Database Systems (CSF212)

BITS Pilani
Hyderabad Campus

Dr.R.Gururaj
CS&IS Dept.

Acknowledgements

The content of the slides (both Text and Figures) are taken from the following source:

http://www.db-book.com/

Some changes are made as per the need.

Chapter 16: Storage & File Structure

Content

1. Overview of Physical Storage Media
2. Magnetic Disks
3. RAID , Storage Access ,File Organization
4. Organization of Records in Files
5. Clustering
6. Data dictionary

Classification of Physical Storage Media

- ❑ Speed with which data can be accessed
- ❑ Cost per unit of data
- ❑ Reliability
- ❑ data loss on power failure or system crash
- ❑ physical failure of the storage device

Can differentiate storage into:

1. Volatile storage: loses contents when power is switched off .

2. Non-volatile storage: Contents persist even when power is switched off.

Physical Storage Media

Cache: Fastest and most costly form of storage; Volatile;

Main memory: Fast access; Generally too small ; Too expensive) to store the entire database ; Volatile

Flash memory

Data survives power failure

Reads are roughly as fast as main memory

But writes are slow

Widely used in embedded devices such as digital cameras;

Magnetic-disk:

Data is stored on spinning disk, and read/written magnetically;

Primary medium for the long-term storage of data;

Typically stores entire database.

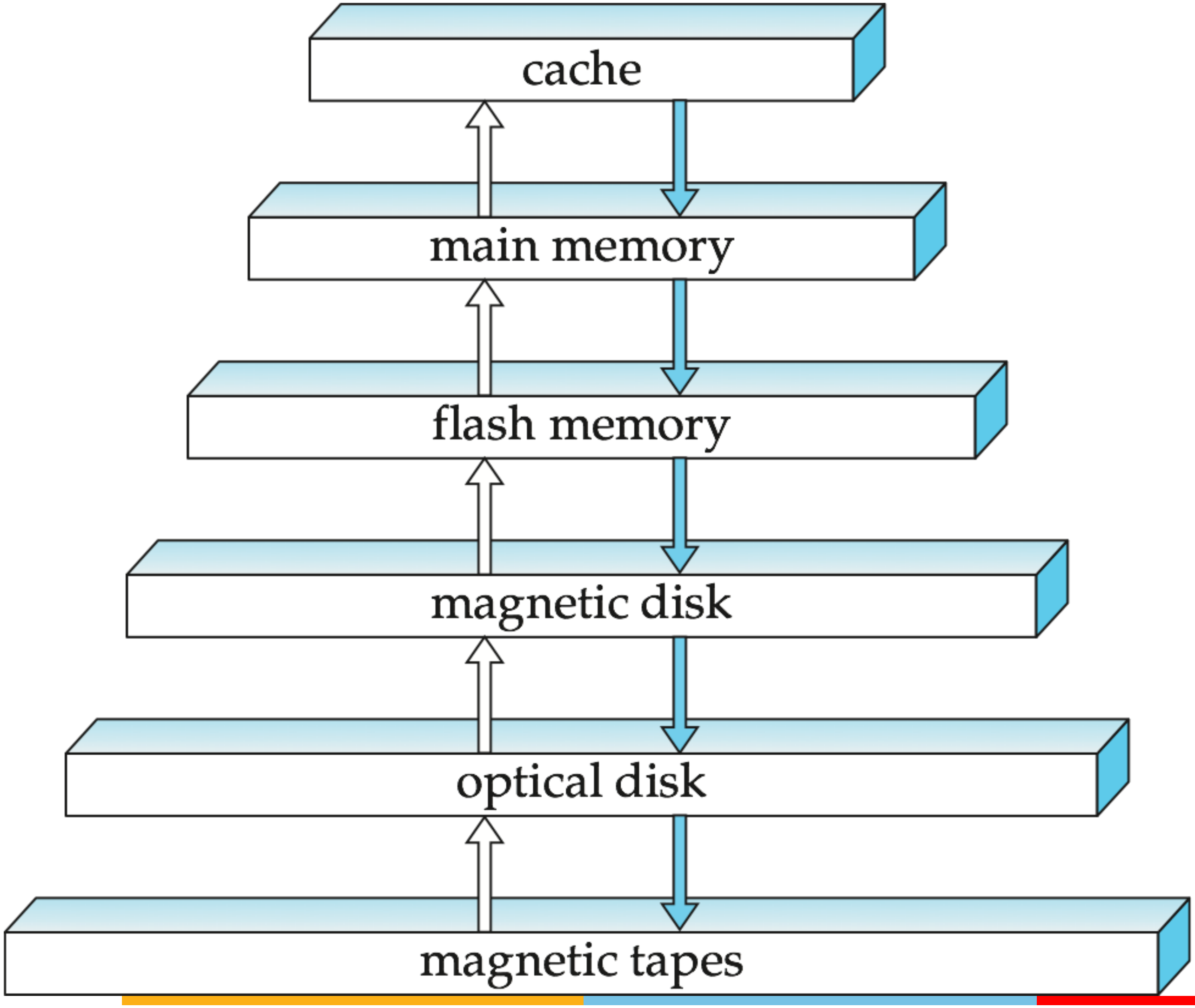
Survives power failures and system crashes

Optical storage: Non-volatile, data is read optically from a spinning disk using a laser ;CD-ROM (640 MB) and DVD (4.7 to 17 GB) most popular forms ; Reads and writes are slower than with magnetic disk ;

Tape storage

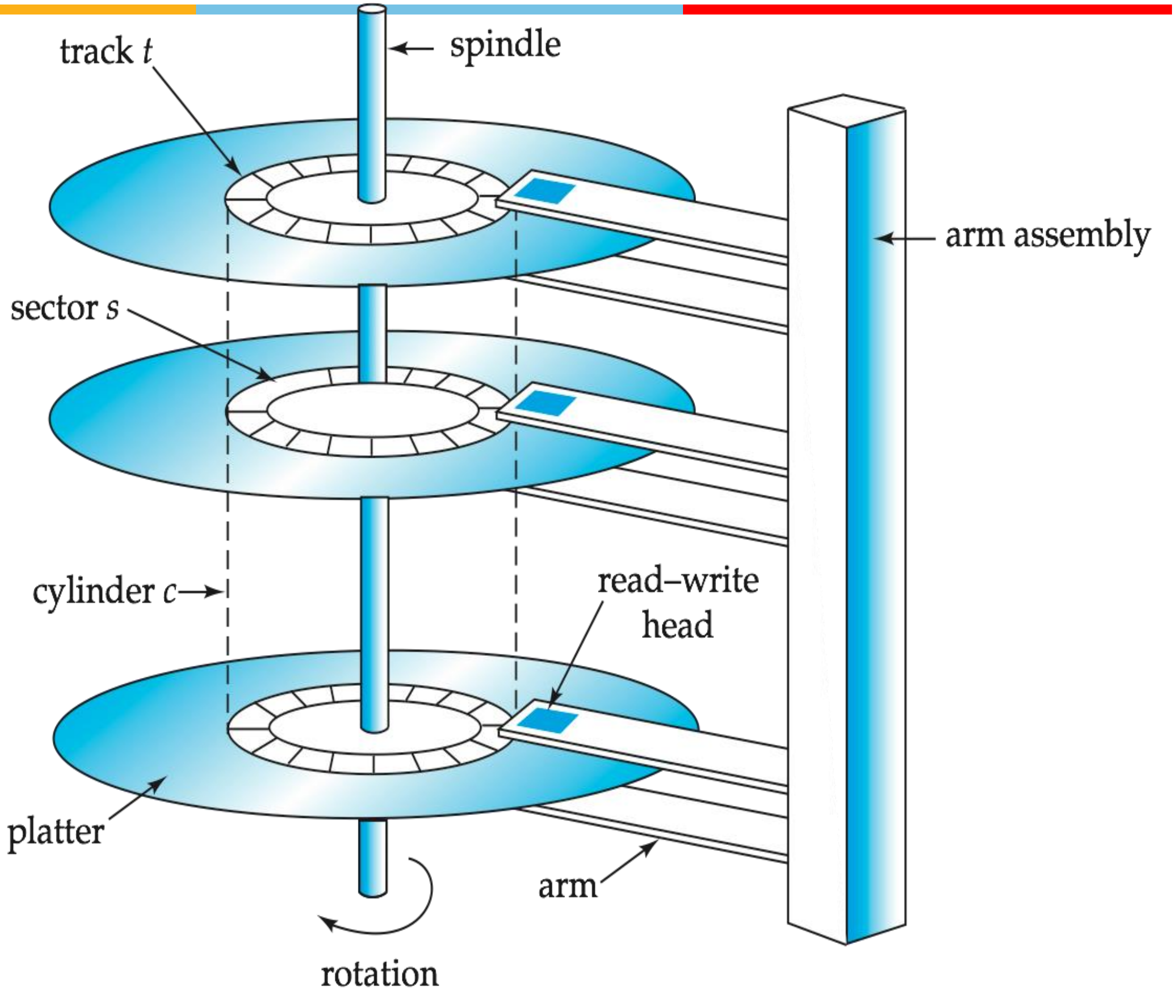
Non-volatile, used primarily for backup (to recover from disk failure), and for archival data; sequential-access – much slower than disk ;very high capacity (40 to 300 GB tapes available)

Storage Hierarchy



- **primary storage**: Fastest media but volatile (cache, main memory).
- **secondary storage**: next level in hierarchy, non-volatile, moderately fast access time
  - also called **on-line storage**
  - E.g. flash memory, magnetic disks
- **tertiary storage**: lowest level in hierarchy, non-volatile, slow access time
  - also called **off-line storage**
  - E.g. magnetic tape, optical storage

Magnetic Hard Disk Mechanism



NOTE: Diagram is schematic, and simplifies the structure of actual disk drives

Read-write head

- Positioned very close to the platter surface (almost touching it)
- Reads or writes magnetically encoded information.

Surface of platter divided into circular tracks

- Over 50K-100K tracks per platter on typical hard disks

Each track is divided into sectors.

- Sector size typically 512 bytes – many KBs
- Typical sectors per track: 500 to 1000 (on inner tracks) to 1000 to 2000 (on outer tracks)

To read/write a sector

- disk arm swings to position head on right track
- platter spins continually; data is read/written as sector passes under head

Head-disk assemblies

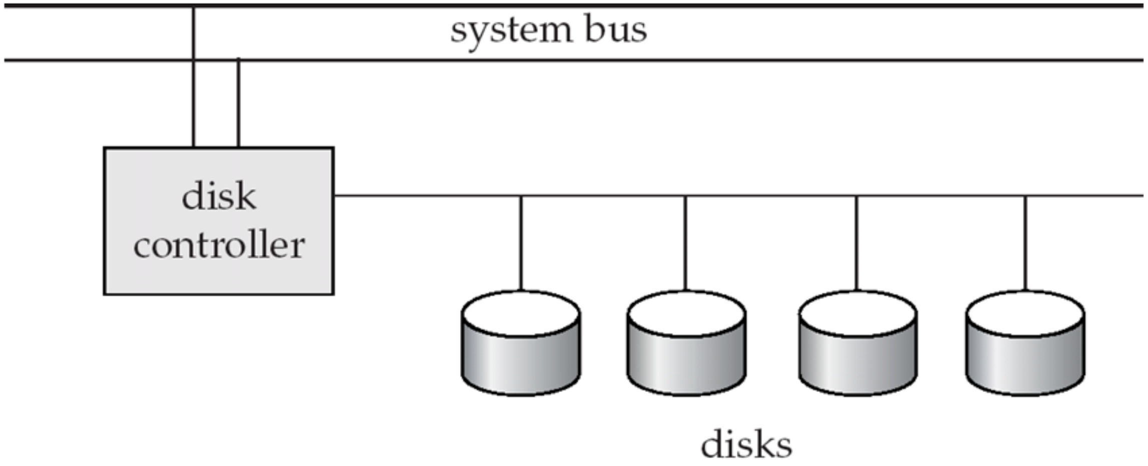
- multiple disk platters on a single spindle (1 to 5 usually)
- one head per platter, mounted on a common arm.

Cylinder *i* consists of *i*th track of all the platters

Disk controller – interfaces between the computer system and the disk drive hardware.

- – accepts high-level commands to read or write a sector
- – initiates actions such as moving the disk arm to the right track and actually reading or writing the data
- – Computes and attaches **checksums** to each sector to verify that data is read back correctly
  - If data is corrupted, with very high probability stored checksum won’t match recomputed checksum

Disk Subsystem



Multiple disks connected to a computer system through a controller.

- Disks usually connected directly to computer system
- In **Storage Area Networks (SAN)**, a large number of disks are connected by a high-speed network to a number of servers

Performance Measures of Disks

- **Access time** – the time it takes from when a read or write request is issued to when data transfer begins. Consists of:
  - **Seek time** – time it takes to reposition the arm over the correct track.
    - 4 to 10 milliseconds on typical disks
  - **Rotational latency** – time it takes for the sector to be accessed to appear under the head.
    - Average latency is 1/2 of the worst case latency.
    - 4 to 11 milliseconds on typical disks (5400 to 15000 r.p.m.)
- **Data-transfer rate** – the rate at which data can be retrieved from or stored to the disk.
  - 25 to 100 MB per second max rate, lower for inner tracks

- **Mean time to failure (MTTF)** – the average time the disk is expected to run continuously without any failure.
  - Typically 3 to 5 years

Optimization of Disk-Block Access

- **Block** – a contiguous sequence of sectors from a single track
  - data is transferred between disk and main memory in blocks
  - sizes range from 512 bytes to several kilobytes
    - Smaller blocks: more transfers from disk
    - Larger blocks: more space wasted due to partially filled blocks
    - Typical block sizes today range from 4 to 16 kilobytes

- **Disk-arm-scheduling** algorithms order pending accesses to tracks so that disk arm movement is minimized
  - **elevator algorithm**:

- **File organization** – optimize block access time by organizing the blocks to correspond to how data will be accessed
  - E.g. Store related information on the same or nearby cylinders.
  - Files may get **fragmented** over time
    - E.g. if data is inserted to/deleted from the file
    - Or free blocks on disk are scattered, and newly created file has its blocks scattered over the disk
    - Sequential access to a fragmented file results in increased disk arm movement

- – Some systems have utilities to defragment the file system, in order to speed up file access

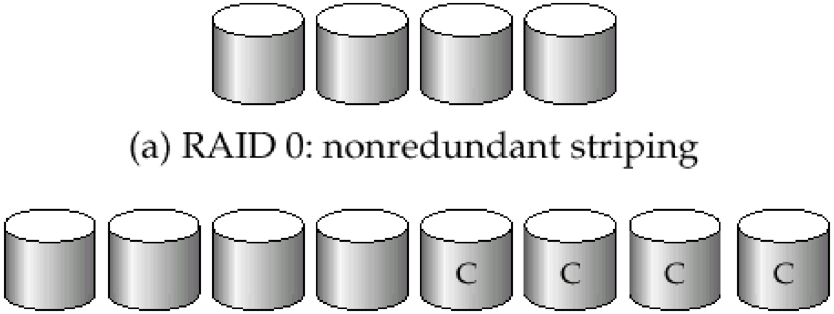
RAID

- **RAID: Redundant Arrays of Independent Disks**
  - disk organization techniques that manage a large numbers of disks, providing a view of a single disk of
    - high capacity and high speed by using multiple disks in parallel,
    - high reliability by storing data redundantly, so that data can be recovered even if a disk fails

RAID Levels

Different RAID organizations, or RAID levels, have differing cost, performance and reliability characteristics.

We have RAID Leve-0 to Level-6



(a) RAID 0: nonredundant striping

(b) RAID 1: mirrored disks

Files and Records

- A **file** is a *sequence* of records, where each record is a collection of data values (or data items).
- Records are stored on disk blocks.
- The **blocking factor (bfr)** for a file is the (average) number of file records stored in a disk block.
- A file can have **fixed-length** records or **variable-length** records.

Record Organization

- ❑ File records can be unspanned or spanned
- ❑ File operations

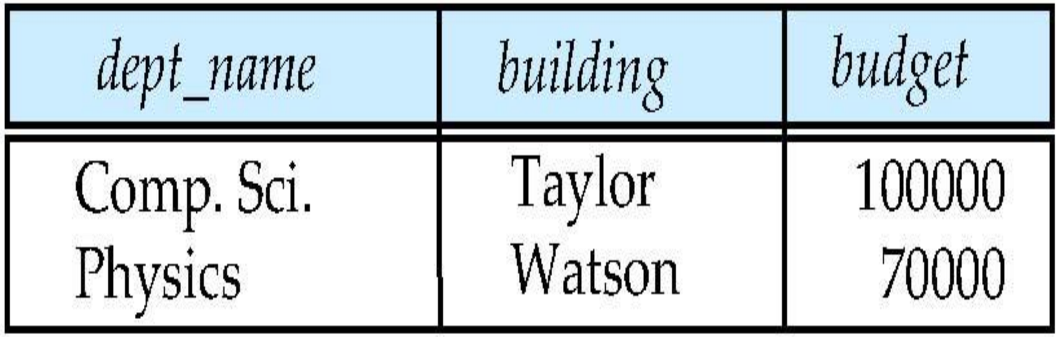
File Organization

- ❖ Unordered Files (heap)
- ❖ Ordered Files (sequential)

Multitable Clustering File Organization

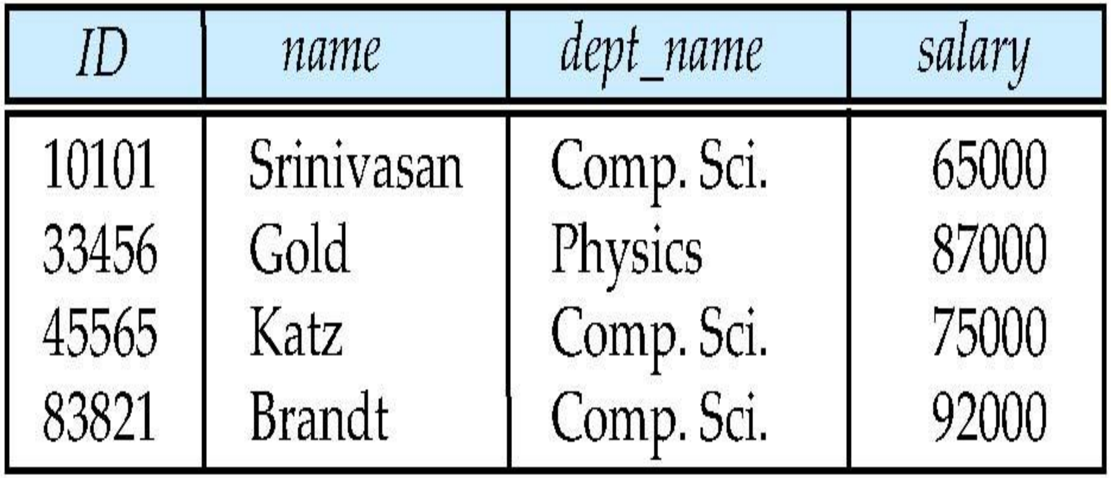
Store several relations in one file using a **multitable clustering** file organization

department

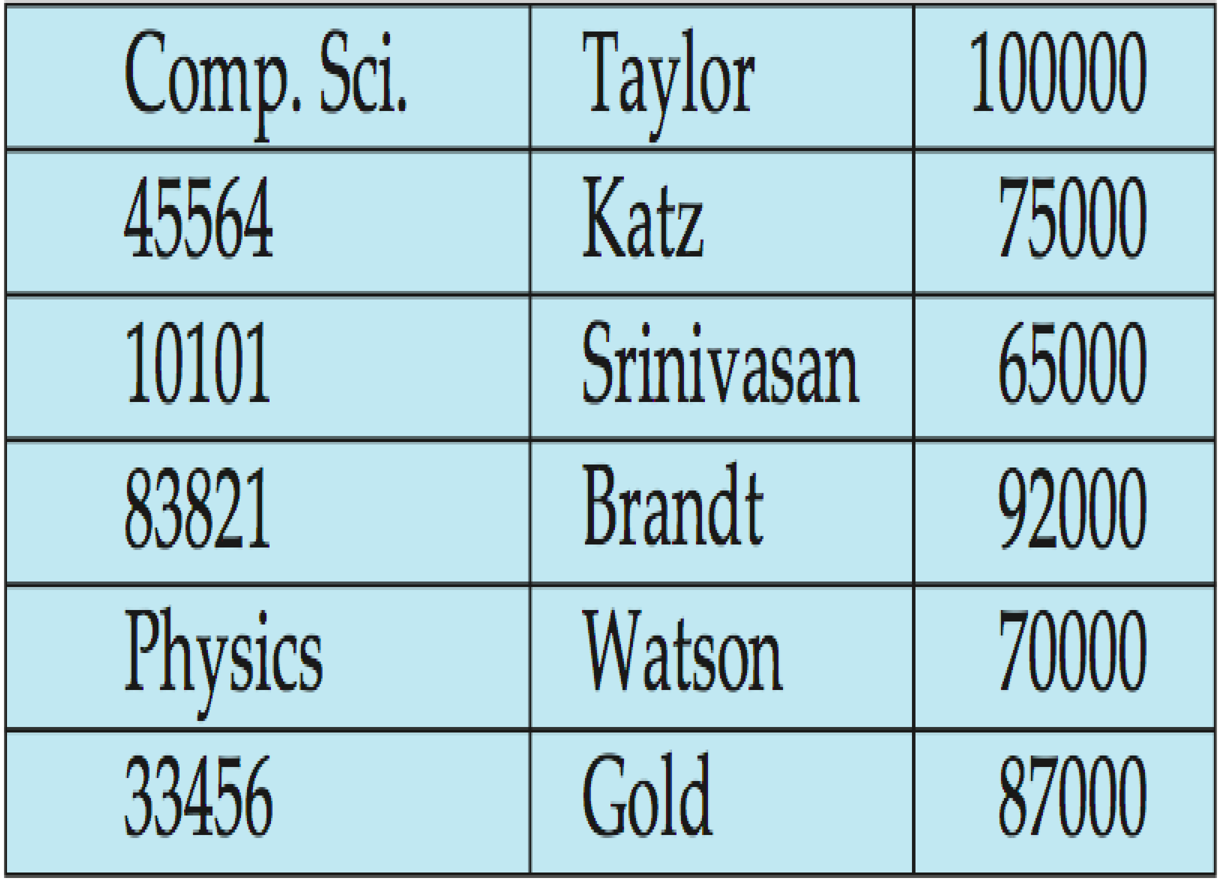


dept_name	*building*	*budget*
Comp. Sci.	Taylor	100000
Physics	Watson	70000

instructor



ID	*name*	*dept_name*	*salary*
10101	Srinivasan	Comp. Sci.	65000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
83821	Brandt	Comp. Sci.	92000



Comp. Sci.	Taylor	100000
45564	Katz	75000
10101	Srinivasan	65000
83821	Brandt	92000
Physics	Watson	70000
33456	Gold	87000

multitable clustering of *department* and *instructor*

- good for queries involving *department instructor*, and for queries involving one single department and its instructors
- bad for queries involving only *department*
- results in variable size records

Data Dictionary Storage

The **Data dictionary** (also called **system catalog**) stores **metadata**; that is, data about data, such as

- Information about relations
  - names of relations
  - names, types and lengths of attributes of each relation
  - names and definitions of views
  - integrity constraints
- User and accounting information, including passwords
- Statistical and descriptive data
  - number of tuples in each relation
- Physical file organization information
  - How relation is stored (sequential/hash/…)
  - Physical location of relation
- Information about indices

Storage Access

- A database file is partitioned into fixed-length storage units called **blocks**. Blocks are units of both storage allocation and data transfer.
- Database system seeks to minimize the number of block transfers between the disk and memory. We can reduce the number of disk accesses by keeping as many blocks as possible in main memory.
- **Buffer** – portion of main memory available to store copies of disk blocks.
- **Buffer manager** – subsystem responsible for allocating buffer space in main memory.

Summary

1. Overview of Physical Storage Media
2. Magnetic Disks
3. Tertiary Storage
4. RAID
5. Storage Access
6. File Organization
7. Organization of Records in Files
8. Data-Dictionary Storage