CS351 – HW #1 – Solution to Question #6:

a) Take 5MB blocks from F1 to memory,
Search these records in F2 by taking 5MB from F2 each time,
If not found, write the record to FD,
Repeat this for all 5MBs of F1 (80/5 = 16 times).

b) Assuming that CPU time is not an issue we can calculate the file processing time.
We read F1 into 5MB space 16 times. Reading the entire F1 takes 28 sec.
We search F2 for each 5MB records of F1. Then we have 28 * 16 = 448 sec
Lastly, writing uncommon records takes 28 * 0.3 = 8.4 sec
Total: 28 + 448 + 8.4 = 484.4 sec

c) 16 times read 5MB of F1: 16*(s+r)
For each 5MB of F1(16 times), read F2 in pieces 5 MB at a time (16 times): 16*16*(s+r),

In order to write, we have to know the distribution of 30,000 uncommon records to 16 memory spaces of 5MB.
In the best case, these uncommon records are distributed u. 100,000/16 = 6250 records per memory space. 30,000/6250 = 4,8 which means 5*(s+r)
In the worst case, these records are distributed over all 16 memory spaces. 16*(s+r)

It is wise to choose the worst case in such questions. The answer is then
16*(s+r) + 16*16*(s+r) + 16*(s+r)

(The following is for the students who are interested in a detailed answer. You do not have to know this solution in the context of this course.)
But the best solution for writing is to know how they are randomly distributed.
Consider the following problem:
Let n be number of records into m blocks. If k records are randomly selected from n records, what is the expected number of blocks hit (blocks with at least one record selected)?
The answer is:

$$m \cdot \left[1 - \prod_{i=1}^{k} \frac{nd - i + 1}{n - i + 1} \right] \text{ where } d = 1 - 1/m.$$

You can read the following paper for details and proof:
Yao, S. B. Approximating block accesses in database organizations. *Communications of the ACM*, 20(4): 260-261, 1977. You can access this article from acm.org/dl.

Let us use this formula for our problem: How many of 16 memory spaces are these 30,000 uncommon records randomly distributed? The answer can be found by using the above formula with: n is 100,000 records in our problem. m = 16 is the number of memory spaces. And we have k = 30,000 randomly selected records.