



Project no. FP6-004265

CoreGRID

European Research Network on Foundations, Software Infrastructures and Applications for large-scale distributed GRID and Peer-to-Peer Technologies

Network of Excellence

GRID-based Systems for solving complex problems

**Condensed Report of the
D.KDM.02 – Proceedings of the First Workshop on
Knowledge and Data Management**

Submission date: October 23, 2006

Start date of project: 1 September 2004

Duration: 48 months

Organisation name of lead contractor for this deliverable: UNICAL

Revision draft

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	PU

Keyword List: Distributed Data Management, Information and Knowledge Management, Data Mining and Knowledge Discovery

1 Introduction

The second deliverable of the CoreGrid Institute on Knowledge and Data Management (D.KDM.02) is given by the proceedings of the “1st CoreGRID Workshop on Knowledge and Data Management in Grids” that has been held in Poznan (Poland) on September 13-14, 2005. In this workshop have been presented the major research results of the KDM Virtual Institute

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 forty-two 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 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 purpose of the workshop was presenting the major research results of the CoreGrid Institute on Knowledge and Data Management and bringing together CoreGRID researchers and invited external scientists doing research in Knowledge and Data Management in Grid and Peer-to-Peer Systems in Europe. 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.

1.1 Workshop Organizers and Committee

Here are listed the workshop organizers belonging to the Institutions involved in the KDM Institute.

Workshop Chair

Domenico Talia, University of Calabria

Organizing Committee

Alvaro Arenas, CCLRC-RAL, UK

Maciej Brzezniak, PSNC, Poland

Paolo Ciancarini, University of Bologna, Italy

Marios Dilaiakos, University of Cyprus, Cyprus

Michael Flouris, FORTH, Crete

Anastasios Gounaris, University of Manchester, UK

Philippe Massonet, CETIC, Belgium

Norbert Meyer, PSNC, Poland

Raffaele Perego, ISTI-CNR, Italy

Domenico Talia, University of Calabria, Italy

Local Arrangements Chair

Maciej Brzezniak, PSNC, Poland

1.2 Workshop Scientific Results

The workshop featured four invited talks by scientists not involved in the CoreGRID network and sixteen regular contributions by CoreGRID researchers. In particular, the invited speakers were as follows:

- R. Wyrzykowski from the Institute of Computer and Information Science of the Czestochova University of Technology. He presented 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.
- J. Smith from University of Newcastle gave a talk on fault-tolerance in distributed query processing (co-author P. Watson). That presentation described and evaluated a new scheme for adding fault-tolerance to distributed query processing through a rollback-recovery mechanism. The high level expression of user requests in a physical algebra offers opportunities for tuning the fault-tolerance provision so as to reduce the cost, and give better performance than employment of generic fault-tolerance mechanisms at the lowest level of query processing. Smith outlined how the publicly-available OGSA-DQP computational grid-based distributed query processing system can be modified to include support for fault-tolerance and presents a performance evaluation which includes measurements of the cost of both protocol overheads and rollback-recovery, for a set of example distributed queries.
- M.E. Gutierrez and his colleagues from Universidad Politécnica de Madrid presented “WS-DAIO: Ontology Access Provisioning in Grid Environments”. Current Grid architecture as defined by OGSA doesn’t explicitly consider ontology usage, nor do there exist protocols or standards in the Grid community for dealing with ontologies. This talk argued 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: the Semantic Grid. That work was performed in the OntoGrid project (FP6-511513).
- M. Koubarakis from Technical University of Crete gave a talk on “Semantic Grid Service Discovery using DHTs”. Koubarakis described the implementation of Atlas, a P2P system for the distributed storage and querying of RDF(S) metadata describing OntoGrid resources and services. Atlas uses state of the art DHT technology to design and implement a distributed system that is able to scale to hundreds of thousands of nodes and to large amounts of RDF(S) data and queries.

Besides the contribution of those invited speakers, the workshop included sixteen presentations given by CoreGRID researchers that compose the major scientific results of the KDM Institute at that time and give a clear picture of future activities. Those talks were arranged in five sessions focused on five different scientific topics:

- Distributed Data Storage,
- Data Access Services,
- Semantic Grid,
- Data Mining on Grids, and
- Metadata and Workflow.

All those sessions are concerned with key topics in the area of knowledge and data management on Grids and are related to research areas included within the scope of the KDM Institute.

In the Distributed Data Storage session was presented a work on Orchestra, a system offering scalable support for shared extensible virtual block devices. Orchestra allows extension storage functionality by providing hierarchies consisting of virtual devices layered over physical storage devices distributed in a commodity cluster. Then has been given a talk that examined the performance features of the iSCSI (Internet Small Computer Systems Interface) protocol and by comparing them to features of FCIP (Fibre Channel Internet Protocol). This work aimed related to the idea of building low-cost, large-scale and scalable distributed storage systems of commodity-based storage components. Finally, an autonomous distributed system built on top of the Violin framework has been proposed that is able to configure and reconfigure the storage hierarchy by detecting service breaches and take actions to eliminate them.

In the Data Access Service session have been presented three papers. The first has been 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. Therefore has been given a talk on data integration and query reformulation in service-based Grids. The XMAP data integration algorithm has been presented and service-based architecture for data integration-enabled query processing on the Grid has been discussed. At the end of the session, has been presented a work focused on the extension of the resource manager of Globus for providing transparent access to inhomogeneous data sources and data source engines such as a relational, flat and xml DBMS, file systems, document stores, and content management systems.

Three talks have been also given in the Semantic Grid session. The first one was concerned with a Reference Semantic Grid Architecture (RSGA) extending the Open Grid Services Architecture, by explicitly defining the mechanisms that allow for the explicit use of semantics. Then a second paper presented a core Grid ontology that defines the fundamental Grid domain concepts, vocabularies and relationships based on an abstract Grid model. The third paper proposed an ontology-based meta-scheduler as a Grid service for co-allocating resources on multiple grid nodes based on semantic information.

Future Grids can be effectively used as an infrastructure for distributed data mining and knowledge discovery in large data sets. To utilize Grids for high performance knowledge discovery, software tools and mechanisms are needed. The session on Data Mining on Grids included four papers discussing distributed data mining with network of sensors, algorithms for distributed computation of frequent itemsets, WSRF-based services for distributed data mining and the architecture of a Grid search engine named Grid@home. All those topics are very significant for the implementation of Knowledge Grids having the ability of discovering and managing distributed knowledge.

The last session on Metadata and Workflow included talks on a scientific metadata model (the CCLRC model) designed to capture the high level information pertaining to scientific studies and the data that they produce. As it is designed to capture metadata across a wide range of scientific disciplines, it is designed to be sufficiently generic to capture standard relationships common across science, allowing interoperability with common level of granularity useful to any study-data. The other two talks concerned Grid workflow optimization with inferential reasoning and resource discovery issues in Grid environments.

All the given lectures gave a wide spectrum of research activities carried out in the KDM Institute and in other European projects in the area of knowledge and data management. As a result of the workshop, the publication of a post-workshop book in the CoreGRID series of Springer has been planned. The book will include a full and revised version of a selection of papers presented at the workshop. It will have a world-wide audience and might become a reference scientific text in the area of Knowledge and Data Management on Grids.

Copies of the slides of presentations at the workshop can be found in the CoreGRID BSCW server at:
<http://www.ercim.org/bscw/bscw.cgi/0/59833>