Sai Saripalli

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

Sai Satipalli

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.

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

Partial differential equations (PDE's) are applied to continuum systems to describe them in proper way.
Discretization in context of numerical simulations is an approximation of PDE's. It is used to reduce the complexity of describing the system behaviour by PDE's and this approximation may lead to error.

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?

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

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.

Sai Sasipalli

Module B: Networks and Data

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

-w (window or buffer size): If more number of packets are lost then the window size is reduced to half of data rate for TCP to improve performance

-P (~~Pass~~ number of parallel streams); more number of parallel stream better is the performance of

-t (time): time taken to transfer data

-i (interval): how often should the performance be printed.

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)

<globus-url-copy> is the server-side data processing plug-in included in GridFTP installation.

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

① It allows applications to get connected to integrated run-time environment.

② It helps to coordinate different operating environments.

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

remote data in distributed application can be accessed by procedural of object oriented (RMI, RPC), CORBA) systems

Remote procedural call (RPC): execution of task in remote address space is called Remote procedural call

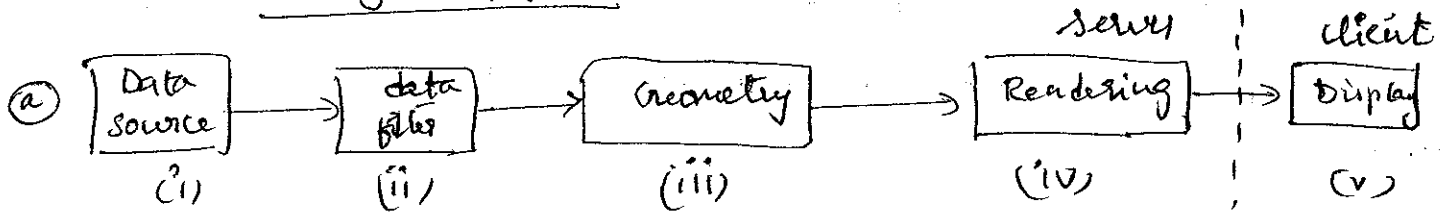
Remote method invocation (RMI): Java remote method invocation implements (java.rmi.remote) and should catch RemoteException in order to handle the method in remote location

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

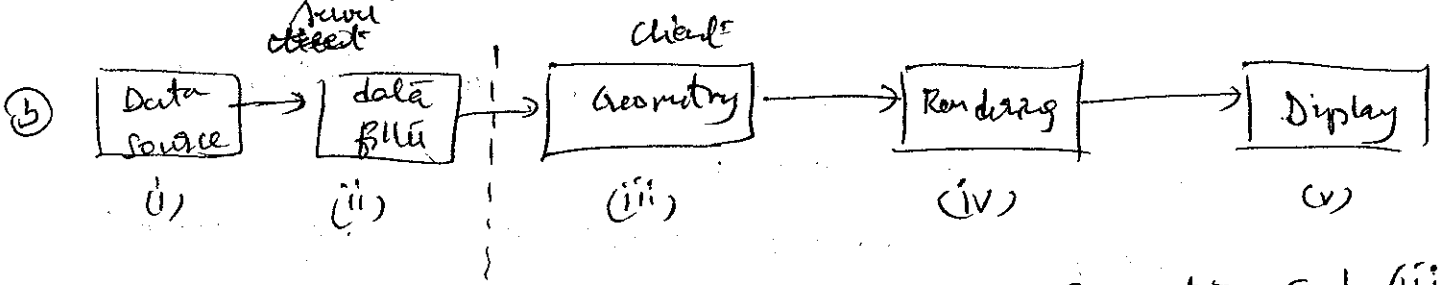
Visualization of data can be done in 2 ways

- remote visualization
- staging (copying data to local location)

Visualization pipeline: remote visualization is done in 2 ways



If we perform all steps (i) → (iv) in the remote location and just display the output (image) to client



If steps (i) - (ii) are performed in the remote location and (iii) - (v) are performed in the client location.

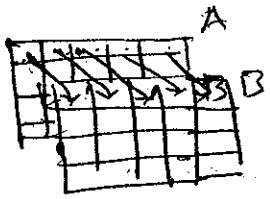
Sai Saripalli

Module C: Simulations and Application Frameworks

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

Identifying a partial differential equation (PDEs) determine the accuracy of simulation. because PDEs help in identifying the behavior of system exactly. Accuracy of simulation can be improved by setting up initial condition and repeating the routine execution over the tiny steps as many times as possible to reduce the error caused by using discretization of PDEs

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 is an High Performance Computing model it is used for Peak performance delivery for given task. Each task is divided into smaller jobs that are executed individually and output is summary of coordination of results of individually executed jobs.

Each process A is divided into small jobs that ~~independently~~ independently access the array element stored on process B.

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

Software framework is ~~set of~~ ~~concepts~~ learn glue of components.

'Cactus' is a software framework.

3 characteristics of elements are

- Interface
- Implementation
- Schedule.

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

CCL files are the flesh of the Cactus. They are

- interface.ccl : provides implementation name and inheritance relationship between routines.
- schedule.ccl : declare the set up routines schedule and variables that are used and synchronized
- Parameter.ccl
  - ↳ Provides the list of parameters that are used for routine communication.

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

sv. einsteintoolkit.org

sv. Cactuscode.org

git. Carpetcode.org

sv.einsteintoolkit.org: helps in simulation of complex applications using distributed environment

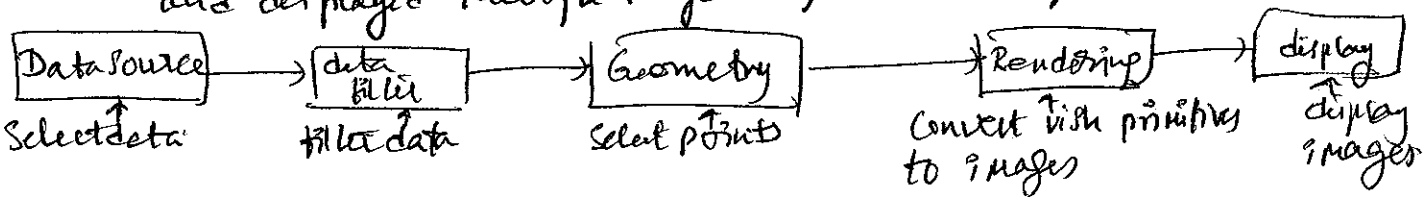
sv Cactuscode.org: helps in communicating variables through during run-time in distributed environment

Sai Sarpali

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

Modular infrastructure ~~that~~ that implements visualization algorithms is called visualization.
Visualization pipeline is the process which shows how data is selected and displayed through image using Vtk software



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

push model	pull model
(i) loads data that is not used	(i) loads data that is only used
(ii) Data to be display is loaded quickly	(ii) Data to be displayed ~~is loaded~~ takes time
(iii) Eg:- Amira software	(iii) Example: Vtk software

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

3 atomic elements of ~~its~~ visualization network are

- Data source : output (select)
- Data Sink : input (select)
- Data filter : both input and output

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

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

Sai Saripalli

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: Image mosaic which is combination of many images that has the pixel data which appears to be from single image from a single telescope

Why distributed: Scale processing is above local limits

How they were distributed:- They use DAG - Enactor for distributed.

Challenges/issues/success: Assigning jobs to resources

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?

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?

Factors:

- resource pooling: we need to pool resources as increase in requirements instead of ~~complex~~ high end systems.
- utility metric: pay for the utilities that are required or used instead of maintaining ~~the~~

Amazon 'EC2' is a real time production Example of cloud Offering This cloud offering is example for IaaS & SaaS. (Software as Service

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

'Usage modes' is one predominant difference between HTC and HPC grids.

HTC ~~Machine~~ operates on multiple usage modes that are ~~different~~ varied in different machines where as HPC operates in single usage mode as it is on individual machine.

HPC - Cray supercomputer is a real production example

HTC - gnumber (Condor) / TACC is a real production example

Sai Sarpotli

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

physical layer: electrical signals and cabling

network layer: routing data between nodes within network specified stations addresses

Transport layer: ensures delivery of messages (TCP | UDP)

Data link layer: transfer data within node based on machine | station address

Session layer: deals with opening | closing of sessions.

Application layer: layer that interacts with GUI for end user

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

TCP	UDP
reliable	unreliable
ordered	unordered

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?

GridFTP is the protocol that I would use for bulk data transmission. This is because it can transfer bulk data from one server to other server within short time.

For video/audio conference we should not look for protocol that can ensure continuity of packets delivery as it is not useful in displaying image. Hence we use Pipeline execution protocol.

- 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 in middleware implementations is used to map local method/file to corresponding method/file where it is called/executed on remote location.

A lookup operation is performed in naming service. It is used for mapping the local file to corresponding file located on a different physical location.

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

Question 4

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

- Identify file/data to be accessed
- Identify naming service
- Setup a routine to access the file using the naming service
- Retrieve the ~~and~~ file if found else report exception.

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

(a) Maximizing the resource usage

(b) Maximizing the application utility

(a) Resource usage can be maximized by equally distributing tasks to all resources available without leaving the resources idle.

(b) Application utility can be maximized by utilization of all utilities that are available for a to expedite execution on a machine

Question 5

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

remote visualization is useful in a scenario where you need only partial file data (information) for visualizing. i.e It is useful when you do not wants to load entire data on to your local machine as the entire data (information) is not necessary to visualize the scenario

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

(a) Distributed visualization helps in visualizing complex/high computation data (information) that requires high end resources which can't be provided on a single machine (source).

(b) It may help in ~~reducing~~ reducing the load on a single machine