Introduction do File Management

Today

- Introduction to the course subject matter
- Discussion of the course syllabus and course organization

Reference : Folk, Zoellick and Riccardi. Sections 1.1 and 1.2.

Course Contents

- Data processing from a computer science perspective
  - Storage of data
  - Organization of data
  - Access to data
  - Processing of data

  This will be built on your knowledge of Data Structures.

- Data Structures vs File Structures
  - Both involve :

    Representation of Data

    +

    Operations for accessing data
  - Difference :

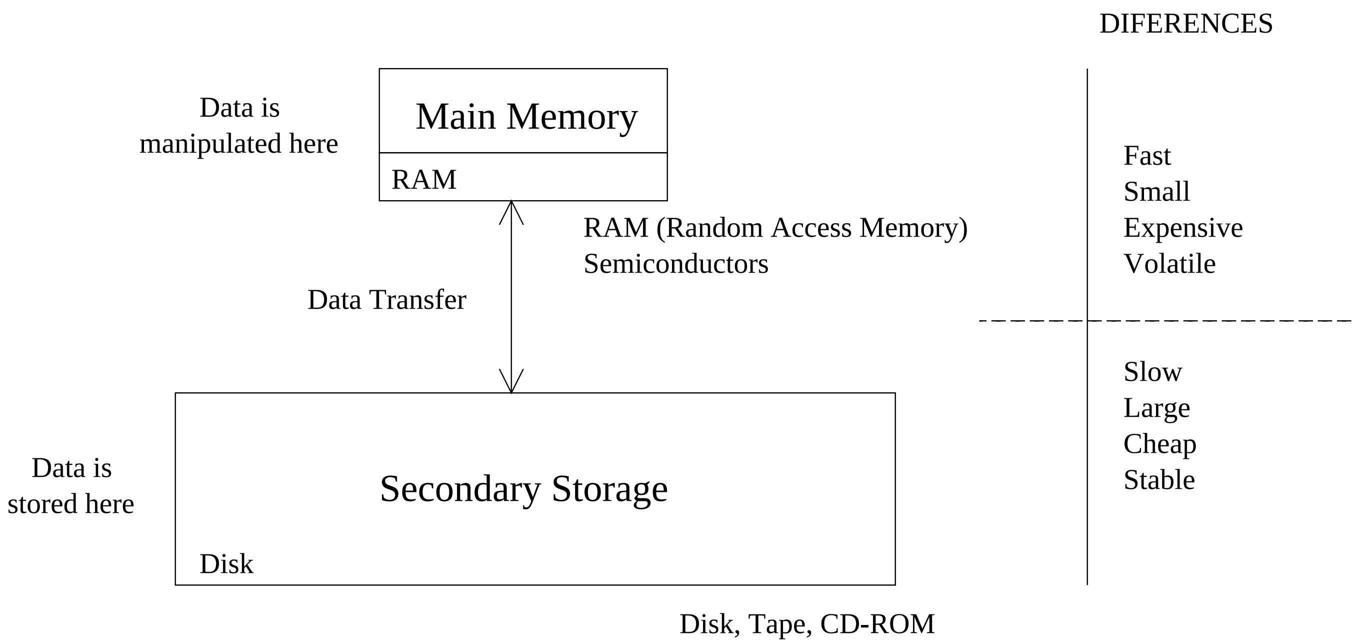
    Data Structures : deal with data in main memory

    File Structures : deal with data in secondary storage (Files)

In order that you appreciate the difference, let us take a brief look at ...

Computer Architecture

L



How fast is main memory in comparison to secondary storage ?

Typical time for getting info from:

main memory:	$\sim$ 12 nanoseconds	$= 120 \times 10^{-9}$ secs
magnetics disks:	$\sim$ 30 milliseconds	$= 30 \times 10^{-3}$ secs

An analogy keeping same time proportion as above:

Looking at the index of a book	20 secs
versus	
Going to the library	58 days

Main Memory

- Fast (since electronic)
- Small (since expensive)
- Volatile (information is lost when power failure occurs)

Secondary Storage

- Slow (since electronic and mechanical)
- Large (since cheap)
- Stable, persistent (information is preserved longer)

Goal of the file structure and what we will study in this course:

- Minimize number of trips to the disk in order to get desired information. Ideally get what we need in one disk access or get it with as few disk accesses as possible.
- Grouping related information so that we are likely to get everything we need with only one trip to the disk (e.g. name, address, phone number, account balance).

History of File Structure Design

1. In the beginning ... it was the tape
   - **Sequential access**
   - Access cost proportional to size of file
     [Analogy to sequential access to array data structure]
2. Disks became more common
   - **Direct access** [Analogy to access to position in array - binary search in sorted arrays]
   - **Indexes** were invented
     - list of keys and pointers stored in small file
     - allows direct access to a large primary file

     Great if index fits into main memory.
     As a file grows we have the same problem we had with a large primary file.
3. Tree structures emerged for main memory (1960's)
   - Binary search trees (BST's)
   - **Balanced**, self adjusting BST's : e.g. AVL trees (1963)
4. A tree structure suitable for files was invented : **B trees** (1979) and **B+ trees**)

   Good for accessing millions of records with 3 or 4 disk accesses
5. What about getting info with a single request ?
   - **Hashing Tables** (Theory developed over 60's and 70's but still a research topic)
     Good when files do not change to much in time
   - **Extendible, dynamic hashing** (late 70's and 80's)
     One or two disk access even if file grows dramatically

Discussion of the course syllabus and course organization

- Discussion of the course list of topics in light of this motivation
- Administrative details: course description
- Form to be completed by students on Tutorial/Lab time preferences