Service oriented coupling framework for data and computational resources to support decision making over integrated data display: GIS Approach

By Ahmet Sayar

Research Committee:

- *Prof. Geoffrey C. Fox (Principal Advisor)*
- *Prof. Randall Bramley*
- *Prof. Kay Connelly*
- *Prof. Melanie Wu*

Indiana University - April, 2007
Computer Science - Community Grids Lab (CGL)

1. Introduction

We investigate the issues pertaining to the traditional Geographic Information Systems (GIS) [2, 9] approaches and propose solutions to these problems based on modern Service Oriented Grids approaches. As in general science domains, GIS requires decision making and situation assessment based on integrated data display. We generally focus on the issues in terms of GIS and geographic data, but findings and recommendations are relevant for any other science domains and data types.

GIS is a system of computer software, hardware, and data used to manipulate, analyze, and graphically present a potentially wide array of information associated with geographic locations. GIS's powerful ability to integrate different kinds of information about a physical location can lead to better informed decisions about public investments in infrastructure and services—including national security, law enforcement, health care, and the environment—as well as a more effective and timely response in emergency situations. However, long-standing challenges to data sharing and integration need to be addressed before the benefits of geographic information systems can be fully realized. Our focus regarding data integration is different from the Database or digital library communities. We deal with the integration at the higher level than they do and we try to utilize their approaches at the bottom level by proposing generic mediator services.

Our work is about developing a Web Services architecture that provides coupling of scientific geophysical applications with archival data through the innovative interactive smart decision making tools. This work can be detailed in a couple of sub-research areas such as: accessing and querying heterogeneous data provided by heterogeneous storages with unified query structures, developing GIS data services, considering performance issues of transferring, parsing and rendering of large geographic data, and composition of GIS services.

In the light of the explanations above we categorize our work as below:

1. Coupling Geo-Science computational Grid with data Grid
    - Integrating Web Map Services with Geo-Science Grid [11, 12]
    - Enabling decision making through integrated data display (3-layered display structure)
    - Creating view-level integration structure. View is abstracted as layers in GIS domain.
    - Creating generic plotting Web Services (Sci-Plot) in order to couple Geo-Science Grid outputs with its inputs (from data grid) at the view level.
2. Handling the heterogeneity in data Grid as a component of the framework
    - Standards (OGC and ISO/TC211), Web Services and mediator services.
    - Different data types
    - Different storage types
3. Interactive and smart decision making tools
    - Coupling interface for browser based remote access
    - Data/information display
    - Interactive querying and mining the data

    - Visualization and analysis of the data and Science Grid simulation outputs
    - Movies and animation tools for the time-series data
4. Performance
    - Accessing remote large data sets provided by geographically distributed data vendors.
    - Transferring, integrating, processing and interpreting data.
    - Proposing: High-performance streaming data services through messaging middleware.
    - Proposing: Advanced pre-fetching, caching and load balancing techniques.

The importance of providing access to “*computational resources”* has been central in many research efforts in Grid community. Another such important issue is distributed access to data stored in various types of “*data resources”*. GIS is especially affected by the developments in both of these areas since these systems are traditionally data-centric; they require access to data from many different sources for creating layers, and tend to use various types of data processing tools for analysis or visualization of the geographic data.

The proposed coupling framework (see Chapter 4) is targeted to a community with a broad and demanding range of functional requirements: the situation understanding and information management systems community. A major challenge in designing and building such a system is not only to develop basic system capabilities but to provide a framework such that the best available tools can be integrated into the system with minimal effort and that each of these components can communicate with each other to create new applications. These components must be able to bi-directionally interact with document management components by sending and receiving information (capability metadata transactions through *getCapabilities* Web Services of the components). With this in mind, we designed the framework as a component based system that will be able to support continuous increase of functionality and probability as new and more sophisticated tools become available.

Distributed data access in GIS is traditionally regarded as dealing with distributed data archives, databases or files. The data storages and data itself can be heterogeneous. As an example of storage or service heterogeneity, we can give three major types of GIS servers used by different Indiana State counties. (1) ESRI [50] ArcIMS and ArcMap Servers are used for Marion, Vanderburgh, Hancock, Kosciusko, Huntington and Tippecanoe counties, (2) Autodesk MapGuide [51] is used for Hamilton, Hendricks, Monroe and Wayne counties and (3) WTH Mapserver Web Mapping Application [52] is for Fulton, Cass, Daviess and City of Huntingburg counties based on several Open Source projects. When a client needs to access these servers he needs separate code or application to access these data. Integrating them is almost impossible without advanced integration techniques.

We also observe the same interoperability problem at the data level. There are numerous ways of describing geospatial data in various formats such as ESRI shape files, ASCII files, XML files, Geography Markup Language (GML) files etc. Depending on the user’s choice of software, applications that digest geospatial data require input in different formats. Users spend significant amount of time converting data from one format to other to make it available for their purposes.

Problems with the data integration mostly come from:

1. Adoption of universal standards: Over the years organizations have produced geospatial data in proprietary formats and developed services by adhering to differing methodologies.
2. Distributed nature of geospatial data: Because the data sources are owned and operated by individual groups or organizations, geospatial data is in vastly distributed repositories.
3. Service interoperability: Computational resources used to analyze geospatial data are also distributed and require the ability to be integrated when necessary.

In order to solve data and service heterogeneities for the GIS computation and data services we use OGC standards in general and mediator services at the lower level of data integration hierarchy. Mediators provide an interface of the local data sources and enable interoperable service interface (WFS interface) to query and access the data. They have mapping rules that express the correspondence between the global schema (GML) and the data source ones. They change depending on the data type kept at the resource. The problem of answering queries is another point of the mediation integration – a user poses a query in terms of a mediated schema (such as getFeature to WFS), and the data integration system needs to reformulate the query to refer to the sources. Mediator services are explained in Chapter 3.

In order to utilize and interact with the coupling framework enabling access to computational resources and data resources in an ordered and synchronized manner, we develop innovative interactive smart decision making tools (see Chapter 6). Tools enable feeding the Geo-Science applications (projects) with the data and association of applications input and output data at the layer (view)-level. The attributes of the data/information which are building these comprehensible representations should be interactively accessed and queried. Resources are organized into projects and they are compliant with some resource schemas. Client portal provides both map-based and project based user interfaces. The former are suitable for geospatial resources with location information while the later caters for all kinds of resources with or without location information. The two interfaces co-exist and are synchronized. That is, resources selected using the map-based interface will also be highlighted in the project based interface, and vice versa.

We build services for supporting scientific GIS applications (mostly *ServoGrid* projects [53, 54 and 55]) that demand high-performance and high-rate data transfers. Since Geo-Science applications require quick response times, the GIS services must enable accessing and processing these large data sets in a reasonable time period. In most cases the amount of collected data reaches to an amount in the order of gigabytes or even terabytes, handling this data becomes a challenge for most users and organizations. Also, simulation and visualization software used in conjunction require high performance computing platforms which are unreachable for common users. We have developed Grid oriented Web Map Web Services in OGC standards and Web Services principles. Grid oriented Map Services enable data integration and layer-overlaid display for the coupling framework by using innovative pre-fetching, load-balancing and caching techniques. It also provides different running modes set before run-time such as streaming or non-streaming data transfer modes. Non-streaming gives better results for applications using small amount of data (less than 20MB). In other cases for large scale applications streaming mode gives high performance results. We give the detailed performance test results in Chapter 7.

These issues are undeniably the crucial points of the numerous research and development efforts [45, 46]. Especially the problems related to the data formats and standards are being addressed by a number of groups and organizations some of which also offer solutions to the application level interoperability issues [47, 48 and 49]. We generally focus on the issues from the GIS and special data point of view, but findings and recommendations are relevant for any other science domains and data types.

The rest of the paper is organized as follows. Chapter 2 highlights the important terms and concepts in our research domain in order for the readers to better understand the discussions in the following chapters. Chapter 3 presents the innovative coupling framework for Geo-Science computational Grid with distributed and heterogeneous data sources. Coupling is done at the view level by introducing 3-layer structured display. Chapter 4 explains the challenges in the heterogeneous data and service integration in the framework and, introduces an innovative approach (mediators) to solve the problem. Chapter 5 presents proposed interactive and smart decision making tools enabling end-users to interact with the coupling framework in a synchronized manner to make decisions over the integrated data display. Section 6 discusses performance issues and proposes innovative pre-fetching, load-balancing and caching techniques for the un-evenly distributed geospatial data. Section 7 explains the key-service architecture in the framework (Grid-oriented Aggregator WMS) enabling of coupling at an acceptable performance level. Section 8 gives the preliminary performance tests. Chapter 9 lists the expected contributions and Chapter 10 presents the future work.