BIDUR BOHARA

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 Different tools or systems implements different encoding standards for texts. So if text generated in editor implementing Unicode-8, standard, the same text may have different characters when opened in editor with just ASCII.

② Different tools in different system have different coding of the End-Of-Line, this also makes the texts look different in different systems.

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

When a continous real world system is fragmented and represented in a Grid structure with numerous grid points, the grid points have state variable values representing the continuous system. This technique to discretize continuous systems & represent by countable points is called Discretization.

In case of simulation, the whole system is discretize by grid structure so that the whole domain is represented by few grid points instead of infinite points if the real continuous system is considered.

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-number generators, as its name implies, generates a set of random number using certain algorithm such that the generated random numbers are not random in real & can repeat the pattern over time.

Disadvantage :- is not real random numbers & repeats the set of numbers
- can be predicted once the pattern is recognize.
- depends upon the seeding condition.

Two reasons - ① they generate numbers that behave close enough to random numbers.

② fast to generate

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

Advantage:-

① Newton-Raphson method has first-derivative term, hence it is 2nd order method, meaning more accurate & the effect of error is highly reduced, compared to Bi-section.

Disadvantage

- can oscillate over the root
- doesn't bracket the root, & hence can be slow

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 case of centralized version control systems, all the files under the project is stored in one centralized server. Where as in case of distributed version control system, the files under version control are distributed accross multiple servers at different location.

↳ "svn" implements centralized version control system.

↳ For centralized system, all the files in a project can be accessed with one address, but if the server is down or crashes the whole project is inaccessible.

↳ For distributed system, even if one server is down the files from others can be accessed. But distributed system requires to commit into & update from all servers.

Module B: Networks and Data

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

-P : is used to specify the number of processors to use for TCP communication, facilitating the parallelism & hence faster transfer.

- w : is used for specifying the window size, which can be used for congestion control also. Increasing window size means more data transfered at a time, hence high trasfer rate. Reducing window size to control 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)

- Globus tool-kit was installed in server side for data processing.
- Used to transfer data from url to url/location.

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

① Freedom for application to run in any platform or OS, because in different OS, the hardware level API & executions is different. So, middleware hides these hard ware level & kernel level complexities for Distributed applications developers.

② Middlewares also handles the communication & cooperation between multiple servers, so the user is not aware of running distributed application, rather it appears a application in single machine.

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

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

(1) In first method, data access from source, processing of data in filters, generating geometries and rendering. are all done in server side. And only the final image is displayed in client.

(2) In second approach, rendering of geometries and display of images takes place in client side, and all the other steps of pipeline is done in server side.

Module C: Simulations and Application Frameworks

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

The closeness of simulation result to the expected output b; to the real world system determines the accuracy of a simulation.

Ways to improve accuracy

① Refine the grid used for solving PDEs. High resolution of grid means more accurate

② Use high-order methods to solve the PDEs such that the propagated errors is very low.

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 stands for Message Passing Interface and is used for writing a parallel code for executing in ~~super~~ multi core systems. The multiple processes or threads communicates by passing messages.

↳ A sends read request to process B; B acknowledges the request & sends the pointer to array to A; array A acknowledges back if it receives the data. All the communication is done via message passing between processes.

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

software framework is like a pillar in tower, it acts as a main application but don't do anything by itself and only provides framework for components to compute & communicate in application via framework.

Flesh in the Cactus is a software framework.
① Component list & platform for components to execute.
② Input parameters.
③ Scheduling strategy.

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

They are the characteristic files in Cactus. And following CCL file exist in Cactus.

① Interface.ccl: defines interface of thorn & its implementation list of variable; functions & its implementation

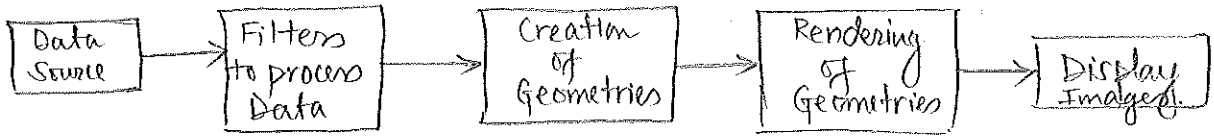
② schedule.ccl: defines the scheduling of routines in thorns. assign & freeing & scheduling of variables.

③ param.ccl: list all the input arguments that the thorn takes & initial values for variables.

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

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".



↳ the data source is processed using filters to represent the data in visual form, thus processed data is used to generate geometries. Those geometries are rendered & finally displayed as Images in screen.

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

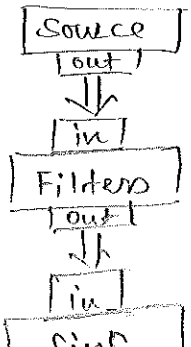
Push Model	Pull Model
- data sent for processing as soon as available from source.	- source gives data only when requested by data sink.
- data is processed even if not required by the system	- data processed only when required to visualize
- data available at earliest	- data available as late as possible.
- Avizo works in push mode	- VISIT work in pull model.

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

① Data Source – generates the data to be processed; only outputs the data.

② Filters – takes data as input; processes the data & outputs a different set of data.

③ Data Sink – takes only inputs of data from Source or filters. Display modules are example of sink.



Visualization Network.

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

A wedge product of two vectors is a bi-vector & represent the plane containing two vectors, an the magnitude of bi-vector giving the are of rectangle.

Bi-vector can be used to find the inverse of vector or the division of a vector by vector.

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

(1) Bundle

(2) Grid.

(3) Field.

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 is distributed – all the files to create a image came from distributed system.

   Nektar – the computation data is huge that it can't be processed in single workstation, so the work is divided in distributed system.

   Ensemble-based/Replica-Exchange – is collection of uncoupled or loosely coupled jobs, so they are highly parallel & were ran in distributed system for high performance & efficiency.

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?

- use of remote resources, otherwise will stay unused most of the time.

- Its an example of 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?

↳ a hierarchical system where layers in lower level supports the functionality of upper layer.

↳ but at any layer, users don't see the lower layer as its contribution.

↳ lower layers are closed to network devices & hardware.

↳ upper layer provides applications to users.

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

	TCP	UDP
①	connection-oriented	① connection-less
②	reliable	② not-reliable.
③	controls congestion with varying window size	③ no congestion control.

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?

↳ Bulk-data protocol; that guarantees the reliability of data transfer without wozzying much of timing. Time doesn't matter but the content of transfer should be unharmed.

↳ Videoconference protocol, that not only takes care of reliability but also fast transfer without data loss. Because the system needs to interactive, without distorted images & lost audio.

- 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.

- 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.

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

① divide jobs into smaller units to be scheduled in distributed system. This conflicts with communication overhead between jobs

Question 5

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

  - when the data to be visualized is huge in size, such that copying & processing data locally is not efficient & practical.
  - in case of terabytes of data to be visualized, the data should be processed in remote servers & only the generated images should be transfered to local machine & displayed.
  - parallel transfer of images results in faster rate of display in case of movie.

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

  - copying of data to local machines is avoided.
  - by generating visulization images remotely, the images can be multi-casted so that the same rendered images can be displayed in multiple local machines with only one computational effort.
  - visualization can be observed by anyone in any part of this world.

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/p time is larger than network transmission.