Exercise 13.1 & 13.2

Exercise 13.1 Suppose you have a file with 10,000 pages and you have three buffer pages. Answer the following questions for each of these scenarios, assuming that our most general external sorting algorithm is used:

(a) A file with 10,000 pages and three available buffer pages.

(b) A file with 20,000 pages and five available buffer pages.

(c) A file with 2,000,000 pages and 17 available buffer pages.

1. How many runs will you produce in the first pass?
2. How many passes will it take to sort the file completely?
3. What is the total I/O cost of sorting the file?
4. How many buffer pages do you need to sort the file completely in just two passes?

Exercise 13.2 Answer Exercise 13.1 assuming that a two-way external sort is used.

Solution 13.1

1. In the first pass (Pass 0), $\lceil N/B \rceil$ runs of B pages each are produced, where N is the number of file pages and B is the number of available buffer pages:
(a) $\lceil 10000/3 \rceil = 3334$ sorted runs.
(b) $\lceil 20000/5 \rceil = 4000$ sorted runs.
(c) $\lceil 2000000/17 \rceil = 117648$ sorted runs.

2. The number of passes required to sort the file completely, including the initial sorting pass, is $\lceil log_{B-1} N1 \rceil + 1$, where $N1 = \lceil N/B \rceil$ is the number of runs produced by Pass 0:
(a) $\lceil log3334 \ / \ log2 \rceil + 1 = 13$ passes.
(b) $\lceil log4000 \ / \ log4 \rceil + 1 = 7$ passes.
(c) $\lceil log117648 \ / \ log16 \rceil + 1 = 6$ passes.

3. Since each page is read and written once per pass, the total number of page I/Os for sorting the file is $2 * N * (\#passes)$:
   (a) 2*10000*13 = 260000.
   (b) 2*20000*7 = 280000.
   (c) 2*2000000*6 = 24000000.

4. In Pass 0, $\lceil N/B \rceil$ runs are produced. In Pass 1, we must be able to merge this many runs; i.e., $B - 1 \geq \lceil N/B \rceil$. This implies that B must at least be large enough to satisfy $B * (B-1) \geq N$; this can be used to guess at B, and the guess must be validated by checking the first inequality. Thus:
   (a) With 10000 pages in the file, $B = 101$ satisfies both inequalities, $B = 100$ does not, so we need 101 buffer pages.
   (b) With 20000 pages in the file, $B = 142$ satisfies both inequalities, $B = 141$ does not, so we need 142 buffer pages.
   (c) With 2000000 pages in the file, $B = 1415$ satisfies both inequalities, $B = 1414$ does not, so we need 1415 buffer pages.

Solution 13.2

Answer Exercise 7.1 assuming that a 2-way external merge-sort is used (as per the description in Section 7.1).

1. In the first pass (Pass 0), N runs of 1 page each are produced, where N is the number of file pages:
   (a) 10000 sorted runs.
   (b) 20000 sorted runs.
   (c) 2000000 sorted runs.

2. The number of passes required to sort the file completely, including the initial sorting pass, is $\lceil log_2 N1 \rceil + 1$, where $N1 = N$ is the number of runs produced by Pass 0:
   (a) $\lceil log 10000 \ / \ log 2 \rceil + 1 = 15$ passes.
   (b) $\lceil log 20000 \ / \ log 2 \rceil + 1 = 16$ passes.
   (c) $\lceil log 2000000 \ / \ log 2 \rceil + 1 = 22$ passes.

3. Since each page is read and written once per pass, the total number of page I/Os for sorting the file is $2 * N * (\#passes)$:
   (a) 2*10000*15 = 300000.
   (b) 2*20000*16 = 640000.
   (c) 2*2000000*22 = 88000000.

4. Using 2-way merge sort, it is impossible to sort these files in 2 passes. Additional buffer pages do not help, since the algorithm always uses just 3 buffer pages.