CSCI5020 External Memory Data Structures: Exercise List 1

In the following problems, B is the block size, and M is the memory capacity. we assume that M is a multiple of B.

Problem 1 (Group-by). Let S be a set of n tuples, each of which has the form (k, v), where k (or v, resp.) is called the *key* (*value*, resp.) of the tuple. We want to report, for each distinct key k that appears in S, the sum of the values of all the tuples whose keys are equal to k. Give an algorithm that achieves this purpose in $O(\frac{n}{B} \log_{M/B} \frac{t}{B})$ I/Os, where t is the number of distinct keys in S.

Problem 2 (f-Splitter). Let S be a set of n elements in $\mathbb{R}$. We want to find f splitters $p_1, p_2, ..., p_f \in S$ in ascending order such that there are $O(n/f)$ elements in the range $(p_{i-1}, p_i]$ for each $i \in [1, f+1]$, defining dummy splitters $p_0 = -\infty$ and $p_{f+1} = \infty$. Describe an algorithm to solve the problem in $O(n/B)$ I/Os for $f = M/B$ (note: the algorithm we discussed in class supports $f = \sqrt{M/B}$).

Problem 3 (k-Partitioning). Let S be a set of n elements in $\mathbb{R}$. Let k be an integer such that n is a multiple of k. We want to partition S into k disjoint subsets $S_1, S_2, ..., S_k$ such that (i) all the elements of S_i are smaller than those of S_j, for any i, j satisfying $1 \le i < j \le k$, and (ii) $|S_i| = n/k$ for each $i \in [1, k]$. It is required that these subsets be output in k arrays: an array for S_1, followed by an array for S_2, and so on. Prove that in the indivisibility model, when $\log_2 n \le B \log_2 \frac{M}{B}$, any algorithm must incur $\Omega(\frac{n}{B} \lceil \log_{M/B} k \rceil)$ I/Os solving this problem in the worst case.