Liu, Ke (kliu14@lsu.edu)

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.

Different systems or tools have different standards to read ~~text~~ files.

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

In numerical simulation, discretization means dividing the domain into discrete elements.
a large number of regular.

Discretization can simplify the solution of a ~~numerical~~ system since the domain is regular, and evolution system behaves ~~in~~ well in a simple way.

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: an algorithm to generator random numbers which follow a certain distribution.

Disadvantages:
(1) Results are not ~~rand~~ really random (computation is deterministic)
(2)
(3)

Reasons they are used: 3
(1) Economical to use than real RNG;
(2) more ~~real~~ easy to reproduce.

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

Advantage:

higher order convergence ~~error~~ rate.

Disadvantage:

require knowledge in $f'(x)$;

root is not always bracketed.

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.

Module B: Networks and Data

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

Window size (-w): ~~Increasing~~ window size will increase the transfer rate.

~~Packet size~~

Parallel stream: Use of parallel stream will increase transfer rate since only one stream will ~~s~~ decrease rate when data is lost.

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)

It ~~stream~~ divide data into multiple streams.

It can be used for partial file transfer.

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

Middleware allows developers to develop applications ~~and~~ without considering different low level API's on different systems.

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

(1) Remote I\O.

(2) ~~Move application or both applied~~ Access data by copying files.

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

(1) Generate geometry on server-side; ~~and~~ render and display on client-side.

(2) Render on server-side and display on client-side.

Module C: Simulations and Application Frameworks

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

Resolution determines the accuracy.
and a correct scheme
To improve accuracy:
increase resolution;
use a better revolution scheme.

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 API in essential. It is the industry standart for parallel HPC computing.
A will also store the array element, a (stored in B)
(1) B compute all the elements other than a;
(2) B compute a;
(3) A get the results on a.
(4) A compute other elements on A.

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

Software framework is the glue between different components
for example. Cactus is a framework.

Three characteristics:
(1) Framework provides glue;
(2) Component (thorn) is independent functional module.
(3) Only users ensemble code, no central control.

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

CCL files are configuration files in Cactus.

(1) interface.ccl:
name of implementation, inheritance; variables; global functions.

(2) schedule.ccl:
When schedule which function; when which variables to be allocated/freed; which ~~par~~ variables should be synchronized.

(3) param.ccl: runtime parameters.

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

Cactus,
MpICC
Eclipse

Ke

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

Visualization pipeline is the ~~def~~ channel data flow through in data visualization.

Data — Filter — Geometry — Render — Display.

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

(1) In push model, data is loaded as soon as it is valid; filter has information about data before rendering; traverse goes through ~~the~~ filter back to source.

(2) In pull model, data is loaded only when there is request; filter has no information about data before rendering; traverse goes through filter to sink.

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

Data source, data link and filter.

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

If $\vec{a}$ and $\vec{b}$ are vectors, $\vec{a} \times \vec{b}$ is a bi-vector.

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

fiber, base

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?

Challenge &/or success of

(a) Montage: find an optimal subset of ~~grids~~ resources; distribute elements to ~~distrib~~ resources.

(b) Climateprediction:

Install SAGA on all the grids; coordination of the distributed ~~grids~~ resources; find an ~~not~~ optimal subset of resources it owns.

(c) Replica-Exchange:

coordination of the distributed resources; find an optimal subset of resources.

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 on WLCG per day: ~~20~~ 2,000.

Number of bytes of generated data: 100TB GB

Cost of LHC experiment: $10 billion.

$$\text{Cost per byte} = \frac{\$10 \times 10^9}{100 \times 10^{12} \text{ bytes}} = \$1/\text{byte} \quad \$0.01/\text{byte}$$

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).

$$\frac{250{,}000 \times 50\% \times 24 \times 2 \text{ /hr}}{2000} =$$

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?

Two driving factors: data centre, increased demand for computing resource.

Example: AZURE. (Is IaaS)

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

Example of

HPC: teragrid.

HTC: OSG.

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

Different layers define different standard and protocol for network implementations.

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

UDP: is unreliable, unordered, packet-oriented, no congestion control.

TCP: reliable, ordered, Byte-oriented, with congestion control.

Ke

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?

TCP for bulk data transfer, since TCP provides high transmission accuracy with little data loss.
UDP for video or audio conference. since UDP has low latency.

- B) Describe circuit network services and their advantage.

~~Orbit~~ Circuit network provides guarantee since it provides certain allocation ~~certa~~ to users.
Other network has no guarantee. It might work, might not work.

Question 3

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

Naming service maps a known abstract names to physical names which virtual points to identical locations where data is stored.

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

function provided at low level is redundant or of little value compared with cost to put it on low level.

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) Maximize ~~utility~~ resource utilization.
(2) Maximize application utility.

Question 5

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

Solution is generated at remote HPC, but we need to visualize it on local machine.

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

Large data size;
Large visualization rate;
Allows collaborative visualization.