GRID TECHNOLOGY FOR INTELLIGENT SYSTEMS AND DATA MINING

Boštjan Brumen
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, Si-2000 Maribor, Slovenia
e-mail: bostjan.brumen@uni-mb.si

Aida Kamišalić
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, Si-2000 Maribor, Slovenia
e-mail: aida.kamisalic@uni-mb.si

David Riaño
Universitat Rovira I Virgili, Departament d'enginyeria informàtica i matemàtiques
Av. Països Catalans, 26, Campus Sescelades
43007 Tarragona, Spain
e-mail: drianyo@etse.urv.es

Tatjana Welzer
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, Si-2000 Maribor, Slovenia
e-mail: welzer@uni-mb.si

ABSTRACT

GRID technology is a new emerging computation paradigm that combines distributed computers to solve a time-consuming problem using idle computing cycles. GRID computing has emerged as an important new field, distinguished from conventional distributed computing by its focus on large-scale inter-organizational resource sharing, innovative applications, and high-performance orientation. It is an evolving area of computing, where standards and technology are still being developed to enable this new paradigm.

Intelligent systems can use the emerging new field in several ways. The most important impact GRID systems can have on intelligent systems is the potential to solve computationally and data intensive problems, which are inherent to the data mining. Secondly, because of the new paradigm, the already solved problems need to be re-addressed in the light of new architecture and new requirements.

In the paper, we outline the open research questions in the field of data mining in the light of the new GRID computing paradigm.

1. INTRODUCTION

Web services, GRID services and the Semantic Web are currently three strands of Internet development. The Internet itself is envisioned to become a programming platform to support real time, fully customized and customizable service creation and development. The Internet will become a utility that binds, connects and mediates all activities and functions of society. Web connects together vast quantities of information.

It has been developed as a medium of sharing the documents among the people rather than for sharing the data and information that can be processed automatically. Today, information tends to be bound up in the data format of the application for which it is designed. Without the right application, the information content of data is inaccessible [XML Semantic Web, 2003].

In the same way that the Web connects together vast quantities of information, the Semantic Web will make the information content of exchanged data accessible to any application that understands semantic web protocols.

The GRID technology will link computing power and storage capacity into a single virtual source of processing capability. Evolution of the GRID started in the early 1990s, when the first metacomputing or GRID environments have emerged. Today we can identify three stages of GRID evolution:

- first generation systems that started as a project to link supercomputing sites,
- second generation systems with a focus on middleware to support large scale data and computation, and
- third generation systems, where emphasis is on distributed global collaboration.

There are addressed issues that can be solved by GRID technology, such as flexible sharing relationships, sharing of varied resources and precise level of control over shared resources. Two key characteristics of third generation systems are solutions involved increasing adoption of a service-oriented model and increasing attention to metadata [Foster, 2001].

Currently, we can talk about GRID computing as a form of distributed computing. It has hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities. As a form of distributed computing, GRID involves coordinating and sharing computers, applications, data, and storage and network resources across dynamic and geographically dispersed organizations. The sharing is necessarily highly controlled, with resource providers and consumers defining clearly and carefully what is shared, who is allowed to share and the conditions under which the sharing occurs. From the aspect of hardware, the GRID is a form of networking, which harnesses unused processing cycles of all computers in a network for solving problems too intensive for any stand-alone machine. As software infrastructure, the GRID comprises a set of protocols, services and tools to achieve secure, reliable and scaleable resource sharing across institutional boundaries.

GRID is considered as a set of additional protocols and services that build on Internet protocols and services to support the creation and use of computation and data enriched

environments. There is focus on areas including security (how to manage credentials, policies across multiple institutions), resource management (remote access to computing and data resources), information query (discovery, configuration and status about resources, services and organizations) and data management (services to locate access, transport and replicate data sets). Groups of organizations that are building experimental and production GRID infrastructures to share resources for specific purposes are called virtual organizations (VO).

GRID computing, however, does not imply unrestricted access to resources; the resources are shared in a controlled manner. Resource owners will typically want to enforce policies that constrain access according to group membership, ability to pay, and so forth. Hence, accounting is important, and a GRID architecture must incorporate resource and collective protocols for exchanging usage and cost information, as well as for exploiting this information when deciding whether to enable sharing.

In the continuation, we first describe Open GRID Service Architecture in Section 2 and continue in Section 3 with a list of the main components of the architecture. Next, in Section 4, we outline the types of the grids. In section 5 we first focus on the experimental data grids, which are of interest in this paper and then we describe the intelligent systems in the light of the new GRID infrastructure. Here we outline the open research questions of data and computationally intensive tasks that constitute intelligent systems. We conclude the paper with final remarks in Section 7.

2. OPEN GRID SERVICE ARCHITECTURE - OGSA

The Open GRID Service Architecture gives an underlying framework on which to build. It supports the creation, maintenance and application of ensembles of services maintained by VOs. Everything is represented by a service (a network enabled entity that provides some capability through the exchange of messages). These services are: computational resources, storage resources, networks, programs, and databases. In a service-oriented view, the interoperability problem can be divided into two subproblems, the definition of service interfaces and the identification of the protocol(s) that can be used to invoke a particular interface and agreement on a standard set of such protocols.

A service-oriented view allows us to address the need for standard interface definition mechanisms, local/remote transparency, and adaptation to local OS services, and uniform service semantics. It also simplifies virtualization (the encapsulation behind a common interface of diverse implementations). Virtualization allows for consistent resource access across multiple heterogeneous platforms with local or remote location transparency, and enables mapping of multiple logical resource instances onto the same physical resource and management of resources within a VO based on composition from lower-level resources. Virtualization allows the composition of services to form more sophisticated services-without regard for how the services being composed are implemented.

Ability to virtualize and compose services depends on more than standard interface definitions. It is required standard semantics for service interactions so that, for example, different services follow the same conventions for error notification.

GRID services are characterized by the capabilities that they offer. A GRID service implements one or more interfaces, where each interface defines a set of operations that are invoked by exchanging a defined sequence of messages, and follows specific conventions.

3. MAIN COMPONENTS OF GRID ARCHITECTURE

GRID architecture is, as mentioned, a set of several protocols that form logical components.

The GRID Security Infrastructure (GSI) is used for secure communication over an open network. It provides a number of services for GRIDs including mutual authentication and single sign-on. GSI is based on public key encryption, X.509 certificates, and the Secure Sockets Layer (SSL) communication protocol. Extensions to these standards have been added for single sign-on and delegation.

A central concept in GSI authentication is the certificate. Every user and service on the GRID is identified via a certificate, which contains information vital to identifying and authenticating the user or service. A GSI certificate includes four primary pieces of information:

- A subject name, which identifies the person or object that the certificate represents;
- The public key belonging to the subject;
- The identity of a Certificate Authority (CA) that has signed the certificate to certify that the public key and the identity both belong to the subject;
- The digital signature of the named CA.

GSI certificated are encoded in the X.509 certificate format, a standard data format for certificates established by the Internet Engineering Task Force (IETF).

Information service component known as Monitoring and Discovery Service (MDS) is used for querying system information from a rich variety of system components, and for optionally constructing a uniform namespace for resource information across a system that may involve many organizations.

The GRID Resource Information Service (GRIS) provides a uniform means of querying resources on a computational GRID for their current status. Such resources include: computational nodes, data storage systems, scientific instruments, network links and databases.

A GRID Index Information Service (GIIS) may be installed and run on one or more systems. Once a GIIS is running, the GRIS running on each system in the VO can be configured to register with the GIIS so that people or applications can search the GIIS for participating systems and query their configuration data. For example, GIIS could list all of the computational resources available within a confederation of laboratories, or all of the distributed data storage systems owned by a particular agency.

Resource Allocation Manager (RAM) processes the requests for resources for remote application execution, allocates the required resources, and manages the active jobs.

4. TYPES OF GRIDS

From an application perspective, there are two types of GRIDs: computational and data GRIDs. A computational GRID is a collection of distributed computing resources, within or across locations that are aggregated to act as a unified processing resource or virtual supercomputer. These computing

resources can be either within or between administrative domains. Collecting these resources into a unified pool involves coordinated usage policies, job scheduling and queuing characteristics, GRID-wide security, and user authentication.

The data GRIDs enable users and applications to manage and efficiently use database information from distributed locations. Data GRIDs rely on software for secure access and usage policies. They can be deployed within one administrative domain or across multiple domains. Here become critical the GRID software and policy management.

The majority of the early GRID deployments have focused on enhancing computation. But, access to distributed data is typically as important as access to distributed computational resources. Distributed scientific and engineering applications often require access in many places by many people, as in virtual collaborative environments, to large amounts of data (terabytes or petabytes). There are data GRID efforts for identifying, prototyping, and evaluating the key technologies required to support data GRIDs for scientific and engineering collaborations.

High-performance, distributed computing applications require two fundamental services: secure, reliable, efficient transfer of data in wide area environments and the ability to register and locate multiple copies of data sets.

There are in use a number of storage systems designed to satisfy specific needs and requirements for storing, transferring and accessing large datasets. These include the Distributed Parallel Storage System (DPSS) and the High Performance Storage System (HPSS), which provide high-performance access to data and utilize parallel data transfer and striping across multiple servers to improve performance. The Distributed File System (DFS) supports high-volume usage, dataset replication and local caching. The Storage Resource Broker (SRB), which connects heterogeneous data collections, provides a uniform client interface to storage repositories, and provides a metadata catalog for describing and locating data within the storage system. Most of these storage systems utilize incompatible and often unpublished protocols for accessing data, and therefore require the use of their own client libraries to access data. To overcome these incompatible protocols, GRIDFTP was proposed as a reliable, secure, high-performance data transfer protocol, and Replica Management for management of multiple copies of files and collections of files.

Distributed scientific and engineering applications require transfers of large amounts of data (terabytes or petabytes) between storage systems, and access to large amounts of data (gigabytes or terabytes) by many geographically distributed applications and users for analysis, visualization, etc. The lack of standard protocols for transfer and access of data in the GRID has led to a fragmented GRID storage community. Users who wish to access different storage systems are forced to use multiple protocols and APIs, and it is difficult to efficiently transfer data between these different storage systems.

As a solution for these problems GRIDFTP was introduced, as a common data transfer and access protocol that provides secure, efficient data movement in GRID environments. It extends the standard FTP protocol. GRIDFTP must be extensible in order to support future innovations by storage system users and developers [Allcock, 2000]. It provides following protocol features:

- GSI and Kerberos support – must support GSI and Kerberos authentication, with user controlled setting of various levels of data integrity and confidentiality.
- Parallel data transfer – must support parallel data transfer between a single client and a single server, and between two servers.
- Partial file transfer – must support transfer of partial files and transfer of regions of a file.
- Third-party control of data transfer – it is necessary to provide third-party control of transfers between storage servers.
- Striped data transfer – make possible partitioning data across multiple servers.
- Support for reliable data transfer – exploits and extends features for restarting failed transfer.

Replica Catalog allows users to register files as logical collections and provides mappings between logical names for files and collections and the storage system locations of one or more replicas of these objects. The catalog registers three types of entries: logical collections, locations and logical files. A logical collection is a user-defined group of files. It is more convenient to register and manipulate groups of files as a collection, rather than requiring that every file be registered and manipulated individually. Location entries in the replica catalog contain all the information required for mapping a logical collection to a particular physical instance of that collection. It may register information about the physical storage system, such as the hostname, port and protocol. The replica catalog includes optional entries that describe individual logical files because users and applications may want to characterize individual files. Logical files are entities with globally unique names that may have one or more physical instances [Allcock, 2000].

Replica Management is the process of keeping track of where portions of the data set can be found. Complete data set may exist in one or possibly several physical locations. It is likely that few universities, research laboratories or individual researchers will have insufficient storage to hold the complete copy. Instead, they will store copies of the more relevant portions of the data set on local storage for faster access.

5. DATA GRIDS, INTELLIGENT SYSTEMS AND DATA MINING

In this section we first briefly list only the GRID projects that are related and relevant to the data mining. The list is far from being complete and serves only for better understanding of the connection between the grids and the intelligent systems, which is the focus of this section.

Data Grids

A scientific collaborations that are currently building data GRIDs for their own use are: the Earth Systems GRID (ESG) [Earth GRID, 2003], the European DataGRID [EU Grid, 2003], the GriPhyN [Physics GRID, 2003], NEESGRID [Nees Grid, 2003] and the Particle Physics Data GRID [Particle GRID, 2003]. The Earth Systems GRID is an experimental data GRID for scientists collaborating on climate studies. The data is collected from ground and satellite-based sensors or generated via simulations. The European DataGRID project is an international project for shared cost research and technological

development. It is focused on solving the data management and analysis needs of the world-wide high energy physics community. The GriPhyN collaboration is a team of experimental physicists and information technology researchers. NEESGRID is a virtual laboratory for the earthquake engineering community. The goal is to develop a systems design for integrating experimental and computational facilities for use by the earthquake engineering community. The Particle Physics Data GRID collaboration focuses on the needs of high-energy physicists. It is working toward the creation of a collaboratory for particle physicists attempting to experimentally verify theories regarding the fundamental nature of matter and energy.

Intelligent Systems and Data Mining

An intelligent system has to perceive its environment, to act rationally towards its assigned tasks, to interact with other agents and with human beings. An intelligent system is usually a software systems and/or a physical machine with sensors and actuators.

The capability to act rationally is covered by topics such as computer vision, planning and acting, robotics, multiagents systems, speech recognition, and natural language understanding. They rely on a broad set of general and specialized knowledge representations and reasoning mechanisms, on problem solving and search algorithms, and on machine learning techniques.

The research in the area of artificial intelligence (AI) has seen its ups and downs since Rosenblatts's introduction of the concept known as perceptron in 1959. In the early 1970's, it became clear that AI would not be the panacea for all computing problems and that a lot of additional work would be needed.

In the 1980's and 1990's, we have seen an enormous expansion of the computing and consequently the rise of the need for intelligent systems that would help humans cope with daily life. At the same time, the computing capacity has grown almost exponentially, making feasible the ideas that have been inapplicable due to the lack of computing power.

The vast amount of data collected by the electronic equipment, operative in such diverse fields as geology, business, astronomy, or medicine is hiding important and valuable knowledge. The amount of data doubles every 15 months [Brumen, 2002].

The knowledge acquisition from data is a very important research area. Namely, the knowledge extracted from the data can actually represent the human decisions. The machines having the ability to learn from humans (and act in the same way) can be regarded as intelligent systems.

The amount of data on the other hand prevents from conducting the analyses and knowledge acquisition on single machines or even on a cluster of machines available within a single organization.

The application that can gain the most form the GRID infrastructure is data mining (DM).

However, several research questions remain open. Let us briefly examine a few we find most relevant.

1. Parallelization of DM tasks for GRID infrastructure

The state-of-the-art parallel algorithms are developed for multi-processor, cluster machines. GRID infrastructure is different from cluster machines; the connection between the GRID machines is not necessarily high-speed network, but rather a typical Internet connection. The parallel algorithms will have to be adapted to this new environment and the task schedulers will need to take into the account the dynamic nature of availability of machine cycles in GRID network of machines.

2. Run-time performance estimation of DM tasks

The run-time performance estimation of DM tasks is crucial for the efficient scheduling. The scheduling broker takes allocation and scheduling decisions, and builds the *execution plan*, establishing the sequence of actions that have to be performed in order to prepare execution (e.g., resource allocation, data and code deployment), actually execute the task, and return the results to the user. The execution plan has to satisfy given requirements (such as performance, response time, and mining algorithm) and constraints (such as data locations, available computing power, storage size, memory, network bandwidth and latency) and to maximize or minimize some metrics of interest (e.g. throughput, average service time). In its decision making process, this service has to exploit a composite performance model which consider the actual status of the GRID, the location of data sources, and the task execution behavior. The broker needs quite detailed knowledge about computation and communication costs to evaluate the profitability of alternative mappings [Orlando, 2002].

3. Accuracy performance estimation of DM tasks

The question of accuracy performance estimation of DM tasks that involve huge amounts of data is quite important. In many cases, large amount of resources may be wasted, but the final user's needs are not satisfied. In classification task of DM, the performance measure is accuracy of the model built. It would be very helpful if the performance estimation were available very quickly. The estimation could be obtained by running the classification task on a sample in a parallel manner on several machines, each with different sample size. The broker would collect the information and the estimation unit would estimate the performance based on the results on the sample.

The problem with performance estimation is that the performance is varying from data set to data set, and depends on the algorithm used. The estimation thus has to be done for each data set and the results can not be directly implied for the next DM task, as opposed to the estimation of the run-time performance of the algorithms, which can be re-used.

4. Privacy of data

The data that is used in data mining tasks is often very sensitive and needs special attention [Brumen, 2003]. The data within GRID infrastructures are often transferred among the nodes. The transfer of data over the open network itself is secure using the services provided by the GRID Security Infrastructure. Once the data are transferred, they are decrypted and stored locally, with local security policies which may differ from the one of the data owner. Additionally, the intruder may attack the (privacy of) data on the GRID node to which the data are transferred; the intruder may even have a full control over the node. To prevent such attacks the GRID services in general and especially the data mining services must provide for means for preserving the privacy of data once the data are transferred to the nodes. The further research of privacy and confidentiality of data within GRID infrastructure is needed, so that the data mining is still possible, but attributes such as privacy and confidentiality are preserved.

6. CONCLUSION

The research on GRID infrastructures has opened a completely new area of research also in the area of the existing intelligent systems, which need to be adapted to the new paradigm.

The core mission of the intelligent systems is to support the end user in daily life. As we have the power tools that enhance the physical power of humans, so we have intelligent systems that enhance the brainpower of humans in tasks that are dull, repeating, require quite much iteration and involve large amounts of data.

The GRID infrastructure enables the execution of tasks that are otherwise infeasible on single machines or even on a cluster of machines within an organization.

The data mining, which is inherently data and processing intensive process, can gain much from the underlying new processing and storage paradigm. However, several problems that were already solved for the existing state-of-the-art, need to be re-addressed in the face of the new requirements and conditions. Successful GRID data mining depends on successful parallel execution of tasks on independent, non-clustered machines. The performance estimation in such setting is crucial for successful completion of the tasks, and the problem needs to be addressed in the future.

REFERENCES

[Allcock, 2000] Allcock, Bill; Foster, Ian; Tuecke, Steven; Chervenak, Ann; Kesselman, Carl: Protocols and Services for Distributed Data-Intensive Science; *ACAT2000 Proceedings*, pp. 161-163, 2000

[Allcock, 2001] Allcock, Bill; Bester, Joe; Bresnahan, John; Chervenak Ann; Foster, Ian; Kesselman, Carl; Meder, Sam; Nefedova, Veronika; Quesnel, Darcy; Tuecke, Steven: Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing, *IEEE Mass Storage Conference*, 2001

[Brumen, 2002] Brumen, Boštjan; Golob, Izidor; Welzer, Tatjana; Rozman, Ivan; Družovec, Marjan; Jaakkola, Hannu: Data Protection for Outsourced Data Mining, Informatica, Vol. 26, No. 2, 2002.

[Brumen, 2003] Brumen, Boštjan; Golob, Izidor; Welzer, Tatjana; Rozman, Ivan; Družovec, Marjan; Jaakkola, Hannu: An algorithm for protecting knowledge discovery data. *Informatica*, Vol. 14, no. 3, pp. 277-288, 2003.

[Chervenak, 2001] Chervenak, Ann; Foster, Ian; Kesselman, Carl; Salisbury, Charles; Tuecke, Steven: The Data GRID: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets, *Journal of Network and Computer Applications*, 23:187-200, 2001

[EU Grid, 2003] European DataGRID Project, http://eu-dataGRID.web.cern.ch/eu-dataGRID/

[Foster, 2001] Foster, Ian; Kesselman, Carl; Tuecke, Steven: The Anatomy of the GRID, Intl. J. Supercomputer Applications, Vol 15, No. 3, 2001

[Foster, 2002] Foster, Ian; Kesselman, Carl; Nick, Jeffrey M.; Tuecke, Steven: The Physiology of the GRID: An Open Grid Services Architecture for Distributed Systems Integration; Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002

[Globus GRID, 2003] Globus Data GRID, http://www.globus.org/dataGRID/

[Nees Grid, 2003] NEESGRID, http://www.neesGRID.org/

[Particle GRID, 2003] Particle Physics Data GRID, http://www.ppdg.net/

[Physics GRID, 2003] GRID Physics Network, http://www.griphyn.org/

[Semantic GRID, 2003] Semantic GRID, http://www.semanticGRID.org/

[XML Semantic Web, 2003] XML, Semantic Web, http://www.xml.com/pub/a/2000/11/07/semanticweb/

[Earth GRID, 2003] Earth System Grid, http://www.earthsystemgrid.org/

[Orlando, 2002] Orlando, S.; Palmerini, P.; Perego, R.; Silvestri, F.: Scheduling High Performance Data Mining Tasks on a Data Grid Environment, In Monien, B.; Feldmann, R. (Eds.): Proceedings of Euro-Par Conference, Lecture Notes in Computer Science Vol 2400, pp. 375–384, 2002