



Memory

- + Memory Organization, hierarchy, RAM, ROM, Memory Address map, Cache, Memory Management

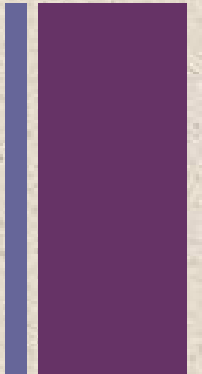
Key Characteristics of Computer Memory Systems

Location <ul style="list-style-type: none">Internal (e.g. processor registers, cache, main memory)External (e.g. optical disks, magnetic disks, tapes) Capacity <ul style="list-style-type: none">Number of wordsNumber of bytes Unit of Transfer <ul style="list-style-type: none">WordBlock Access Method <ul style="list-style-type: none">SequentialDirectRandomAssociative	Performance <ul style="list-style-type: none">Access timeCycle timeTransfer rate Physical Type <ul style="list-style-type: none">SemiconductorMagneticOpticalMagneto-optical Physical Characteristics <ul style="list-style-type: none">Volatile/nonvolatileErasable/nonerasable Organization <ul style="list-style-type: none">Memory modules
---	--

Table 4.1 Key Characteristics of Computer Memory Systems



Characteristics of Memory Systems



■ Location

- Refers to whether memory is internal and external to the computer
- Internal memory is often equated with main memory
- Processor requires its own local memory, in the form of registers
- Cache is another form of internal memory
- External memory consists of peripheral storage devices that are accessible to the processor via I/O controllers

■ Capacity

- Memory is typically expressed in terms of bytes (1 byte=8 bits or words)
- Common word lengths are 8, 16, and 32 bits.

■ Unit of transfer

- For internal memory the unit of transfer is equal to the number of electrical lines into and out of the memory module
- This may be equal to the word length, but is often larger, such as 64, 128, or 256 bits.

Method of Accessing Units of Data

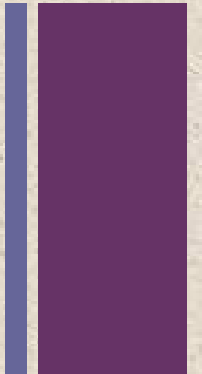


Capacity and Performance:





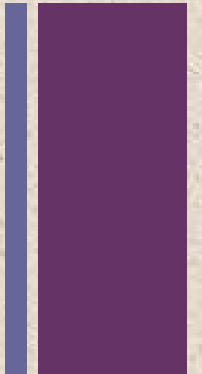
Memory



- The most common forms are:
 - Semiconductor memory
 - Magnetic surface memory
 - Optical
 - Magneto-optical
- Several physical characteristics of data storage are important:
 - Volatile memory
 - Information decays naturally or is lost when electrical power is switched off
 - Nonvolatile memory
 - Once recorded, information remains without deterioration until deliberately changed
 - No electrical power is needed to retain information
 - Magnetic-surface memories
 - Are nonvolatile
 - Semiconductor memory
 - May be either volatile or nonvolatile
 - Nonerasable memory
 - Cannot be altered, except by destroying the storage unit
 - Semiconductor memory of this type is known as read-only memory (ROM)
- For random-access memory the organization is a key design issue
 - Organization refers to the physical arrangement of bits to form words

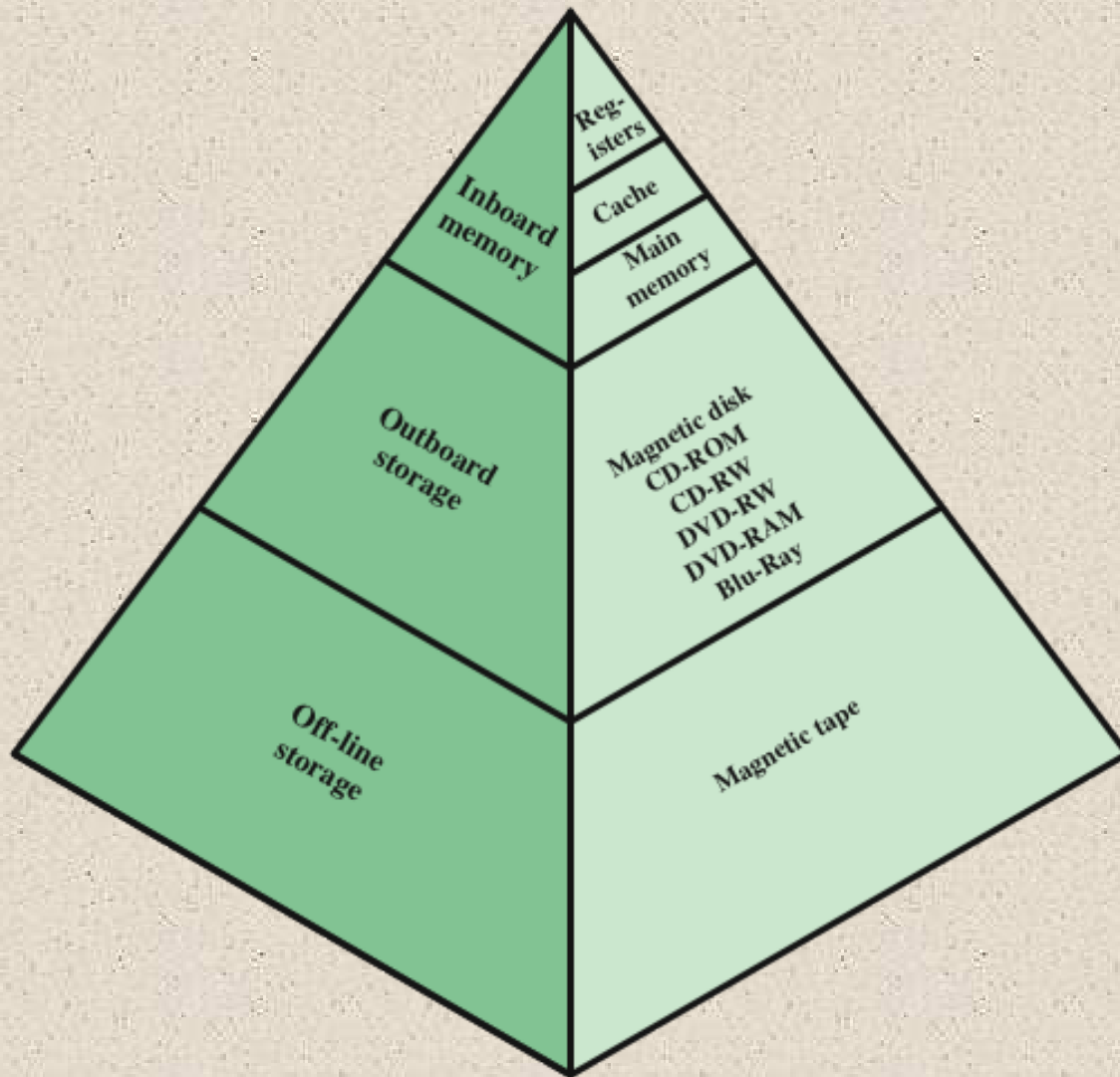


Memory Hierarchy



- Design constraints on a computer's memory can be summed up by three questions:
 - How much, how fast, how expensive
- There is a trade-off among capacity, access time, and cost
 - Faster access time, greater cost per bit
 - Greater capacity, smaller cost per bit
 - Greater capacity, slower access time
- The way out of the memory dilemma is not to rely on a single memory component or technology, but to employ a memory hierarchy

+ Memory Hierarchy - Diagram



As one goes down the hierarchy, the following occur:

- a. Decreasing cost per bit
- b. Increasing capacity
- c. Increasing access time
- d. Decreasing frequency of access of the memory by the processor

Figure 4.1 The Memory Hierarchy



+ Internal Memory

ROM, DRAM, and SRAM

+ Memory Cell Operation

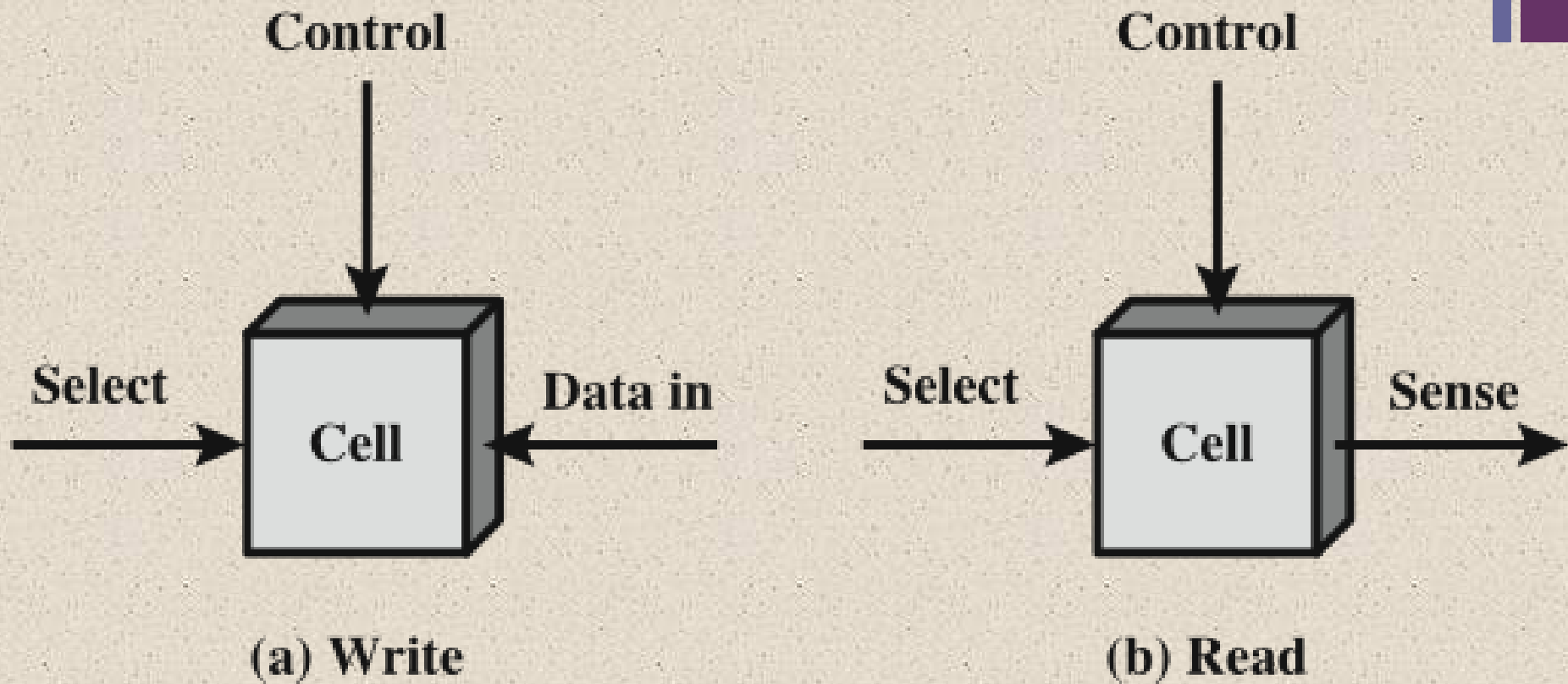


Figure 5.1 Memory Cell Operation

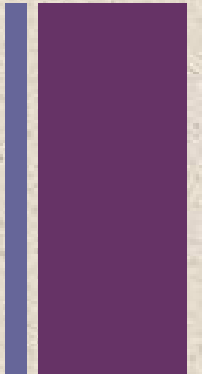
Semiconductor Memory Types

Memory Type	Category	Erasure	Write Mechanism	Volatility	
Random-access memory (RAM)	Read-write memory	Electrically, byte-level	Electrically	Volatile	
Read-only memory (ROM)	Read-only memory	Not possible	Masks	Nonvolatile	
Programmable ROM (PROM)			Electrically		
Erasable PROM (EPROM)		UV light, chip-level			
Electrically Erasable PROM (EEPROM)		Electrically, byte-level			
Flash memory		Electrically, block-level			

Table 5.1 Semiconductor Memory Types



Dynamic RAM (DRAM)

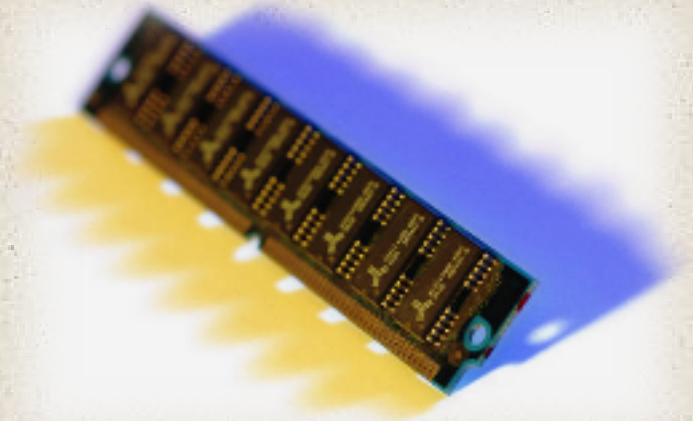


- RAM technology is divided into two technologies:
 - Dynamic RAM (DRAM)
 - Static RAM (SRAM)
- DRAM
 - Made with cells that store data as charge on capacitors
 - Presence or absence of charge in a capacitor is interpreted as a binary 1 or 0
 - Requires periodic charge refreshing to maintain data storage
 - The term *dynamic* refers to tendency of the stored charge to leak away, even with power continuously applied



Static RAM (SRAM)

- Digital device that uses the same logic elements used in the processor
- Binary values are stored using traditional flip-flop logic gate configurations
- Will hold its data as long as power is supplied to it

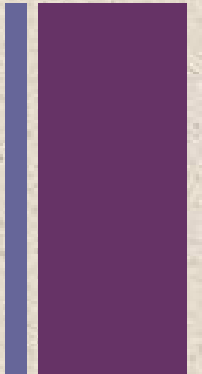


SRAM versus DRAM

- Both volatile
 - Power must be continuously supplied to the memory to preserve the bit values
- Dynamic cell
 - Simpler to build, smaller
 - More dense (smaller cells = more cells per unit area)
 - Less expensive
 - Requires the supporting refresh circuitry
 - Tend to be favored for large memory requirements
 - Used for main memory
- Static
 - Faster
 - Used for cache memory (both on and off chip)

SRAM

DRAM

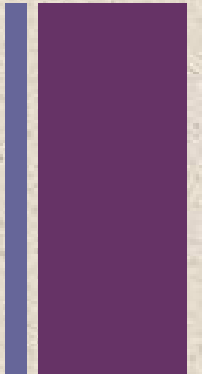


Read Only Memory (ROM)

- Contains a permanent pattern of data that cannot be changed or added to
- No power source is required to maintain the bit values in memory
- Data or program is permanently in main memory and never needs to be loaded from a secondary storage device
- Data is actually wired into the chip as part of the fabrication process
 - Disadvantages of this:
 - No room for error, if one bit is wrong the whole batch of ROMs must be thrown out
 - Data insertion step includes a relatively large fixed cost



Programmable ROM (PROM)



- Less expensive alternative
- Nonvolatile and may be written into only once
- Writing process is performed electrically and may be performed by supplier or customer at a time later than the original chip fabrication
- Special equipment is required for the writing process
- Provides flexibility and convenience
- Attractive for high volume production runs

Read-Mostly Memory

EPROM

Erasable programmable read-only memory

Erase process can be performed repeatedly

More expensive than PROM but it has the advantage of the multiple update capability

EEPROM

Electrically erasable programmable read-only memory

Can be written into at any time without erasing prior contents

Combines the advantage of non-volatility with the flexibility of being updatable in place

More expensive than EPROM

Flash Memory

Intermediate between EPROM and EEPROM in both cost and functionality

Uses an electrical erasing technology, does not provide byte-level erasure

Microchip is organized so that a section of memory cells are erased in a single action or "flash"

Typical 16 Mb DRAM (4M x 4)

RAS= Row Address Select

CAS= Column Address Select

WE= Write Enable

OE= Output Enable

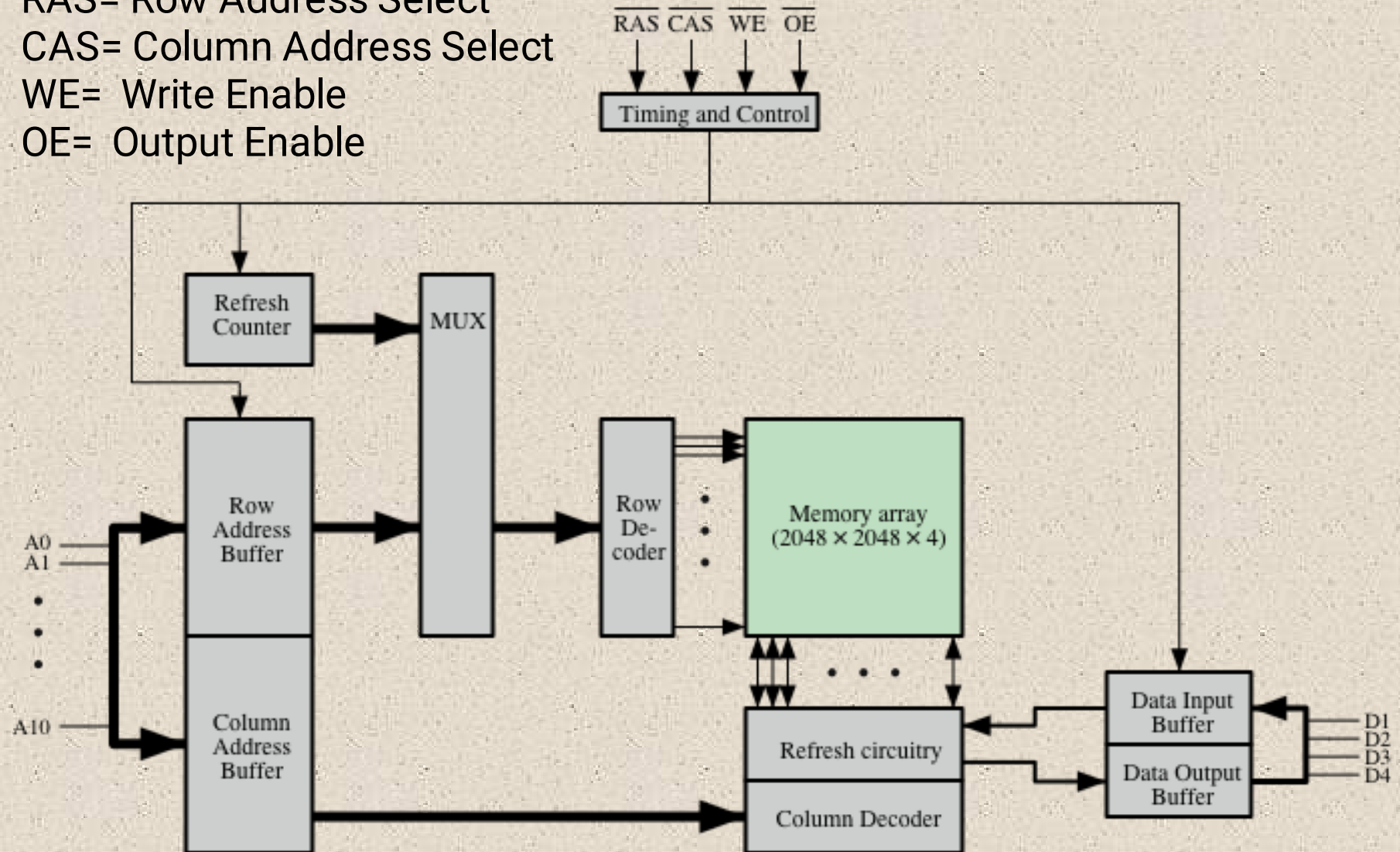
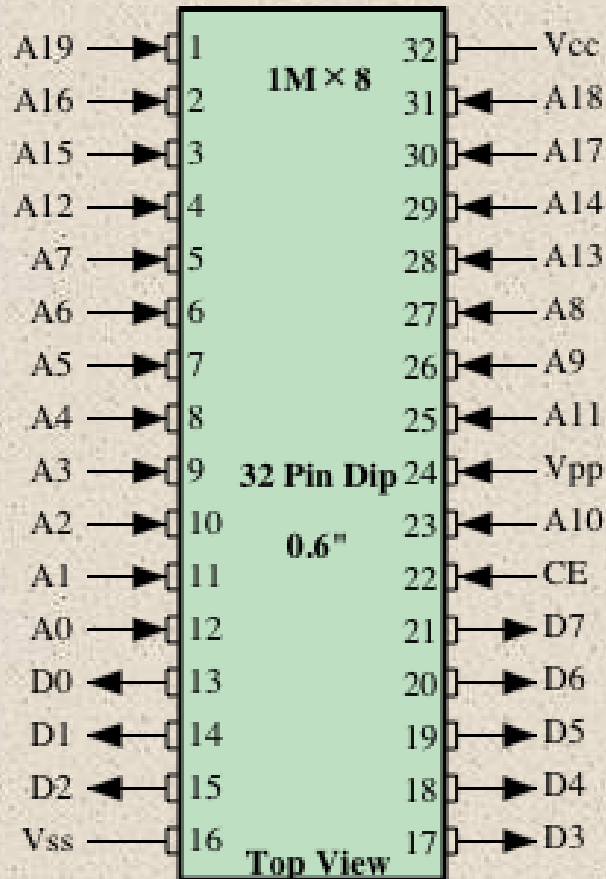
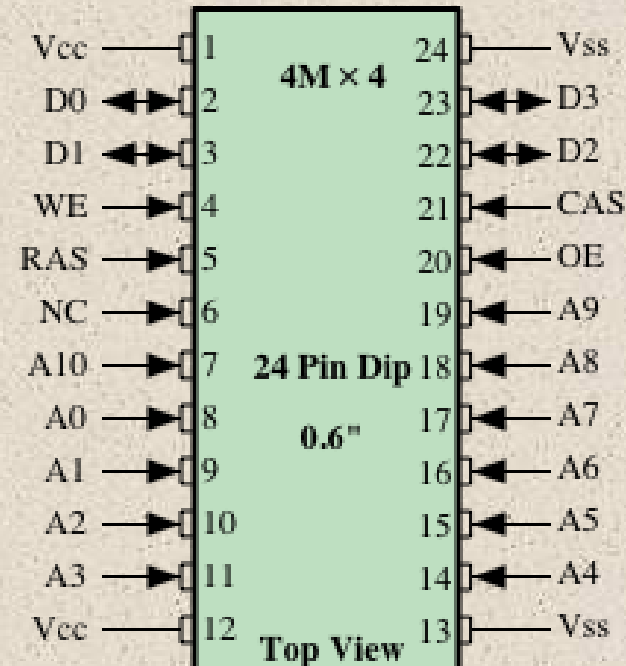


Figure 5.3 Typical 16 Megabit DRAM (4M x 4)

Chip Packaging



(a) 8 Mbit EPROM



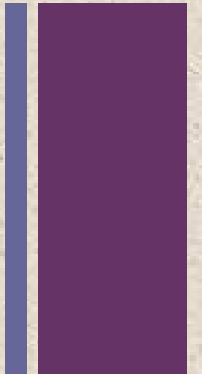
(b) 16 Mbit DRAM

The power supply to the chip (V_{cc}).

A ground pin (V_{ss}).

A program voltage (V_{pp}) that is supplied during programming (write operations).

Figure 5.4 Typical Memory Package Pins and Signals



Error Correction

■ Hard Failure

- Permanent physical defect
- Memory cell or cells affected cannot reliably store data but become stuck at 0 or 1 or switch erratically between 0 and 1
- Can be caused by:
 - Harsh environmental abuse
 - Manufacturing defects

■ Soft Error

- Random, non-destructive event that alters the contents of one or more memory cells
- No permanent damage to memory
- Can be caused by:
 - Power supply problems
 - Alpha particles due to radioactivity falloff which is common in all materials

Advanced DRAM Organization

Synchronous
SDRAM

Double Data
Rate
DDR-DRAM

RamBus
RDRAM

- One of the most critical system bottlenecks when using high-performance processors is the interface to main internal memory
- The traditional DRAM chip is constrained both by its internal architecture and by its interface to the processor's memory bus
- A number of enhancements to the basic DRAM architecture have been explored:

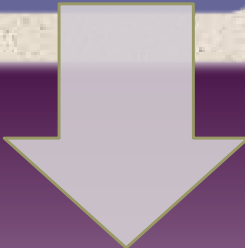
+

	Clock Frequency (MHz)	Transfer Rate (GB/s)	Access Time (ns)	Pin Count
SDRAM	166	1.3	18	168
DDR	200	3.2	12.5	184
RDRAM	600	4.8	12	162

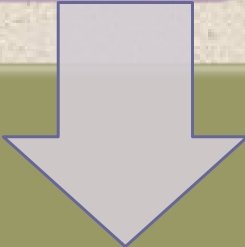
Table 5.3 Performance Comparison of Some DRAM Alternatives

Synchronous DRAM (SDRAM)

One of the most widely used forms of DRAM



Exchanges data with the processor synchronized to an external clock signal and running at the full speed of the processor/memory bus without imposing wait states



With synchronous access the DRAM moves data in and out under control of the system clock

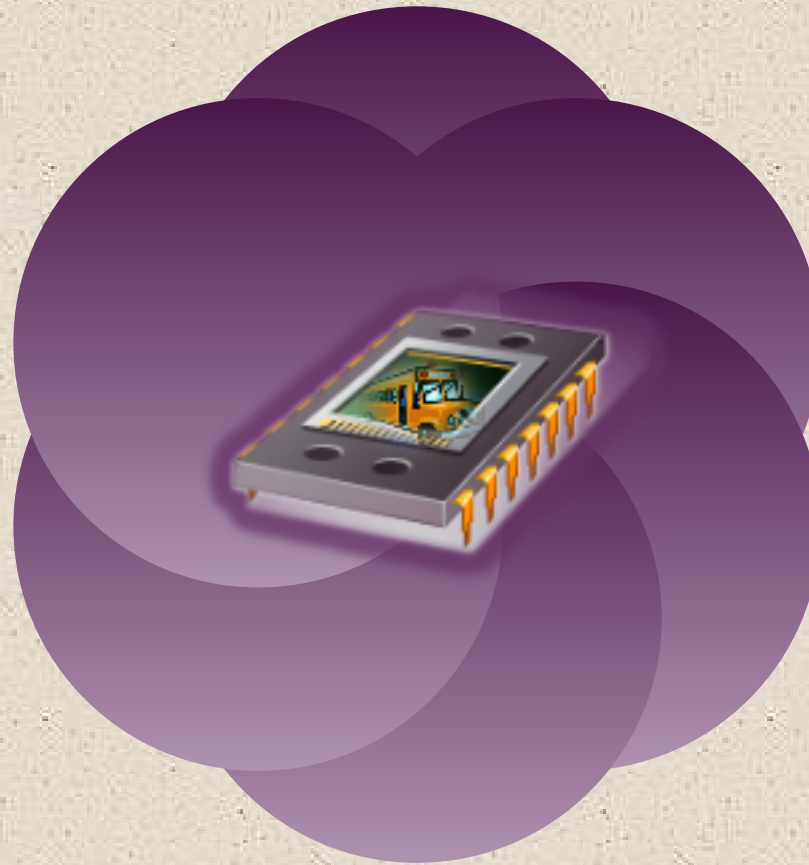
- The processor or other master issues the instruction and address information which is latched by the DRAM
- The DRAM then responds after a set number of clock cycles
- Meanwhile the master can safely do other tasks while the SDRAM is processing

RDRAM

Developed by Rambus

Bus delivers address and control information using an asynchronous block-oriented protocol

- Gets a memory request over the high-speed bus
- Request contains the desired address, the type of operation, and the number of bytes in the operation



Adopted by Intel for its Pentium and Itanium processors

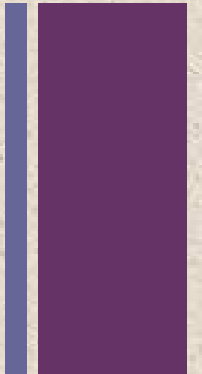
Has become the main competitor to SDRAM

Chips are vertical packages with all pins on one side

- Exchanges data with the processor over 28 wires no more than 12 centimeters long



Double Data Rate SDRAM (DDR SDRAM)

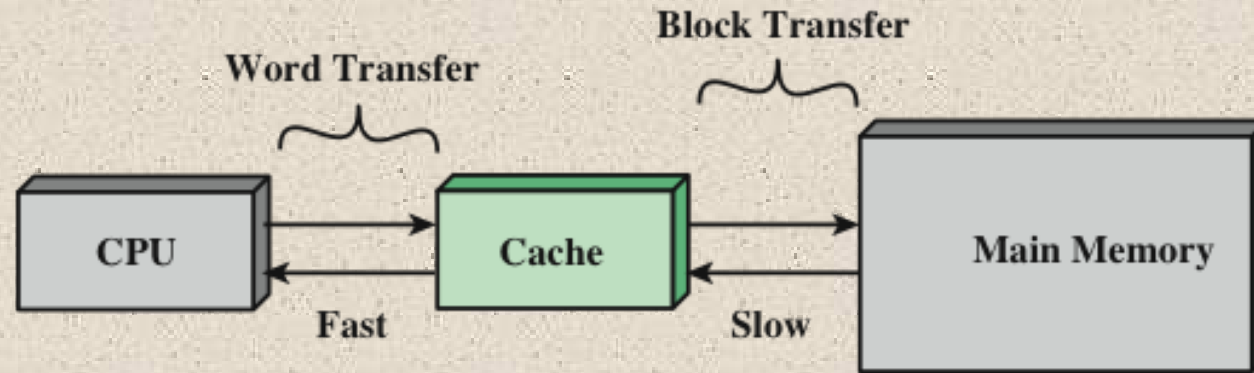


- SDRAM can only send data once per bus clock cycle
- Double-data-rate SDRAM can send data twice per clock cycle, once on the rising edge of the clock pulse and once on the falling edge

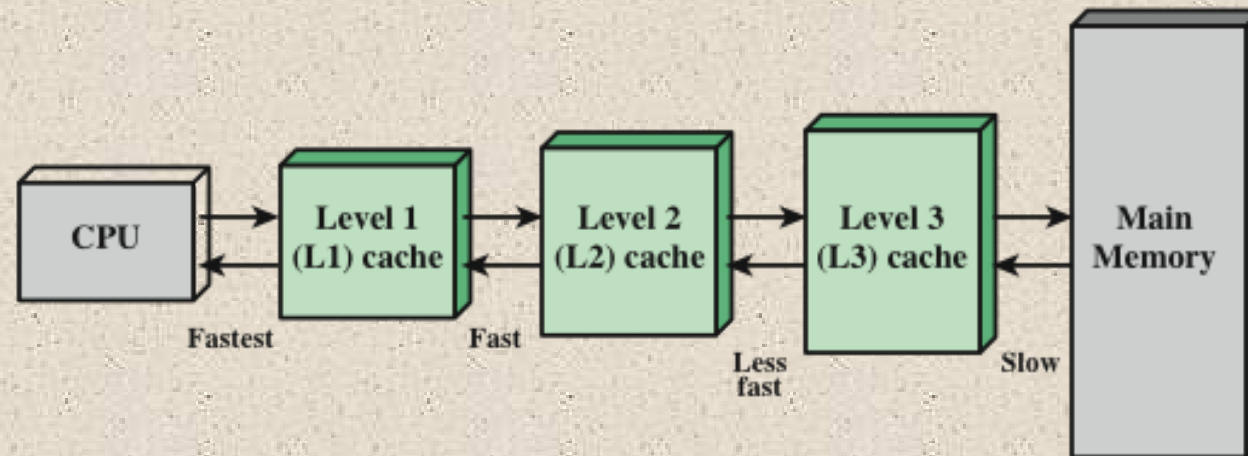
Cache Memory

Cache memory is designed to combine the memory access time of expensive, high-speed memory combined with the large memory size of less expensive, lower-speed memory.

Cache and Main Memory



(a) Single cache



(b) Three-level cache organization

Figure 4.3 Cache and Main Memory

Cache and Main Memory

Suppose that the processor has access to two levels of memory. Level 1 contains 1000 words and has an access time of 0.01 μ s; level 2 contains 100,000 words and has an access time of 0.1 μ s. Assume that if a word to be accessed is in level 1, then the processor accesses it directly. If it is in level 2, then the word is first transferred to level 1 and then accessed by the processor. For simplicity, we ignore the time required for the processor to determine whether the word is in level 1 or level 2.

Suppose 95% of the memory accesses are found in the cache. Then the average time to access a word can be expressed as

$$(0.95)(0.01 \mu\text{s}) + (0.05)(0.01 \mu\text{s} + 0.1 \mu\text{s}) = 0.0095 + 0.0055 = 0.015 \mu\text{s}$$

The average access time is much closer to 0.01 μ s than to 0.1 μ s, as desired.

Cache/Main Memory Structure

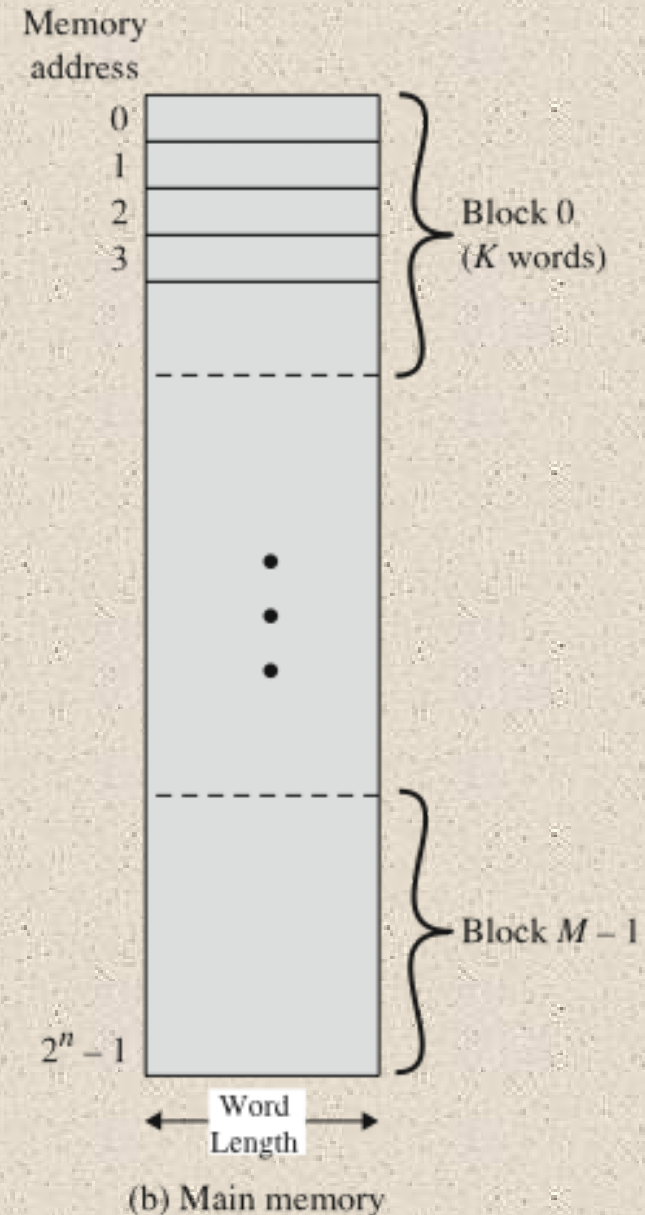
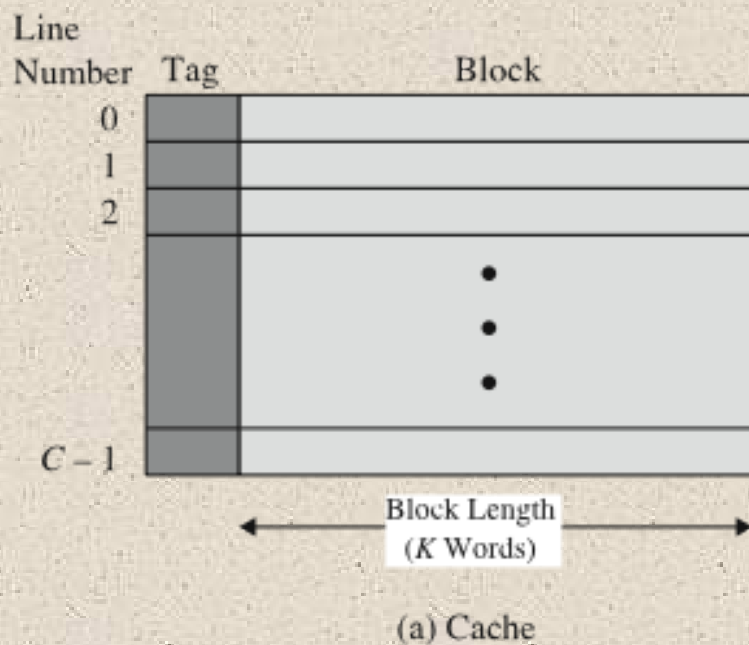


Figure 4.4 Cache/Main-Memory Structure

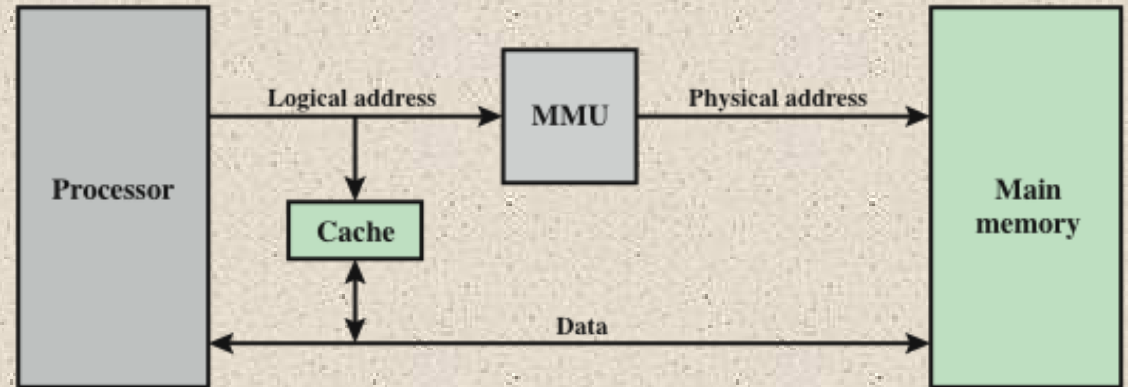
Elements of Cache Design

Cache Addresses	Write Policy
Logical	Write through
Physical	Write back
Cache Size	Line Size
Mapping Function	Number of caches
Direct	Single or two level
Associative	Unified or split
Set Associative	
Replacement Algorithm	
Least recently used (LRU)	
First in first out (FIFO)	
Least frequently used (LFU)	
Random	

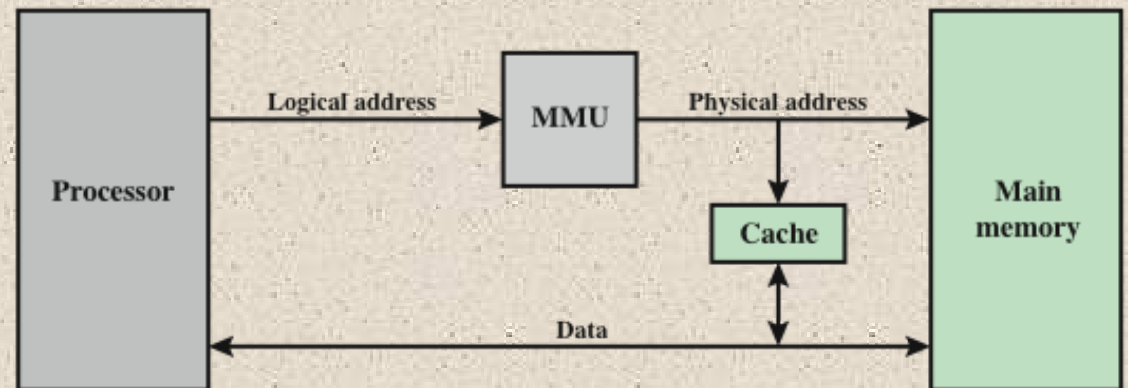
Table 4.2 Elements of Cache Design



Logical and Physical Caches



(a) Logical Cache



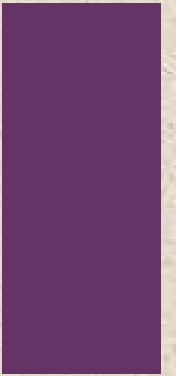
(b) Physical Cache

Figure 4.7 Logical and Physical Caches

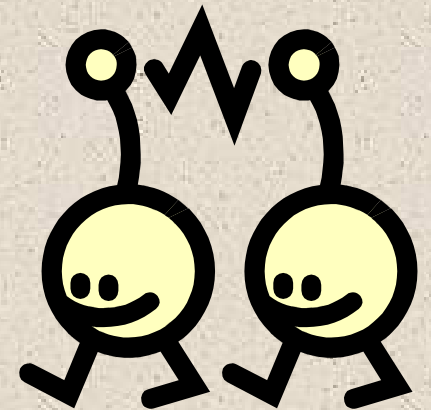


Cache Addresses

Virtual Memory

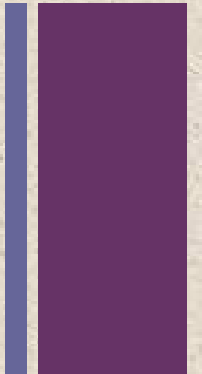


- Virtual memory
 - Facility that allows programs to address memory from a logical point of view, without regard to the amount of main memory physically available
 - When used, the address fields of machine instructions contain virtual addresses
 - For reads to and writes from main memory, a hardware memory management unit (MMU) translates each virtual address into a physical address in main memory



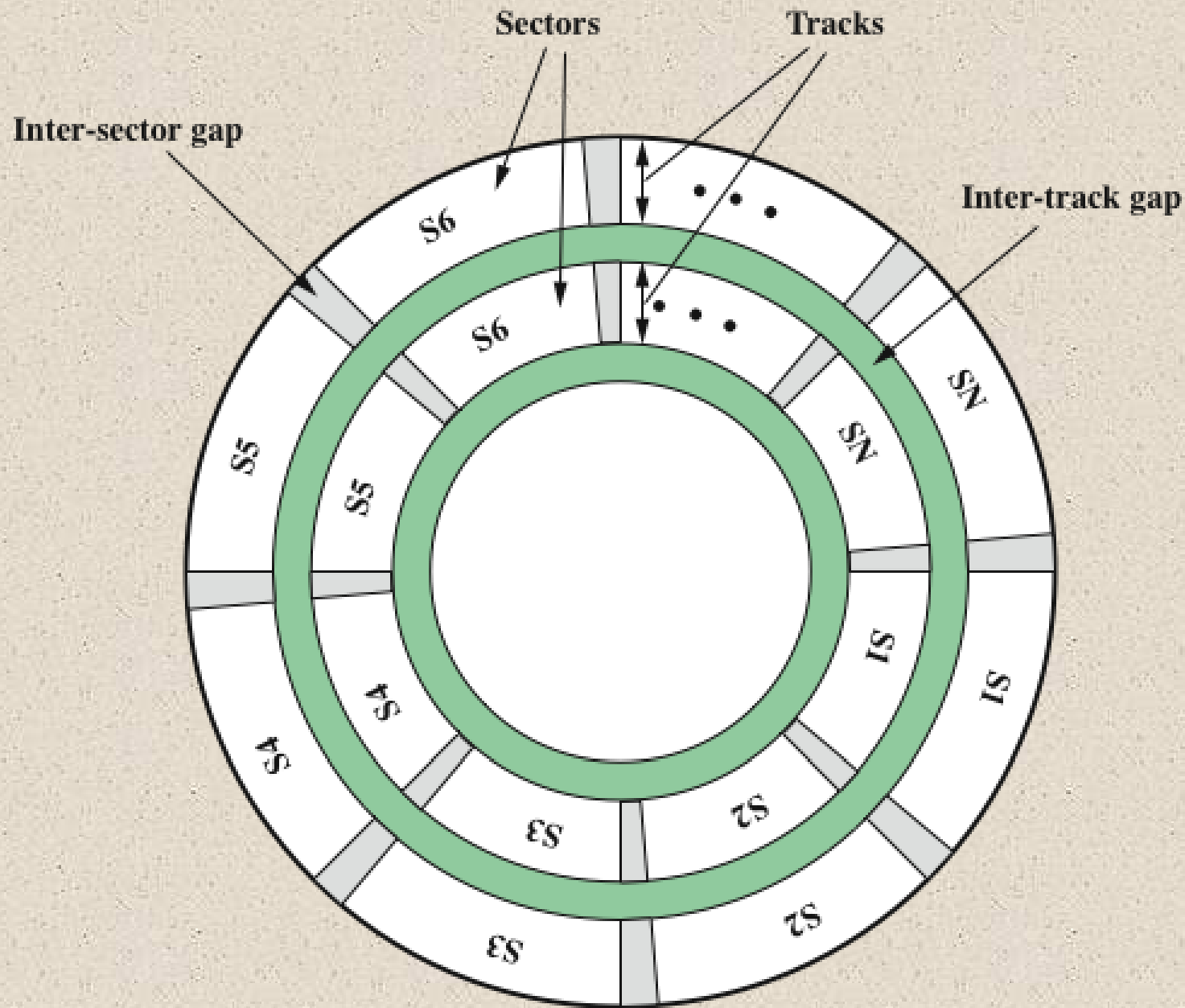


Magnetic Disk



- A disk is a circular *platter* constructed of nonmagnetic material, called the *substrate*, coated with a magnetizable material
 - Traditionally the substrate has been an aluminium or aluminium alloy material
 - Recently glass substrates have been introduced
- Benefits of the glass substrate:
 - Improvement in the uniformity of the magnetic film surface to increase disk reliability
 - A significant reduction in overall surface defects to help reduce read-write errors
 - Ability to support lower fly heights
 - Better stiffness to reduce disk dynamics
 - Greater ability to withstand shock and damage

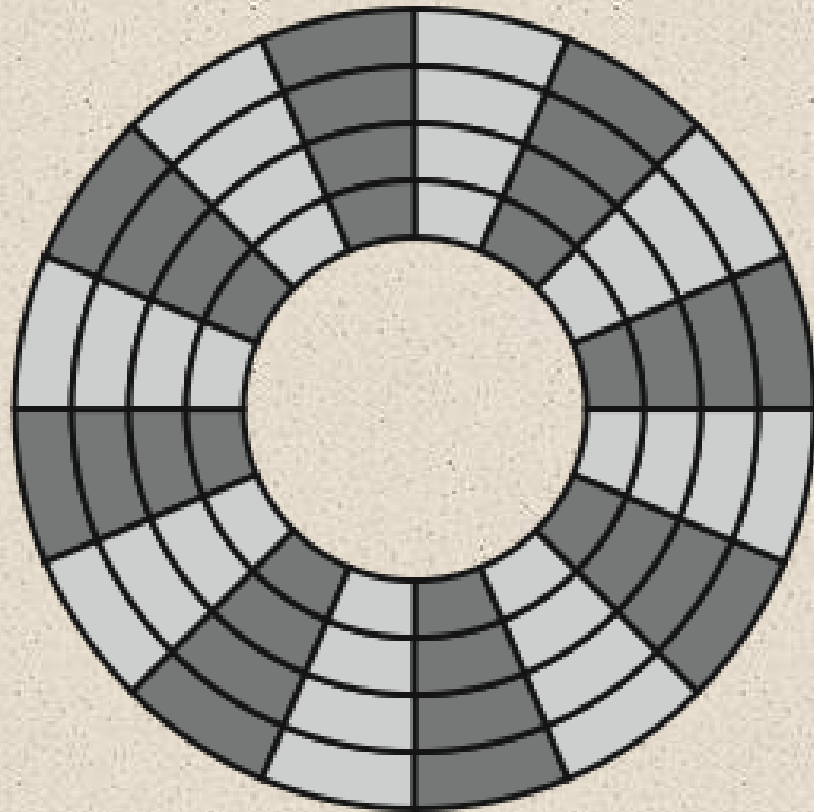




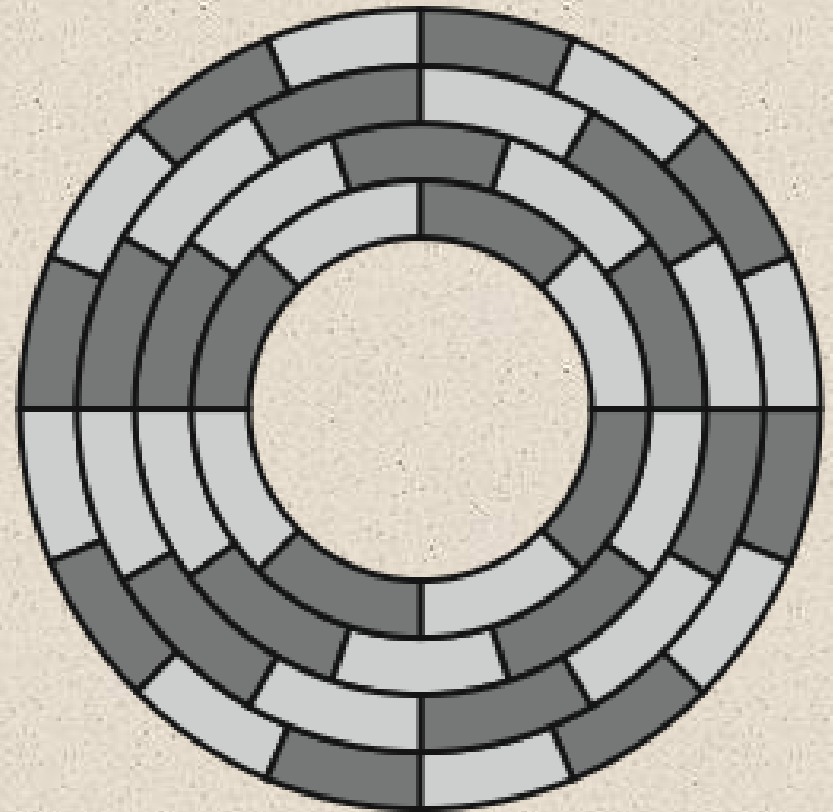
Disk Data Layout

Figure 6.2 Disk Data Layout

Disk Layout Methods Diagram



(a) Constant angular velocity



(b) Multiple zoned recording

Figure 6.3 Comparison of Disk Layout Methods