Name: _____

CSE 4502/5717 Big Data Analytics
Exam II; November 12, 2020

Note: You are supposed to give proofs to the time and processor bounds of your algorithms. Read the questions carefully before attempting to solve them.

1. (20 points) Input are two sets X and Y of integers of size n each. These sets are stored in sorted order striped across D disks. The problem is to compute $X \cap Y$. Show that this can be done in $O\left(\frac{n}{DB}\right)$ parallel I/O operations.

2. (20 points) Input are sorted sequences $R_1, R_2, \ldots, R_\ell$ each of length n. These sequences are striped across D disks. Show how to merge these sequences in $O\left(\frac{\ell n}{BD} \log \ell\right)$ parallel I/O operations.

3. (20 points) The input for this problem is a string S of length n. The goal is to find the longest substring of S that occurs eactly three times in S (starting from three different positions). Present an $O(n)$ time algorithm to solve this problem. For instance, if $S = atgcaacgcagctat$, then the answer is gc (since it occurs exactly three times). Even though t is also a substring that occurs three times, it is not the longest.

4. (20 points) Input are k strings $S_1, S_2, \ldots, S_k$ (with $\sum_{i=1}^{k} |S_i| = M$). The problem is to identify a *signature* for every input string. U_i is a signature for S_i, if U_i is a substring of S_i of minimum length that does not occur in any of the other strings ($1 \leq i \leq k$). Present an $O(M)$ time algorithm for solving this problem.

5. (20 points) Input are a text T of length m and a pattern P of length n. We are also given the suffix array for T and the LCP values. T and P are strings from an alphabet Σ, with $\sigma = |\Sigma|$. The problem is to find all the occurrences of P in T within a Hamming distance of 1. If s_1 and s_2 are strings of the same length, then the Hamming distance between them is defined to be the number of places in which they differ. For example, the Hamming distance between *aacgtt* and *agcgat* is 2. Present an algorithm for this problem that runs in $O(\sigma n(n + \log m))$ time.