Pradeep Kumar Martha.

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. Depends on the editor how it indents the file.
Some editor Consider tabs as 4 Spaces & Some editor Consider 8 Spaces as a tab.

2. when a text file is opened in windows it does not show any ^M characters. Whereas when the Same text file is ftped to unix it ends up showing ^M characters. So it depends on operating System ~~other mode~~ in which you view the file.

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

The ~~PDEs~~ partial differential equations describe Continuum Systems. ~~and~~ which have infinite degree's of freedom. To reduce the Complexity, Discretization is performed.

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?

random number generates ~~the~~ numbers.
We have to use Seed so that it initializes the random number Generator function.

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.

A Version Control System helps to manage the Soure Code files accross multiple developers or users ~~not~~ without Conflict.

A Centralized Version Control will have a single reposity of Code base. where every user will get the files from that repositry.

Adv: There is Synchronization between users & Consistency ~~between~~ in Code base.

Disadvantage: when a file is checked out by a person. the other person has to wait until the first file has been checked in

example: svn.

Distributed Version Control: It is difficult to implement.

Advantage: Multiple access to the files Can be provided

Disadvantage: The Latest file ~~also~~ ~~modified~~ Copy should be replicated on all the distributed ~~System~~. Version Server.

example: Implement svn on multiple machines.

Pradeep Kumar Mantha

Module B: Networks and Data

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

The two TCP parameters are.

-w = window size - As the window size decreases the bandwidth used will becomes less. If a packet is lost then window size is halved. After Successfull transmission of packet add a segment to window for each RTT.

-b = Speed of transfer - Less Speed will decrease the bandwidth utilization & increases congestion.

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)

gridftp is used for bulk data transfer. The Syntax is

globus-url-copy <Source> <Target>

The plugin used is gsiftp.

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

1. Naming Service.

2. Communication: Sockets are not implemented Same on different Operating Systems. So Middleware will Convert the Communication API to Comunication API which the Corresponding Operating System Can understand.

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

1. grid-ftp: Standared ftp protocol defined by Open grid forum. It Can be used to ~~set~~ manage the remote data

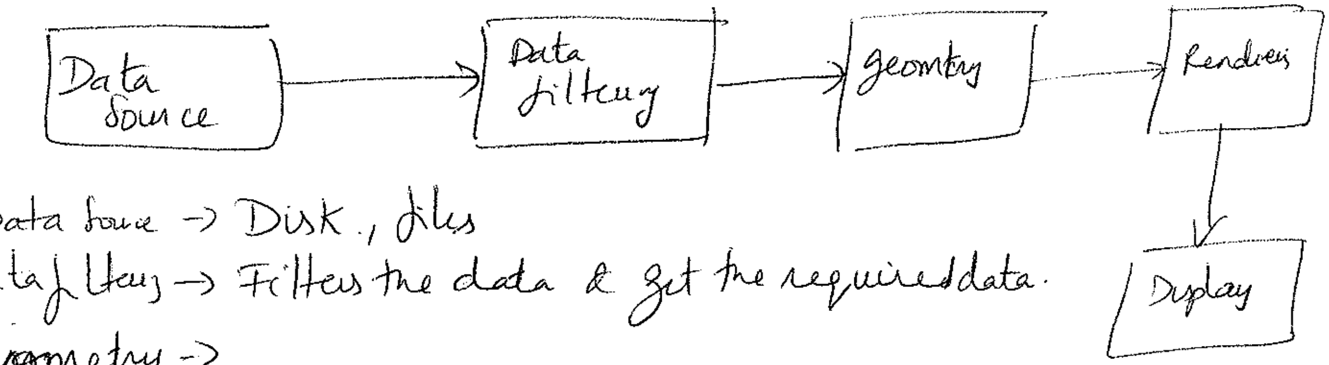
~~2. Remote I/O.~~
~~2. Video Confrencing.~~

2. petashare - provides global namespace for distributed resources - gives access to remote data.

3. Irods. - Integreate rule oriented data system. ~~conta~~ provides interface to access remote data.

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

Visualization pipeline



Data source -> Disk, files

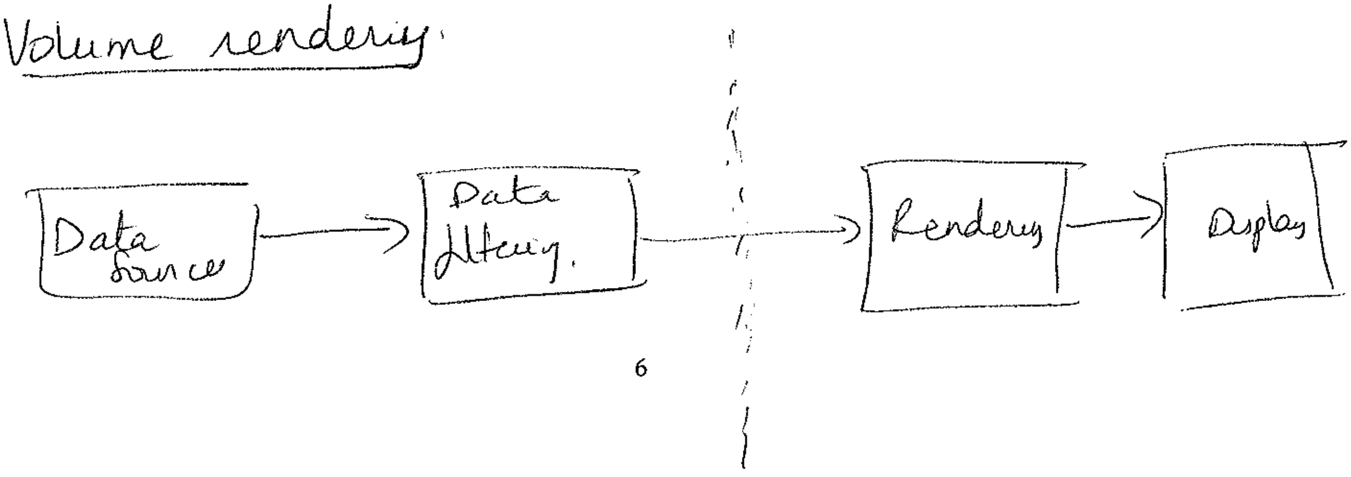
Data filtering -> Filters the data & get the required data.

geometry ->

Rendering -> produces images

Display -> group the images & display.

Volume rendering.



Pradeep Kumar Mantha

Module C: Simulations and Application Frameworks

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

(1). Discretization: ~~PDEs~~ partial differential equations describe Continnum systems ~~and~~ which have infinite degree of freedom. To reduce the complexity Discretization is done on the PDE's. which results in approximation. Approximation result in error.

(2) The accuracy Can also be improved by setting Correct ~~to~~ initial & boundary Conditions.

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 a parallel programming language. In MPI, a copy of program is given to all the processes. Two processes communicate using message passing. So when a process A needs to access an array element stored on process B it Sends a message to process B. process B handles the message sent by A & responds.

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

Cactus is a software framework where

(2) each computational task is a Component and Can developed by a group of developers.

(1) The framework provides glue. i.e it assemble all the Component. It provides main function, libraries & provides Communication interface between Components.

(3) The end user will assemble all the Component once he has all Components ready from all the developers.

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

① Interface.ccl – implements the thorn, inherits the thorns which are ~~used~~ required for this thorn. Provide info abt all the procedures & variables to be used a provided by this thorn.

② Schedule.ccl - Decide which function to be executed at what time.

③ param.ccl - define the runtime execution a use of thorn. It can also be used in extending the thorn ~~to~~ be used by other thorns.

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

Software framework - which provides most of software development for ~~Soft~~ Code development.

The Components can be divided ~~to~~ into

ⓐ editor - to write the Code.

ⓑ Source Code Verson Control

ⓒ Compilers & linKers.

ⓓ Testing tools.

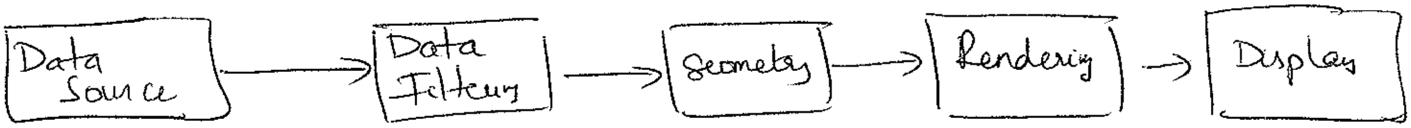
ⓔ deployment tools

Pradeep Kumar Mantha.

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

A Visualization pipeline Can be represented pictorially as



Data Source -> Disks, Files
Data Filters -> using iso-surface level
Geometry -> Colors, transparencies.
Rendering -> images
Display -> Movies

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

Push model.	pull model.
① Data is made available as early as possible	Data is made available as late as possible
② Data is available even if it is not required	② Data is accessed only when It is required.

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

① Data source - only output

② Data Filtering - ~~only output~~ Both input & output

③ Data sink - only input

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

A bi-vector is two dimensional & is used to store the state of data.

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

(a) Datasource ~~[illegible]~~

(b) Data sink

(c) Date Filtering. (Iso-Surface level)

Pradeep Kumar Mantha

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?

~~Application~~ Application	Why distributed	How distributed.	Challenges
Montage	Processing > local availability	DAG is created & executed by DAGMAN	Coordination.
Nektar		MPI	Coordination
Ensemble-based/RE	Many Computational tasks.	Saga-advert.	~~[illegible]~~ Implementing Saga on multiple machines. ~~[illegible]~~ Coordination

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?

① 1 Million jobs are executed / day.

② 1 Petabyte of data is generated.

Cost of generating a byte of data ≃ 0.01$

1+2+3+4

2+4+8+10

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).

1M - 250,000

Consider each job takes 1 unit of time for its execution

1 Million Jobs should be distributed over 2,50,000 cores. each core has a utilization factor of 50%

1 Job - 1 core - ~~0.5 unit of time~~ 50%.. utilization

~~So each job takes~~ So first 2,50,000 Jobs take 2 unit

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?

(1) Resource pooling.

(2) pay on ~~demand~~ usage.

example: Amazon. Web Service. eucalyptus, Nimbus. Azure.

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

HPC grids:

More number of machines with less Computational Capability

example = Loni.

HTC grid.

less number of ~~highly~~ ~~Configured machine~~ high Computational Capability.

example: EGI.

Pradeep Kumar Mantha

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

① physical layer: used to define physical & electrical specification of the network media used to carry data bits.

② DatalinK Layer: adds addressing & transfer the data within a network.

③ Network Layer: used to transfer data between network from source to destination. Routing of ~~packet~~ data is done. path Decision Making algorithm are used

④ Transport Layer: used to provide Congestion control protocols, Flowcontrol. reliable transmission of data ⑤ Session Layer ⑥ Presentation Layer ⑦ Application Layer.

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

TCP	UDP
Connection oriented	Connection less
Byte oriented	packet oriented.
reliable	unreliabled
ordered	un-ordered

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 used for bulk data transmission. It is ~~file~~ a standard file transfer protocol ~~was~~ defined by OpenGrid forum.

for Video or audio Conference - latency should be minimum. So reliable data transfers are not used. Compression technique introduce latency but bandwidth Can be Utilized effectively. Distribution tree networks Can be used. as the data is not transmitted to each user. (multicasting)

- B) Describe circuit network services and their advantage.

Circuit network Services:

When the data is transmitted to ~~the~~ a network. the data is distributed ~~automatically~~ within the network to all the nodes by Circuit network.

Question 3

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

~~Naming service:~~

~~It is used to find a node & a service over a network.~~

~~These implementations~~ ~~defse~~ ~~differ on different infrastructure. So Middleware takes care of all the implementation details & hide them from user.~~

- 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.

First the metadata is verified. Two types of metadata are available ①. System metadata (user details, file details)

② user defined metadata (contain domain specific details)

After getting the details of file from meta data then data is retrieved from the corresponding data server or node.

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

① . Maximum Resource utilization

Utilize resources efficiently while scheduling.

② Maximize Application Utilization

Run the application as fast as possible.

Question 5

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

(a) when there is lack of resources (memory, software infrastructure) on local system.

(b) when the simulation output is huge to ~~visualize~~ to move from one location to another.

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

(1). effective Utilization of I/O resources.

(2) Improved throughput.