Ling Zhu

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. because of different text editors used. For example, some editors can recognize encoding. but some cannot.

2. because of different systems. for example. Unix treat ~~text~~ file as plain text file so cannot display the .doc file as windows does.

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

1. discretization is a kind of approximation. It discretizes a system (domain) into elements, and each element. is consistent with the system (PDEs)

2. because the original problems don't have analytical solution so numerical method is used. Discretization is one of steps of simulation procedures.

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?

PRNG (pseudo random-number generator) is an algorithm to generate quasi-random numbers.

Disadvantages:

1. high correlation of these random numbers
2. short period of these numbers.
3. lack of uniformity.

Still used. because:

1. low cost
2. reproducable

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

disadvantage of Newton - Raphson method

1. only applicable and fast when well initialized. Bisection can always find.

2. need to know first derivation of the function. Bisection doesn't need.

Advantage:

1. fast, when a good initial value choosed. (2-order of convergence) can be used to polish the results from other methods, like bisection method.

Bisection method only has 1-order of convergence.

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.

distributed version control systems: (svn)

advantage: easy to update.

disadvantage: may get confliction.

Centralized version control Systems.

advantage: Not easy to introduce bugs / confliction.

disadvantage: Not easy to update. (users tend to bypass it).

Module B: Networks and Data

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

-w. set window size : the higher window size the better performance, until it approaches the best performance

(-i. the interval between output get printed. no obvious influence on TCP performance.)

-t. the time set. for transmitting data. The Longer time, the better performance. until it approaches the best performance.

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)

this server-side data processing plug-in used for transferring data. and allowes operations on data remotely.

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

1. "Everyone" can use middleware

2. effectively utilize resources.

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

1. generic middleware (eg. SOAP)
2. Remote I/O.
3. Parrot

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.

Resolutions determines the accuracy

Two ways.

1. use finer meshes.

2. improve scheme. for example: For ODE problem Runge-Kutta method has 4th-order accuracy. which Euler method only has 1-order accuracy. The results from Euler will be less accurate than the results from Runge-Kutta.

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 message passing interface. It is used in parallel computing to establish the communication and interface between different components.

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

Software framework is a lean and provide glue to combine different components together.

Framework Example: Cactus.

3 characteristic elements :

reliability

usability

portability

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

ccl files are interface of a thorn.. There are 3 .ccl file in a thorn.

param.ccl. — gives parameters that can be used in this or other thorns

interface.ccl. – defines interface with other thorns.

schedl. ccl – schedule the thorn. when to execute.

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

Ling

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

Data → Data of Interest~~s~~ → geometry → image ⟶ visualization

(From Simulation) (Filter) image rendering

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

push model - render data

pull model - filter data

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

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

bi-vector is the result of the production of 2-vectors
the magnitule is the area of the formed parallelism
and the direction is normal to the formed 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.	Nektar	Climateprediction.net
why	processing scale above local limit.	① processing scale above limit ② fundamental challenge with 3-D problem	high-demand, need quick response.

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 PG / day in WLCG

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).

250,000 x 50% = 125,000 cores in work

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. Technological =
   persuit of powerful computing / storage / simulation.
   Non-technological :
   Connect to real-life.

2. Azure .
   It's an example of PaaS

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

The difference is whether it's aimed to peak - performance or the fully - utilization

HPC: TeraGrid.

HTC: OSG

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

physical layer. point to point data transmission

data-link layer: add station address

network layer: transmit data in bytes ,

transport layer: transmit data in packet .

session layer:

presentation layer:

application layer:

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

TCP:	UDP:
connection-oriented.	Connectionless
reliable	unreliable
ordering	no-ordering
Congestion control	
byte-oriented	packet-oriented

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?

- 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.

a naming service can be used to find file in its logical name instead of by host name.

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

end-to-end argument:

evaluate the network performance at end-point.

Example helps to describe:

consider "vedio conference": the reliability transmission will lead to latency.

Question 4

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

Two types <
1. Staging. access data where it is located
2. Remote I/O.

sequence <
1. quere data in metadata catlog
2. use pointer to find data in the distributed file system.

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

1. Data scheduling , for example, make sure the data don't get overloaded.
2. Network scheduling.

Question 5

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

Scenarios: When facing data with huge size (TB or PB) and the network performance is not good.

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

benefits. 1. no need to transfer data through network
2. not affected by poor network performance, latency.
3. Can use supercomputer to render more images when dealing with large size of datas