A Grid Based Infrastructure for Environmental Applications

Andrea Guerriero§, Ciriaco Pasquale§, Francesco Ragni
§Department of Electric and Electronic, Technical University of Bari
Via Re David 200, 70100 Bari (BA) Italy

Abstract: Environmental monitoring data should be transformed, by scientific analysis, into assessments of current ecosystem condition and in evolution trends in time and space. The local, regional and global scales scientific data analysis, involves information mining from several data sources, i.e. multispectral satellite remote sensing data, urban environmental monitoring stations and oceanic or coastal buoys. The huge data amount to process, the complexity of the environmental models adopted and the necessary distributed collaboration between scientists and government agencies, require specific cyberinfrastructures.

In this paper we present a Web Grid Portal prototype for environmental data processing, that aspires to provide distributed powerful computational resources able to process and to manage immense data set by a friendly graphical interface. Portal users can access to environmental data collected by different sources, process data using distributed computing resources and web applications or upload specific data processing applications.

Key-Words: - environmental monitoring, cyberinfrastructure, e-Science, Science Gateway, GRID, Globus, distributed database, web portal..

1. Introduction

Environmental monitoring systems can be designed to detect global climatic change, pollution risk trend in time and space, anthropogenic impacts on natural resources, desertification, glacier retreat, sea level rise, changes in rainfall patterns, reductions in the ozone layer, increased intensity and frequency of hurricanes and extreme weather events. The extreme wide range of different and often immense data set that should be acquired and the complexity of the mathematical model used to analyze, to process, to calibrate and to correlate them, impose scientists to realize huge computational and storage infrastructures often limited to one or few applications.

A committee of the National Science Foundation (NSF) coined term "cyberinfrastructure" to describe a new research environments in which advanced computational, collaborative, data acquisition and management services are available to researchers through highly distributed network [1]. This distributed environment should supply, to researchers and institutions, hardware and services to develop new applications, to interoperate across institutions and disciplines, to insure that data are preserved and easily available and empowers collaboration over disciplines.

Several Grid based multi discipline cybeinfrastuctures projects have been developed all over the world, i.e. Teragrid [2] and UK e-Science programme [3]. These projects include Science Gateways, web portals that provide a common interface to high performance computing (hpc) users.

In this work we present a web portal prototype with users in front and Grid services, for studying urban and coastal environment, in back.

In order to study these environments several phenomena should be analyzed and correlated such as surface temperature distribution, anthropogenic CO and CO_2 emission, land use changes and nutrient pollution. In our prototype we have considered:

- MODIS data (Moderate Resolution Imaging Spectroradiometer) that is a key instrument aboard EOS satellites that observe the entire Earth's surface, acquiring data in 36 spectral bands [4];
- data acquired from environmental urban stations performing continuous automatic

measures both chemical substances (SO_2, NO, NO_2, CH_4, NMHC, THC, PM10, Benzene, Toluene, Xilene, IPA, O_3, RAD, PAN, CO, Co_2) and meteorological quantities (temperature, velocity and direction of the wind and pressure). These stations are placed in various typologies of urban areas (parks, residential areas, sub-urban areas), oriented to long-term evolution of pollutant biological effects;

- Sidimar information, relative to SST (sea surface temperature), wind direction and velocity, pH, salinity, oxygen, chlorophyll, nitrogen;
- boiler dataset collected according to Italian prescription laws [5].

The proposed web portal allows, by a graphical interface, the access to environmental data collected by different sources; process data using available web application; upload specific data processing application.

A security access service, a data selection service, a management data service, data calibration services and graphical representation tools for processed data have been realized.

This paper is structured as follows: section 2 describes Data Structure to explain databases organization, section 3 outlines the main characteristics of the Grid services Infrastructure and the platform and tools utilized, section 4 shows the available Web Grid Services for environmental applications. Finally, section 5 closes the paper discussing the system performances and future improvements.

2. Data Structure

Data structure is composed by four kinds of different resources: Modis data, Sidimar data, Air quality environmental data, Boiler environmental pollution dataset.

Data acquired using Modis sensors that provide high radiometric sensitivity (12 bit) in 36 spectral bands ranging in wavelength from 0.4 µm to 14.4 µm [6] are packed in Hierarchical Data Format (HDF), a data file format designed by the national center of Supercomputing Application to assist user in the storage and manipulation of scientific data across diverse operating system and machine.

For this environmental application, we have considered the following MODIS product: MOD02, MYD02, MOD03, MYD03, MOD28, MOD35, MYD35, relative to TERRA and AQUA satellites.

HDF file is characterize by a great dimension, for example MOD02 is a file of about 327 MB and it represents information of a single acquisition for a limited area. Considered MODIS data set requires about 350 GB per year, for this reason and for improving recovery performance, data set are spitted and-or replicated in ten storage nodes.

Sidimar data, are environmental marine data acquired to oceanographic monitoring . They collect five millions of measures acquired in the last ten years; tree millions of these measures with APAT stations [7], that classify information about temperature, pH, salinity, oxygen, chlorophyll and others parameters, that characterized coastal Mediterranean sea.

We have organized the structure of this dataset in a Matlab matrix structure. Due to the compact Matlab data representation, the Sidimar database is allocated in a single storage node.

The air quality environmental data, are provided by a network of monitoring stations installed in various typologies of urban sites (parks, residential areas, high road traffic area, suburban zones), according to the national prescriptions. urban areas, For our prototype, we have considered data relative to a single city,

The measurement network is characterized at least by seven stations in a city with 250.000 people, such as Taranto. All the equipments carry out approximately 50 measures/minute (raw data), that are subsequently elaborated to generate hourly medium values (valid data). These values are considered valid if at least 75% of the raw data are right. This network is able to measure both chemical substances (SO_2, NO, NO_2, CH_4, NMHC, THC, PM_{10}, Benzene, Toluene, Xilene, IPA, O_3, CO) and other meteorological quantities. So, these stations record sets of data (one for each station; everyone stores sequences of measured values for every considered pollutant) at regular time, typically every hour [8].

The structure of this dataset are organized in a Matlab matrix structure.

Due to the compact Matlab data representation, the database is allocated in a single storage node. Urban environmental information are integrated by boilers emissions distribution in Taranto city area . This data are not acquired by a data acquisition system, but they are a collection of parameters: flue gases temperature, air temperature, CO and CO_2, and efficiency, that according with UNI 10389, boilers users must communicate to its own Italian Provincia every two year, in order to respect the Italian law for energy conversion.

Collected data represent information about 70000 boilers distributed on Taranto city area. In our experimentation they are used to integrate and validate local information.

3. Grid Infrastructure

Grid computing has emerged as a global platform to support organizations for coordinated sharing of distributed data, applications, and processes. Globus Toolkit [9] is one of the most widely used middleware for building Grid system and applications. This toolkit is a collection of open source software services and libraries that enable distribute infrastructures and applications, providing:

· Information services, by Monitoring and Discovery Services (MDS4): identifying available resources or services;

· Execution management services, the Grid Resource Allocation and Management (GRAM) supports: initiation, monitoring, management, scheduling, and/or coordination of remote computations;

· Data management services: GridFTP server, Reliable File Transfer service RFT, Replica Location Service RLS are services supporting high performance file transfer, management of replicated file and framework for accessing and integrating data resources, including relational and XML databases and semistructured files; a powerful security infrastructure for authentication and authorization. For our application we use an open source relational database, MySQL, as data repository. To improve grid infrastructure, shown in figure 1, a system performances data are distributed among twelve nodes, therefore the RLS is used to find in which node data are located, according with [10].

The prototype of proposed Grid portal has been realized using Xampp: a free, cross-platform standalone server, consisting mainly of the Apache HTTP Server, MySQL database, and interpreters for scripts written in the PHP and Javascript (Ajax) programming languages.

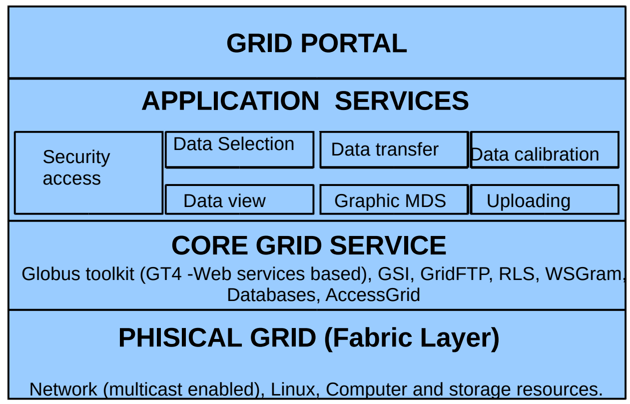

Figure 1 A concept of grid infrastucture

4.Web architecture and services

High parallel computing (hpc) solutions are usually complex and in continuous evolution [11-14], for these reasons their users should learn new Command Line Interfaces (CLI) with several options for each command, dedicated languages, file transfer protocols, shell scripts and security management.

These constrains may discourage many users that cannot focus on their applications. A generic user would like to use the computing infrastructure as an extension of own desktop, in the same trivial way he uses Web, without knowing grid infrastructure.

For this reason one of the main research activity regards frameworks development for a user friendly management. These should allow developers to easily create and maintain new applications and make hardware and software resources available to users in a simple and intuitive way.

We have employed a Grid Portal that represents a solution of the above problems, providing: access control and user management, job submission and management, file transfer, resources managements.

The graphical interface provides powerful services such as MyProxy and MDS interfaces.
The employed service are:

- security access with X.509 certification protocol;
- data selection, that provides a list of suitable products on grid databases;
- data calibration, to obtain thematic map;
- data transfer, to download processed data on client machine;
- data view, that provides a graphical representation of elaboration results;
- uploading specific data processing applications, that allow user to execute own script or web application on the grid;
- graphic MDS, that shows grid resources and allows the nodes selection to use in the processing.

4.1 Security access

Security services are fundamental to assure: authenticity of the user, correct use of GRID resources and the integrity and privacy of data exchanged between user and GRID Services.
All portal users must have a portal account and a valid PKI/GSI certificate. MyProxy server authenticates user to portal access. MyProxy provides a solution for delegating credentials to Grid portals to allow the portal to authenticate to Grid services on the user's behalf. To use MyProxy with a Grid portal, users first store a Grid credential on a MyProxy server. User credentials are stored on the MyProxy server, in this way user can "login" to the Grid portal with their MyProxy username and password.
Three main levels of user are managed to access to Grid Portal: root (portal, customers and web services administrator), user (authorized generic customer that can submit own jobs or demo applications to the Grid infrastructure), guest (generic customer that can only execute demo applications).

4.2 Data Selection

Aim of this web service is to supply a fast and user friendly interface in order to find all data for own elaboration. User can select an area using Google Maps API [15] to define a region of interest.

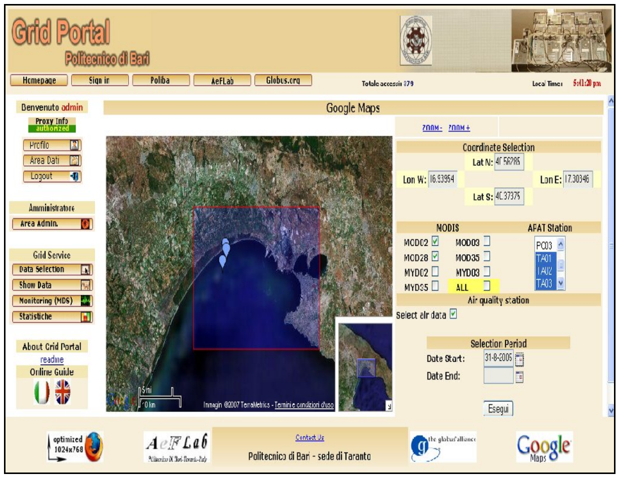

Figure 2 Selection area of interest

The map data using Google Maps is sourced largely from NAVTEQ and TeleAtlas. Similarly, the imagery found in "Satellite" mode is sourced predominantly from DigitalGlobe and MDA Federal. The extracted coordinate are the first parameter of RLS query. Moreover user can select a time period of interest and data source products(eg: MOD28 in date 01/01/2005). All this parameters are used to get data from different grid storage with RLS. The Replica Location Service, used to associate Logical File Name (LFN) with Physical File Name (PFN), keeps track of where replicas exist on physical storage systems. This research is based on Local Replica Catalog (LRC) and Replica Location Index (RLI).
The LRC, installed in each storage node, associates the LFN with physical locations on the node. LRC periodically sends information about logical name mappings to one or more RLIs. RLI collects information about the logical name mappings stored in the LRCs and answers queries about those mappings. An example of user friendly selection is shown in figure 2.

4.3 Data Calibration

Services available at moment allow graphical queries data such as MODIS database. Considering for example sea surface temperature, standard MODIS products must be calibrated. We have employed a service to produce a map along coastal area combining MOD28 products with Sidimar station. An examples of data processing allowed by the proposed system and a conventional web browser is reported in the following figure 3.

Data calibration is necessary for Air quality monitoring stations too, because raw data are affected by lacking or incorrect data. We used techniques shown in [16].

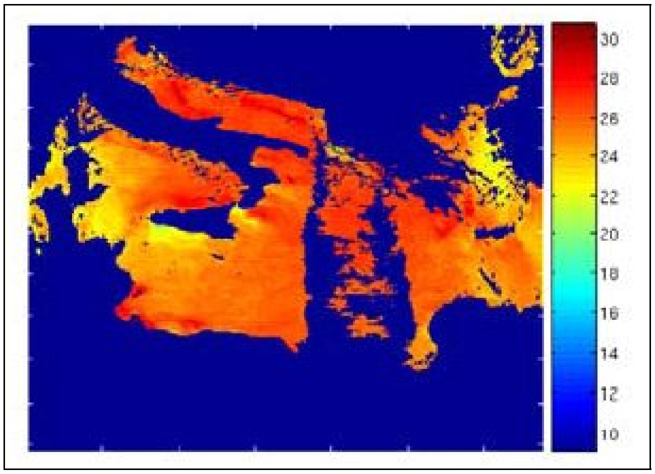

Figure 3 Calibrated Sea Surface Temperature map

4.4 Data Transfer

Grid Portal supplies to customers a temporary memory area in order to store the elaboration results. When Data Selection service or Data Calibration Service produce results, all processed data are temporary stored in this area. User can download only final results of elaborations, text files, Matlab structures or images.

4.5 Data View

Results of different data processing, obtained by available services, applied in specific geographical areas, can be shown using a layering system. Different elaboration results can be shown overlapping a transparent color maps upon selcected Google maps, or showing result using marker in which geographical information are described.

An example of an urban temperature distribution map layer is shown in figure 4.

4.6 Uploading

Uploading is a service to allow researchers and institutions to develop and test new applications. The portal permits the uploading of user scripts or web service in assigned local area.

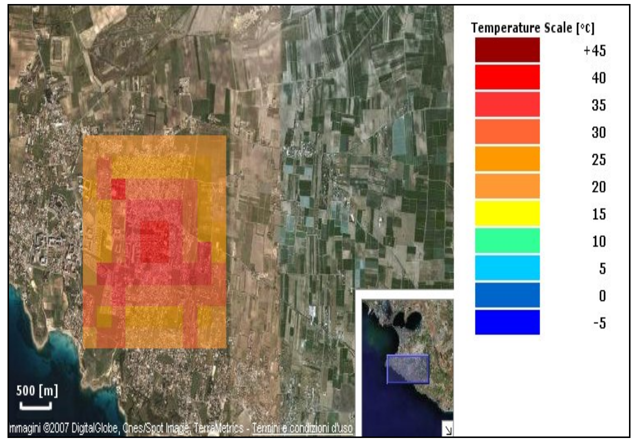

Figure 4 Example of Temperature Map

Grid portal adds the interface modules with security database in order to qualify the service to be executed. The user guide, provides information about the data format to use, This script life depends on the time period chosen in login session; after that time all temporary data, uploading script and web service will destroy.

4.7 Graphic MDS

User can shows data structure and visualize an interface that shows MDS services (Monitoring and discovery systems), for example as shown in figure 5. The goal of Web Service is to allow user to know available Globus resources. This service shows the list of active machines and their features. This information is shown in real time with fast socket connection to grid service. In this way, user can know which nodes are active, free, and where to execute own job by a simple click on node.

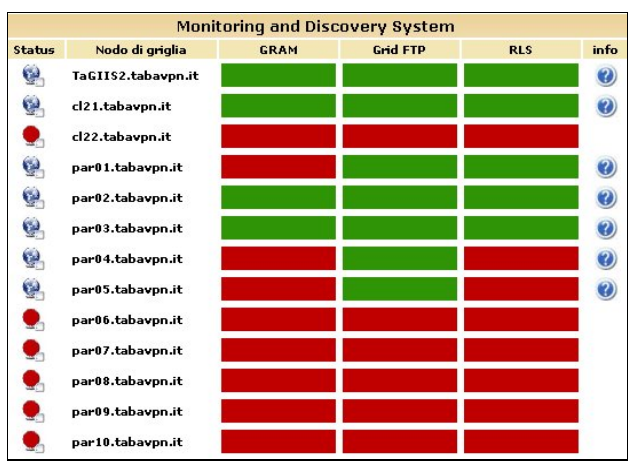

Figure 5 Graphical interface of MDS.

5 Conclusion and future works

This paper describes a prototype of Web Grid Portal for urban and coastal environmental area monitoring. A graphical interface allows access and management of different environmental data, data processing by basic services, upload of specific data processing applications and download or graphical displaying of processed data.

The system, based on Grid services and conventional web browser technology, integrates heterogeneous computing systems, data from different sources and some processing and display tools, to provide a powerful interface.

We are working to improve the web portal realizing a grid GIS for environmental monitoring.

Acknowledgment

This work was partly supported by European Regional Development Found (ERDF), “New.Ton” Project (project code A.1.180 - Interreg IIIB Archimed programme).

References

[1] www.nsf.gov
[2] www.teragrid.org
[3] www.rcuk.ac.uk/escience
[4] Modis design concept: http://daac.gsfc.nasa.gov/MODIS/products.shtml
[5] Italian laws: L. 10/1991, Dlgs.vo 192/2005]
[6] D. T. Llewellyn-Jones, P. J, Minnett, R. W. Saunders, and M.A. Zavody, “Satellite multichannel infrared measurements of sea surface temperature of the N. E. Atlantic Ocean using AVHRR/2”, Quarterly Journal of the Royal Meteorological Society, 110, 613-631.
[7] Programma di Monitoraggio dell'Ambiente Marino costiero 2001 – 2003, under the grant of Ministero dell’Ambiente e della Difesa del Territorio http://www.sidimar.ipzs.it/
[8] A. Amato, G. Andria, T. Delvecchio, V.Di Lecce, A. Guerriero, A.M.L.Lanzolla, C. Pasquale, V.Piuri, “Analysis and metrological characterization of the state of atmospheric pollution by means of mathematical models and agents.”, CIMSA 2003 – IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications Lugano – Switzerland , 29-31 July 2003 , 140-145.
[9] http://www.globus.org/toolkit/docs/4.0/
[10] A. Guerriero, R. Matarrese, A. Morea, C. Pasquale, F. Ragni, K. Tijani, "Grid Services for SST Measures”, to be published in VECIMS 2007 – IEEE International Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems Ostuni - Italy, 25-27 June 2007.
[11] Y. Wang, Y. Xue, S. Zhong, “Grid Service applied to remote Sensing processing”, Geoscience and Remote Sensing Symposium, 2005. IGARSS 2005 IEEE July 2005, Volume: 5, 3488- 3489.
[12] I. Foster, H. Kishimoto, A. Savva, D. Berry, A. Djaoui, A. Grimshaw, B. Horn, F. Maciel, F. Siebenlist, R. Subramaniam, J. Treadwell, J. Von Reich, “The Open Grid Services Architecture, Version 1.0.”, Informational Document, Global Grid Forum (GGF), January 29, 2005.
[13] Open Grid Services Architecture–Data Access and Integration Project(OGSA-DAI),http://www.ogsadai.org.uk/
[14] I. Foster, C. Kesselman, S. Tuecke, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations”, International J. Supercomputer Applications, 15(3), 2001.
[15] http://www.google.com/apis/maps/
[16] A. Guerriero, R. Matarrese, A. Morea, C. Pasquale, F. Ragni, K. Tijani, "A Grid Portal to improve SST Maps”, International Workshop on Advances in Sensors and Interfaces Bari - Italy, 26-27 June 2007.