Uni Freiburg, Web Science Group
Prof. Peter Fischer
Systems Infrastructure for Data Science - Winter 2014/15

Exercise Sheet #9: Parallel Databases

January 8, 2015

A. Define the terms scale-up and speed-up. Why is a shared-nothing architecture attractive for parallel database systems?

B. Consider the following relations from an employee database :

`EMP(ENO, Ename, Title)`

`PROJ(PNO, Pname, Budget, Category)`

`ASSIGN(ENO, PNO, Role, Duration)`

Suppose that the database is partitioned and stored across ten nodes as follows:

- The EMP relation is partitioned to three fragments E1, E2, E3, based on a hash function applied to the attribute ENO and stored at sites S1,S2, S3 respectively.
- The PROJ relation is partitioned to three fragments P1, P2, P3, based on a hash function applied to the attribute `Category` and stored at sites S4,S5, S6 respectively.
- The ASSIGN relation is partitioned to four fragments A1, A2, A3, A4 based on a hash function applied to the attribute ENO and stored at sites S7,S8, S9 and S10 respectively.

Assume that the nodes S1 to S10 are pairwise connected with point-to-point links (i.e. simultaneous broadcast is not possible). You are given the following information about the relations: size(EMP)=300, size(PROJ)=600 and size(800). Consider the various parallel join algorithms discussed in the class. Which algorithm would you use for performing a join of EMP with ASSIGN based on the attribute ENO? Which algorithm would you use for joining PROJ with ASSIGN based on the attribute PNO? Give reasons for your answer.