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.

89-956-3415

Chou, Chui-hui

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 systems have different definitions for some characters. For example, on unix-like systems, '\n' stands for the new line character, but the equivalent characters on MS Windows are '\r' '\n'.

② Different terminals have different alignment settings. For example, a screen on a terminal may be 80×24, but on another terminal, a screen may be 100×25.

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

① Using discrete values to approximate continuous values.

② Because a computer stores all values as discrete values. A user has to choose some sample values for computers to process.

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?

disadvantages:

① It's not perfectly normal distributed.

②

③

reasons:

① It's easy to acquire.

② It's reproduceable with the same random seed.

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

advantage: It approaches the root faster.

disadvantage:

①

②

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

advantage: easy to manage.

disadvantage: single point of failure.

example: SVN

distributed

advantage: enhanced accessibility because of replicatio

disadvantage: hard to sychronize.

Module B: Networks and Data

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

① -w , buffer size

if the size is two small, the sender can not send sufficient data to utilize the whole bandwidth.

② -P , parallel streams

the more streams two ends use, the more bandwidth they get.

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)

① It filters or selects the data for transferring at the sender side. Only the desired data are transferred.

② Usage:

A user can speeify that only the records with certain timestamps are transferred.

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

① Uniform programming interface.

② Naming. A service which maps logical names to physical names on physical locations.

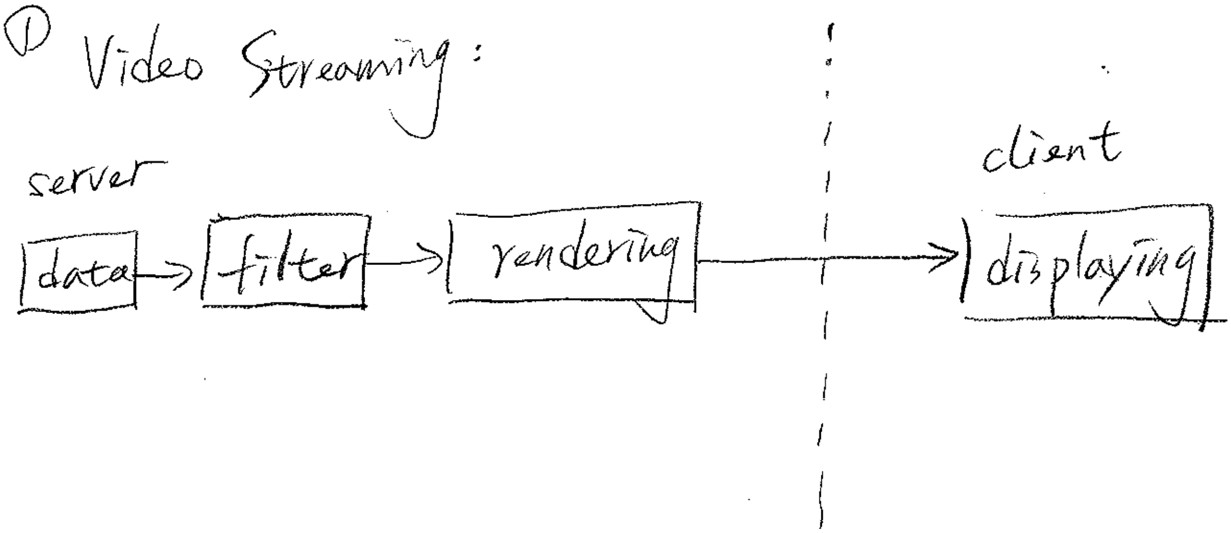
4. Briefly outline two methods for accessing remote data in a distributed application.

① Staging: copy the whole data to local storage devices on demand.

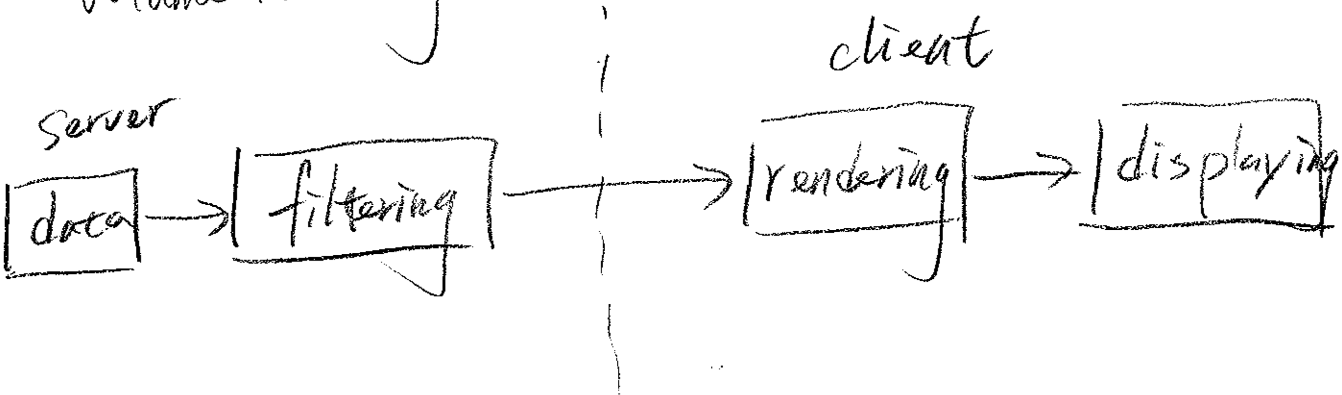
② Remote I/O: retrive required blocks of data via middle wares.

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

① Video Streaming:



② Volume Rendering



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 PDE used. Using a PDE which better fits the behavior of the simulated targets.

② The level of discretization. Providing more detailed data of the simulated targets.

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?

① Message Passing Interface. It's used for information exchage between processes running on different processors.

② a. A specifies the array name and the index, arr[x]

b. x is transferred to the physical address on the memory of process B.

c. The element is retreived and returned to A as arr[x].

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

An infrastructure which glues many independent components for a simulation.

software framework: Cactus.

characteristics:

① Does nothing but coordinating components

②

③

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

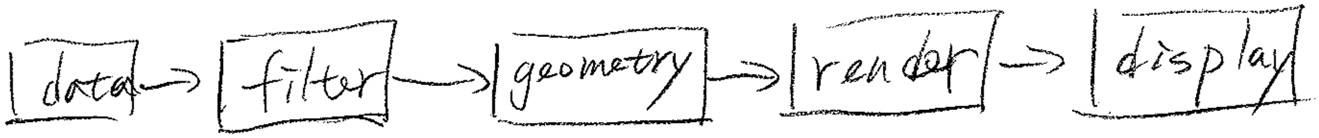
5. Name and briefly describe five tools that support code development in large, distributed, international collaborations.

Cactus

Carpet

Module D. Scientific Visualization

1. Define and describe a "Visualization Pipeline".

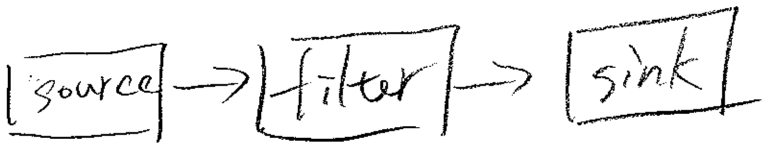


2. What is the difference between the "push model" and the "pull model"?

Push: the data source sends data down streams for rendering.

Pull: the sink sends requests upstream via filter to the source to ask for data for rendering.

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



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

An area spanned by two vectors.
It's value is $|a||b| \sin \alpha$, α is the angle between the two vectors.

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

① Vector field
② tensor field
③ scalar 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?

→ Nektar: ① A single host cannot provide sufficient resources.
② Use distributed version of MPI.
③ Latency for communication is too high.
→ ClimatePrediction.net: ① Data reside in distributed locations.
②
③
→ SCOOP: ① Owning all resources for infrequent use is not economi
② Universities use their HPC machines to compute different data of interests
③ Co-scheduling.

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?

factors : ① Scalability ② Fault Tolerance

Cloud offering : Amazon - IaaS

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

example : HPC - Tera Grid.

Part II

Networks and Data

Question 1

- A) How are layers used in network implementations?

① Physical Layer: For actual point-to-point connection.

② Data Link Layer: For finding paths for nodes within a LAN.

③ Network Layer: For finding paths for nodes in different LAN

④ Transport Layer: For finding the receiver process on a host.

⑤ Application Layer: For running a network application.

Upper Layers use lower layers functions.

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

TCP:

① connection-oriented

② byte-stream based

③ reliable

④ with congestion control

⑤ with flow control

UDP:

① connectionless

② packet based

③ unreliable

④ without congestion control

⑤ without flow control

0

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?

① TCP. Because I need every bit to be transffered correctly. So, I use a reliable data transfer protocol.

② UDP. Because guaranteeing that the media stream plays smoothly is the major concern. A reliable data transfer protocol introduces too much delay. So, I use a connectionless unreliable data transfer protocol.

- B) Describe circuit network services and their advantage.

Once a router between the two communication ends is found, the bandwidth in each device in between is reserved for that communication.

The advantage:

Bandwidth between two ends is guaranteed.

It's especially useful for bandwidth sensitive services, such as HD Video Conferencing.

Question 3

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

A naming service is the service which translates a logical name of a resource into a physical name and location.

It's used for directing a service request from applicatio an to its physical service provider.

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

If a service is based on a communication system, you cannot provide a complete new service with some characters which are not supported by the underlying communication system. Partial implementation of such a service is possible.

It may work in some extent.

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.

① The data are distributed and not suitable for aggregate to a single location. For example, the LHC output is too large for a single site.

② The visualization cannot be done locally.

For example, the Cactus Simulation output is too large for a laptop to visualize

③ Many audiances request the visualization result at the same time. The data only have to be rendered once and broadcasted to all audiances.

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

① High speed. Because data are rendered on some specifically designed servers.

② High Accessibility. Because there may be many copies of data being rendered, viewers can choose one which is the most easy for them to see.