Data Management and File Structure

Lab Work

External Sorting

External sorting is used when the size of the data file is larger than the available memory. In this case, the data is loaded into memory partly, sorted and written into a temporary file named a run. Then the runs are merged into a sorted file.

In choosing a sort algorithm, two characteristics of sort algorithms are considered:

1- The sort algorithm should be able to overlap sorting with I/O
2- The sort algorithm should be able to create large runs

In this lab work we will examine heap sort and its capability to create large runs.

Assume the available memory to load data in as large as 20 records (each record is an integer in our example)

Data file is a set of 50 random numbers. Read 20 numbers and create a heap tree. Then delete the root node of the heap tree and write it in a second file. At this stage we can read a new number from data file and insert it into the heap tree IF it is larger than the last number deleted from the heap tree. Repeat until no more data can be inserted into the heap (the next integer is smaller than the last deleted number). At this stage delete the remaining nodes from the heap tree and write them into the second file. Count the number of integers in the output file and report it (generally it is larger than the size of the available memory)

Repeat the procedure for 20 times and measure the average size of the output file (the run) and report it.