

Name: _____

CSE 4502/5717 Big Data Analytics

Exam I; March 3, 2026

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. (17 points) Input is an array $A[1 : n]$ of arbitrary real numbers. The array could only be of one of the following two types: 1) **Type I:** A has $\frac{n}{3}$ elements that are in the range $[1, 5)$, another $\frac{n}{3}$ elements in the range $[5, 10)$, and another $\frac{n}{3}$ elements in the range $[10, 15]$; or 2) **Type II:** A has $\frac{n}{3}$ elements that are in the range $[5, 10)$, another $\frac{n}{3}$ elements in the range $[10, 15]$, and another $\frac{n}{3}$ elements in the range $(15, 20]$. Present a Las Vegas algorithm that determines the type of the array in $\tilde{O}(\log n)$ time.

2. (16 points) Input is an array $A[1 : n]$ of real numbers. There are 5 elements in this array that have $\frac{n}{10}$ copies each. The other elements occur exactly once each. The problem is to output any one of these 5 elements. Present an $\tilde{O}(\log n)$ time Las Vegas algorithm to solve this problem.

3. (17 points) Input is a sequence $X = k_1, k_2, \dots, k_n$ of real numbers and an integer m , $1 \leq m \leq n$. The problem is to compute the average A_i of the elements $k_i, k_{i+1}, k_{i+2}, \dots, k_{i+m-1}$, for every i , $1 \leq i \leq (n - m + 1)$. Present an algorithm to solve this problem that uses $\frac{n}{\log n}$ CREW PRAM processors and runs in $O(\log n)$ time. (**Hint:** Use prefix computations).

4. (17 points) Input is a sequence X of n keys where each key is an arbitrary real number. It is known that the total number of distinct keys in X is only a constant. Present an algorithm to sort X that takes $O(\log n)$ time using $\frac{n}{\log n}$ arbitrary CRCW PRAM processors.

5. (17 points) Input is a sequence X of n arbitrary real numbers residing in a disk, with $n \gg M$, M being the main memory size. The problem is to identify the B smallest keys of X (where B is the block size). Assume that $M = cB$, for some constant c . Present an algorithm to solve this problem that makes only one pass through the data.

6. (16 points) Input are two sequences X and Y of real numbers in a disk. Each sequence is of length n . The problem is to check if they have a common element. Show how to solve this problem in $O\left(\frac{n}{B} \frac{\log(n/M)}{\log(M/B)}\right)$ I/O operations. Here M is the main memory size and B is the block size.