Grid Data Management Systems & Services

Arun Jagatheesan^{1, 2}, Reagan Moore¹, Norman W. Paton³ and Paul Watson⁴

¹San Diego Supercomputer Center University of California at San Diego La Jolla, CA 92093, USA {moore,arun}@sdsc.edu

²Institute for High Energy Physics and Astrophysics University of Florida Gainesville, FL 32611, USA

Summary

The Grid is an emerging infrastructure for providing coordinated and consistent access to distributed, heterogeneous computational and information storage resources amongst autonomous organizations.

Data grids are being built across the world as the next generation data handling systems for sharing access to data and storage systems within multiple administrative domains. A data grid provides logical name spaces for digital entities and storage resources to create global identifiers that are location independent. Data grid systems provide services on the logical name space for the manipulation, management, and organization of digital entities.

Databases are increasingly being used within Grid applications for data and metadata management, and several groups are now developing services for the access and integration of structured data on the Grid. The service-based approach to making data available on the Grid is being encouraged by the adoption of the Open Grid Services Architecture (OGSA), which is bringing about the integration of the Grid with Web Service technologies.

The tutorial will introduce the Grid, and examine the requirements, issues and possible solutions for integrating data into the Grid. It will take examples from current systems, in particular the SDSC Storage Resource Broker and the OGSA-Database Access and Integration project.

Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003 ³Department of Computer Science University of Manchester Manchester, UK norm@cs.man.ac.uk

⁴School of Computing Science University of Newcastle-upon-Tyne Newcastle-upon-Tyne, UK Paul.Watson@ncl.ac.uk

Presenters

Arun Jagatheesan is an Adjunct Researcher at the Institute for High Energy Physics and Astrophysics at the University of Florida and a visiting scholar at the San Diego Supercomputer Center (SDSC). His research interests include Data Grid Management, Internet Computing and Workflow Systems. He leads the SDSC Matrix Team and is involved in Research and Development of multiple data grid projects at SDSC.

Reagan Moore is a Distinguished Scientist and the Co-Program Director of the Data and Knowledge Systems Group at the San Diego Supercomputer Center. His research interests include datagrids, digital libraries, and persistent archives. Dr. Moore manages multiple research projects including the NSF National Science Digital Library, NARA, NASA, Library of Congress, DOE Particle Physics Data Grid, NSF National Virtual Observatory, and NSF NPACI program.

Norman Paton is a Professor of Computer Science at the University of Manchester, where he co-leads the Information Management Group. He works principally on distributed information management, spatiotemporal databases, and genome data management. He is Co-Chair of the Database Access and Integration Services Working Group of the Global Grid Forum.

Paul Watson is a Professor of Computer Science at the University of Newcastle and Director of the North-East Regional e-Science Centre. His research has mainly been in the area of high performance database systems, including the design of a number of parallel database servers in both academia and industry. He and Paton are co-leaders of a number of projects on databases and the Grid, with a focus on distributed query processing.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment