Distributed Computing and Big Data

Chennai Mathematical Institute

Duration: 90 Mins. Max: 25 Marks.

Instructions

- This is a closed book exam.
- This is an individual task. Do not discuss with anyone.
- No electronic devices are allowed. Wherever heavy calculation is involved, you need not evaluate it to the final number unless it is explicitly asked for. For example, it is acceptable to leave the answer as $\frac{1}{1+\frac{5}{32}}$. You need not evaluate it to 0.865.
- Clearly mention your name and roll number in your answer sheet.

Section 1: Correct answers carry 1 mark each. Answer True/False. Wrong answers carry -0.5 marks each.

Question 1. In the model of the distributed system as discussed in the class, there is no common global memory. True/False?

Question 2. While computing average access time, head switching time is often considered negligible. True/False?

Question 3. Data lakes are schemaless. True/False?

Question 4. Grid computing infrastructure refers to the use of heterogeneous systems. True/False?

Question 5. Scalar time is strongly consistent. True/False?

Question 6. Hadoop uses a Write Once and Read Once model. True/False?

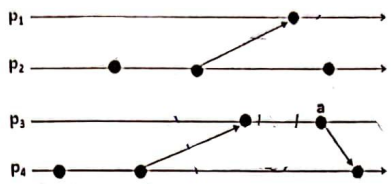
Question 7. Hadoop data nodes send periodic heartbeat signals and block reports to name node. True/False?

Section 2: Correct answers carry 2 marks each. No negative marks.

Question 8. Assume disk size = 512 GB, block size = 8 KB. How much space (in MB) will we need to store the free space bitmap?

Question 9. As per Amdahl's law, What is the best achievable speed up if only 25% of the job can be parallelized, and we have 4 processors?

Question 10. If we were to annotate the following space-time execution diagram with vector time stamps, how would we annotate the event marked as 'a'?



Question 11. For the same diagram given above, annotate the events in p_3 with scalar time.

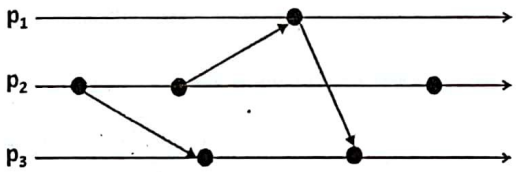
Question 12. For the same diagram given above, identify an inconsistent cut not involving the event marked as 'a' i.e., 'a' must be in the future of the cut.

Question 13. In the muddy children puzzle, as discussed in the class, what would the children say during the first and second rounds if n=4 and k=3? i.e., there are four children, and three of them have muddy forehead and to start with, they are told that at least one of them have muddy forehead. Assume that each child can only say 'yes', 'no' or 'dont know'. Use the following format to provide your answer.

```
Round1: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
Round2: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
```

Section 3: Question carries 3 marks. No negative marks.

Question 14. Annotate all the events in the following diagram with matrix time.



Question 15. You are given a large text file. You need to find all the distinct words that have no vowels. For example, if the input text file has the content, "my phone is ringing and ringing again", there are six distinct words (my, phone, is, ringing, and, again). The word without vowel is only one, "my". Therefore, the output should be a single word, "my".

Describe an approach using map reduce logic. No need to write any code.

CMI/2023/Mid-Semester Exam 24/02/2023

DISTRIBUTED COMPUTING AND BIG DATA

DURATION: 90 MINS. Chennai Mathematical Institute MAX: 25 MARKS.

Instructions

- You are allowed to carry a single A4 size paper with hand written contents. You should not exchange these cheat sheets with other students.
- This is an individual task. Do not discuss with anyone.
- Calculators (that do no access internet) are allowed. But no other electronic devices are allowed. Wherever heavy calculation is involved, you need not evaluate it to the final number unless it is explicitly asked for. For example, it is acceptable to leave the answer as $\frac{1}{1+\frac{5}{32}}$. You need not evaluate it to 0.865.
- Clearly mention your name and roll number in your answer sheet.
- Please submit your cheat sheet along with your answer sheet.

Section 1: Correct answers carry 1 mark each. Wrong answers carry -0.5 marks each.

Question 1. In 1928, IBM introduced a new version of the punched card with rectangular holes and 80 columns. It turned out to be one of IBM's most important technological innovations, propelling the company to the forefront of data processing. We classify such computers that used punched cards for storing programs as Von Neumann machines? True/False?

Question 2. IBM Summit, the fastest super computer of 2018, was capable of computing 200,000 trillion calculations per second. Oak Ridge National Laboratory has a super computer that could do 2000 teraFLOPS. IBM Summit is faster (purely based on the given data alone). True/False?

Question 3. The inode (index node) is a data structure in a Unix-style file system that describes a file-system object such as a file or a directory. True/False?

Question 4. A key benefit to STaaS is that you are offloading the cost and effort to manage data storage infrastructure and technology to a third-party cloud service provider. True/False?

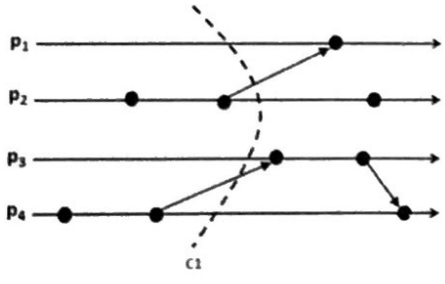
Question 5. Your company needs a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions. In such situations, your company can use a data lake. True/False?

Question 6. Ram bought a hard disk that has 500 GB of hard disk space and 5000 RPM rotation rate. Shyam bought a hard disk that has only 100 GB of hard disk space and it also has the same rotation rate. Ram's disk will have lesser rotational delay when compared to Shyam's disk. True/False?

Question 7. There is always one reducer in every map-reduce program. True / False?

Question 8. Map code is executed by the namenode in the hadoop cluster. True/False?

Question 9. The cut C1 is consistent. True/False?



Question 10. A solution to the General's Paradox is sending large number of messengers in each direction to guarantee a messenger gets through to the other side. True/False?

Question 11. Some technologists have estimated that all the words ever spoken by mankind would be equal to five Exabytes (an extraordinarily large unit of digital data). How much Gigabytes that (5 Exabytes) would be?

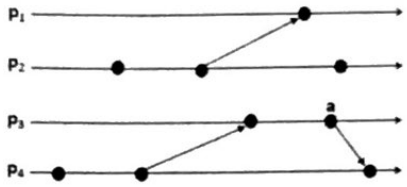
Question 12. The basic model of a distributed system, as discussed in the class, has a set of disconnected processes with no shared memory and no global clock. True/False?

Section 2: Correct answers carry 2 marks each. No negative marks.

Question 13. As per Amdahl's law, What is the best achievable speed up if only 20% can be parallelized, and we have 8 processors?

Question 14. Draw a space time execution diagram that provides an example of an inconsistent cut with exactly three processes, four events in the past and four events in the future.

Question 15. If we were to annotate the following space-time execution diagram with vector time stamps, how would we annotate the event marked as 'a'?



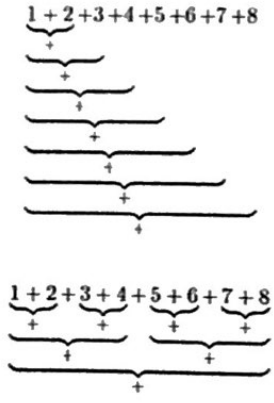
Question 16. In the muddy children puzzle, as discussed in the class, what would the children say during the first and second rounds if n=4 and k=2? i.e., there are four children, and two of them have muddy forehead and they are told that at least one of them have muddy forehead. Assume that each child can only say 'yes', 'no' or 'dont know'. Use the following format to provide your answer.

```
Round1: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
Round2: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
```

Question 17. Assume disk size = 256 GB, block size = 16 KB. How much space (in MB) will we need to store the free space bitmap?

Section 3: Question carries 3 marks. No negative marks.

Question 18. Our ability to write parallelizable programs decides the speed up we can achieve through scaling. Consider two programs to add large list of numbers. The first program adds the first two numbers, remembers the result, and adds that result to the next number. It continues doing this until the end of the list is reached. The second program adds two numbers at a time in parallel. It recursively does so until the final results are arrived at. The following figure explains their logic with an example of eight numbers.



Assume that the list is large and the numbers may be unordered. As per Gustafson's law, assuming each addition is an operation, how much speedup (approximately) can these programs achieve if there are four processors?

CMI/2022/Mid-Semester Exam Mar 2022

Chennai Mathematical Institute

Distributed Computing and Big Data for Uploading. Max Marks: 20. Duration: 60 Mins + 30 Mins

---

Roll No.: _____ Date: 17/03/2022

Name: _____

Instructions

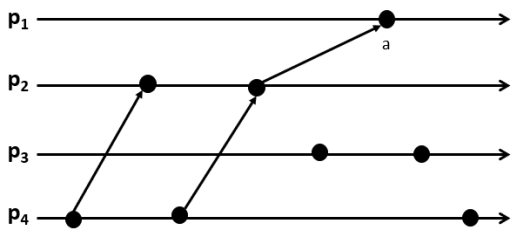
- Submit a single pdf file carrying your answers on moodle under "Mid Term" assignment. For any reason, if you cannot upload to moodle, email your work to vvtesh.cmi@gmail.com.
- A penalty of 1 mark applies for every two minutes of late submission.
- This is an individual task. Do not discuss with anyone.
- Please stop writing after 60 minutes. Uploading may take time. We apply late penalty strictly. If you make several submissions, the last submission will be taken for grading.

Section 1: Questions carry 3 marks each.

Question 1. Ramesh bought a hard disk with rotational delay of 3ms. With what RPM does the disk spin? If it had 20 sectors per track, what is its read time?

Question 2. Ram bought a new hard disk and configured it so that the number of bytes per inode is r. Prem too bought a new hard disk of same size and configured it to have the number of bytes per inode as p. Ram and Prem stored f number of files each in their respective disks. Given that $r > p$ and Ram had relatively smaller sized files, is it possible that Ram ran out of disk space while Prem did not? Explain with an example.

Question 3. If we were to annotate the following space-time execution diagram with matrix time stamps, how would we annotate the event marked as 'a'?

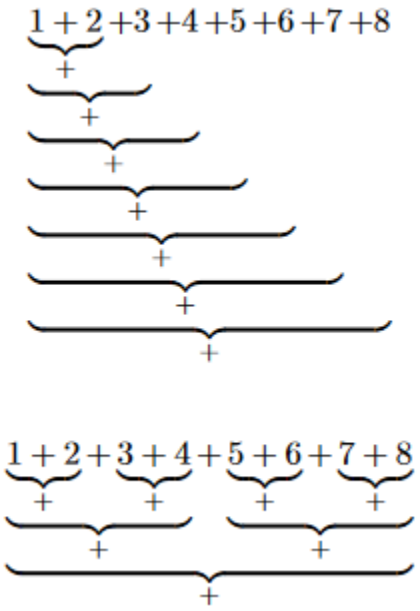


Question 4. In the muddy children puzzle, as discussed in the class, what would the children say during the first two rounds if n=4 and k=3 i.e., there are four children and they are told that at least three of them have muddy forehead? Assume that each child can only say 'yes', 'no' or 'dont know'. Use the following format to provide your answer.

```
Round1: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
Round2: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
```

Section 2: Questions carries 4 marks each.

Question 5. Our ability to write parallelizable programs decides the speed up we can achieve through scaling. Consider two programs to add numbers. The first program adds the first two numbers, remembers the result, and adds that result to the next number. It continues doing this until the end of the list is reached. The second program adds two numbers at a time in parallel. It recursively does so until the final results are arrived at. The following figure explains their logic.



As per Gustafson's law, assuming each addition is an operation, how much speedup (approximately) can these programs achieve if there are four processors?

Question 6. You are provided with a large text file containing the names of millions of chess players and their world rank. The file format is as shown below:

```
Viswanathan Anand, 15
Magnus Carlsen, 1
Venkatesh Vinayakarao, 1029388
...
```

Provide the design of a map-reduce job to pick top 10 players from this file. You do not need to write code. Explain clearly, the logic behind the mappers and reducers.

CMI/2021/Mid-Semester Exam June 2021

Chennai Mathematical Institute

Distributed Computing and Big Data Duration: 90 Mins. Max Marks: 20.

Roll No.: _____ Date: 02/06/2020

Name: _____

Instructions

- Submit a single pdf file carrying your answers on moodle under "Mid Term" assignment. For any reason, if you cannot upload to moodle, email your work to vvtesh.cmi@gmail.com.
- A penalty of 1 mark applies for every minute of late submission (beyond 13:40 Hrs).
- This is an individual assessment. Do not discuss with anyone.
- This paper refers to a variable z. If the last digit in your roll number is i, then $z = (i\%4) + 1$ where % is the modulo operator. For example, if the last digit is 1, then $z = 2$.

Section 1: Questions 1 to 5 carry 3 marks each.

Question 1. Ramesh has a file of size $4z$ terabytes. Ramesh wishes to send this file to Ria. He can send the file over a 400 Mbps direct dedicated network connection.

(1) How much time will it take for Ria to receive the file through the network?
(2) What is the maximum file size for which Ramesh will prefer to use the dedicated network channel over a overnight (i.e., within 24 hours) courier?

Explain your answer in detail. Include relevant calculations.

Question 2. Ram bought a new hard disk and configured it so that the number of bytes per inode is r. Prem too bought a new hard disk of same size and configured it to have the number of bytes per inode as p. Ram and Prem stored f number of files each in their respective disks. Given that $r > p$ and Ram had relatively smaller sized files, is it possible that Ram ran out of disk space while Prem did not? Explain with an example.

Question 3. A drive spins at 4800RPM and has average seek time of 12ms. The disk has $20 + 4z$ sectors per track. What is the average access time?

Question 4. Assume a disk size of $4z$ TB with block size of 8 KB.

(1) How much space will you need (in MB) to store the free space bitmap?
(2) If this free space bitmap needs to store additional information on whether the block is corrupt or not, how can you do it? How much space will you need to store this extended free space bitmap?

Question 5. If we have 100 processors and 5% of the total jobs cannot be parallelized, what is the scaled speedup achievable as per Gustafson's law?

Section 2: Question 6 carries 5 marks.

Question 6. You are provided with the following facts about a model of execution of a distributed system that uses global vector time stamps.

(1) $(1, 1, 1, 1) \rightarrow (2, 1, 1, 1)$ is a happens-before relation.
(2) Exactly two events occurred in each process.
(3) Between the two events of every process, at least one event occurred in another process.
(4) The first event occurred in the process p_3.

Agreeing to the above facts:

(1) Draw the space-time execution diagram annotated with global vector time stamps. (1.5 Marks)
(2) Draw the corresponding hasse diagram. (2 Marks)
(3) Draw the same space-time execution diagram annotated with matrix time. (1.5 Marks)