Knowledge and Data Management in GRIDs

Knowledge and Data Management in GRIDs

edited by

Domenico Talia
University of Calabria
Italy

Angelos Bilas
ICS-FORTH
Greece

Marios D. Dikaiakos
University of Cyprus
Cyprus

Springer

Domenico Talia
Università Calabria
Dipto. Elettronica Informatica
Sistemistica (DEIS)
via P. Bucci,41 c
87036 RENDE
ITALY

Angelos Bilas
ICS-FORTH
P O BOX 1385
711 10 HERAKLION
GREECE

Marios D. Dikaiakos
75 Kallipoleos Str.
University of Cyprus
Dept. Computer Science
P.O.Box 20537
1678 NICOSIA
CYPRUS

Library of Congress Control Number: 2006935054

Knowledge and Data Management in GRIDs
edited by Domenico Talia, Angelos Bilas and Marios D. Dikaiakos

ISBN-13: 978-0-387-37830-5
ISBN-10: 0-387- 37830-8
e-ISBN-13: 978-0-387-37831-2
e-ISBN-10: 0-387- 37831-6

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Foreword

While Grids have initially emerged from the need to get access to more computing power by combining several high-performance computers, it has been quickly evident that there is a similar need to get access to data, such as databases, file systems, and digital libraries which are widely spread in the Internet. By accessing these distributed data and processing them using Grid computing resources to produce knowledge, we can expect to extend the scope of Grid Technologies to new applications. In fact, research activities in this area are being pursued by several research teams and, at the same time, big companies are very active in the area.

The huge amount of dispersed data and information repositories have arisen new challenges in the field of Grid computing. Grids are evolving towards flexible and knowledge-based infrastructures, in which services will be dynamically composed, allowing applications to access heterogeneous resources to be exploited in complex distributed applications. The field of Grid computing can take advantage of related paradigms, such as Workflows, Services, and Ontologies in order to provide an infrastructure with mentioned features. This new approach can be referred as the Knowledge and Data Management in Grids, and it must address issues related to data services composition, knowledge discovery, data and knowledge integration to provide the ability for extracting useful knowledge from unmanageable volume of data, by exploiting storage management, database and data mining techniques in a Grid context.

The strategic importance of Data and Knowledge Management in the context of Grid Technologies, have led CoreGRID, the only one Network of Excellence in Grid and P2P technologies funded by EU 6th Framework Programme, to have a dedicated institute to investigate research issues in this area. This book is the result of the efforts carried out by researchers involved in this CoreGRID Institute during the first year. While the CoreGRID ambition is to foster integration and collaboration, the first year was mainly to let CoreGRID researchers to know one each other. Several meetings and workshops were organized to give the opportunity to researchers to exchanged and confront their ideas. This was the goal of the first Workshop on Knowledge and Data Management in Grids that has been held in Poznan (Poland) on Septem-

ber 13-14, 2005. I would like to take this opportunity to express my gratitude to the organizers of this workshop as well as to all contributors.

Thierry Priol, CoreGRID Scientific Co-ordinator

Preface

Data and knowledge play a key role in current and future Grid applications and services. The issues concerning representation, querying, discovery, and integration of data and knowledge in dynamic distributed environments can be addressed by exploiting features offered by Grid Technologies. Current research activities are leveraging the Grid for the provision of generic- and domain-specific solutions and services for data management and knowledge discovery. The goal is to promote a wide diffusion and use of knowledge-based Grid services for the Semantic Grid and the Knowledge Grid. To this end, researchers are focusing on problems related to (i) providing commodity-based distributed frameworks for storing, accessing, and handling data, (ii) developing semantic-based techniques and tools for supporting data intensive applications, and (iii) designing distributed data analysis techniques and services for information and knowledge extraction on Grids.

The CoreGRID Network of Excellence aims at strengthening and advancing scientific and technological excellence in the area of Grid and Peer-to-Peer technologies. To achieve its objectives, CoreGRID brings together a critical mass of well-established researchers from more than forty European institutions active in the fields of distributed systems and middleware, models, algorithms, tools and environments.

In the CoreGRID NoE, the Institute on Knowledge and Data Management (KDM) has the objective to improve integration of research activities in the fields of data management, knowledge discovery and Grid computing for providing knowledge-based Grid services for the Semantic Grid through a tight coordination of European researchers active in those areas. The research tasks undertaken in the context of the KDM Institute compose a unified scenario of the data and knowledge management in GRIDs through a layered approach that starts from efficient data storage techniques up to information management and knowledge representation and discovery. At the same time, joint research activities pursued by the Institute partners are providing single solutions for data and knowledge management that will bring benefits to research and industry in GRID technology. Within its activities. the KDM Institute organized the first Workshop on Knowledge and Data Management in Grids that has been held in

Poznan (Poland) on September 13-14, 2005. The purpose of the workshop was bringing together CoreGRID researchers and invited external scientists doing research in Knowledge and Data Management in Grid and Peer-to-Peer Systems. The workshop provided a forum for the presentation and exchange of views on the latest Grid Technology research in the area of knowledge and data management.

This book is the third volume of the CoreGRID series and, as a result of that workshop and some additional invited papers, it brings together scientific contributions by researchers and scientists working on storage, data, and knowledge management in Grid and Peer-to-Peer systems. The book chapters present the latest Grid solutions and research results in key areas of knowledge and data management such as distributed storage management, Grid databases, Semantic Grid and Grid-aware data mining. All the addressed topics are discussed in the context of recent research projects.

The book includes four parts: Grid Data Management, Grid Data Storage, Semantic Grid, and Distributed Data Mining. All those sections are concerned with key topics in the area of knowledge and data management on Grids. They provide a general view of the main challenges in implementing data- and knowledge-intensive applications in a Grid computing scenario.

The first part includes four chapters. The first one presents an overview of the OGSA-DAI (Open Grid Service Architecture - Data Access and Integration) software, which provides a uniform and extensible framework for accessing structured and semi-structured data and provide some examples of its use in significant Grid projects. The second chapter discusses data integration and query reformulation in service-based Grids. The XMAP data integration algorithm is presented and service-based architecture for data integration-enabled query processing on the Grid is discussed. In the third chapter are evaluated the benefits of using OGSA-DAI in bioinformatics GRIDs by establishing communication between OGSA-DAI and GRID project developers as well as through practical case studies involving current projects. The last chapter of this part discusses fault-tolerance in Grid-based distributed query processing. A new scheme for adding fault-tolerance to distributed query processing through a rollback-recovery mechanism is evaluated in the context of the OGSA-DQP system.

The Grid Data Storage part includes a chapter on Conductor, a rule-based production system providing the ability to configure storage systems to meet resource constraints and application requirements. Conductor is able to evaluate alternatives and minimize system costs based on certain criteria. Then an autonomous distributed system built on top of the Violin framework is presented that is able to configure and reconfigure the storage hierarchy by detecting service breaches and take actions to eliminate them. The third chapter of this part presents the Clusterix Data Management System (CDMS), a solution

for data management on Grids. Taking into account Grid specific networking conditions - different bandwidth, current load and network technologies between geographically distant sites, CDMS tries to optimize data throughput via replication and replica selection techniques.

The third part includes five chapters discussing key topics in the Semantic Grid area. The first chapter describes the dynamic aspects of the Semantic Grid reference architecture, S-OGSA, by presenting the typical patterns of interaction among these services. The next chapter describes a science metadata model developed at CCLRC providing interoperability of scientific information systems in the organization and form a specification of the type and categories of metadata that studies should capture about their investigations. Then the Semantic Grid part includes a chapter that argues that providing the appropriate means for accessing and using ontologies effectively is a key factor in enriching current Grid with semantic technologies and supporting progress towards the next generation Grid. That work was performed in the OntoGrid project. The fourth chapter in this part proposes an ontology-based metascheduler as a Grid service for co-allocating resources on multiple grid nodes based on semantic information. Finally, the part finishes with a chapter that presents the implementation of Atlas, a P2P system for the distributed storage and querying of RDF(S) metadata describing OntoGrid resources and services.

The last part of the book includes contributions on Distributed Data mining in Grids. The first chapter describes the composition of distributed knowledge discovery services according to the WSRF model by using the Knowledge Grid environment. The chapter focuses in particular on the application modeling. Applications are designed using a UML model, which is translated into a BPEL representation, in turn processed by the Knowledge Grid services for its execution. The second chapter addresses the problem of mining frequent closed itemsets in a highly distributed setting like a Grid. Authors show how frequent closed itemsets, mined independently at each site, can be merged in order to derive globally frequent closed itemsets. The last chapter reports progress made by using data mining techniques in the TELEMAC project concerned with enhancing the efficacy of anaerobic digestion in potentially unstable digesters. After placing the specific TELEMAC situation in a generic Grids context, authors present a classification approach to attributes for metadata and indicate some examples of model resource discovery.

From recent developments we can see the Grid moving from a computation platform to a data and knowledge management infrastructure. This trend needs new models, tools and solutions for enabling Grid computing to support advanced Grid applications. This book discusses some of the key technologies needed to support this trend and presents solutions recently designed to implement scalable applications.

We would like to thank all the participants for their contributions to making the KDM workshop a success. The workshop program committee for reviewing the submissions; the PSNC colleagues in Poznan for their support, and all the authors that contributed chapter for publication in this volume. A special thank to the Springer staff, Vladimir Getov and Paolo Trunfio for their assistance in editing the book.

Our thanks also go to the European Commission for sponsoring under grant number 004265 this volume of the CoreGRID project series of publications.

Domenico Talia, Angelos Bilas, Marios D. Dikaiakos

Contributing Authors

Pinar Alper School of Computer Science, The University of Manchester, United Kingdom (penpecip@cs.man.ac.uk)

Mario Antonioletti EPCC, The University of Edinburgh, JCMB, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, United Kingdom (mario@epcc.ed.ac.uk)

Malcolm Atkinson National e-Science Centre, 15 South College Street, Edinburgh, EH8 9AA, United Kingdom (mpa@nesc.ac.uk)

Sean Bechhofer School of Computer Science, The University of Manchester, United Kingdom (seanb@cs.man.ac.uk)

Angelos Bilas Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas, P.O.Box 1385, Heraklion, GR 71110, Greece (bilas@ics.forth.gr)

Neil P. Chue Hong EPCC, The University of Edinburgh, JCMB, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, United Kingdom (N.ChueHong@epcc.ed.ac.uk)

Carmela Comito DEIS, University of Calabria, Via P. Bucci 41C, 87036 Rende (CS), Italy (ccomito@deis.unical.it)

Antonio Congiusta DEIS, University of Calabria, Via P. Bucci 41C, 87036 Rende (CS), Italy (acongiusta@deis.unical.it)

Oscar Corcho School of Computer Science, The University of Manchester, United Kingdom (ocorcho@cs.man.ac.uk)

Shirley Crompton CCLRC, Daresbury Laboratory, Warrington WA4 4AD, United Kingdom (s.y.crompton@dl.ac.uk)

Maurice Dixon Computing, Communications Technology and Mathematics, London Metropolitan University, 31 Jewry Street, London, EC3N 2EY, UK (M.Dixon@Londonmet.ac.uk)

Michail D. Flouris Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada (flouris@cs.toronto.edu)

Julian R. Gallop e-Science, CCLRC Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, OX11 0QX, UK (J.R.Gallop@rl.ac.uk)

Carole Goble School of Computer Science, The University of Manchester, United Kingdom (carole@cs.man.ac.uk)

Asunción Gómez-Pérez Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660, Boadilla del Monte, Madrid, Spain (asun@fi.upm.es)

Anastasios Gounaris School of Computer Science, University of Manchester, UK (gounaris@cs.man.ac.uk)

Alex Gray Cardiff School of Computer Science, Cardiff University, Cardiff CF24 3AA, United Kingdom (w.a.gray@cs.cardiff.ac.uk)

Miguel Esteban Gutiérrez Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660, Boadilla del Monte, Madrid, Spain (mesteban@fi.upm.es)

Stratos Idreos CWI, Amsterdam, The Netherlands (S.Idreos@cwi.nl)

Andrew Jones Cardiff School of Computer Science, Cardiff University, Cardiff CF24 3AA, United Kingdom (Andrew.C.Jones@cs.cardiff.ac.uk)

Zoi Kaoudi Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece (zoi@di.uoa.gr)

Konstantinos A. Karasavvas National e-Science Centre, 15 South College Street, Edinburgh, EH8 9AA, United Kingdom (kostas@nesc.ac.uk)

Konrad Karczewski Institute of Computer and Information Sciences, Czestochowa University of Technology (xeno@icis.pcz.pl)

Ioannis Kotsiopoulos School of Computer Science, The University of Manchester, United Kingdom (ioannis@cs.man.ac.uk)

Manolis Koubarakis Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece (koubarak@di.uoa.gr)

Lukasz Kuczynski Institute of Computer and Information Sciences, Czestochowa University of Technology (lkucz@icis.pcz.pl)

Renaud Lachaize Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas, P.O.Box 1385, Heraklion, GR 71110, Greece (rlachaiz@ics.forth.gr)

Simon C. Lambert e-Science, CCLRC Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, OX11 0QX, UK (S.C.Lambert@rl.ac.uk)

Erietta Liarou Dept. of Electronic and Computer Engineering, Technical University of Crete, Greece (erietta@intelligence.tuc.gr)

Claudio Lucchese Dept. of Computer Science, Ca' Foscari University of Venice, Via Torino 155, 30172 Venezia, Italy (clucches@dsi.unive.it)

Matoula Magiridou Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece (matoula@di.uoa.gr)

Brian Matthews CCLRC, Rutherford-Appleton Laboratory, Didcot, Oxfordshire OX11 0AX, United Kingdom (b.m.matthews@rl.ac.uk)

Iris Miliaraki Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece (iris@di.uoa.gr)

Paolo Missier School of Computer Science, The University of Manchester, United Kingdom (pmissier@cs.man.ac.uk)

Zsolt Németh MTA SZTAKI Computer and Automation Research Institute, P.O. Box 63, Budapest, H-1518, Hungary (zsnemeth@sztaki.hu)

Raffaele Perego HPC Laboratory, ISTI-CNR of Pisa, via G. Moruzzi 1, 56124 Pisa, Italy (perego@isti.cnr.it)

Salvatore Orlando Dept. of Computer Science, Ca' Foscari University of Venice, Via Torino 155, 30172 Venezia, Italy (orlando@dsi.unive.it)

Rizos Sakellariou School of Computer Science, University of Manchester, UK (rizos@cs.man.ac.uk)

Claudio Silvestri Dept. of Computer Science, Ca' Foscari University of Venice, Via Torino 155, 30172 Venezia, Italy (silvestri@dsi.unive.it)

Jim Smith Newcastle University, Newcastle upon Tyne, UK (Jim.Smith@ncl.ac.uk)

Shoaib Sufi CCLRC, Daresbury Laboratory, Warrington WA4 4AD, United Kingdom (s.a.sufi@dl.ac.uk)

Domenico Talia DEIS, University of Calabria, Via P. Bucci 41C, 87036 Rende (CS), Italy (talia@deis.unical.it)

Paolo Trunfio DEIS, University of Calabria, Via P. Bucci 41C, 87036 Rende (CS), Italy (trunfio@deis.unical.it)

Paul Watson Newcastle University, Newcastle upon Tyne, UK (Paul.Watson@ncl.ac.uk)

Richard White Cardiff School of Computer Science, Cardiff University, Cardiff CF24 3AA, United Kingdom (r.j.white@cs.cardiff.ac.uk)

Philipp Wieder Grid Computing and Distributed Systems Group, Research Centre Jülich, 52425 Jülich, Germany (ph.wieder@fz-juelich.de)

Wolfgang Ziegler Fraunhofer Institute SCAI, Department of Bioinformatics, 53754 Sankt Augustin, Germany (wolfgang.ziegler@scai.fraunhofer.de)