CS5225 Parallel & Distributed Databases

Take-Home Assignment 1 (Due: 17 Sept 2007)

1. Consider an Engineering database that maintains three tables:

    EMP (<u>ENO</u>, NAME, TITLE, SAL)
    PROJ (<u>PNO</u>, PNAME, BUDGET, LOC)
    ASG (<u>PNO, ENO</u>, RESP, DUR)

EMP stores employee information, including the name, title and salary of the employee. PROJ stores project information, including the name, budget and location of the project. ASG keeps information about the assignment between projects and employees. RESP is the responsibility of the employee, while DUR is the duration that the employee works on the project. ENO is the employee id. PNO is the project id. The keys are underlined.

The following predicates appear in the majority of queries.

LOC = "New York", LOC = "Montreal", LOC = "Paris"
SAL > 5000, SAL $\leq$ 5000

a) Write down the fragments if we perform primary fragmentation on EMP and PROJ, given the predicates above.
b) Write down the fragments if we perform derived horizontal fragmentation on ASG based on the fragments of PROJ in a).
c) Write down the fragments, if we further fragment EMP based on fragmented ASG. Is this kind of fragmentation (on the three tables) a good one? Why? (Hint: Analyze the desired properties of fragmentation.)

2. Assume that a) the cost of storing fragments at sites is negligible, and b) the cost of accessing a local fragment is 0 unit and c) the cost of READing a remote fragment is 1 unit and the cost of WRITEing a remote fragment is 2 units.

   The access profiles of six applications are as follows:

Applications	Fragment											
	1		2		3		4		5		6	
	R	W	R	W	R	W	R	W	R	W	R	W
1	-	30	-	10	10	-	30	-	-	-	50	10
2	20	-	40	-	-	30	-	50	10	40	-	-
3	10	20	-	50	-	-	20	-	30	-	30	-
4	20	10	20	-	-	50	10	-	-	10	-	30
5	-	-	10	10	100	-	-	20	50	80	10	40
6	10	60	10	-	-	-	20	40	-	-	-	-

   The frequency of applications on sites is as follows:

Applications	Sites			
	A	B	C	D
1	20	-	-	30
2	-	10	10	-
3	10	30	-	20
4	-	-	50	20
5	20	10	-	-
6	-	10	20	-

   Allocate the six fragments to the four sites using the “best-fit” method.

3. Consider the engineering database in Q1. Draw the reduced query trees for the following fragments:

a) Assume that relation PROJ is horizontally fragmented in

$$PROJ_1 = \sigma_{PNO \leq "P2"}(PROJ)$$

$$PROJ_2 = \sigma_{PNO > "P2"}(PROJ)$$

Reduce the following query to fragments:

```
SELECT BUDGET
FROM   PROJ, ASG
WHERE  PROJ.PNO = ASG.PNO
AND    ASG.PNO = “P4”
```

b) Assume that ASG is horizontally fragmented as

$$ASG_1 = \sigma_{PNO \leq "P2"}(ASG)$$

$$ASG_2 = \sigma_{"P2" < PNO \leq "P4"}(ASG)$$

$$ASG_3 = \sigma_{PNO > "P4"}(ASG)$$

Reduce the following query to fragments (PROJ is fragmented as in 3.1):

```
SELECT RESP, BUDGET
FROM   PROJ, ASG
WHERE  PROJ.PNO = ASG.PNO
AND    PNAME = “CAD/CAM”
```

c) Assume that ASG is now indirectly fragmented as

$$ASG_1 = ASG \ \triangleright\!<_{PNO} \ PROJ_1$$

$$ASG_2 = ASG \ \triangleright\!<_{PNO} \ PROJ_2$$

And EMP is vertically fragmented as

$$EMP_1 = \Pi_{ENO,ENAME}(EMP)$$

$$EMP_2 = \Pi_{ENO,TITLE}(EMP)$$

Reduce the following query to fragments (PROJ is fragmented as in 3.1):

```
SELECT ENAME
FROM   EMP, PROJ, ASG
WHERE  PROJ.PNO = ASG.PNO
AND    EMP.ENO = ASG.ENO
AND    PNAME = “CAD/CAM”
```

4. Consider the multi-join query $R(A,B) \bowtie S(A,C) \bowtie T(A,D) \bowtie U(A,E)$. Each of the relations is stored on a different site. The goal is to minimize the amount of data transmitted over the network. The size of relations are as follows:

Relation	Number of tuples	Bytes per tuple
R	3000	40
S	2000	30
T	2000	70
U	2500	40

Assume that for any relations X and Y (|X| denotes the number of tuples in X), we have the following estimates:

i) $|X \bowtie Y| = 0.5 \times \max(|X|,|Y|)$

ii) $|X \rhd< Y| = 0.5 \times \min(|X|,|Y|)$

iii) Column A is 10 bytes. The tuple size of $|X \bowtie Y|$ is the sum of the tuple size of X and Y minus the common attribute.

a) What is the minimal cost if running all computation on only one site (in terms of bytes transmitted)? On which site?

b) Given the initial plan in a), use hill climbing optimization strategy to find the final plan and the total cost. Only join operations are considered.