Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

Cloud Computing (BUCI029H7)

CREDIT VALUE: 15 credits

Date of examination: Thursday, 21st May 2015
Duration of paper: 10:00am – 12:00pm (2 hours)

RUBRIC

1. *This exam paper contains* six *questions each of which is worth* 20 *marks. Students should attempt to answer* **any five** *of them.*
2. *Only the first five answered questions will be marked. No mark will be awarded for answering additional questions.*
3. *The use of non-programmable electronic calculators is permitted, but programmable electronic devices such as smart phones must be switched off.*
4. *This paper is not prior-disclosed.*

1. **(20 marks)**

Give brief answers to the following questions.

(a) A data mining program consists of four consecutive parts, P1, P2, P3 and P4 with the percentages of runtime being 10%, 30%, 40% and 20% respectively on a single processor. It is known that P1 and P3 can be parallelised while P2 and P4 cannot. If the problem/dataset size is fixed, how much speedup can this program achieve at most through parallel computing, according to Amdahl's Law? If the problem/dataset size can be arbitrarily large, how much speedup can this program achieve at most through parallel computing, according to Gustafson's Law? (5 marks)

(b) What does SPMD stand for? What does it mean in the context of parallel computing? Does it belong to data parallelism or task parallelism? (5 marks)

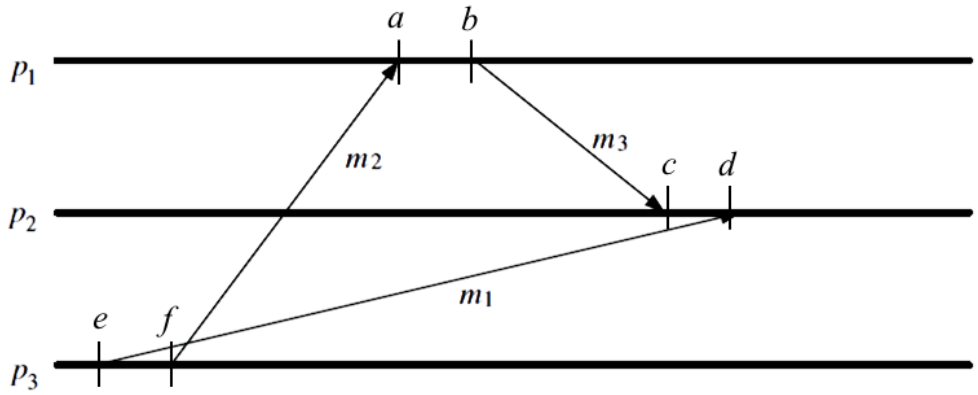
(c) What is a race condition? What is a deadlock? What is a livelock? (5 marks)

(d) What is eventual consistency? Why don't we insist on strong consistency in all distributed systems? (5 marks)

2. **(20 marks)**

Give brief answers to the following questions.

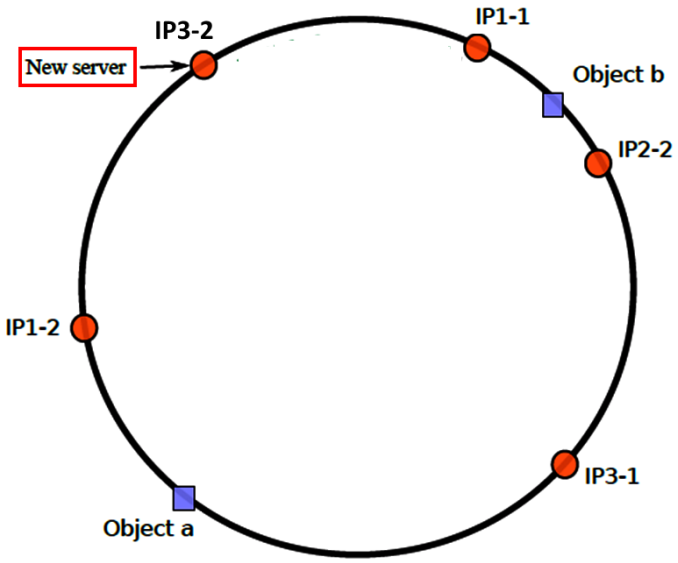
(a) What are the Lamport timestamps of the events a, b, c, d, e and f respectively in the following space-time diagram? (5 marks)



(b) What is the Chandy-Lamport snapshot algorithm used for? How many steps does it have? For a distributed system consisting of 4 processes that are fully connected with each other, how many messages would be exchanged to obtain a snapshot using this algorithm? (5 marks)

(c) In RESTful APIs, what HTTP methods should be nullipotent and what HTTP methods should be idempotent? Which HTTP method should be used to add new items to a collection at the given URI? (5 marks)

(d) Why is it better to use consistent hashing rather than standard hashing for distributed indexing? Ignoring data replication (for failure recovery and load balancing etc.), in the following schematic diagram of consistent hashing, where should "object a" be stored, and what will happen when a new server "IP3-2" joins the cluster? (5 marks)



3. **(20 marks)**

Give brief answers to the following questions.

(a) What is the order of execution of the following five MapReduce components in Hadoop: shuffling & sorting, reducer, partitioner, mapper, combiner? (5 marks)

(b) What are the advantages and disadvantages of using the "in-mapper combining" design pattern instead of combiners? (5 marks)

(c) What are the advantages and disadvantages of using the "stripes" design pattern instead of the "pairs" design pattern? (5 marks)

(d) When can we use map-side join? When can we use in-memory join? (5 marks)

4. **(20 marks)**

There is a large text file of computer science bibliography data held in an HDFS over a number of machines.
Each line of this file describes the details of one paper in the following format.
$authors \mid title \mid conference \mid year$
The different fields are separated by the | character, and the list of authors are separated by commas. An example line is given below.
D Zhang, J Wang, D Cai, J Lu | Self-Taught Hashing for Fast Similarity Search | SIGIR | 2010
You can assume that there are no duplicate records, and each distinct author or conference has a different name.
Write a MapReduce program (in pseudo-code) to calculate for each conference the average number of authors per paper since year 2000.
A combine function should be implemented to accelerate the computation.

5. **(20 marks)**

There is a large text file of computer science bibliography data held in an HDFS over a number of machines. It is the same file as described in the previous question.
Each line of this file describes the details of one paper in the following format.
$authors \mid title \mid conference \mid year$
The different fields are separated by the | character, and the list of authors are separated by commas. An example line is given below.
D Zhang, J Wang, D Cai, J Lu | Self-Taught Hashing for Fast Similarity Search | SIGIR | 2010
You can assume that there are no duplicate records, and each distinct author or conference has a different name.
Write a MapReduce program (in pseudo-code) to calculate for each author the total number of papers he/she has published.
The "in-mapper combining" pattern should be implemented to accelerate the computation.

6. **(20 marks)**

There is a large text file of computer science bibliography data held in an HDFS over a number of machines. It is the same file as described in the previous two questions.
Each line of this file describes the details of one paper in the following format.
$authors \mid title \mid conference \mid year$
The different fields are separated by the | character, and the list of authors are separated by commas. An example line is given below.
D Zhang, J Wang, D Cai, J Lu | Self-Taught Hashing for Fast Similarity Search | SIGIR | 2010
You can assume that there are no duplicate records, and each distinct author or conference has a different name.
Write a MapReduce program (in pseudo-code), using the "stripes" pattern, to calculate for each author the total number of papers he/she has co-authored with each different collaborator (i.e., researcher who has co-authored papers with him/her before). For example, the output corresponding to the author *D Zhang* could be as follows.
D Zhang — D Cai: 3, WS Lee: 13, J Lu: 6, J Wang: 8
A combine function should be implemented to accelerate the computation.

