Name: _____

Student ID: _____

Department of Computer Science

Final Exam, CS 4411a — Databases II

Prof. S. Osborn
April 22, 2010
3 Hours

No aids. No electronic aids
Answer all questions on the exam page
This paper contains 19 pages; the last page is for rough work.

Question	Maximum	Your Mark
1	25	
2	7	
3	18	
4	14	
5	15	
6	14	
7	18	
8	16	
9	12	
10	10	
11	15	
Total	164	

1. (25 marks) For each of the following statements, state whether it is **true** or **false**. If it is false, **correct the statement** without changing the underlined text. (Note: there might be more than one correction to make!)

   (a) Shredding is a technique used to convert pointers in XML databases when data is brought into main memory. **T** **F**

   (b) Indexing can be used to speed up query processing in XML databases. **T** **F**

   (c) The sort-merge join can join two unsorted relations of n tuples in O(n) page fetches. **T** **F**

   (d) Semijoins are used in one possible execution strategy for distributed relational database joins. **T** **F**

   (e) Path expressions are used in distributed relational databases to drill down into the fragments. **T** **F**

   (f) Concurrency Control looks after the A part of ACID. **T** **F**

   (g) Recovery looks after the A part of ACID. **T** **F**

   (h) Wait-Die is a concurrency control scheme. **T** **F**

(i) Wound-Wait is used with centralized databases. **T** **F**

(j) Write Ahead Logging writes to the database and then to the log. **T** **F**

(k) Shadow paging is a distributed recovery technique which may require redo and may require undo. **T** **F**

(l) Multiple granularity locking in relational databases allows the database to avoid phantom locks. **T** **F**

(m) Two-phase commit is a locking mechanism used in centralized database systems. **T** **F**

(n) The access control provided by DB2 is an example of Mandatory Access Control. **T** **F**

(o) In the SeaView model for mandatory access control, one rule is the security level of the database must be greater than the security level of all the relations in the database. **T** **F**

(p) Polyinstantiation occurs in a relational database when we have too many tuples for the disk. **T** **F**

(q) One of the requirements for privacy is that personal data can only be collected for a predefined, legitimate purpose. **T** **F**

2. (7 marks)

   (a) In the following, assume that relation R(A, B, C) with primary key A, has a clustering index on attribute A with depth 5, and an index on attribute B with depth 4. For which of the following can the time to execute the query, in terms of local disc accesses, be proportional to O(the depth of the index), or essentially a constant? CIRCLE ALL STATEMENTS WHICH ARE CORRECT.

   i) $\sigma_{A=7}(\mathrm{R})$

   ii) $\sigma_{A=7 \text{ and } B \neq \text{"}Smith\text{"}}(\mathrm{R})$

   iii) $\sigma_{A \neq 7}(\mathrm{R})$

   iv) $\pi_{B}(\mathrm{R})$

   (b) When an object is brought into main memory in an OODB, which of the following might be required? CIRCLE ALL STATEMENTS WHICH ARE CORRECT.

   i) references to other objects will need to be converted to their main memory format

   ii) record header formats might need to be changed

   iii) string formats might need to be changed

3. (18 marks) The data for this question is the data from your assignment 2. Here is a small sample of the two files, authors.xml and paperdata.xml, respectively:

```
<authorRoot>
<author>
   <Fname>P.-A.</Fname>
   <Lname>Larson</Lname>
   <Location>Microsoft</Location>
</author>
<author>
   <Fname>Peter</Fname>
   <Lname>Boncz</Lname>
   <Location>CWI Amsterdam</Location>
</author>
...
</authorRoot>

<documents>
<dblp>
     <inproceedings key="conf/sigmod/ChapmanJ09" mdate="2009-07-01">
          <author>Adriane Chapman</author>
          <author>H. V. Jagadish</author>
          <title>Why not?</title>
          <pages>523-534</pages>
          <year>2009</year>
          <booktitle>SIGMOD Conference</booktitle>
          <ee>http://doi.acm.org/10.1145/1559845.1559901</ee>
          <crossref>conf/sigmod/2009</crossref>
          <url>db/conf/sigmod/sigmod2009.html#ChapmanJ09</url>
          <cite>conf/ipaw/ChapmanJ08</cite>
         ...
    </inproceedings>
    <article key="journals/cacm/Codd70" mdate="2003-11-20">
          <author>E. F. Codd</author>
          <title>A Relational Model of Data for Large Shared Data Banks.</title>
          <pages>377-387</pages>
          <year>1970</year>
          <volume>13</volume>
          ...
     </article>
...
</dblp>
</documents>
```

You may assume the following context file is available for Galax:

```
declare variable $authors := doc("authors.xml");
declare variable $papers := doc("paperdata.xml");
```

(a) (3 marks) Give an XPath expression to return the title of all papers which have been published in 2008.

(b) (7 marks) Give an XQuery FLWOR query to return the titles of all publications where the year is 2008 and at least one of the authors has a first name of "Frank". Make your answer a valid XML document.

(c) (8 marks) Give an XQuery FLWOR query to create, from the authors data, a list of all authors organized by their location. Each location should only occur once. Sort the output by the location. Your output should be formatted as follows:

```
<authorsByLoc>
   <Place>
      <Location>Microsoft></Location>
      <authorList>
         <author>
            <Fname>P.-A.</Fname>
            <Lname>Larson</Lname>
         </author>
         <author>
            <Fname>Jim</Fname>
            <Lname>Gray</Lname>
         </author>
         ...
      </authorList>
   </Place>
   <Place>
      <Location>CWI Amsterdam</Location>
      <authorList>
       ...
      </authorList>
   </Place>
...
</authorsByLoc>
```

4. (14 marks) Consider the following relations which hold some of the data contained in the XML documents of the previous question:

```
Papers(DocId, Title, WherePub, Year)   primary key is {DocID}
Authors(FName, LName, Location)        primary key is {FName, LName}
PaperAuths(FName, LName, DocId, position)  primary key is {FName, LName, DocID}
```

Also consider the following statistics on these tables:

```
For Papers:                              For Authors:
No. of tuples in Papers: 10              No. of tuples in Authors: 40
No. of bytes in DocID: 8                 No. of bytes in FName: 20
No. of bytes in Title: 100               No. of bytes in LName: 20
No. bytes in WherePub: 50                No. of bytes in Location: 30
No. bytes in Year: 4                     Distinct values in FName: 35
Distinct values in DocId: 10             Distinct values in LName: 39

For PaperAuths:
No. of tuples in PaperAuths: 20
No. of bytes in FName: 20                No. of bytes in LName: 20
No. of bytes in DocID: 8                 No. of bytes in position: 2
Distinct values of DocId: 10             Distinct values of FName: 19
Distinct values of LName: 19
```

(1 mark each unless otherwise stated)

(a) How many bytes per tuple are in relation Papers?

(b) How many bytes per tuple are in $\sigma_{DocID=12345}$(Papers)?

(c) How many tuples are in $\sigma_{DocID=12345}$(Papers)?

(d) How many distinct values of attribute Title are in $\sigma_{DocID=12345}$(Papers)?

(e) How many distinct values of attribute Title are in $\sigma_{DocID\neq12345}$(Papers)?

(f) How many bytes per tuple are in $\pi_{FName,LName}$(Authors)?

(g) How many tuples are in $\pi_{FName,LName}$(Authors)?

(h) How many distinct values of FName are in $\pi_{FName,LName}$(Authors)?

(i) (2 marks) How many bytes per tuple are in Authors $\bowtie$ PaperAuths?

(j) (2 marks) How many tuples are in Papers $\bowtie$ PaperAuths?

(k) (2 marks) How many distinct values of DocID are in Papers $\bowtie$ PaperAuths?

5. (15 marks) Consider these relations again:

```
Papers(DocId, Title, WherePub, Year)   primary key is {DocID}
Authors(FName, LName, Location)        primary key is {FName, LName}
PaperAuths(FName, LName, DocId, position)  primary key is {FName, LName, DocID}
```

Furthermore, assume we have created fragments so that we can store information about old papers on one site, and information about newer papers on another site in a distributed database.

Old = $\sigma_{Year<1995}$(Papers)
New = $\sigma_{Year>=1995}$(Papers)
OldPaperAuths = PaperAuths $\ltimes$ Old
NewPaperAuths = PaperAuths $\ltimes$ New

(a) (2 marks) What attributes are included in the fragment NewPaperAuths?

(b) (3 marks) We want to execute the following SQL query over the distributed data above:

```
Select p.title
From Papers as p, PaperAuths as a
Where p.DocID = a.DocId and a.LName = "Boncz"
```

Translate the query to a relational algebra on global relations. Use a join operator if appropriate.

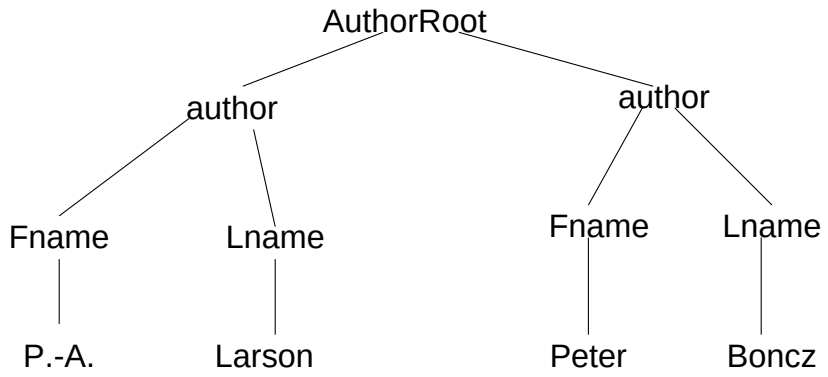
(c) (4 marks) Show the query tree corresponding to your algebra query just above. (put your answer on the next page.)

(d) (6 marks) Replace any relations in your tree with the fragments defined above. Also, express the relations as qualified relations of the form

`[R:F]`

where F contains any predicates you know to be true about the fragment. After doing this, perform any further optimizations possible on your algebra tree.

6. (14 marks) Consider the following XML Document:



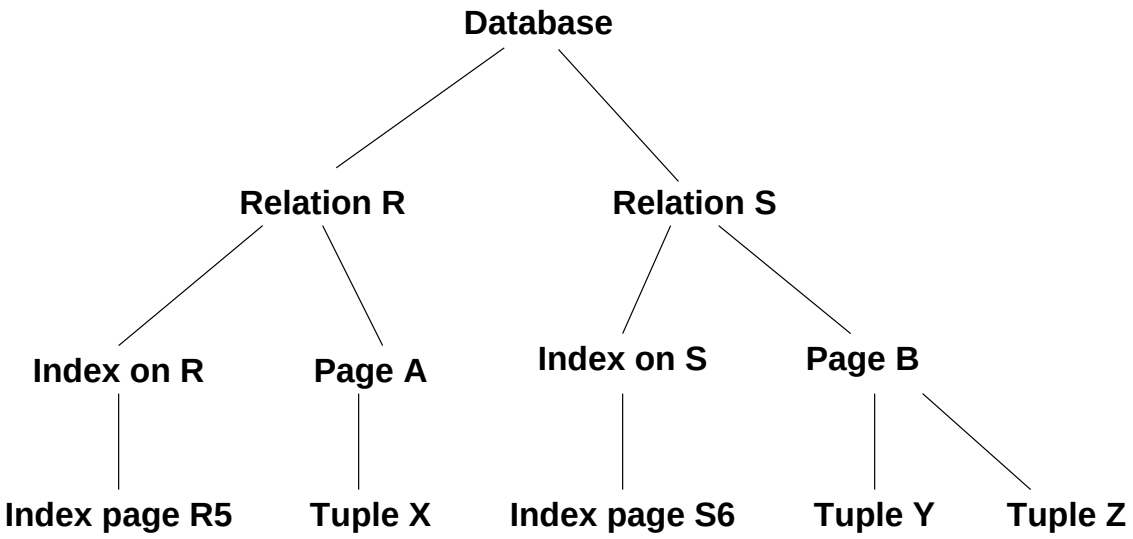
(a) (7 marks) The following table, called Node, represents one way of storing the XML in a relational database. Fill in the rest of the entries, including the one marked ?? (the column titled "Flag" can contain only Element, Attribute or Value).

Node					
Label	Start	End	Level	Flag	Value
AuthorRoot	1	??	0	Element	null

(b) (7 marks) Give an SQL expression which extracts from the above table, Node, the data corresponding to the following XPath expression
/author[/Lname = "Gray"]/Fname

7. (18 marks) Put your answer to this question on the next page.

   Consider the following (simplified) tree of granules for a relational database:



   (a) (6 marks) Transaction T1 wants to do the following in this order:

   read all the tuples on page A of relation R, not touching the index
   read index page S6 searching for tuple Z, in relation S
   then delete tuple Z on page B of relation S
   then update index page S6

   Show a sequence of lock and unlock instructions which follow the rules for multiple granularity locking, and which allows this transaction to execute.

   (b) (6 marks) Transaction T2 wants to do the following in this order:

   read index page R5 looking for tuple X in relation R
   update an attribute value in tuple X of relation R
   read index page S6 looking for tuple Y
   read tuple Y on page B of relation S

   Show a sequence of lock and unlock instructions which follow the rules for multiple granularity locking, and which allows this transaction to execute.

   (c) (6 marks) Show a valid interleaving of T1 and T2 such that there are no deadlocks. **Have as much interleaving in your schedule as you can manage**. Use the following lock compatibility matrix. Put your answer on the next page.

Lock Compatibility Matrix

	Already Granted					
Requested	None	IS	IX	S	SIX	X
IS	yes	yes	yes	yes	yes	-
IX	yes	yes	yes	-	-	-
S	yes	yes	-	yes	-	-
SIX	yes	yes	-	-	-	-
X	yes	-	-	-	-	-

Put answer to part (a) ↓	Put answer to part (c) ↓	Put answer to part (b) ↓

8. (16 marks) Suppose the following timestamps are recorded for the following data items:

data item	read TS	write TS
w	15	11
x	11	16
y	16	18
z	10	11

For each of the following operations, state whether or not it would be allowed using the revised timestamp ordering algorithms for concurrency control, and, if allowed, **whether or not any timestamps change**. Assume all these operations are independent of each other (i.e., they all refer to the original timestamps for w, x, y, and z). If any of them uses the Thomas Write Rule, say that.

(a) read y on behalf of transaction T whose timestamp is 15.

(b) write y on behalf of transaction T whose timestamp is 15.

(c) read z on behalf of transaction T whose timestamp is 15.

(d) write z on behalf of transaction T whose timestamp is 15.

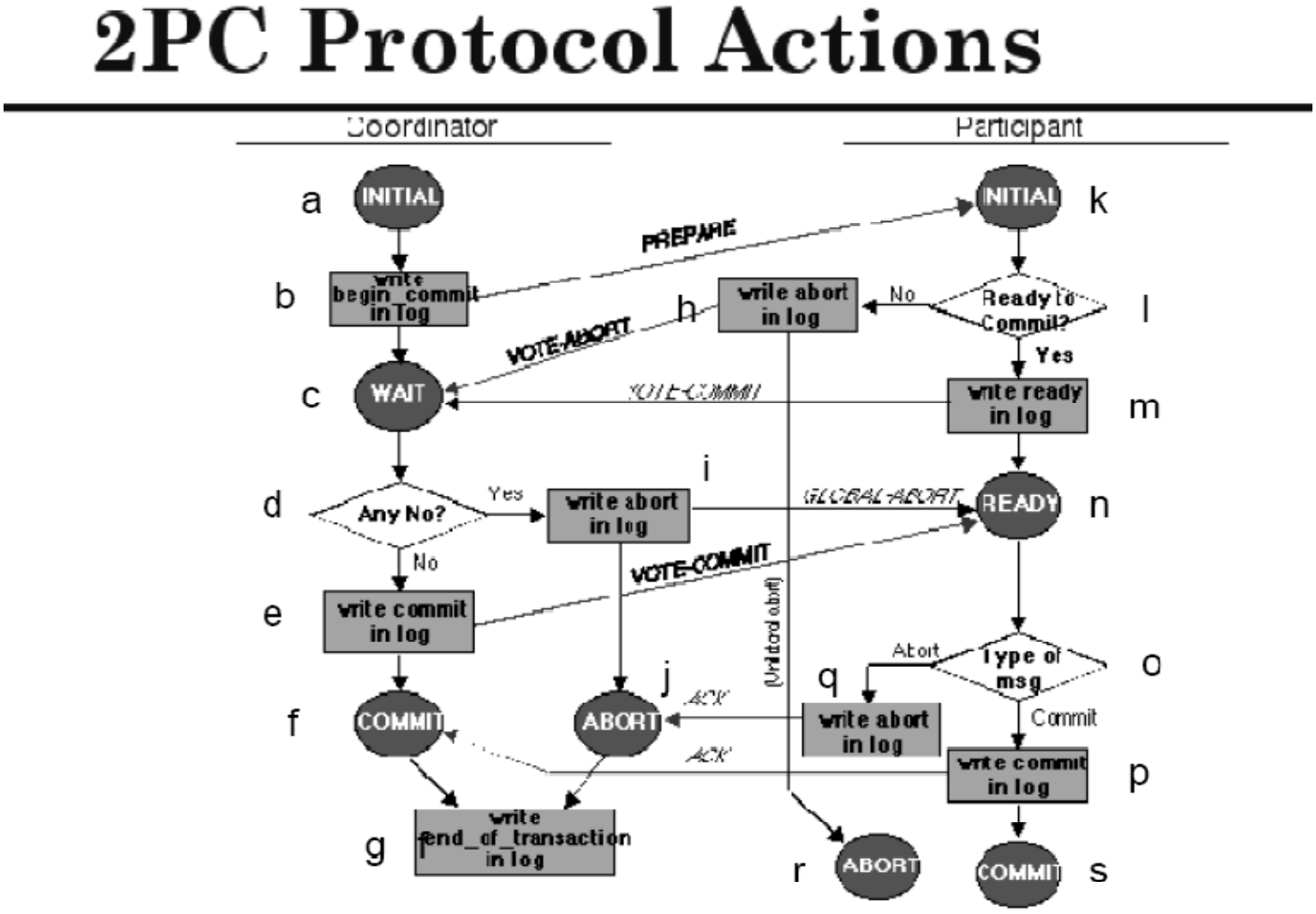
(e) read x on behalf of transaction T whose timestamp is 15.

(f) write x on behalf of transaction T whose timestamp is 15.

(g) read w on behalf of transaction T whose timestamp is 15.

(h) write w on behalf of transaction T whose timestamp is 15.

9. (12 marks) Consider the diagram for Two-Phase commit reproduced here from Özsu's web page:



For each of the following, circle the correct answer.

(a) (1 mark) The coordinator and participants must have completed all their reads and writes to decide to vote to commit.

i. True

ii. False

(b) (1 mark) It is possible for one of the participants to decide to commit, when the coordinator has decided to abort.

i. True

ii. False

(c) (1 mark) A participant P might remain blocked if some other participants crash but the coordinator and P do not crash.

i. True

ii. False

(d) (1 mark) The coordinator might become blocked.

i. True

ii. False

(e) (1 mark) Participants are said to be uncertain after they receive the message containing the result of the vote.

i. True

ii. False

(f) (1 mark) It is possible for a participant to be in its READY state and the coordinator to be in its COMMIT state at the same time.

i. True

ii. False

(g) (1 mark) It is possible for the coordinator to be in its WAIT state and one of the participants to be in the ABORT state at the same time.

i. True

ii. False

(h) (1 mark) If the coordinator crashes, the participants can still decide to commit if (**Choose one**):

i. At least one of them is in the ABORT state

ii. All of them are in the INITIAL state

iii. At least one of them is in the READY state

iv. At least one of them is in the COMMIT state

(i) (4 marks) Give the sequence of nodes a participant passes through if the final decision is to abort.

10. (10 marks) Consider this multilevel relation for the SeaView model, where the security levels are TS > S > C > U. The primary key of the underlying relation is {ItemNo}.

ItemNo	C_{ItemNo}	ItemName	$C_{ItemName}$	Cost	C_{Cost}	SellingPrice	$C_{SellingPrice}$	TC
123	U	nails	U	2.50	U	2.75	U	U
123	TS	spyphone	TS	2.50	TS	2.75	TS	TS
123	S	phone	S	1.75	S	1.75	TS	TS

Some of the SeaView rules are given here:

The Tuple Class Property says that, for all i, the class of attribute i must be $\leq$ the class of the whole tuple, TC.

Relation Class Integrity: The access class of the relation scheme must be dominated by ($\geq$) the access class of the lowest data that can be stored in the relation.

Polyinstantiation integrity: given primary key attributes AK and key class CK, for each attribute Ai not in AK, there is a functional dependency: AK,CK,$C_i \rightarrow A_i$

(a) (4 marks) Give the S-instance of this relation, i.e. all data readable by an S user.

(b) (3 marks) Can a user cleared at level C insert the following data into the above relation: item number 123 with item name "hammer", cost 2.50 and selling price 1.50? If so, show all the values for this tuple in all the columns above. If not say why.

(c) (3 marks) Can a user cleared at level S insert the following data into the above relation: item 123 with item name "hammer", cost 1.75 and selling price 2.50? If so, show all the values for this tuple in all the columns above. If not say why.

11. (15 marks) Answer **Three** of the following (only the first 3 answers will be marked).

    (a) In a relational database, the schema is stored as tables in the database. Suppose an XML database has a schema expressed using XML Schema (i.e. in an XML format). What are some options for the XML database package to store such a schema?

    (b) Suppose we have relation R(A, B, C) stored on one site and relation S(B, D, E) stored on another site. Describe the semijoin algorithm for computing the join of R and S, so that the answer ends up at the site of R.

    (c) Explain why it is easier to insert new nodes into an XML document whose node labels are given by Dewey numbers, than it is with the (start, end) labels used in question 6 on this exam.

    (d) One of the levels of isolation offered in DB2 is degree 2 level (degree 3 is called serializable) where only write locks are two-phase but read locks are also used. For what types of transactions is this degree of isolation useful?

    (e) Explain what phantom locks are and how they can be avoided.

    (f) Explain what phantom deadlocks are and how they can be avoided.

(for rough work or answers to Question 10)