SANDEEP KHURANA
8994667 75

LSU Department of Computer Science
Fall 2010 Final Exam
CSC7700 Scientific Computing
December 6th 2010, 5.30pm to 7.30pm

General Instructions

- This is a closed book exam.
- No calculators or electronic devices.
- Part I of the exam covers all the five course modules and is designed to take 80 minutes to complete. Part II of the exam is for the Networks and Data module and is designed to take 40 minutes to complete.
- Part I is worth 20% of the final grade. Each module includes 5 questions. All questions have equal weight. Answer all questions.
- Part II is worth 10% of your final grade. Answer only four out of five questions. If you answer all five, only the lowest graded four will be taken into consideration. Questions have two parts, you need to answer both parts of the four questions you select.

Part I

Module A: Basic Skills

1. Provide two reasons why the same text file can look different when viewed on different systems or within different tools.

① End line/CR/LF Characters: These may happen as the property is operating system dependent.

② Tab spacing: The number of spaces used for a single tab can vary though it can be set in a file

2. In the context of numerical simulations, explain what is meant by discretization and why it is used.

Discretization:- When doing simulations we need to define the grid and the working space. Also, we may want to increase the resolution and do the computations at discrete steps.

The more we discretize and increase the resolution more time would it take to do the simulation.

3. Briefly describe what a pseudo random-number generator is, and name three disadvantages over real random-number generators. Name two reasons why pseudo random-number generators are often used despite these disadvantages?

Pseudo random generators are the systems which implement Hardware or software required to give numbers which are randomly distributed.

Disadvantages -

① Speed
② Complexity of implementation.
③ Randomness - since these are Pseudo random generator

These are used because of the fact that -

① Different algorithms (H/W or S/W) have different efficiency
② Fulfills random

4. Name one advantage and two potential disadvantages of the Newton-Raphson method over the bi-section method for root-finding.

Bisection method may do oscillations between the root findings. Thus Newton-Raphson has a defined way.

Newton-Raphson will potentially be expensive in some cases where there is complex equations involved

5. Explain the difference between centralized and distributed version control systems, including one advantage and one disadvantage for each. Name one software implementation example for each kind of system.

Centralized Version Control Systems (VCS) is means of managing the code in which a single server is the only place where commits are done. where as in distributed VCS these servers are distributed across different places.

Distributed VCS are often used by scientists spread globally since they work on different implementations of the application

Centralized VCS are often used by software developers who contribute to the same product. This helps them to do (a) versioning (b) reverting (c) tracking changes, thus reviewing.

Software Implementation:-

Centralized VCS - SVN, Git, CVS

Module B: Networks and Data

1. List two TCP parameters used in `iperf` and briefly describe their influence on the performance of TCP.

(1) Window size: Increasing this generally increases the speed of transmission though there should be an agreement between sender and receiver and optimum must be used.

(2) Threads: As we learnt from the assignment that increasing the number of parallel threads increase the speed. but after a number of threads the performance remained same.

2. Briefly describe what the server-side data processing plug-in included in the standard GridFTP installation does and what it can be used for (hint - you used it in your homework)

3. List two benefits that middleware provides to developers of distributed applications.

(1) Implementation can be developer defined: these provide the flexibility to developer so that they can write their own implementation for the interface. For eg. in Distributed Data management

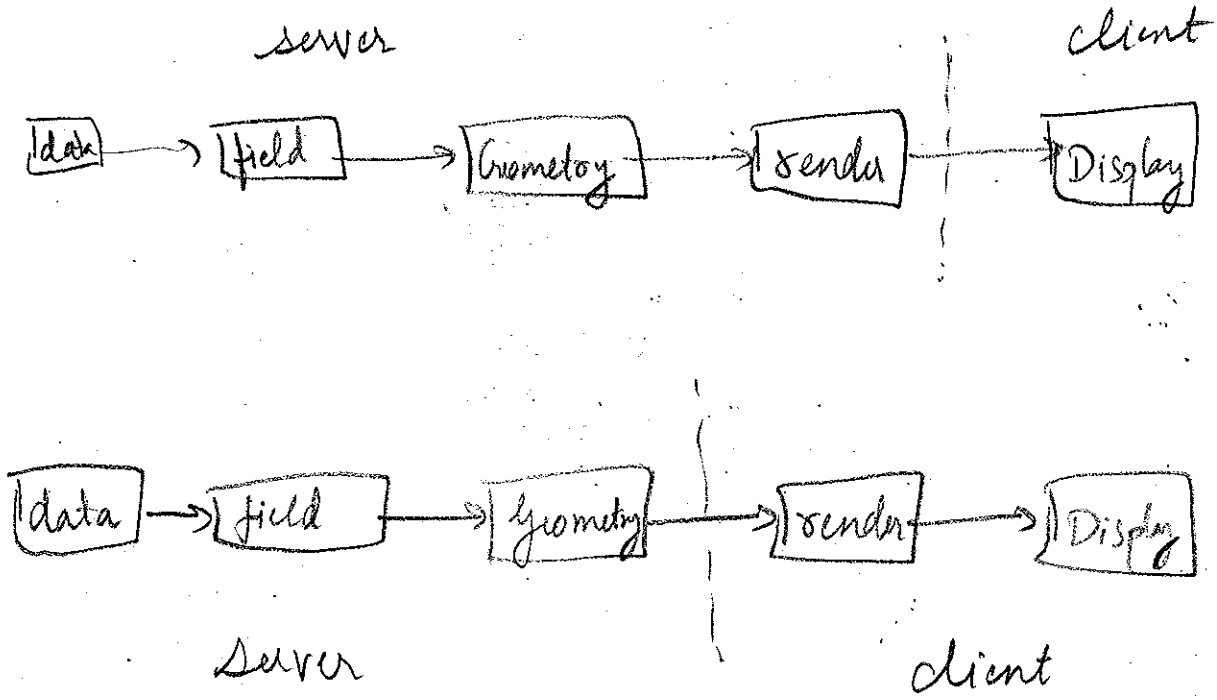
(2) Hide the low level details - Well this may be taken as an advantage since the middleware abstracts the ground level OS calls / routines.

4. Briefly outline two methods for accessing remote data in a distributed application.

① Staging :- This is copying the entire required data to the place where application is running. Advantageous when most file is required.

② Remote I/O :- Remotely accessing the data. We used grid FTP in our assignment for this. Used where latency is less.

5. Briefly outline two methods of doing remote visualization (based on distribution of the visualization pipeline)



Module C: Simulations and Application Frameworks

1. What determines the accuracy of a simulation? List two ways in which accuracy can be improved.

Accuracy of a simulation is defined by the number of grid points or discretization. If we take less number of intermediate steps/points the simulation will not give a better result even if the model is correct. Also, if we consider various forces acting on a system to get a better model, the simulation will be more accurate. For example, in case of astrophysics there are many forces acting, need to consider all.

2. What is MPI, and what is it used for? Assume there are two processes, and process A needs to access an array element stored on process B. Schematically, how does this work?

MPI (Message Passing Interface) is used to do parallel programming. It involves changing the algorithm and making it to work in parallel so that they can be submitted to different process for execution simultaneously.
If process A and B are tightly coupled then they can still communicate. process A will have to send a message to process B, asking it to return the required element. MPI has API's to make this feasible for programmers.

3. What is a software framework? Name one software framework, and provide three characteristic elements of a software framework.

Software framework: It is a design of the software which enables the development of the software or application built on it to develop faster. This can be clearer with its elements:-

(1) Pluggable architecture. - new development to come in fast
(2) Scalable. -
(3) Lifetime - It should increase the life of the software used.

Example - We used Einstein toolkit which is a framework it has flesh & thorns. Makes it easy for developers & users.

4. What are CCL files in Cactus? List which CCL files exist, and what they define.

CCL files define the descriptions of the thorn. We used this in our assignments to define class, their member functions (name, type, level). This was the interface.ccl file.

The definition of these functions are defined in a separate file. This separates interface and its implementation.

5. Name and briefly describe five tools that support code development in large, distributed, international collaborations.

(1) Version control system - Used for versioning, tracking changes, viewing contributions from specific user.

(2) Bug trackers - Used to track the bugs raised in the software, Good to track the stability of software

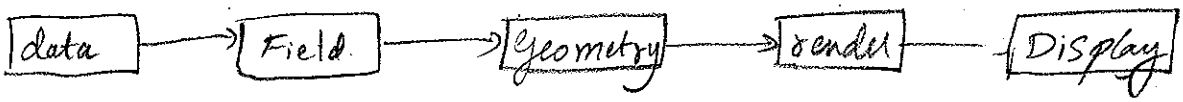
(3) Meetings - Essential for management of the software.

(4) Reviewing - Essential to validate check ins and suggestions from experienced developers.

(5) Feature documentation - Very much helpful for testers and building building test cases where testing team is at different locations.

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".



This is the very basic visualization pipeline. It involves taking in the data, selecting a field, selecting the geometry, rendering it and then display. In case of distributed visualization, we can either do rendering and display at only display at remote end.

2. What is the difference between the "push model" and the "pull model"?

Push model is the kind of model where the data is available as soon as the objects are created where as in the Pull model there should be an object to collect the data from the sink.
Objects are created and given the data as soon as they are available in Push model but in Pull model data will be pulled only when required. Vish uses the Pull model.

3. Describe the three atomic elements ("building blocks") in a visualization network.

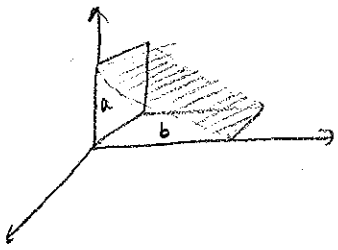
(1) Fiber : This explains the basic element in the visualization. There must be data to visualize it and Fiber connect the data from file to the form such that it can be used

(2) Rendering : This is the most crucial part in the network. This gives the output to visualize the data.

(3) Bounds - This specifies the bounds on the input field. Probably there may be infinite space to render.

4. Define and describe the purpose of a bi-vector.

bivectors are generated from the wedge (exterior) product of two vectors. they define the area of the two vectors with a plane.



5. Which are the three property objects ("communication types") in the "F5" fiber bundle data model that are visible to the end user?

Module E: Distributed Scientific Computing

1. We discussed five applications – Montage, Nektar, Climateprediction.net, SCOOP and Ensemble-based/Replica-Exchange simulations. For any THREE of these (you choose which three), answer any ONE of the following: Why they were distributed? How they were distributed? The Challenges &/or success in distributing them?

Montage – Collection of various scanning of sky. It is based on M-W, can be distributed. Challenge:- task distribution

Climate prediction :- It is naturally distributed processes can take up different parts. Challenge; failure, joining of the result.

Nektar: Distributed because of the fact that various parts can be analyzed separately

Challenge:

2. Estimate to within an order of magnitude the number of jobs that are executed in the Worldwide LHC Computing Grid (WLCG) *per day*. Estimate to within an order of magnitude the number of bytes of data generated (overall) by the WLCG. Estimate the cost of the LHC Experiment. Therefore what is the cost of generating a byte of data from the LHC experiment?

Data generated by LHC experiment is in the order of TB. The jobs that executed are around 46K jobs/day

3. Using your estimate (whatever it was) of number of jobs (on the WLCG) from the previous answer, given that there are approximately 250,000 cores as part of the WLCG, and that it has a typical utilization factor of 50%, estimate the average time each job takes. (assume: each job is a single-core job).

4. List two factors – technological or non-technological, driving Cloud Computing. Provide a "real production" example of a Cloud offering. Is the Cloud offering an example of IaaS, PaaS or SaaS?

Techno logical - ① Virtualization ②

Non-technological - ① Reduction in cost

Microsoft Azure : real example of cloud offering.

PaaS : Amazon's service, used for.

SaaS :

IaaS :

5. Provide one difference between predominantly HTC and HPC Grids. Provide a "real production" example of a HPC and HTC Grid.

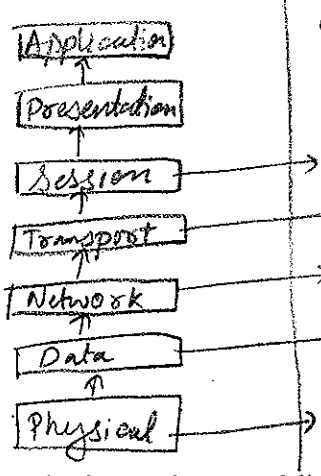
Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

OSI Layer



Each layer in a network has specific function associated to it. There are different protocols working at each level.

Session → Sessions associated with each communication
Transport → End to end, Reliable communication, ordering
Network → Routing, Logical paths.
Data → Frames,
Physical → Transfer bits

- B) What are the major differences between TCP and UDP?

TCP

(a) Reliable
(b) ordered
(c) Connection oriented
(d) Packet based

UDP

(a) Unreliable.
(b) Unordered.
(c) Connection less
(d) Data based.

Question 2

- A) What data transmission protocol would you use for bulk data transmission and why? What protocol would you use for video or audio conference and why?

For bulk transmission I would use TCP. This is the most trivial protocol for file transfer. It offers high throughput and is reliable. Since this is file transmission the transmitted file should be exactly same as original. For video or audio conference UDP would be good option since reliability is not a matter of concern. And even if one or more frames are not received it wont give any impact.

- B) Describe circuit network services and their advantage.

Question 3

- A) Describe what a naming service is (in middleware implementations) and what is it used for.

Naming service is a service which resolves the request to a particular resource. In middleware systems we saw that the IP and Port were not the only required parameter for communication. They are useful when -

(1) Too many Endpoints are generated. (eg. IP and Port)
(2) Dynamic nature of mapping.

- B) In your own words, describe the "end-to-end" argument.

"End to end" is a w

Question 4

- A) List the usual sequence of operations for accessing data in a distributed file system.

- B) Briefly describe the two possible (and sometimes conflicting) optimization goals of a scheduling system.

Question 5

- A) Describe use case scenarios where remote visualization is useful or needed.

Remote visualization is useful when:-
(1) There is a low latency of disk access.
(2) Local hardware does not support rendering (saw an example in the class)
(3) Data is available remotely, otherwise staging/remote I/O will have to ~~taken~~ into place.

- B) Describe some of the possible benefits of distributed visualization.

(1) Sharing of the resources, no need for high end rendering systems for every user.
(2) Reduce the I/O in a remote fashion (data is generally in TBs)
(3) helpful when the disk I/O time is larger than network transmission.