

Assignment 2

COP5711

Due: 07/03/2024

This assignment is to demonstrate your understanding of the parallel algorithms discussed in class. Each of these parallel algorithms consists of a number of computation phases. A new phase cannot start until every processing node completes the current phase. We discussed a PowerPoint slide in class how to estimate the computation cost for the Hash Phase of the Grace Algorithm. We clarified that this computation cost for a given phase is determined by the slowest processing node (last to finish its workload), not the sum of the local computation costs of all the processing nodes. The estimation of the other phases of this and other multi-phase algorithms is the same. You need to understand each phase of these algorithms in order to do such estimation. Do the following exercises to demonstrate your knowledge of these techniques.

1) (30 pts.) We apply the GRACE algorithm to perform $R \bowtie S$ on a small shared-nothing system with four processing nodes (PNs). R has 4,000 pages and S has 8,000 pages. Each relation is evenly divided among the four PNs. Thus, each PN has 3,000 pages of tuples. The Hashing Phase results in data skew as follows:

- 50% of the data in the first 8 bucket pairs: $R_0/S_0 - R_7/S_7$
- 20% of the data in the second 8 bucket pairs: $R_8/S_8 - R_{15}/S_{15}$
- 15% of the data in the third 8 bucket pairs: $R_{16}/S_{16} - R_{23}/S_{23}$
- 15% of the data in the fourth 8 bucket pairs: $R_{24}/S_{24} - R_{31}/S_{31}$

For each of the three parallel phases, estimate the read cost, the write cost, and the total computation cost. Show and explain the derivation of your mathematical analysis.

2) (70 pts.) We apply the ABJ algorithm for the same setting described in Question 1. Estimate the read cost, the write cost, and the total computation cost for each of the four phases of ABJ. Show and explain the derivation of your mathematical analysis.