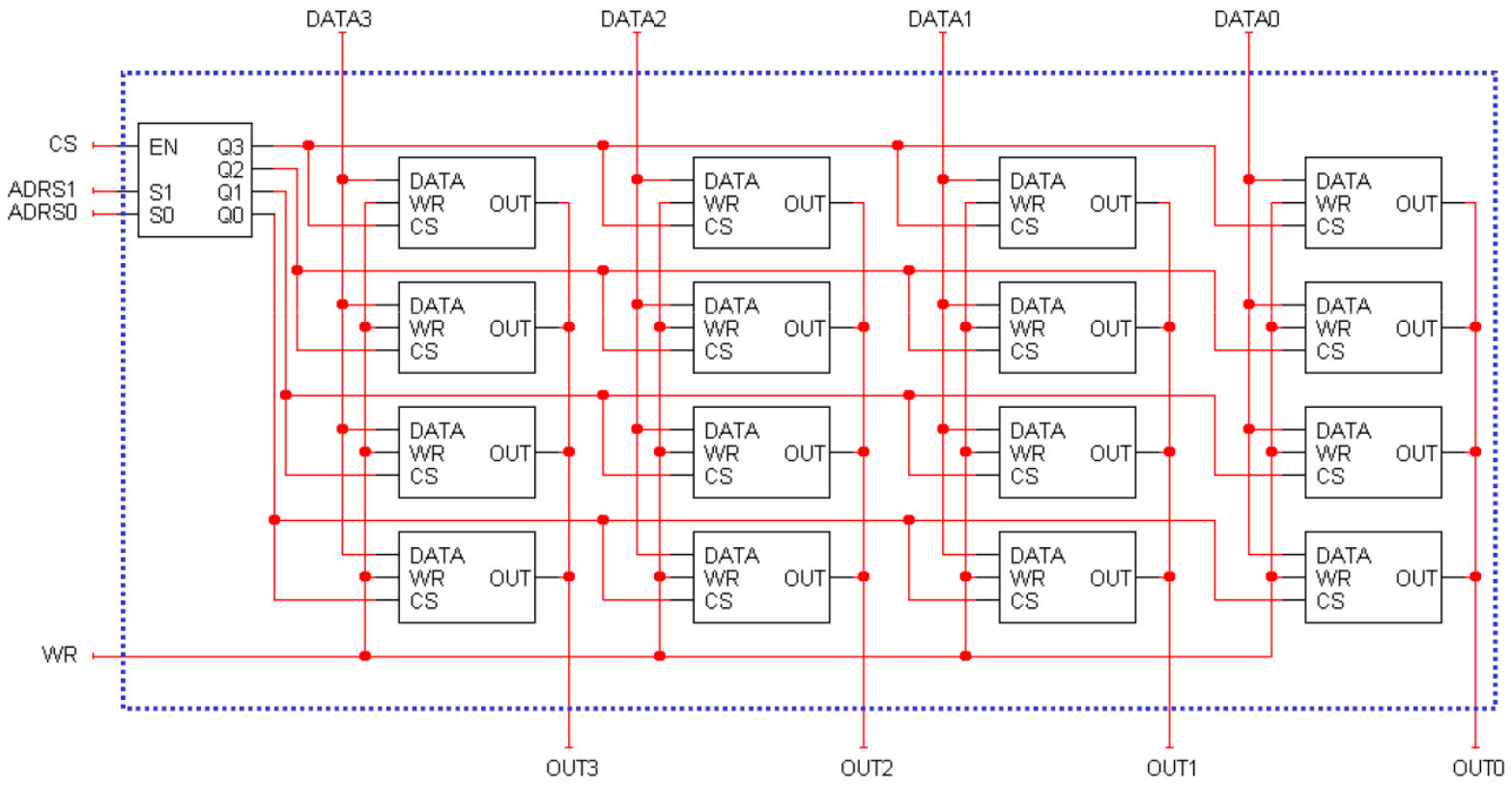


Memory Enable	Read/Write	Memory Operation
0	X	None
1	0	Write to selected word
1	1	Read from selected word

$2^2 \times 1$ bits Memory

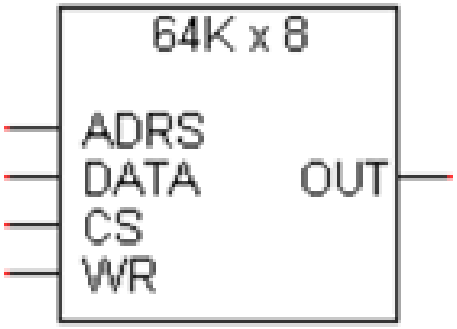


$2^2 \times 4$ bits Memory



Building Memory Blocks

- We have a 64K×8 bit Memory Chip

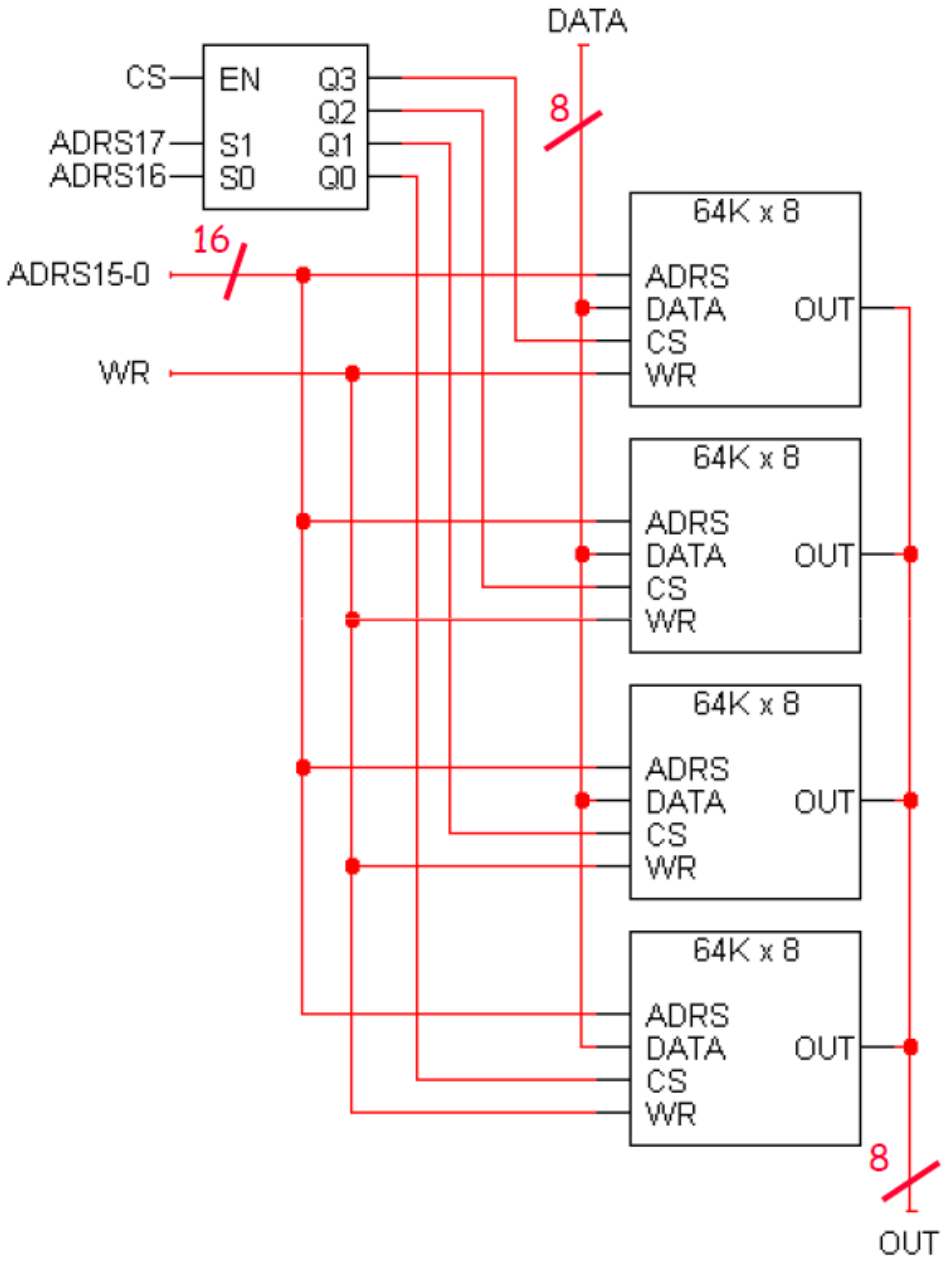


- We want to build:
  - a 256K×8 bit memory block
  - a 64K×16 bit memory block

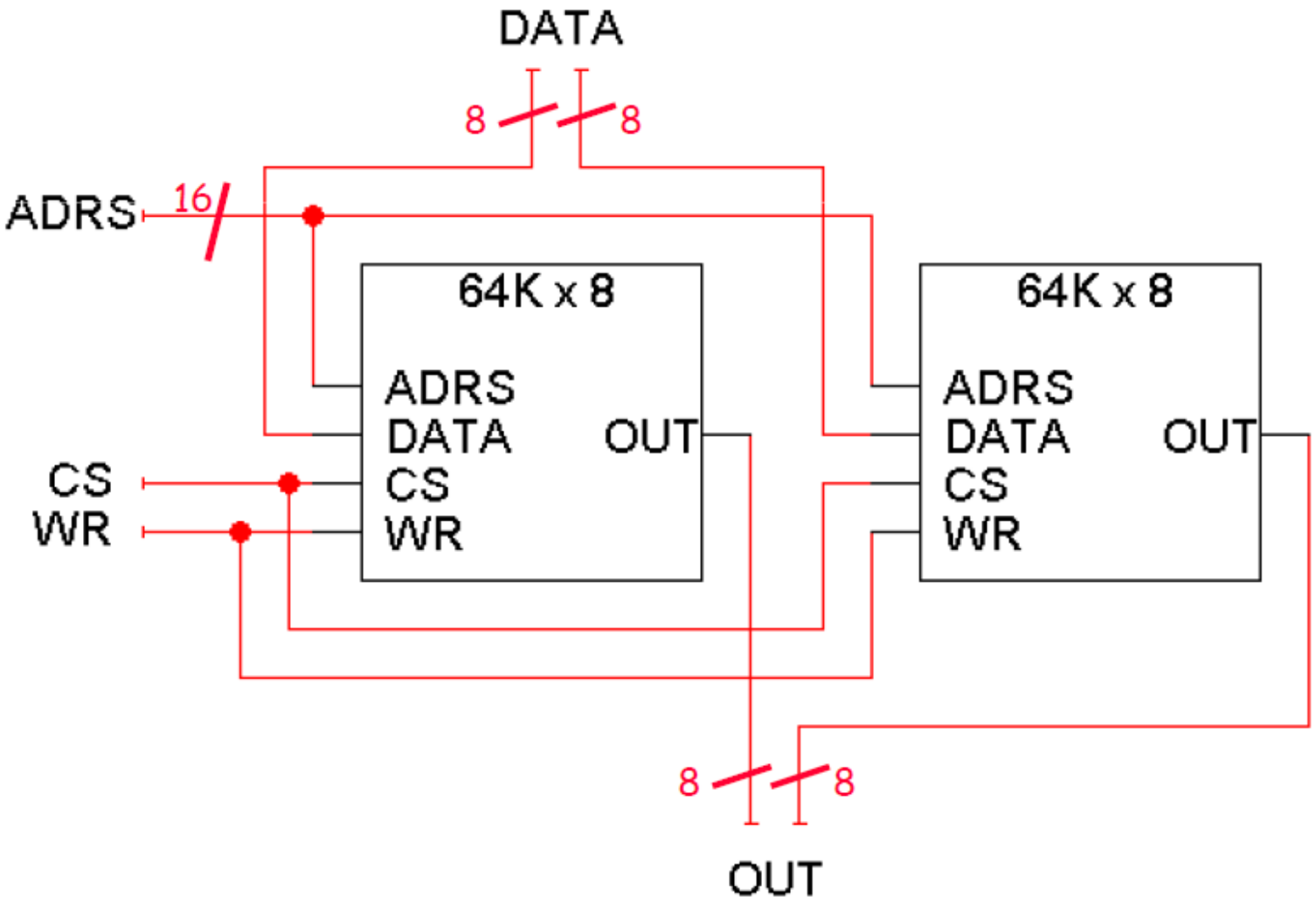
256K×8 bits

Address Range

11 1111 1111 1111 1111	(0x3ffff)
to	
11 0000 0000 0000 0000	(0x30000)
10 1111 1111 1111 1111	(0x2ffff)
to	
10 0000 0000 0000 0000	(0x20000)
01 1111 1111 1111 1111	(0x1ffff)
to	
01 0000 0000 0000 0000	(0x10000)
00 1111 1111 1111 1111	(0x0ffff)
to	
00 0000 0000 0000 0000	(0x00000)



64K×16 bits


DATA
8
8
ADRS
16
64K x 8
ADRS
DATA
CS
WR
OUT
64K x 8
ADRS
DATA
CS
WR
OUT
CS
WR
8
8
OUT

Exploiting Memory Hierarchy

Ideally one would desire an indefinitely large memory capacity such that any particular ... word would be immediately available. ... We are ... forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.

A. W. Burks, H. H. Goldstine, and J. von Neumann

Preliminary Discussion of the Logical Design of an Electronic Computing Instrument, 1946

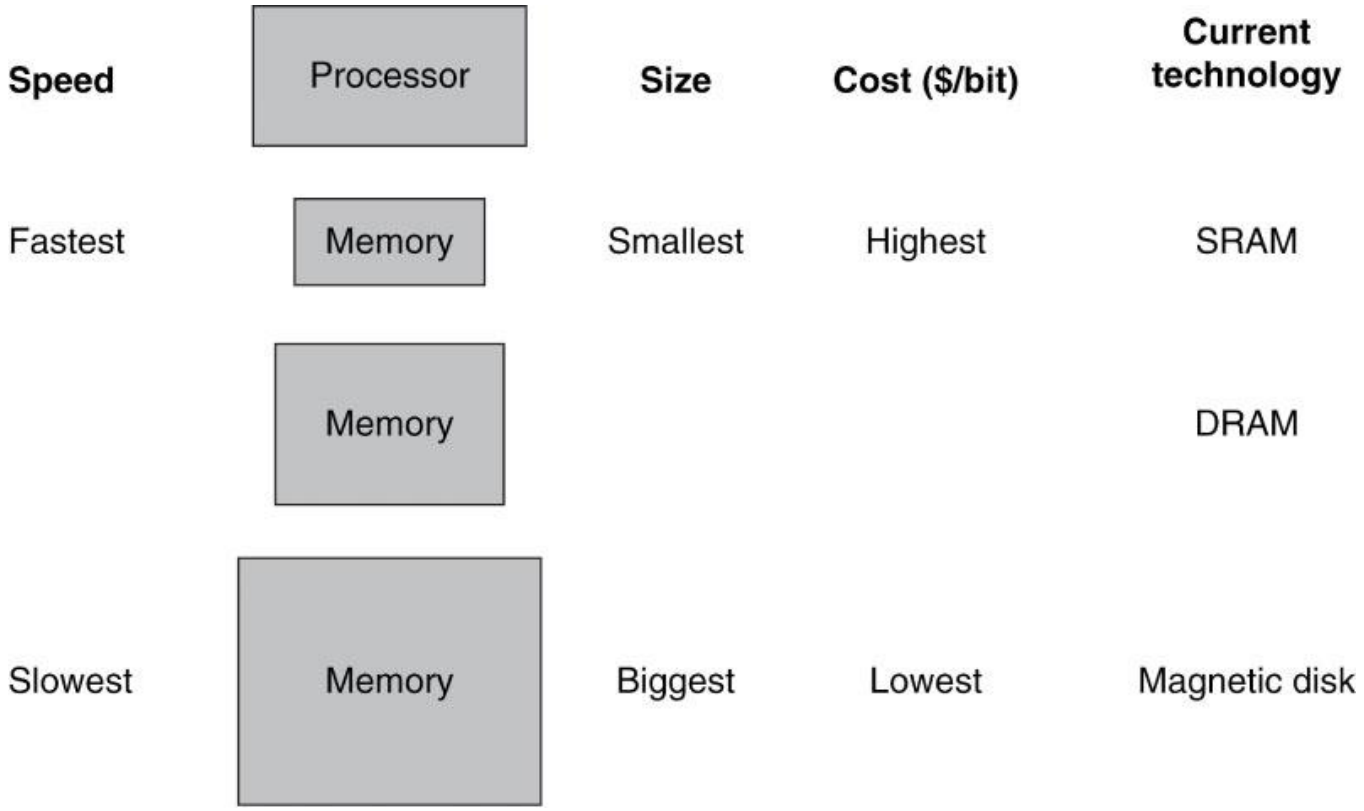
Principle of Locality

- Programs access a small proportion of their address space at any time
- Temporal locality
  - Items accessed recently are likely to be accessed again soon
  - e.g., instructions in a loop, induction variables
- Spatial locality
  - Items near those accessed recently are likely to be accessed soon
  - e.g., sequential instruction access, array data


Western Carolina

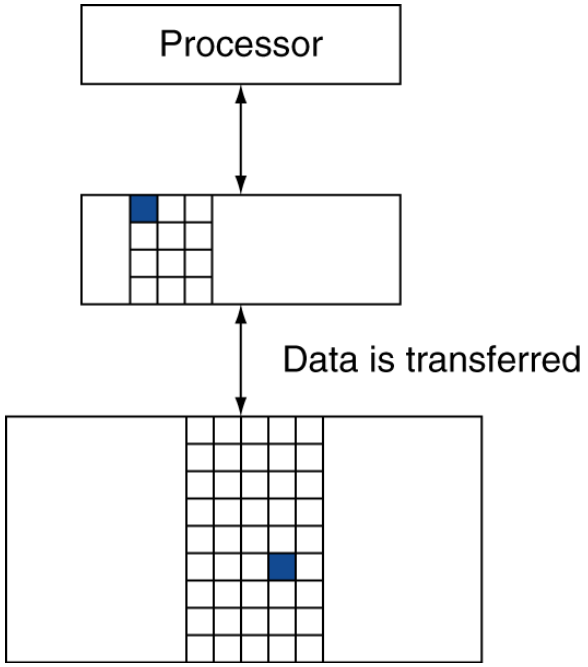
Taking Advantage of Locality

- Memory hierarchy
- Store everything on disk
- Copy recently accessed (and nearby) items from disk to smaller DRAM memory
  - Main memory
- Copy more recently accessed (and nearby) items from DRAM to smaller SRAM memory
  - Cache memory attached to CPU



The basic structure of a memory hierarchy. By implementing the memory system as a hierarchy, the user has the illusion of a memory that is as large as the largest level of the hierarchy, but can be accessed as if it were all built from the fastest memory. Flash memory has replaced disks in many personal mobile devices, and may lead to a new level in the storage hierarchy for desktop and server computers.

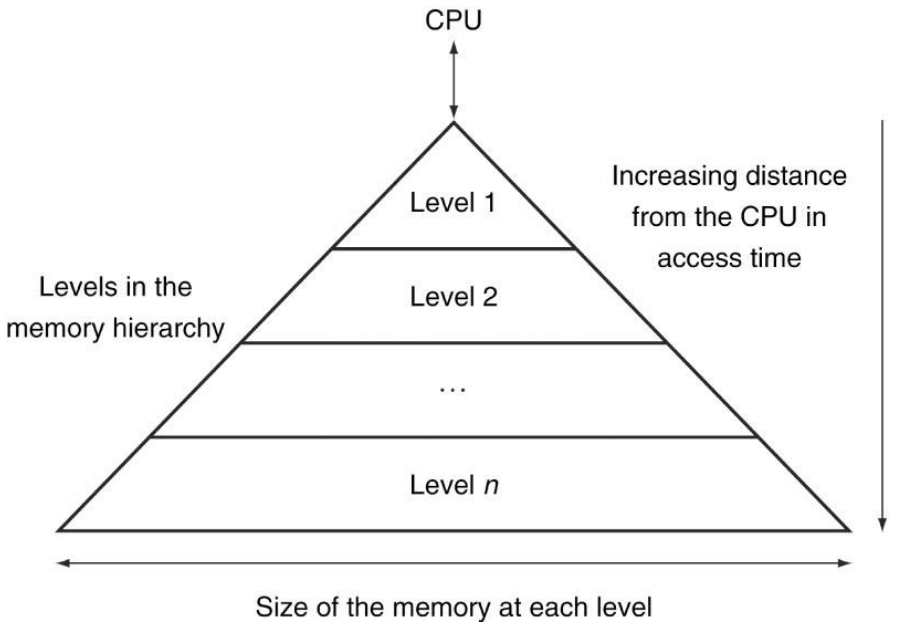
Memory Hierarchy Levels



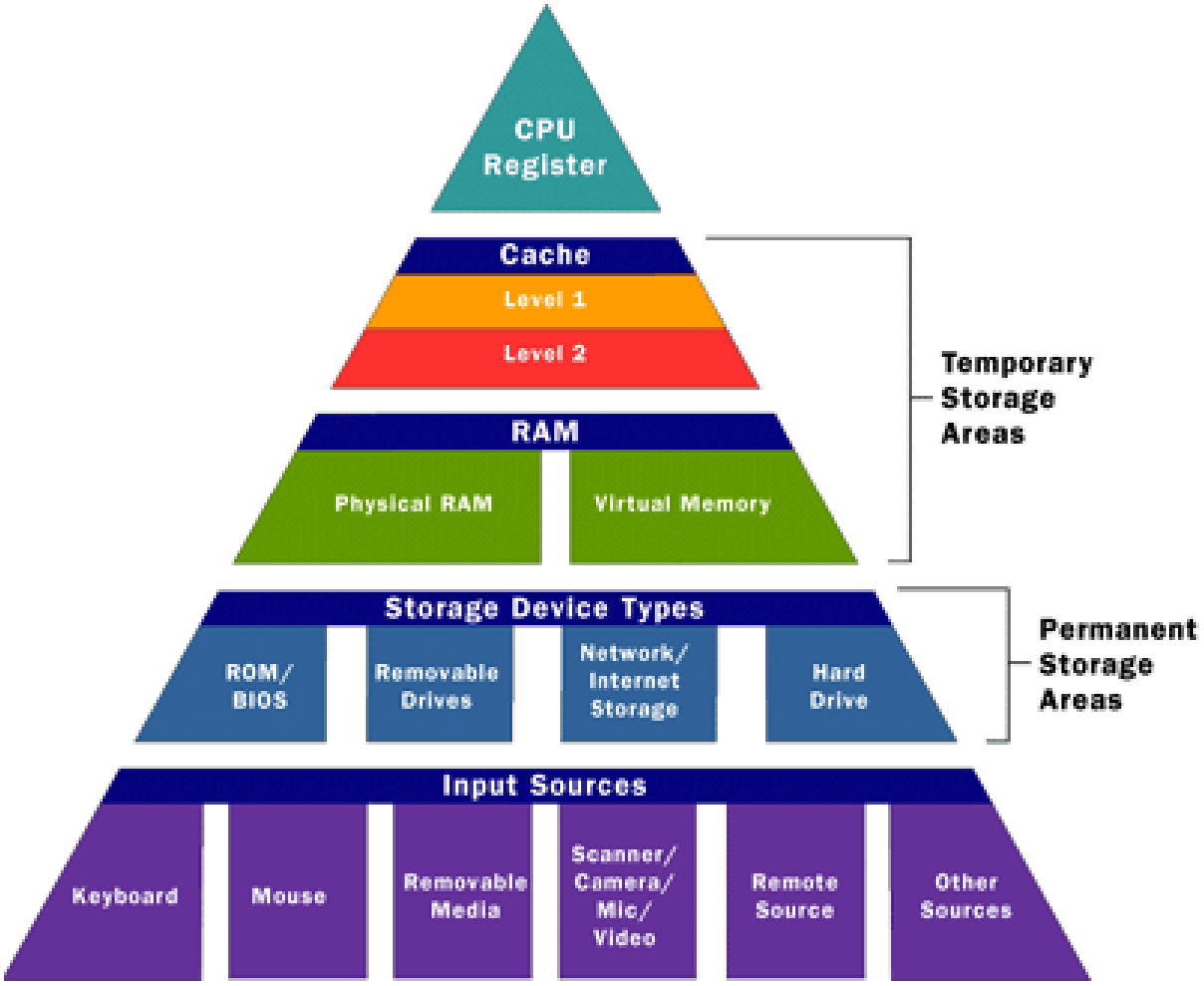
- Block (aka line): unit of copying
  - May be multiple words
- If accessed data is present in upper level
  - Hit: access satisfied by upper level
    - Hit ratio: hits/accesses
- If accessed data is absent
  - Miss: block copied from lower level
    - Time taken: Miss penalty
    - Miss ratio: misses/accesses
      = 1 – hit ratio
  - Then accessed data supplied from upper level

Structure of a Memory Hierarchy

- Fast memories are small, large memories are slow
  - We really want fast, large memories ☹
  - Memory hierarchy gives this illusion ☺



All-in-One


CPU
Register
Cache
Level 1
Level 2
RAM
Physical RAM
Virtual Memory
Temporary Storage Areas
Storage Device Types
ROM/ BIOS
Removable Drives
Network/ Internet Storage
Hard Drive
Permanent Storage Areas
Input Sources
Keyboard
Mouse
Removable Media
Scanner/ Camera/ Mic/ Video
Remote Source
Other Sources