Ethan Wang
305319001

CS 143 HW7

1. R(A, B, C)
   - 8000 tuples
   - disk block 4,000 bytes
   - tuple is 40 bytes
   - each disk block is full

a) $\frac{4000}{40}$ = 100 tuples per block

$\frac{8000}{100}$ = **80 blocks**

b) Find the smallest value for M that we can use.

To finish in two passes:

# of sorting runs ≤ available memory blocks − 1

$\lceil 80 / M \rceil \leq M - 1$

M = 10 blocks

2. R(A, B) 1000 blocks

S(A, C) 100 blocks

3300 I/Os

Total cost for Hash Join: $3(b_R + b_S)$

(bucketizing + join steps)

- Entire bucket of smaller table fits in main memory
  - one block reserved for larger table
  - one block reserved for writing output
- Size of S bucket: $\lceil \frac{100}{M-1} \rceil$
  - Size of S bucket ≤ M − 2
    - ↳ $\lceil \frac{100}{M-1} \rceil \leq M - 2$

M = 12 blocks

3. R(A, B, C)
S(B, C, D, E)

- 500 tuples of R per block on disk
- |R| = 750,000
- 100 tuples of S per block on disk
- |S| = 250,000
- For every R tuple, roughly 5 tuples in S with R.B = S.B

① To read R (first two lines): $\frac{750,000}{500} = 1,500$

Blocks of S to read: $750,000 \times 5 = 3,750,000$

Total 3,751,000

- 5000 tuples in S with R.C = S.C, stored sequentially

② To read R (first two lines): $\frac{750,000}{500} = 1500$

Blocks of S to read: $750,000 \times \left(\frac{5000}{100}\right) = 37,500,000$

(sequential access)

The first plan has 10 times less I/Os than the second plan. The second plan should be chosen if random I/O is over 10 times slower than sequential access. Otherwise, the second plan will complete faster despite needing more I/O.