Service-Oriented Grid Computing for SAFORAH

Ashok Agarwal[3], Patrick Armstrong[3], Andre Charbonneau[4], Hao Chen[2], Ronald J. Desmarais[3], Ian Gable[3], David G. Goodenough[1, 2], Aimin Guan[2], Roger Impey[4], Belaid Moa[1], Wayne Podaima[4], Randall Sobie[3, 5]

[1]Department of Computer Science, University of Victoria, Victoria, BC
[2]Canadian Forest Service, Natural Resources Canada (NRCan)
[3]Department of Physics and Astronomy, University of Victoria, Victoria, BC
[4]Information Management Services Branch, National Research Council Canada, Ottawa, Ontario
[5]Instittue for Particle Physics, Canada

Abstract. The SAFORAH project (System of Agents for Forest Observation Research with Advanced Hierarchies) was created to coordinate and streamline the archiving and sharing of large geospatial data sets between various research groups within the Canadian Forest Service, the University of Victoria, and various other academic and government partners. Recently, it has become apparent that the availability of image processing services would improve the utility of the SAFORAH system. We describe a project to integrate SAFORAH with a computational grid using the Globus middleware. We outline a modular design that will allow us to incorporate new components as well as enhance the long-term sustainability of the project. We will describe the status of this project showing how it will add a new capability to the SAFORAH forestry project giving researchers a powerful tool for environmental and forestry research.

Keywords: forestry, remote sensing, computing, grids

1. Introduction

Research projects are building very sophisticated instruments to study the world we live in and the universe surrounding us. Telescopes, satellites, accelerators, and ocean laboratories are collecting vast amounts of data for basic and applied research. Dealing with such large data sets is a challenge and many projects are adopting emerging data grid techniques. The Canadian Forest Service's SAFORAH (System of Agents for Forest Observation Research with Advanced Hierarchies) is an example of a project that collects data from a variety of land and spaced-borne remote sensing sources [1]. SAFORAH was created to coordinate and streamline the archiving and sharing of large geospatial data sets between various research groups within the Canadian Forest Service, and various other academic and government partners. Currently, only the raw or unprocessed images and data sets are made available on the SAFORAH data grid. Recently it has become apparent that providing researchers with processed images would significantly enhance SAFORAH. We describe a project to add the ability to process data in SAFORAH to create new information products

using a service-oriented computational grid based on the Globus Toolkit Version 4 (GT4) middleware [2].

The new functionality in SAFORAH is supported by integrating the GT4 computational grid provided by the Gavia Project as shown in Figure 1. The Gavia Project is a developmental computational grid using a service-oriented architecture initiated in 2006. It is based on GT4 and uses a metascheduler based on Condor-G [3]. The Gavia Project provides transparent access from the SAFORAH system to a set of distributed computational facilities in Canada. The architecture of the system also allows the use of other metaschedulers such as Gridway which are currently under development by other groups.

Gavia integrates the task brokering services provided by the Condor-G package with GT4. Prior to Gavia, the project team established the GridX1 Project, a computational grid in Canada using older versions of Condor-G and GT2 software. GridX1 was used to run production simulation jobs for particle physics experiments [4].

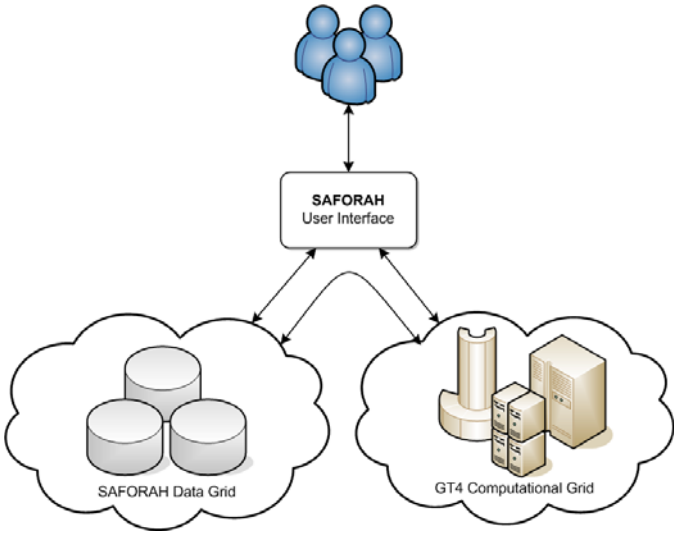

Fig. 1. High-level overview showing the integration of a GT4 computational grid into the SAFORAH system. The SAFORAH data grid and the GT4 computational grid are pre-existing systems that are united through a service.

2. Overview of the SAFORAH System

SAFORAH was originally created to coordinate and streamline the archiving and sharing of large geospatial data sets between various research groups within the Canadian Forest Service, and various other academic and government partners within a secure framework. SAFORAH makes optimum use of distributed data storage to enhance collaboration, streamline research workflows and increase the return on the Canadian government's investment in Earth Observation (EO) data by freeing researchers from focusing on data storage and distribution issues.

The current SAFORAH data grid provides significant benefits to the participating organizations and partners who work on the large national projects to monitor and report on the state of Canada's forests as required by Parliament and international agreements. Land cover mapping is often the primary source for determining current status of an area and is used as a baseline for future changes. Increasingly, remote sensing is being turned to as a timely and accurate data source for land cover mapping. Many provincial and territorial mapping agencies have recognized the importance of Landsat remotely-sensed land cover mapping. They have started utilizing remote sensing information products in SAFORAH to improve their decision-making processes on sustainable development and environmental assessment.

SAFORAH presents the user with a web portal to the data distributed in the various storage facilities across multiple organizations and government agencies. Users, depending upon their access privileges, can download or upload any predefined EO imagery from anywhere on the data grid. Currently, four Canadian Forest Service centres (Victoria, Cornerbrook, Edmonton, and Québec City), Environment Canada's Canadian Wildlife Service in Ottawa, the storage facility at University of Victoria and the University of Victoria Geography Department are operationally connected to the SAFORAH data grid. Two sites at the Canadian Space Agency (CSA) and the University of Victoria's Computer Science Department are also linked to SAFORAH. The metadata is stored in either the Catalogue and User Data Ordering System (CUDOS) or the Open Geospatial Consortium (OGC) Catalogue Service for Web (CS/W) for catalogue, data archiving and ordering.

SAFORAH requires new computing and storage resources to handle both the growing demands for remote sensing data from hyperspectral, multispectral and radar satellites and the demands for processing capabilities for derived information products. Grid computing and high-speed networking offers a solution to these challenges by allowing grid users to access not only heterogeneous storage facilities but also processing resources distributed across administrative domains.

SAFORAH offers two main components for users to interact with the data grid shown in Figure 2: the Catalogue and User Data Ordering System (CUDOS) from MacDonald, Dettwiler and Associates (MDA), and the more recent web-services interface based on the OGC standards and specifications. The main difference between the CUDOS catalogue and the OGC service is that CUDOS allows users to access a concatenated granule in a zipped

format containing the complete image and associated files. OGC allows users to access EO data, stored as individual channels or bands for each EO image, such as a 400-band hyperspectral image. OGC provides users with a simple HTTP interface for requesting registered map images or accessing geospatial coverage objects. The two systems provide access to complementary data sets. The OGC interface can interoperate with other OGC compatible geospatial information systems via CUDOS.

CUDOS, which conforms to the US metadata standard established by the Federal Geographic Data Committee (FGDC), uses pre-Web-service components of Globus Toolkit 4 (GT4) and has its own security mechanisms for user authentication and access controls to SAFORAH. The connection between CUDOS and the SAFORAH data grid is authenticated by using Grid Canada credentials.

The OGC interface uses GT4 web services. The implementation enhances information interchange with other geospatial information systems in the Canadian Geospatial Data Infrastructure (CGDI), such as the CFS National Forest Information System (NFIS). The services include the Catalogue Service, Coverage Service, Web Map Service, and other supporting grid services. Other OGC clients or GIS-based information systems may also gain access to these grid-enabled services to obtain EO data through the standard OGC Web service portals.

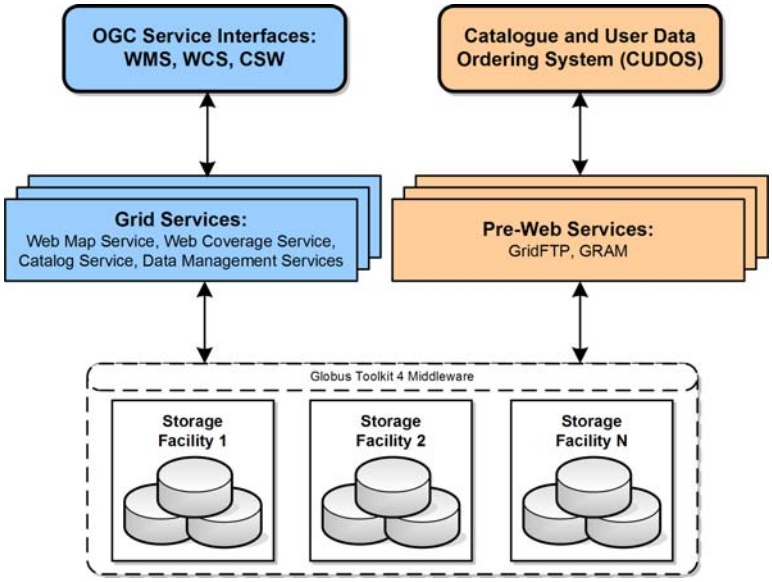

Fig. 2. Architecture of the SAFORAH system showing the two components CUDOS and the OGC interface. CUDOS uses pre-WS components while the OGC interface utilizes the WS components in the Globus Toolkit V4.

3. Overview of the Computational Grid Service

The OGC interface of SAFORAH is built using the Web Services Resource Framework (WSRF) standard, which is the same standard used by GT4. It was therefore a logical choice to integrate it with a GT4 computational grid for SAFORAH's computation requirements.

Members of the project team have constructed a small GT4 computational grid running on a number of Canadian computing resources. The components of a GT4 grid are shown in Figure 3. The two key components are the metascheduler and registry service. The role of the metascheduler is to broker access to the distributed computational facilities, which schedule jobs to individual worker nodes: the metascheduler acts as a scheduler of schedulers. As a central grid service, the metascheduler provides the functionality to submit, monitor, and manage jobs across the facilities. The metascheduler discovers available resources by querying the registry. The registry is a second grid service which stores and accesses resource advertisements from the computing facilities. The type and number of resources that are available are included in each advertisement. Additional services to manage data and security are also components of the grid.

The GT4 computational grid was designed in a modular fashion to allow it to use components from other projects. This group has previously developed metascheduler and registry services based on Condor-G. The metascheduler is called Gavia, and is a Globus incubator project [5].

Gavia is built using WSRF, a product of the Open Grid Forum, and is implemented using GT4. The WSRF includes standards for job management and for resource registry: Web Services Grid Resource Allocation and Management (WS-GRAM) is a standard which was created to allow the management of job execution on grid resources, and Web Services Monitoring and Discovery System (WS-MDS), a standard for the advertisement and discovery of grid resources. Since WS-MDS is implemented as part of GT4, the Gavia registry is simply a deployment of this technology. However, during the initial development, there was no mature and stable WS-GRAM-enabled metascheduler. As such, a metascheduler was developed using Condor-G at the core for making allocation decisions. The Gavia metascheduler was constructed by developing a WS-GRAM interface to Condor-G, as well as a mechanism to discover available resources in the registry. This system allows users or automated clients to submit jobs to a single service, which in turn performs an allocation decision and resubmits the job to the selected facility. By providing transparent access to many geographically-distributed facilities, users are able to focus on their applications and research, rather than having the burden of managing their executions at many facilities at once.

There is an alternative metascheduler developed by the GridWay project [6]. This project's computational grid can operate using either the Gavia or GridWay metascheduler.

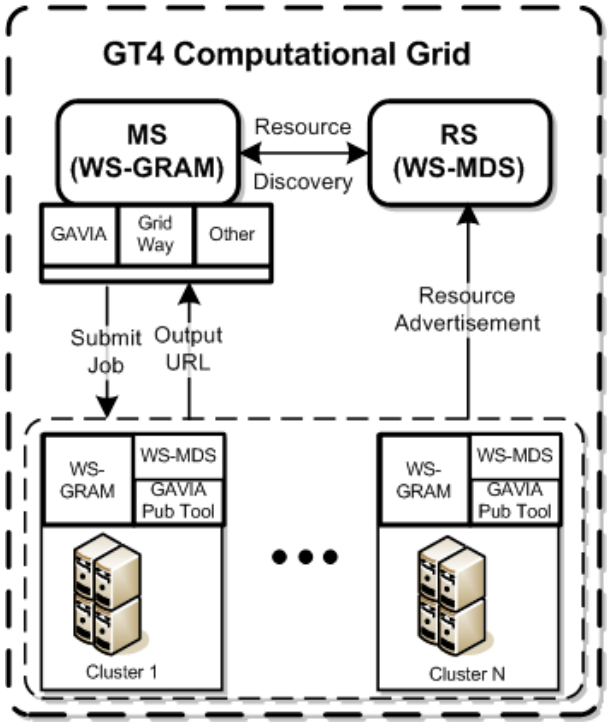

Fig. 3. An overview of the components of a GT4 computational grid showing the metascheduler service (MS) and the registry service (RS). Each cluster publishes its status to the registry, which is used by the metascheduler to schedule jobs.

4. SAFORAH Computational Grid Service

Fig. 4 shows the integrated SAFORAH and computational grid systems. The existing SAFORAH system is shown on the left, and the GT4 computational grid on the right. A new component, the Integration Service (IS), is shown in the center. The purpose of this service is to coordinate the interaction between the existing data grid and the new computational grid. The decisions made by the Integration Service are listed within the red box.

We give a brief overview of how the system will respond to a user request: The user submits a request via the SAFORAH web portal. If the request is for existing data, the system then emails the user a link to the data. If the request is for computed data, the request is sent to the IS, which then locates the input data, generates the necessary scripts, creates the job description and sends the job to the metascheduler. The metascheduler finds the appropriate resource after querying the registry service and submits the job to one of the clusters, which then internally submits it to one of its worker nodes. After the job is completed, the data is written to a SAFORAH node and a link is returned to the user via the Integration Service or email.

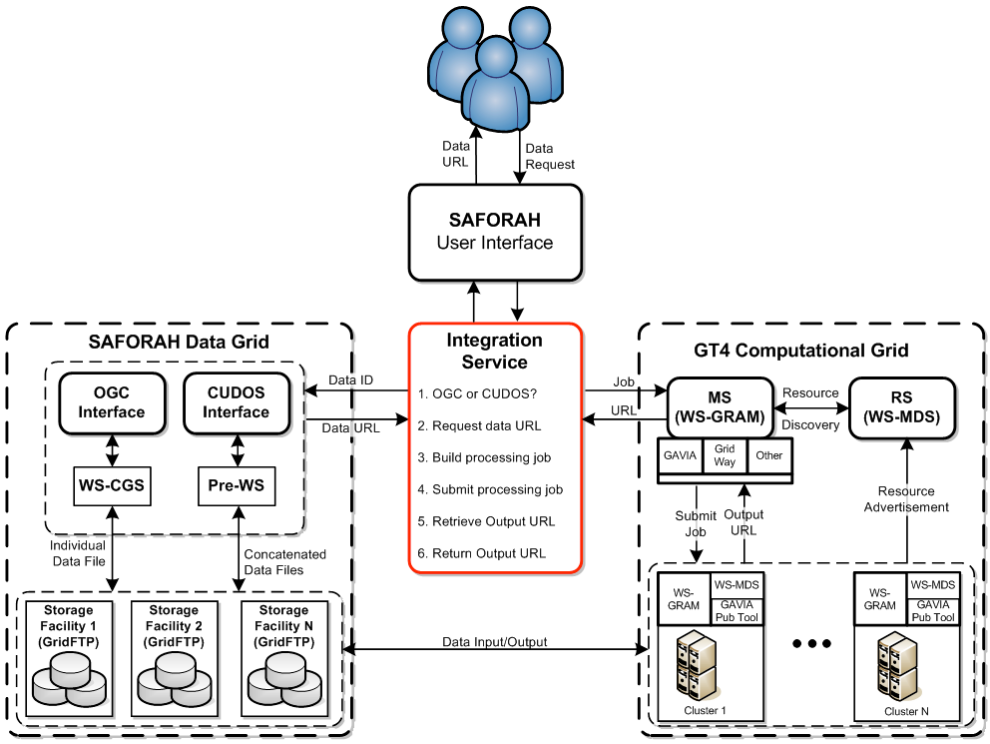

Fig. 4. A detailed figure showing the integration of the SAFORAH data grid with a GT4 computational grid. The researcher access the system via SAFORAH user interface and the Integration Service (IS) retrieves the data location from SAFORAH, runs the requested application on the grid and returns the links for the output to the researcher.

5. Status of the Project

The project is ending its first year of development as part of a two-year project. A beta prototype for the system has been implemented and tested. Currently, authorized users have access to two virtual products: a hyperspectral calibration product (force-fit) and Shane Cloude RADAR decomposition method (SCM). For more information about these two products, the reader may refer to [7] and [8, 9], respectively. The force-fit product is a very lightweight product and takes only several minutes to run. It allows us to test the system and check its state quickly. SCM takes much longer to complete (around one-half of an hour for a 300MB input data) and allows the users to generate a real forestry product. Fig. 5 shows an example of the virtual product computed using SCM.

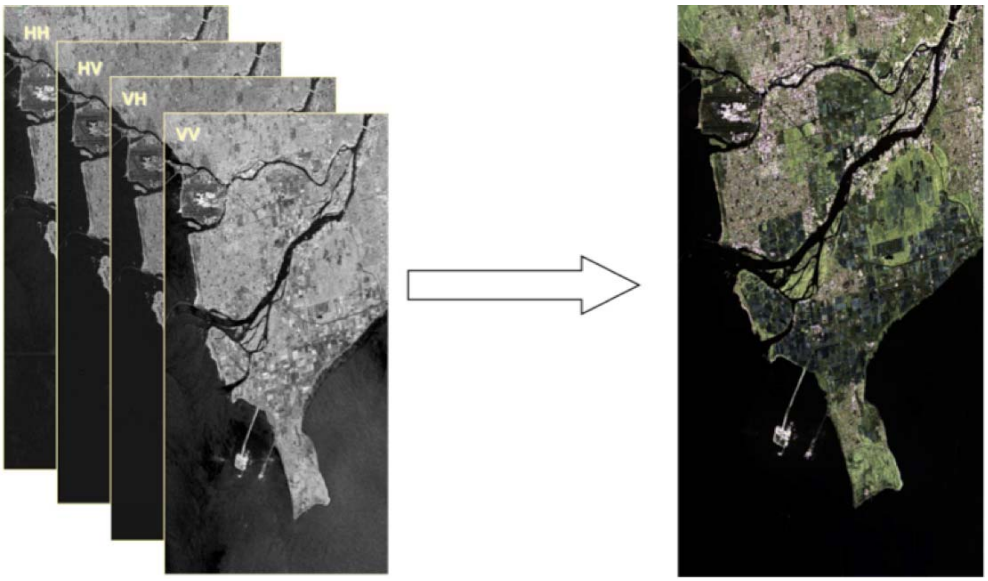

Fig. 5. SCM takes four polarization channels of Radarsat-2 FQ15 data, and produces an HSV land cover product image. The image shown highlights the Vancouver area.

6. Summary

The integration of SAFORAH with a GT4 computational grid has been designed and implemented in a development platform. The goal is to release a production level service within the next 12 months. We have highlighted the use of the system with a two examples. Work is ongoing to enable the system to generate other forestry virtual products.

We thank Natural Resources Canada, the University of Victoria, National Research Council Canada, Natural Sciences and Engineering Research Council of Canada, CANARIE Inc. and the Canadian Space Agency for their financial support. We are very grateful for the technical support of the Center for Spatial Information Science and Systems of George Mason University for the grid SOA development. We appreciate the high-bandwidth connectivity provided by Shared Services BC and the Radarsat-2 data provided by MDA.

References

[1] D. G. Goodenough, H. Chen, L. Di, A. Guan, Y. Wei, A. Dyk, and G. Hobart, “Grid-enabled OGC Environment for EO Data and Services in Support of Canada's Forest Applications”, Proceedings of IGARSS 2007. IEEE International. July 2007, 4773.

[2] The Globus Alliance, "The Globus Toolkit", http://www.globus.org.

[3] Condor – high throughput computing. http://www.cs.wisc.edu/condor.

[4] A. Agarwal, et al. “GridX1: A Canadian computational grid. Future Generation”, Computer Systems, Volume 23, Issue 5, June 2007, Pages 680-687.

[5] The Gavia Project. http://dev.globus.org/wiki/Incubator/Gavia-MS

[6] GridWay Metascheduler - Metascheduling Technologies for the Grid. http://www.gridway.org/doku.php

[7] D.G. Goodenough, A. Dyk, G. Hobart, and H. Chen, “Forest Information Products from Hyperspectral Data - Victoria and Hoquiam Test Sites”, In Proc. IGARSS 2007, pp. 1532 - 1536, Barcelona, Spain.

[8] S.R. Cloude, “Radar target decomposition theorems”, Inst. Elect. Eng. Electron. Lett., vol. 21, no. 1, pp. 22-24, Jan. 1985.

[9] S.R. Cloude, E. Chen, Z. Li, X. Tian, Y. Pang, S. Li, E. Pottier, L. Ferro-Famil, M. Neumann, W. Hong, F. Cao, Y. P. Wang, K. P. Papathanassiou, “Forest Structure Estimation Using Space Borne Polarimetric Radar: An ALOS-PALSAR Case Study”, ESA Dragon Project, 2007.