The University of New South Wales
COMP9315 DBMS Implementation
Sample Exam

[Instructions] [Notes] [PostgreSQL] [C]
[Q1] [Q2] **[Q3]** [Q4] [Q5] [Q6] [Q7] [Q8]

Question 3 (12 marks)

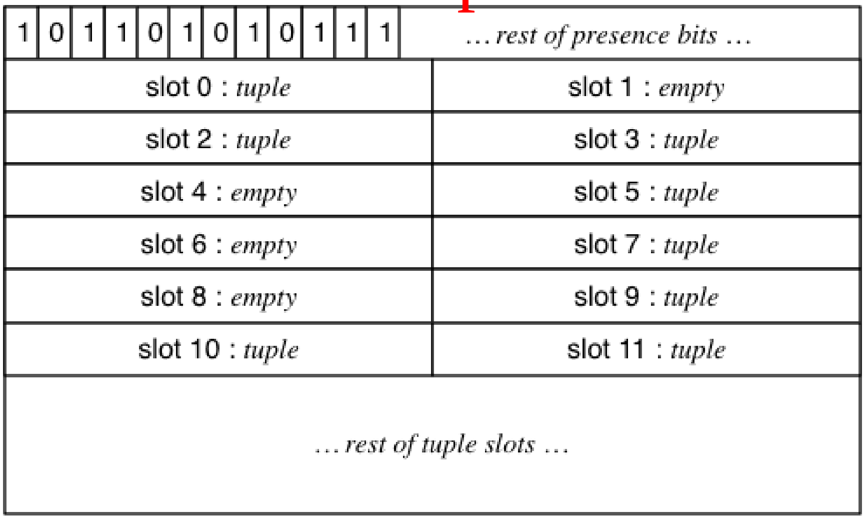
Consider the following query on the table from Question 2:

```
select distinct name from People;
```

Assume that the `People` table has the following characteristics:

P	4096	size of pages/buffers in bytes
r	1,234,567	total number of tuples
R	32	size of each tuple in bytes (fixed-sized records)
B	101	number of available memory buffers

Each page contains a page directory as well as the tuples. The page directory consist of 1 presence **bit** for each tuple, and the bits are packed as tightly as possible into a sequence of bytes at the start of the page. The remaining space is used for tuples. All pages, even those in intermediate files, have the same page structure. The following diagram shows what such a page might look like after some tuples have been inserted and others removed:



For this problem, you can assume that tuples are added consectively starting from slot one, so the first *k* presence bits will be one if there are *k* tuples in the page, and the remaining presence bits will be zero.

For the above scenario, you should:

a. Calculate the number of tuples per page

b. Calculate the total number of pages for the `People` table

c. Calculate the cost of duplicate removal using sorting

d. Calculate the cost of duplicate removal using hashing

e. What would be the effect on the hashing approach if only 91 buffers were available? (you do not need to do a complete detailed cost calculation for this part; just describe the effect in words)

The costs should give the total number of page reads/writes and include all page reads and page writes *except* the page writes for the final output stage. Note the because we are not considering the size of the final output, the costs do not rely on the percentage of duplicates.

You can ignore any effects from buffer management and assume that all buffers will be used in the most effective manner possible. For the hashing case, assume that the hash functions partition the tuples uniformly. Your cost calculations must show clearly the number of pages in all intermediate files.

Show all working.

Instructions:

- Type your answer to this question into the file called `q3.txt`
- Submit via: **`give cs9315 sample_q3 q3.txt`**
  or via: Webcms3 > exams > Sample Exam > Submit Q3 > Make Submission

End of Question