Name: ____________________________________

CSE 4502/5717 Big Data Analytics
Exam II; November 9, 2021

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 is a sequence X of n real numbers striped across D disks. It is known that the number of distinct elements in X is only C (where C is a constant). The goal is to sort X. Show that this can be done in two read passes through the data. Assume that $M = \Theta(BD)$, where M is the core memory size and B is the block size.

2. (20 points) Input is a sequence X with n elements that is residing in D disks. The problem is to identify the i^{th} smallest element of X. Assume that the core memory size is $M = \Theta(BD)$, B being the block size. Show how to solve this problem in $O\left(k\frac{n}{BD}\right)$ parallel I/O operations, where $k = \left\lceil \frac{i}{BD} \right\rceil$. (You can assume that the elements of X are distinct).

3. (20 points) Input are two strings S_1 and S_2 of length n each from an alphabet Σ of constant size, and an integer l. The problem is to check if there exist a substring X in S_1 and a substring Y in S_2 such that $|X| = |Y|$, $|X| \geq l$, and the Hamming distance between X and Y is no more than 1. (Given two strings of the same length, the Hamming distance between them is defined as the number of positions in which they differ. For instance, if $A = gacgta$ and $B = gccatc$, then the Hamming distance between A and B is 3 since they differ in positions $2, 4$, and 6.) Present an algorithm to solve this problem in $O(n^2)$ time.

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 the longest substring common to all the k input strings. Present an algorithm to solve this problem in $O(kM)$ time.

5. In this problem we are given a text T, a pattern P, and the suffix array S for T. The problem is to identify all the occurrences of P in T. Let $|T| = m$ and $|P| = n$.

   (a) (12 points) Present an algorithm to solve this problem in $O(1)$ time using mn arbitrary CRCW PRAM processors.

   (b) (8 points) Present an algorithm to solve this problem in $O(1)$ time using $n\sqrt{m}$ arbitrary CRCW PRAM processors. (Assume that the number of occurrences of P in T is no more than 1).

   Specifically, the output should be an array $A[1:m]$ such that $A[i] = 1$ if $P = T_i$; (If $T = t_1 t_2 \cdots t_m$ then $T_i = t_i t_{i+1} \cdots t_{i+n-1}$); Also, $A[i] = 0$ if $P \neq T_i$, for $1 \leq i \leq m$.