Lin Xue

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. encoding. different encoding scheme will create different view

2. big-endian and small-endian

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

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 generator is often used to generate random numbers based on some seed.

Disadvantage: 1. Sometimes is not purely random because of the seed chosen. 2 sometimes the periodic of one word is not random.

Advantage: fast and cheap

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

Centralized version control is that all the changes are made local, but the decision (merge and trunk) will be made by same central server.

Distributed version control is that all the changes, ope and decisions are made locally, everyone will take re s its self.

SVN is an example for centralized version control

Module B: Networks and Data

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

-W, the tcp sider receiver buffer size, used to buffer the tcp packets. Large value can accomodate high latency network which will consume more buffer

-P, the tcp port number, (1 ...), specify the tcp port number you may want to use

GRID FTP IS AND SOME OF ITS

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)

FEATURES

Grid-ftp is a file transfer tool based on ftp protocol between different sites

Using Grid-ftp, you can get/put files remotely between different sites which is belonged to the grids.

Comare to usual ftp tool, it can help you directly transfer files within grid once you have the access to both sites.

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

1. middleware can accomodate different upper layer applications to act the same to lower layer, such that can make different languages and applications work without changing for the low platform.

2. middleware can make distributed applications work together transparently without knowing they are actually separated

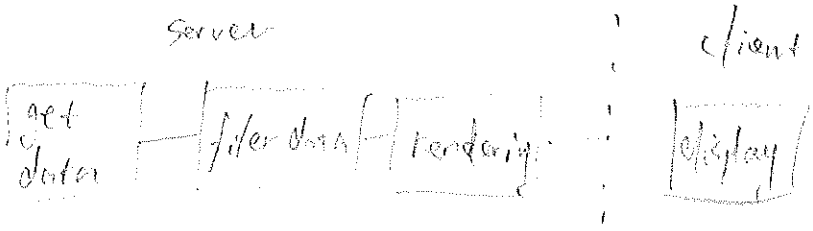
4. Briefly outline two methods for accessing remote data in a distributed application.

1. Gridftp. use ftp to directly transfer data between different sites within the grid.

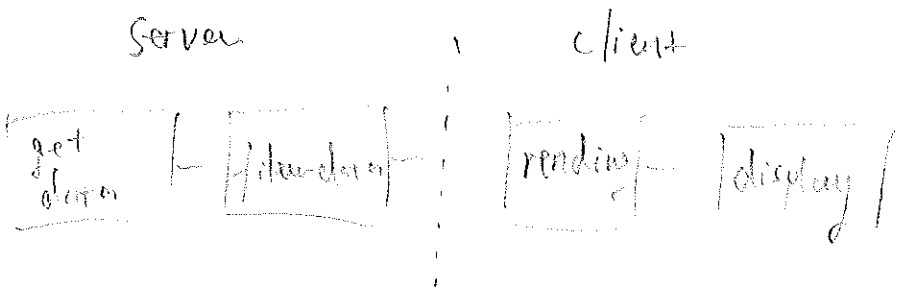
2. SCP. can accessing remote data when you have the access.

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

1 Doing visualization and rendering at server side while just doing display at the client.



2. doing the rendering at client site, that help to decrease the load on 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.

1. the understanding of the simulated problem, including the accurate physical problems and the math equations.

2. carefully design the architecture of the 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 a API for multiprocessor parallely computing, it is used for doing computing parallelly on processors in a multi-processor architecture

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

A software framework is an architecture for design, develop, test, and run a bunch of software. for example, Cactus is a software framework.

Characters: 1. uniform format. 2. easy to extend

3. develop fast

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

CCL files are some specific files in Cactus which help to compile and link inside Cactus framework.

interface.ccl : define interfaces between different thorns

scheme.ccl : define how the thorns and the flesh will be

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

Cactus, developed by CCT LSU to do mainly physics simulati

Einstein toolkit, developed long time ago for scientific sim

.Net framework, developed by MS, only on MS windows.

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

Visualization Pipline includes the steps of get visulized data, filte the data rendering and display.
It is used for pipelining the large amount of visualized data processing to create the visualization for people. We can choose to rendering data on server side or on client side.

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

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

1. filte: help to filte the visual data from raw data.
2. rendering. create the real visualization, which refires most of computational power.
3. display, display the visualization.

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

The bi-vector is used for visualize the multi-dimension image.

equation for bi-vector

$a \wedge b = (a \times b) \wedge b$

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

HDF5

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: every part will have its own work and result, so that the combination will be a big powerful picture.

Nektar can be used for blood chart of human.

climateprediction.com: for climate prediction it should be quick and need a lot of power, so every distributed site can contribute to the computation.

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?

number of jobs per day : 1000

number of bytes by WLCG : 1 Terabytes

Cost of LHC experiment : $500

Cost of one byte : $ 50 Cents

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).

for one day:

cores active : 125000 core

cores/job : 125000 / 1000 = 125 core/job

time for each job = $\frac{24 \times 60}{125} \approx 12$ min.

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. end user do not need to have great powerful pcs, they can just use a terminal to access through the cloud.
2. Inside cloud, there will be very powerful servers do everything so that user do not need to care

Example: SAGA, Amazon EC2

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

HPC Grids, a grid of distributed sites which can communicate with one each other and provide a cloud computing environment for end users.

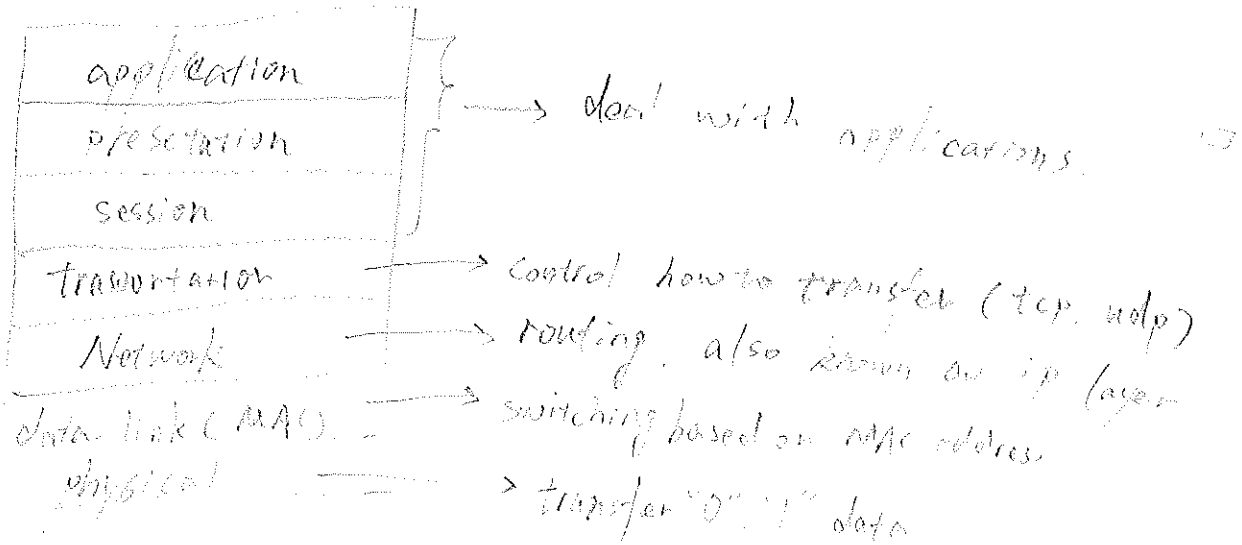
Teragrid, is an example of HPC Grids.

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?



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

TCP: reliable, garentee the transmission, has congestion control

UDP: unreliable, just transfer for its best effort, donot have 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?

For bulk data transmission, use tcp, because tcp can provide reliable tragmission, once there are some loss, it will help to recovery

For video/audio conference, use udp, because video/audio suffer to lantency, udp will send as fast as possible, while tcp may increase the lantency.

- 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 is the service help you to get the specific name inside the network.

Only with correct naming can the host be recognized and the data can be accessed, because within middle ware implementation all the distributed servers will have different name and that they can not communicate

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

In contrast to "host-to-host".

"end-to-end" means port to port which is mainly the concept in transport layer, and application layer.

While "host to host" is mainly network layer

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.

Visualization always requires very big amount of data and computational power, so that if end user who only have less powerful pc and small amount of disk/ram may want to use the remote visulization, In remote visulization the remote server will take care of the computing and the disk, the use will only care about the easy display.

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

Distributed visualization is:

powerful: every distributed end may contribute to the computation

transparent: server side will take care of all the Computation and will provide Powerful hardware so that the distributed ends can just use the resources.