Documento Descriptivo Sobre las Áreas Temáticas de e-Ciencia

Acción Especial IRISGRID

IFCA, IFIC, RedIRIS, CAB, CNB, CIEMAT, IMEDEA, IAA, UCM, GAC, CIRI, IFAE/PIC, AOSO, GRyCAP, UAM, UAL, UM, UAL, UNIOVI, UPM, UGR, INTA, ARCOS

Contenidos

1 Introducción

1.1 El Grid

Un Grid es un conjunto de recursos hardware y software distribuidos por Internet que proporcionan servicios accesibles por medio de un conjunto de protocolos e interfaces abiertos y estandarizados (gestión de recursos, gestión remota de procesos, librerías de comunicación, seguridad, soporte a monitorización…) y organizados por medio de unos procedimientos y guías de "buenas prácticas" bien definidas. Las organizaciones virtuales que se interconectan por medio de un Grid mantienen sus propias políticas de seguridad y gestión de recursos. La tecnología usada para construir un Grid es complementaria a otras tecnologías para aprovechar los recursos distribuidos en la intranet de una organización.

1.2 IRISGrid

IRISGrid es una iniciativa cofinanciada por el Ministerio de Ciencia y Tecnologia en la que participan casi 200 investigadores de 23 centros españoles en la investigación en el área de tecnologías Grid. IRISGrid pretende aportar los protocolos, procedimientos y guías de "buenas prácticas" necesarios para construir dentro de España un Grid de investigación coordinando a los diferentes Grupos y Centros interesados en investigación sobre tecnologías Grid. Esta iniciativa pretende unir recursos distribuidos geográficamente para que los Grupos involucrados tengan un banco de pruebas o "Test-bed" que soporte la investigación en cualquiera de las áreas de aplicación del Grid.

En la iniciativa IRISGrid participan los centros reflejados en la siguiente tabla

CODIGO	Centro	**URL**
IFCA	Instituto de Física de Cantabria, CSIC-UC, Santander	**grid.ifca.unican.es**
IFIC	Instituto de Física de Corpuscular, Centro Mixto CSIC-UV, Valencia	**alpha.ific.uv.es/grid**
RedIRIS	Centro de Comunicaciones CSIC-RedIRIS, Madrid-Sevilla	**www.rediris.es/**
CAB	Centro de Astrobiología, CSIC-INTA, Torrejón de Ardoz - Madrid	**www.cab.inta.es/~CABGrid**
CNB	Unidad de Biocomputación, Centro Nacional de Biotecnología, Madrid	**www.biocomp.cnb.uam.es**
CIEMAT	Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Madrid	**wwwae.ciemat.es/grid**
IMEDEA	Instituto Mediterráneo de Estudios Avanzados, CSIC-UIB, Palma de Mallorca – Islas Baleares	**www.imedea.uib.es**
IAA	Instituto de Astrofísica de Andalucía	**www.iaa.csic.es**
UCM	Grupo de Arquitectura de Sistemas Distribuidos y Seguridad, Dpt. Arquitectura de Computadores y Automática, Universidad Complutense de Madrid	**www.ac.dec.usc.es**
GAC	Departamento de Electrónica y Computación, Universidad de Santiago	**www.ac.dec.usc.es**
CIRI	CEPBA-IBM Research Institute, Universidad Politécnica de Cataluña, Barcelona	**www.ciri.upc.es**

IFAE/PIC	Instituto de Física de Altas Energías, Consorcio Generalitat Catalunya-Universitat Autònoma de Barcelona / División Port d'Informació Científica	
AOSO	Arquitectura de Ordenadores y Sistemas Operativos, Universidad Autónoma de Barcelona, Cerdanyola del Vallés, Barcelona	**www.caos.uab.es**
GRyCAP	Grupo de Redes y Computación de Altas Prestaciones, Dpt. Sistemas Informáticos y Computación, Universidad Politécnica de Valencia	**www.grycap.upv.es**
UAM	Física Experimental de Altas Energías de la Universidad Autónoma de Madrid	**heppc11.ft.uam.es**
UAL	Universidad de Santiago de Compostela, Centro de Supercomputación de Galicia	**www.usc.es/gaes, www.cesga.es**
UM	Universidad de Murcia	**www.ants.dif.um.es y www.ditec.um.es**
UAL	Universidad de Almeria, Supercomputación: Algoritmos Departamento de Arquitectura de Computadores y Electrónica	**www.ace.ual.es/Investigacion**
UNIOVI	Universidad de Oviedo, Asturias	
UPM	Laboratorio de Mecánica de Fluidos Computacional. Dpto. Motopropulsión y Termofluidodinámica, ETSI Aeronáuticos, Universidad Politécnica de Madrid	
UGR	Grupo de Circuitos y Sistemas para Procesamiento de la Información, Dpt. Arquitectura y Tecnología de Computadores, Universidad de Granada	
INTA	Instituto Nacional de Tecnología Aeroespacial, Torrejón de Ardoz - Madrid	**www.inta.es/**
ARCOS	Arquitectura de Computadores, Comunicaciones y Sistemas, Universidad Carlos III de Madrid, UCIIIM	**arcos.inf.uc3m.es**

IRISGrid no pretende dar servicio técnico, sino fijar las normas, protocolos, procedimientos que regulen el Grid. El objetivo de IRISGrid es facilitar a los grupos interesados la unión de sus recursos en Grid, definiendo los procedimientos relacionados principalmente con autenticación y monitorización de recursos. La potestad de los recursos está enteramente bajo control de los centros que los administran, sin que su inclusión en el Grid deba suponer un cambio en sus políticas de seguridad o de gestión de sus recursos locales.

1.3 Estructura del Documento

Este documento se estructura en cuatro partes principales. En primer lugar, la sección 1 contiene una breve descripción de los objetivos, participantes y términos de IRISGrid. La sección segunda describe el estado actual del Middleware, la sección tercera contiene una descripción por áreas de aplicación, de la situación actual de las tecnologías de Grid. Finalmente, la sección cuarta acaba con las conclusiones.

2 Estado del Middleware

El objetivo principal de esta sección es proponer los aspectos principales que se deberían incluir en el área temática de *middleware* dentro de la propuesta para la creación de un Programa de e-Ciencia. En primer lugar se describe de forma general el estado actual de la tecnología, y se analizan las diferentes líneas de investigación para cada uno de los niveles del *middleware* relevantes en un Grid. A continuación se indica como se contempla el desarrollo de tecnología Grid en el Plan Nacional de I+D+I actual, y se enumeran los proyectos de investigación en *middleware* Grid en España. Por último se realiza una propuesta de las líneas de investigación que se deberían incluir en el área temática de *middleware*.

2.1 Descripción de la Tecnología Grid

El objetivo de esta sección es proporcionar una visión global de la tendencia actual de las diferentes tecnologías que permiten aprovechar de forma conjunta los recursos disponibles en sistemas interconectados por red. Los siguientes modelos de computación en red aportan mecanismos para aprovechar al máximo los recursos distribuidos que generalmente se encuentran infrautilizados:

- **Cluster Computing**: Diseño de un cluster dedicado de equipos como alternativa a la adquisición de un equipo multiprocesador. Su ventaja fundamental es la mejor relación coste/rendimiento. Sus inconvenientes son: dificultad de programación y mantenimiento. Los clusters suelen estar gestionados por sistemas que se encargan de ejecutar las aplicaciones de los usuarios sobre las distintas máquinas en función de diferentes criterios de planificación fijados por el sistema. Estos sistemas de gestión pueden ser sistemas integrados de planificación como MOSIX (www.mosix.cs.huji.ac.il) o gestores de colas batch como PBS (versión libre) y PBS-Pro (versión comercial) de Veridian Systems (www.openpbs.org), LSF de Platform Computing (www.platform.com), SGE de Sun Microsystems (www.sun.com/Gridware) o Condor de la Universidad de Wisconsin-Madison (www.cs.wisc.edu/condor).
- **Intranet Computing**: Unión de la potencia computacional desaprovechada en los recursos hardware distribuidos en una red de área local (un único dominio de administración). Su principal ventaja es que puede proporcionar rendimientos semejantes a los ofrecidos por los sistemas de alto rendimiento con un coste económico casi nulo. La mayoría de los gestores de colas para clusters suelen ofrecer soluciones para unir múltiples clusters independientes dentro de una red local y mover los trabajos desde los clusters más ocupados a los más desocupados. Algunos de ellos también ofrecen facilidades para usar de forma oportunista recursos individuales que no están integrados en un cluster. También existen empresas como GridSystems, Avaki, Entropía o United Devices que comercializan software de Intranet Computing específico para aplicaciones paramétricas.
- **Internet Computing**: Aprovechamiento de la potencia de los recursos distribuidos por Internet siguiendo el modelo cliente/servidor. Actualmente casi todas estas herramientas se limitan a ejecución de aplicaciones paramétricas. Su ventaja es el gran rendimiento que se puede obtener. Sus principales inconvenientes son debidos al bajo ancho de banda y a la escasa seguridad en Internet. Avaki, Entropía o United Devices mantienen versiones de sus herramientas que permiten su uso en Internet. Probablemente el ejemplo más típico de Internet Computing es el proyecto

seti@home (setiathome.ssl.berkley.edu). Algunos gestores de colas como Condor ofrecen también la posibilidad de aprovechar recursos sobre Internet ofreciendo mayor seguridad y confiabilidad aunque requiere que todas las máquinas involucradas usen dicho gestor.

El uso de las tecnologías descritas anteriormente posibilita el aprovechamiento eficiente de los recursos dentro de una misma organización. Algunas de ellas como SGE, LSF o Condor permiten incluso unir diferentes departamentos u organizaciones pero con la condición de que sea su software el que gestione los recursos internos. Sin embargo, ninguna de estas tecnologías permite unir dominios de administración diferentes manteniendo la política de seguridad de cada centro y las herramientas de planificación ya en uso. Por otro lado, los interfaces y protocolos básicos que utilizan las herramientas anteriores no están basados en estándares abiertos, condición imprescindible para que la tecnología Grid se extienda, y en pocos años sea tan habitual como actualmente es la tecnología Web.

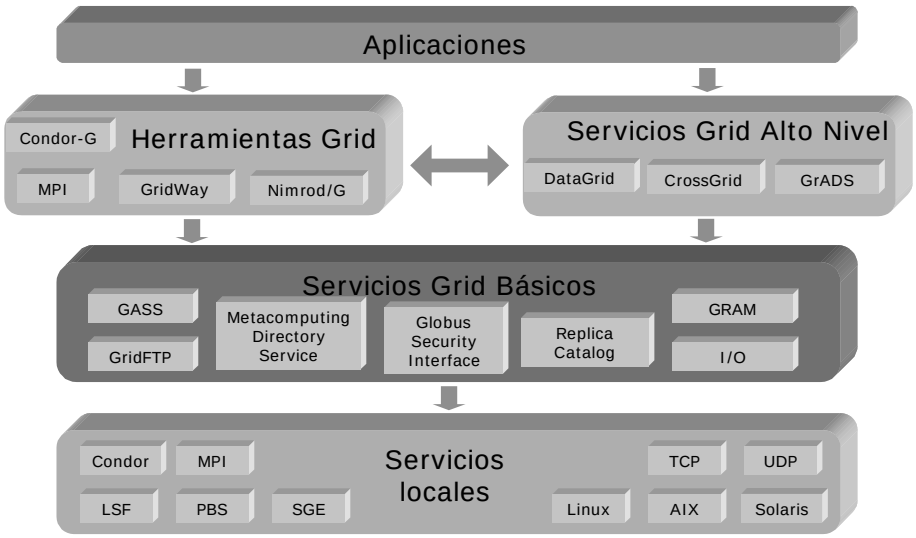
La necesidad de aprovechar los recursos disponibles en los sistemas informáticos conectados a Internet y simplificar su utilización ha dado lugar a una nueva forma de tecnología de la información conocida como *Grid Computing*. Esta nueva tecnología es análoga a las redes de suministro eléctrico: la idea es ofrecer un único punto de acceso a un conjunto de recursos distribuidos geográficamente (supercomputadores, clusters, almacenamiento, fuentes de información, instrumentos, personal...). De este modo, los sistemas distribuidos se pueden emplear como un único sistema virtual en aplicaciones intensivas en datos o con gran demanda computacional.

Las tecnologías descritas en la sección anterior son casos especiales de un nuevo paradigma de computación distribuida que en poco tiempo estamos seguros que va a revolucionar no sólo la computación de altas prestaciones sino Internet en general. Esta nueva tendencia denominada *Grid Computing* supone un cambio radical en la colaboración de sistemas conectados a Internet y en particular en la computación de altas prestaciones debido a su enorme potencial respecto al intercambio y gestión de recursos. Un sistema Grid se caracteriza por carecer de un control centralizado, estar basado en estándares abiertos y proporcionar calidad de servicio (www.Gridtoday.com/02/0722/100136.html). **Es importante resaltar que la tecnología Grid no pretende sustituir las tecnologías anteriores ya que su ámbito de aplicación es diferente**. El objetivo de la tecnología Grid es unir de forma segura los recursos de diferentes dominios de administración, respetando sus políticas de seguridad y herramientas de gestión internas. Además, su objetivo es unir todo tipo de recursos y no únicamente capacidad de procesamiento y almacenamiento.

2.2 Modelo de Capas del Middleware

Dentro de los sistemas *Grid* podemos diferenciar tres niveles de investigación relacionados con *middleware*:

- Servicios locales
- Servicios Grid básicos
- Servicios Grid alto nivel
- Herramientas Grid



2.2.1 Servicios Locales

En este nivel se incluye el desarrollo de herramientas para explotar los recursos en la Intranet del centro investigación. En este campo existen en la actualidad muchos servicios disponibles para realizar un uso eficiente de los recursos locales. Podemos destacar:

- Herramientas de monitorización de los diferentes recursos físicos del cluster (Ganglia: ganglia.sourceforge.net, Hawkeye: www.cs.wisc.edu/condor/hawkeye),
- Gestores de colas batch como los mencionados en la sección 1,
- Librerías de programación paralela (MPI: www-unix.mcs.anl.gov/mpi/, OpenMP: www.openmp.org/ o PVM: www.csm.ornl.gov/pvm/pvm_home.html),
- Herramientas de depuración y monitorización de aplicaciones (TotalView:www.etnus.com/Products/TotalView/, Paradyn:www.cs.wisc.edu/~paradyn/, Vampir:www.pallas.com/pages/vampir.htm, Paraver: http://www.cepba.upc.es/paraver).

Las **líneas abiertas de investigación** más representativas relacionadas con los servicios locales se centran en los siguientes aspectos:

- Escalabilidad de servicios y herramientas. A medida que los clusters locales van aumentando en tamaño, se hace también necesario adaptar las herramientas y servicios para que puedan usarse también sobre sistemas de miles de componentes.
- Automatización de políticas de gestión, configuración, control e instalación que garanticen el funcionamiento de clusters de gran tamaño sin la necesidad de una intervención humana constante y que puedan reaccionar a condiciones dinámicas.
- Mecanismos de tolerancia de fallos, que garanticen, por ejemplo, la ejecución de las aplicaciones o el funcionamiento de los servicios ante la presencia de fallos temporales en distintos elementos del cluster.

2.2.2 Servicios Grid Básicos

El *Grid Forum* (www.Gridforum.org) es el organismo encargado de crear los estándares de servicios y protocolos necesarios para crear la infraestructura o tecnología *Grid*. A pesar de que existen otras tecnologías *Grid* como Legion (www.cs.virginia.edu/~legion/), Unicore (www.unicore.org) o MOL (www.uni-paderborn.de/pc2/projects/mol/), la mayoría de los proyectos *Grid* actuales se están construyendo basados en los servicios y protocolos proporcionados por el ***Globus Toolkit*** (www.golbus.org). La tecnología Globus ha sido seleccionada como estándar de facto por las 12 compañías (Compaq, Cray, SGI, Sun Microsystems, Entropia, IBM, Microsoft, Platform Computing y Veridian en Estados Unidos; y Fujitsu, Hitachi y NEC en Japón) más importantes del sector de computación de altas prestaciones (http://www.globus.org/developer/news/20011112a.html). La próxima versión de Globus denominada **OGSA** (*Open Grid Services Architecture*) muestra una clara convergencia hacía la tecnología de Web Services en el campo de e-Business. El Globus Toolkit 3 (GT3), del que ya existe una versión alfa, apuesta por servicios Grid basados en Web Services. Esta evolución representa una gran oportunidad para lograr una amplia aceptación y difusión de la tecnología Grid, que puede extenderse, al igual que lo hizo el WWW, desde su ámbito original en el área de la computación científica, al de las aplicaciones comerciales.

Globus Grid toolkit es una colección de componentes software *open-source* y *open-architecture* diseñados para soportar el desarrollo de aplicaciones de alto rendimiento sobre entornos distribuidos tipo *Grid*. Realmente se trata de un conjunto de componentes autónomos que permiten al diseñador construir un *Grid*. Cada componente proporciona un servicio básico como autenticación, asignación de recursos, información, comunicación, detección de fallos y acceso remoto a datos. Los sistemas y aplicaciones Grid se pueden desarrollar empleando estos servicios y protocolos como elemento básico:

- GRAM (Globus Resource Allocation Manager): Este servicio proporciona un API para solicitar el inicio de ejecuciones en un recurso de computación y para gestionar estas ejecuciones una vez invocadas
- MDS (Metacomputing Directory Service): Este servicio proporciona un API para averiguar la estructura y estado de los recursos
- GSI (Globus Security Infrastructure): Este servicio proporciona una solución de autenticación global
- GASS (Global Access to Secondary Storage): Este servicio proporciona mecanismos de acceso y APIs para almacenar ficheros en diferentes sistemas

A partir de estos servicios básicos de middleware, las **líneas de investigación abiertas** se encuentran orientadas mejorar los protocolos y servicios básicos aportados por Globus para adaptarlos a nuevas tecnologías o ambientes. Por ejemplo:

- Añadir a Globus el soporte para IPv6 y por añadidura las funcionalidades de autoconfiguration, multihoming y seguridad.
- Mejorar la infraestructura de clave pública de Globus para hacerla más flexible e integrar extensiones actuales de las PKI para gestión de autorizaciones y credenciales.
- Implantar un esquema de autenticación y autorización más flexible y acorde con los estándares actualmente en uso en la infraestructura middleware de las redes académicas, con el objetivo de facilitar el uso de estos recursos a los usuarios finales del Grid e integrarlo en el uso normal de los recursos Internet

El movimiento hacia los Web Services que está experimentando Globus hace que también haya que replantear de nuevo los servicios ofrecidos, unido a la novedad del mundo de los servicios Web, que abre a la vez un campo de investigación muy interesante y potencialmente con mucho futuro.

2.2.3 Servicios Grid Alto Nivel

A pesar del tremendo esfuerzo realizado por la comunidad científica, la ejecución y gestión de trabajos en un Grid resulta una tarea ardua y difícil, debido principalmente a la naturaleza dinámica y compleja que caracteriza los Grids. Habitualmente el usuario ha de encargarse manualmente de todas las todos los pasos involucrados en la ejecución de un trabajo, a saber: descubrimiento y selección de recursos, inicialización, envío, monitorización, migración y finalización. Por lo tanto es necesario desarrollar servicios Grid de alto nivel y herramientas que realicen de forma automática y eficiente los pasos anteriores; además de adaptar la ejecución de un trabajo a las condiciones dinámicas del Grid (disponibilidad, carga, coste ....), así como a las demandas dinámicas de la aplicación (tiempo máximo de ejecución, presupuesto, necesidad de hardware/software específico,...).

En este nivel se incluye el desarrollo de middleware que, sobre el middleware básico proporcionado por Globus, proporciona servicios adicionales para aplicaciones y usuarios. Se trata de una capa de servicios reutilizables de más alto nivel que puedan hacer uso de los existentes para ofrecer servicios más flexibles en áreas como planificación, gestión de datos, visualización o colaboración. Por ejemplo, *trading* inteligente de componentes basado en especificaciones (mejoras al lenguaje de especificación RSL de Globus), introducción de mecanismos de calidad de servicio (QoS) para ofrecer servicios con ciertas garantías de servicio (soportando así servicios en tiempo real tales como las aplicaciones colaborativas), gestión de *scheduling* con prioridades, etc.

Entre los proyectos en curso cabe resaltar *European Data Grid* (www.eu-dataGrid.org), *Grid Physics Network* (www.griphyn.org), GrADS (nhse2.cs.rice.edu/grads/) y *CrossGrid* (http://www.crossGrid.org/), que tratan de desarrollar, entre otros, métodos eficientes de computación y acceso a datos distribuidos por medio de servicios basados en componentes Globus.

Por ejemplo, una aportación significativa del proyecto DataGrid es la implantación de un servicio *Resource Broker* que se encarga de la **gestión de trabajos**, buscando recursos del Grid y ejecutando el trabajo en los mismos. Este *broker* se encarga de llevar un seguimiento del trabajo, lanzándolo en otros recursos si alguno falla, hasta la finalización del mismo. Esto facilita el lanzamiento de trabajos al usuario, que sólo se ha de encargar de especificar los requerimientos del mismo y no de los recursos donde ejecutarlo. Además proporciona información sobre el estado del trabajo, haciendo de repositorio central de información y de los datos producidos por el mismo. El proyecto CrossGrid prevé hacer modificaciones sobre el *Resource Broker* anterior, para soportar algunas características hasta ahora no soportadas hasta ahora como el lanzamiento de aplicaciones MPI sobre varios 'sites', o mejorar el diseño del sistema pasando de un *Resource Broker* centralizado a una arquitectura distribuida basada en Agentes de *Scheduling*. Otra de los puntos fuertes de CrossGrid es el desarrollo de un portal Web sobre el cual los usuarios podrán lanzar sus trabajos, mejorando la interfaz con el sistema Grid subyacente.

Por tanto, los siguientes **aspectos pendientes por resolver** serían:

- Desarrollo de servicios de descubrimiento y selección de recursos

- Desarrollo de servicios de planificación y adaptación a las condiciones dinámicas de un Grid y de un trabajo
- Desarrollo de servicios de inicialización, envío, monitorización, migración y finalización de trabajos

Para la mayoría de las aplicaciones a ejecutar en entornos Grid el acceso a los datos es tan importante como el acceso a los recursos de cómputo, puesto que la mayoría de las aplicaciones científicas y de ingeniería requieren el acceso a grandes volúmenes de datos (terabytes o petabytes). En otras ocasiones los datos que generan las aplicaciones exceden de los recursos locales disponibles. Además, muchas de las aplicaciones futuras como lo entornos colaborativos virtuales también requerirán el acceso a datos ampliamente distribuidos.

Un **Grid de datos** proporciona un entorno en el que poder manipular y acceder a datos almacenados en sistemas ampliamente distribuidos. Su objetivo está orientado al intercambio y procesamiento de información de forma segura y eficiente, para lo cual hay que desarrollar servicios que permitan agrupar diferentes sistemas de almacenamiento locales donde poder almacenar, replicar e incluso fragmentar los datos. Los requisitos fundamentales que se imponen en un Grid de datos son:

- Espacio de nombres global.
- Técnicas de reducción de latencia.
- Consistencia en el acceso a los datos.
- Mecanismos de acceso común para localizar y acceder a los datos.
- API para el acceso a los datos.

Muchos de estos requisitos son similares a los de un sistema de ficheros local, sin embargo, presentan una problemática diferente, como son la necesidad de almacenar grandes cantidades de datos, existencia de diferentes protocolos de acceso a los datos locales (GridgFTP, NFS, HTTP-Webdav, etc.), problemas de seguridad y posibilidad no solo compartir datos, sino también distribuir y fragmentar los datos a través de los diferentes recursos de almacenamiento locales.

Los resultados que se han obtenido hasta la fecha en los Grids de datos se puede resumir en los siguientes puntos:

- Desarrollo de servicios de directorios que permiten la búsqueda de datos en entornos ampliamente distribuidos.
- Empleo de técnicas de replicación como forma de mejorar las prestaciones en el acceso a los datos. Estas técnicas, sin embargo, no son apropiadas en entornos colaborativos o cuando lo que se quiere es fragmentar los datos a través de diferentes recursos de almacenamiento.
- Desarrollo de protocolos de acceso a datos como GridFTP.
- Desarrollo de API para acceder a ficheros remotos, como el servicio GASS (Global Access to Secondary Storage) que proporciona Globus Grid toolkit.

Quedan, sin embargo, aspectos pendientes por resolver:

- Conseguir un verdadero espacio de nombres global, donde sea sencillo localizar los datos.
- Desarrollar técnicas de almacenamiento de altas prestaciones que permitan mejorar el acceso a los datos en un Grid, como por ejemplo, el empleo de técnicas de E/S paralela y distribución de datos a través de los diferentes recursos de almacenamiento

- Desarrollar servicios para integrar diferentes protocolos y sistemas de almacenamientos locales.
- Desarrollar interfaces de acceso a datos adecuadas para computación de altas prestaciones, como por ejemplo, MPI-IO para entornos Grid.
- Explotar las posibilidades de mecanismos de búsqueda e indexación distribuidos, que empleen las mismas tecnologías Grid para ayudar en la localización de recursos.
- Integrar estos procedimientos con los mecanismos de autorización, de manera que todas las interacciones sean susceptibles de personalización
- Mecanismos de tolerancia a fallos, que garanticen el acceso a los datos en presencia de fallos en el Grid.

En general**, la investigación en servicios de alto nivel se desarrollará en dos líneas**:

- Servicios de utilidad para todo el conjunto de aplicaciones (servicios horizontales), como por ejemplo, servicios de gestión de trabajos y datos, reserva anticipada de recursos, accounting distribuido, suscripción a eventos, tolerancia a fallos, etc.
- Servicios específicos para un dominio de aplicación (servicios verticales), por ejemplo, servicios de visualización 3D para los ámbitos científicos que los requieran, servicios de simulación distribuida genérica, etc.

Estos servicios se pueden ofrecer a través de Servicios Web o utilizando cualquier otro mecanismo RPC como CORBA. En este ámbito, la tecnología de componentes surge como una tecnología de desarrollo muy adecuada. El desarrollo basado en componentes (iniciado, aunque en un estado primitivo todavía, por la nueva versión de Globu 3) permite que los desarrolladores de aplicaciones creen componentes que pueden ser utilizados por cualquier otra aplicación. El desarrollo de aplicaciones se convierte entonces en una actividad de conexión de los distintos componentes que están disponibles para el desarrollo de las mismas y que están distribuidos por todo el Grid. La gestión de qué componentes existen en cada centro de computación se puede realizar de forma automática por herramientas añadidas a las básicas de Globus, además de otras herramientas de desarrollo de aplicaciones basadas en el descubrimiento automático de componentes, la conexión y la puesta en ejecución. El uso de componentes para programación Grid se ha iniciado en diferentes proyectos, como el Common Component Architecture (cca-forum.org), GridCCM (www.irisa.fr/paris/Gridccm/welcome.htm), la especificación Lightweight CCM, Globus en su versión 3 y el propio proyecto PIRAMIDE con el modelo de componentes CORBA *Lightweight Components* (CORBA-LC).

2.2.4 Herramientas Grid

En este nivel se incluyen herramientas de más alto nivel como librerías de programación, entornos especializados para la resolución de problemas y otras herramientas de ayuda al desarrollo de aplicaciones. Estas herramientas se basan en las componentes básicas y de alto nivel.

2.2.4.1 Librerías y Herramientas para Tipos Específicos de Problemas

Tradicionalmente, existe un conjunto de aplicaciones cuyas demandas computacionales no pueden ser satisfechas por los supercomputadores actuales. La resolución de estos problemas requiere un estudio de cómo el Grid puede mejorar los paradigmas de supercomputación actuales, la investigación de nuevos paradigmas y el desarrollo de herramientas de alto nivel y API's (Application Programmer Interface) que permitan a

los científicos e ingenieros expresar de forma sencilla los problemas en un entorno Grid. En particular, ejemplos de problemas que potencialmente pueden beneficiarse del uso de Grids son:

- **Aplicaciones distribuidas de Alto Rendimiento** (High Performance Computing, HPC); cuyos requisitos computacionales únicamente puede satisfacerse mediante la unión de múltiples supercomputadores. Ejemplos de estas aplicaciones son, la dinámica de fluidos computacional, la simulación numérica de procesos físicos complejos, o simulaciones meteorológicas. Entre los proyectos actualmente en desarrollo cabe destacar: MPICH-G2 (www3.niu.edu/mpi/), y Cactus (www.cactuscode.org).
- **Aplicaciones de Alta Productividad** (High Throuput Computing, HTC); problemas que requieren del análisis de todas las posibles soluciones en un espacio de parámetros (problemas NP-Completos). Aunque actualmente estos problemas se resuelven reduciendo el espacio de parámetros mediante alguna heurística (enfriamiento simulado, algoritmos genéticos,...), el Grid ofrece una infraestructura computacional más adecuada para resolver estos problemas de alta productividad débilmente acoplados. Entre los proyectos actualmente en desarrollo cabe destacar: Nimrod/G (www.csse.monash.edu.au/~rajkumar/ecoGrid/), MW (www.cs.wisc.edu/condor), AppLeS (grail.sdsc.edu/), GridWay (www.dacya.ucm.es/asds)
- **Aplicaciones de Ejecución Auto-Adaptativa**: La aparición de la tecnología Grid ha dado lugar a un nuevo paradigma de aplicaciones capaces de adaptar su ejecución de acuerdo a sus requisitos dinámicos. Por ejemplo, los métodos numéricos de refinamiento adaptativo de malla, aumentan de forma sistemática la resolución de la malla numérica en aquellas zonas del dominio con elevados errores de discretización. De esta forma no es posible conocer a priori la cantidad de memoria que necesitará la simulación. Las aplicaciones auto-adaptativas son capaces de buscar recursos adicionales a medida que evoluciona su ejecución para satisfacer sus requisitos. Entre los proyectos actualmente en desarrollo cabe destacar: Cactus Worm (www.cactuscode.org), GridWay (www.dacya.ucm.es/asds).

Otros ejemplos de entornos especializados incluirían, que permite el la resolución de problemas que sigan un modelo de master-worker, Netsolve (http://icl.cs.utk.edu/netsolve), herramienta basada en una arquitectura cliente/servidor para el uso de librerías matemáticas sobre entornos distribuidos, o DAGman (www.cs.wisc.edu/dagman), servicio para controlar aplicaciones compuestas por múltiples trabajos que exhiben relaciones de dependencia entre ellos.

2.2.4.2 Herramientas de ayuda al desarrollo de aplicaciones

Durante el ciclo de desarrollo de las aplicaciones, es necesario realizar prototipos, depurarlos, sintonizarlos, etc. Cuando las aplicaciones van a ser ejecutadas en un entorno en Grid, la complejidad de este proceso es mucho mayor. Herramientas de desarrollo tales como depuradores, y sintonizadores de la eficiencia son esenciales para que los usuarios puedan entender porque su aplicación no esta dando los resultados o la eficiencia esperados. Otro tipo de herramientas altamente útiles son aquellas que permiten de manera automática obtener programas ejecutables en Grid.

En la actualidad existen pocas herramientas que realicen las tareas indicadas previamente. Algunos ejemplos existentes son:

- P2D2: depurador "Grid-enabled" (NASA)

- Guard: depurador paralelo "relativo" (Monash)
- Ygdrasil toolkit: permite integración a nivel de línea de comando de depuradores como Ladebug, gdb, etc (HP)
- SimGRID: simulador de planificaciones de aplicaciones distribuidas (UCSD, USA)
- Dimemas: simulador de la eficiencia de programas MPI ejecutados distribuidamente (CEPBA-UPC)
- Paraver (CEPBA-UPC), Vampir (Pallas): Herramientas de visualización de la ejecución de aplicaciones distribuidas.

A pesar de que existen algunas herramientas, la mayoría de ellas están en un estado preliminar de desarrollo. **Es necesario extender las funcionalidades de dichas herramientas, robustecer las versiones actuales y lo mas importante hacer conocer estas herramientas a los posibles usuarios.** El campo en el que claramente existe un vacío es el de modelos de programación que faciliten el desarrollo de aplicaciones en GRID de manera que el usuario pueda especificar la funcionalidad de su aplicación sin necesidad de saber con detalle la topología ni características del Grid en el que se ejecutará.

2.3 Conclusión

A pesar de haberse realizado un gran esfuerzo durante los últimos años, la madurez del *middleware* actual es relativa. Todavía queda bastante camino por recorrer hasta contar con *middleware* que permita una explotación eficiente y sencilla de un entorno Grid. Proponemos las siguientes líneas de actuación:

1. Servicios locales
    - Herramientas de monitorización,
    - Gestores de colas batch
    - Librerías de programación paralela
    - Herramientas de depuración y monitorización de aplicaciones
2. Servicios Grid Básicos
    - Mejora de protocolos y servicios básicos
3. Servicios Grid de Alto Nivel
    - Servicios de gestión de trabajos y datos para todo el conjunto de aplicaciones (servicios horizontales)
    - Servicios específicos para un dominio de aplicación (servicios verticales)
    - Uso de componentes
4. Herramientas Grid
    - Herramientas para aplicaciones de alta productividad
    - Herramientas para aplicaciones de alto rendimiento
    - Herramientas para aplicaciones adaptativas
    - Herramientas de ayuda al desarrollo de aplicaciones
    - Herramientas para aplicaciones intensivas en datos

Por otro lado, no debemos olvidar que es necesario estar en contacto con los proyectos orientados al desarrollo de aplicaciones en las diferentes áreas temáticas.

3 Áreas de Aplicación

3.1 Área de Meteorología

3.1.1 Motivación y Necesidades

Históricamente, la Meteorología ha sido una de los principales usuarios de las nuevas tecnologías de la Computación, tanto en lo relativo a la capacidad de cálculo, como al almacenamiento de grandes volúmenes de información y a su rápida distribución mediante redes de alto rendimiento. En el pasado, muchas de las tareas involucradas en este área (integración de modelos numéricos de predicción, mantenimiento de bases de datos operativas, etc.) eran exclusivas de grandes centros meteorológicos que disponían de la tecnología necesaria. En la actualidad, la situación es distinta debido al abaratamiento de la tecnología, y diversos grupos de investigación públicos y privados llevan a cabo costosas simulaciones meteorológicas que utilizan distintas bases de datos para realizar tareas tan diversas como: estudios climatológicos y de cambio climático, pronóstico meteorológico local, predicción de viento para la gestión de parques eólicos, difusión de contaminantes en el mar y en la atmósfera, etc.

Sin embargo, el tipo de estudios que pueden emprender estos grupos está todavía limitado por los recursos computacionales de que disponen. Asimismo, la colaboración entre distintos grupos para abordar proyectos comunes se ve dificultada por la heterogeneidad de bases de datos y formatos de uso común en este ámbito. Por tanto, la tecnología GRID puede proporcionar un doble beneficio en este Área permitiendo abordar problemas más complejos y facilitando la colaboración y acceso/compartición de datos a los mismos.

La propuesta de Centros de e-Ciencia en la que se encuadra esta iniciativa parece el marco más apropiado para implantar esta tecnología, proporcionando un soporte técnico apropiado para emprender iniciativas en este Área.

A continuación se describen más en detalle las características de estos procesos y bases de datos involucrados con la actividad de esta área:

3.1.1.1 Procesos que requieren computación de altas prestaciones

La integración de cualquier modelo numérico de circulación atmosférica u oceánica es un proceso computacionalmente costoso, tanto en tiempo de cómputo como almacenamiento de información. Estos modelos constituyen la herramienta fundamental para numerosas aplicaciones prácticas. Por otra parte, la aplicación de distintas técnicas estadísticas (correlación canónica, regresión múltiple, técnicas de clasificación, etc.) a campos atmosféricos almacenados requiere también un enorme esfuerzo computacional debido a las dimensiones de los datos.

En la actualidad las técnicas de predicción basadas en conjuntos requieren aún mayor esfuerzo computacional ya que los modelos se integran varias veces, a partir de condiciones iniciales perturbadas. Este tipo de aplicaciones es un ejemplo típico de aplicación paramétrica, que puede explotarse fácilmente en un entorno GRID obteniendo una mayor productividad.

Por otra parte, los proyectos de reanálisis son necesarios para llevar a cabo estudio climatológicos en zonas de interés. En este caso, los modelos se integran durante largos períodos de tiempo guardando principalemente los análisis obtenidos; esta tarea da lugar

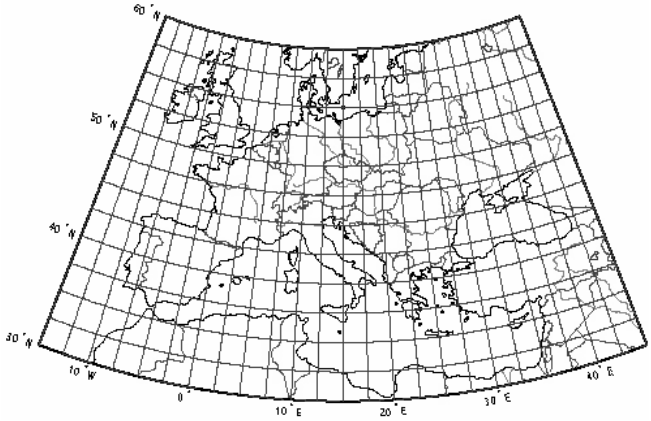
a enormes bases de datos que definen las climatología de la zona de interés (a la resolución dada por el modelo integrado). Éste es otro ejemplo de aplicación paramétrica en la que el entorno GRID permitiría incrementar la productividad.

3.1.1.2 Bases de datos

Una de las principales características de esta área es la gran cantidad y diversidad de datos que se manejan en cualquier aplicación real. En esta sección tratamos de describir las fuentes más importantes.

3.1.1.3 Salidas de modelos numéricos

Los modelos numéricos de circulación atmosférica son la principal herramienta en la Meteorología moderna. Estos modelos proporcionan una predicción del estado de la atmósfera a distintas resoluciones temporales (de horas a meses) y espaciales (entre los cientos y las decenas de Km de resolución horizontal). Por ejemplo, la figura siguiente muestra una rejilla horizontal utilizada para simular la dinámica mensual de la atmósfera sobre Europa, con una resolución horizontal de 300 Km (esta rejilla tiene otra componente vertical, con 15 niveles de altura distintos).



En cada uno de estos puntos de la rejilla 3D se simula el valor de distintas variables primitivas (temperatura, humedad, geopotencial, viento) y derivadas (espesores, vorticidad, precipitación, etc.). Todo este volumen de información se genera en cada paso de integración y se almacena horaria o diariamente, dependiendo de la resolución temporal de la integración. Por tanto el volumen de información generado en cada simulación es enorme.

Aparte del uso operativo de estos modelos, en la actualidad resultan de gran utilidad los proyecto de Re-análisis, donde un mismo modelo es integrado un largo período de tiempo (decenas de años) para obtener salidas homogéneas de un mismo modelo válidas para estudios Climatológicos o Estadísticos. El volumen de datos almacenados en estas bases de datos de reanálisis es del orden de Tera Bytes. Algunas de estas bases de datos de reanálisis son públicas, como la del NCEP/NCAR y otras están disponibles para trabajos de investigación (como la del ECMWF).

3.1.1.4 Observaciones y datos de estaciones meteorológicas

Las estaciones situadas en tierra, en barcos y en aviones, proporcionan observaciones de variables atmosféricas en todo el mundo. En España distintos organismos disponen de numerosas estaciones de observación a lo largo y ancho de la geografía. El número de estaciones que se instalan para distintas aplicaciones (estudio de cuencas hidrológicas, etc.) crece cada día proporcionando una tupida red de datos, algunos de los cuales son

públicos y otros privados para los que el entorno GRID tendrá que proporcionar medidas de privacidad apropiadas.

3.1.1.5 Datos de satélites y radares

Aparte de los datos ya mencionados, existe numerosa información pública relativa a mediciones de satélites y radares, que proporcionan datos interesantes sobre la dinámica de las nubes y la fenomenología asociada. En el corto plazo estos datos (imágenes, etc.) pasarán a formar parte de la cadena operativa, siendo asimilados por los modelos, o siendo post-procesados con técnicas inteligentes de minería de datos.

3.1.2 Casos de uso

La iniciativa propuesta en el área de la Meteorología está vertebrada sobre distintas casos de uso en los que se podría aplicar la tecnología GRID en esta área mejorando la eficiencia de los sistemas actuales y permitiendo llevara a cabo proyectos que hasta ahora no son viables por la limitación computacional. La implementación de estas aplicaciones requerirá una estrecha colaboración con los grupos encargados del desarrollo de middleware (software intermedio entre las aplicaciones y el entorno GRID) que proporcionarán las herramientas apropiadas para que el acceso a datos, ejecución de procesos en distintos clusters, etc., sean tareas casi-transparentes para los grupos que desarrollen aplicaciones.

Por otra parte, una iniciativa de este tipo tiene el beneficio añadido de fomentar el **trabajo colaborativo** entre distintos grupos de investigación en Meteorología Españoles (compartición segura de observaciones, salidas de modelos numéricos, reanálisis, predicciones operativas, etc.), así como el desarrollo de proyectos comunes aunando esfuerzos computacionales.

Algunos casos de uso de interés para esta comunidad son los siguientes:

1) Un problema de enorme interés en la actualidad es la **predicción por conjuntos**. En la actualidad la tecnología GRID está suficientemente madura para permitir la ejecución de aplicaciones paramétricas, permitiendo incrementar su productividad. La predicción por conjuntos es un ejemplo típico de este tipo, ya que se trata de ejecutar la misma aplicación (modelo atmosférico) con distintas condiciones iniciales (parámetros); por tanto, un primer ejemplo realista de aplicación Meteorológica en el entorno GRID sería el desarrollo de un prototipo de sistema de predicción por conjuntos que integrase los distintos modelos y métodos de perturbación utilizados por los distintos grupos nacionales.
2) **Integración de modelos de área limitada en regiones de interés**: Un denominador común a los distintos grupos que realizan desarrollo e investigación en el ámbito de la Meteorología es la necesidad de simular la circulación de la atmósfera en determinadas condiciones (incluso diariamente, de forma operativa). En la actualidad existen distintas simulaciones globales de baja resolución disponibles, tanto de reanálisis (integraciones en tiempo pasado para un período largo de tiempo, 10-50 años), como operativas (que proporcionan los campos atmosféricos previstos con unos días, semanas, o incluso meses de antelación). En la mayoría de las ocasiones, la resolución de estos modelos no es suficiente para analizar diversos problemas locales y, por tanto, es necesario integrar modelos de mayor resolución sobre una zona de interés concreta (modelos de área limitada, o modelos regionales), utilizando como condiciones iniciales y/o de contorno de los modelos globales.

Uno de los modelos regionales de dominio público más populares es el MM5 (por ejemplo, en la Península distintos grupos de investigación que componen la "Red Ibérica para la investigación y desarrollo de aplicaciones en base al modelo atmosférico MM5" http://redibericamm5.uib.es/ integran separadamente este modelo en distintas regiones peninsulares).

La integración eficiente de un modelo atmosférico aprovechando el entorno grid involucra el análisis de la escalabilidad de la implementación paralela dependiendo de los recursos computacionales y de la latencia de la red (recursos que, a diferencia de la ejecución en un cluster local, varían de una ejecución a otra). Por tanto, una aplicación piloto de gran interés científico sería estudiar el rendimiento de distintas paralelizaciones de **MM5 en el entorno GRID**.

3) Otra aplicación piloto de gran interés para la comunidad científica es la realización de proyectos de **reanálisis de alta resolución**. En este caso, se tratar de integrar el mismo modelo atmosférico durante un largo período de tiempo (representativo de la climatología del problema que se quiera abordar). En este caso, se podría utilizar el modelo MM5 como aplicación paramétrica que sería ejecutada con distintas fechas (parámetros) en el entorno GRID aumentando la productividad. Esto permitiría disponer de reanálisis regionales en áreas de interés en un tiempo razonable. El modelo puede inicializarse a partir de condiciones de reanálisis del NCEP o del ECMWF.
4) Una vez que en el GRID haya sido almacenada información de reanálisis (global y regional), podrían implementarse **técnicas inteligentes de acceso a la información (minería de datos)**. En este caso se trata de obtener patrones de información (Empirical Ortogonal Functions EOFs, clasificaciones, etc.), en lugar de la cantidad ingente de datos crudos. Por ejemplo, en muchas ocasiones sólo es necesario acceder a las componentes principales de ciertos campos, a las correlaciones canónicas de campos y observaciones, o incluso a los campos análogos a un campo dado (los vecinos); en estos casos, sería de gran utilidad disponer de servicios de acceso a la información que proporcionasen directamente la información solicitada. Este tipo de aplicaciones se encuadra en el marco más general de la Minería de Datos, necesaria para gestionar de forma apropiada complejas bases de datos con ingente información.

3.1.3 Middleware actual y específico

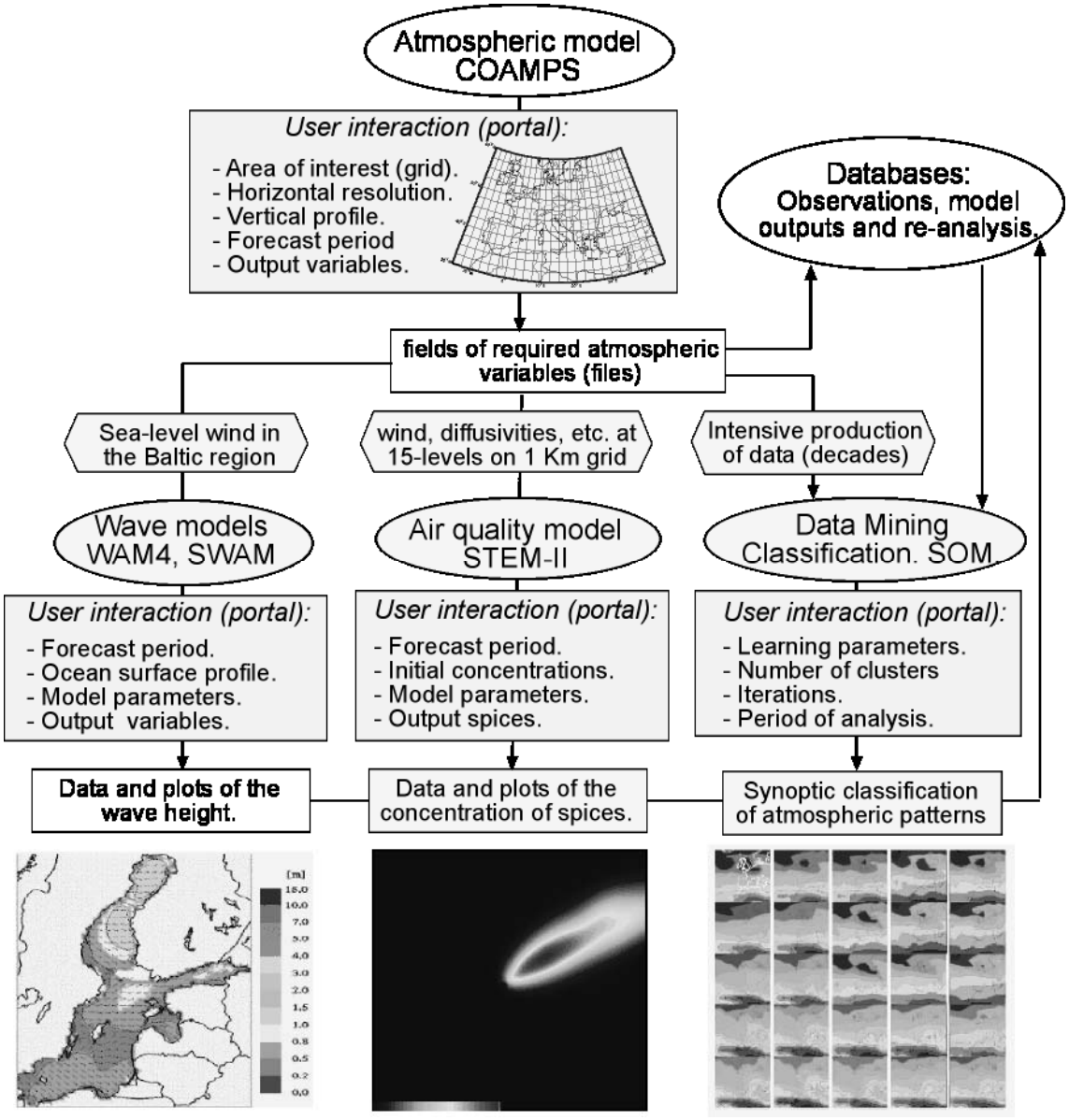
El middleware utilizados en los proyectos piloto en esta área se ha restringido hasta la fecha al uso de las distintas versiones de Globus. La parte más específica de esta área es la relacionada con los distintos formatos de los datos meteorológicos que se utilizan. Por ello, para un acceso distribuido transparente a los datos será necesario desarrollar algún módulo específico de codificación/decodificación automática en cualquier formato.

3.1.4 Proyectos piloto

Los primeros desarrollos y proyecto piloto de aplicaciones meteorológicas en entornos GRID se están llevando a cabo en el marco de proyectos Europeos. Por ejemplo, el proyecto Europeo del V Programa Marco CROSS-GRID (http://www.crossgrid.org), cuenta con una activa participación de grupos Españoles en el ámbito de la dispersión de contaminantes en la atmósfera y en la implementación de herramientas de minería de datos para bases de datos meteorológicas. Fruto de estas iniciativas se están resolviendo distintos problemas de migración y adaptación de las técnicas y productos existentes al

nuevo entorno GRID, facilitando el trabajo para futuras iniciativas en este campo. Otra de las experiencias piloto en este proyecto es la paralelización y migración a GRID del modelo COAMPS "Coupled Ocean/Atmosphere Mesoscale Prediction System" en forma de aplicación web en la que el usuario pueda seleccionar interactivamente un área de interés, la resolución horizontal y vertical, y un período de predicción, obteniendo los campos meteorológicos solicitados resultado de la integración del modelo. Está planeado integrar este servicio web con aplicaciones específicas de cálculo de contaminates en la atmósfera, cálculo de oleaje, y procesamiento y análisis de proyectos de re-análisis regionales. En la figura siguiente se muestran más detalles sobre los proyectos piloto involucrados en este proyecto.

Por otra parte, también se pretende utilizar tecnología GRID en el proyecto Intergrado del VI Programa Marco ENSEMBLES, que comenzará en el 2004 y que generará grandes volúmenes de información (salidas de modelos regionales y climáticos en distintos escenarios de cambio climático).



3.1.4.1 Barreras para el despliegue de aplicaciones

A nivel nacional todavía no existe ninguna experiencia piloto de que haga uso de esta tecnología. El principal problema para ello es la falta de recursos computacionales y de personal especializado de apoyo que permita llevar a la práctica alguno de los casos de uso descritos. Por ello, el fomento de centros de e-ciencia que den soporte computacional y técnico ayudaría a llevar a cabo estas tareas.

3.1.4.2 Grupos participantes en el área temática

La propuesta de colaboración en el desarrollo de una infraestructura GRID en el ámbito de la Meteorología cuenta con el interés de distintos grupos de investigación. Hasta ahora se ha tratado únicamente de recabar las expresiones de interés de algunos de los grupos que desarrollan tareas relacionadas con los casos de uso descritos anteriormente, pero no se ha iniciado ninguna colaboración a nivel nacional dadas las barreras comentadas en el punto anterior.

- Inteligencia Artificial en Meteorología http://grupos.unican.es/ai/meteo
  Universidad de Cantabria / CSIC
  Contacto: José Manuel Gutiérrez (gutierjm@unican.es)
- Grupo de Fisica No Lineal http://chaos.usc.es
  Unidade de Observacion y Prediccion Meteoroloxica http://meteo.usc.es
  Universidad de Santiago de Compostela
  Contacto: Vicente Pérrez Muñuzuri (uscfmvpm@cesga.es)
- Grupo de Meteorología
  Universidad de las Islas Baleares
  Contacto: Juan M. Torres (joanm.torres@uib.es)
- Grupo de Predicción y caracterización del recurso eólico
  CIEMAT
  Contacto: Jorge Navarro Montesinos (jorge.navarro@ciemat.es)
- Meteorology and Climate Applications and Modelling (MCAM)
  Universidad Complutense - Universidad de Murcia
  Contacto: Juan Pedro Montavez (jpmontav@ucmail.ucm.es)
- Grupo de Meteorología
  Universidad del País Vasco
  Contacto: Jon Saenz (jsaenz@wm.lc.ehu.es)

3.1.5 Conclusiones

En el área de Meteorología existe suficiente masa crítica y problemas de interés como para que el desarrollo de aplicaciones Grid pueda tener éxito. La experiencia inicial y participación futura en distintos proyectos Europeos por parte de grupos nacionales así lo acredita. Sin embargo, no se dispone de una infraestructura computacional y técnica apropiada para plantear proyectos Grid a nivel nacional. El desarrollo de centros de e-Ciencia que den este soporte a los grupos nacionales permitiría llevar a la práctica proyectos en este área con garantías de éxito.

3.2 Área de Física de Altas Energías

La Física Experimental de Partículas Elementales, muy a menudo llamada Física de Altas Energías (en adelante, FAE), es un campo de bastante actividad en España, considerado prioritario desde 1983 y actualmente parte del Programa Nacional de Física de Partículas 2004-2007.

Actualmente, la comunidad trabaja organizada en "experimentos", cada uno de los cuales es una actividad colaborativa a nivel mundial para diseñar, construir, operar y realizar el análisis de datos de un detector. Los detectores están localizados en un número bastante reducido de sitios, tradicionalmente Laboratorios que operan aceleradores (esencialmente en Europa, USA, Japón y China) o en lugares idóneos para detectar rayos cósmicos o realizar otros tipos de experimentos sin acelerador (esta modalidad está teniendo un gran auge en los últimos 10 años y se espera que esta tendencia siga, con experimentos en lugares generalmente bastante remotos en Argentina, Antártida, el desierto del Sudoeste de USA, el fondo del Mar Mediterráneo, Siberia, Australia, etc. o cerca de observatorios astronómicos en Chile, Canarias, etc.).

Sin preocuparnos de factores de orden 2 y con la intención de generar una pincelada fácil de recordar, podemos hablar de los siguientes órdenes de magnitud:

Número de científicos en FAE…….	10^4
Tamaño promedio de un grupo colaborador en un experimento dado…….	10
Número de institutos con grupos en FAE……	10^3
Número de países con grupos en FAE…….	10^2
Duración en años de un experimento (desde concepción hasta desmantelamiento)……...	10

Una tendencia importante es que un experimento en particular tiene una duración total cada vez mas larga, y que viene a ser del mismo orden que la longitud de una carrera profesional. Por otro lado, tanto los detectores como los aceleradores se construyen espaciados en el tiempo, por tanto en cualquier momento se puede encontrar alguno en cualquier fase (conceptualización, diseño, construcción, toma de datos, análisis, upgrade, desmantelamiento). Por tanto, un físico o ingeniero o técnico particular tiende a trabajar simultáneamente en unos cuantos experimentos, con una constante de tiempo de cambio de contexto que varía entre unos pocos días y unos pocos años. Otra característica importante es que aunque los físicos de FAE se especializan, lo hacen de una manera plural cubriendo aspectos de las fases anteriormente descritas, y por tanto participan íntimamente en muchas de dichas fases. Esto quizás sea algo diferente de lo que ocurre en otros campos, donde hay una diferenciación más aguda entre los "científicos" y los "instrumentalistas". Por último, hay esencialmente una completa descorrelación entre las localidades geográficas de la institución de un físico particular, de las instituciones de sus colaboradores y de sus experimentos.

Dado este panorama, no será ninguna sorpresa que FAE utiliza de manera generalizada el soporte digital en modo colaborativo, cubriendo toda la gama desde correo electrónico, Webs colaborativos, repositorios con acceso remoto de diseños mecánicos,

electrónicos y de software, hasta la videoconferencia y la gestión y operación remota de dispositivos y sistemas de adquisición de datos. Dado el gran número de países, instituciones y dominios de administración que intervienen, la única manera práctica de realizar todas estas actividades es utilizando herramientas con estándares abiertos a través de Internet. El resultado neto es que, excluyendo el análisis colaborativo de datos, cada colaborador en FAE es un "e-Físico" que colabora continuamente a escala mundial utilizando un ordenador personal que es necesario renovar aproximadamente cada 3 años, una infraestructura de correo, Web y repositorios colaborativos que contiene unos 100s de GB y que, en condiciones ideales, consumiría individualmente el equivalente nominal de una Ethernet tradicional, es decir 10 Mbps. Es interesante notar que 10^4 personas x 10 Mbps = 100 Gbps nominales, aunque evidentemente la estructura de picos y valles de utilización es muy aguda. La falta de red académica de calidad es un factor crítico limitante a la colaboración, que sorprendentemente no sólo se manifiesta en los llamados "países menos desarrollados" sino también en Europa, USA, Canadá y Japón debido a deficiencias de "último Kilómetro" o incluso "últimos metros" en campus universitarios.

3.2.1 Motivación y Necesidades

El principal problema de estos "e-Físicos" es la necesidad de acceder continuamente de un dominio de administración/seguridad a otro y su sueño es tener un "*single sign-on*" global como individuo por un lado, y por el otro el poder cambiar su *contexto o rol* de una manera fácil y relativamente dinámica para facilitar el trabajo cuasi-simultáneo en varios experimentos. Estos deseos se están haciendo realidad poco a poco, gracias al despliegue de Infraestructuras Grid como EGEE. Se espera que gracias al despliegue coordinado de un conjunto interoperable de Autoridades de Registro y de Certificación X.509 todas las personas de la comunidad FAE tengan en breve un certificado reconocido a nivel mundial. Luego, gracias a sistemas todavía en fase experimental, como el "*Virtual Organization Management System*" desarrollado en el proyecto EU DataGrid, se puede gestionar el acceso a recursos asignados a cada experimento (es decir, a cada "Organización Virtual") según la rol asignada a cada individuo.

3.2.1.1 El caso distinguido: tratamiento y análisis de los datos:

El tratamiento de los datos de FAE en sí requiere atención especial por su magnitud y dificultad. Es difícil generalizar, pero se puede decir que casi cualquier experimento actual está generando al menos 10-100 GB *por día* de datos a analizar, y en muchos casos 1 TB por día es habitual. Como se verá mas abajo, el Large Hadron Collider tendrá un volumen muy superior. Una figura muy aproximada es que el tratamiento de estos datos requiere entre 100 y 10000 "instrucciones" por byte, utilizando la palabra "instrucción" en el sentido de la definición de MIPS. El número de individuos que apliquen dichos tratamientos está generalmente anti-correlacionado con los recursos necesarios, es decir que hay muchos tratamientos de pocas "instrucciones" por byte (el análisis individual de datos muy filtrados o muy reconstruidos) y pocos tratamientos de muchas "instrucciones" por byte (re-procesamientos o simulaciones organizados por la colaboración experimental). Una figura extremadamente aproximada es que en total se necesitan unas 20000 "instrucciones" por byte por año, y que el flujo integrado global de datos alcanza al menos el triple del flujo que emana del sistema de adquisición de datos.

Estas cifras son bastante superiores a las de décadas anteriores y desbordan el modelo habitual en la década de los 90 basado en realizar gran parte del tratamiento de datos en estaciones de trabajo RISC y ordenadores personales conectados por Ethernet.

Afortunadamente, el hecho de que los datos de FAE son mayoritariamente un enorme conjunto de registros independientes cuyo tratamiento y análisis constituye un ejemplo de computación "vergonzosamente paralela", sigue vigente, y por tanto soluciones basadas en clusters de, por ejemplo, PCs con Linux y Gigabit Ethernet son posibles. La única peculiaridad, que afecta más a gestores que a científicos, es el hecho de que el número de ordenadores pasa a ser superior al número de físicos.

Para dar números mas concretos, aunque con la advertencia de que están en proceso de revisión y que pueden haber cambios de factores de 2 o mas, se citan en la Tabla 1 algunas de las estimaciones recientes para la computación de los cuatro experimentos del Large Hadron Collider (LHC). Estas necesidades probablemente representen del orden de la mitad de las necesidades globales de FAE en 2006-2010.

Si las capacidades de almacenamiento, proceso y transferencia de datos necesarias para el LHC son de por sí bastante altas, es importante mencionar también que dichos datos se corresponden a unas 10^{11} a 10^{12} colisiones, entre las cuales los 6.000 físicos del LHC distribuidos en cientos de institutos buscarán subconjuntos quizás tan pequeños como 10^3 o 10^2 colisiones, utilizando características iterativamente derivadas de los datos y no conocidas *a priori*. Es evidente que un nivel importante de gestión automática y fiable de subconjuntos de datos será muy necesario. Dicho sistema de gestión aún queda por desarrollar y deberá de ser el objeto de intenso desarrollo en un futuro inmediato.

La evolución de FAE en las últimas décadas ha estado dominada por el diseño, construcción y explotación de detectores de partículas (en adelante, experimentos) cada vez mas grandes y complejos, los cuales operan en aceleradores que colisionan haces de partículas cada vez mas energéticos y/o intensos. Esta tendencia se puede explicar a nivel divulgativo a través de tres observaciones relacionadas a la física fundamental: a) la relación inversa entre longitud de onda y energía (mecánica cuántica), b) el incremento del número de partículas producidas cuando se incrementa la densidad de energía (similaridades entre la teoría de campos y la termodinámica) y c) el que la existencia de partículas con masa mayor que la energía disponible pueda modificar ligeramente observables (teoría cuántica de campos). Mas coloquialmente, para avanzar el conocimiento en FAE o se observan colisiones con mucha energía para explorar distancias muy pequeñas o se observa un gran número de colisiones para explorar efectos muy sutiles, o las dos cosas a la vez. Esto a su vez conlleva la construcción de experimentos de gran tamaño, ya que para poder distinguir con suficiente precisión las trayectorias individuales de un gran número de partículas producidas y también medir su momento o energía, por curvatura en un campo magnético si son cargadas o por métodos calorimétricos si no lo son, se han de vencer las limitaciones de resolución de detección de posición, máximo campo magnético generable en condiciones prácticas y cantidad de material necesario para absorción calorimétrica.

Además, particularmente en la última década, ha habido un gran desarrollo de experimentos de partículas que no utilizan aceleradores, observando partículas elementales que inciden sobre la tierra, tales como rayos cósmicos, neutrinos o rayos X. Este desarrollo está desdibujando las fronteras entre FAE, astrofísica, astronomía y cosmología. Una particularidad de estos experimentos es que muy a menudo requieren localizarse en zonas bastante remotas, tales como desiertos, el fondo marino, Antártida, etc., en los cuales no hay una infraestructura de soporte similar a la de un laboratorio de aceleradores. Además muchas veces la conectividad de red puede ser inadecuada, obligando a resucitar técnicas abandonadas por FAE hace muchos años, como por ejemplo la grabación y envío de cintas magnéticas con los datos experimentales. En

España, por ejemplo, este es el caso del Observatorio del Roque de los Muchachos en La Palma, Canarias.

Las colaboraciones experimentales han crecido en tamaño a medida que se construye un menor número de detectores más complejos, más costosos y con duración de proyecto más larga. La típica colaboración experimental ha crecido en tamaño de menos de una docena de físicos provenientes de dos o tres instituciones en los años 70 a cientos de físicos de docenas de instituciones en los años 90 hasta las macro-colaboraciones del Large Hadron Collider, que típicamente tienen unos 1.500 colaboradores procedentes de cientos de instituciones. Dado que en el período 1970-presente el número de universidades y centros de investigación por país no ha cambiado significativamente, y el número de investigadores por centro tampoco, el resultado neto es una completa internacionalización de las colaboraciones experimentales. Además hay un fenómeno de descorrelación entre las localidades geográficas de la institución de un físico particular, de las instituciones de sus colaboradores y de sus experimentos.

La comunidad de FAE ha invertido mucho esfuerzo en desarrollos de técnicas Grid. La razón es análoga a la adopción universal de redes compartidas TCP/IP: Al poder "infraestructurizar" parte de los servicios por un lado, y al "universalizarlos" por el otro, se pueden obtener grandes economías de escala. Por ejemplo, no tiene sentido que FAE mantenga su propio sistema de Autoridades de Registro y Certificación, cuando puede ser compartida con otros campos científicos por un lado, y además la mayoría de Universidades y Centros de Investigación también desean dicho sistema para sus propias aplicaciones. En este sentido una política coherente hacia el desarrollo de Infraestructura Grid en España es muy importante para FAE.

Localización	Parámetro	∑capacidad 2006-2008	Incremento anual >2008
CERN	Almacenamiento en disco	5.000 TB	2.500 TB
	Almacenamiento en cinta	20.000 TB	10.000 TB
	Transferencia de cinta	4.000 MB/s	2.000 MB/s
	CPU en cluster	20.000 KSI2000	10.000 KSI2000
	Red agregada del cluster	1.000 Gbps	500 Gbps
	Conectividad externa	80 Gbps	40 Gbps
Conjunto	*Número de centros*	*10*	-
De centros de	Almacenamiento en disco	20.000 TB	10.000 TB
transformación	Almacenamiento en cinta	20.000 TB	10.000 TB
De datos	Transferencia de cinta	1.000 MB/s	500 MB/s
	CPU en cluster	40.000 KSI2000	20.000 KSI2000
	Red agregada del cluster	2.000 Gbps	1.000 Gbps
	Conectividad externa	200 Gbps	100 Gbps
Conjunto	*Número de centros*	*25-40*	*?*
De centros de	Almacenamiento en disco	10.000 TB	5.000 TB
análisis de datos	Almacenamiento en cinta	5.000 TB	3.000 TB
y simulación	Transferencia de cinta	No evaluada	No evaluada
	CPU en cluster	40.000 KSI2000	20.000 KSI2000
	Red agregada del cluster	No evaluada	No evaluada
	Conectividad externa	100 Gbps	50 Gbps

Tabla 1

Además, podría ser el caso (aunque esto queda por probarse), que los recursos computacionales y de almacenamiento se logren utilizar de una manera mas eficiente si se incorporan a Infraestructuras Grid. De hecho, es la automatización de la replicación y gestión automática de complejos subconjuntos de datos lo que es realmente crucial para FAE (y lo que desafortunadamente no ha recibido énfasis y comprensión adecuados por parte de la prensa o por autoridades de política científica, que tienden a enfatizar los aspectos de "supercomputación" del Grid).

Aunque para un Físico de Altas Energías de casi cualquier edad el término "e-Ciencia" resulte un tanto curioso, ya que da por hecho el ser un "e-Físico", es muy probable que

se pueda realizar una beneficiosa transferencia de conocimientos de la comunidad FAE a otras comunidades científicas. En el caso español, la única Infraestructura Grid que opera en modo producción 24x7 es la derivada del banco de pruebas para el LHC, que ahora se está extendiendo a otras ciencias a través del proyecto europeo EGEE. Se considera muy conveniente que acciones de fomento de la "e-Ciencia" en España re-utilicen estas experiencias para maximizar los beneficios a corto y medio plazo.

3.2.1.2 Grupos de Trabajo

Los grupos de Altas Energías españoles están participando en experimentos del LHC y en otros experimentos de otras características (por ejemplo, en experimentos de Astroparticulas – AMS, Antares, Magic- o en experimentos en Estados Unidos y Japón (CDF, BaBar, K2K). A continuación se dan los grupos que participan en experimentos LHC:

- Grupo de Altas Energías del CIEMAT (Madrid) que participa en CMS
- Grupo de Altas Energías del IFCA (Santander) que participa en CMS
- Grupo de Altas Energías de la UAM (Madrid) que participa en ATLAS
- Grupo de Altas Energías del IFAE (Barcelona) que participa en ATLAS
- Grupo de Altas Energías del IFIC (Valencia) que participa en ATLAS
- Grupo de Altas Energías de la UB (Barcelona), que participa en LHCb
- Grupo de Altas Energías de la USC (Santiago de Compostela), que participa en LHCb

Además esta el PIC (Puerto de Información Científica) de Barcelona cuya finalidad es, entre otras, el suministrar servicios GRID a los tres experimentos del LHC (ATLAS, CMS y LHCb).

Por otro lado existen otros grupos de Altas Energías en España que no participan en ningún experimento del LHC :

- Grupo de AAEE de la Universidad de Zaragoza
- Grupo de AAEE de la Universidad Complutense de Madrid
- Grupo de AAEE de la Universidad de Granada

Un aspecto de esencial importancia dentro de las Altas Energías en España es nuestra participación en el CERN del que nuestro país es miembro desde 1984. Podemos decir que la comunidad de físicos de Altas Energías de grupos españoles esta mayoritariamente en colaboraciones o experimentos del CERN (concretamente en experimentos de LHC: ATLAS, CMS y LHCb) , que es el Laboratorio Europeo de Altas Energías que actúa, entre otras cosas, como gran facilidad de suministro de aceleradores y de física de detectores. Debido a la excelencia de este centro, el CERN es también el centro líder de proyectos GRID tales como el DATAGRID, el EGEE, etc.

La participación española en la puesta a punto del sistema GRID internacional para el LHC (que es el llamado Proyecto LHC Computing GRID) está estimada en el 5% del esfuerzo total

Y como se ha comentado, también hay grupos que están participando en proyectos de aceleradores de EEUU y de Japón y están interesados en utilizar los avances en GRID para el procesado de datos. Por ejemplo están las participaciones en CDF y en BaBar que son experimentos que están en fase de toma de datos (running experiments) que pueden servir con banco de pruebas de algunos aspectos de procesado de datos ‘a la GRID’.

3.3 Área de Astrofísica

El objetivo principal de esta sección es destacar la necesidad actual de crear una infraestructura Grid en el ámbito nacional, y más concretamente el gran impulso que esto supondría en el área de Astrofísica.

Inicialmente se analizarán las principales motivaciones para la aplicación de esta tecnología en Astrofísica. Posteriormente se enumerarán los proyectos nacionales e internacionales con interés Grid en marcha y se detallarán posibles "use cases", así como potenciales proyectos pilotos para un programa Grid en Astrofísica.

Finalmente se mencionará la perspectiva de cara al VI Programa Marco (6PM) y los fundamentos básicos para una futura transferencia de tecnología, visibilidad y difusión de proyectos.

3.3.1 Motivación y Necesidades

Aunque las tecnologías Grid nacieron en el ámbito de la física de Altas Energías, rápidamente se expandieron a otras áreas científicas (meteorología, salud, biología, etc.), como fruto de su inmenso potencial y de las nuevas posibilidades que abre en prácticamente cualquier campo científico e industrial.

Concretamente en el ámbito de la Astrofísica se pueden mencionar tres aspectos donde la aparición de esta tecnología ha supuesto una autentica revolución: acceso y tratamiento de archivos astronómicos, solución de problemas numéricos complejos y programas de observación remota.

3.3.1.1 Acceso y tratamiento de archivos astronómicos

La presente y futura generación de telescopios, antenas, satélites y misiones espaciales aseguran la generación de cientos de Gigabytes de datos astrofísicos cada día. Estos cubren distintos tipos de objetos (galaxias, estrellas, atmósferas planetarias, etc.), distintas longitudes de onda (radio, óptico, rayos X, etc.), hasta diferentes fenómenos (explosiones gamma, asterosismología, fenómenos cosmológicos, etc.). Además los detectores de última tecnología nos ofrecen imágenes de cada vez mejor resolución y mayor campo, con el consecuente aumento del "peso informático" de dichos datos (Ej. : el interferómetro ALMA podrá generar señales de hasta 0.01 arcsec de resolución y un volumen de datos de hasta 42Gbytes/hora.)

Toda esta información se almacena en un número cada vez mayor de bases de datos esparcidas por todo el globo, con una enorme variedad de formatos, formas de acceso, y políticas de uso. Son los llamados archivos astronómicos.

Estos archivos se están convirtiendo en la autentica materia prima de la investigación actual del astrónomo, ya sea como fuente de selección para solicitar observaciones más concretas, ya sea para el análisis directo sobre un determinado "dataset" (un 10% del archivo ISO se descarga cada mes). Además su existencia está permitiendo que emerja un nuevo estilo de astrofísica que requiere del manejo, tratamiento y análisis masivo de vastos "datasets" simultáneamente, como por ejemplo, los estudios estadísticos del gran catálogo estelar que generará la misión COROT, donde será necesario comparar cientos o miles de curvas de luz en búsqueda de planetas extrasolares.

Es evidente que este panorama abre un nuevo horizonte de enormes desafíos científicos, pero también exige superar un gran número de limitaciones actuales.

Por un lado, existe un claro problema de almacenamiento. Los archivos crecen de manera ilimitada (el archivo del HST aumenta del orden de 1-2Tbyte/año) y las nuevas misiones en marcha (Herschel, XMM-Newton, etc.) realizarán muestreos del cielo más grandes y completos, que harán que los archivos alcancen volúmenes del orden de los Petabytes en un plazo medio de tiempo.

Por otro lado, la información contenida en estos archivos no es independiente la una de la otra. Cada vez es más necesario consultar varios archivos simultáneamente, por ejemplo, para obtener información de un mismo objeto en diferentes longitudes de onda y que han sido recogidas por distintas misiones. Los astrónomos emplean una enorme cantidad de tiempo en la tediosa y compleja labor de correlacionar la información de diversos archivos. En este sentido, uno de los retos actuales más importantes y ambiciosos de la Astrofísica es el de integrar el mayor número de archivos en un único "observatorio virtual". Este observatorio tendría un acceso único y sencillo, además de contener "agentes" capaces de interoperar entre los diferentes archivos a petición del usuario y de manera transparente para éste.

Por último, hasta la fecha, el "data-mining" de archivos completos de gran volumen ha quedado limitado a un pequeño grupo de especialistas, debido a la necesidad de un alto conocimiento técnico, acceso restringido en los archivos y una alta exigencia de recursos computacionales.

Es obvio que todas estas limitaciones exigen nuevos planteamientos tecnológicos.

3.3.1.2 Solución de problemas numéricos complejos

La Astrofísica no es una ciencia exclusivamente observacional, y necesita de modelos teóricos que puedan ser contrastados con las observaciones. Además cuanto mayor y más precisa es la información que contienen éstas, más exigentes y complejas son las revisiones de dichos modelos, y más elementos debemos incluir en ellos. Por ejemplo, es necesario introducir la presencia de campos magnéticos y crear modelos magneto-hidrodinámicos, que a su vez se acoplen con modelos de emisión, para así poder reproducir con la mayor exactitud posible los "jets" relativistas en numerosas galaxias activas.

Parejo al aumento de la complejidad física de estos modelos, se encuentra la de su resolución que ha implicado el desarrollo de nuevos métodos numéricos más potentes – refinamiento adaptativo de malla, técnicas de Fourier, modelos de mezcla gaussianos, etc. – y por supuesto una cada vez mayor exigencia en recursos computacionales. En este sentido la aparición de las tecnologías de paralelización y de intercambio de mensajes han permitido el desarrollo de "clusters" de decenas a miles de nodos, dedicados exclusivamente a la resolución de modelos numéricos. Pero, incluso las configuraciones de cálculo más potentes hoy en día quedan limitadas ante la complejidad de muchos de los problemas existentes, que demandan un salto cualitativo en la tecnología de cómputo actual.

3.3.1.3 Observación remota

Otro aspecto donde la astronomía actual esta sufriendo importantes cambios es en el terreno de la observación remota. Los últimos avances en la tecnología de las comunicaciones, y el gran aumento de ancho de banda en las redes científicas, permiten el desarrollo de los llamados observatorios robóticos, cuyos telescopios e instrumentos pueden ser manejados remotamente por los observadores desde sus centros de investigación. Esto evita el desplazamiento físico al observatorio, lo que reduce los costes de construcción (no es necesario construir áreas de servicios para el observador)

y produce un mayor aprovechamiento de las horas útiles de observación. Además, desde un único centro se pueden realizar observaciones coordinadas entre varios telescopios situados a diferentes latitudes, con el objeto de monitorizar, incluso 24 horas al día, distintos objetos. Esto resulta enormemente beneficioso, por ejemplo, en el caso del estudio de la pulsación estelar, donde es necesario disponer de observaciones con una base de tiempo lo más grande posible. Por último, una red de telescopios robóticos ofrece a la comunidad un sistema de respuesta rápida para la observación de fenómenos transitorios como novas, supernovas o explosiones de rayos gamma.

Es evidente, a la vista de lo expuesto, que la Astrofísica se encuentra en un importante punto de inflexión, donde se enfrenta a fascinantes retos científicos que exigen superar toda una serie de importantes limitaciones tecnológicas.

3.3.1.4 Justificación de la necesidad de uso y desarrollo de la tecnología Grid

Al igual que en otros muchos entornos, la implantación de la tecnología Grid en el ámbito de la Astrofísica ha supuesto una autentica revolución que ha abierto la posibilidad de enfrentarse a muchos retos inabordables hasta la fecha. Además, la comunidad astrofísica ofrece un entorno colaborativo (observatorios, centros de investigación, archivos astronómicos, etc.) que se ajusta perfectamente a la filosofía Grid.

Entre las ventajas que el uso de la tecnología Grid supone en Astrofísica se pueden destacar:

- El manejo y acceso remoto a los grandes volúmenes que conforman los archivos astronómicos requiere de sofisticados sistemas de almacenamiento (configuraciones RAID, entornos SAN, etc.), altos niveles de transferencia I/O, redes con gran ancho de banda, construcción de bases de datos bien indexadas, y sobre todo, de un personal cualificado y especializado que mantenga y administre estos sistemas. Un entorno Grid implica la existencia de una serie de centros especializados interconectados en una red de alta velocidad y con los recursos necesarios, tanto humanos como informáticos, para ofrecer servicio de acceso a estos archivos al resto de centros usuarios.
- Un tejido Grid de datos (ver definición en la sección dedicada al área temática de Middleware) conforma la base perfecta para mantener un “Observatorio Virtual” formado por diferentes archivos distribuidos geográficamente, interrelacionados entre sí y accesibles desde un único punto común.
- En este sentido, la existencia de un “middleware” específico para Astrofísica permite la construcción de herramientas básicas (navegación, búsquedas, presentación de información, interfases, etc.), que de una manera transparente para el usuario correlacionen la información de los diferentes archivos, mostrándola con un formato único y estándar.
- Además, la existencia de este “middleware” astrofísico permite pensar en más sofisticados niveles de explotación de los archivos – visualización de grandes volúmenes de datos, selección interactiva de “datasets”, herramientas de análisis estadístico, herramientas de análisis masivo de fotometría y espectroscopia, etc. –, o incluso en el desarrollo de herramientas capaces de buscar entre todos los archivos, determinados patrones de selección – “discovery tools” -, fundamentales para detección de objetos exóticos, eventos transitorios, etc.
- Una estructura Grid implica que todas estas herramientas se ejecuten de modo remoto y distribuido. Esto abre la posibilidad a que grupos de investigación

puedan realizar estudios que de otra manera exigirían una gran inversión en tiempo y en recursos computacionales y humanos.

- Los problemas numéricos actuales – modelos cosmológicos, relatividad numérica, interacciones galácticas, etc.– y las técnicas empleadas – análisis de Fourier, mallados adaptativos, etc.– requieren de grandes exigencias computacionales – clusters de cientos de nodos, máquinas de memoria distribuida, máquinas vectoriales, etc. – equipos muy costosos, y en algunos casos insuficientes. La tecnología Grid supone alcanzar una nueva escala de capacidad computacional y la posibilidad de enfrentarse a problemas numéricos complejos hasta ahora inabordables.
- Además, un entorno Grid conlleva una red de centros especializados ofreciendo un uso colectivo de recursos computacionales de manera remota, a bajo coste, y con una única interfaz, transparente para el usuario y capaz de adaptarse a sus necesidades.
- Promueve la creación de una comunidad para compartir y desarrollar proyectos científicos entre diversos centros, basándose en el concepto Grid de Organización Virtual.
- Mantiene en todo momento los criterios de seguridad y autenticación que la explotación remota de los recursos exige.
- De igual forma que un tejido Grid integra recursos computacionales y de almacenamiento, también puede integrar diferentes tipos de dispositivos, como puedan ser telescopios e instrumentos de observación, distribuidos geográficamente. Por tanto un Grid puede conformar la estructura ideal para una red de observatorios robóticos diseñados para observación remota.

3.3.1.5 Estudio de necesidades y proyectos de la comunidad.

La creación de un programa de e-ciencia nacional implica una importante inversión a diferentes niveles.

Siguiendo la arquitectura de capas planteada para este programa de e-ciencia nacional, el primer requerimiento para una infraestructura Grid es disponer de un conjunto de centros conectados en una red de alta velocidad. En este sentido la red académica nacional (RedIRIS) conforma un magnífico punto de partida al ofrecer anchos de banda que van desde los 155mbps a los 2.5Gbps.

Por otro lado, los centros que conformen este tejido deben estar provistos de altos recursos computacionales para poder ofrecer un servicio Grid de calidad. En consecuencia, es necesario realizar una importante inversión, probablemente llevada a cabo en varias fases, y con el objetivo de crear una infraestructura computacional potente y distribuida geográficamente por todo el estado español.

Sobre esta infraestructura debe descansar un “middleware” básico que ofrezca un punto de acceso común a las distintas áreas específicas, y que aproveche de forma efectiva los recursos. Actualmente existen varios centros con experiencia en desarrollo y mantenimiento de este tipo de “middleware”, pero se necesitará apoyo para diseminar este conocimiento entre los posibles nuevos centros especializados, a través de estancias, colaboraciones y cursos, y a su vez para mantener y reforzar las líneas de desarrollo existentes.

El acceso temático a los recursos Grid (Astrofísica, Altas Energías, Salud, etc.) se realizará a través de organizaciones virtuales (VO). Estas organizaciones serán las encargadas de desarrollar, mantener y promover el “middleware” y aplicaciones

específicas de cada área, y en concreto en el área de Astrofísica. En consecuencia, será necesario dotar a estas organizaciones de personal que desarrolle estas tareas, así como las de mantenimiento, soporte y acceso a los recursos Grid en cada área.

3.3.2 Casos de Uso y Proyectos Pilotos

Dentro de las primeras iniciativas de IRIS-GRID se encontrará el diseño de unos "use-cases" a partir de una estrecha interacción con la comunidad astrofísica, y que servirán de directrices para el desarrollo de futuros proyectos dentro del Grid nacional.

Sin entrar en detalles, dos posibles "use cases" pueden tratar sobre la resolución de problemas complejos en Astrofísica y el desarrollo de un observatorio Virtual Nacional.

3.3.2.1 Resolución de problemas numéricos complejos en Astrofísica

Como se ha mencionado anteriormente, la Astrofísica actual se enfrenta a problemas de una Física muy compleja y que requieren de tratamientos numéricos elaborados y con un alto coste computacional.

Muchos de ellos son revisiones más completas de problemas conocidos, como puedan ser refinamientos en los modelos de interior estelar, donde se incluye transporte convectivo, lo que exige la introducción de métodos numéricos no lineales; o modelos de evolución galáctica con síntesis de formación estelar más complejas, o simulaciones de interacción entre galaxias con condiciones iniciales y de contorno más próximas a las observadas, etc. Por otro lado, existe toda una línea de investigación denominada Relatividad Numérica, basada en el tratamiento numérico de la teoría de la Relatividad General en Astrofísica, y que sirve de base para la elaboración de complejos modelos cosmológicos sobre el origen y evolución del Universo. Las misiones en marcha, como Planck, dan muestra del enorme interés en este campo existente en la comunidad científica.

Enfrentarse a estos problemas exige de una altísima capacidad de cálculo y de recursos informáticos y humanos tan sólo al alcance de unos pocos centros especializados, debido a su alto coste económico, y que en bastantes problemas concretos es, a día de hoy, todavía insuficiente. Como ejemplo, una simulación de una colisión de dos estrellas de neutrones, un posible evento generador de ondas gravitatorias, requiere de más de 1024 Gigabytes de memoria y un pico de 100000 Gigaflops.

La creación de un tejido Grid que aglutine toda una serie de recursos distribuidos geográficamente para dar servicio a la comunidad astrofísica, no sólo aumentaría la capacidad computacional en varios órdenes de magnitud, sino que además conllevaría toda una serie de ventajas que se pueden enumerar en:

- Promover la creación de una comunidad para compartir y desarrollar códigos numéricos y resultados científicos, a través del concepto de Organización Virtual.
- Ofrecer un acceso transparente y remoto a una red de recursos (clusters, almacenamiento, etc.) distribuidos geográficamente para la ejecución de códigos y simulaciones numéricas de alta exigencia computacional, que puede llegar al orden de los Teraflops en potencia de calculo, y de los Terabytes en capacidad de almacenamiento.
- Estimular el desarrollo de servicios y aplicaciones dentro de un entorno Grid en el área de Astrofísica, tales como el desarrollo de códigos de simulación para resolución de problemas concretos, o de aplicaciones Grid avanzadas (visualización, estadística, etc.) para uso general de la comunidad astrofísica.

- Permitir la interacción remota con los recursos de cálculo, de forma que los centros usuarios puedan instalar, compilar, chequear, ejecutar e interactuar con sus propios códigos.
- Mantener en todo momento los criterios de autenticación, seguridad y fiabilidad durante el proceso.
- Permitir el acceso a los clientes del área a los resultados totales, parciales o colaterales (animaciones, gráficas, etc.) generados durante la ejecución de la simulación.

Así pues, la idea básica es que una serie de centros especializados y con altas prestaciones computacionales y humanas conformen una organización virtual capaz de ofrecer, no solamente una inmensa capacidad de cálculo, sino también el desarrollo de una serie de aplicaciones Grid enfocadas a la resolución de simulaciones numéricas de interés en Astrofísica, y que puedan ser explotadas por cualquier centro usuario. De igual manera, debe ofrecer a cualquier grupo científico la posibilidad de poder desarrollar y ejecutar sus propios códigos numéricos en dicho entorno Grid.

Como ejemplo de la potencia de cálculo de un entorno Grid, se lanzaron simulaciones de Astrofísica Relativista en un grid formado por maquinas del SDSC (USA), NCSA (USA) y el Max Planck (Germany), hasta alcanzar más de 1500 procesadores unidos entre ambos continentes por una red de 622mbps. Se utilizó Cactus y una adaptación Grid de las librerías de paralelización MPICH (MPICH-G2), todo ello sobre Globus. El resultado fue mejorar en un 70% el rendimiento de ejecución habitual.

3.3.2.2 Desarrollo de un observatorio Virtual Nacional

Un observatorio virtual es un conjunto de archivos astronómicos conectados e interrelacionados a través de una red de alta velocidad, y de herramientas de software para la explotación y análisis de dichos archivos por parte de la comunidad científica.

Manteniendo la analogía con uno real, en un observatorio virtual los “telescopios” estarían representados por la red de archivos, mientras que los “instrumentos” sería la colección de aplicaciones software empleadas en la petición y análisis de los datos, sin que sea necesario que estos se muevan físicamente del archivo, y respondiendo a las solicitudes realizadas por los “observadores”.

Un Observatorio Virtual Nacional estaría conformado por archivos astronómicos desarrollados y/o mantenidos por grupos de investigación propios de un país.

Entre las numerosas ventajas que ofrecen los Observatorios Virtuales, se pueden enumerar:

- Los datos pueden ser explotados múltiples veces y por diferentes grupos de investigación. Además, son almacenados de una manera controlada y con una interfaz uniforme, lo que asegura su utilización por largo tiempo.
- Permite el acceso y explotación de los archivos a instituciones que de otra manera no podrían, por carecer de recursos económicos suficientes.
- Evita al astrónomo tener que correlacionar diferentes archivos cuando, por ejemplo, el estudio implique datos en diferentes longitudes de onda.
- Facilita el análisis masivo y simultaneo de diferentes archivos, en busca de correlaciones físicas entre los datos, búsqueda de nuevos objetos, estudios estadísticos, etc.
- Asegura los requerimientos de autenticación y confidencialidad del acceso a los archivos.

Genera un entorno colaborativo de servicio, etc.

Las necesidades de almacenamiento, cómputo y acceso que exige la existencia de un observatorio virtual y su explotación por parte de la comunidad, ha hecho que su formación haya ido paralela a la aparición y desarrollo de las tecnologías Grid, hasta el punto que la práctica totalidad de los Observatorios Virtuales nacionales e internacionales que están en marcha o en formación se sustentan sobre un grid de datos.

Las ventajas de un Grid en esta materia son obvias:

- Ofrece una red de centros especializados distribuidos geográficamente con los suficientes y escalables recursos de almacenamiento, cálculo y humanos para el mantenimiento de los archivos astronómicos.
- Implica la existencia de un "middleware" común donde pueden construirse las aplicaciones necesarias para la interacción con los datos.
- Su alta capacidad computacional permite abordar programas que exijan procesado y análisis masivo de diferentes archivos correlacionados, y que de modo individual serían imposibles de realizar por cada centro.

Así pues, un primer y fundamental paso para la formación de un Observatorio Virtual Nacional, que pueda en el futuro integrarse en las iniciativas internacionales abiertas como AVO o IVOA, es el de generar una infraestructura Grid adecuada.

3.3.2.3 Definición de posibles proyectos piloto

El objetivo básico de un programa Grid en el área de Astrofísica consiste en la construcción del "middleware" específico para dicho área. En esta línea, y como paso inicial, habría que interactuar con la comunidad científica y realizar un detallado estudio de "use cases" concretos. Este estudio serviría de guía para la definición de los proyectos piloto iniciales. Una lista de potenciales aplicaciones piloto podría ser:

Investigar las posibilidades de adaptación a la arquitectura Grid de software de uso común en Astrofísica (IDL, GAIA, IRAF, STARLINK, MIDAS, AIPS, FTOOLS, etc.).

Construir o adaptar herramientas de visualización y manipulación de datos astrofísicos.

Estudio piloto de la integración y correlación de diferentes archivos astrofísicos en un entorno Grid.

Construcción de herramientas de acceso, búsqueda, navegación y presentación de resultados de un conjunto de archivos astronómicos integrados.

Construcción o adaptación de librerías básicas de cálculo numérico para su empleo en un entorno Grid, etc.

3.3.3 Proyectos Grid en Marcha.

Como anexo a este documento se adjunta una lista detallada de algunos de los grupos nacionales de Astrofísica que suscriben el interés en la formación de un programa de e-ciencia nacional.

3.3.3.1 Proyectos Grid en marcha en el área de Astrofísica.

En la actualidad existen numerosos proyectos Grid en marcha y que abarcan distintas áreas de investigación. En USA podemos citar proyectos como **PDG** y **GriPhyN** (Física de Partículas), **DOE ScienceGrid, Earth System Grid** (Meteorología), **NEESGrid** (Sismología), **Fusion Collaboratory** (Fusión nuclear), así como el **International**

Virtual Data Grid Laboratory (iVDGL), y el proyecto **TeraGrid** que pretende unir cuatro centros de supercomputación a 40 Gbps.

Por parte europea, en el año 2000 se lanzó el proyecto **European DataGrid (EDG)** coordinado por el CERN y que desarrolla nuevo *"middleware"* para construcción de aplicaciones para tratamiento de grandes volúmenes de datos en las áreas de Física de Partículas, de Bioinformática, y de Observación de la Tierra. Este proyecto cuenta con participación española, concretamente en el desarrollo de un testbed distribuido por toda Europa, a través de los grupos de Física de Altas Energías: IFAE, IFIC, IFCA, Universidad de Oviedo, UAM y CIEMAT. Otros proyectos Europeos son **GridLab**, **DataTag** y **CrossGrid**, este último también cuenta con una alta participación española en centros como RedIris, IFCA, IFIC, UAB, USC, CESGA y la Universidad de la Coruña.

Más detalladamente, y con implicaciones en el área de Astrofísica podemos mencionar:

- ***Astrophysical Virtual Observatory (AVO):*** Actualmente en fase A y consistente en un estudio para el diseño de un observatorio virtual para la comunidad astronómica europea basándose en el paradigma Grid. Los principales participantes son: ESO, ST-ECF, Consorcio ASTROGrid, CNRS y el Jodrell Bank Observatory (Victoria University of Manchester)
- ***US National Virtual Observatory*** **:** Financiado por el US National Science Foundation y cuyo objetivo, entre otros, es la realización de una serie de prototipos para demostrar el interés y eficiencia del uso de las tecnologías Grid en la construcción de un observatorio virtual. Estos tres prototipos ya implementados son:
    - Búsqueda de gamma-ray burst.
    - Búsqueda de candidatas a enanas marrones.
    - Servicio de análisis de morfología galáctica.
- ***AstroGrid*****:** Proyecto de e-ciencia en Reino Unido cuya principal motivación es la de construir un tejido Grid para la interconexión de los archivos astronómicos de cinco misiones con participación inglesa: WFCAM, VISTA, XMM-SSC, e-MERLIN, SOHO y Cluster.
- ***The International Virtual Observatory Alliance (IVOA):*** Fusión de los tres proyectos anteriores con el objetivo de conformar un observatorio virtual mundial. A este consorcio se están uniendo otros proyectos de observatorios virtuales nacionales (Ej. *Australian Virtual Observatory (Aus-VO), Japanese Virtual Observatory (JVO), IDGAR (Italia), etc.).*
- ***European Grid of Solar Observations (EGSO):*** Consiste en un testbed Grid para la formación de un observatorio virtual solar, financiado bajo el programa comunitario IST del 5PM y con tres años de duración.
- ***eStar:*** Un proyecto realizado por las Universidades de Liverpool y Exeter. Desarrollar infraestructura software para la formación de una red de telescopios robóticos que realicen monitorizaciones de 24 horas, y de agentes de búsqueda inteligentes. Esta red utiliza *Globus* como *"middleware"* básico.
- ***iAstro:*** Una *"COST action"* lanzada en el 2001 con la motivación de asegurar la calidad de aplicaciones Grid en el ámbito de la Astrofísica. Dentro de iAstro se encuentra el proyecto **CABGrid**, dirigido por el CAB y la UCM, y con el objetivo de generar un laboratorio virtual para Astrobiología tomando *Globus* como punto de partida.

- ***GridLab :*** IST del 5PM que está realizando todo un conjunto de aplicaciones Grid (GAT) para ejecución de códigos de simulación numérica en escenarios de Relatividad Numérica y detección y análisis de ondas gravitacionales. Estas aplicaciones están basadas en *Cactus*, una aplicación *"open source"* para cálculo numérico en Grid. Varios centros científicos y empresas, como Sun Microsystems, participan en el proyecto.

A tenor de estos proyectos se hace palpable el enorme interés mundial que suscitan las tecnologías Grid, y su aplicación al mundo de la ciencia y de la industria.

3.3.3.2 Perspectivas de participación en el 6PM.

Continuando con lo ya establecido en el 5PM, la Comisión Europea sigue apostando fuertemente por el desarrollo y uso de la tecnología Grid, como lo demuestra que en el 6PM existan tres líneas prioritarias donde se promueve explícitamente el uso de esta tecnología. Concretamente en Information Society Technology (IST), donde aparece en dos apartados, Complex Problem Solving y e-Health, en Life Sciencies, Genomics and Biotechnology for Health, y en Research Infraestructures del programa Structuring the European Research Area.

Dentro de este marco, existen numerosas propuestas con implicación Grid, pero en concreto una con participación española, Enabling Grids for e-Science and Industry in Europe (EGEE), lleva asociado un "workpackage" con aplicación en el área de Astrofísica, concretamente para aplicaciones Grid en simulaciones numéricas del fondo de microondas.

En este sentido, es fundamental de cara a futuros proyectos europeos, la formación de una infraestructura Grid nacional, como la que representa IRIS-GRID.

3.3.4 Transferencia de tecnología, visibilidad y difusión de los proyectos.

La implantación de un programa de e-ciencia nacional y con conexión Europea implica grandes beneficios tanto para el mundo científico, como para el empresarial.

Un grid es un enorme sistema de cálculo distribuido geográficamente y que puede ofrecer a la comunidad una potencia de cálculo de pico del orden de los Teraflops, así como una capacidad de almacenamiento del orden de los Petabytes.

- Esta infraestructura permite que grupos de investigación, así como pequeñas y medianas empresas puedan abarcar proyectos que de otra manera les exigiría una inversión en infraestructura generalmente inabordable.
- Además la existencia y funcionamiento de esta infraestructura genera toda una serie de fuentes de financiación inerciales a distintos niveles: institucionales, locales, nacionales, europeas.
- Por estos motivos es necesario promover el uso de estas tecnologías, dando a conocer al mayor número de potenciales centros y empresas usuarias las ventajas de su aplicación. Este proceso de diseminación y transferencia tecnológica debe realizarse a través de:
- Encuentros con empresas con el objetivo de fomentar la colaboración y el desarrollo conjunto de proyectos relacionados con las tecnologías Grid. En el área de astrofísica en concreto, pueden ejecutarse programas conjuntos entre empresas y centros de investigación relacionados con la fabricación y puesta a punto de instrumental de alta tecnología para telescopios y misiones espaciales.

- Participación y organización de “workshops” nacionales e internacionales para intercambio de ideas sobre los distintos aspectos de las tecnologías Grid y de e-ciencia.
- Construcción y mantenimiento de un portal Web único, donde se concentre toda la información sobre IRIS-Grid. Desde documentación de uso, proyectos en marcha, material de difusión, monitorización de aplicaciones, etc.
- Motivar a través de seminarios, estancias y cursos en centros de investigación el uso del Grid nacional.

3.4 Área de Salud

Este documento pretende fomentar la discusión en la preparación de la propuesta para el área temática de la salud en el programa de e-Ciencia de RedIRIS. El documento constituye un primer borrador de los puntos a destacar en la propuesta final.

3.4.1 Motivación y Necesidades

3.4.1.1 Situación de la TIC en el área de la salud

La informatización de los servicios asistenciales es un proceso complejo y lento pero inexorable. Numerosos servicios se encuentran en la actualidad total o parcialmente informatizados (administración, laboratorio, radiodiagnóstico,…) en gran parte de la sanidad pública y privada española. La necesidad de disponer de un sistema seguro, ágil, robusto y eficiente para el almacenamiento, proceso y transmisión de información relacionada con la salud es absolutamente aceptada por toda la comunidad usuaria, científica y empresarial.

Sin embargo, existen numerosas barreras que limitan la velocidad con la que se están adaptando los sistemas, más aún si se compara con otros procesos del mundo empresarial. La naturaleza multimedia de la información de la salud (imágenes, texto, señales, vídeos, procesos, etc.), la complejidad de su tratamiento (proceso de imágenes, análisis de señales, extracción de información en textos, etc.) y su dificultad de transmisión (numerosos formatos incompatibles, gran volumen, etc.) constituyen un factor limitador técnico importante. Más aún, otros factores, como la privacidad de la información, la disponibilidad de los recursos o el manejo del enorme volumen de datos históricos, suponen un freno aún mayor.

Las tecnologías GRID suponen una importante oportunidad para la solución de numerosos de estos problemas.

3.4.1.2 Justificación de GRID como solución

El mundo de la salud constituye un ejemplo perfecto para la implantación de un entorno GRID. La existencia de una comunidad virtual de personal médico proveniente de varios centros (asistencia primaria, especializada, urgencias) que cubren un área asistencial se adapta a la estructura GRID. La información de los pacientes se encuentra distribuida y su acceso y proceso como un conjunto global es absolutamente deseable. Además, el acceso y proceso al gran volumen de datos médicos puede ser abordable desde un conjunto de recursos GRID.

La seguridad, autenticación y fiabilidad que ofrece la arquitectura GRID responde a los requerimientos que el proceso de este tipo de información necesita.

3.4.1.3 e-Ciencia en Salud

La madurez actual de los entornos GRID es relativa. Si bien estos entornos se encuentran en un estado admisible para la comunidad científica, aún existen numerosos conceptos que deben solucionarse para el despliegue del GRID en salud a nivel generalizado.

Por una parte, se necesita adaptar las aplicaciones para que funcionen en un entorno GRID, con un énfasis especial en la seguridad y confidencialidad de la información. Muchas aplicaciones GRID en salud compartirán similares requerimientos, por lo que

es importante definir una capa de 'GRID en salud' que ofrezca la funcionalidad común a las aplicaciones.

Además existen problemas de índole tecnológico, como los requerimientos de infraestructura local en los centros asistenciales que necesitan ser analizados detenidamente.

Por tanto, la aplicación de GRID en salud requiere de un importante esfuerzo investigador, y este proyecto pretende ofrecer el entorno adecuado para que éste se desarrolle.

3.4.1.4 Estudio de Necesidades y Proyectos por Regiones

La I+D+I de las TIC en el ámbito de la salud implica a una gran variedad de entes de naturaleza diversa, como centros tecnológicos, usuarios médicos y empresas proveedoras. Es por tanto importante que el ámbito de este proyecto transcienda de los centros de investigación y abarque a todos entes implicados en la medida de lo posible.

Los centros hospitalarios, usuarios de la tecnología, son a la vez grandes proveedores de contenido. Esto puede implicar que se requiera analizar en determinados casos la conexión entre las redes informáticas hospitalarias y Red IRIS. En el caso de redes hospitalarias públicas, existen importantes dotaciones de infraestructura muy apropiadas para este proyecto (p.e. red ARTERIAS de la Comunidad Valenciana). En el caso de redes hospitalarias privadas, existe una importante tendencia a la conexión de alta velocidad entre los centros pertenecientes a grupos asistenciales privados (como el grupo NISA).

El diseño, planificación e implementación de un Middleware básico para aplicaciones de GRID en salud requerirá disponer de gran experiencia en aplicaciones biomédicas y de aplicaciones GRID. Es necesario por tanto identificar centros y experiencias en el área biomédica que puedan proporcionar la información que se necesita, al mismo tiempo que coordinarse con otras iniciativas similares que puedan estar desplegándose en otros países.

3.4.1.4.1 Proyectos en el Área de la Salud con Interés en GRID

La comunidad investigadora de las TIC en salud es muy numerosa en España. Recientemente, y para coordinar las acciones de investigación, se han puesto en marcha diversas redes de excelencia bajo el Fondo de Investigación Sanitaria del Instituto de Salud Carlos III que pretenden facilitar la cooperación entre los centros especialistas en proceso de imágenes, telemedicina o tecnología sanitaria.

Además de estas acciones, existen numerosas iniciativas cuya cooperación puede proporcionar una gran cantidad de aplicaciones susceptibles de beneficiarse de las tecnologías GRID (aplicaciones computacionalmente intensivas, con grandes requerimientos de almacenamiento de datos, colaborativas, etc.).

3.4.1.4.2 Proyectos en GRID

La definición de un middleware de GRID en salud requiere, además de la experiencia relativa a las aplicaciones de las TIC en salud, el conocimiento de los middleware GRID genéricos. Los entornos actuales no están absolutamente maduros, y se espera que evolucionen en los próximos años notablemente. Este carácter dinámico de la tecnología básica debe ser tenido en cuenta para asegurar que los desarrollos que se realicen durante el proyecto sean de aplicación tanto al principio como al final del proyecto.

Este conocimiento requerirá del contacto con los comités de desarrollo y estandarización y los proyectos de middleware genéricos existentes.

3.4.1.4.3 Perspectivas de participación en el 6PM

Si bien el proyecto pretende la creación de un GRID en salud a nivel nacional, las características intrínsecas de la tecnología GRID permiten su extensión a áreas de aplicación más grandes, como el entorno europeo.

La apuesta que la Comisión Europea está realizando sobre las tecnologías GRID en el VI Programa Marco es notable. No sólo existen líneas prioritarias para proyectos específicos de GRID, sino que en muchas de otras áreas hay una recomendación del uso de estas tecnologías. Más precisamente, la línea prioritaria de 'e-Health' recomienda el uso de esta tecnología en la medida de lo posible.

Existen numerosas acciones en el VI Programa Marco cuya conexión parece necesaria. Existen varias propuestas de infraestructura (EGEE por ejemplo) cuyo objetivo es desplegar una infraestructura de GRID y avanzar en el desarrollo del middleware, con interés en las áreas de Física de Altas Energías y Salud. Por otro lado, existen otras propuestas de redes de excelencia en el área de GRID aplicado a la salud, como la Propuesta de Red de Excelencia 'HEAlth GRID VENture' (HEAVEN). Es importante destacar que tanto las propuestas de infraestructura con en la red de excelencia HEAVEN tienen entre sus objetivos el contacto con desarrolladores de aplicaciones GRID en varias áreas, como la salud.

3.4.2 Casos de Uso

El área de la Salud, como se ha comentado, constituye un área multidisciplinar en la que la información es de carácter multimedia y los procesos que sufre son variados y complejos. Es por tanto necesario antes de describir los casos de uso, el definir con una mínima precisión los tipos de datos y procesos que intervienen de forma genérica (y sin pretender ser exhaustivos) en el área.

3.4.2.1 Tipos de Datos

Los datos relativos a la salud pueden estructurarse en función del objetivo de las aplicaciones. Atendiendo a éste objetivo, se pueden clasificar en tres niveles: población, paciente y órgano. Si bien se consideran más niveles inferiores (celular y biomolecular), éstos se caracterizan por una problemática más específica y se tratarán en el área de bio-computación.

Datos a nivel población

El interés de los datos poblacionales es extraer conocimiento que corresponda de forma genérica a grandes grupos de pacientes, con el objetivo de diagnosticar, planificar una terapia o definir las características endémicas de una determinada población.

La información relevante a éste nivel la constituyen conjuntos que resumen o totalizan la información de los pacientes representativos. No son relevantes a este nivel los datos que caracterizan de forma única la información (datos personales), si bien lo son los datos que permiten su agrupamiento (área de residencia, sexo, rango de edad). A nivel médico, es relevante disponer de información sobre patologías y terapias, así como de los datos que han conducido a un diagnóstico (imágenes de radiodiagnóstico, señales vitales, análisis).

Datos a nivel paciente

La información relevante a nivel de paciente es la contenida en la historia clínica. La historia clínica informatizada constituye un gran reto en la gestión hospitalaria actual. Numerosos esfuerzos se han realizado en la interconexión de datos médicos (HL7, DICOM, Vital, etc.).

La información relativa a los pacientes adolece de los siguientes problemas:

- Se encuentra distribuida, fruto de la visita a lo largo de la vida de un paciente a diferentes centros de atención sanitaria.
- Es confidencial, por lo que su acceso debe estar restringido a nivel electrónico, al menos al mismo nivel que a nivel impreso.
- Tiene un gran volumen, por lo que su transferencia considerarse con cuidado para evitar la congestión del tráfico.
- Es multimedia, por lo que su almacenamiento requiere de un tratamiento especial.

Son datos a este nivel:

- Información demográfica: Datos personales generales y médicos del paciente.
- Información radiológica: Imágenes radiológicas, medicina nuclear.
- Historial clínico: Episodios, diagnósticos, tratamientos, alergias, intolerancias.
- Información analítica: Análisis de muestras, medidas bioeléctricas.

Datos a nivel órgano

La información relevante al nivel de órgano la constituye aquella que refleja el estado o caracteriza a un órgano o sistema funcional. Esta información puede ser parte de la información del paciente o bien constituir modelos y parámetros funcionales de órganos.

La información a nivel de órgano puede considerarse:

- Imágenes anatómicas y funcionales de órganos.
- Señales y muestras.
- Modelos computacionales del comportamiento funcional de órganos-

3.4.2.2 Tipos de Procesos

Dada la naturaleza de los datos que intervienen en el GRID en salud expuesta anteriormente, es importante describir los procesos que intervendrán en su gestión para poder definir más claramente los casos de uso que aparecen.

Procesos que requieren computación de altas prestaciones

La simulación de los modelos funcionales de órganos o los modelos de comportamiento de pacientes y poblaciones requiere de una gran cantidad de recursos dedicados a la resolución conjunta de un único problema. Ejemplos de este tipo de procesos son: la simulación multimodal de la actividad funcional de sistemas orgánicos, como el vascular o el respiratorio; la identificación de unidades anatómicas o funcionales en imágenes 3D; la identificación de arritmias y la localización precisa de los focos en señales cardíacas; etc.

Procesos que requieren de alta productividad

El carácter experimental de la investigación en salud requiere en muchos casos realizar numerosas simulaciones variando un conjunto de parámetros. De los resultados de la experimentación se ajustan experimentalmente tratamientos o diagnósticos.

El GRID puede proporcionar una gran cantidad de recursos para la ejecución eficiente, robusta y simultánea de diferentes experimentos.

Procesos que requieren el acceso a grandes volúmenes de datos

El proceso de las bases de datos médicas es importante tanto para la consolidación de la información perteneciente a poblaciones, la integración de toda la información perteneciente a un paciente concreto o la extracción de conocimiento a partir de técnicas de minería de datos. Es importante permitir la búsqueda de información a partir de conceptos complejos y no indexados, como similitud de imágenes, muestras, diagnósticos, etc.

Procesos de colaborativos o de comunicación igual a igual.

La información médica relativa a pacientes concretos se almacena de forma organizada en los centros asistenciales. Esta información está sujeta a una serie de normas que garantizan la confidencialidad e integridad de los datos.

Sin embargo, la investigación en cualquier área de salud requiere disponer de casos representativos de una determinada patología o colectivo. Estos casos pueden anonimizarse de forma que sea irreconocible su origen, y su compartición puede ser de gran utilidad a la comunidad científica. Si bien estos casos podrían hacerse disponibles a partir de las grandes bases de datos distribuidas, es habitual que los facultativos dispongan en repositorios locales colecciones de datos seleccionados por su interés.

La posibilidad de compartir de manera eficiente y segura estos datos a través de Internet permitiría facilitar la colaboración entre especialistas y el acceso a una mayor cantidad de datos de calidad. La comunicación mediante arquitecturas P2P resulta especialmente apropiada, ya que se aumenta el número de réplicas, se reducen los tiempos de descarga y se trabaja sobre datos preseleccionados.

3.4.3 Middleware Actual y Previsto

La tecnología GRID básica en la que se soporta el middleware específico de la salud deberá elegirse teniendo en cuenta su eficiencia y aceptación. La tecnología GRID todavía no se encuentra en su total madurez y por tanto la capa de servicios específica de la salud deberá ser lo suficientemente independiente para poder adaptarse a cambios en la estructura básica del Middleware.

Actualmente, OGSA (Open Gris Service Architecture) proporciona una estructura suficientemente conceptualizada. La decisión deberá tomarse teniendo en cuenta las iniciativas europeas con las que se pretenda establecer conexiones.

3.4.3.1 Desarrollo Previsto de Middleware Específico del Área

La puesta en marcha de aplicaciones piloto permitirá identificar requerimientos, componentes y procesos que sean comunes a gran número de aplicaciones de GRID en salud. Estos componentes y requerimientos constituyen la capa intermedia que permitirá desarrollar más rápidamente las aplicaciones GRID en el área de la salud.

Entre estos componentes, se encontrarán:

- Módulos de interfaz, capaces de cargar, transmitir y convertir los datos provenientes de diferentes fuentes en diferentes formatos.
- Módulos de proceso de imágenes, capaces del filtrado, segmentación, proyección 3D, reformateo de planos, etc.

- Módulos de cálculo intensivo, capaces de resolver problemas numéricos específicos, como análisis mecánicos de prótesis, dinámica de fluidos, transmisión de potenciales eléctricos, etc.

La ilustración 1 muestra la relación entre las aplicaciones en salud y el GRID genérico. Esta ilustración muestra como ejemplo tres potenciales componentes de esta capa intermedia.

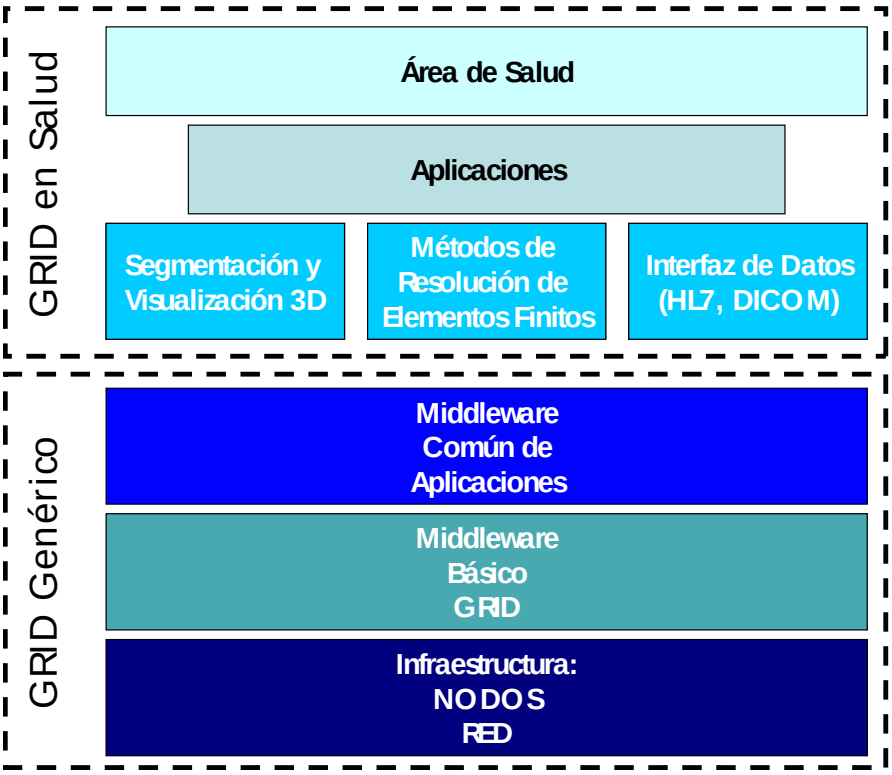


Ilustración 1: Esquema de un GRID en salud.

3.4.3.2 Recursos de Infraestructura y Humanos Disponibles, Dimensión

Un programa de e-Ciencia en salud es intrínsecamente multidisciplinar, tanto en lo relativo a personal (personal médico, ingeniería, informática, telecomunicaciones, químico, etc.) como en lo relativo a los centros (centros de investigación, laboratorios, hospitales, etc.).

Esta característica, unida a los requerimientos técnicos de un programa de e-Ciencia, conlleva a la necesidad de infraestructuras de comunes entre centros de investigación y proveedores de información (hospitales principalmente). La estructura GRID de cómputo puede ubicarse en centros de investigación y de servicio, pero la conexión de los centros proveedores de información con el entorno GRID debe permitir grandes transferencias de datos. Los requerimientos variarán en función de los proyectos piloto que se quieran implantar.

3.4.4 Definición de Posibles Proyectos Piloto

El proyecto pretende definir una capa de servicios genéricos para las aplicaciones GRID en salud. Esta capa se construirá a partir del análisis de requerimientos de aplicaciones y estándares existentes o en desarrollo identificados durante la etapa exploratoria del proyecto. A partir del análisis de estas aplicaciones se determinarán servicios, componentes, interfaces y formatos que darán lugar a la capa mencionada. Sobre esta

capa se desarrollarán experiencias piloto que demuestren los beneficios de la tecnología y de los resultados del proyecto.

Potencialmente, las aplicaciones piloto que se desarrollarán abordarán:

- Proceso de imágenes: Identificación de órganos y entidades funcionales, proyección 3D, generación de modelos, herramientas de diagnóstico asistido por computador.
- Simulación funcional de órganos y sistemas: Dinámica de fluidos, transporte de potenciales eléctricos, deformaciones visco-elásticas, análisis estructural.
- Acceso inteligente a grandes volúmenes de datos: Extracción de conocimiento y minería de datos sobre bases de datos poblacionales, búsqueda de patrones en imágenes y señales, localización de similitudes en diagnósticos y pruebas.
- Trabajo colaborativo: Compartición segura de datos médicos, video-conferencia, asistencia remota.

3.4.5 Transferencia de Tecnología, Visibilidad y Difusión de Proyectos

El proyecto persigue la puesta en marcha de un programa de e-Ciencia, pero con el objetivo de que este entorno sea beneficioso tanto para la comunidad científica como para el mundo empresarial.

La disponibilidad de un sistema de cómputo con una potencia pico del orden del Teraflop y una capacidad de almacenamiento de varios Terabytes permitirá abordar problemas tipo gran reto que actualmente son inabordable, máxime para los centros que no disponen de suficiente infraestructura para abordar problemas de tamaño medio.

Por tanto, es de especial interés para mantener la viabilidad del proyecto una vez terminada la duración propuesta, el disponer de un gran número de usuarios y clientes que puedan justificar la continuación de las operaciones. Más aún, el acceso a gran escala justifica la participación de empresas que puedan beneficiarse de este concepto tanto como usuarios como servidores específicos.

Para conseguir un gran impacto y un gran número de usuarios se plantea la difusión de resultados a tres niveles:

- Foros científicos. Atendidos principalmente por investigadores que desarrollan nuevas aplicaciones en el área de la salud, con el objetivo de atraer nuevas aplicaciones al área GRID.
- Foros de salud. Atendidos principalmente por usuarios (facultativos, enfermería, gerencia, etc.), pretenden fomentar el uso de las aplicaciones GRID, a la par de involucrar a los centros usuarios en el entorno GRID como proveedores de contenido.
- Foros empresariales y tecnológicos. Atendidos principalmente por empresas, con el objetivo de desarrollar y experimentar modelos de negocio que puedan beneficiarse de la tecnología GRID.

La difusión consistirá en:

- Elaboración de una imagen corporativa acorde con el proyecto global.
- Preparación de material de difusión impreso y electrónico.
- Mantenimiento de un sitio Web con la información actualizada, disponiendo de áreas públicas y privadas.
- Organización de presentaciones y demostraciones en eventos.

3.5 Área de Bioinformática

3.5.1 Motivación y Necesidades

3.5.1.1 Objetivos

Establecer las premisas para la constitución de un entorno de e-ciencia en el área de Bioinformática. Este documento se circulará entre los interesados para su comentario y refinamiento posterior.

3.5.1.2 Resumen:

Definición del área:

En el contexto de este documento, usaremos el término de *Bioinformática* para incluir todas las *aplicaciones de la tecnología de la información (TI) en las Ciencias de la Vida*. Por tanto incluimos la aplicación de la TI en otras disciplinas conocidas con denominaciones afines como son la Biocomputación o la Biología Teórica cuya delimitación estricta no es objeto de este documento.

La integración de diversas aplicaciones de TI en Ciencias de la Vida bajo una denominación común está justificada por la creciente interrelación de las diferentes ramas de conocimiento y su progresiva confluencia hacia puntos comunes de encuentro.

Por motivos prácticos, se hace exclusión de la aplicación de la TI en las *Ciencias de la Salud* en esta sección. Si bien estas disciplinas están confluyendo e incrementando su sinergia recientemente y se espera una mayor imbricación en el futuro próximo, el problema de las Ciencias de la Salud es todavía suficientemente diferenciado como para garantizar un tratamiento separado, y por lo tanto es abordado en un Área Temática separada.

Las Tecnologías de la Información en las Ciencias de la Vida:

El desarrollo de la Bioinformática ha ido estrechamente relacionado con el de las tecnologías experimentales. En este sentido, el desarrollo de la Biología Molecular ha supuesto un reto constante para las Tecnologías de la Información que han logrado a duras penas mantener un ritmo de crecimiento similar al de las Ciencias de la Vida en los últimos 20 años.

El desarrollo de las técnicas experimentales ha ido parejo a su popularidad en el ámbito científico y sobre todo a su repercusión e impacto social y económico, potenciado por hitos públicos de gran repercusión popular (como la secuenciación del genoma humano, las vacas locas y otros similares).

Ha sido precisamente uno de estos avances, la secuenciación del Genoma Humano, el detonante de la explosión de tecnologías experimentales conocida generalmente como *"Era Post-genómica"* (o "revolución de las *-ómicas"). En este contexto, se está abriendo un paradigma de trabajo nuevo que pretende abordar organismos enteros en lugar de genes o productos aislados: las tecnologías de genómica, proteómica, transcriptómica, etc... permiten ahora realizar un abordaje holístico del estudio de los procesos que codifican la vida. Este cambio de paradigma supone un nivel de complejidad adicional al basarse en una integración masiva de enormes cantidades de datos sin parangón hasta la fecha.

Este salto tiene dos vertientes, una social y otra práctica. La primera y más seria es la toma de *conciencia social* sobre la factibilidad de abordar experimentalmente una serie

de problemas hasta ahora inasequibles, con la consecuente presión sobre gobiernos y científicos para el desarrollo de estas tecnologías. La segunda y más grave es de índole práctica y significa un salto cuántico de varios ordenes de magnitud en el tamaño de los datos recopilados y otro aún mayor en el tratamiento de los mismos para alcanzar el grado de comprensión requerido para integrar datos de diversas fuentes en un enfoque sistémico.

El reto actual de todas las áreas de Bioinformática es responder a este cambio en la demanda de las necesidades analíticas en todas las áreas de las Ciencias de la Vida. Para ello no es posible continuar con los abordajes "clásicos" cuya capacidad analítica podía ir pareja al crecimiento exponencial de la tecnología.

El cambio está siendo progresivo, lo que está conduciendo a la mayoría de los grupos de Bioinformática a buscar soluciones a medida que los métodos en uso se iban quedando cortos. La primera fase de contención ha consistido en la instauración de sistemas de clusters y adaptación de algoritmos y procedimientos para su ejecución paralela a pequeña escala. Estas soluciones, si bien están permitiendo desarrollar y probar nuevos algoritmos y soluciones, resultan insuficientes para el tratamiento de datos experimentales masivos.

Las necesidades básicas se pueden expresar como una mayor demanda de capacidad de almacenamiento y tratamiento de información, y un crecimiento desmesurado de la capacidad de cálculo.

Los costes asociados al tratamiento de información derivada de las nuevas técnicas experimentales superan con creces la capacidad de prácticamente cualquier grupo o entidad aislada, aún recurriendo a soluciones paralelas de bajo coste como las granjas de PCs. En estas condiciones, la única solución viable consiste en la compartición de recursos entre grupos de forma solidaria hasta reunir recursos suficientes para abordar los problemas experimentales.

Los primeros resultados de estos esfuerzos están poniendo de manifiesto un problema adicional: la comprensión de las ingentes cantidades de información disponibles empieza a precisar de nuevos sistemas de *minería de datos* que permitan desarrollar métodos de inferencia sensibles al contexto capaces de extraer información relevante integrando fuentes *diversas y dispersas* de forma automatizada.

En otras palabras, *la forma más efectiva en coste y recursos de resolver los problemas actuales de las Ciencias de la Vida y por extensión responder a las crecientes demandas sociales sobre ellas es recurrir a sistemas masivamente distribuidos de tecnología Grid*.

3.5.1.3 e-Ciencia de la Vida

La aplicación de sistemas de e-Ciencia en Bioinformática no se justifica solamente por el crecimiento de las demandas computacionales. Además, precisamente por sus características, los problemas tratados proyectan perfectamente sobre una estructura computacional distribuida. Estas características son fundamentalmente dos:

- *Computación de grano muy grueso*: la mayoría de los problemas son divisibles en grandes subprocesos de intensa demanda computacional y ejecutables de forma independiente y en paralelo

- *Experimentos de muy alto rendimiento*: las grandes colecciones de datos de características similares generadas por las "*-ómicas" demandan un tratamiento uniforme (secuenciación en masa, análisis de estructuras masivo, etc..) y por tanto son tratables como una enorme colección de problemas independientes.

A esto debe unirse la enorme y creciente popularidad de estas técnicas experimentales reflejada en un gran número de laboratorios que aplican estas técnicas y son generadores independientes de grandes colecciones de datos. Esta situación aboga por una *organización descentralizada del almacenamiento, consulta y tratamiento de datos* apoyando con más fuerza la idoneidad de las tecnologías de Grid.

3.5.1.4 Necesidades básicas

Si bien hay una serie de problemas (procesamiento repetitivo de problemas masivos) que pueden empezar a beneficiarse inmediatamente de un entorno masivamente distribuido, aún no estamos en condiciones de explotar esta tecnología en toda su capacidad, precisamente por la novedad del problema experimental y consecuente escasez de abordajes informáticos.

Es de destacar que *ya existen grupos trabajando con tecnologías Grid*, especialmente en colaboración con otros grupos extranjeros y usando infraestructura foránea, pero sería irreal hacer depender las necesidades analíticas de la comunidad española de Ciencias de la Vida de Grids pertenecientes a otros países.

La situación de la mayoría de los grupos de Bioinformática en general puede calificarse como intermedia: han dado el salto a la *computación distribuida en clusters y granjas* de estaciones de trabajo, pero aunque en principio pasar los problemas a un entorno de mayor amplitud como es una Grid podría ser natural, no es posible afirmar que vaya a resultar trivial hasta que se aborde, y esto solo puede hacerse creando una infraestructura previa.

Finalmente, los usuarios, la principal fuerza de tracción sobre la Bioinformática plantean de por sí una problemática adicional: por un lado, pueden ya empezar a usar de forma inconexa algunos de los servicios desarrollados de forma distribuida sobre clusters, y en escasa medida sobre Grid. Además de necesitar nuevas herramientas y la adaptación de muchas más de las existentes, también precisan *puntos de entrada* que proporcionen un ambiente de acceso unificado a los diversos servicios disponibles, con *interfaces de trabajo intuitivos y versátiles*.

3.5.2 Proyectos de Grid en la Actualidad

La Unión Europea ha definido las tecnologías Grid como áreas prioritarias para su desarrollo en el 6º Programa Marco. Más aún, tras el análisis de las expresiones de interés, ha dado especial relevancia a las iniciativas Grid en Ciencias de la Vida y la Salud, incluyéndolas en varias áreas del Programa, en ocasiones con precedencia sobre iniciativas Grid de más amplio interés. Esta decisión resulta llamativa cuando se considera la escasa representación de EoI en Ciencias de la Vida, dos en total, una de ellas remitida desde España en representación de la Red Europea de Biología Molecular (EMBnet).

Ya hemos mencionado la presencia de diversos grupos españoles de Bioinformática en iniciativas Grid extranjeras, tanto americanas como europeas. La participación de los mismos en éstas refleja la madurez de los grupos y la existencia y aprovechamiento de colaboraciones con grupos e instituciones pioneros de otros países, pero sobre todo refleja la existencia inmediata de necesidades de computación que no pueden ser satisfechas efectivamente en nuestro país.

Las fuertes relaciones existentes entre grupos españoles y extranjeros sitúan a los grupos de Bioinformática españoles en una situación bastante apropiada para la participación en iniciativas similares del 6º Programa Marco. De hecho, varios grupos están participando ya activamente en la preparación y presentación de iniciativas Grid

para las primeras llamadas (como las propuestas EGEE y HealthGrid).

En otro orden, se están estableciendo las bases para constituir iniciativas Grid que incluyan a grupos de Bioinformática de países del entorno iberoamericano, bien dentro de colaboraciones auspiciadas por la UE como en otros marcos de colaboración internacional.

Por supuesto, es recomendable mantener y potenciar estas colaboraciones, fortaleciendo los vínculos que nos unen tanto a la UE, como EEUU, Iberoamérica y otros países, lo que conlleva una necesidad adicional: la de asegurar la compatibilidad e integración de cualquier iniciativa de e-Ciencia en España con sus equivalentes en otros países.

Un último factor a considerar es la razonable salud de la comunidad bioinformática española, favorecida por las actividades de la Red Temática Nacional de Bioinformática, en especial en relación con otros países del entorno, que la sitúa en un lugar de especial relevancia e influencia internacional; una situación temporalmente ventajosa que no sería aconsejable desaprovechar.

3.5.3 Casos de Uso y Aplicaciones Piloto

Como ya se ha mencionado, nuestro objetivo no es realizar un ejercicio intelectual en bioinformática distribuida, sino responder a las necesidades de la ciencia experimental. En este sentido podemos replantear el problema en base a las distintas disciplinas y sus necesidades específicas:

- *Biología teórica*: modelización de sistemas biológicos complejos.
- *Bioinformática "tradicional"*: análisis de secuencias, y su extensión a la genómica.
- *Biocomputación*: análisis de estructuras, y su extensión al análisis de alto rendimiento.
- *Bioinformática "moderna"*: tecnologías de arrays y chips.
- *Tratamiento de información*: gestión de la información generada en el laboratorio: LIMS y bases de datos.
- *Provisión de servicios*: consolidación de servicios Grid de cara al usuario.

A éstas hay que añadir necesidades específicas de la Bioinformática en sí misma: a pesar de la familiaridad con tecnologías distribuidas en cluster, el entorno de Grid es suficientemente diferente como para precisar un apoyo de adaptación. Adicionalmente, el middleware de Grid aún no soporta de forma suficientemente transparente entornos de desarrollo avanzados como los precisos para poder asegurar un desarrollo eficiente en Ciencias de la Vida. En resumen, esto supone unas necesidades adicionales de:

- *Desarrollo de middleware y herramientas de desarrollo en Grid*
- *Formación y adaptación de desarrolladores a los nuevos entornos Grid*
- *Soporte en la instalación y mantenimiento de las facilidades locales*

Es de destacar que todas las áreas de aplicación descrita están representadas entre grupos españoles de Bioinformática de reconocido prestigio, así como que los intereses de los grupos a menudo se extienden sobre varias áreas simultáneamente. En la exposición siguiente mostramos las líneas de trabajo e interés de diversos grupos siguiendo de forma laxa la enumeración previa:

Rafael La Hoz -- Dpto. Biología Teórica, UCM

El Dpto. de Biología Teórica viene desarrollando desde hace tiempo modelos

computacionales de estructuras macromoleculares y celulares complejas mediante autómatas finitos. La nueva disponibilidad de ingentes cantidades de datos a nivel de genomas o proteomas enteros y resolución estructural de alto rendimiento requiere para su comprensión la elaboración de modelos de simulación mucho más complejos y con un número muchísimo mayor de componentes.

El interés principal del grupo es la elaboración de experimentos de simulación *in silico* del comportamiento de estructuras celulares complejas. Por su alto número de componentes este tipo de experimentos de modelización son susceptibles de ser mejorados mediante técnicas eficientes de paralelización.

La disponibilidad de un entorno de computación distribuída proporcionaría mayores posibilidades de cálculo para los experimentos de simulación, permitiendo la realización de modelos con un número mucho mayor de componentes moleculares y por consiguiente mucho más completos, incrementando la comprensión de los procesos biológicos a escala celular.

Grupo de Julio Rozas, U. Barcelona

Este grupo trabaja principalmente en Evolución y Genética de Poblaciones Molecular. La evolución fué una de las primeras ciencias en usar tecnologías de la información, reflejo de su relevancia en la misma. El grupo trabaja en el desarrollo de aplicaciones para su uso aplicado y dispone ya de varios proyectos que requieren cómputo pesado:

1) Análisis de patrones evolutivos en genomas completos.

2) Estimación de parámetros poblacionales mediante simulaciones de ordenador basadas en métodos de Montecarlo.

El interés por tanto gira en torno a la adaptación de técnicas de análisis clásico a las nuevas tecnologías de genómica y la construcción de modelos avanzados para la comprensión de la dinámica de poblaciones.

Instituto Cavanilles de Biodiversidad y Biologia Evolutiva, Dept. de Genetica / Serv. Bioinformatica. Universitat de Valencia

Este grupo ha empezado a trabajar en entornos distribuidos con el sistema de InnerGrid, aplicado a temas de reconstruccion filogenetica y genomica comparativa.

Aunque el proyecto se ha iniciado hace relativamente poco tiempo, cuenta con un estudiante de doctorado que va a trabajar especialmente con ello.

Laboratorio de Bioinformática. Centro de Astrobiología (CSIC - INTA).

Las líneas de trabajo de este grupo se centran, por un lado, en el estudio de la estabilidad termodinámica de proteínas y su relación con evolución molecular y modelos de plegamiento; y, por otro, en el análisis de interacciones entre proteínas y proteína-ligando.

El grupo se beneficia de la interacción con el laboratorio de Computación Avanzada del CAB y el departamento de Informática del INTA lo que está permitiendo el desarrollo de dos iniciativas en entorno Grid:

- Una colaboración en curso, entre el Laboratorio de Computación Avanzada y el Laboratorio de Bioinformática, para realizar computaciones masivas con el programa PROTFINDER, desarrollado en el centro, sobre predicción de estructuras y propiedades termodinámicas de proteínas mediante métodos de alineamiento estructural (threading). El programa ha sido adaptado para su ejecución en un entorno Grid usando la herramienta GridWay, desarrollada también en el centro, y

ya se han realizado las primeras pruebas. Actualmente, se están realizando algunas mejoras en el método para luego aplicarlo, mediante tecnología Grid, a grandes bases de datos de secuencias de proteínas, o bien, a genomas completos.

- Un proyecto conjunto del Laboratorio de Bioinformática del CAB, el departamento de Informática del INTA y la empresa GridSystems para el uso de un gran número de ordenadores del campus del INTA en un proyecto de análisis masivo de interacciones entre proteínas implicadas en procesos de división celular bacteriana. En la actualidad se están realizando pruebas piloto utilizando el software InnerGrid sobre plataformas Windows, Solaris y Linux simultáneamente y una versión para Grid del programa AutoDock.

Grupo de Modesto Orozco. Univ. De Barcelona y Parque Científico de Barcelona

El grupo posee una larga tradición en la utilización de herramientas de cálculo intensivo enmarcadas dentro de la química cuántica, la dinámica molecular o el docking, a las que más recientemente se han incorporado de herramientas de bioinformática estructural que incluyen la construcción y manipulación de bases de datos estructurales. La principal limitación actual de este tipo de técnicas no es tanto el fundamento teórico de las mismas sino la capacidad de cálculo masivo. No tiene sentido, en la actualidad, analizar un número limitado de situaciones, cuando se dispone de información estructural suficiente para ampliar el estudio a situaciones generales.

Para ello, los sistemas paralelos y en particular la informatica distribuida (GRID) constituye una solución excelente.

La disponibilidad de tecnología Grid ofrece la posibilidad tanto de optimizar la utilización de los recursos de computación ya disponibles, como de extender el uso de herramientas tradicionales a sistemas de tamaño mucho mas realista.

Xavier Messeguer -- CIRI, Barcelona; R. Guigó, Alfonso González -- IMIM, Grupo de informática biomédica

En el entorno del IMIM y el Parque Científico de Cataluña han confluído una serie de grupos con un interés común por el desarrollo de aplicaciones y servicios en Bioinformática y la colaboración mutua. Todos ellos son grupos de sólida reputación y larga experiencia en la distribución de cómputo, con clusters de computadores instalados y en producción desde hace tiempo.

Algunos de estros grupos, como es el caso del de Roderic Guigó, están ya estableciendo relaciones con otras iniciativas Grid nacionales en otros países para desarrollar colaboraciones como una extensión natural a la distribución de sus trabajos (como blast o geneid) en clusters.

Evidentemente los grupos de Biologia Estructural también tienen un interés genuino en aprovechar iniciativas de computación distribuída: las aplicaciones de modelización molecular se cuentan entre las primeras en haber sido paralelizadas, con una larga casuística y experiencia en este problema, y la disponibilidad de una infraestructura masivamente distribuída multiplica el tamaño de los problemas que son abordables y -por consiguiente- extiende seriamente nuestra capacidad de comprensión del funcionamiento de los seres vivos.

J. M. Carazo -- Unidad de Biocomputación, CNB, CSIC

La Unidad de Biocomputación es un grupo de reconocido prestigio internacional con una larga experiencia en computación paralela (imparte esta asignatura en la UAM) que mantiene estrechas colaboraciones con otros grupos de computación de altas

pestaciones nacionales y extranjeros, en especial con iniciativas Grid tanto europeas como Norteamericanas.

El interés del grupo se centra en la comprensión de procesos biológicos que requieren la interacción de grandes números de componentes que experimentan cambios estructurales dinámicos. Para ello es preciso desarrollar técnicas eficientes de resolución de estructuras de grandes complejos macromoleculares en especial mediante microscopía electrónica de transmisión.

La Unidad de Biocomputacion trabaja en el desarrollo de algoritmos de reconstrucción tridimensional que presenten un comportamiento robusto frente al ruido promediando un gran número de imágenes, desarrollando métodos iterativos que constituyen una alternativa ventajosa a los métodos tradicionales, pero que tienen demandas computacionales considerablemente superiores tanto en cómputo como en la necesidad ejecutar el mismo programa cientos de veces para obtener validaciones estadísticas.

La metodología de reconstrucción 3D constituye una de las aplicaciones que más fácilmente puede beneficiarse de las posibilidades de las grids, puesto que el empleo de las grid hace posible considerar estudios vedados hasta la fecha.

Adicionalmente la Unidad de Biocomputación trabaja en el desarrollo de sistemas para el acceso integrado a datos biológicos con una aproximación de mediación semántica. La resolución de heterogeneidades semánticas es esencial para la integración de información cuando se manejan datos complejos y cuando los usuarios y aplicaciones provienen de distintas disciplinas.

Grupo de J. L. Oliver -- U. Granada

Este es un grupo interdisciplinar de biólogos y físicos que estudia el genoma desde la perspectiva de los sistemas complejos. El grupo posee una sólida reputación como desarrolladores de aplicaciones de segmentación genómica que permiten descomponer un genoma en regiones de composición homogénea (isocoras). La delimitación precisa de las fronteras entre isocoras es muy útil para la anotación de genomas, aumentando sensiblemente la eficiencia de los programas de predicción computacional de genes y otros elementos funcionales (islas CpG, promotores, elementos repetidos, frecuencia de splicing alternativo, etc). Además, el grupo está también interesado en las correlaciones y en la estructura a gran escala del genoma, con la vista puesta en el análisis de las interacciones entre las partes que componen este sistema complejo. En ambas vertientes, el grupo viene experimentando una creciente necesidad de recursos computacionales distribuidos: por un lado están procediendo a paralelizar antiguos algoritmos secuenciales con el objeto de poder atacar en un tiempo razonable el análisis y la comparación de los genomas completos actualmente disponibles. Por otro lado, los nuevos algoritmos que están diseñando son ya paralelos desde el principio, utilizando la escasa infraestructura de computación distribuida disponible actualmente. El grupo mantiene colaboraciones con otros grupos, principalmente el Laboratorio de Evolución Molecular de la Stazione Zoologica Anton Dohrn de Nápoles, dirigido por Giorgio Bernardi, descubridor de las isocoras, y el Dpto. de Física de la Universidad de Boston, dirigido por Eugene H. Stanley, descubridor de las correlaciones de largo alcance en el ADN. Asimismo, mantiene una colaboración muy estrecha desde hace años con Wentian Li, miembro del Instituto de Santa Fe, Nuevo México, para estudios de la complejidad. Finalmente, el grupo tiene varias de sus aplicaciones disponibles en línea a través de la web, habiéndose convertido su servidor en el núcleo de la recién constituida Red de Bioinformática de Andalucía, por lo que precisan fuertes recursos para poder ofrecer acceso a algoritmos distribuidos a un número creciente de usuarios.

Javier de las Rivas -- U. Salamanca, Centro de Investigación del Cáncer, CSIC

El interés del CIC en las tecnologías de procesamiento paralelo le ha llevado a desarrollar entornos de producción distribuídos sobre las máquinas del Centro, llegando a convertirse en un prototipo de pruebas de la tecnología de GridSystems. Las áreas de aplicación de mayor interés son:

- Manejo de Datos de resultados de Genómica, especialmente de datos derivados de tecnología Affymetrix.
- Cálculo computacional sobre datos de expresión de microarrays genómicos aplicando herramientas y algoritmos con el paquete estadístico público R.
- Calculo computacional bioinformático, especialmente implementación de algoritmos tipo BLAST, Psi-BLAST, FASTA, HMM en entornos paralelos.
- Calculo computacional con paquetes de manejo biomolecular 3D: software de visualizacion, software de docking, software de dinámica molecular.

Alfonso Valencia -- Grupo de diseño de proteínas, CNB, CSIC

El Grupo de Diseño de Proteínas es uno de los grupos más representativos del país, con estrechas relaciones con muchos otros grupos de trabajo dentro y fuera de España. Desde hace años trabaja sobre entornos paralelos con granjas linux, y más recientemente con clusters Linux y HP/Tru64 y con un sistema Paracel BlastMachine de 20 procesadores. El grupo colabora con otras iniciativas Grid europeas (como UK-grid) para poder correr algunos de los servicios que ofrece públicamente.

Además de el desarrollo de aplicaciones paralelas, el interés del grupo en tecnologías Grid se basa en la posibilidad de ofrecer servicios avanzados que requieren computación pesada:

- Un metaservidor de estructura tridimimensional, llamado Libellula.
- Un sistema de cálculo de las interacciones entre dos genomas, llamado ECID.
- Un sistema de agrupación de secuencias en familias de proteínas, llamado FunCUT.
- Un sistema de bases de datos para el análisis de genomas completos, llamado ORFandDB.
- Dos servicios de análisis de abstracts de Medline, para encontrar los artículos relacionados con los genes proporcionados de entrada. Estos servicios son Geisha y HCAD.

Algunos de estos servicios son susceptibles de ser directamente integrados en entornos Grid desde el primer momento, como Libellula y ECID, asociados a la granja de PC's, y FunCUT, asociado tanto a la BlastMachine como la granja de PC's.

Dpto. de Arquitectura de ordenadores, U. de Málaga

Este Departamento posée una larga tradición de colaboración con grupos de investigación en Ciencias de la Vida y Bioinformática, trabajando en la paralelización de algoritmos pesados de análisis de datos biológicos. El Dpto. ha realizado una fuerte inversión en el desarrollo de aproximaciones basadas en multiprocesadores y procesamiento paralelo en diversas áreas. Sus líneas de trabajo básicas son:

- Diseño de entornos paralelos de computación -- arquitecturas distribuídas (Grid)
- Desarrollo de componentes software -- middleware
- Desarrollo de aplicaciones y paralelización de algoritmos.

En consecuencia, el interés es múltiple, centrado principalmente en el desarrollo de middleware y entornos paralelos, pero también en el desarrollo de aplicaciones y, en

tanto que usuarios y proveedores de acceso a las mismas, consumidores importantes de fuertes recursos computacionales.

Servicio EMBnet/CNB

El Servicio EMBnet/CNB funciona como un servicio nacional a toda la comunidad investigadora del país, constituyendo el nodo español de la Red Europea de Biología Molecular (EMBnet) desde hace más de una década. Desde su proyección internacional mantiene estrechas relaciones con otras iniciativas Grid europeas y juega un papel central en la promoción y coordinación de iniciativas Grid en Ciencias de la Vida.

Los intereses fundamentales del Servicio se centran en la provisión de servicios de cómputo científico al usuario final investigador, la asistencia en el uso de las herramientas, la formación y el apoyo al desarrollo de aplicaciones.

La demanda de herramientas con alta demanda de prestaciones por parte de los usuarios empieza a notarse ya, y a medida que las nuevas técnicas experimentales se popularicen y extiendan es de prever un cambio consecuente: por un lado un incremento en el número de usuarios, y por otro, la incorporación de nuevos métodos de análisis altamente exigentes desarrollados por grupos de Bioinformática como los expuestos en éste documento.

En este sentido, es de notar que las herramientas desarrolladas en grupos de investigación suelen precisar una elaboración posterior para facilitar su uso e integrarlas con las demás aplicaciones que requiere el usuario. El uso de aplicaciones diseñadas para Grid requerirá un esfuerzo importante para elaborar portales de acceso uniformes e integrar herramientas dispares en un entorno cómodo al investigador.

Finalmente el servicio constituye un punto de referencia y apoyo para los desarrolladores de Bioinformática. En este sentido se ha solicitado en el proyecto EGEE a la UE la asignación de dos ingenieros para proporcionar un servicio de formación, asesoría, soporte y adaptación de aplicaciones a entornos Grid que se ofrecerá a la comunidad bioinformática a nivel europeo que sería deseable incrementar con especialistas dedicados a apoyar específicamente a la comunidad española.

Esta lista no pretende ser exhaustiva, otras instituciones (p. ej. el Instituto Cavanilles de Biodiversidad y Biologia Evolutiva Dept. de Genetica / Serv. Bioinformatica, Universitat de Valencia, el Institut de Biotecnologia i Biomedicina Vicent Villar Palasi, UAB, el Instituto de Investigaciones Biomédicas del CSIC, etc..) y grupos (p. ej el grupo de Guillermo Thode en la U. Málaga) han expresado su interés y apoyo por una iniciativa de e-Ciencia para Ciencias de la Vida, bien como una extensión natural de sus recursos en cluster existentes, bien como consumidores de recursos avanzados de computación a distintos niveles.

3.5.4 Transferencia de Tecnología, Visibilidad y Difusión de Proyectos

Muchos de los grupos mencionados han demostrado ya su capacidad para transferir los resultados de su trabajo a otras entidades y generar patentes y productos comerciales, e incluso para formar el germen de empresas de nueva creación.

El creciente interés de la Biotecnología depende en buena medida para su éxito de la existencia de herramientas analíticas capaces de procesar las ingentes cantidades de información que se están generando. En estos momentos existe una gran escasez de aplicaciones, por lo que los proyectos enumerados son en su mayoría de gran interés tanto académico como comercial si pueden ser llevados a buen término. Esta es una razón más para fundamentar la importancia de la instauración de una iniciativa Grid: su presencia permitirá el desarrollo de aplicaciones necesarias y competitivas y reducirá

nuestra dependencia tecnológica de grupos y empresas extranjeros.

El interés por las herramientas que hay que desarrollar procede de su utilidad y conlleva una repercusión importante: las aplicaciones desarrolladas deberán ser soportadas y mantenidas para su uso en producción. Este mantenimiento al no ser un trabajo de investigación no puede ser financiado con cargo a los presupuestos de investigación, lo que tiene dos efectos notables:

Por un lado supone un incremento en la relevancia de los servicios de apoyo a la investigación (como es el caso de EMBnet/CNB, que actúa como punto de referencia para otros servicios) en el papel de formación del usuario final, apoyo y asesoría técnica.

Por otro lado supone la necesidad de encontrar una entidad que pueda garantizar el mantenimiento y mejora progresivos de las aplicaciones más relevantes y populares, un papel en el que la labor empresarial es fundamental.

En resumen, para que las iniciativas de desarrollo en Ciencias de la Vida tengan éxito en el futuro próximo hacen falta herramientas que aún hay que desarrollar y para las cuales resulta necesario disponer de una infraestructura de e-Ciencia. Para que esta iniciativa tenga éxito es preciso garantizar la transferencia de conocimientos y tecnología en varias direcciones:

- transmisión de experiencia y conocimientos de desarrollo paralelo a los desarrolladores desde servicios de referencia y otros grupos de desarrollo
- transmisión de software desarrollado a servicios y empresas, junto con el conocimiento preciso para explotarlo y mantenerlo
- transmisión del conocimiento de explotación y asesoría al usuario final desde servicios y empresas de apoyo.

Para garantizar estos procesos de comunicación conviene aumentar los intercambios que ocurren naturalmente con la dotación de recursos par la realización de reuniones y congresos que permitan el libre intercambio de ideas e información entre grupos, servicios y empresas (como por ejemplo las reuniones anuales de las Redes Temáticas), actividades de diseminación (como la creación de un portal WWW de referencia) así como la realización de cursos de formación para desarrolladores y cursos en el uso de aplicaciones para usuarios, y sobre todo, actividades de coordinación general de las iniciativas a través de uno o más coordinadores.

3.6 Área de Química Computacional

La Química Computacional consiste en la modelización cuantitativa de fenómenos de interés químico usando métodos y técnicas computacionales. Las líneas más relevantes inmersas en este área son:

- Descripción cuántica molecular (englobando la tradicional Química Cuántica): Estructura electrónica, movimiento nuclear, reactividad química, caracterización molecular, espectroscopia electrónica y roto-vibracional, semejanza molecular cuántica.
- Dinámica molecular (Resolución de las ecuaciones del movimiento de sistemas moleculares con aproximaciones cuasiclásicas, semi-clásicas o cuánticas): Dinámica de reacciones, cálculo de secciones eficaces de reacción, reactividad química.
- Mecánica molecular (Aplicación de la mecánica clásica con potenciales interatómicos adecuadamente parametrizados): Estructura 3D de macromoléculas, interacciones ligandoreceptor

Los métodos y técnicas de la Química Computacional se aplican (entre otras posibilidades) a:

- Predicción e interpretación de estructuras moleculares.
- Modelización de reacciones químicas.
- Predicción e interpretación de espectros electrónicos y rotovibracionales.
- Identificación de especies en el espacio interestelar por medio de sus patrones roto-vibracionales.
- Modelización de reacciones complejas en modelos de química atmosférica, de combustión de hidrocarburos, o de nubes interestelares.
- Molecular docking entre ligandos bioactivos y sus receptores macromoleculares.
- Diseño de novo de ligandos bioactivos.
- Estudios de estado sólido.

3.6.1 Motivación y Necesidades

Existe una importante necesidad de recursos para computación de alto rendimiento, que permitan una optimización de recursos hardware y software existentes. Además la posibilidad de crear organizaciones virtuales resulta muy adecuada para la aglutinar y coordinar el acceso a los recursos para la resolución de los grandes problemas del área. Además, el acceso al Grid permitiría abordar nuevos problemas de mayor complejidad.

3.6.2 Casos de uso

En el área de Computación de alto rendimiento se han identificado las siguientes necesidades:

- Necesidad de altas exigencias computacionales en cuanto a potencia de cálculo, y en algunos casos en cuanto a espacio en disco (como en problemas de estructura electrónica).
- Utilización de herramientas ya paralelizadas: Se sacaría partido de su instalación en el entorno grid, pero la heterogeneidad inherente en el sistema y su paralelización, obedeciendo a un paradigma de paralelización fina, limitaría su uso a nivel de los sistemas multiprocesadores (incluyendo cluster de computadores) integrados en el grid. Ejemplos de estas herramientas son los paquetes de estructura electrónica

Gamess, NorthWest Chem o Siesta.

Por otro lado se persigue la optimización de recursos hardware y software existentes, dado

- Los recursos computacionales no se utilizan al 100% el 100% del tiempo.
- Un grid con un alto número de nodos proporciona una gran cantidad de recursos libres a lo largo del tiempo, tantos más cuanto mayor es el grid.
- El entorno grid, por lo tanto, permitiría abordar el tratamiento de problemas con alta exigencia en recursos computacionales a grupos muy capacitados pero con insuficiente infraestructura propia.

3.6.2.1 Grupos Participantes

A continuación se citan algunos de los grupos más relevantes en el área:

- Grupo de Química Computacional y Computación de Alto Rendimiento de la Universidad de Castilla-La Mancha (Camelia Muñoz Caro, Alfonso Niño Ramos, Sebastián Reyes Ávila)
- Grupo de Dinámica Molecular de Reacciones Químicas de la Universidad del País Vasco (Ernesto García Para, Maite Martínez González)
- Departamento de Química-Física, Universidad de Valencia (Raúl Crespo Crespo)

3.6.3 Middleware actual y específico

Además del desarrollo de aplicaciones piloto, existe una necesidad de herramientas middleware para:

- Permitir la creación de portales web personalizados para cada organización virtual (como el GridPort toolkit de NPACI) .
- Programación paralela sobre el grid, con la ayuda de herramientas como GRID superscalar y GridWay.
- Adecuación al estado dinámico del sistema (sistemas autoadaptativos)
    - Monitorización dinámica de trabajos, recuperación automática de fallos, recopilación automática de resultados parciales en el grid, integración y análisis automático de los mismos.
    - Punto clave de los nuevos modelos: Generación de grandes cantidades de información sobre el grid que debe ser transferida, filtrada, interpretada y organizada de forma automática.

3.6.4 Proyectos piloto

La utilización de Grid en este ámbito no sólo permitiría compartir recursos sino cohesionar a la comunidad científica. Por un lado, sería deseable la creación de organizaciones virtuales, con el objetivo de hacer disponible un entorno virtual que permita la colaboración de grupos con similares líneas de investigación. De esta forma se obtendría mayor eficacia y difusión de los esfuerzos individuales al integrarse en una base de conocimiento común y una mayor facilidad para la organización de proyectos coordinados a nivel nacional y europeo.

En lo referente a nuevos problemas de alta complejidad, potenciales proyectos pilotos son:

- Nuevos problemas complejos exigen el desarrollo de nuevas herramientas de modelización.
- La programación sobre paralelismo grueso saca partido de la estructura granular y

heterogénea del grid. Ejemplos de aplicación:

- Exploración de hipersuperficies de energía potencial para dinámica molecular, reactividad química o estructura roto-vibracional.
- Cálculo de trayectorias en dinámica de reacciones.

3.6.5 Conclusiones

Finalmente, se pueden resumir las acciones a realizar en los siguientes pasos:

- Desarrollo de prototipos de herramientas en paralelismo grueso para realizar pruebas de carga reales sobre un entorno grid.
- Para los casos anteriores, desarrollo de los modelos teóricos de rendimiento del sistema. Estos modelos serían de utilidad para el análisis de los resultados obtenidos y la planificación de tareas.
- Integración de varios grupos de investigación en un modelo de organización virtual. El punto clave sería la modelización de una base de información común de la organización (consideraciones de autenticación, duplicidad de información y procesos de actualización).
- Desarrollo de un prototipo modelo de portal web que actúe de interfaz transparente entre los usuarios de la organización virtual y los recursos del sistema.

3.7 Área temática de sistemas complejos

3.7.1 Motivación de un entorno Grid

3.7.1.1 Necesidades computacionales en el estudio de sistemas complejos.

No existe una definición precisa de sistemas complejos pero por este tipo de sistemas se suelen entender aquellos en los que la dinámica esta gobernada por ecuaciones no lineales presentando comportamientos caóticos, a menudo de alta dimensionalidad. Estos sistemas pueden estar formados por muchos elementos en interacción y es característico que el comportamiento del sistema global trascienda las expectativas generadas por la inferencia directa a partir de los comportamientos individuales.

Típicamente este comportamiento cooperativo emergente viene condicionado más por características globales del sistema, como son sus simetrías, dimensionalidad, topología de la interacción entre componentes, … que por la dinámica individual.

Por ello sistemas pertenecientes a áreas de conocimiento diversas como es el caso de sistemas hidrodinámicos, sistemas ópticos no lineales, medios granulares, sistemas y circuitos electrónicos, sistemas optoelectrónicos, reacciones químicas, sistemas biológicos, redes de comunicación, … pueden mostrar dinámicas globales, e intestabilidades parecidas. Surge entonces de forma natural la pretensión de intentar comprender este tipo de fenomenologías en su conjunto, intentando extraer comportamientos universales que puedan ser luego aplicados a otros sistemas parecidos.

Es por tanto esta una ciencia interdisciplinar que desde su propia definición de objetivos pretende romper con las divisiones que tradicionalmente separan y delimitan las diversas áreas de conocimiento y que aspira a tender puentes por los que ideas de disciplinas dispares circulen de forma fluida y beneficien al desarrollo de cada disciplina especifica.

Dentro de este contexto, la computación juega un papel crítico en la comprensión de los fenómenos emergentes globales. Dada la naturaleza no lineal y la complejidad de la dinámica en muchos casos sólo la computación numérica intensiva es capaz de aportar luz sobre comportamientos a gran escala. La dependencia espacial típicamente presente en estos sistemas así como la necesidad de abordar comportamientos a escalas de tiempo muy diversas supone un reto computacional importante. Movimientos de interfases, frentes o ondas de choque o formación de vórtices, estructuras localizadas o solitones en sistemas extendidos requieren por ejemplo de una elevada resolución espacial al tiempo que el estudio de la dinámica de estas estructuras a gran escala y su interacción con otras estructuras similares requieren sistemas de gran tamaño. Y si se quiere entender el comportamiento global, no basta con integrar las ecuaciones dinámicas para un valor dado de los parámetros. Lo crucial es el conocimiento de las inestabilidades y bifurcaciones que surgen al cambiar los parámetros de control del sistema.

Por otro lado no existe el sistema complejo prototípico como tal ni por tanto un programa valido para el estudio de los comportamientos cooperativos en cualquier sistema complejo. La diversidad subyacente en esta ciencia ha motivado descripciones en términos diversos, desde mapas acoplados a ecuaciones en derivadas parciales, pasando por ecuaciones diferenciales estocásticas. Y es que si bien el comportamiento emergente global puede ser común, el modelaje numérico de cada sistema particular tiene que ser adecuado y realista para el mismo si se quiere cumplir con la pretensión de

ayudar al desarrollo de la ciencia en disciplinas específicas y en el abordaje problemas concretos.

3.7.1.2 Justificación del GRID como solución

Son muchos los grupos científicos en distintas partes del mundo trabajando en este tipo de problemas. Para ello solo basta con ver el crecimiento experimentado en los últimos años por una revista netamente interdisciplinar y enfocada hacia este tipo de sistemas como es Physical Review E. A pesar de ello no existe una base común de aplicaciones a fin de realizar el modelaje de los distintos sistemas. Además la potencia computacional de los distintos grupos de investigación suele ser siempre limitada y en muchos casos insuficiente para abordar los problemas más complejos o explorar en detalle el espacio de parámetros. El GRID se presenta entonces como una solución dado que:

a) En un entorno GRID se podrían crear y compartir aplicaciones de tipo general que pudiesen ser utilizadas por distintos grupos de investigación en el estudio de ciertos tipos de comportamientos.
b) El entorno GRID permitiría desarrollar de forma eficiente cálculos intensivos o repetitivos, como es el caso de exploración del espacio de parámetros, permitiendo el aprovechamiento al máximo de los recursos computacionales existentes en distintos centros.
c) Los datos generados en bruto podrían pasar a formar parte de una base de datos distribuida a fin de facilitar su intercambio y acceso tanto por parte de grupos propiamente dedicados al estudio de sistemas complejos, como por parte de grupos dedicados al estudio de sistemas específicos, los cuales podrían utilizar esos datos para análisis, visualización o comparación con otros resultados.

3.7.2 Desarrollo previsto de middleware específico

Como ya se ha dicho el estudio de la dinámica de sistemas complejos puede hacerse en base a diversos tipos de modelos cuya implementación computacional tiene a su vez requerimientos muy diversos. A pesar de ello ciertos componentes parecen ser claramente necesarios como serian:

- Módulos de interfaz que permitiesen a los usuarios definir el modelo concreto a estudiar y el rango de parámetros en los que se quiere estudiar ese modelo.
- Módulos de cálculo que permitiesen la integración numérica de las ecuaciones diferenciales o el análisis de estabilidad mediante procedimientos semianalíticos.
- Módulos de visualización y análisis de datos. Típicamente, la gran cantidad de datos generados imposibilita su interpretación directa. La visualización y el tratamiento de estos datos son por tanto herramientas imprescindibles en la comprensión de la dinámica.

El desarrollo de distintos proyectos piloto y el interés despertado en la propia comunidad científica en las aplicaciones a sistemas concretos indicaran en que dirección es más adecuado perfilar los componentes.

3.7.3 Casos de Uso y Proyectos Piloto

3.7.3.1 Use Cases

Es esta área cabe distinguir distintos tipos de procesos:

3.7.3.1.1 Procesos que requieren computación de altas prestaciones.

En este tipo de procesos entrarían las simulaciones numéricas de la dinámica caótica en sistemas con dependencia en 2 o 3 dimensiones espaciales, en especial aquellos en los que se requiere la descripción simultánea de escalas espaciales muy diversas. Típicos ejemplos podrían ser la descripción de distintos régimenes de caos espacio-temporal en ecuaciones prototípicas (Ginzburg-Landau, Swift-Hohenberg) o en modelos de sistemas concretos (formación de estructuras e inestabilidades en reacciones químicas o en sistemas ópticos no lineales, fenómenos de turbulencia…).

3.7.3.1.2 Procesos en los que se requiere alta productividad.

Este es el caso de la exploración en el espacio de parámetros de la dinámica que presentan distintos modelos. Ello permite identificar las inestabilidades que dan lugar a la formación de estructuras espaciales por rotura espontánea de simetría, las estructuras emergentes de esta inestabilidad y en algunos casos determinar inestabilidades subsecuentes de estas estructuras. El GRID proporcionaría acceso a recursos remotos que permitirían la ejecución simultánea de muchas de estas simulaciones al mismo tiempo que se minimiza el tiempo muerto de esos recursos.

3.7.3.1.3 Acceso a grandes volúmenes de datos compartidos.

Los dos procesos anteriores serian los básicos en la primera etapa de funcionamiento de un GRID en sistemas complejos. A partir de la experiencia adquirida se podría abordar el punto c) descrito en el apartado 1.2, es decir, la constitución de una base de datos distribuida con los resultados en bruto obtenidos en simulaciones intensivas, los cuales podrían ser utilizados para visualización y análisis por diversos grupos científicos. Para ello seria necesario adoptar unos criterios básicos en la forma de almacenamiento de estos datos así como proveer de herramientas específicas que facilitasen su consulta por parte de grupos científicos trabajando en áreas específicas.

3.7.3.2 Proyectos piloto

Los proyectos piloto de más fácil implementación son aquellos relacionados con el punto b) desarrollado en el apartado 1.2, es decir las simulaciones numéricas en aquellos casos en los que se pretende explorar la dinámica para distintos valores de los parámetros. Es ese contexto surgen dos tipos de cálculos: los tradicionales cálculos de integración numérica de las ecuaciones de evolución dinámica y los potentes métodos semianalíticos que permiten el análisis riguroso de estabilidad en soluciones conocidas solo numéricamente.

Proyectos involucrando computación de altas prestaciones requieren de una velocidad de comunicación entre ordenadores mucho mas elevada de la que parece ser disponible en estos momentos, por lo que el cálculo paralelo en maquinas localizadas en distintos lugares se desarrollaría más adelante en función de las disponibilidades. En cualquier caso el GRID puede facilitar el acceso a este tipo de maquinas a investigadores localizados en distintas partes del mundo.

3.7.4 Transferencia de tecnología visibilidad y difusión de proyectos.

La clara interdisciplinaridad de este programa de e-Ciencia lo hace abierto a un gran número de interacciones con otras ramas del mundo científico, la empresa y la sociedad en general, por lo que la difusión de los conocimientos adquiridos tanto en la elaboración del programa como en los resultados científicos que pueda dar es de gran importancia.

Para que esta difusión de conocimientos se dé propondríamos la creación de sitios web, listas de correo, foros de discusión y entornos colaborativos en los que tanto profesionales como personas simplemente interesadas pudieran acceder al conocimiento de que es objeto la ciencia de los sistemas complejos.