Tuesday, May 10, 2005
2.30p.m. - 4.15p.m.

University of Glasgow

DEGREES OF M.Sc. and PG Dip. in Adv. CS, M.Sci., M.Eng., B.Eng., B.Sc., M.A. and M.A. (Social Sciences)

COMPUTING SCIENCE M:
GRID COMPUTING

(Answer 2 out of 3 questions.)

1. **(a)** Grid technologies are often used to establish virtual organisations (VOs). Explain what is meant by this term. Describe the capabilities that a Grid infrastructure should provide in order to dynamically establish and subsequently manage a VO. Explain the difficulties that arise in establishing and enforcing global policies across a given VO.

[12]

(b) What are the advantages and disadvantages of public key infrastructures for the establishment of VO security? Outline briefly the responsibilities of a certification authority. Describe the benefits and drawbacks of centralized certification authorities versus local certification authorities.

[16]

(c) Role Based Access Control (RBAC) technologies such as PERMIS can provide an increased level of security for given VO sites. Describe the principles behind RBAC technologies and their advantages and disadvantages when applied across large scale multi-institutional VOs.

[10]

(d) Explain the process by which a Grid service can be made secure using RBAC infrastructures, including a description of the generic mechanisms by which policy decisions can be made and enforced. In your answer, you may wish to refer to Grid services based upon recent releases of the Globus toolkit (version 3.3+). Describe the differences when securing a Grid service with GSI and with RBAC technologies such as PERMIS.

[12]

2. Fundamental properties of Data Grid services are:

(i) Secure, reliable, efficient data transfer; and

(ii) Ability to register; locate, and manage multiple copies of datasets

(a) What are the tools being developed by the Globus Alliance that will provide these services? List some of the key functions of these tools.

[6]

(b) What functionality should a Grid infrastructure provide to make a data resource available and usable in a Grid environment? Your answer should include examples of the motivation behind providing such functionality.

[11]

(c) Explain in detail the architecture of a Grid data service, describing the composite services that may exist, their stateful or stateless nature, and their relationship with one another. You may reference a specific implementation to explain this architecture, and illustrate your answer with a diagram if appropriate.

[12]

(d) Using the model you have described in (c), describe the steps involved in a typical data access service invocation.

[4]

(e) With the sequencing of the human genome (and numerous other organisms), the life science community is experiencing an unprecedented growth in the production and consumption of a broad array of data sets. Outline the challenges in developing Grid solutions to access and use these data sets. What does the life science community need to do to facilitate Grid based solutions?

[10]

(f) Outline the challenges associated with the long-term management and curation of scientific and non-scientific data. Explain the benefits and potential drawbacks in capturing data provenance information.

[7]

3. The concepts of Grid Computing have found widespread application in support of large-scale scientific computing. An example of this is in the high-energy physics community, which is working towards using a computational grid to distribute and analyse data from experiments at CERN (the European particle physics research centre) and elsewhere. These experiments will produce many terabytes of data over the next decade, and that data will need to be analysed by physicists at universities and research laboratories worldwide.

(a) A key challenge in that analysis will be scheduling computational jobs, since the processing of results will involve the cooperative execution of tasks on many computing systems, distributed across many different organizations. For example, a PhD student in Glasgow might wish to access data stored at CERN in Geneva, analyse the data on high-performance computational clusters in Glasgow and Manchester, and visualize the results on her desktop machine. Describe the job scheduling and management challenges that must be overcome before such a system can be implemented. Explain why these challenges occur and discuss how they might be resolved. Your answer should focus on the conceptual challenges in concurrently executing a set of related jobs across several organizations, not on the specifics of particular implementation technologies.

[20]

(b) One of the issues with scheduling jobs across multiple organizations is access to the large amounts data that Grid Computing systems frequently require. Considering the example from part (a), assuming the data storage site and computational clusters are connected via the Internet, and ignoring network data transfer performance issues, explain why it might be difficult to transfer the data between the storage site and the hosts where the computation is to be performed. Explain what changes you might make to the network, systems attached to the network, or organizational management procedures in order to remove or limit the effects of these problems.

[10]

(c) Finally, considering the example from parts (a) and (b), discuss possible reasons why the network data transfer performance might be less than desired. Explain what could be done to improve the data transfer rates seen by applications, and discuss whether such changes are deployable in the short- to medium-term.

[20]