Thesis Proposal

View-level coupling of Information Systems and Science Applications: Applications to GIS

Ahmet Sayar
asayar@cs.indiana.edu

1. Introduction

Geographic Information System (GIS) [1,2] is a system for creating, storing, sharing, analyzing, manipulating and displaying spatial data and associated attributes. It can be used for creating or capturing geographic information from various sources in digital form or for viewing available geographically referenced data in human recognizable formats. Perhaps the simplest example of Geographic Information Systems is widely available map viewers which process layers of geospatial data to create map images. GIS are used in a wide variety of tasks such as urban planning, resource management, environmental impact assessment, emergency response planning in case of disasters, crisis management and rapid response etc. Although these seem to be relatively independent and different areas, a common feature of almost all GIS use cases is the need for the system to relate information from different sources.

In a relatively short period of time the Internet has dramatically changed how scientists, industry specialists and the public access, exchange and process information. As in most other cases the geospatial data access and dissemination methods also significantly evolved. This helped academia, governments and businesses to have easy access to substantial amount of geospatial data.

Most of the traditional distributed Information System approaches (involving GIS) are based on more traditional client-server models and lack the potential of easily linking distributed computational components and moreover, coupling of data and computation sources to solve the problems in science domains.

Like any other information systems, GIS journey from centralized mainframe systems to desktop systems and finally to distributed systems [3]. Today a modern GIS requires distributed systems support at two levels; first for accessing various geospatial databases to execute spatial queries and second for utilizing remote geographic analysis, simulation or visualization tools to process spatial data (science applications). This thesis proposes an architectural performance efficient framework for coupling these two levels of Information System's requirements.

2. Motivation

The traditional GIS systems conventionally used to access and analyze local data do not have the ability to interact with online data sources and with other remote spatial analysis applications. To be able to interact with online geospatial resources the traditional GIS programs are evolving into distributed applications, compatible with various distributed systems architectures. Besides having many advantages of distributed systems we face some problems regarding data and service formats and standard to make the services and data interoperable [10, 11].

Science applications can be summarized as set of data mining software and simulation processes and requiring access to various data sources and run several services in an ordered and synchronized way. They help Information Systems to extract much more complex information and knowledge from the data. For example, we might need to see the state based probability of an earthquake happening in United States for the duration of upcoming 5 years. In order to make this service publicly available and usable by common users, data should be represented in a comprehensible format such as images or graphs. The systems complexities should be hidden and applications should be accessed easily. Another example might be average temperature plotted over the map images produced by GIS. Plotting and average calculations etc all are scientific processes undertaken by the science applications. In order to see the average temperature or the probability of the earthquake happenings related to the earth location in a comprehensible fashion, Information systems and science applications should be coupled, accessed and managed synchronously and seamlessly.

To be more practical, we can approach to the issues from the human point of view. The users need to access and query the heterogeneous data remotely and seamlessly and, even run the science application simulations and synchronize the results with the data grids interactively. The data might be coming from different software and hardware platforms provided by different vendors located geographically separate places. The heterogeneity comes from the data storages and/or data itself. In order to answer users' request, sometimes, the heterogeneous data sets must be integrated, interpreted and, human comprehensible data representations must be created. Since the information system users don't have to be expert people like the system developers and scientists, they need to interact with the system through interactive and easy to use tools and needs to get the result in a reasonable time period. Otherwise, the system becomes useless from the user point of view.

We focus on the issues from the GIS point of views, but findings and recommendations are relevant for any other science domains and data types.

3. Research Issues

We group our research issues into two. First group of research issues are related to the well-known ***performance*** and ***interoperability*** issues of the distributed Information systems. Second group of research issues are related to the innovative view-level (layered) **coupling** of Information Systems and science applications.

Information Systems demand high-performance and high-rate data transfers and, require quick response times. In most cases the amount of collected data reaches to an amount in the order of gigabytes or even terabytes. Therefore, the GIS services must enable accessing and processing these large data sets in a reasonable time period. Handling large data becomes a challenge for most users and organizations. Specifically for the GIS, this picture will be even worse when the map animations and map movies need to be created.

Since the proposed GIS system is a distributed system and due to the limited bandwidth and network speed, sometimes it is impossible to make large scale geo-science applications feasible. Information Systems utilize and handle large size data structures and need high performance data transfer and rendering. Instead of dealing with the hardware and network issues, we work on software solutions to improve the system performance such as streaming data transfer, caching and pre-fetching.

The distributed Information systems face interoperability and heterogeneity problems. These are mostly result from adoption of the universal standards, distributed nature and heterogeneous formats of science data and, service interoperability. In order to make the system interoperable, the services and data types should be standardized. Service standardization includes defining standard service functionalities and corresponding service interfaces. After defining the service standards we need to define common request formats and expected responses

The importance of providing access to science applications and, associating their outputs with the corresponding inputs from Information Systems has been central in many research efforts. 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 and, creating more complex view-level displays.

We identify the following research questions:

- How to couple GIS Information Systems (data and computation sources) with the science applications at the view level.

- How to define standard Plotting Web Services to couple science applications output with GIS Information Systems outputs (map layers).

-How to create user oriented interactive display and querying tools to interact with the Information Systems and scientific applications in synchronized way.

- Can we implement unified data-centric Information System architecture to provide common interfaces for accessing archival geospatial data sources?

- How can we incorporate widely accepted geospatial industry standards (OGC and ISO-TC211) with Web Services?

- How to build Web Services for data and computation source for supporting scientific GIS applications demanding high-performance and high-rate data transfers.

- How to make unified access and querying of heterogeneous data provided by geographically distributed sources

- How to make parallel processing for un-evenly distributed and large size geospatial data (workload cannot be guessed before the results come)

4. Methodology

Our proposed GIS architecture is Web Service based Service Oriented Architecture (SOA) implemented in JAVA. In order to develop and evaluate our architecture, we chose Apache Axis 1.x [13] version to deploy Web Services. We exploited Apache Tomcat [14] as a servlet container. Apache Tomcat is developed in an open and participatory environment. It implements Java Server Pages (JSP) and the servlet specifications from Sun Microsystems.

For the creation of interactive browser-based system, we use AJAX (Asynchronous JAVA +XML), Dynamic HTML, JavaScript and Web Service client tools.

In order to achieve interoperable GIS system we use universally accepted OGC [15] and ISO/TC211 standards specifications for online services and data model. As online standard services we develop Web Map Services (WMS) [10] and Web Feature Services (WFS) [11] (by Ayding G.) and, convert them into Grid-enabled Web Services by extending. We also utilize standard definition of capability metadata (RDF like structures) of the services and enable inter-service communications.

As data model we use semi-structured data defined by widely accepted OGCTC-211 standards. It is called Geographic Markup Language (GML) [4]. GML is basically an XML encoding for the modeling, transport and storage of geographic information including the spatial and non-spatial properties of geographic features. Feature is an entity related to earth with a geographic location such as road, river, states etc. GML has separate schema standards for both content part and presentation part. That enables display and query of the data. Display is related to presentation part of the data schema and query is related to attributes of the data schema.

5. Architecture in Brief

We investigate the issues for coupling the Information Systems with the science applications. General concept is illustrated in Figure 1. From now on, we focus on GIS and Geo-Science application coupling. In order to achieve this we develop innovative Aggregator Web Map Services and Sci-Plotting services providing information in comprehensible data layers (Figure 2). We propose a framework enabling interactive data display and analysis and, data may be coming from Information System (triangle in Figure 3) and/or science applications.

The coupling is done at the view-level and based on two key services, Map Services and Sci-Plotting services. Since we focus on Information System architecture and general coupling issues, we take science applications as a black box and integrate to the coupling framework through a service oriented workflow system.

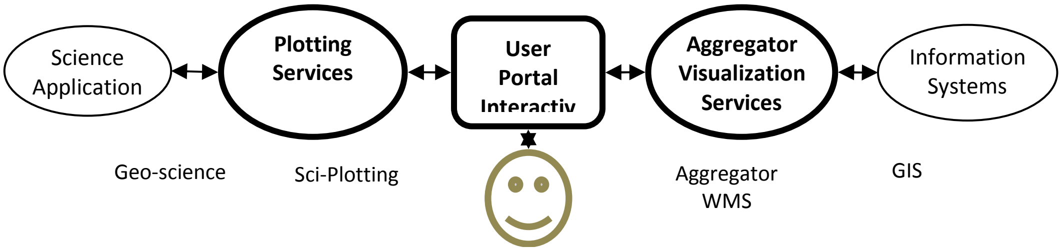

Figure 1: General coupling concept for any domain

The architecture utilizes 3-layered structure display. The bottom layer is to show the underlying maps such as satellite images. The middle layer is to display the features of the map by using vector data. For example, earthquake data or state boundaries can be illustrated in this layer. Finally, the third layer is to associate the previous two layers (coming from Map Servers) with science application's outputs.

In summary, Figure 2 shows the seen part of the ice-berg and the Figure 3 is the ice-berg. In Figure 2, the bottom two layers come from Information System and Top layer come from the geo-science applications. Bottom two layers are provided by Grid-enabled WMS and top layer is provided by Grid-enabled Sci-plotting services (see also Figure 3).

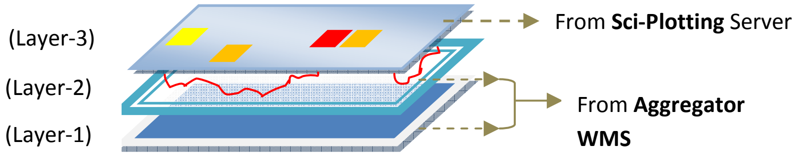

Figure 2: Output structure of the integration framework (Figure 3). Example plot shows possibilities of earthquake happenings; red is for the highest possibility regions.

We investigate the issues related to the traditional Geographic Information Systems (GIS) [2, 9] and, propose an architectural framework as solutions based on modern Service Oriented Grids approaches. Developing such an "Information System" has also some other issues due to the characteristics of distributed and heterogeneous nature of the data and computation sources.

Components of the architecture (Figure 3**)**: The proposed Information System (triangle in Figure 3) is composed of chains of WMS and WFS services. **Aggregator WMS** is actually a WMS with some extensions providing high performance mapping services by using innovative pre-fetching, load-balancing and caching techniques (they will be explained briefly in the following chapters). **WFS** (implemented by Aydin G.) access various geospatial databases to retrieve and present the data to the users in a standard (GML feature collections). WMS interact with WFS by submitting structured queries and in compliance with OGC's Filter Encoding and OGC Common Query Language. **WMS** Enables visualizing, manipulating and analyzing geospatial data through maps shown on browser based interactive GUI. Map Servers typically compose maps in layers. **User Portal** and smart map tools enable end-users to interact with the proposed system. Resources are organized into projects and they are compliant with some resource schemas. The client portal is capable of supporting capabilities metadata of WMS and WFS. The client portal provides both map-based and project based user interfaces. The two interfaces co-exist and are synchronized.

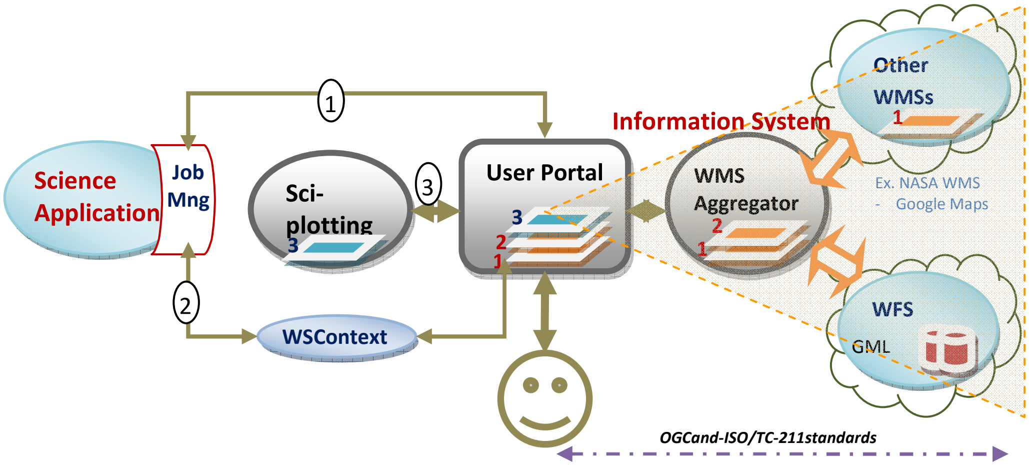

Figure 3: ***Information Grid and science application coupling through Web Map Service and interactive decision making tools.***

Sci-Plotting is the key service enabling the overlaying Geo-Science applications' output as layers in 3-layer structured displays. Its core functionality is provided by Dislin scientific-data plotting libraries. We wrap them as Web Services and integrate into the proposed system as a layer providing service such as WMS. **Job Manager** (implemented by Gadgil H.) is actually a scripting technique for managing distributed workflows. Different Geo-Science applications (such as Pattern Informatics [19] and Virtual California [21]) require different set of parameters for the application users to trigger the job manager. This set of parameters and their order are defined earlier by the Job manager and user portal knows how to invoke it. **WS-Contex**'s (implemented by Aktas M.) specifications defined by OASIS (Organization for the Advancement of Structured Information). When multiple Web services are used in combination, the ability to structure execution related data called context becomes important. WS-Context provides a definition, a structuring mechanism, and a software service definition for organizing and sharing context across multiple execution endpoints. We use it for implementation of asynchronous service runs. In our framework we take science application as a black box. All the communication with this black box is handled by the Job Manager. We integrate ServoGrid [16, 17 and 18] science application to the system.

In order to utilize and interact with the coupling framework enabling access to computational and data resources in an ordered and synchronized manner, we develop innovative interactive smart decision making tools. Interactive decision making tools enable feeding the scientific applications with the data and associate the input and output of the applications at the layer (view)-level. The decision making tools also enable the accessing and querying of the attributes of the data/information which are building these comprehensible representations in overlaid layers interactively. Resources are organized into projects and they are compliant with some resource schemas defined in Aggregator Web Map Server (WMS). The client portal provides both map-based and project based dynamically updated user interfaces based on the capability metadata of initially connected Aggregator WMS. 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.

Addressing heterogeneity issues: In order to solve data and service heterogeneities for the proposed GIS internally (triangle in Figure 3), we use universally accepted and widely used OGC and ISO/TC-211 standards. This enables us to extend the system with other compatible data sources (WFS) and Map Services (WMS). Since we develop Information system with SOA approach using Web Services we need to solve the problems of inter-service communications of services using different communication protocols. OGC's standard protocol is HTTP-GET/POST. In contrast, Web Service architecture we implement uses SOAP+HTTP protocol. We propose online handlers containing service specific adapters for mapping request and response heterogeneity at the service interface and communication level. For example Web Service version of OGC compatible services accept the requests in structured format but original OGC services request are in attribute value pairs. Since the semantic of the queries are well-defined by OGC converting them to each other will be done the standard wrappers. Each different service type needs its own type of handler.

We use OGC defined semantics for service definitions (capability file) and data definitions (GML). We use two classes of services. These are WMS and WFS. Heterogeneous data sources are integrated to this system through WFS services. Each WFS has its own type of adaptor for their own data holdings and, it converts local data schema to general data model (GML). WFS is a kind of front end to the heterogeneous data-sources enabling data access and query. We currently use only databases for storing and handling the data.

Addressing Performance Issues: We build services for supporting scientific GIS applications that includes computation and data resource demanding high-performance and high-rate data transfers.

We take the performance into consideration from two respects. First one is related to structured data handling such as parsing, rendering and displaying. The second one is related to data transfer issues at the application (or software) level. Since the views are composed of layers provided by Aggregator WMS and Sci-Plotting Servers (see Figure 2 and Figure 3), we mostly focus on these servers' performances and architectures.

We implement streaming data transfer using publish-subscribe based messaging middleware called Naradabrokering. Parsing and rendering of the structured GML data we use Pull Parsing technique.

Since all the data in the system (geo-data) is defined and queried in ranges by bounding boxes, we do range query partitioning to implement parallel processing to increase the performance of the data access, display and querying. Parallel processing is done for handling the data changing very often. Parallel processing technique is applied together with caching, cached data extraction and rectangulation processes. Alternatively, for the archived data not changing often, we apply the pre-fetching approach to store the data in the intermediary Map Servers to fasten the rendering and the response times.

6. Tests and Analysis

We will perform (1) usability tests and (2) performance tests. We will be testing our proposed system with our creation of interactive decision making tools over real ServoGrid science applications. These are Pattern Informatics (PI) [19] and Virtual California (VC) [20].

PI application is used to produce the well-publicized "hot-spot" maps. PI analyzes earthquake seismic records to forecast regions with high future seismic activity, i.e. earthquake happenings.

VC is earthquake simulation model for the California. The simulation takes into account the gradual movement of faults and their interaction with each other.

(1) The tests will be done for the **usability** of the coupling framework:

- Testing basic interactive data (map) display tools, zoom-in, panning, moving etc.
- Interactive unified data querying through 3-layer structured display. click on the map image and get the attribute of the data in a pop-up widow.
- Map animations. Creating map movies with user provided parameters interactively. Animations are composed of series of static map images in 3-layered structure format.
- Coupling GIS System with **PI** science application
  - We run PI code through the user portal and plot the possibilities of the earthquake happenings in color-coded grid over the previously created seismic and earth map.
- Coupling GIS system with **VC** science application.
  - We run VC code through the user portal and play the result forecast values as a movie streams. Streams can be captured and watched by JMF Client software. Each frame in the stream is actually a 3-layer structured static map created based on user provided data and simulation outputs and rendered by Sci-Plotting Web Services.

Measurement of success for the usability: sending the link to some newsgroups and email groups and getting feedbacks.

(2) The tests will be done for the **performance**

For different data size

- Rendering and display of data *(bottom and middle layer in* Figure 2*)* ***(T)***
  - Caching and parallel processing (for the data change very often)
  - Pre-fetching (for the data do **not** change often)
- Science application output rendering and display timing *(top layer in* Figure 2*)* ***($t_{2.1}$ and T)***
  - Sci-Plotting service performance
  - Data sizes vs rendering (or plotting) time
- Structured data (GML) transfer issues (from WFS to WMS –see Figure 3) ***($t_{1.1}$-$t'_{1.3}$)***
  - Streaming data transfer (through publish/subscribe based messaging middleware -NaradaBrokering [22])
  - Non-streaming data transfer
- Stress tests over Virtual California science application for streaming map movies applications requiring high performance data transfer and rendering.

Measurement of success for the performance: since the system is user-oriented and interacted through browser based interactive tools, the practical measurement is browser time-outs. Each browser (such as IE and mozilla) has its internally defined http connection time-out limits. We will also compare the performances of the system with the other well-known deegree project [22] and Minnesota MapServer [23] mapping tools and GML data rendering system.

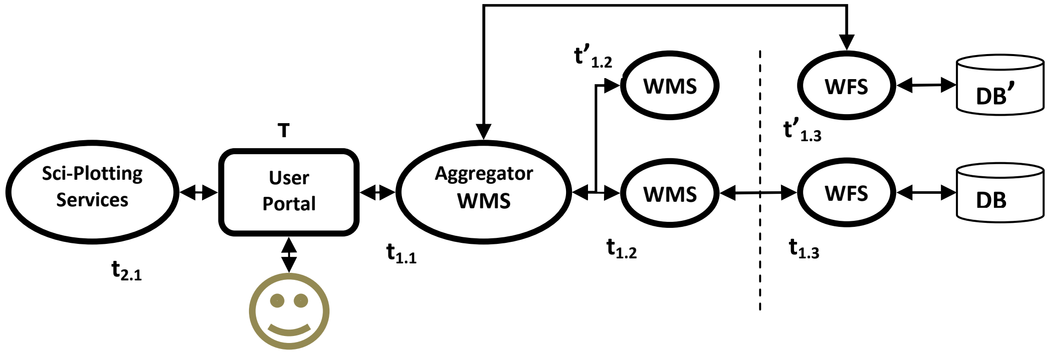

Figure 4: Performance test illustrations.

7. Related Work

Linked Environments for Atmospheric Discovery (LEAD) [25] is a large scale project funded by NSF Large Information Technology Research grant for addressing fundamental IT and meteorology research challenges to create an integrated framework for analyzing and predicting the atmosphere [27]. On the other hand we aim same things for the earth-related Geo-science. LEAD and our architecture both use SOA for the utilization of distributed sources.

At the application level, LEAD supports for adaptive analysis and prediction of mesoscale meteorological events. They call it forecast mode using available observations or model generated data and manages necessary resources. They mostly focus on automated data management, scalable data archiving system and easy search and access interfaces via GUI and underlying ontology. LEAD has MyLEAD concept for enabling users to process their own data.

Users can interactively explore the weather as it evolves, create custom scenarios or acquire and process their own data. Our approach is based on accessing querying and displaying the data not management.

As we do they also provide a web-portal as the entry point for students, users or advanced researchers to the meteorological data, services, models, and workflow, analysis and visualization tools related to the project.

Like LEAD, GEON [26] is also SOA based architecture adopted with a portal developed for end-users and, NSF funded large scale project involving the development of a distributed, services-based system that enables geoscientists to publish, integrate, analyze, and visualize their data.

In summary, compared to the related projects our contributions are: (1) Introducing the innovative techniques for high performance structured data rendering and (2) interactive and seamless geo-data access and querying architectures and, (3) general issues of view-level coupling of geo-data and corresponding output data processed through science applications. We also work on creating interoperable Geographic Information System in SOA and universally accepted standards and, make it extendable via the WMS mediators to create more complex information at the view level.

8. Timeline

General blueprint arch and implementations are done. We have developed coupling framework in SOA architecture. We have developed streaming and non-streaming versions of WMS and WMS Aggregator services. We have developed interactive browser based tools to interact the system such as static map tools, map animation tools and coupling tools. We still need to work on Sci-Plotting server to make it more general.

The remaining is mostly related to performance issues and tests. We have finished implementing the pre-fetching and streaming data transfer techniques but need to work on caching, range query partitioning and parallel processing.

Internally we developed an interoperable system using universally accepted standards (OGC). Since the proposed system is in SOA, our standard OGC services are extended as Web Services. In order to make the system extensible with other OGC compatible services, we will develop proposed a standard distributed handler for Map Services to solve interface level heterogeneity. Handler enable integrating-layering third party OGC Map Servers' layers by solving interface and protocol level heterogeneity and, enable creation of more complex information. We will test the system for usability and performance.

References

[1] GIS Research at Community Grids Lab, Project Web Site: http://www.crisisgrid.org.

[2] Ahmet Sayar, Marlon Pierce, Geoffrey Fox OGC Compatible Geographical Information Services Technical Report (Mar 2005), Indiana Computer Science Report TR610

[3] Peng, Z.R. and M. Tsou, Internet GIS: Distributed Geographic Information Services for the Internet and Wireless Networks. 2003: Wiley

[4] Cox, S., Daisey, P., Lake, R., Portele, C., and Whiteside, A. (eds) (2003), OpenGIS Geography Markup Language (GML) Implementation Specification. OpenGIS project document reference number OGC 02-023r4, Version 3.0.

[5] ESRI, ArcIMS, 9 Architecture and Functionality, J-8694. ESRI White Paper, http://downloads.esri.com/support/whitepapers/ims_/arcims9-architecture.pdf. 2004.

[6] Autodesk. MapGuide http://usa.autodesk.com. [cited.

[7] MapServer, W. http://www.wthengineering.com/GIS/web_gis.htm. [cited.

[8] Di, L., et al., The Integration of Grid Technology with OGC Web Services (OWS) in NWGISS for NASA EOS Data, in GGF8 & HPDC12 2003: Seattle, USA. . p. 24-27.

[9] Fox, G. and M. Pierce. Web Service Grids for iSERVO. in International Workshop http://www.eps.s.u-tokyo.ac.jp/jp/COE21/events/20041014.pdf on Geodynamics: Observation, Modeling and Computer Simulation University of Tokyo Japan October 14 2004. 2004.

[10] de La Beaujardiere, J., Web Map Service, OGC project document reference number OGC 04-024. 2004.

[11] Vretanos, P. (2002) Web Feature Service Implementation Specification, OpenGIS project document: OGC 02-058, version 1.0.0. Volume,

[12] Cox, S. (2003) Observations and Measurements. Volume, DOI: OGC 03-022r3.

[13] Apache Axis, http://ws.apache.org/axis/.

[14] Apache Tomcat, http://tomcat.apache.org/.

[15] OGC (Open Geospatial Consortium) official web site http://www.opengeospatial.org/

[16] Fox, G. and M. Pierce. Web Service Grids for iSERVO. in International Workshop http://www.eps.s.u-tokyo.ac.jp/jp/COE21/events/20041014.pdf on Geodynamics: Observation, Modeling and Computer Simulation University of Tokyo Japan October 14 2004. 2004.

[17] Aydin, G., et al. SERVOGrid Complexity Computational Environments (CCE) Integrated Performance Analysis. in Grid Computing, 2005. The 6th IEEE/ACM International Workshop on. 2005: IEEE.

[18] Fox, G. and M. Pierce, SERVO Earthquake Science Grid, in summary of iSERVO technology October 2004 in January 2005 report High Performance Computing Requirements for the Computational Solid Earth Sciences edited by Ron Cohen and started at May 2004 workshop on Computational Geoinformatics.

[19] Tiampo, K. F., Rundle, J. B., McGinnis, S. A., & Klein, W. Pattern dynamics and forecast methods in seismically active regions. Pure Ap. Geophys. 159, 2429-2467 (2002).

[20] Rundle, JB, PB Rundle, W Klein, J Martins, KF Tiampo, A Donnellan and LH Kellogg, GEM plate boundary simulations for the Plate Boundary Observatory: Understanding the physics of earthquakes on complex fault systems, Pure and Appl. Geophys., 159, 2357-2381 2002.

[21] Rundle, P.B, J.B. Rundle, K.F. Tiampo, A. Donnellan and D.L. Turcotte, Virtual California: Fault Model, Frictional Parameters, Applications, PAGEOPH, submitted

[22] Pallickara, S. and G. Fox. NaradaBrokering: A Middleware Framework and Architecture for Enabling Durable Peer-to-Peer Grids. in Lecture Notes in Computer Science. 2003: Springer-Verlag.

[23] Deegree projects home page http://deegree.sourceforge.net/

[24] UMN MapServer project home page http://mapserver.gis.umn.edu/

[25] Beth Plale, Dennis Gannon, Dan Reed, Sara Graves, Kelvin Droegemeier, Bob Wilhelmson, Mohan Ramamurthy, “Towards Dynamically Adaptive Weather Analysis and Forecasting in LEAD”, *To appear ICCS workshop on Dynamic Data Driven Applications*, Atlanta, Georgia, May 2005.

[26] GEON (Geosciences Network): A Research Project to Create Cyberinfrastructure for the Geosciences. http://www.geongrid.org

[27] Kelvin K. Droegemeier, et al. Linked environments for atmospheric discovery (LEAD): A cyberinfrastructure for mesoscale meteorology research and education. in 20th Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, . 2004. Seattle, WA.