International Journal of Computer and Communication Technology

Volume 6 | Issue 3 Article 6

July 2015

KNOWLEDGE DISCOVERY IN DATA GRID WITH ADVANCE RESERVATION FOR METEOROLOGY

M. THANGAMANI
Dept. CSE, Kongu Engineering College, manithangamani2@gmail.com

J.MALAR VIZHI
Dept. CSE, Kongu Engineering College, vizhijegan@gmail.com

G. NANDHINI
Dept. CSE, Kongu Engineering College, nandhujoshna@gmail.com

KNOWLEDGE DISCOVERY IN DATA GRID WITH ADVANCE RESERVATION FOR METEOROLOGY

M. THANGAMANI[1], J.MALAR VIZHI[2] & G.NANDHINI[3]

[1,2&3]Dept. CSE, Kongu Engineering College
Email:manithangamani2@gmail.com, vizhijegan@gmail.com, nandhujoshna@gmail.com.

Abstract- In weather forecasting as well as in other scientific domains, large-scale and distributed data collections are emerging as critical community resources. With the development of Grid technologies, data management and sharing can be exploited in such an efficient way. In this paper, we present our approach in constructing a portal-based Meteorological Data Grid System with first application in Weather Forecasting .Our system architecture has three layers. The first layer is Modeling System that uses Numerical Weather Prediction (NWP) models to generate forecast data automatically in GRIB/NetCDF format. The second layer is a Data Grid which provides users with secure and easy access to distributed meteorological datasets. It also addresses authentication / authorization for secure transfers, mechanisms for scalable data replication, and technologies for searching relevant datasets regarding metadata provided by users. A Grid Portal, the third layer, is built for purpose of easy using the system. We also allow advance resource reservation to have exclusive access to Grid resources

1. INTRODUCTION

Nowadays, Grid Computing has emerged as a new technology for efficient resource sharing among many organizations known as Virtual Organization. The grid computing paradigm essentially aggregates the view on distributed hardware and software resources. By Grid Computing, various resources such as computers, networks and storage facilities (databases, file systems, etc) are organized in a uniform framework. In the Grid Computing environment, scientists from different geo-graphical locations have fully access to shared computing and data resources. In this environment, many scientific, collaborative applications can be executed efficiently. Grid Computing is one of the most important technologies in the future.

Currently, Grid Computing is applied to solve distributed, large-scale, and collaborative problems in many domains such as: High Energy Physics, Bio-Informatics, Virtual Observation, Environmental Science, etc. Weather forecasting is also such a domain where Grid Computing's advantages would show the best.

Weather forecasting plays a crucial role in our society nowadays. The prediction of changes in weather conditions is especially important in a tropical country like India where 70% of citizens relies their livings on agricultural activities. A precise knowledge of the evolution of weather situation is really helpful in order not only to issue civil protection warnings in case of crisis such as flooding, heavy rains, tornadoes, tsunami, etc, but also provides useful information that helps farmers to take appropriate actions in agricultural works.

Today, Numerical Weather Prediction (NWP) method is popularly used to produce meteorological forecast. It uses mathematical models of the atmosphere to predict the weather. An NWP model, in this context, is a software that generates meteorological information for the future times at given positions and altitudes. Its manipulation of huge datasets and complex computations requires powerful supercomputers or computer clusters. Currently there are a number of regional NWP models being used for making short-term weather forecast such as MM5 (Fifth-Generation Mesoscale Model), WRF (Weather Research and Forecasting Model), and HRM (High Resolution Model). These models are separately deployed in different organizations such as High Performance Computing (HPC) .However, there is a lack of a common framework for managing, sharing the computing and storage resources as well as local observations and forecast data among these institutes.

With the advent of Grid Computing Technology, there is a feasible answer for this problem: the data grid. Data Grid is a grid computing system that deals with data - the controlled sharing and management of large amounts of distributed datasets. Building a data grid that connects several HPC Centers (called Meteo-Data Grid) for sharing meteorological datasets is an essential, crucial, and suitable solution because of the following reasons. First, the meteo-data grid infra-structure will help exploiting hardware and software resources in diverse HPC centers. Second, the data grid will provide researchers a central access point to all resources of the grid via a grid portal. Through the portal, scientists from different locations will gain an easy access to all large-scale, dispersed meteorological datasets. Additionally, the data grid will offer an efficient and reliable mean for storing and querying meteorological datasets via many useful services such as: a metadata service for querying datasets by description attributes; a replica service that creates replicas in order to improve availability and access performance of the datasets; a fast, secure and consistent file transfer service for moving large-size datasets from one site to another. Finally, the meteo-data grid will be a virtual organization of meteorology community. Every meteorologist can get

easy access to various resources in the grid. Instead of storing all of the datasets, a HPC center will exploit the capacity of others HPC centers' storage facilities. Thus, the cost for maintaining data in each meteorology center will reduce.

In this paper, we present the proposed meteo-data grid architecture and its first implementation using Grid Technology for weather forecasting. We also describe the prototype of the grid portal that enables researchers to access data resources easily via a web browser.

The rest of this paper is structured as follows. Section 2 briefly explains the background knowledge of grid computing and grid portal. The detailed description of our Meteo-Data Grid System is discussed in section 3. The prototype implementation of the system and some technical issues is presented in section 4. Finally, section 5 presents a survey of some similar weather grid projects and discussion about future works.

2. BACKGROUND

1.1 Grid Computing

Grid Computing System is an incorporated computing environment with some specific characteristics as the followings: distributed, adaptable, extensible, and heterogeneous. There are two types of grid system: Computational Grid and Data Grid. The former provides services that address the management of scattered computing resources; and the latter deals with the management of distributed data resources such as databases and large-size datasets at grid nodes. Strictly speaking, this categorization is relative. Data Grids are often combined with Computational Grids in order to solve real computational science applications and to increase performance. In any grid computing environment, data management is a vital job, because data is one of the most crucial resources of applications. Data Grid provides data needed by Computational Grid. And vice versa, output data that results from computational processes of Computational Grid will be stored in Data Grid. Figure 1 illustrated this relationship between Data Grid and ComputationalGrid.

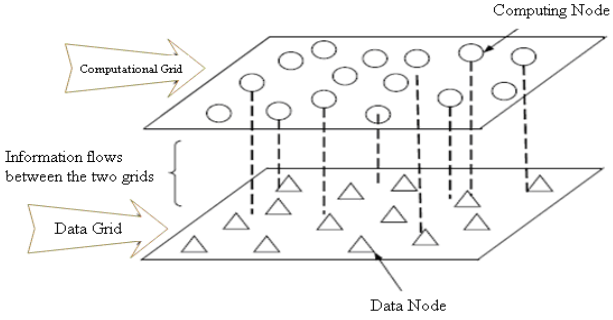


Figure 1 Relationship between Data Grid and Computational Grid

Grid Computing System Architecture has five layers, as follows:

- Fabric: this lowest layer comprises the basic resources such as computing and storage systems, from which Grid Computing System is built.
- Connectivity: this layer deals with communication and authentication issues.
- Resource: services at this layer concern with providing secure remote access to various resources in the grid.
- Collective: the synchronized management of multiple resources problem is addressed at this layer.
- Application: this layer contains applications of specific disciplines that exploit grid computing power provided by the lower layers.

At present, Globus Toolkit is considered a defacto standard grid middleware. Globus Toolkit contains a variety of services and APIs that provide both distributed data and computing resource management abilities. Using Globus Toolkit, you can easily deploy a grid system as well as develop grid applications.

1.2 Grid Portal

In the Grid Computing Environment, resources and services are complex and distributed, and usually change. Furthermore, almost all grid users lack of skills to deal with the complicated, heterogeneous grid environment. Thus, it is difficult for nonprofessional users to work directly with dispersed and varied grid resources.

Grid Portal is one of the most efficient and easiest mechanisms for grid users to interact with a Grid Computing System. Grid Portal works as an interface between grid users and lower grid services. Via Grid Portal, scientists from various domains can **utilize** distributed grid services, applications, datasets and tools whenever they have access to the Internet. Grid Portal hides grid users from the complexities of underlying systems. By using Grid Portal, users do not need to download any specific grid software or configure network policies.

You can develop a grid portal by utilizing various Portal Toolkits such as GridSphere and Grid Port. In this paper, we use GridSphere as a framework to develop the Meteo-Data Grid Portal.

2. THE METEO-DATA GRID

Everyday, several High Performance Computing Systems run NWP models (HRM, WRF and MM5) to produce meteorological datasets in GRIB or NetCDF file format. These datasets are automatically stored in data servers. A visualization application server takes datasets of each day as input data, and then creates images depicting corresponding temperature, humidity, wind velocity, and rainfall as output. These daily images are available to nonprofessional users -

who are novices in the meteorological fields - in the grid portal.

Moreover, meteorology researchers can submit searching queries for desired datasets via the portal. Whenever an expected dataset is found, researchers can issue a visualization job to the visualization server. Equivalent images of the corresponding dataset are then produced, and sent back to the portal server in order to display to the end users.

Figure 2 is the illustration of our system architecture. Our system has four main components:

Weather forecast modeling systems. These systems use computer clusters or supercomputers to run regional NWP models such as HRM, WRF and MM5, etc. They are datasets providers. The total amounts of data produced by these systems each day can be over 1 GByte.

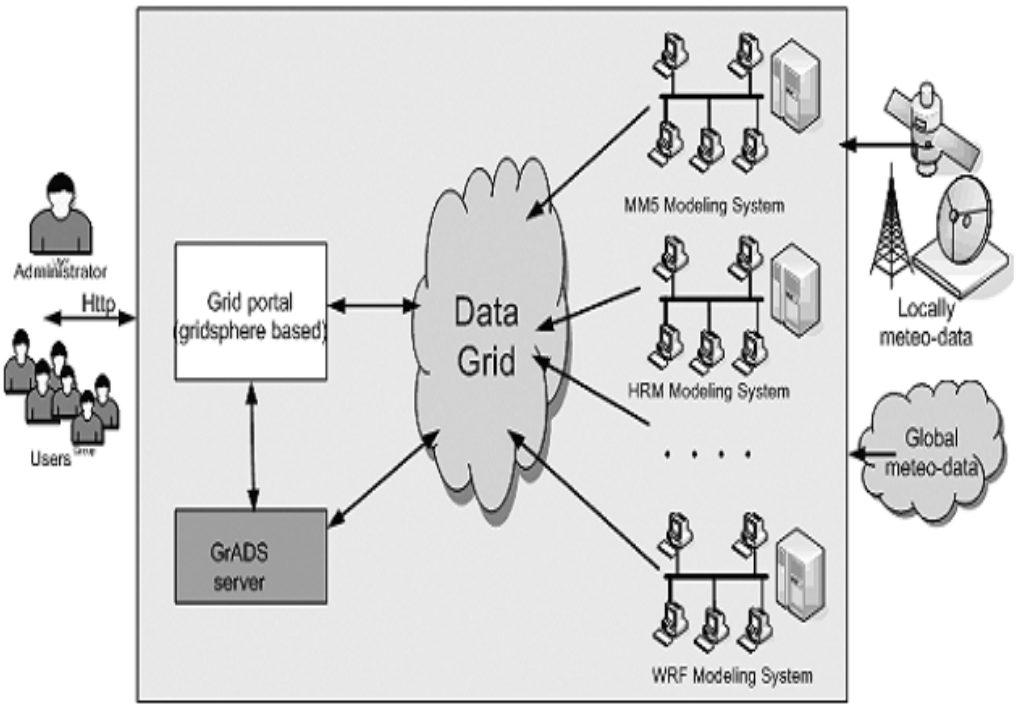
Data grid services. Data Grid plays the role of a virtual repository for storing distributed meteorological datasets that can be accessed later. It addresses authentication, authorization for secure transfers, mechanisms for scalable data replication, and technologies for searching relevant datasets regarding metadata provided by users.

Visualization System. The main function of Visualization System is to generate daily images of temperature, rainfall, wind velocity, etc, from meteorology datasets. It also has the ability of responding to users' visualization request.

Grid Portal: Grid Portal is the gateway that connects end-users with the Data Grid. It offers a single access point to all distributed data and computing resources of the system.

Figure 3 describes the MM5 modeling system flow chart. The primary programs work in a process as follows. Terrestrial and isobaric meteorological data are horizontally inter-polated - by TERRAIN and REGRID - from a latitude-longitude mesh to a variable high-resolution, rectangular domain on a Mercator, Lambert conformal, or polar stereographic projection. Since the inter-polation does not provide much mesoscale detail, the interpolated data may be enhanced (program RAWINS or little_r) with observations from the standard network of surface and rawinsonde stations using either a successive-scan Cressman technique or multiquadric scheme. Program INTERPF performs the vertical interpolation from pressure levels to the sigma coordinate system of MM5. Program MM5 – the numerical weather prediction part of the modeling system – gets initial and boundary conditions from INTERPF (and TERRAIN if needed) to initialize itself and produce forecast data.

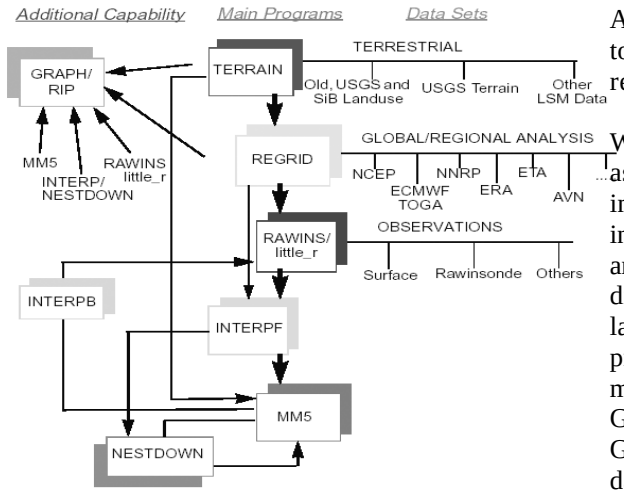
We now use GFS (Global Forecast System – available at NCEP site) analysis, combining with local observation data as input for the MM5 modeling system. The system updates automatically 4 times per day and generates forecast for 78 hours to the future in 1 hour step



.

Figure 2 The MM5 Modeling System Flow Chart

4. SYSTEM PROTOTYPE



We can use Globus Toolkit 4.0 to implement the Data Grid layer. Several components of GT4 used in our system are:

- ❖ GridFTP – Grid File Transfer Protocol and RFT - Reliable File Transfer: reliable file transfer services that enable fast, secure, and efficient transfer of large and numerous data files between data servers.
- ❖ RLS – Replica Location Service: a service that provides the management and tracking of replicated files at multiple sites to reduce data access latency and the demands on network bandwidth.
- ❖ DRS – Data Replication Service: a high level service that use RFT and RLS to replicate datasets.
- ❖ MCS – Metadata Catalog Service: a service that helps end-users to discover the desired files via the specification of metadata.
- ❖ GSI and CAS: a Grid Security Infrastructure used for addressing authentication and authorization in the grid environment.

The datasets computed from MM5 application are automatically stored in the data grid with relevant metadata. The metadata is generated by additional tools such as WGRIB and NCDump. The description attributes are extracted from the datasets themselves and associated with some additional information such as generating date, publisher, and model, etc. Then, they are stored in the MCS server.

There are several technical issues that need to be considered. First, because the MCS Server of Globus Toolkit 4.0 provides a central point access to metadata of all datasets in the grid, then if this server is down, the system will not function correctly. To prevent this bottle-neck failure of the single access point for MCS, a mirrored MCS Server should be used to automatically replace the failed one. Second, in order to select a suitable data server for storing and/or replicating data provided by a data producer, our system also need a monitoring service that determines the capacity and availability of each node in the grid. And finally, the ability to easily reconfigure the topology of the grid when a certain node is added or removed is vital to our system.

We used GrADS – Grid Analysis and Display System as the visualization server for rendering corresponding images regarding relevant datasets. GrADS is an interactive desktop tool for accessing, manipulating and visualizing earth science data. It uses a 4-dimensional data environment including longitude, latitude, vertical level, and time. It also provides a programmable interface to support the development of more sophisticated visual applications. In our system, GrADS server takes meteorological datasets (such as GRIB file, NetCDF, etc) as inputs; and then it creates daily images for weather forecasting as outputs. After receiving users' request parameters such as a particular field (temperature, winds, etc), level (1000mb, 950mb, etc), and time slice (0h, 6h, etc), GrADS server generates corresponding images. Additionally, based on GrADS API, we have developed an on-demand service that receives user's requests to create a graphic visualization for a given GRIB file.

Our Data Grid Portal is built on GridSphere framework. Using this framework, we can easily integrate a new portlet into the portal. We have been building five portlets for the grid portal, including: (i) Information Portlet shows the newest information of weather forecasting in form of images; (ii) Search Portlet provides an interface for querying meteorological data of the grid; (iii) GrADS Portlet visualizes meteorological datasets; (iv) User management Portlet manages user accounts; and (v) System management Portlet is used to setup and configure operational parameters of the data grid.

Figure 4 describes the Information Portlet. These three-day-forecasting images displayed in this portlet are automatically generated based on metadata extracted from daily forecast datasets. The extracted metadata includes the highest temperature, the lowest temperature, relative humidity, wind direction, rainfall, and cloud thickness. Based on this information, the Portlet will create appropriate images.

Figure 5 illustrates images displayed by the GrADS portlet. Whenever receiving users' requests regarding a particular dataset, GrADS portlet will produce corresponding images of wind, temperature, humidity, and precipitation.

1 DISCUSSION

So far, we have introduced the architecture and discussed implementation of our meteorological data grid system. There are many grid systems currently working in the meteorology major.

Earth System Grid (ESG) [5] provides meteorologists with access to large datasets generated by computational models of the Earth's climate. Currently, the total amount of data in ESG storage devices is approximately several hundred Petabytes. The ESG infrastructure is a system made up of physical devices and software services, including: archival and disk storage systems; storage resource managers and GridFTP servers, Metadata and Replica Services, and a Web Portal that is the user interface to the system.

The Korean Meteo-Data Grid, a project of K*Grid [6], integrates and systemizes widespread meteorological data servers. Although sharing the same idea of building a meteorology portal much like the Korean Meteo-Data Grid in K*Grid project, we follow different approaches. In K*Grid, LAS (Live Access Server) is used to integrate data servers. K*Grid also uses an agent-based searching mechanism to query data through multiple data servers. In our system, we exploit the data grid services of GT4 for setting up a virtual database of meteorology.

In the future, to fulfill the development of our system, several activities have to be completed. We will deploy other NWP models such as HRM and WRF as well as some Climate models to work with Meteo-Data Grid. In order to obtain accurate forecasting, the integration of various datasets produced by different models needs to be considered. Moreover, the next generation of data grid that provides seamless, reliable, secure and inexpensive access to distributed resources will be Semantic Grid – the combination of Data Grid and Semantic Web [8]. Semantic Web technologies should be exploited to facilitate geographically distributed meteorological scientists to resolve complex scientific problems corporately.

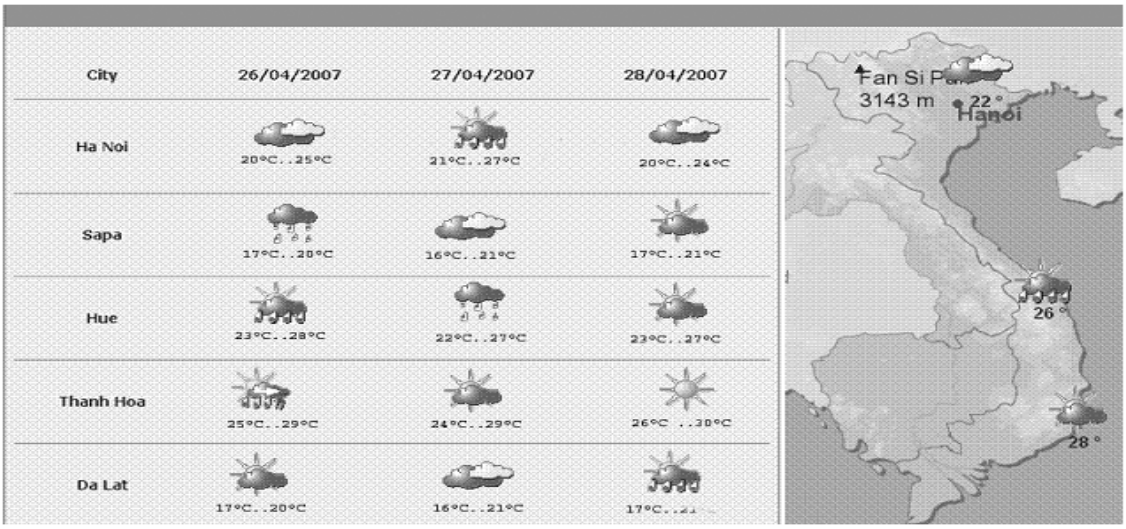


Figure4.Information Portlet

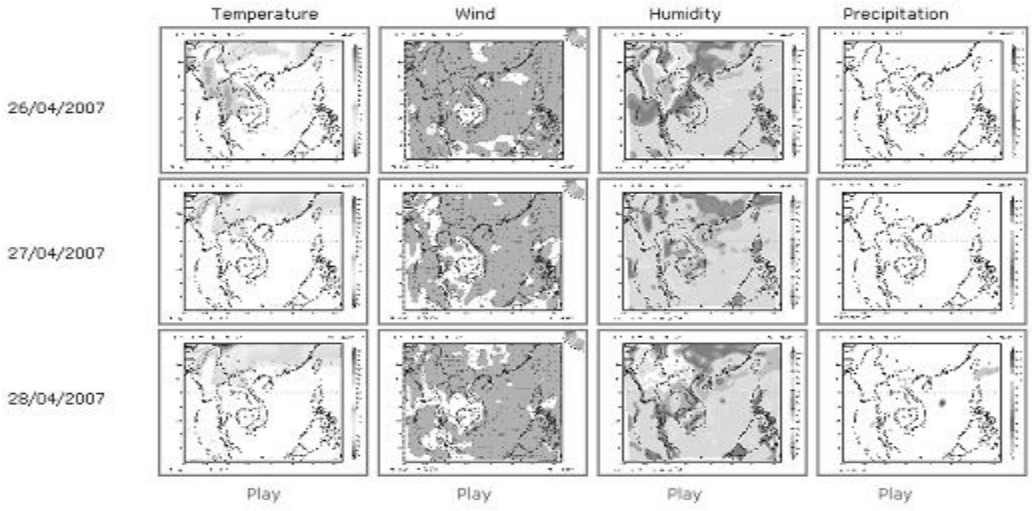


Figure 3.GrADS Portlet

REFERENCES

[1] Ian Foster, Carl Kesselman, and Steven Tuecke, "The Autonomy of the Grid, Enabling Scalable Virtual Organizations," In t. J. Supercomput. Appl. 2002.

[2] Ian Foster, Carl Kesselman, "The Grid: Blueprint for a New Computing Infrastructure". Morgan Kaufmann Publishers, 1998.

[3] Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, and Steven Tuecke, "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets", Journal of Network and Computer Applications, 2001.

[4] W. Allcock, A. Chervenak, I. Foster, L. Pearlman, V. Welch, M. Wilde, "Globus Toolkit Support for Distributed Data-Intensive Science" . Proceedings of Computing in High Energy Physics (CHEP '01), 2001.

[5] The Earth System Grid Project, http://www.earthsystemgrid.org/

[6]K*Grid,http://doi.ieeecomputersociety.org/10.1109/WETICE.2006.19

[7] MM5 Community Model Homepage, http://www.mmm.ucar.edu/mm5/

[8] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American Magazine, May 2001.