Daniel Kogler

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

Daniel Kogler

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.

Text files can be represented in binary using different encodings on different systems or even be displayed with different formats for different tools

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

We cannot simulate a continuous phenomenon to arbitrary precision. Discretization is an approximation of the continuous phenomenon so that we can represent the data using finite memory.

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 produce what appear to be random values but in fact are a repeatable sequence of values.

Disadvantages: -Not completely random
-Algorithm can eventually be cracked.
-Large amounts of produced values will eventually yield repetition of the sequence.

Reasons used: -Inexpensive compared to true RNGs
-easy to implement

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; SVN
advantage; Easier to make the changes you make compatible with other users' changes
disadvantage; If someone else updates the repository before you, you must make your code work with their changes before you can submit

Distributed; Branch/Trunk
advantage; Can make changes to your local copy without worrying about other users' changes.
disadvantage; When finally merging your branch to the trunk, the process can be very difficult

Daniel Kogler

Module B: Networks and Data

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

Packet size - increasing the packet size generally improved performance.

window size - increasing window size could also increase performance, but usually produced little benefit over default

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.

It provides abstraction and encapsulation of the low level interfaces to the hard ware.

Allows applications to be more portable.

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

Remote procedure calls

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

Method 1: Copy the data and perform the rendering locally. Useful if local hardware has necessary graphics support and memory size/transfer time is not an issue.

Method 2: Let processing occur near the data and move the processed data to a server specialized for rendering.

Daniel Kogler

Module C: Simulations and Application Frameworks

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

Accuracy is determined by the accuracy of the model and the amount of approximation of values allowed (see below)

Can be improved by: -Making the simulation more fine grained.

-Running the simulation using smaller time steps

Both approximate the modeled phenomenon more closely

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?

Message-Passing-Interface, used to parallelize applications

Using MPI, process B would send a message to process A containing the necessary data (send/receive or broadcast)

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

A software framework is used to make putting together a code easier.

CACTUS - a software framework

1) Thorns/modules are libraries

2) Flesh is thin (does nothing but coordinate interactions between thorns).

3) Ideally thorns are stand-alone, and no thorn is more important than any other thorn.

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

interface - defines how other thorns see the thorn

implementation - defines how the thorn works, i.e. the actual C code.

param - declares what other thorns should be activated and when

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

Cactus

Linux

GCC

Eclipse

SVN

Daniel Kogler

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

A visualization pipeline represents the stages of computation which lead from raw data to the final rendered image.

Data → filter → process → render image
data filtered data

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

Push model "pushes" the data to where it will next be used.
Pull model "pulls" the data as it is needed.

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

Raw data → data used to construct the image.
Data processor → processes the data into a form which can be visualized
Renderer → takes processed data and uses it to create an image.

4. Define and describe the purpose of 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?

Daniel Kogler

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 - distributed so that different subimages could be processed simultaneously on different machines

Nektar - One challenge was degradation of performance caused by adding too many processors spread too far apart.

Replica-Exchange simulations - Used to verify results of each parallel simulation.

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?

~ 10 million jobs/day

~ 1 Peta byte

~ 10 billion dollars

$$\frac{10 \text{ billion}}{1 \text{ quadrillion}} = \frac{10 \cdot E10}{10 \cdot E15} = \frac{1}{10^5} = .001 \text{ cent/byte}$$

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).

$$\frac{10{,}000{,}000/\text{day}}{250{,}000 \text{ core}} = 40 \frac{\text{jobs}}{\text{core day}}$$

50% ⇒ 40 = 12 hours

1 job = .3 hours

1 job = 18 minutes

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?

One factor is hype, regardless of anything else.
Another is a good work environment.

Amazon.

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

Daniel Kogler

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

Different layers provide management capabilities for different levels of complexity.

i.e. Physical layer handles point to point communication, switching layer allows communication to different addresses in the same LAN
Routing Layer connects multiple networks
and there is a final layer specifying connection ports to communicate with specific applications

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

TCP is much more reliable, but UDP can potentially be faster. With UDP, however, there is a chance that the data will arrive out of order or even not at all, so for most cases TCP is preferable, as it protects against such occurrences.

Daniel Kogler

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 - it depends on the nature of the data but almost certainly TCP to avoid any errors being caused by out-of-order transfer or dropped packets.

Video/audio conference - UDP usually, as the human brain can usually compensate for the occassional communications glitch, so even with a couple of errors in the transfer the conference will be minimally affected.

- 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.

Request data → Server queries other servers →
Server with data replies/sends data to first server (not necessarily according to some research for certain scenarios)

→ Send data to requestor

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

1) Maximize work performed; attempt to maximize utilization of resources in order to get as much done as possible, even if it slows down some applications by doing so.

2) Minimize run time of a particular application. Try to get a particular job compled as fast as possible.

Question 5

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

If the data size is too large for local systems, or if local systems do not have graphics support, OR if the time needed to transfer data is too long, then remote visualization is necessary.

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

Possible benefits include - performing the visualization faster than possible on a single machine by utilizing specialized servers.

- being able to put forth a collaborative effort in the visualization process