Institut for Matematik og Datalogi
Syddansk Universitet

November 18, 2013
JFB

Introduction to Computer Science
E13 – Discussion Sections 10 – Week 48

Bring your computer to your discussion section.

1. Consider the following problem from last week:

   Assume that the database relations A and B each are stored as sequential files, but now no longer ordered on the X attribute.

   Describe an algorithm based on nested loops for executing the statement

   $$C \leftarrow \text{JOIN } A \text{ and } B \text{ where } A.X = B.X$$

   How many comparisons between tuples are performed (as a function of $|A|$ and $|B|$, the numbers of tuples in each relations)?

   New part: Describe how to speed up the algorithm by first using hashing on each relation.

2. (Work in groups of 2 or 3.) Write a program to compute the probability of at least one collision when hashing is used with m records and n buckets. (See the calculation on page 427 of your textbook and generalize it.) Assume that the the hash function spreads data out essentially randomly. Use your program to answer problem 7 on page 428 and problem 57 on page 436. How did you use your program?

3. Question 6 on page 428.

4. Explain how a poorly chosen hash function can result in a hash storage system becoming little more than a sequential file.

5. Suggest some possible designs for hash functions, for example taking the middle half of the bits and squaring modulo the number of bins. Can you detect any weaknesses?

6. Read about SHA-1 and SHA-3 (maybe on Wikipedia). Why do these look like better hash functions than the ones you proposed? Is there anything about them which is less good than what you proposed? These are used for cryptographic and security purposes, which we will discuss later in this course.