

Introduction

Information Technology (IT) is changing all aspects of life, including how scientific research is conducted. Over the years, a few grass-roots efforts developed in the geosciences that started to take advantage of these new resources. However, the impact of IT in the geosciences has not been at a high level yet. To improve understanding of IT and familiarize a group of students and researchers with these new tools, we have designed a Summer Institute for Geoscientists lecture series primarily focused on immediate needs of the community in the area of information technology. The goal of this Summer Institute is to educate a group of earth scientists and expand the community of IT users in earth science research.

With the broad and rapid adoption of IT in science, and the advent of major initiatives such as the NSF *Cyberinfrastructure* program and UK's E-science activity, it is essential that the community at large be in a good position to take full advantage of new opportunities to further improve research and educational activities via IT-enabled approaches. This Summer Institute is a step in this direction and designed as an educational and outreach activity in the geosciences primarily focusing on immediate needs. We have chosen a format that provides a quick introduction to six key IT topics that are relevant to the overall science goals of the community in the immediate term. Each course will be taught by an expert in that field and assisted by other experts from the San Diego Supercomputer Center and NCAR. Lectures will be followed with hands-on lab exercises. In the following sections we describe each of these courses in detail.

Description of Workshop Courses

1. Data Management

Scientific applications and research are increasingly becoming interdisciplinary. As a result already complex data management issues in the geosciences become even harder to deal with using classical data integration mechanisms. However, data integration is a well-established and vibrant area of research in computer sciences, and knowledge transfer to earth scientists in this area will be essential for successful continuation of leading-edge earth science research. The goal of this one-day course is to introduce scientists to the basic concepts, techniques, and tools for database integration and knowledge representation in the context of scientific data modeling and integration.

The course will include a theoretical, lecture-style part on the fundamentals of data modeling and integration—which will take about two-thirds of the day—and a practical "hands-on" part of the course, which will take the remaining time.

The course will provide fundamentals of data integration approaches, such as layered database federation architecture (3-layered vs. 5-layered), *mediator* architecture, and the concepts of *global-as-view* versus *local-as-view* approaches to data integration. The course will also include an introduction to the topic of, "*Knowledge Representation with Ontologies.*" Topics covered will include what ontologies are; how they are used and what standards and tools exist. The lectures will also provide an introduction to conceptual data modeling and structural representations. Modeling with Entity-Relationship (ER) and/or UML diagrams, and representing data using relational and XML schemas will also be discussed. Finally, we will end with an introduction to *semantic data registration* and address the issue of how one can "register" one's "neatly modeled" dataset/database to a data federation registry. We will consider the question of how the data mediation system will utilize this information to support "smart" data discovery and integration.

The "hands-on" part of this lecture will include practical exercises, in teams of two or three persons, to explore the concepts introduced earlier in the course, to encourage active participation and further the understanding of the new material. Planned exercises include:

- modeling a mini-domain: ontology, conceptual model, relational and XML data model
- creating your own ontology in an ontology editor (e.g. Protégé)
- understanding the *consequences* of the ontology (using description logic reasoners and Formal Concept Analysis (FCA))
- demonstration of a "behind the scenes" operation of a database mediator (including, for example, view unfolding, reasoning to check query containment, and chasing queries with integrity constraints)

2. Web Services

The geosciences are a collection of primarily data and observation driven disciplines, yet a mechanism to share collected data and developed software tools has not been widely established. The data collected are stored in several different formats on different platforms. Software developed in the community employs a variety of mechanisms for accessing such data and conduct analysis on them, with little or no collaboration and standards. To make these data available to the larger scientific community and to stimulate integration of resources, earth scientists need to get involved in an infrastructure for integrating this information. In order for scientists to have a common mechanism of accessing and querying their data and tools, they need to have a uniform interface for retrieving information.

Web Services provide this common platform. The Web Services technology allows, for the first time, a standards-based approach to a workable interoperation framework. The World Wide Web consortium (W3C – <http://www.w3c.org>) describes Web Services as: "*A software system designed to*

support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards”.

This section of the Summer Institute will provide a day-long course on Web services: what they are, how to build them, and how to use in them in scientific research. The goal of this course is to help participants become scientific data and/or tool providers for the community, thereby enabling overall research integration efforts. Participants should be able to go back to their respective institutions and be able to employ some of these techniques and implement Web services interfaces in order to “publish” their own research resources.

In this course we will cover:

- The understanding of service oriented architectures (SOA)
- Web services and their benefits
- Hands-on training for creating web services
- Using “real” data sets, help attendees build sample data web services
- Use real data sets to build mapping web services
- Pick Some analysis tools/application and create some analysis web services

3. Grid Technology

Grid technology is a growing new field in computer sciences that promises major benefits to science applications, including in the earth sciences. In this portion of our Summer Institute we will provide a day-long overview of Grid computing technology and how it can benefit the sciences. The day-long lecture will focus on Data Grid technologies. We will review various existing technologies that enable data to be accessed via the grid, archived and stored in collections, and moved from one physical location to another. Grid Security mechanisms will also be discussed.

We will provide a high-level overview of Grid Computing and its relationship to other technologies. Basic ideas and rationale will be discussed along with a short history of Grid computing. We will then continue with data access methods in the grid. Emerging standards such as OGSA-DAI (Open Grid Services Architecture – Data Access and Integration) will be discussed.

We will continue with an examination of data storage and management mechanisms. We will focus on databases (DB2, Oracle), data federation (including, DB2 Information Integrator), GPFS (IBM’s General Parallel File System), PVFS (Parallel Virtual File System), SRB (SDSC Storage Resource Broker), and the Virtual Data Toolkit (VDT). Data movement will be discussed using GridFTP, RFT (reliable file transfer service), SRB and GMR (IBM Grid Movement and Replication).

An important component of Grid technology is security mechanism. Grid security is based on GSI (Grid Security Infrastructure) public/private key pairs. However, managing key pairs (management and distribution of keys) is non-trivial. We review the available tools including MyProxy servers and role-based authentication with CAS (Community Authentication Service) and VOM (Virtual Organization Management).

Having reviewed the grid technology we will take a look into existing grids. In particular, we would like to show how tools described above are used in today's grids, such as TeraGrid, GEON, GriPhyN (Grid Physics Network), the Earth System Grid (ESG), and BIRN (the Biomedical Informatics Research Network).

Finally, the lectures will conclude with a review of upcoming new standards and tools. In particular, we will briefly introduce OGSA (Open Grid Service Architecture) and the new WSRF (Web Services Resource Framework).

4. Geographic Information Systems

Geographic Information Systems (GIS) provide the capability to manage and manipulate geospatial information, including rendering of maps and map-based information. Increasingly, GIS has become a critical tool for managing spatial data in a variety of disciplines. Indeed, creating, managing, analyzing and presenting spatial data is now a fundamental component of many scientific data analyses and decision-support systems, including in the geosciences. The main thrust of GIS is integration of heterogeneous data based on co-location in space. Joining data based on spatial relationships (proximity, adjacency, containment, intersection, etc.) requires that data layers are converted into a common coordinate system. Additional issues critical for spatial data interoperability include standard registration of data set schemas and semantics, format and spatial data type conversion, management of resolution and uncertainty, identity of spatial objects, etc. These issues, along with techniques and standards for spatial data representation and online mapping, are the focus of the workshop.

The GIS course will begin with an introduction to the capabilities of a typical GIS, and review some applications in the geosciences. We will address common misconceptions about GIS and, conclude the lecture with research issues in GIS. We will discuss common GIS data structures and analysis techniques and provide use case scenarios. We will also introduce a variety of GIS-related software products: ArcGIS (desktop GIS from ESRI) and spatial databases such as PostgreSQL/PostGIS, ArcSDE, Internet map servers (e.g., ArcIMS) and Web client rendering technologies (e.g. SVG). A special focus will be on online mapping and integration of heterogeneous spatial data from multiple data servers.

The hands-on portion of the workshop will introduce the popular ArcGIS desktop GIS system, and software for registration and discovery of spatial information, and for serving spatial data online (with ArcIMS). The course will briefly discuss geospatial data and metadata: discovery and integration, GIS-related data standards (GML and other OpenGIS standards; SVG and other W3C standards). Our examples will include service-based infrastructure for managing geosciences data from the project GEON, assembling maps as well as hands-on exercises (ESRI Virtual Campus introduction; ArcGIS; ArcIMS; ArcExplorer Web)

5. Parallel Computing

Our introduction to parallel computing will start with basic definitions and descriptions, and how it can help science applications. We will discuss the advantages and disadvantages of parallel computing along with its limits and pitfalls.

At the high-end of computing, parallel computing can be viewed as enabling "Big Science." We will introduce a motivating example (e.g., the Stommel model), where a single processor could solve small versions of the model, but large problems require parallel computing.

We will discuss the general methodology for taking the serial version of the Stommel model and converting it to run as a parallel application. The speed up of the Stommel model for various problem sizes and numbers of processors will be shown and discussed. This will also further illustrate the advantages and limits of parallel computing.

There are two primary methods of parallelizing programs. We will show how the Message Passing Interface (MPI) is used for programming distributed memory machines using message passing. OpenMP will be used to show how programs running on shared memory machines can be parallelized.

The basic subroutines of MPI will be explained. We will see how these routines are used to do message passing programming with simple "*hello world*"-style programs. After the introduction we will go back to a serial version of the Stommel model and show how MPI can be used to parallelize it to run on multiple processors using message passing.

OpenMP is a directives based parallelization technique designed for shared memory processors. We will again discuss the basic syntax and semantics of OpenMP. After the introduction with simple examples we will go back and parallelize the Stommel model using OpenMP. Source code for the examples will be available for the students.

6. Scientific Workflows

Due to their intuitive nature, scientific model building and execution environments, and scientific workflows, are rapidly becoming the preferred way for scientists to design and execute local and/or distributed calculations, using a graphical user interface. Traditionally, workflows are used in streamlined calculations and processes. However, workflows can be used dynamically to study and examine scientific tasks and hypothesis testing.

A Scientific Workflow System (SWS) allows a user to integrate data integration and analysis steps into larger "analytical pipelines", possibly for (semi-) automatic execution. This requires an understanding of design methodologies for scientific workflows, component architectures, models of component interaction etc. The Summer Institute lecture on this topic will include:

- Presentations of state-of-the-art approaches to modeling scientific workflows, e.g., actor-oriented modeling and design: AMD vs. object-oriented modeling and design, existing component architectures, e.g. Common Component Architecture (CCA).
- Hands-on exercises to provide a brief introduction to the Kepler Scientific Workflow System. Participants will be provided with the Kepler-2-GO CD and
 - install the system locally
 - run pre-design workflows with different parameter settings
 - design their own mini workflow