CPSC 404 Quiz 4 Question

Zhe Li

November 2020

1 Sort Merge Join Question

Given the following relations:
Voter(voterID, name, address, regisDate)
Vote(voterID, candidate, date, state)
Voter table has 60000 pages, and Vote table has 80000 pages. The primary keys
of both relations are underlined. Available buffer pool has $B = 400$ pages.

1.1 Naive Implementation

Using a naive implementation, estimate the I/O cost of computing the join of
the two relations using SMJ. Show your steps. Use SSLs of size B pages each.

Solution:
Sort Voter:

$$\lceil \frac{60000}{400} \rceil = 150 SSLs$$

Phase I takes $2 \times 60000 = 120000$ I/Os.
Because $150 \leq 400$, these SSLs can be merged in one pass.
Phase II takes $2 \times 60000 = 120000$ I/Os.

Sort Vote:

$$\lceil \frac{80000}{400} \rceil = 200 SSLs$$

Phase I takes $2 \times 80000 = 160000$ I/Os.
Because $200 \leq 400$, these SSLs can be merged in one pass.
Phase II takes $2 \times 80000 = 160000$ I/Os.
Merge Join takes $60000 + 80000$ I/Os.
Therefore, total SMJ cost $= 5 \times (60000 + 80000)$ I/Os.

1.2 Optimization Thoughts

The above implementation is not optimized and is wasting many I/Os. Please
state how to improve the implementation, and what is the prerequisite of making

such improvement.

Solution: We can improve the naive implementation by replacing the merge part of each table in their phase II of sorting by directly merging to the resulting table. To make this improvement, we require the number of SSL is less than the buffer size.

1.3 Improved Implementation

Use an improved implementation to make SMJ as efficient as possible and estimate the resulting I/O cost. Show your steps. Still use SSLs of size B pages each.

Solution:
To merge the two relations in one pass, we require $150 + 200 \leq 400$. There is true in our case. Therefore, we can cut down the cost from $5 \times (60000 + 80000)$ to $3 \times (60000 + 80000)$