NAGABANDI KARTHIK KUMAR

89-960-2766.

(NKK)

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

HEK

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)=> one reason on why a text file can look different is due to different type of text editors used.

→ some text editors include aligning, coloring the data types etc, which may not be present in other text editors.

(2)=>

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

→ Discretization is a method which is generally used in place of PDE(partial differential Equation).

=> Discretization is generally used to minimize the errors, which cannot be easily done by partial differential equations.

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.

=> The only advantage of Newton-Raphson method is that, the root-finding is faster than using bisection method.

=> The disadvantages are
(1) The calculation of root using Newton-Raphson is more complicated than using bisection method.

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.

=> In centralized system the controlling is done by only system, all the decisions and permissions are handled only by one system.

-Advantage: Redundancy is less and security is more in centralized system.

Disadvantage: single point of failure.

=> In distributed control system, the control is distributed across the network, the distributed systems function independently to ~~goa~~ achieve a common goal and the final Job is done.

Advantages: since the functionality is distributed, the processing is faster and since the data is distributed any failure of system can be replaced with other.

Disadvantages: security is less in distributed control system, Data is lost easily & redundant data.

MAGABANDI
KARTHIK
KUMAR

Module B: Networks and Data

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

① -w => window size. Syntax: -w < size of window >

=> There is always a limit to the window size, the larger is the window size, the fast would be the transfer of data.

② -p => number of parallel process Syntax => -p < no of ||le >

=> As the number of parallel streams increases, the more is the data transfer, hence the data transfer is faster.

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.

=> Middleware provides library services for distributed application.

=> Middleware is a layer of software between API and running Infrastructure.

=> Middleware provides an Interace for application, it generally contains NOS (network operating sys) which takes care of the data transfer taking place.

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

(1) Java RMI => Remote Method Invocation,

- create a class which extends Unified Remote Object and inheits its previous Interface.
- start the system manager.

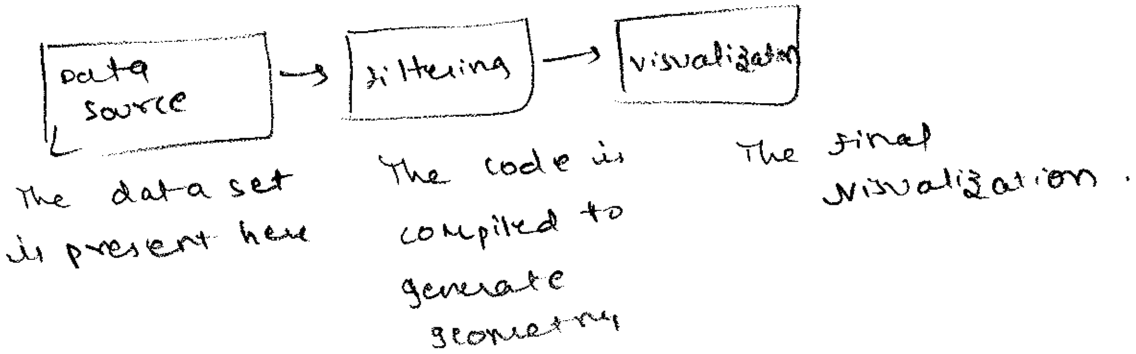
→ Create Naming Service for reference by clients.

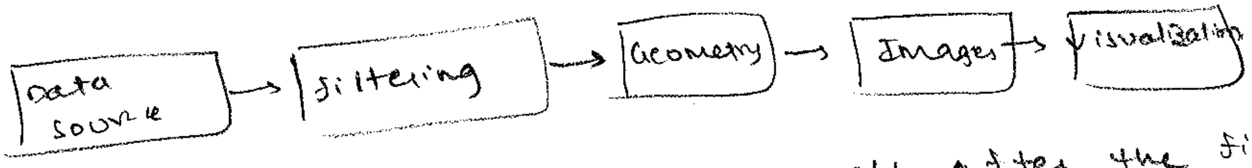
(2) CORBA → common object Remote Brokerage Architecture.

→ ORB is initiated, server creates instance, name the instance & stores it in ORB for referency by the client.

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

=> one method for doing remote visualization is to use the data set remotely, Instead of downloading & using it in local system, using the data sets already available would be faster.


Data source
filtering
visualization
The data set is present here
The code is compiled to generate geometry
The final visualization.



→ The same works in this Model as well, After the filtering of the data set is done, the geometry of the data set is generated, and then this geometry generates Images. All these Images together generate visualization.

NKK

Module C: Simulations and Application Frameworks

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

→ GNU plot generally determines the accuracy of the simulation.

→ Using GNU plot we can find how deviating it is from the normal simulation.

→

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 generally used for parallel simulations.

→

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

→ A software framework is a application which is used to merge the components.

→ cactus is an example of software framework, and

(1) The thorns in cactus are the components

(2) the flesh is the framework in cactus.

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

⇒ CCL files are configuration files in Cactus.

The CCL files which exist in Cactus are

(1) interface.CCL ⇒ Includes interfaces & implementation classes, lists all the parameters and methods used.

(2) schedule.CCL ⇒ schedule.ccl gives when the thorn is scheduled to be executed and other scheduling processes.

(3) Param.ccl ⇒ Param.ccl gives all the Parameters used by the thorn.

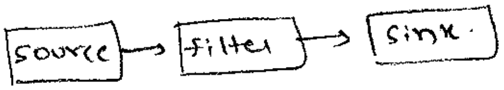
5. Name and briefly describe five tools that support code development in large, distributed, international collaborations.

NEK

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

vizualization pipe is a model which includes source , filter and sink.



=> the source file consists of the code which is to be visualized and

=> the filter , filters the methods for geometrical structure.

=> The sink, is a location where ~~dat~~ visualization is stored.

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

push model	pull model
(1) Filter modole receives data initially	(1) Filter modole receives data at the end of simulation.
(2) Data is given even if not reqvired.	(2) Data is given only if req.
(3) The viz. pipeline is ~~@~~ accessed in creation phase	(3) The viz. pipeline ~~should be~~ is accessed in Rendering phase.

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

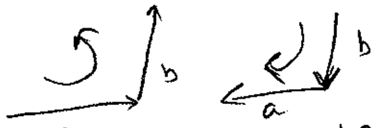
→ The source, is where the data set is present, which is used for visualization.

→ the filter , is where compiling of the code is done

→ the sink is where the visualization is finally stored.

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

⇒ Bi veltor is used to give the product and the angle of rotation of the vectors.



$a \wedge b = ab$ $\quad$ $a \wedge b = -ba$

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

Bounding Box

ortho slice

NKK

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?

Challenges:

(1) Montage: collection of resources & proper placement.

(2) Nextar: Installing saga on all systems,

(3) climateprediction.net ⇒ collecting appropriate data from all centers.

(4) SCOOP: not-robust,

(5) Ensembled based/Replica-Exchange

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?

→ Infrastructure limitation is one factor for driving cloud computing. since data to be stored is increasing day by day, there is a scarcity of Infrastructure.

→ Purchasing softwares according to the use is also responsible for clould computing.

⇒ Amazon EC2 is an real production example, and this is a Saas (software as a service) cloud offering.

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

→ The one main difference between HTC & HPC comes with the amount of data, used by the respective grids.

→ The amount of data on which operations are performed and the data transfer taking place is very large in HTC as compared to HPC.

HTC ⇒ Heavy neutron collider is a real production of HTC

HPC ⇒ Tera grids is a real production example of HPC

✓

NAGABANDI
KARTHIK
KUMAR.

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

→ The physical layer is used to gather information from resources in the form of bytes.

→ The Data link layer →

→ The network layer → The network layer deals with what kind of transport protocol most be used for data transfer

→ The transport layer → The transport layer includes TCP & UDP & deal with these transfers.

→ session Layer → The session layer holds session Id for each process(or) transactions being held by system

→ Application Layer. → The Application layer include services such as http | SMTP | FTP etc.

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

TCP	UDP
→ connection oriented protocol.	→ connection less protocol.
→ Reliable	→ un reliable
→ byte - oriented	→ Message oriented.
→ Medium for data-transfer is essential.	→ Air is the medium.
→ Congestion can be caused in a path	→ Since there is no medium, ~~[illegible]~~ Congestion is rarely created.
→ Acknowledgement can be received.	→ No acknowledgement can be received.

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?

⇒ GRID FTP is generally used for bulk data transfer,

⇒ since the data transfer is taking place in the form of grids and grids have large space, Grid FTP is used.

⇒ 'TCP' would be better for video or audio conference because, the loss of data is minimum in TCP since it is a connection oriented protocol.

- B) Describe circuit network services and their advantage.

→ A Network circuit is generally generated for data transfer between systems. The different circuit network services are

(1) Internet 2 ION: Internet 2 Ion is a service which has dedicated circuit, for a Network, Reservations have to be made to get a Network circuit.

(2) Lightpath switching: optical fibers are used for transfer of data, the data transfer takes place in the form of lambdas. There are 2 types static & dynamic, static has fixed path, dynamic ⇒ Mirrors are used. Adv: Faster mode of transfer.

(3) GLIF/GOLE: This service has administrative Network circuit, which is used for faster data transfer.

(4) wide path x. Indi band ⇒

(5) coallocation & distributed ⇒ since it is distributed, this services are faster than other services.

Question 3

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

→ once the Instance is generated by a server, each Instance must have a name, so that it can be accessed.

⇒ once the Instance is named, it is stored in ORB server.

~~to~~ → Naming service comes to use here, where if a client request for the Instance generated by server, it can be accessed only through the name given by Naming service. First ORB is initrated and then the objects are given names so that they can be accessed.

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

→ end-to-end, refers to the data transfer taking from a one end of the server to other end of the client.

→ This end-to-end data transfer involves various interfaces, Network operating systems (NOS), Routing, security,

⇒ All these have to be taken care so that the data transfer is being taken place between legitimate users.

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 may be needed, if the data set which is used to visualize the object is larger in size. In this case, instead of dowloading the entire dataset, the visualization can be done remotely saving space & time.

→ Remote visualization also may be helpful, if the local system does not have enough resources to perform visualization, this is like sharing of resources to get the visualization.

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

→ Visualizations generally use large data sets, compiling these and then visualizing these may be time consuming, therefore, by distributing the data set, the visualization can be performed and then merged to give resultant visualization, which save space and time.

⇒ Another benefit of distributed visualization is that, since the data set are smaller, More accuracy is generated.

→ Minimum utilization of resources, better visualization and avoiding redundancy.