Teledetección: Agua y desarrollo sostenible. XIII Congreso de la Asociación Española de Teledetección.
Calatayud, 23-26 de septiembre de 2009. pp. 457-460. Editores: Salomón Montesinos Aranda y Lara Fernández Fornos

DESARROLLO E IMPLEMENTACIÓN DE UN LABORATORIO VIRTUAL PARA LA TELEDETECCIÓN OCEANOGRÁFICA BASADO EN GRID

D. Mera (*), J. M. Cotos (*), C. Cotelo (**), Y. Sagarminaga (***), J. Pérez (****).

(*) Instituto de investigaciones Tecnológicas. Universidad de Santiago de Compostela. Campus Universitario Sur. 15782 Santiago de Compostela, España. david.mera@usc.es.
(**) Centro de Supercomputación de Galicia. Avda de Vigo, s/n Campus Sur. 15705 Santiago de Compostela, A Coruña, España.
(***) Fundación Azti. Txatxarramendi Ugartea z/g 48395 Sukarrieta (Bizkaia), España.
(****) Instituto Canario de Ciencias del Mar. Carretera de Taliarte, s/n. 35200 Telde - GranCanaria – Islas Canarias, España.

RESUMEN
En este artículo presentamos el desarrollo e implementación de un laboratorio virtual para la comunidad oceanográfica llamado Retelab. El principal objetivo de este proyecto es desarrollar una herramienta útil y de fácil acceso para los investigadores donde los conocimientos informáticos no sean necesarios. Retelab es un sistema basado en tecnología Grid, lo cual proporciona una potencia de cálculo y de almacenamiento que no es posible conseguir utilizando tecnologías clásicas. El desarrollo de Retelab tiene cinco fases. La primera es la de gestión de usuarios, donde se ha buscado una solución equilibrada entre seguridad y facilidad de uso. El despliegue de una base de datos virtual y distribuida para la gestión y almacenamiento masivo de datos es acometido durante la segunda fase. La integración de herramientas para la visualización y análisis de los datos almacenados es el núcleo de la tercera fase. La siguiente fase es la del desarrollo de un sistema automático para el envío de trabajos al Grid y para finalizar, la última fase, consiste en el despliegue de aplicaciones oceanográficas con el objetivo de integrarlas en el sistema para probarlo y mejorarlo.

ABSTRACT
We propose in this paper the development of a virtual laboratory for the Ocean research community called Retelab. The project main objective is to build a useful tool for the researchers where the computer skills are not necessary. Retelab is based on Grid technology. This technology provides more storage and computational power than classic technologies. Retelab development has several phases. The first phase is the user access and registration phase where we have develop a well balanced solution between simplicity and security. The deployment of a distributed virtual data base to management of large amount of datasets is faced on the second phase. Different tools for analyzing and visualizing the data stored are integrated in the Retelab project during the third phase. Next phase is the development of automatic tool for job submission. Finally, the last phase is the deployment of testbeds to test and improve the system.

Palabras clave: Grid, base de datos virtual, Data Grid, metadatos.

INTRODUCCIÓN

El lanzamiento de nuevas misiones para la observación de la tierra se esta incrementando cada año. Durante el año 2009 la Agencia Espacial Europea enviará 3 satélites para la exploración terrestre y otros tres están en periodo de construcción. La gran cantidad de datos generados por estas misiones hacen que su análisis y almacenamiento sea una tarea ardua y tediosa y en muchos casos inabarcable. Concretamente, el estudio del océano es una tarea interdisciplinar donde biólogos, físicos, meteorólogos, oceanógrafos e informáticos necesitan trabajar juntos para procesar la información de una forma eficiente.

La tecnología Grid fue desarrollada con el objetivo de permitir computación colaborativa compartiendo recursos de hardware (clusters, PCs, sensores…), software y conjuntos de datos a través de una red (por ejemplo Internet). Un sistema Grid nos aporta posibilidades computacionales y de almacenamiento mucho mayores que las tecnologías clásicas pero su uso es complejo y requiere ciertos conocimientos informáticos.

Este trabajo presenta el desarrollo e implementación de un laboratorio virtual para proyectos multidisciplinares relacionados con la teledetección oceanográfica llamado Retelab (Sagarminaga et al 2007). Con este proyecto

pretendemos acercar los beneficios de la tecnología Grid a la comunidad oceanográfica a través de un sistema sencillo, útil y que no necesite excesivos conocimientos informáticos para usarse.

DESCRIPCIÓN DEL PROYECTO

Retelab ha sido estructurado en diferentes fases y todas ellas han seguido un desarrollo centrado en la simplicidad y facilidad de uso. Las fases principales del proyecto se detallan a continuación.

Gestión de Usuarios

Un sistema Grid comparte recursos de diversas organizaciones y esto genera un problema de seguridad. Actualmente, uno de los sistemas más extendidos para identificar usuarios es la Infraestructura de Clave Pública (PKI). La Infraestructura de Seguridad Grid (GSI) es una piedra angular en el desarrollo de cualquier Grid y está basada en la PKI. La forma de acceder a un sistema basado en PKI es poseyendo un certificado digital (DC). A pesar de que hoy en día términos como certificado digital y clave pública empiezan a ser conocidos, la obtención de un DC y su uso no es algo trivial para la mayoría de la gente. Hemos integrado diversas tecnologías para desarrollar un método compatible con la GSI y que fuese cómodo de utilizar por los usuarios de Retelab.

Una vez solucionada la identificación de los usuarios, necesitábamos un subsistema para administrar las políticas de seguridad y que resolviese cuestiones como. ¿Quién puede acceder al sistema?, ¿Con qué recursos puede trabajar? o ¿Qué acciones puede ejecutar sobre un recurso concreto? Finalmente, se decidió que un sistema de Control de Acceso Basado en Roles (RBAC) era el que mejor se adaptaba a nuestros requerimientos.

Capacidad de Almacenamiento

Actualmente, proyectos como el Worldwide LHC Computing Grid (Shiers 2007) generan tal cantidad de datos que su análisis y almacenamiento es una tarea prácticamente imposible.

Una de las aportaciones mayores que Retelab puede dar a la comunidad Oceanográfica es la de proporcionar un almacén de datos oceanográficos que se pueda utilizar, analizar y alimentar con nueva información. El sistema de almacenamiento de Retelab es el de una Base de Datos Virtual Distribuida (BDVD) que en terminología Grid denominaríamos Data Grid.

Esta BDVD permitirá a los usuarios analizar, trabajar y buscar en grandes conjuntos de datos. Para facilitar el uso de esta BDVD, Retelab usa los metadatos como medio para identificar los recursos almacenados.

Visualización y Análisis de Datos

Esta fase está íntimamente ligada con el sistema de almacenamiento y búsqueda. Una vez que un usuario realiza una búsqueda y obtiene unos resultados, el sistema posee medios para visualizar y analizar los datos online y de esta forma comprobar si los datos son útiles para usarlos como parámetros en sus trabajos Grid.

Capacidad de Procesado

Existen muchos sistemas que permiten enviar trabajos a un sistema Grid pero necesitan interacción con el usuario final lo que los hace menos simples y transparentes. Retelab se ha desarrollado pensando en usuarios con pocos conocimientos informáticos y por ello se ha creado un procedimiento que automatiza el proceso de envío de trabajos seleccionando las opciones óptimas en cada momento.

Aplicaciones de Prueba

Una vez se hayan finalizado el resto de las fases del proyecto, integraremos diversas aplicaciones de prueba para probar y mejorar el sistema.

DISEÑO E IMPLEMENTACIÓN

La interfaz de usuario de Retelab ha sido desarrollada como un portal Web basado en portlets. Los portlets son componentes modulares de interfaz de usuario gestionados y visualizados a través de un portal Web. Los portlets son muy útiles dentro de los sistemas Grid porque permiten una alta interoperatibilidad y reutilización. Se pueden desarrollar y añadir a un portal bajo demanda.

Cada fase del proyecto fue finalizada con el desarrollo de un prototipo. Los siguientes subapartados muestran brevemente su arquitectura.

Gestión de Usuarios

La gestión de usuarios de Retelab incluye su registro, control y política de acceso. La Figura 1 muestra los diferentes módulos y tecnologías usadas en la implementación de este prototipo.

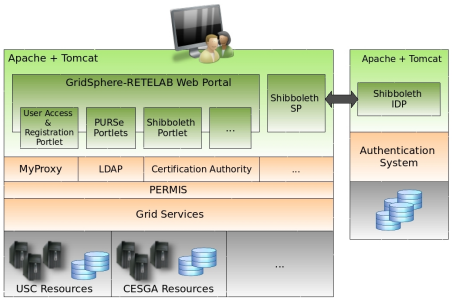

Figura 1. - Gestión de Usuarios.

La gestión de usuarios puede verse desde 3 escenarios diferentes. El primero de ellos sería el registro de usuarios. El registro se realiza a través de los portlets de registro. Estos portlets están basados en una versión del sistema PURSe (Foster et al. 2006) que hemos mejorado para incluirle soporte para roles. Cada vez que un usuario se registra se envía una notificación al administrador. Una vez que el administrador comprueba la petición, el sistema genera automáticamente un DC para ese usuario, evitando de esta forma la tediosa tarea de adquirirlo, y además se crea un Certificado de Atributos (AC) que almacena el rol del usuario y que será usado dentro del sistema RBAC para gestionar su acceso a Retelab. El almacenamiento de las credenciales de usuario se realiza en un repositorio online.

El segundo escenario es el del acceso al sistema de un usuario registrado. Para acceder es necesario introducir el nombre de usuario y la contraseña proporcionada durante el registro. Una vez autentificado, el sistema recuperará el AC y el DC del usuario y, dependiendo de su rol, le aplicará una política de seguridad u otra. PERMIS (Chadwick and Otenko 2003) es el software utilizado para integrar los roles dentro del Grid.

El último escenario describe el acceso de los usuarios no registrados. Retelab tiene diferentes socios participantes. Es lógico pensar que si se confía en esos socios se debería dejar acceder a sus usuarios sin que necesitasen registrarse en el portal. Shibboleth es un software que permite crear una red de confianza entre diferentes centros, de forma que, con un único proceso de autentificación, se permite acceder a los diferentes recursos de cada centro. Cada usuario se autentifica en su centro y una vez validado, Shibboleth comparte atributos del usuario (nombre, correo, rol, etc.) con el resto de los centros. De esta forma cada centro tiene información del usuario para decidir si permite su acceso o no. En Retelab se ha integrado Shibboleth dentro del proyecto quedando el proceso de la siguiente manera: Cuando un usuario de confianza quiere acceder al portal, usará el portlet de Shibboleth. Este portlet le mostrará los centros en los que Retelab confía para que el usuario seleccione el suyo. Una vez seleccionado el centro, se redirigirá al usuario al sistema de autentificación del centro escogido. Una vez validado, el centro enviará a Retelab los atributos del usuario y se permitirá o denegará el acceso dependiendo de esa información. Si el usuario es aceptado, el sistema generará un DC y un CA utilizando sus atributos.

Capacidad de Almacenamiento

La arquitectura utilizada para el desarrollo del prototipo de la BDVD está dividida en tres capas (Figura 2). La capa superior es la de interfaz de usuario. La interfaz de acceso a la BDVD es un portlet desplegado en el portal. En la capa intermedia o capa de middleware se encuentra el núcleo del sistema. Cada host que forma el Data Grid tiene instalado un subsistema llamado “Locate Replica Catalog” (LRC) que almacena una lista de pares que contienen la ruta de un dato físico alojado en ese host y el nombre lógico para ese dato Cada LRC informa de los datos que contiene al subsistema Replica Location Index (RLI) instalado en el servidor central que mantiene una lista con los nombre lógicos de los datos y el LRC que los almacena. Para finalizar, el servidor central posee un catalogo de metadatos con una lista de atributos asociada a cada nombre lógico. Con esta estructura es posible localizar un dato físico a través de sus metadatos. En la capa inferior o capa de recursos se encuentran los recursos físicos de Retelab (servidores, clusters, etc.). Cada organización mantiene su propia configuración de sus recursos ya que un Grid permite unir recursos heterogéneos.

Visualización y Análisis de Datos

La Figura 2 muestra, en la capa de interfaz, dos subsistemas para la visualización y análisis de datos online. El Live Access Server (LAS), accesible a través de un portlet y el Unidata Integrated Data viewer (IDV). Para la utilización del IDV se ha desarrollado un procedimiento que genera un enlace por cada elemento recuperado tras una búsqueda sobre la BDVD. Este enlace descarga e instala temporal y automáticamente, en el ordenador del cliente, el IDV. Una vez instalada, la aplicación se ejecuta con los datos asociados al enlace.

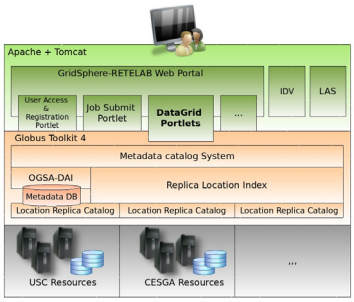

Figura 2.- Data Grid.

Capacidad de Procesado

Se ha desarrollado un prototipo para gestionar el envío de trabajos al Grid (Figura 3). En un Grid típico el usuario es el encargado de parametrizar los envíos mientras que en Retelab se ha integrado el metaplanificador GridWay para automatizar el proceso. GridWay, selecciona en cada momento las mejores opciones disponibles para ejecutar el trabajo y lo envía al recurso del Grid óptimo. En todo momento el usuario podrá ver el estado de su trabajo y al finalizar podrá acceder a los datos resultantes a través de un enlace.

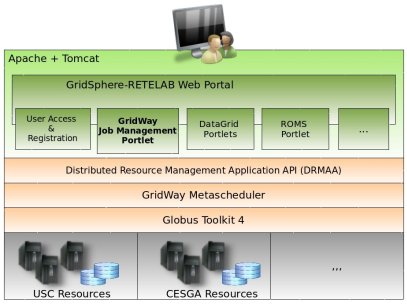

Figura 3.- Sistema de Envío de Trabajos.

CONCLUSIONES Y TRABAJOS FUTUROS

En este artículo hemos presentado el desarrollo de un sistema Grid, centrado en el estudio Oceanográfico, llamado Retelab. Se ha implementado un sistema de fácil acceso y uso sencillo pero sin sacrificar la seguridad. Se han desarrollado prototipos para la gestión de usuarios, almacenamiento y búsqueda de datos y para el procesado de tareas, consiguiendo lo que esperemos sea, un sistema útil para la investigación. El desarrollo del proyecto ha sido posible gracias al trabajo de socios tan heterogéneos como la Universidad de Santiago, el CESGA, el Instituto Canario de Ciencias Marinas o el AZTI.

Retelab no está todavía accesible. La siguiente fase en su desarrollo será la de pruebas y mejoras en la que desplegaremos algunos "testbeds" en el sistema. El primero será ROMS (Shchepetkin and McWilliams 2005). ROMS es un proyecto utilizado para simular sistemas oceánicos regionales. ROMS es un software ya desarrollado que queremos integrar y mejorar. Actualmente, puede ejecutarse usando los métodos de paralelización de MPI o de OpenMP. Nuestro objetivo es desarrollar un método híbrido (openMP + MPI) para ejecutarlo sobre Retelab.

El proyecto Sentinazos que se encuentra en una fase muy temprana de desarrollo, será el segundo testbed. El objetivo de esta aplicación será detectar de forma automática vertidos de fuel en el océano utilizando imágenes de satélite. Para conseguirlo se aplicarán sobre las imágenes diferentes algoritmos basados en procesado de imágenes, redes neuronales o lógica fuzzy.

BIBLIOGRAFÍA

Chadwick, W.D. and Otenko, A. 2003. The PERMIS X.509 role based privilege management infrastructure. *Future Generation Computer Systems*, v.19:2, pp. 277-289.

Foster, I., Nefedova, V., Ahsant, M., Ananthakrishnan, R., Liming, L., Madduri, R., Mulmo, O., Pearlman, L. and Siebenlist, F. 2006. Streamlining Grid operations: Definition and deployment of a portal-based user registration service. *Journal of Grid Computing*. v. 4(2), pp. 135–144.

Sagarminaga Y., Pérez J., López I., Cotos J.M. 2007. Retelab: Desarrollo de un laboratorio Virtual. *XII Congreso Nacional de Teledetección*, pp. 391-398.

Shchepetkin, A.F. and McWilliams J.C. 2005. The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, v. 9, issue 4, pp. 347-404.

Shiers, J. 2007. The Worldwide LHC Computing Grid (worldwide LCG), *Computer Physics Communications*, v. 177, Issues 1-2, pp. 219-223.

AGRADECIMIENTOS

Este proyecto está siendo financiado por el Ministerio de Educación y Ciencia (ESP2006-13778-C04).