CE/CZ 4123 Big Data Management Tutorial 7

More Practice for Preparing for Quizzes

1. Given the following three tables (primary keys are underlined):

   Employee(<u>EID</u>, Salary)

   Manager(<u>MID</u>, Salary)

   Employee-Manager(<u>EID</u>, MID)

   Each manager supervises at least one employees. Employee-Manager is a table that contains the manager ID (i.e., MID) for each employee (i.e., EID). How to convert the relational data model to a key-value data model? Consider that the main purpose of the conversion is for the query "Given an employee ID, find the salary of the employee's manager". The conversion should retain the information as much as possible.

2. Consider reading data from memory hierarchy consisting of L1 Cache, L2 Cache, and main memory with the following parameters.

   - L1 Cache:

     Read access time: 2 nanoseconds

     Miss ratio: 0.4

   - L2 Cache:

     Read access time: 10 nanoseconds

     Miss ratio: 0.2

   - Main memory:

     Read access time: 100 nanoseconds.

   Estimate the average data read cost and explain your answer. (Note: consider L1, L2 caches and main memory only).

3. In the lecture, we introduced a cost-free magic function telling which pages locate the qualified data for a query. Consider a disk page size is 1024 integers. There are 10240 integers, which are 1, 2, 3, …, 10240, sequentially stored at 10 consecutive disk pages.

Consider the following 4 queries using 4 scans over the data, where each query range is decided by three integers x, y, z, and 1<x<y<z<10240.

Query 1: searching values in the range [1, *x*] (i.e., values at least 1 and at most *x*)
Query 2: searching values in the range [*x*+1, *y*]
Query 3: searching values in the range [*y*+1, *z*]
Query 4: searching values in the range [*z*+1, 10240]

List **all possible** total number of read I/Os needed for the 4 scans, *with* the magic function. Please explain your answer.

4. We have a 32-integer array *A* in the main memory. Let cache size be 16 (integers), and cache line size be 4 (integers). Suppose that initially the cache is empty, and the cache replacement policy is the same as the one introduced in the lectures, i.e., first cached first evicted. Let the *f-trip* and *b-trip* scanning be the following.

```
f-trip(){
        for (int j=0; j<32; j++){
                Access A[j];// Access does not change the data
        }
}
b-trip(){
        for (int j=0; j<32; j++){
                Access A[31-j];
        }
}
```

If we need to do 99 scans of the array, and we can select each scan to be either f-trip or b-trip. Please give one best selection strategy that gives the minimum number of misses and explain your answer. Please also compute the number of cache hits and cache misses in the best strategy.