The Development of Geospatially-enabled Grid Technology for Earth Science Applications

Liping Di
Laboratory for Advanced Information Technology and Standards (LAITS)
George Mason University
9801 Greenbelt Road, Suite 316-317
Lanham, MD 20706
ldi@gmu.edu

Abstract – **This paper discusses a research project that integrates OGC and Grid technologies for making NASA EOSDIS data easily accessible to Earth science modeling and applications communities. OGC web service technology is developed for providing interoperable access and services of geospatial data. The built-in OGC geospatial services include subsetting, resampling, georectification, reprojection, reformatting, and visualization. Grid technology is developed for sharing data, storage, and computational powers of high-end computing within a virtual organization. The integration of the two technologies makes Grid technology geospatially enabled and OGC standard compliant and makes OGC technology Grid enabled. The integration allows researchers to focus on science and not issues with data receipt, format, and manipulation.**

I. Introduction

Geospatial data are those describing natural and social phenomena or events on Earth. Because their importance to socio-economic activities, huge volumes of geospatial data have been collected by both public and private sectors. Due to their multi-disciplinary nature, geospatial data are very diverse. The data products may differ in spatial/ temporal extent and resolution, origination, format, name convention, and map projection. In order for geospatial data to be useful, the data have to be processed to extract information and knowledge based on requirements of individual applications. A typical Earth science application requires access to and manipulation of data and information from multiple sources provided by multiple data and information systems. In order to facilitate the interoperability, the geospatial community has developed a set of technologies, standards, and interface protocols. The particular important ones are those developed by the Open GIS Consortium (OGC) for interoperability of geospatial data, information, and systems over the web. The OGC web service technology and standards are widely accepted and used by the geospatial community for interoperability among geospatial systems. However, they do not provide a mechanism for securely sharing the distributed computational resources. Meanwhile, because of the large volumes of geospatial data and geographically distributed stakeholders, geospatial data and associated computational resources are naturally distributed. Grid technology, because of its security and distributed resource sharing capabilities, is the ideal technology for filling the technology gap. Therefore, the integration of Grid technology with OGC Web Service (OWS) technology will greatly benefit the geospatial community. However, Grid technology currently is geospatially unable because it has no aware of the characteristics of geospatial data. The paper discusses a project that extends Globus Toolkit to make Grid geospatially enabled through the integration with OGC technology. The extension provides OGC WCS, WMS, and WRS interfaces so that any OGC-compliant clients can access to Grid-managed geospatial data. It also enables the virtual geospatial data products and allows OGC clients to access the virtual products through WCS interfaces without knowing the data products are generated on the fly. The integration allows researchers to focus on science and not on issues with data receipt, format, and manipulation.

This project is funded by NASA ESTO AIST program

II. The OGC Web Data Services Technology

OGC is an international organization promoting the interoperability and sharing of geospatial resources and services in the distributed environment through the development of consensus-based implementation specifications [1]. OGC specifications are widely used by geospatial communities for sharing data and resources and are becoming ISO standards. The recently developed OGC web-services specifications allow seamless access to geospatial data in a distributed environment, regardless the format, projection, resolution, and the archival location [2]. The fundamental ones include Web Coverage Services (WCS) [3], Web Feature Services (WFS) [4], Web Map Services (WMS) [5], and Web Registries Services (WRS) [6]. OGC WCS defines web interfaces for accessing on-line multi-dimensional, multi-temporal geospatial data in an interoperable way. Coverage data include gridded geospatial data and remote sensing images. Both HDF-EOS Swath and Grid data, representing more than 95% of EOS data holdings, are coverage data. OGC WFS defines web interfaces for accessing feature-based geospatial data. Feature data are traditionally called vector data. HDF-EOS Point data are feature data. WCS and WFS together cover

all geospatial data. They form the foundation for OGC web-based interoperable data access. OGC WMS defines interfaces for assembling maps from multi-sources over Web. A WMS server normally converts data to a visualized form (map) based on requirements from the client. OGC WRS defines web interfaces for finding data or services from the registries.

The OGC technology allows users to specify the requirements for the data they want. An OGC compliant server has to preprocess the data on-demand based on users' requirements and then returns the data back to users in the form specified by users. At the end, users get the data that exactly match their requirements in both the contents and the structure (e.g., format, projection, spatial and temporal coverage, etc) [7]. This will significantly reduce the time needed for users to acquire and preprocess the data before they can be used in models or analysis packages.

The Laboratory for Advanced Information Technology and Standards (LAITS) of George Mason University has developed NWGISS to test OGC interfaces in NASA's data environment. Funded by NASA ESTO, NASA ESDISP, and OGC, NWGISS provides interoperable, personalized, on-demand data access and services (IPODAS) to EOSDIS data with built-in georectification, reprojection, subsetting, resampling, reformatting, and visualization functions [7, 8]. Currently, NWGISS consists of five components: a Map Server, a Coverage Server, a Catalog Server, a multi-protocol geoinformation client (MPGC), and a Toolbox. The map server serves HDF-EOS data as maps to any OGC-compliant map clients. The coverage server allows clients to access multi-dimensional data at user specified geographic coverage, parameters, projection, and formats. The catalog server provides the catalog search capabilities to catalog clients. MPGC enables users to search WRS server and to access data served by OGC web coverage, map, and feature servers. It also provides a set of data manipulation, processing, and analysis functions at user's desktop. The tool component consists of two-way translators between HDF-EOS and major GIS formats and the CreateCapabilities program. Figure 1 shows the NWGISS architecture.

III. The Grid Technology

The Grid is a rapid developing technology, originally motivated and supported from sciences and engineering requiring high-end computing, for sharing geographically distributed high-end computing resources [9, 10, 11]. The vision of the Grid is to enable resource sharing and coordinated problem solving in dynamic, multi-institutional virtual organizations [10]. It provides on-demand, ubiquitous access to computing, data, and services and constructs new capabilities dynamically and transparently from distributed services. With grid technology, new applications, such as distributed collaboration, distributed data access and analysis, distributed computing, are enabled by the coordinated use of geographically distributed resources.

Currently, dozens of major Grid projects around the world in scientific and technical computing for research and education have been either deployed in the operational use or demonstrated technically. Considerable consensuses on key concepts and technologies have been reached. The key for the GRID success is the open source middleware called Globus Toolkit [12, 13]. It has become a de facto standard for all major Grid implementation. Although far from complete or perfect, the Grid technology is out there, evolving rapidly, and has large tool/user base. The Global Grid Forum is a significant force that coordinates the development of the technology in the world [14].

IV. The Integration of Grid and OGC Technologies

The main work of this project is to integrate Grid and OGC technologies. Based on the analysis of the two technologies and the EOSDIS data environment, it is decided that the integration takes place between the backend of the NWGISS OGC servers and front-end of data Grid services. The key is to make Grid-managed geospatial data accessible through any OGC compliant clients by making NWGISS OGC servers working with Globus.

Based on the original project plan, the integration is conducted in three phases. The *first phase* is the testbed and initial integration, which include the setup of the development environment, preliminary design of the integration, and implementation of WCS access to Grid-managed data. The *second phase* is the data naming and location transparency, which include investigating the use of data Grid and Replica Services (metadata catalogues, replication location management, reliable file transfer services, and network caches) to provide naming and location independence for data used by NWGISS and revising NWGISS to invoke such Grid services. The approach to investigating the data Grid and Replica Services will be to configure a data Grid testbed. This will be followed by the integration of NWGISS data catalogs into a data Grid catalog and the investigation of naming approaches, followed by interfacing NWGISS with data generators and data Grid Replica Location service.

The *third phase* is the virtual dataset research and development. Virtual datasets are those the Grid knows how to produce on-demand, but not produced (materialized) yet. The concept of virtual datasets has been implemented in the Grid Physics Network (GriPhyN) project [15], and is being tested in Earth sciences [16] . This project will investigate the feasibility of using the virtual data services (materialized data catalog, virtual data catalog, abstract planner, concrete planner) to provide the on-the-fly data transformation services needed by NWGISS. The approach to investigating the virtual data Grid services is to conduct research necessary to represent the virtual Earth science data products in terms of the

virtual data catalog so that they may be re-materialized on-the-fly. This involves research and development on following items:

1. How to represent a credible variety of Earth Science virtual data products in a Virtual Data Generator (VDG) compatible prescription form and to modify the NWGISS MPGC to edit and submit the data generation prescriptions to the data Grid.
2. Developing the Abstract and Concrete planners that can convert the ES virtual data product prescriptions into workflow representations suitable for a Grid workflow engine.
3. Using the Data Grid services to manage the resulting re-materialized data so that the NWGISS servers can have rapid and transparent access to all of the data needed by users.

The heart of the work for phase III is to design and implement the XML metadata descriptions for all of the key structures and transformation prescriptions for generating ES datasets, that we believe is more complicated than generating high-energy physics products. ISO 19115 metadata standard [17], and FGDC remote sensing metadata standard [18] are the starting point for such a work. Figure 2. shows the architecture after the three phase integration.

V. The Test and Demonstration Environment

The integration has been taken place at development environment at GMU, NASA Ames, and LLNL, and tested and demonstrated at EOSDIS Data Pools and DOE's ESG testbed. The Data Pools is an EOSDIS project for providing large volumes of EOS data on-line for users to directly and rapidly access. Each data pool provides discipline-specific EOSDIS data. Currently there are four operational data pools, located at GSFC, Langley, EDC, and NSIDC respectively. At present, those data centers determine which kind of data to be on-line but NASA intends eventually to put most of EOSDIS data on the pools for users to directly access. This means that Data Pools will eventually hold multi-petabytes of EOS data. The intended users of data pools include climate and environmental modeling, and application communities. Most of data in the Data Pools are in HDF-EOS, the standard format for EOSDIS. Because of huge volumes of data, distributed high-performing computers and storages, high-speed networks, and the fact that both Grid and OGC technologies are used, Data Pools are the ideal place for testing and demonstrating the integration. This project is accessing and manipulating the data pool data through OGC WCS interface currently implemented in data pools to show how well the integration serves diverse user groups ranging from modelers to value-added service providers.

In addition to Data Pools, we will also use DOE's Earth System Grid (ESG) as testbed. Funded by the Scientific Discovery through Advanced Computing (SciDAC), ESG seeks a new paradigm in the climate change modeling community evolving from centralized data sharing to distributed data-sharing. ESG enables distributed teams of researchers to effectively and rapidly acquire knowledge and understanding of massive amounts of climate data holdings. However, it has no OGC interfaces currently. The ESG testbed will test and demonstrate the access to satellite remote sensing data by modeling communities for model output validation and integration.

VI. Current Implementation Status

This project was started in May 2003. In the past one year, the project has finished the tasks in phase I. The project designed the architecture of the OGC Data Grid, built the testbed/development environment at GMU, NASA Ames Research Center, and Lawrence Livermore National Lab, designed and implemented a Grid Services-based reliable data transfer service, modified NWGISS WCS server to use the service for accessing Grid-managed data, used the Grid Security Infrastructure design as the basis of restructuring NWGISS to be a secure Grid framework, designed a metadata catalogue for a data replica catalogue, produced a catalogue architecture scalable to Data Pools, and implemented an OGC WRS interface to the Grid catalog managers. The results of the first phase integration were demonstrated at the SC 2003 Conference.

VII. Conclusions and Future Work

Both OGC and Grid technologies are very promising for applications in Earth sciences for providing interoperable sharing of geospatial data, information, knowledge, and computational resources. The two technologies match each other very well in the Earth observation (EO) environment. The EO community will be benefited by the integration of the two technologies.

In the coming years, we will mainly work on the second and third phases of the planned research. The major research efforts will be concentrated on the implementation of virtual geospatial products. This research work is a part of the large efforts to make the geospatial Grid concept discussed in [16] a reality.

References

[1] OGC (2003) The Open GIS Consortium Homepage, http://www.opengis.org.

[2] Di, L., 2004. "The Open GIS Web Service Specifications for Interoperable Access and Services of NASA EOS Data." In Qu, J. etc eds, *Earth Science Satellite Remote Sensing*. Springer-Verlag (in press).

[3] Evans JD (ed) (2003) Web Coverage Service (WCS), Version 1.0.0. OpenGIS© Implementation Specification. Open GIS Consortium Inc. http://www.opengis.org/docs/03-065r6.pdf.

[4] Vretanos PA (ed) (2002) Web Feature Service Implementation Specification, Version 1.0.0. OGC 02-058. Open GIS Consortium Inc. http://www.opengis.org/docs/02-058.pdf.

[5] de La Beaujardière J (ed) (2001) Web Map Service Implementation Specification, Version 1.1.1. OGC 01-068r2, Open GIS Consortium Inc. http://www.opengis.org/docs/01-068r2.pdf.

[6] Reich L (ed) (2001) Web Registry Server Discussion Paper, OpenGIS Project Document 01-024r1. Open GIS Consortium Inc., http://www.opengis.org/docs/01-024r1.pdf.

[7] Di, L., W. Yang, D. Deng, and Ken. McDonald, 2002. "Interoperable, Personalized, On-demand Geospatial Data Access and Services Based on OGC Web Coverage Service (OWS) Specification", *Proceeding of NASA Earth Science Technology Conference*, CDROM, Pasadena, California. 3pp.

[8] Di, L., 2004. "The NASA HDF-EOS Web GIS Software Suite (NWGISS)." In Qu, J. etc eds, *Earth Science Satellite Remote Sensing*. Springer-Verlag. (in press).

[9] Foster I., C. Kesselman, J. M. Nick and S. Tuecke, 2002. The Physiology of the Grid: An open Grid services architecture for distributed systems integration. Open Grid Service Infrastructure WG, Global Grid Forum. http://www.globus.org/research/papers/ogsa.pdf.

[10] Foster I., C. Kesselman and S. Tuecke, 2001. The Anatomy of the Grid – Enabling Scalable Virtual Organizations. Intl. J. of High Performance Computing Applications, 15(3), 200-222.

[11] Foster, I. and C. Kesselman, editors, 1999. The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann Publishers.

[12] Foster I. and C. Kesselman 1998. The Globus Project: A Status Report. *Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop*, pp. 4-18. ftp://ftp.globus.org/pub/globus/papers/globus-hcw98.pdf

[13] Globus homepage, http://www.globus.org.

[14] Global Grid Forum (GGF). http://www.ggf.org.

[15] **GriPhyN** - Grid Physics Network. http://www.griphyn.org/

[16] Di, L., 2004. "The Geospatial Grid." In S. Rana and J. Sharma eds, *Frontiers of Geographic Information Technology*, Springer-Verlag. (In press)

[17] ISO 19115: 2003- Geographic Information – Metadata. http://www.iso.org.

[18] Di, L., B. Schlesinger, etc, eds. 2002, *The FGDC Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata*, *FGDC-STD-012-2002*. U.S. Federal Geographic Data Committee. Reston, VA (144 p).

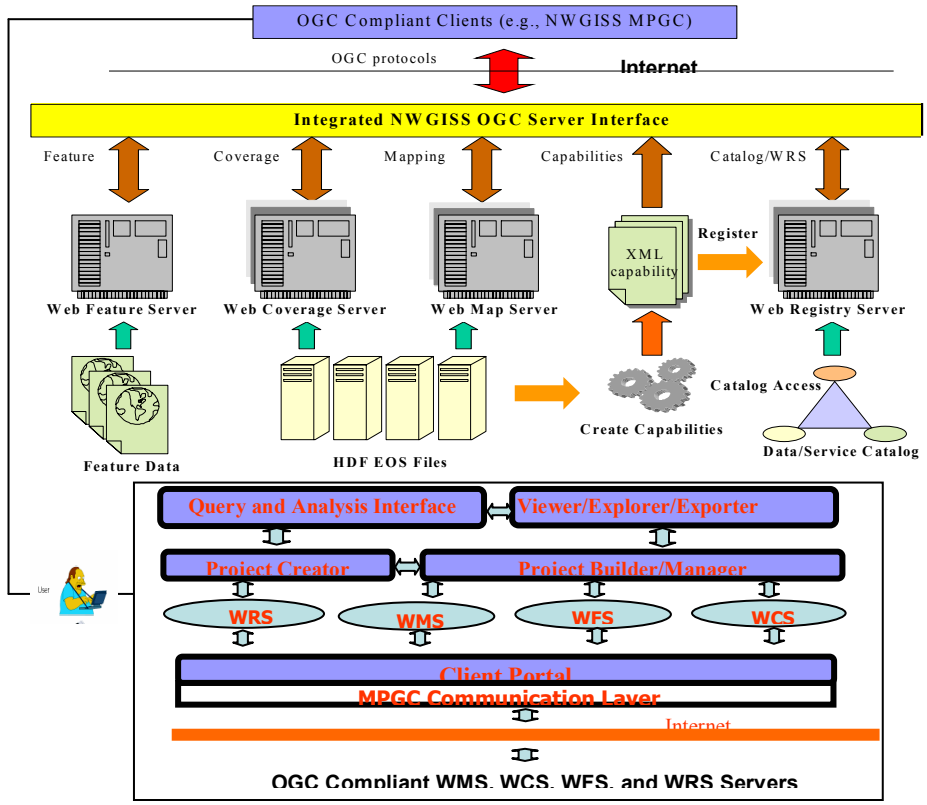

Figure 1. The NWGISS Architecture

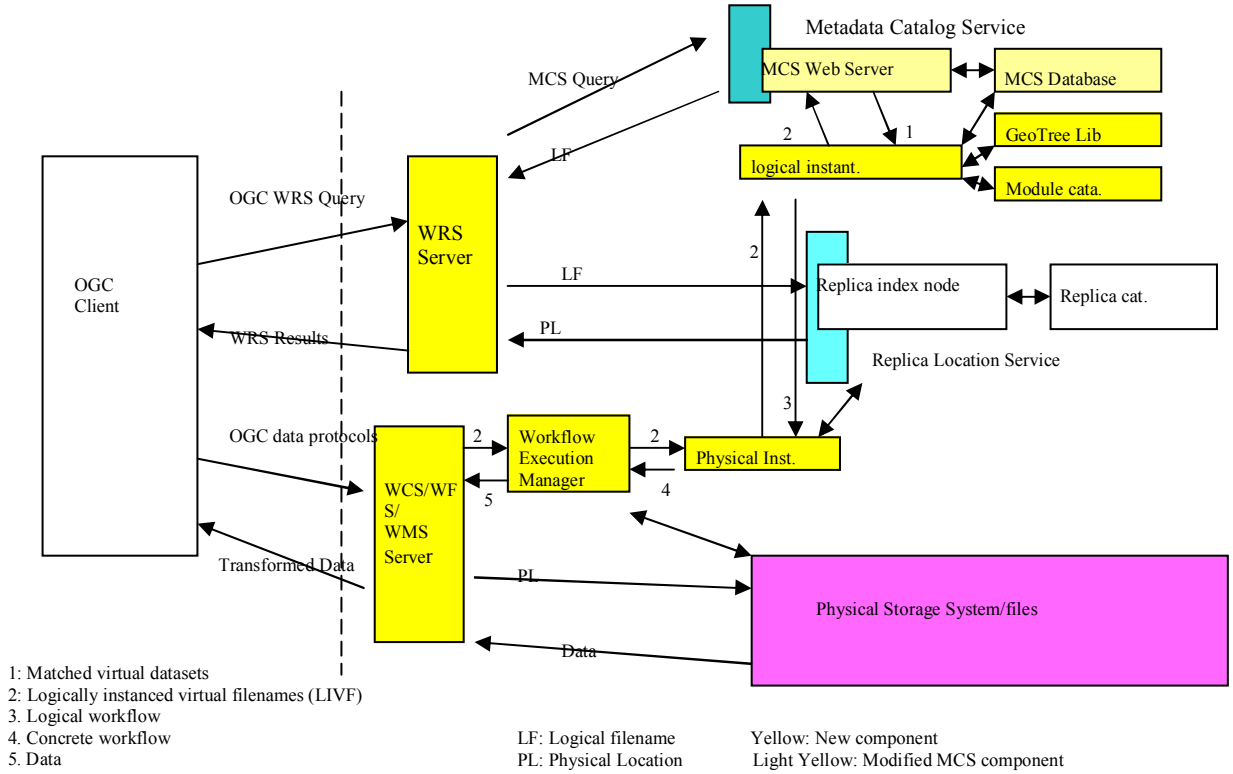

Figure 2. The Architecture of the Integration