La Computación en GRID se abre camino en el IFIC

Este artículo pretende clarificar el concepto de computación GRID, la utilización de las tecnologías GRID para construir el modelo de computación de la era del LHC y mostrar las actividades que se llevan a cabo en el IFIC.

Los proyectos científicos de comienzo de este siglo abordan objetivos cada vez más ambiciosos que requieren la resolución de problemas computacionales complejos, tanto por el volumen de los cálculos a realizar como por el tamaño y complejidad de la información a procesar. Por ejemplo, para la investigación en Física de Partículas se está construyendo el acelerador LHC (Large Hadron Collider) del CERN, en el cual se realizarán cuatro experimentos (ATLAS, CMS, ALICE y LHCb), entre los cuales se encuentra el experimento ATLAS , experimento en el que participa el IFIC de Valencia y cuyo objetivo principal es la búsqueda del bosón de Higgs.

El LHC, que entrará en funcionamiento el año 2007, requerirá el almacenamiento y procesamiento del orden de 6-8 Petabytes (10^{15} bytes) de datos cada año durante un periodo de 10 años. La figura 1 muestra la complejidad de un suceso que se producirá en el LHC.
Se espera una producción de 100 sucesos por segundo en ATLAS, siendo el tamaño de cada suceso de aproximadamente 1 MByte. Así pues, el reto tecnológico es proporcionar un acceso rápido y transparente a muestras de datos del tamaño del TB a los recursos de computación distribuidos por el mundo.

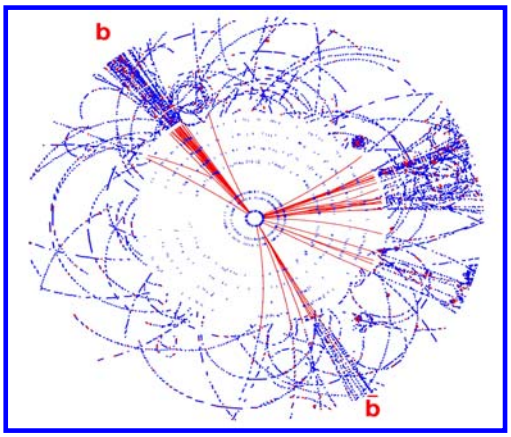

Figura 1: Trazas producidas por un suceso del LHC

La figura 2 representa los filtros que sufren los datos desde su producción hasta su análisis. Los recursos necesarios para procesar el volumen de datos que hemos comentado se estiman en el equivalente de una "*fábrica*" de 200.000 PC´s, un orden de magnitud por encima de los mayores supercomputadores actuales, y con claras dificultades técnicas, operativas y de financiación. Por otro lado, los equipos científicos son en muchos casos colaboraciones internacionales, con miembros distribuidos por todo el mundo. Áreas científicas como la Física de Altas Energías basan su desarrollo en estos proyectos, por lo que la organización de los correspondientes recursos de computación, es un desafío.

Al atacar este desafío existen varias soluciones entre las que se encuentran aquellas que se basan en las tecnologías GRID que propone agregar y compartir recursos de computación distribuidos entre diferentes organizaciones e institutos, a través de redes de alta velocidad.

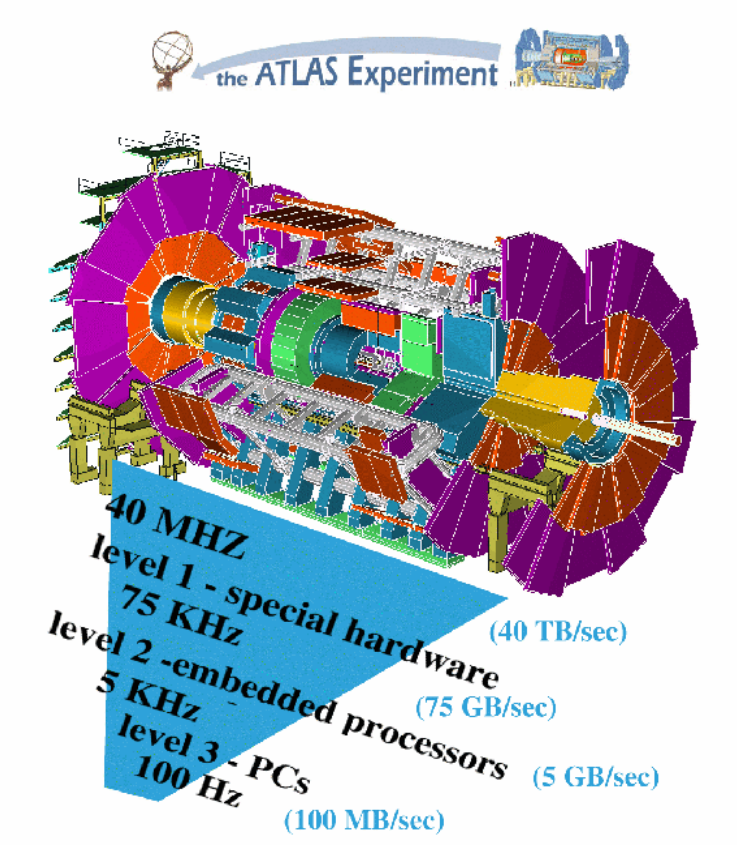

Figura 2: Tasas de datos producidos en los 3 niveles del Trigger

¿Qué es la Computación GRID?

El principio de la computación (en) GRID es la agregación de recursos computacionales heterogéneos distribuidos entre distintas organizaciones para formar un metaordenador[1]. Para hacer más fácil la utilización de estas infraestructuras distribuidas, surge la necesidad de establecer una arquitectura global que sea plasmada en la práctica en una serie de servicios básicos en forma de *middleware*, y que simplificarán el modo de desarrollar aplicaciones que puedan hacer uso de estas infraestructuras. Estos servicios básicos deben ofrecer entre otras cosas un acceso eficaz a los recursos computacionales, con la idea de compartir recursos individuales para proveer un valor añadido. La figura 3 representa la arquitectura del *middleware* definida en el proyecto DataGrid, y con ella se pretende ilustrar la estructura en capas: el middleware desarrollado sirve como pieza de conexión entre la capa de organización y gestión de la infraestructura (Fabric) y la capa de aplicaciones.

El GRID proporciona servicios accesibles por medio de un conjunto de protocolos e interfaces abiertos. Entre ellos se encuentran los siguientes servicios: gestión de recursos, gestión remota de procesos, librerías de comunicación, seguridad y soporte a monitorización. Dentro del GRID las personas que comparten los mismos fines y necesidades se pueden agrupar formando las organizaciones virtuales. Estas organizaciones virtuales son las que definen los privilegios de uso y de accesos a los

[1] Metaordenador se refiere a una "colección de ordenadores unidos como uno" para ofrecer una gran capacidad de cálculo y servicios.

recursos del GRID. Además, mantienen sus propias políticas de seguridad y gestión de recursos.

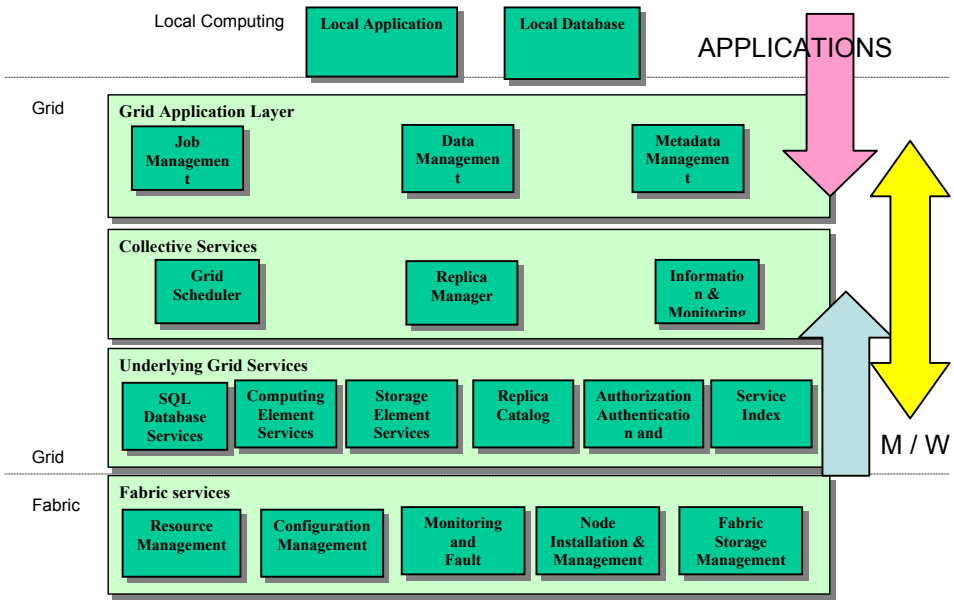

Figure 3: arquitectura del middleware

Líneas de trabajo en el IFIC

En el IFIC se llevan a cabo varios proyectos de I+D en tecnologías GRID dentro del *Grupo GRID y Computación de ATLAS*. Los temas de dicho grupo se están realizando dentro de Colaboraciones Internacionales gracias a proyectos Europeos como *CrossGrid* y *DataGrid*. Las líneas de trabajo relacionadas con la computación GRID pueden agruparse de la siguiente forma:

A) proyectos relacionados con el desarrollo y la implementación de tecnologías GRID:
- desarrollo del middleware en especial en nuevas herramientas y servicios GRID
- integración de todo middleware desarrolado en el proyecto.
- puesta a punta del banco de ensayos (testbed) del proyecto: desde el mantenimiento del testbed, pasando por las versiones de los programas, hasta un sistema de ayuda a los usuarios.
- utilización de dicho testbed en la ejecución de los diferentes aplicaciones (interactivas) incluidas en el proyecto (Meteorología, Física de Altas Energías, Inundaciones, etc.)

B) proyectos orientados hacia el computing del LHC (ATLAS)
- disponer de una infraestructura de calculo "a la GRID" para la producción de datos simulados con el fin de contrastar los datos con los modelos teóricos de Física de Partículas.

- comprobar y validar el modelo de computación, almacenamiento y de procesamiento masivo de datos con el fin de que la comunidad de físicos del LHC puedan realizar sus trabajos de análisis de una forma eficaz y flexible[2].

Infraestructura del IFIC

El IFIC cuenta con una infraestructura (Grupo de Ordenadores para el GRID (GOG)) significativa en recursos de computación gracias al esfuerzo de cooperación entre el CSIC y la Universitat de Valencia para construir una infraestructura de cálculo masivo. Su puesta a punto se ha conseguido gracias al trabajo elaborado por el personal de los Servicios Informáticos del IFIC, siendo responsables de su instalación y mantenimiento. En estos momentos (septiembre 2003) el GOG está realizando la transición hacia el modo de computación GRID aprovechando la experiencia adquirida en la instalación y gestión de 2 minigranjas (6-7 PC´s cada una) que están sirviendo de banco de pruebas de los desarrollos GRID que se realizan en los proyectos europeos DataGrid y CrossGrid..

GOG es una granja de ordenadores (ver figura 4) para la implementación de tecnologías GRID y se compone de 192 PCs Athlon (134 procesadores en el IFIC y 58 en el ICMOL (Instituto de Ciencia Molecular)) con formato 2U para rack comunicados con FastEthernet mediante equipos de conmutación de Gigabit Ethernet CISCO.

Figura 4: vista frontal del GOG

Cada PC posee un procesador AMD Athlon de 1.2 GHz y de 1.4 GHz con 1 GByte de memoria SDRAM y un disco duro UDMA100 de 40 GBytes.
El objetivo es conseguir una plataforma de computación para cálculo intensivo con paralelismo en el programa o en los datos que se integre en las iniciativas GRID. Esta infraestructura ha de servir para probar a una escala razonable las herramientas presentes y futuras para granjas masivas, tanto locales como en el GRID.

[2] Este avance en computación no será sólo utilizado por los físicos del LHC sino que se espera que sea utilizado por gran parte de físicos de Altas Energías.

Proyectos

El IFIC comenzó sus actividades GRID participando en el testbed del proyecto DataGrid. Además de este proyecto nuestro instituto está participando en:
- CrossGrid
- Proyecto Trienal Coordinado (LCG-ES)

En el año 2000 el programa IST lanza el proyecto European DataGrid (EDG) coordinado por el CERN, con el objetivo de construir la próxima generación de infraestructura de computación que permita el cálculo intensivo y análisis de datos compartidos a gran escala, desde cientos de Terabytes a Petabytes, entre comunidades científicas ampliamente distribuidas. Su interconexión esta basada en la red (Gigabit) Europea GEANT, en funcionamiento desde finales del año 2001. Uno de los objetivos básicos del proyecto es el desarrollo de un testbed distribuido por toda Europa, en el que España participa a través de los grupos de Física de Altas Energías: IFAE (Barcelona), IFIC (Valencia), IFCA (Santander), Universidad de Oviedo, UAM y CIEMAT (Madrid). Una visión general de los proyectos lanzados por la Information Society Technologies (IST) se encuentra en la figura que se muestra adjunta a la izquierda.

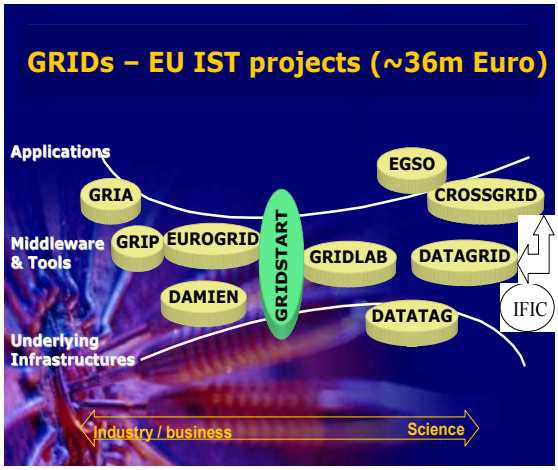

Figura 6: Relación de proyectos Grid internacionales clasificados según su orientación y etapa de implementación

En el caso de CrossGrid, uno de sus objetivos es modificar el middleware adaptándolo para poder ejecutar aplicaciones interactivas en un entorno Grid. En este proyecto se han definido 4 aplicaciones que utilizarán desarrollos Grid comunes, éstas son:

1) Simulación interactiva y visualización de un sistema biomédico
2) Sistema de apoyo a un equipo de crisis por inundaciones
3) Análisis de datos distribuidos en Física de Altas Energías
4) Previsión meteorológica y modelización de la contaminación atmosférica

En las figuras pueden observarse los diagramas correspondientes a la aplicación 1) (Simulación biomédica) y 4) (Predicción meteorológica).

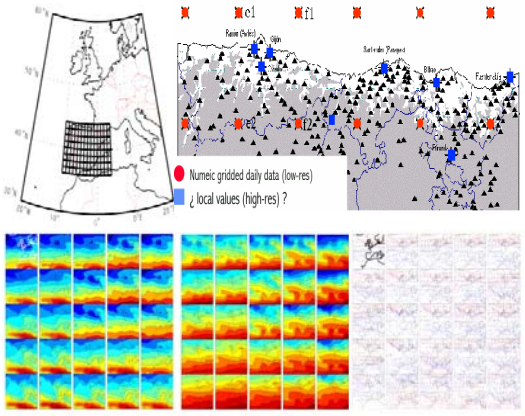

Figura 7: Meteorología, predicción meteorológica local en una zona del norte de España

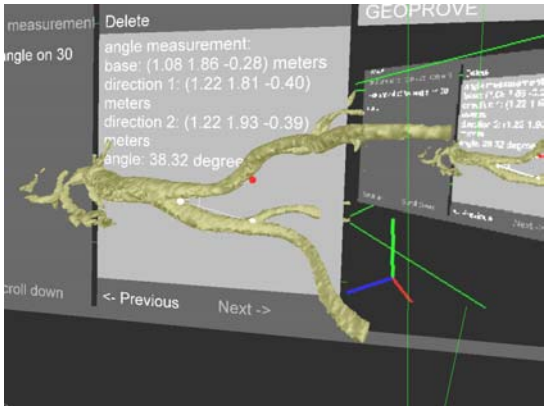

Figura 8: Vista de la estructura arterial en 3D

Otro objetivo es la extensión del testbed del proyecto DataGrid, que cuenta con una notable participación española. El CSIC (IFIC, IFCA y RedIRIS) es uno de los socios principales, responsable del apartado de "testbed" en el que participan también los siguientes centros españoles: la UAB (Barcelona) y la USC (Santiago) en colaboración con el CESGA. En la figura 9, parte izquierda, se refleja el banco de ensayo (testbed) del proyecto CrossGrid, el cual está coordinado por el CSIC, que incluye 15 centros de computación unidos a través de la red académica europea de alta velocidad (Gigabit) llamada Géant.

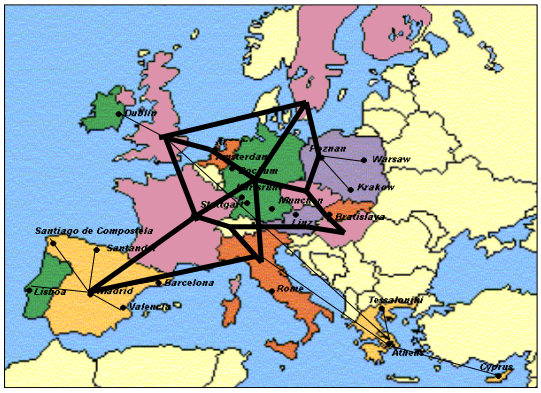

Figura 9: Banco de ensayo (testbed) del proyecto CrossGrid

El apoyo del programa nacional de Física de Altas Energías ha sido fundamental, y actualmente se cuenta con un proyecto nacional trienal, LCG-ES (MCyT), paralelo al proyecto LCG (LHC Computing Grid) del propio CERN. Este proyecto, LCG-ES, tiene como objetivo el desarrollo de una infraestructura DataGrid para la simulación y análisis de datos y, un Centro de Disponibilidad de Computación y Almacenamiento para el LHC. Las instituciones pertenecientes al Proyecto Nacional Trienal, LCG-ES, se pueden observar en la figura 10.

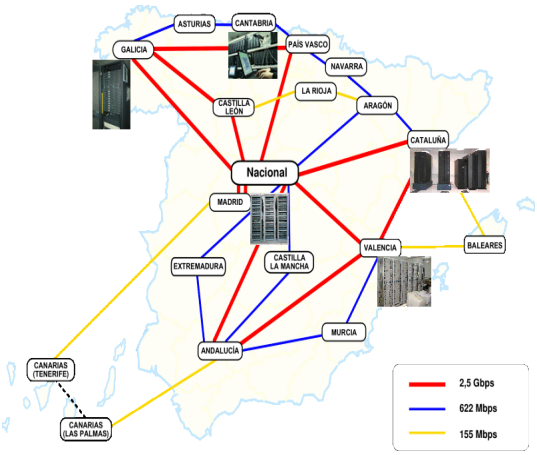

Figura 10: Grupos de AAEE y estructura de la red en España

IRIS-Grid y e-Ciencia

En las Jornadas Técnicas de RedIRIS que se desarrollaron en Salamanca (Noviembre del 2002) se lanzó el Grupo de Iniciativas para el Grid a nivel nacional (IRIS-GRID). Entre los integrantes de este Grupo se encuentran todos los pertenecientes al Proyecto LCG-ES (figura 9) anteriormente citados. En Septiembre de 2003 el IFIC organizó la III Reunión de IRIS-GRID en la que se pasó revista a los diferentes proyectos e iniciativas de Gris y e-Ciencia en nuestro país. Por e-Ciencia, se entiende una actividad científica de cualquier especialidad que se realiza mediante colaboraciones globales facilitada por Internet, que usa grandes colecciones de datos, recursos de computación del orden de TeraFLOPS y visualización de altas prestaciones. Es de destacar que el Programa Nacional de Tecnologías de la Información y las Comunicaciones (TIC) ha contemplado y apoyado en su última convocatoria explícitamente el arranque de estas iniciativas.

Por su parte el VI Programa Marco comunitario incluye un epígrafe específicamente destinado a la resolución de problemas científicos complejos, mediante tecnologías Grid. Entre las Expresiones de Interés (EoI) para dicho programa, la propuesta de un proyecto integrado denominado EGEE (Enabling Grids for e-Science and Industry in

Europe) coordinada por el CERN ha contado con el respaldo de gran número de instituciones, incluyendo en nuestro país varios centros del CSIC, Universidades (UAB, UAM, UCM, USC, UC, UM) y centros de investigación (IFAE, PIC, CIEMAT, o CESGA). La filosofía del EGEE se puede ilustrar en la siguiente figura.

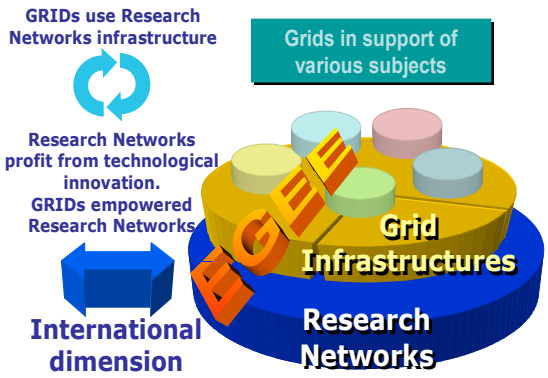

Figura 11: Esquema de la filosofía del EGEE

Este proyecto podría posibilitar la explotación de una infraestructura Grid común europea sobre la evolución de la red Géant, y a su vez proporcionaría el marco para las redes de excelencia o proyectos orientados a las diferentes aplicaciones en e-Ciencia. La organización de la participación española, incluyendo el planteamiento de un sistema de centros de e-Ciencia es una de las actividades en marcha.

Conclusiones

El Grid[3] va a suponer la siguiente revolución en la Sociedad de la Información de comienzos del siglo XXI (si tenemos en cuenta que INTERNET ha supuesto la revolución anterior) y ello va a influir en nuestra forma de trabajar y de concebir las infraestructuras computacionales, en particular, para los proyectos científicos (multidisciplinario, computación de recursos, etc).

Los beneficios que se obtienen de la utilización de las tecnologías Grid son tanto científicos como tecnológicos y es un entorno especialmente propicio para colaboraciones entre comunidades científicas y el sector empresarial e industrial.

[3] Para más información consultar la siguiente dirección:
http://ific.uv.es/grid