A Three Tier Architecture for LiDAR Interpolation and Analysis

Efrat Jaeger-Frank[1], Christopher J. Crosby[2], Ashraf Memon[1], Viswanath Nandigam[1], J. Ramon Arrowsmith[2], Jeffery Conner[2], Ilkay Altintas[1], and Chaitan Baru[1]

[1] San Diego Supercomputer Center, University of California, San Diego,
9500 Gilman Drive, La Jolla, CA 92093, USA
{efrat, amemon, viswanat, altintas, baru}@sdsc.edu

[2] Department of Geological Sciences, Arizona State University,
Tempe, AZ 85281, USA
{chris.crosby, ramon.arrowsmith, jsconner}@asu.edu

Abstract. Emerging Grid technologies enable solving scientific problems that involve large datasets and complex analyses. Coordinating distributed Grid resources and computational processes requires adaptable interfaces and tools that provide a modularized and configurable environment for accessing Grid clusters and executing high performance computational tasks. In addition, it is beneficial to make these tools available to the community in a unified framework through a shared *cyberinfrastructure*, or a portal, so scientists can focus on their scientific work and not be concerned with the implementation of the underlying infrastructure. In this paper we describe a scientific workflow approach to coordinate various resources as data analysis pipelines. We present a three tier architecture for LiDAR interpolation and analysis, a high performance processing of point intensive datasets, utilizing a portal, a scientific workflow engine and Grid technologies. Our proposed solution is available through the GEON portal and, though focused on LiDAR processing, is applicable to other domains as well.

1 Introduction

With improvements in data acquisition technologies comes an increase in the volume of scientific data The demand for efficient processing and management of the data have made Grid infrastructures an essential component in a wide range of scientific domains. Grid infrastructure technologies enable large scale resource sharing and data management, collaborative and distributed applications and high performance computing, for solving large scale computational and data intensive problems. However, the distributed and heterogeneous nature of Grid clusters, such as various hardware platforms and software systems, access and interaction interfaces and data and resource management systems, make the Grid environment difficult to use by the layman, and thus require additional management to coordinate the multiple resources. In this paper we propose a

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part III, LNCS 3993, pp. 920–927, 2006.

coordination of Grid resources in a workflow environment as part of a three tier architecture. The workflow system provides a modularized and configurable environment. It gives the freedom to easily plug-in any process or data resource, to utilize existing sub-workflows within the analysis, and easily extend or modify the analysis using a drag-and-drop functionality through a graphical user interface.

The Geosciences Network (GEON) [1] is an NSF-funded large Information Technology Research (ITR) project to facilitate collaborative, inter-disciplinary science efforts in the earth sciences. GEON is developing an infrastructure that supports advanced semantic-based discovery and integration of data and tools via portals (the GEON portal), to provide unified and authenticated access to a wide range of resources. These resources allow geoscientists to conduct comprehensive analyses using emerging web and Grid-base technologies in order to facilitate the next generation of science and education. One of the challenging problems GEON is currently focusing on is distribution, interpolation and analysis of LiDAR (Light Distance And Ranging) [2] point cloud datasets. The high point density of LiDAR datasets pushes the computational limits of typical data distribution and processing systems and makes grid interpolation difficult for most geoscience users who lack computing and software resources necessary to handle these massive data volumes. The geoinformatics approach to LiDAR data processing requires access to distributed heterogeneous resources for data partitioning, analyzing and visualizing all through a single interactive environment. We present a three tier architecture that utilizes the GEON portal as a front end user interface, the Kepler [3] workflow system as a comprehensive environment for coordinating distributed resources using emerging Grid technologies, and the Grid infrastructure, to provide efficient and reliable LiDAR data analysis. To the best of our knowledge, there exists no previous work on utilizing a scientific workflow engine as a middleware behind a portal environment for coordinating distributed Grid resources.

The rest of this paper is organized as follows. Section 2 provides an introduction to LiDAR data and describes the traditional processing approach. Section 3 gives a brief overview of the Kepler scientific workflow system. The novel approach for LiDAR processing, utilizing the Kepler workflow engine through the GEON portal is described and analyzed in Section 4. We conclude in Section 5 and discuss expansions of the current work aimed at making GEON a leading portal for LiDAR processing.

2 Introduction to LiDAR and Previous Approach

LiDAR (Light Distance And Ranging, a.k.a. ALSM (Airborne Laser Swath Mapping)) data is quickly becoming one of the most exciting new tools in the Geosciences for studying the earth's surface. Airborne LiDAR systems are composed of three separate technologies: a laser scanner, an Inertial Measurement Unit (IMU) and a Global Positioning System (GPS) all configured together to calculate the absolute location for the earth's surface based upon each individual laser return. The systems typically record one or more returns per square meter on the ground with an absolute vertical accuracy of better than 15 cm. Capable of generating digital elevation models (DEMs) more than an order of magnitude more accurate than those currently available, LiDAR data offers geologists the opportunity to study the processes the shape the earth's surface at resolutions not previously possible. LiDAR data is currently being utilized by earth scientists for a wide variety of tasks, ranging from evaluating flooding hazards to studying earthquake faults such as the San Andreas [2].

Unfortunately, access to these massive volumes of data generated by LiDAR is currently difficult, and the average geoscience user is faced with the daunting task of wading through hundreds or thousands of ASCII flat files to find the subset of data of interest. The distribution, interpolation and analysis of large LiDAR datasets, currently performed on a desktop PC with software packages available and familiar to most earth scientists, also presents a significant challenge for processing these types of data volumes. In the current state of affairs, the popularity and rate of acquisition of LiDAR data far outpaces the resources available for researchers who wish to work with these data. The geoinformatics approach to LiDAR processing described herein represents a significant improvement in the way that geoscientists access, interpolate and analyze LiDAR data. The improved, internet-based approach, acts to democratize access to these exciting but computationally challenging data.

3 Kepler: A Scientific Workflow System

Kepler [4, 5] is a cross-project, multi-disciplinary collaboration to build open source tools for scientific workflows that provide domain scientists with an easy-to-use, yet powerful system for capturing and automating their ad-hoc process. Kepler is built on top of the PtolemyII system developed at UC Berkeley, which provides a set of java APIs for modeling heterogeneous, concurrent and hierarchical components by means of various models of computations [6, 7]. Kepler provides the scientists with a repetitive and configurable environment available through a graphical user interface and as a command-line tool. It combines high-level workflow design with execution and runtime interaction, access to local and remote data and legacy applications, and local and remote service invocation along with a built-in concurrency control and job scheduling mechanism.

Computational units in Kepler are called *actors*, which are reusable components communicating with each other via input and output ports. The control of flow of actors is orchestrated by a *director* that specifies the model of computation. Kepler uses the Modeling Markup Language (MoML), inherited from the underlying PtolemyII system, as its workflow description language. MoML is a modularized and extensible XML modeling language where actors can be defined as place holder stubs to be set prior to the workflow execution. In the next section we describe how utilizing the Kepler features enhances LiDAR processing.

4 A Scientific Workflow Based Approach for LiDAR Processing

In the following section we present a three tiered Kepler scientific workflow solution for facilitating distribution, interpolation and analysis of LiDAR datasets.

4.1 Coordinating Distributed Resources in a Single Environment

LiDAR processing requires three main computational steps each deployed on a distributed resource: querying point cloud datasets, processing the data using various interpolation algorithms, and visualizing the results. Coordinating these steps in a Kepler scientific workflow provides scheduling and monitoring of each task and communication between the various resources. Furthermore, the scientific workflow environment gives us modularity and extensibility through reusable actors. The LiDAR processing workflow, depicted in Figure 1, provides a *conceptual* workflow where each of the components can be dynamically customized by the availability of the data and processing algorithms and is set on the fly prior to the workflow execution. This modularized approach can be captured as a workflow pattern of *subset*, *analyze*, and *visualize*. Below we elaborate on each of the processing steps.

The massive amount of LiDAR data (currently two datasets at ∼1 billion points each) were uploaded and organized in a DB2 spatial database on *DataStar* to provide a unified structure to the collected data. The subset query returns all points (X,Y,Z) that reside within a user selected bounding box. The database connection information along with the query are specified as workflow parameters and are set on the fly prior to the workflow execution. The query is then performed on DataStar where the result is also stored on an NFS mounted disk.

The analysis step consists of an interpolation algorithm. As shown in the figure, the query response is shipped to the *analysis* cluster and is then interpolated

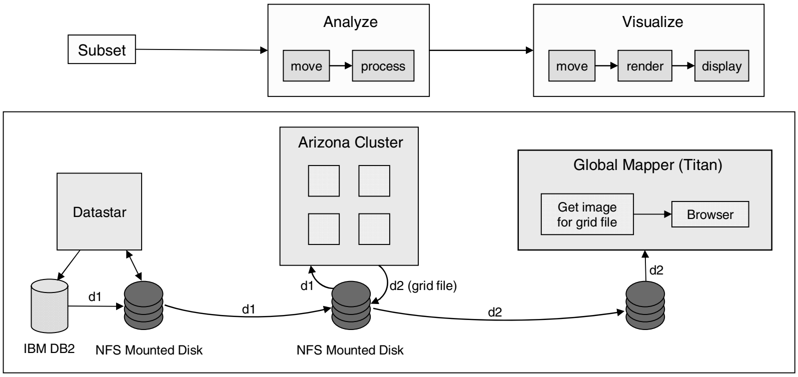

Fig. 1. Coordinating distributed heterogeneous resources in a single scientific workflow environment

into a regularized grid. Currently we use the GRASS spline interpolation [8] algorithm deployed on the Arizona GEON four nodes cluster. Other interpolation algorithms, for example, Inverse Distance Weighted (IDW) or Krieging algorithms may be plugged in as well. Interpolation of the high-point density LiDAR data currently constitutes the bottleneck in the overall process. Parallelization of the interpolation code or alternative interpolation algorithms will likely result in improved performance.

The interpolation results may be visualized and/or downloaded. At present, the Global Mapper imaging tool[1] is used to create a display from the interpolated results. Global Mapper, available to the GEON community through a web service, takes an ASCII grid file as an input and returns a URL to the resulting image. The image can be displayed on a web browser and requires no specific visualization components. Other visualization methods may be applied as well.

4.2 A Three Tier Architecture

The GEON project is developing a web-based portal for gathering GeoScience applications and resources under a single roof. In order to make the workflow based approach for LiDAR processing uniformly accessible through the GEON portal, our proposed solution is based on a three tier architecture (depicted in Figure 2): the *portal layer*, the *workflow layer* or control layer, which is the Kepler workflow system, and the *Grid layer* as the computation layer.

The *portal layer*, a portlet, serves as a front end user interface. It enables the user to partition the data using an interactive mapping tool and attribute selection through a WMS map[2]. Algorithms, processing attributes and desired derivative products, are also chosen using a web interface. Within Kepler, one can design a predefined parameterized workflow template which is modularized and configurable using place holder stubs to be set prior to the workflow execution. The aforementioned *"subset, analyze, visualize"* workflow pattern serves as a conceptual workflow template, defined in the Kepler workflow description language, MoML. A workflow instance is created on the fly from the conceptual workflow based on the user selections. The instantiated workflow is then scheduled to be executed by the workflow layer.

The *workflow layer*, also referred to as the main control layer, communicates both with the portal and the Grid layers. This layer, controlled by the Kepler workflow manager, coordinates the multiple distributed Grid components in a single environment as a data analysis pipeline. It submits and monitors jobs onto the Grid, and handles third party transfer of derived intermediate products among consecutive compute clusters, as defined by the workflow description. In addition, it sends control information (a.k.a. tokens) to the portal client about the overall execution of the process. The workflow is executed by the Kepler engine in a batch mode. Once a job is submitted, the user can detach from the system and receive an email notification after the process has completed.

[1] http://www.globalmapper.com/

[2] OpenGIS Web Mapping Specification (http://giserver.esrin.esa.int/quickwms/).

As the LiDAR processing workflow involves long running processes on distributed computational resources under diverse controlling authorities, it is exposed to a high risk of component failures, and requires close monitoring. In order to overcome these failures with minimal user involvement, Kepler provides a data provenance and failure recovery capability by using a job database and smart reruns. The job database is used for logging the workflow execution trace and storing intermediate results along with the associated processes/components that were used to produce them. The workflow engine maintains information about the status of each intermediate step, and this can be used to initiate a smart rerun from a failure point or a checkpoint. These advanced features thus eliminate the need to re-execute computationally intensive processes.

The *Grid layer*, or the execution layer is where the actual processing implementations are deployed on the distributed computational Grids. Currently a simple submission and queueing algorithm is used for mapping jobs between various resources based on the number of allocated tasks and the size of the data to be processed. In the near future we plan to utilize the Pegasus [9] Grid scheduler to benefit from mapping jobs based on resource efficiency and load, thus making the process more robust. We also plan to extend this further by deployment of the computationally challenging processes on a higher performance machine, for example, DataStar, which consists of 32 P690 processors with 128GB of memory running at 1.7GHz, each with a gigabit connection to an underlying SAN disk infrastructure, making it an ideal machine for compute intensive tasks.

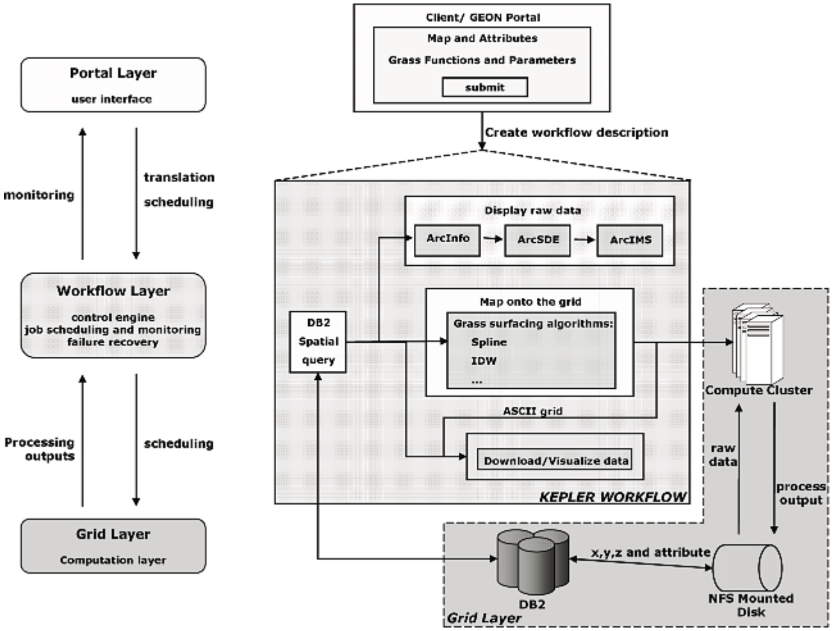

Fig. 2. Three tier LiDAR processing architecture

4.3 Data Upload and Access

As initial pilot datasets, LiDAR data collected along the Northern San Andreas Fault (NSAF) in Sonoma and Mendocino Counties, California and data from the Western Rainier Seismic Zone in Pierce County, Washington were utilized. The approximately 1.2 billion points in each of these datasets were stored in a DB2 Spatial Extender Relational Database on DataStar. To optimize query performance, and due to its sheer volume, the data were grouped by USGS Quarter Quadrants[3] and organized in the database as a table per quarter quadrant. The multiple data segments are accessible through the WMS map using a metadata table that specifies the bounding box coordinates for each segment. Each record in the datasets consists of a coordinate, elevation and corresponding attributes. The Spatial Extender feature in DB2, leveraging the power of standard SQL for spatial data analysis, enables storage, access and management of the spatial data. Furthermore, it provides a grid indexing technology for indexing multi-dimensional spatial data, dividing a region into logical square grids, thus facilitating region based subset queries.

4.4 Analysis of the Approach

The three tier based architecture utilizes the Kepler workflow system to reengineer LiDAR data processing from a Geoinformatics approach. The data and processing tools are accessible through a shared infrastructure and are coordinated using the Kepler engine. Kepler provides an incremental design and development environment which allows users to create incomplete workflow templates, which can be filled on demand prior to the workflow execution. Seamless access to resources and services is enabled through existing generic components such as SSH, Web Service, database access and command line processor actors. These generic building block components offer the flexibility to plug-in any applicable data or process, thus providing a customizable and modularized environment. The LiDAR workflow currently has access to two datasets both stored in a DB2 database on DataStar. Other databases sharing a similar structure may be used simply by pointing to a different database location. Currently we use GRASS spline interpolation, available via a web service deployed on the Arizona GEON cluster. Other interpolation algorithms such as GRASS IDW can be applied by updating a WSDL URL parameter. The visualization tools are interchangeable as well. Global Mapper can be replaced with FlederMaus[4] or ArcIMS visualization tools, both deployed on the GEON portal as web services.

The LiDAR workflow's main processing bottleneck is in the interpolation of the high point density datasets. Currently, the GRASS spline interpolation is limited to processing 1,600,000 points. This problem is being addressed by the GRASS development community and we anticipate a solution to this problem in the near future. Ultimately however, to improve performance, a parallel interpolation algorithm is required along with deployment on a higher performance

[3] http://www.kitsapgov.com/gis/metadata/support/qqcode.htm

[4] An interactive 3D visualization system (http://www.ivs3d.com/products/fledermaus/).

machine. We are also testing the Pegasus system [9] in coordination with the Pegasus group at the University of Southern California for a more efficient mapping of the interpolation sub-workflow onto the Grid.

5 Conclusion

In this paper we describe a three tier architecture for LiDAR data processing using a comprehensive workflow system, a shared cyberinfrastructure and evolving Grid technologies. The first version of this effort is available at the GEON portal (https://portal.geongrid.org:8443/gridsphere/gridsphere), and has already been incorporated as a public tool. We plan to extend this work in progress by making additional datasets available, and improving the overall performance with advanced processing tools such as parallel interpolation algorithms and enhanced visualization methods. We also intend to utilize the Kepler provenance system to link the workflow execution trace to the portal interface in order to provide extended user monitoring of the workflow execution status. Our goal is to provide a centralized location for LiDAR data access and interpolation that will be useful to a wide range of earth science users.

Acknowledgement. GEON is supported by NSF ITR 0225673. Special thanks to Dogan Seber and Ghulam Memon for their support in this work.

References

1. NSF/ITR: GEON: A Research Project to Creat Cyberinfrastructure for the Geosciences, www.geongrid.org
2. Carter, W.E., Shrestha, R.L., Tuell, G., Bloomquist, D., and Sartori, M., 2001, Airborne Laser Swath Mapping Shines New Light on Earth's Topography: Eos (Transactions, American Geophysical Union), v. 82. p. 549
3. Kepler: An Extensible System for Scientific Workflows, http://kepler.ecoinformatics.org
4. Ludäscher B., Altintas I., Berkley C., Higgins D., Jäger-Frank E., Jones M., Lee E.A., Tao J., Zhao Y.: Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows, to appear, 2005
5. Altintas I., Berkley C., Jäger E., Jones M., Ludäscher B., Mock S.: Kepler: Towards a Grid-Enabled System for Scientific Workflows, in the Workflow in Grid Systems Workshop in The Tenth Global Grid Forum, Germany, 2004
6. Lee E. A. et al.: PtolemyII Project and System. Department of EECS, UC Berkeley, http://ptolemy.eecs.berkeley.edu/ptolemyII
7. Liu X, Liu J., Eker, J., Lee E. A.: Heterogeneous Modeling and Design of Control Systems, in Software-Enabled Control: Information Technology for Dynamical System, Tariq Samad and Gary Balas, Wiley-IEEE Press, 2003
8. Mitasova, H., Mitas, L. and Harmon, R.S., 2005, Simultaneous Spline Interpolation and Topographic Analysis for LiDAR Elevation Data: Methods for Open Source GIS, IEEE GRSL 2(4), pp. 375- 379
9. Blythe J., Jain S., Deelman E., Gil Y., Vahi K., Mandal A., Kennedy K.: Task Scheduling Strategies for Workflow-based Applications in Grids. CCGrid 2005