Contents

Overview of caGrid

Collaborative biomedical research studies, which involve participation of multiple institutions and require integration of disparate, heterogeneous data and analytical resources, have long been hindered by paucity of interoperable resources and lack of systems to link them. As a result, rich collections of distributed information resources and complementary expertise of research groups are underutilized in most biomedical research domains. Recognizing these obstacles and problems in cancer research, the National Cancer Institute (NCI) has established the cancer Biomedical Informatics Grid (caBIG?) program -- for an overview of the caBIG? initiative, please visit http://cabig.cancer.gov. The objective of this program is to develop enabling informatics technologies for collaborative, multi-institutional biomedical research and to create a voluntary network of cancer centers and research laboratories with the overarching goal of accelerating translational cancer research.

The caGrid infrastructure is an integral component of caBIG?. The Grid technologies and methodologies adopted for caBIG?, and implemented in caGrid, provide a loosely coupled environment wherein local providers are given freedom of implementation choices and ultimate control over access and management, but harmonize on community accepted virtualizations of the data they use, and make them available using standardized service interfaces and communication mechanisms.

While caGrid enables numerous complex usage scenarios, in its simplest base, its goals are to: enable universal mechanisms for providing interoperable programmatic access to data and analytics to caBIG?, create a self-describe infrastructure wherein the structure and semantics of data can be programmatically determined, provide the core infrastructure for federating data and analytical resources and applications deployed at different institutions within the guidelines and policies accepted by the caBIG? community, and implement mechanisms for enabling researchers to discover, query, integrate, and synthesize information from existing distributed resources as well as contribute new resources and applications to the caBIG? environment.

The main application community motivating the need for caGrid is the cancer research community. However, the infrastructure has been designed and implemented as a general middleware system which can support other biomedical application domains. caGrid is built as a service-oriented architecture on top of Grid Service technologies; more specifically, on the Web Services Resource Framework (WSRF) standards. It also draws from the model driven architecture paradigm to enable syntactic and semantic interoperability among resources and rich metadata driven discovery and query of distributed resources. To this end, it leverages and supports the concepts of controlled vocabularies, strongly-typed services, common data elements, published information models, and rich service metadata. In essence, caGrid combines Grid computing, service oriented

architecture, and model driven architecture in an integrated framework.

caGrid has been released to the community through several versions. Starting with version 1.0, the caGrid infrastructure has been deployed in the caBIG? environment for production use and as an Enterprise Grid middleware system. The latest version is caGrid 1.2, released in March 2008. The caGrid 1.2 release represents a backwards compatible release of caGrid, with a focus on increased usability, bug fixes, and various feature enhancements. caGrid development is an ongoing effort, improving and adding new functionality based on feedback from caGrid users and new requirements from the community.

caGrid Architecture

caGrid is a service-oriented Grid software infrastructure. It leverages Grid Services technologies and Grid systems, including the Globus Toolkit and Mobius, and tools developed by the NCI such as the caCORE infrastructure. Each data and analytical resource in caGrid is implemented as a Grid Service, which interacts with other resources and clients using Grid Service protocols. caGrid services are standard WSRF (version 1.2) services and can be accessed by any specification-compliant client. The caGrid infrastructure also consists of coordination services, runtime environment to support the deployment, execution, and invocation of data and analytical services, and tools for easier development of services, management of security, and composition of services into workflows.

A salient characteristic of caGrid, which differentiates it from other Grid middleware systems, is the focus on syntactic and semantic interoperability, driven by the guidelines and requirements developed by the caBIG? community. caGrid adopts a model-driven architecture approach. Client and service APIs in caGrid represent an object-oriented view of data and analytical resources. These APIs operate on registered domain models, expressed as object classes and relationships between the classes. caGrid leverages existing NCI data modeling infrastructure to manage, curate, and employ the data models. Domain models are defined in Unified Modeling Language (UML) and converted into common data elements. These common data elements are in turn registered in the Cancer Data Standards Repository (caDSR). The definitions of these data elements draw from vocabulary registered in the Enterprise Vocabulary Services (EVS). The concepts of data elements and the relationships among the data elements thus are semantically described.

Clients and services communicate through the Grid using messages encoded in XML. In caGrid, when an object is transferred over the Grid between clients and services, it is serialized into a XML document that adheres to a XML schema registered in the Mobius Global Model Exchange (GME) service[7]. As the caDSR and EVS define the properties, relationships, and semantics of caBIG? data types, the GME defines the syntax of their XML materialization.

caGrid provides extensive support for researchers to discover distributed resources by taking advantage of rich structural and semantic descriptions of data models and services. Each caGrid service is required to provide service metadata. The base metadata contains information about the service-providing cancer center, such as the point of contact and the institution?s name providing the service. It also describes the objects used as input and output of the service?s operations. The definitions of the objects themselves are described in terms of their underlying concepts, attributes, attribute value domains, and associations to other objects being exposed as extracted from the caDSR. In addition, the service metadata specifies the operations or methods the service provides, and allows semantic concepts extracted from the EVS to be applied to them. This base metadata is extended for different types of services. For instance, data services comply with an additional ?domain model? metadata standard, which details the domain model, including associations and inheritance information, from which the objects being exposed by the service are drawn. When a service is deployed, its

service metadata is registered with an indexing registry service, called the Index Service. A researcher can discover services of interest by looking them up in this registry. caGrid provides a series of high-level APIs for performing searches on these metadata standards. For instance, a client can search for analytical services that provide operations that take a data type representing a given concept as input.

Security is essential for successful deployment of the caGrid environment because of the need to protect intellectual property of researchers and to ensure protection and privacy of patient related information. A comprehensive set of services are provided in caGrid to support secure and controlled access to resources based on policies set forth by the owners of the resources. These services enable Grid-wide management of user credentials, support for grouping of users into virtual organizations for role based access control, and management of trust fabric in the Grid.

caGrid Use Cases and Pattern Templates

caGrid is an e-Science infrastructure that builds on a model driven and service oriented architecture foundation. The key architecture features of caGrid are based on a wide range of biomedical research pattern templates and use cases. Pattern templates describe the common characteristics and components of research projects and motivate the design of biomedical informatics tools. The following paper presents examples of pattern templates arising in translational research projects. It describes how caGrid is motivated by the requirements of biomedical research and illustrates how caGrid along with another e-Science tool, called caIntegrator, can be used to implement the translational research templates.

- e-Science, caGrid, and Translational Biomedical Research, Technical Report, No. OSUBMI_TR_2008_n01, Biomedical Informatics, 2008.

Papers on caGrid

The following papers provide an overview of the caGrid infrastructure, the GAARDS security infrastructure, and the Introduce toolkit -- if a link does not work (e.g., because of journal access restrictions), please contact us:

caGrid Architecture Design

- J. Saltz, S. Oster, S. Hastings, S. Langella, T. Kurc, W. Sanchez, M. Kher, A. Manisundaram, K. Shanbhag, and P. Covitz, "caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid", Bioinformatics 2006 22(15):1910-1916. A paper presenting an overview of version 0.5 of caGrid and the design principles behind the caGrid infrastructure.

- S. Oster, S. Hastings, S. Langella, D. Ervin, R. Madduri, T. Kurc, F. Siebenlist, I. Foster, K. Shanbhag, P. Covitz, and J. Saltz, "caGrid 1.0: A Grid Enterprise Architecture for Cancer Research", the 2007 AMIA Annual Symposium, November, 2007. A short overview of caGrid 1.0.

- S. Oster, S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. Covitz, K. Shanbhag, I. Foster, and J. Saltz, "caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research", J Am Med Inform Assoc. 2008;15:138-149. This paper presents a longer overview of caGrid 1.0 and its core components.

Pattern Templates for Translational Biomedical Research

- J. H. Saltz, S. L. Hastings, S. Langella, S. Oster, D. W. Ervin, A. Sharma, T. C. Pan, M. N. Gurcan, S. Madhavan, R. Madduri, E. Caserta, J. D. Permar, R. A. Ferreira, P. R. Payne, U. V. Catalyurek, G. Leone, M. Ostrowski, K. H. Buetow, K. Shanbhag, E. Siegel, I. Foster, and T. M. Kurc, "e-Science, caGrid, and Translational Biomedical Research", Technical Report, the Department of Biomedical Informatics, The Ohio State University, 2008. This paper presents pattern templates for using caGrid and caIntegrator for Translational Biomedical Research.

Introduce

- S. Hastings, S. Oster, S. Langella, D. Ervin, T. Kurc, and J. Saltz, "Introduce: An Open Source Toolkit for Rapid Development of Strongly Typed Grid Services", Journal of Grid Computing, Vol.5, No.4, pp. 407-427, March, 2007. A paper describing the architecture and design of the Introduce toolkit.

caGrid Security

- S. Langella, S. Oster, S. L. Hastings, F. Siebenlist, J. Phillips, D. W. Ervin, J. D. Permar, T. M. Kurc, and J. H. Saltz, "The Cancer Biomedical Informatics Grid (caBIG) Security Infrastructure", the 2007 AMIA Annual Symposium, November, 2007. A short overview of the architecture of the caGrid security infrastructure, GAARDS.

- S. Langella, S. Hastings, S. Oster, T. Pan, A. Sharma, J. Permar, D. Ervin, B. Cambazoglu, T. Kurc, and J. Saltz, "Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment", J Am Med Inform Assoc. 2008;15:363-373. A longer overview of the GAARDS infrastructure with application example.

- S. Langella and K. Modi. GAARDS Security Infrastructure Updates. Technical Report, the Department of Biomedical Informatics, The Ohio State University, 2008. An overview of recent updates to the GAARDS infrastucture.