ACCRA TECHNICAL UNIVERSITY

FACULTY OF APPLIED SCIENCES

COMPUTER SCIENCE DEPARTMENT

Computer Organization and Architecture BCP 203

MEMORY

Presentation Content

1. Introduction
2. Cache Memory
   - Pentium 4 Cache Organization
   - ARM Cache Organization
3. Internal Memory
   - Semiconductor Memory
4. External Memory
   - Magnetic Memory
   - RAID
   - Solid State Drives
   - Optical Memory

1. Introduction

- A typical computer system is equipped with a hierarchy of memory subsystems
- A memory subsystem is considered ***internal*** to the computer if it is directly accessed by the processor
- It is considered ***external*** if it is accessed by the processor via an I/O module
- At the highest level (closest to the processor) are ***registers***
- Next comes one or more levels of ***cache***. If there are multiple levels they are denoted L1, L2, L3 and so on
- Next comes ***main memory***
- Next level is typically a ***fixed hard disk*** and one or more levels below that consisting of removable media

1.1 Characteristics of Computer Memory

Location
- Internal (e.g. processor registers, main memory, cache)
- 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

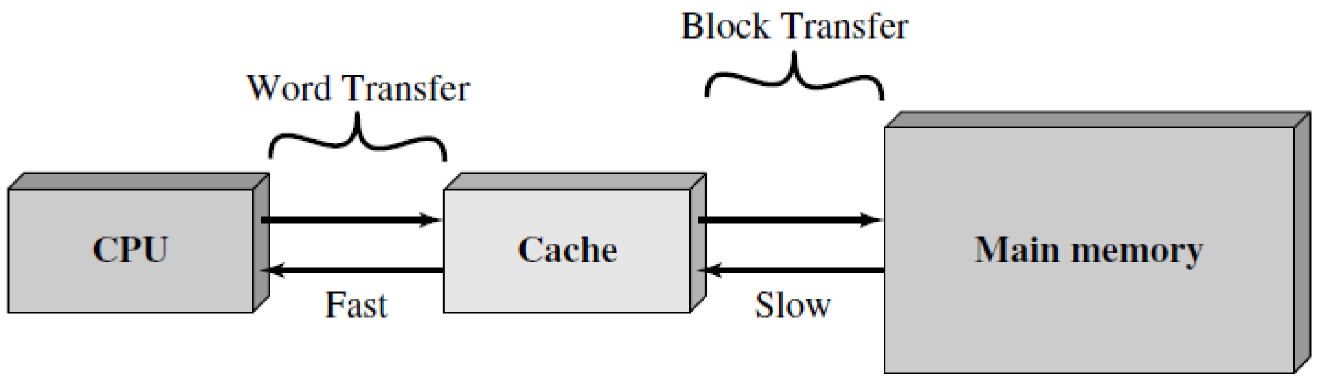
Performance
- Access time
- Cycle time
- Transfer rate

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

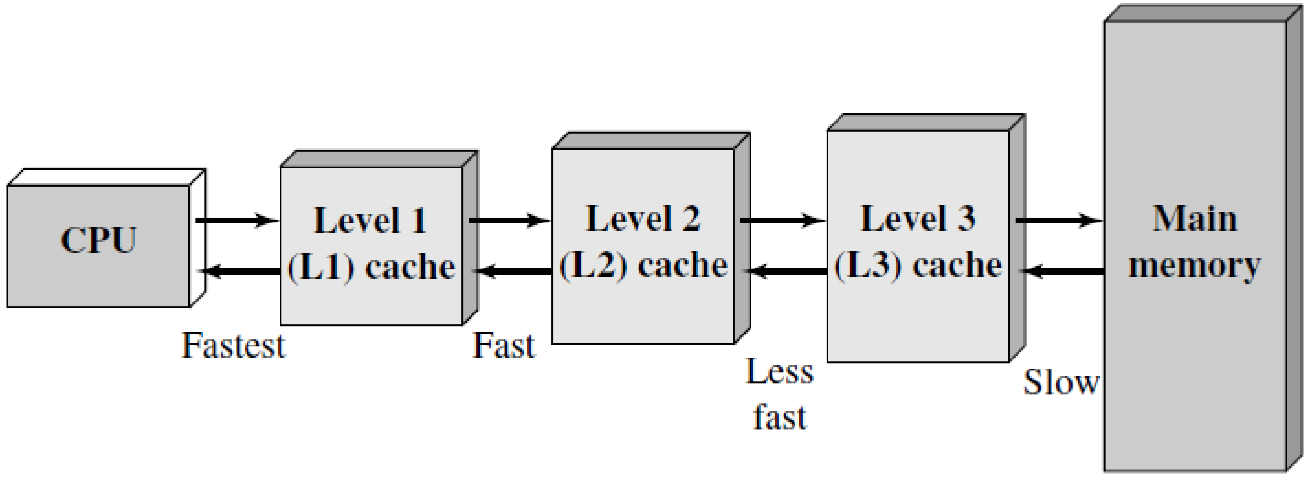
Physical Characteristics
- Volatile/nonvolatile
- Erasable/nonerasable

Organization
- Memory modules

2. Cache Memory

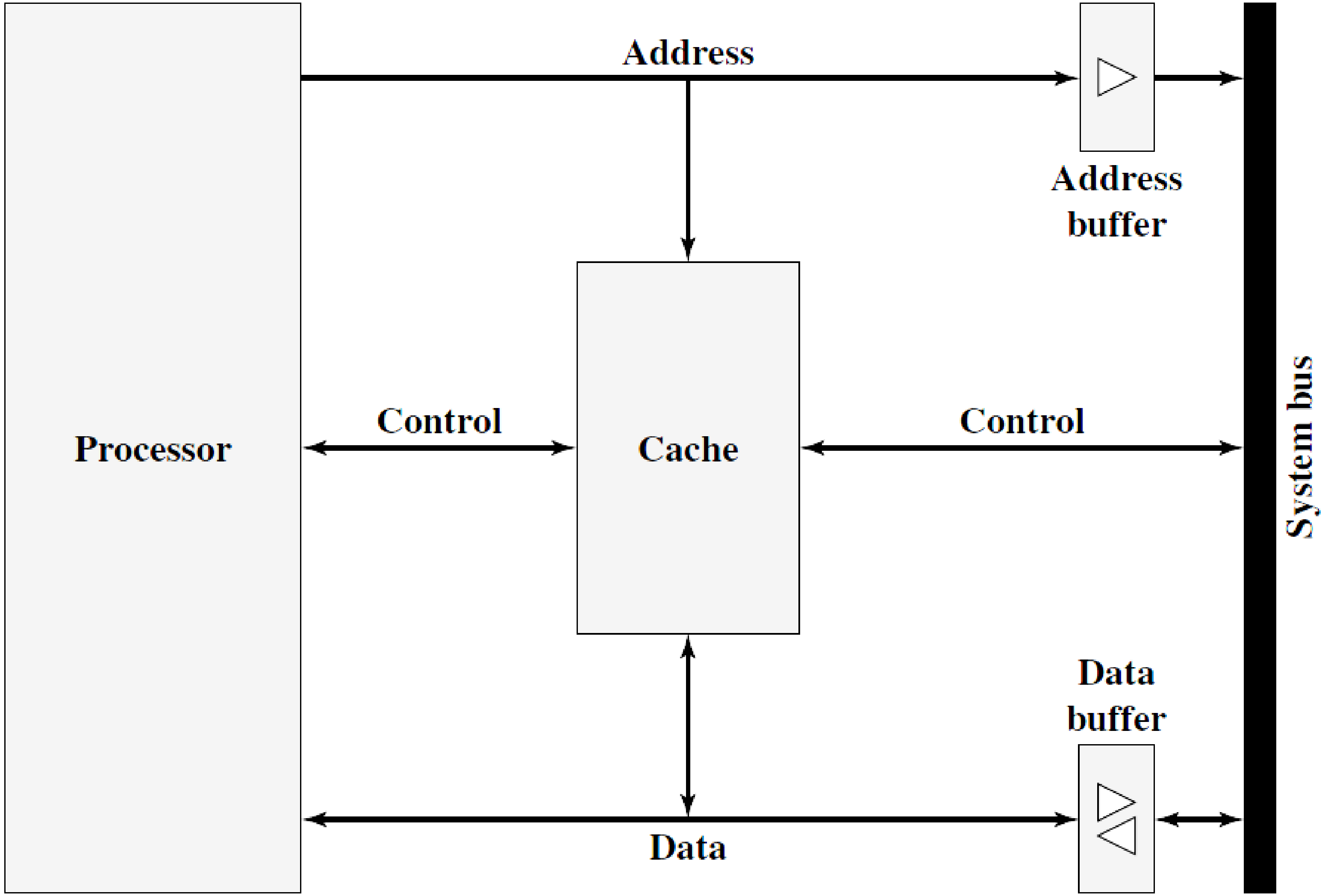


(a) Single cache



(b) Three-level cache organization

2.1 Typical Cache Organization



2.2 Elements of Cache Design

Cache Addresses
- Logical
- Physical

Cache Size

Mapping Function
- Direct
- Associative
- Set Associative

Replacement Algorithm
- Least recently used (LRU)
- First in first out (FIFO)
- Least frequently used (LFU)
- Random

Write Policy
- Write through
- Write back
- Write once

Line Size

Number of caches
- Single or two level
- Unified or split

2.3 Pentium 4 Cache Organization

- Evolution of Cache organization in intel microprocessors

Problem	Solution	Processor on which Feature First Appears
External memory slower than the system bus.	Add external cache using faster memory technology.	386
Increased processor speed results in external bus becoming a bottleneck for cache access.	Move external cache on-chip, operating at the same speed as the processor.	486
Internal cache is rather small, due to limited space on chip	Add external L2 cache using faster technology than main memory	486
Contention occurs when both the Instruction Prefetcher and the Execution Unit simultaneously require access to the cache. In that case, the Prefetcher is stalled while the Execution Unit's data access takes place.	Create separate data and instruction caches.	Pentium
Increased processor speed results in external bus becoming a bottleneck for L2 cache access.	Create separate back-side bus that runs at higher speed than the main (front-side) external bus. The BSB is dedicated to the L2 cache.	Pentium Pro
	Move L2 cache on to the processor chip.	Pentium II
Some applications deal with massive databases and must have rapid access to large amounts of data. The on-chip caches are too small.	Add external L3 cache.	Pentium III
	Move L3 cache on-chip.	Pentium 4

2.3 Intel Core CPUs

- Evolution of Cache organization intel microprocessors

	Max Cores	Max Base Clock	Max Cache
Core m3	2	1.1 GHz	4 MB
Core i3	5*	4.0 GHz	8 MB
Core i5	6	4.1 GHz	12 MB
Core i7	8	4.0 Ghz	16 MB
Core i9	10	3.7 GHz	20 MB
Core X	18	3.8 GHz	20.75 MB

Based on available CPUs, March 2021. E&OE. *The i3-L13G4 has 5 cores.

- Intel Xeon CPUs are used in Servers and high-end workstations

2. ARM Cache Organization

- ARM Cache features

Core	Cache Type	Cache Size (kB)	Cache Line Size (words)	Associativity	Location	Write Buffer Size (words)
ARM720T	Unified	8	4	4-way	Logical	8
ARM920T	Split	16/16 D/I	8	64-way	Logical	16
ARM926EJ-S	Split	4-128/4-128 D/I	8	4-way	Logical	16
ARM1022E	Split	16/16 D/I	8	64-way	Logical	16
ARM1026EJ-S	Split	4-128/4-128 D/I	8	4-way	Logical	8
Intel StrongARM	Split	16/16 D/I	4	32-way	Logical	32
Intel Xscale	Split	32/32 D/I	8	32-way	Logical	32
ARM1136-JF-S	Split	4-64/4-64 D/I	8	4-way	Physical	32

- ARM7 models used a unified L1 cache, while all subsequent models use a split instruction/data cache
- ARM architecture uses a small first-in-first-out (FIFO) write buffer to enhance memory performance

3. Internal Memory

- In the early 1950s, Jay W. Forrester and his group at MIT developed magnetic core storage to replace *vacuum tubes* as the ***main computer storage***.
- ***Main memory*** was often referred to as **core**, a term that persists to this day
- Invented in1948 transistors became the memory of choice by 1960s
- Transistors are tiny electrically operated switches that can alternate between ‘on’ and ‘off’ states many millions of times per second
- At the start of the twenty first century, IC-based semiconductor memory was the primary memory device for main computer storage
- The use of semiconductor memories has increased the speed and decreased the price of main computer storage
- Categories of semiconductor memories exist; RAM (**S**RAM/**D**RAM), ROM, PROM, EPROM, EEPROM
- SRAM is faster, more expensive, less dense than DRAM and is used for cache memory
- DRAM is used for main memory

3.1 Semiconductor Memory Types

ACCRA TECHNICAL UNIVERSITY

COMPUTER SCIENCE DEPARTMENT

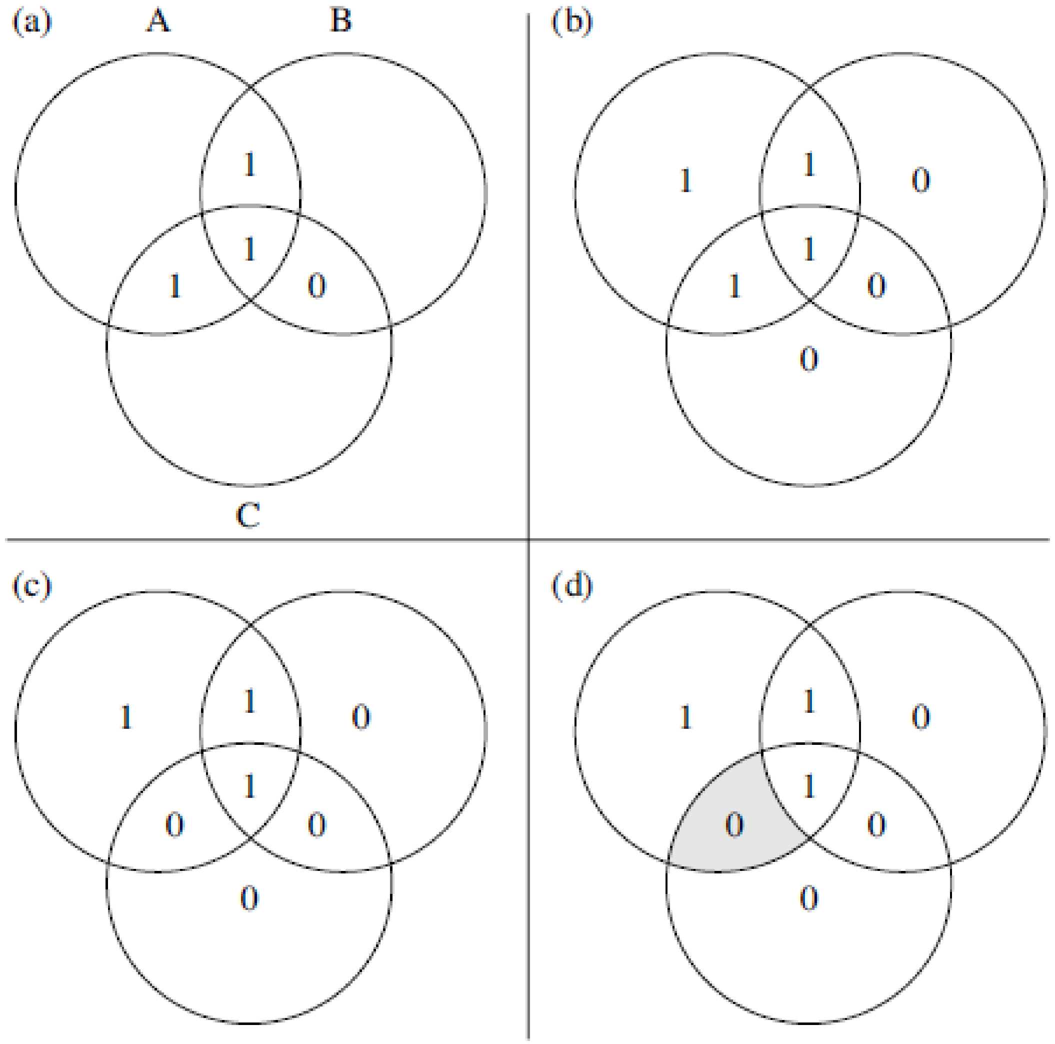
- Read only memories contain permanent patterns of data that cannot be changed
- A variation of ROM is PROM
- Another variation of the read only memory is the *read mostly memory*
- Forms of read mostly memory are EPROM, EEPROM and flash memory

<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</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</td><td rowspan="2">Not possible</td><td>Masks</td><td rowspan="5">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</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>

3.2 Error Detection and Correction

- Semi-conductor memories are prone to two kinds of errors
  1. Hard failure
  2. Soft errors
- Hard failure is permanent physical defect so that the memory cell affected cannot reliably store data but become stuck at 0 or 1 or switch erratically between 0 or 1
- Soft error is a random, non-destructive event that alters the content of one or more memory cells without damaging the memory system and may be caused by power supply problems or radioactive particles
- If error can be corrected, then use
  1. Single Error Correction (SEC) i.e. ***Hamming code***
  2. Single Error Correction - Double Error Detection (SEC-DED) code i.e. ***Extended Hamming code***

3.2 Hamming Error-Correcting Code



3.3 Advanced DRAM Organization

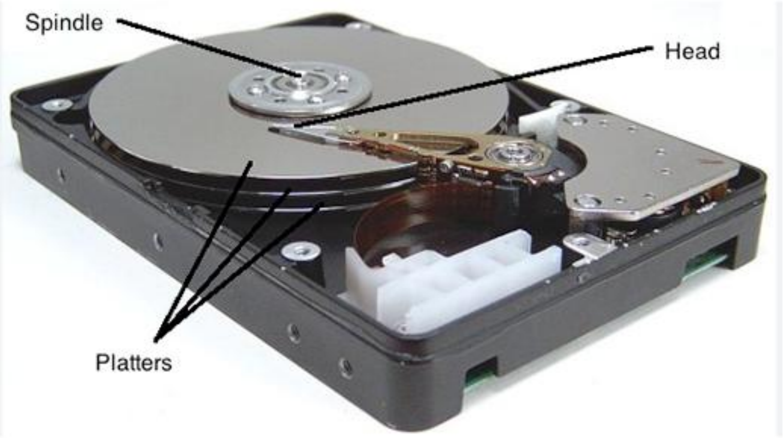
- The basic building block of main memory remains the DRAM chip
- Traditional DRAM is constrained both by its internal architecture and its interface to the processor memory bus
- Hence insertion of one or more levels of high-speed SRAM cache between main memory and processor
- But SRAM is more expensive than DRAM and expanding cache size beyond a certain limit yields diminishing returns
- In recent years a number of enhancements to the basic asynchronous DRAM architecture have been explored
- These enhancement schemes include; SDRAM (DDR-SDRAM), DDR-DRAM (DDR2, DDR3), RDRAM (CRDRAM, DRDRAM)
- Cache DRAM (CDRAM) is also receiving considerable attention

4. External Memory

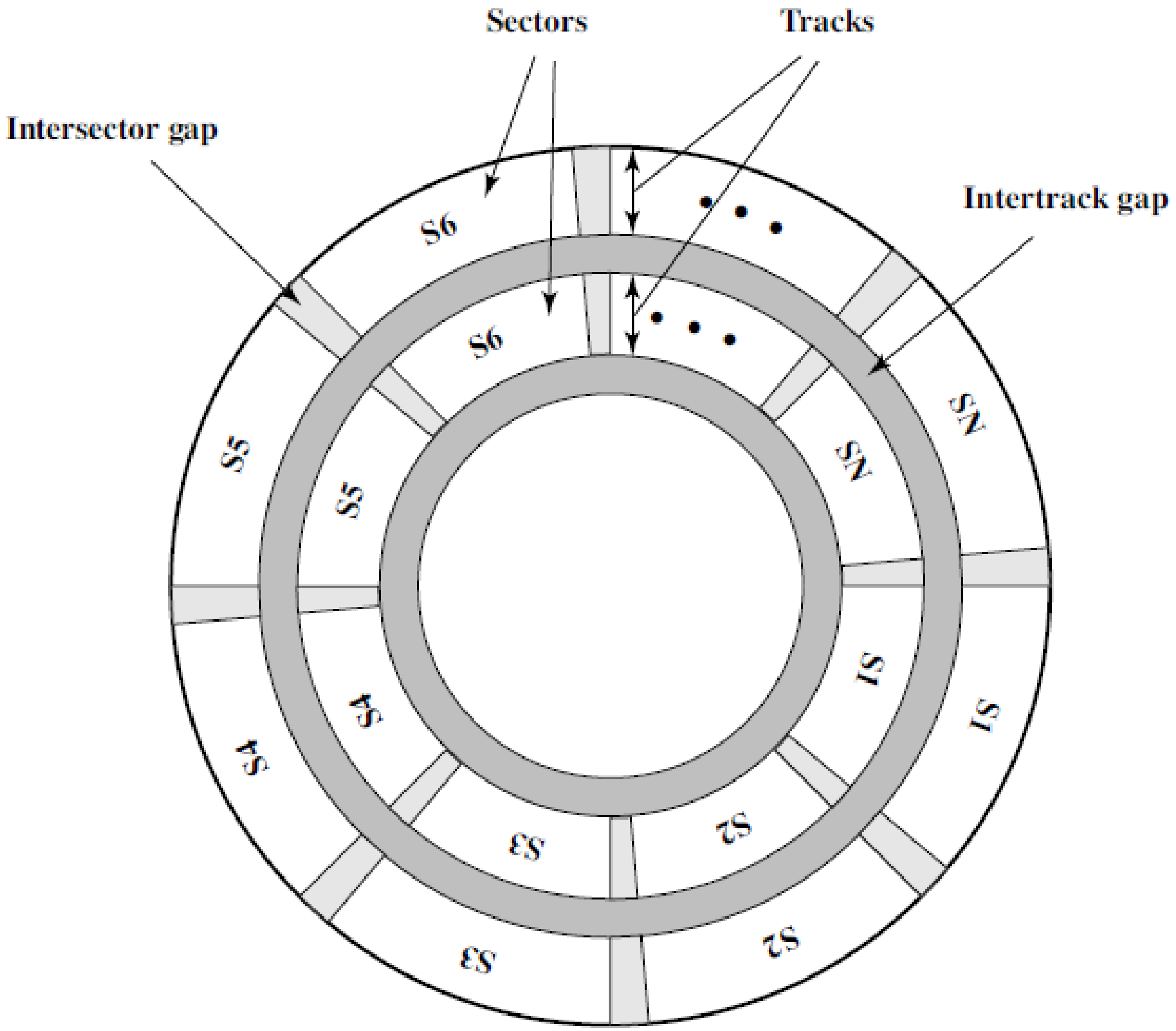
- Also known as secondary memory or auxiliary memory refers to storage facility other than the ***main memory*** and on the internet – *in the cloud*
- It will be nice to use only the fastest memory, but because that is the most expensive memory, we trade off access time for cost by using more of the slower memory
- No one technology is optimal in satisfying the memory requirements of the computer system.
- Current trends make use of
  1. magnetic surfaces,
  2. optical and
  3. Magneto-optical technologies
- Magnetic disks are the foundation of external memory on virtually all computer systems
- A disk is a circular platter constructed of nonmagnetic material, called the ***substrate*** coated with a '**magnetizable**' material

4.1 Magnetic Memory

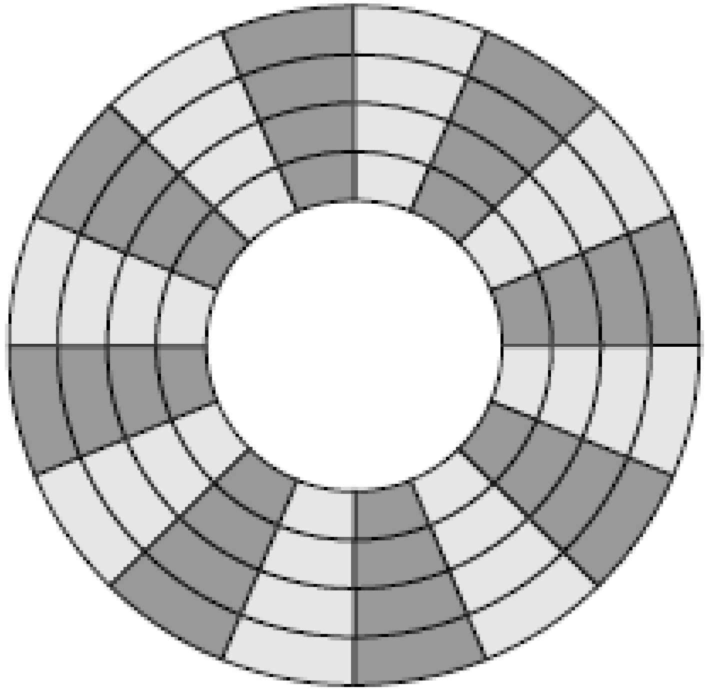
- Data are recorded on and retrieved later from the disk via a conducting coil named the ***head***;
- In many systems there are two heads; a *read head* and a *write head*
- The read head consists of a magneto-resistive sensor (**MR**)
- By passing current through the MR sensor, resistance changes are detected as voltage signals
- The head is a relatively small device capable of writing to or reading from a platter rotating beneath it
- Data is organized in concentric set of rings called ***tracks***
- Each track is the same width as the head
- There are thousands of tracks per surface
- Data are transferred to and fro the disk in ***sectors***



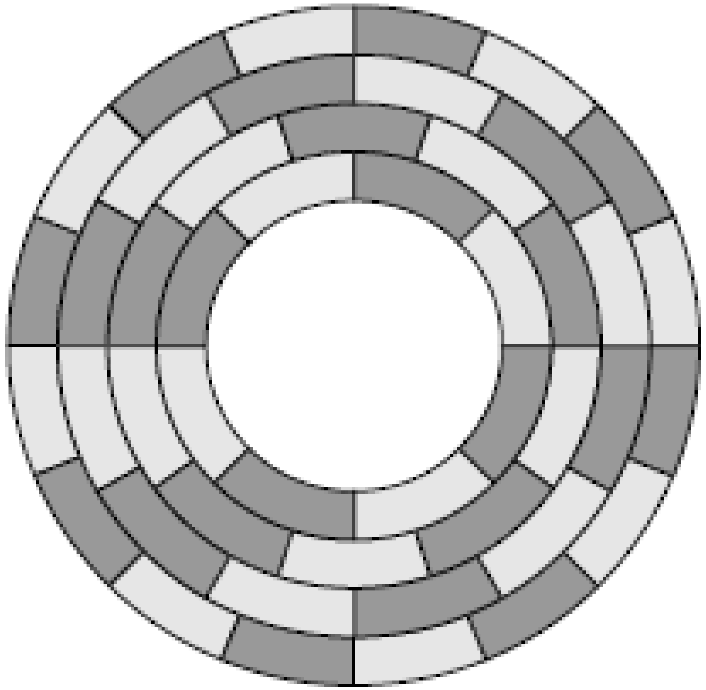
4.1.1 Disk layout



4.1.2 Disk Layout Methods



(a) Constant angular velocity



(b) Multiple zoned recording

4.1.3 Physical Characteristics of Disks

Head Motion
- Fixed head (one per track)
- Movable head (one per surface)

Disk Portability
- Nonremovable disk
- Removable disk

Sides
- Single sided
- Double sided

Platters
- Single platter
- Multiple platter

Head Mechanism
- Contact (floppy)
- Fixed gap
- Aerodynamic gap (Winchester)

4.1.4 Disk Performance Parameters

- On a movable system, the time it takes to position the head at the track is known as the ***seek time***
- The time it takes for the beginning of a sector to reach the head is known as ***rotational delay*** or ***rotational latency***
- The sum of the seek time, if any and the rotational delay equals the ***access time*** which is the time it takes to get into position to read or write
- Once the head gets into position, the read or write operation is performed as the sector moves under the head; this is the data transfer portion of the operation; the time required for the transfer is the ***transfer time***

4.2 Redundant Array of Independent Disks

- RAID is a standardized scheme for multiple-disks database design
- The RAID strategy employs multiple disk drives and distributes data in such a way as to enable simultaneous access to data from multiple drives, thereby improving I/O performance and allowing easier incremental increases in capacity
- Multiple disks increase the risk of failure
- RAID makes use of stored parity information that enables recovery of data lost due to disk failure
- The RAID scheme is made of seven levels, zero to six that designate different design architectures
- There are also nested RAID levels such as RAID10 (RAID 1+0), RAID03 (RAID 0+3), RAID50 (RAID 5+0)
- And non-standard RAID levels like RAID 7, Adaptive RAID, JBOD

4.2.1 RAID Levels

Category	Level	Description	Disks Required	Data Availability	Large I/O Data Transfer Capacity	Small I/O Request Rate
Striping	0	Nonredundant	N	Lower than single disk	Very high	Very high for both read and write
Mirroring	1	Mirrored	$2N$	Higher than RAID 2, 3, 4, or 5; lower than RAID 6	Higher than single disk for read; similar to single disk for write	Up to twice that of a single disk for read; similar to single disk for write
Parallel access	2	Redundant via Hamming code	$N + m$	Much higher than single disk; comparable to RAID 3, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
	3	Bit-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
Independent access	4	Block-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 5	Similar to RAID 0 for read; significantly lower than single disk for write	Similar to RAID 0 for read; significantly lower than single disk for write
	5	Block-interleaved distributed parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 4	Similar to RAID 0 for read; lower than single disk for write	Similar to RAID 0 for read; generally lower than single disk for write
	6	Block-interleaved dual distributed parity	$N + 2$	Highest of all listed alternatives	Similar to RAID 0 for read; lower than RAID 5 for write	Similar to RAID 0 for read; significantly lower than RAID 5 for write

N = number of disks; m proportional to log$N$

4.3 Optical Memory

CD
Compact Disk. A nonerasable disk that stores digitized audio information. The standard system uses 12-cm disks and can record more than 60 minutes of uninterrupted playing time.

CD-ROM
Compact Disk Read-Only Memory. A nonerasable disk used for storing computer data. The standard system uses 12-cm disks and can hold more than 650 Mbytes.

CD-R
CD Recordable. Similar to a CD-ROM. The user can write to the disk only once.

CD-RW
CD Rewritable. Similar to a CD-ROM. The user can erase and rewrite to the disk multiple times.

DVD
Digital Versatile Disk. A technology for producing digitized, compressed representation of video information, as well as large volumes of other digital data. Both 8 and 12 cm diameters are used, with a double-sided capacity of up to 17 Gbytes. The basic DVD is read-only (DVD-ROM).

DVD-R
DVD Recordable. Similar to a DVD-ROM. The user can write to the disk only once. Only one-sided disks can be used.

DVD-RW
DVD Rewritable. Similar to a DVD-ROM. The user can erase and rewrite to the disk multiple times. Only one-sided disks can be used.

Blu-Ray DVD
High definition video disk. Provides considerably greater data storage density than DVD, using a 405-nm (blue-violet) laser. A single layer on a single side can store 25 Gbytes.

4.3.1 Optical Memory Characteristics