Beneficios del uso de la tecnología grid computing en bioinformática usando la infraestructura de IRISGrid

Benefits Achieved in Bioinformatics by Using Grid Computing Technology within IRISGrid Infrastructure

A. Fuentes, J. L. Vázquez, E. Huedo, R. S. Montero e I. M. Llorente

Resumen

La computación Grid ha tenido un fuerte auge en los últimos años, aunque esta evolución todavía no se ha plasmado en un uso real y amplio por parte de la comunidad científica y académica. La falta de aplicaciones para su ejecución en Grid es un hecho, debido fundamentalmente a que la migración de éstas hacía este nuevo modelo de computación debe venir precedida por el convencimiento de los beneficios de esta migración. En este artículo, pretendemos demostrar las mejoras que supone el uso de la tecnología grid al ejecutar una aplicación del campo de la Bioinformática. Como veremos, la mejora del rendimiento usando grid es considerable respecto a su ejecución en un cluster local. Además, explicaremos las infraestructuras y middleware involucrados en este experimento, así como de todos los elementos que han intervenido en él.

Palabras clave: computación Grid, IRISGrid, computación distribuida.

Summary

Grid Computing has gained importance in the last years, although this evolution has not been seen in real use by the Academic Comunity. The lack of aplications to be executed in Grid is a real fact due basically to the lack of confidence in the benefits this migration may carry with. What is pretended to be shown in this paper is the improvement achieved by using grid computing when executing a Bioinformatic application. As shown the improvement is important with respect to the execution in a local cluster. Apart from this, the infrastructures and middleware involved in this test is also explained as well as the elements involved.

Keywords: Grid Computing, IRISGrid, Distributed Computing

La falta de aplicaciones para su ejecución en Grid es un hecho, debido fundamentalmente a que su migración hacía este nuevo modelo de computación debe venir precedida por el convencimiento de su beneficio

1.- Introducción a la computación Grid

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*. De este modo, los sistemas distribuidos se pueden emplear como un único sistema virtual en aplicaciones intensivas en datos o con gran demanda computacional.

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 (gestión de recursos, gestión remota de procesos, librerías de comunicación, seguridad, soporte a monitorización,...). Las organizaciones virtuales que se interconectan por medio de un Grid tienen que mantener sus propias políticas de seguridad y gestión de recursos. Esto significa que la tecnología usada para construir un Grid es complementaria a otras tecnologías aprovechando los recursos distribuidos en la intranet de una organización.

IRISGrid, la Iniciativa Nacional de Grid, proporciona las infraestructuras necesarias para que los diferentes recursos distribuidos puedan ser usados de forma colaborativa y coordinada para la ejecución distribuida de aplicaciones entre las diferentes organizaciones académicas y de investigación en España. Además, diferentes grupos españoles están participando en otras iniciativas Grids en Europa, tales como EGEE, CrossGrid y LCG, situando a España en una posición de conocimiento privilegiada en esta tecnología emergente.

A pesar de los esfuerzos que se están realizando en torno al desarrollo de la tecnología Grid, no podemos hablar de madurez tecnológica, de hecho, observando el panorama internacional, los problemas de interconexión y compatibilidad de Grids son patentes. Las infraestructuras Grids de las diferentes iniciativas a nivel mundial integran middleware y software diferentes, que en términos generales, no suelen ser compatibles e interoperables. En esta línea, los esfuerzos que se están realizando a nivel mundial están encaminados a conseguir unos estándares que nos permitan esta compatibilidad tecnológica. Además, todavía existen pocas aplicaciones preparadas para usar los recursos Grids de forma óptima, y por tanto, el trabajo de portar aplicaciones de las diferentes áreas de conocimiento para ser ejecutadas en las infraestructuras Grids tiene que ser una línea prioritaria.

El objetivo pretendido en este artículo es mostrar los resultados de la ejecución de una aplicación Grid, del ámbito de la Bioinformática, en las infraestructuras proporcionadas por IRISGrid, teniendo los centros de recursos distribuidos a lo largo de la geografía española. Para ello, comenzaremos describiendo los diferentes actores que intervienen durante la ejecución, como IRISGrid, la iniciativa nacional de Grid, introducida en el siguiente apartado, la distribución de recursos y la descripción del experimento, explicada en el punto 4, para luego pasar a una descripción sobre la aplicación que será ejecutada, punto 5. En el apartado 6, se muestran los resultados de la ejecución de ésta aplicación, y se realizan algunas comparativas en relación a su ejecución en un entorno local, lo cual nos llevará a esquematizar las conclusiones derivadas de los resultados obtenidos.

IRISGrid cuenta con la participación de más de 50 grupos en España, dentro de todos los ámbitos de conocimiento. Coordinada por RedIRIS, los objetivos de IRISGrid son

2.- IRISGrid: iniciativa nacional de Grid

IRISGrid, la iniciativa española en Grids, nació en el año 2002, a partir de la proposición de diversos grupos nacionales interesados en esta tecnología en España. A día de hoy, IRISGrid cuenta con la participación de más de 50 grupos en España, dentro de todos los ámbitos de conocimiento. Coordinada por RedIRIS, los objetivos de IRISGrid son:

1.- Integrar a los diferentes grupos interesados en las tecnologías grid en España, y su interés.
2.- Mantener unas infraestructuras nacionales Grid, que permitan la correcta operatividad y uso de un Grid de investigación, y que facilite el desarrollo de middleware y aplicaciones por parte de los grupos españoles y que asegure asimismo la integración y acercamiento a esta nueva infraestructura.
3.- Coordinación de los diferentes proyectos Grid en España y sus infraestructuras con el objeto de asegurar la interoperatividad de estos.

España cuenta con diferentes grupos con bastante experiencia en la tecnología Grid, debido a su participación en proyectos europeos e internacionales, lo cual permite contar a IRISGrid con un alto nivel de asesoramiento. En esta línea, la situación actual del estado de la tecnología Grid en las diferentes áreas de conocimiento y su uso, fue uno de los objetivos iniciales de IRISGrid.

Actualmente, IRISGrid dispone de unas infraestructuras operativas, pero en estado de redefinición, con el objetivo de abordar retos futuros. El trabajo se está realizando por un comité de personas integrantes de la Acción Especial de Middleware en el marco de IRISGrid, que tiene como objetivo el desarrollo de nuevo Middleware, evaluación y asesoramiento en la implantación de éste en el TestBed de IRISGrid.

Los recursos disponibles en IRISGrid y mostrados a continuación en el siguiente punto, muestran la verdadera naturaleza del concepto de computación Grid: su naturaleza descentralizada.

3.- Descripción del experimento y ejecución distribuida

El experimento Grid, realizado en el marco de IRISGrid implicó un gran número de recursos, tanto materiales como humanos que pudieron hacerlo factible, además de contar con la colaboración de las diferentes instituciones participantes en IRISGrid, y en concreto, de aquellas que participaron directamente en él.

Durante la preparación de los experimentos, uno de los retos importantes fue el no incidir en las configuraciones particulares de aquellas instituciones que ya trabajan en Grids, y poseen sus recursos en producción integrados en diferentes proyectos. Así, se consiguió realizar este experimento conservando uno de los puntos claves que debe poseer cualquier entorno Grid, la heterogeneidad. Las infraestructuras usadas fueron:

Durante la preparación de los experimentos, uno de los retos importantes fue el no incidir en las configuraciones particulares de aquellas instituciones que ya trabajan en Grids

1.- Infraestructuras nacionales de IRISGrid (Autoridad de Certificación, Sistemas de Información Global (GIIS), Sistemas de Monitorización).

2.- Autoridad de Certificación de DataGrid.

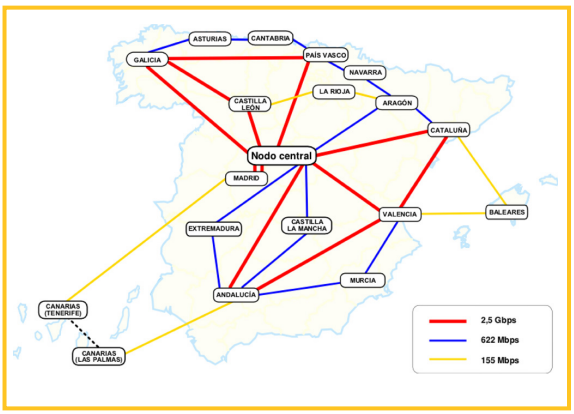
3.- Recursos de instituciones participantes en los proyectos EGEE, CrossGrid y LCG.

4.- Recursos de centros integrados en IRISGrid.

5.- GridWay. El planificador de procesos distribuidos entre los diferentes centros participantes y que permitió asegurar la heterogeneidad del experimento.

6.- Aplicación cedida por el Centro de Astrobiología para el cálculo de proteínas.

7.- Todos los centros que aparecen en la siguiente tabla están conectados a RedIRIS, la Red nacional y de investigación española.



INSTITUCIÓN	ARQUITECTURA	VELOCIDAD	S.O	Nº NODOS	JOB. MGR.	PROYECTO
RedIRIS	Intel	2,6Ghz	Linux	4	Fork	IRISGrid
Dacya-UCM	Intel	2,5Ghz	Linux	5	Fork	IRISGrid
CAB-INTA	Alpha	450Mhz	Linux	30	PBS	IRISGrid
CESGA	Intel	3,2Ghz	Linux	80	PBS	IRISGrid
IFCA	Intel	1,3Ghz	Linux	34	PBS	EGEE/CrossGrid
IFIC	Intel	1,2Ghz	Linux	117	PBS	EGEE
CNB	Intel/Xeon	2Ghz	Linux	8	PBS	EGEE
IMEDEA	AMD	800Mgz	Linux	14	PBS	IRISGrid
FDI-UM	Intel Xeon	2,4Ghz	Linux	3	Fork	IRISGrid
CIEMAT	Intel/Xeon	2,8Ghz	Linux	6	PBS	LCG
DFT-UAM	Intel	2,6Ghz	Linux	49	PBS	EGEE
PIC	Intel	2,8Ghz	Linux	69	PBS	EGEE
BIFI-UNIZAR	Intel	3,2Ghz	Linux	50	SGE	EGEE

Tabla 1

Un aspecto importante y común entre todos los recursos participantes, es que todos tienen como middleware básico Globus, en sus diferentes versiones pre-webservices. El detalle de las máquinas participantes, pueden ser observadas en la siguiente tabla.

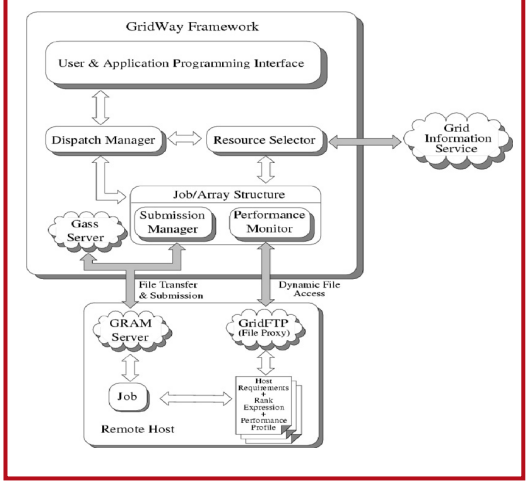
4.- Arquitectura de GridWay

El núcleo de la herramienta *GridWay* es un *agente de envío* que realiza automáticamente todas las fases de la planificación de un trabajo y vela por que su ejecución sea correcta y eficiente:

- La ejecución adaptativa del trabajo se realiza mediante una *planificación dinámica*. Una vez que el trabajo es asignado inicialmente a un recurso se re-planifica periódicamente para descubrir otros más idóneos, cuando se detecta un deterioro en su rendimiento o cuando se produce un fallo.
- El rendimiento real de la aplicación se evalúa periódicamente comprobando el tiempo de suspensión acumulado, y mediante un programa externo (*performance evaluator*) que examina un perfil de rendimiento generado por la aplicación.
- La selección y descubrimiento de recursos se realiza mediante otro programa (*resource selector*) que construye una lista de recursos factibles, según los requisitos del trabajo, ordenados atendiendo a las preferencias de la aplicación.

La arquitectura del agente de envío se muestra en la figura adjunta. El usuario interactúa con la herramienta a través de un interfaz de usuario, que maneja sus peticiones (*submit*, *kill*, *stop*, *resume*...) y las reenvía al *dispatch manager*. El *dispatch manager* se despierta periódicamente en cada intervalo de planificación e intenta enviar los trabajos pendientes a recursos del Grid y es también responsable de decidir si la migración de los trabajos re-planificados es factible y merece la pena. Una vez que un trabajo es asignado a un recurso, se arranca un *submission manager* y un *performance monitor* para vigilar su ejecución correcta y eficiente.

Fig. 3: ARQUITECTURA DEL AGENTE DE ENVÍO



5.- Aplicación

La aplicación usada en el experimento se engloba en el campo de la Bioinformática. Ésta tiene por objeto el cálculo de la predicción de la estructura y las propiedades termodinámicas de una proteína partiendo de sus secuencias de aminoácidos. El algoritmo, presentado en "5ª Critical Assesment of Techniques for Protein Structure Prediction (CASP5)" alinea intervalos de la secuencia a analizar con todas las 6150 diferentes estructuras almacenadas en la Protein Data Bank (PDB). A continuación se realiza una comparativa entre secuencia y estructura en base a un modelo simplificado de una función de energía libre más un intervalo de error. La comparativa con menor puntuación se considerada una predicción si satisface unos requerimientos de calidad. Es en estos casos cuando el

La aplicación usada en el experimento se engloba en el campo de la Bioinformática y tiene por objeto el cálculo de la predicción de la estructura y las propiedades termodinámicas de una proteína partiendo de sus secuencias de aminoácidos

algoritmo puede ser utilizado para estimar parámetros termodinámicos de la secuencia a estudiar, como por ejemplo, la energía libre de *Gordin* y el intervalo de energía normalizado.

Para aumentar la velocidad de análisis y reducir los datos necesitados, los ficheros PDB son preprocesados para extraer las matrices características que proporcionan una representación reducida de la estructura de las proteínas. El algoritmo, aplicado dos veces, primero para seleccionar las 100 mejores estructuras candidatas, y segundo con parámetros específicos que permitan una búsqueda más fina del alineamiento óptimo.

El uso de la tecnología Grid Computing, como se demuestra en el experimento, es muy interesante para esta clase de problemas de Bioinformática. Como la búsqueda de coincidencias no es secuencial, podemos enviar partes a diferentes recursos computacionales.

El experimento aplica el algoritmo para la predicción de las propiedades termodinámicas de familias de proteínas heterogéneas, es decir, proteínas que tienen la misma función en diferentes organismos.

Los resultados conseguidos en el experimento muestran los beneficios que se pueden obtener por diferentes áreas de conocimiento al usar de forma habitual esta tecnología

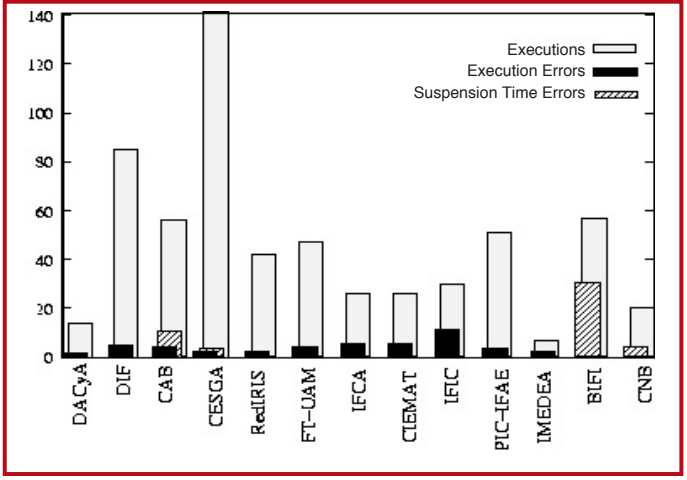
6.- Resultados

Los resultados conseguidos en el experimento, que consistió en la ejecución de la aplicación explicada anteriormente, con las infraestructuras detalladas muestra, como veremos, los beneficios que se pueden obtener por diferentes áreas de conocimiento al usar de forma habitual esta tecnología.

Así, mostraremos el número de ejecuciones que fueron planificadas en los diferentes centros y los fallos de ejecución en cada uno de ellos, el tiempo ganado por cada una de estas instituciones usando el Grid en relación a sus recursos particulares, así como la distribución de tiempos por institución en el envío y ejecución de los diferentes Jobs.

En la figura 4 que aparece a continuación, observamos las distintas ejecuciones que se produjeron durante el experimento en los diferentes recursos que el planificador GridWay eligió de forma dinámica. Al mismo tiempo también podemos observar el número de errores que se produjeron en cada uno de estos recursos, que implican una replanificación de estas ejecuciones en nuevos recursos por parte del planificador. Como podemos observar en la figura 5 que aparece en la siguiente página el impacto de la transferencia de ficheros a la hora de replanificar estos trabajos es también una cuestión importante, ya que supone, una parte adicional en el tiempo de ejecución global de los trabajos realizados.

FIG. 4: DIFERENTES EJECUCIONES



Los ficheros que implican la ejecución de esta aplicación son:

1.- Ejecutable, de 0.5MB, implementando en las diferentes arquitecturas existentes en IRISGrid

2.- Los ficheros PDB comprimidos, con el objeto de reducir el tiempo de transferencia, que asciende a 12MB

3.- Fichero de parámetros, de 1KB

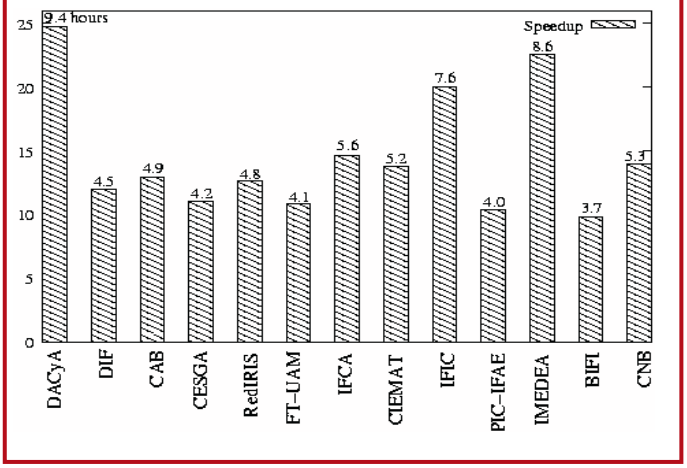
La ganancia que supondría el uso de la tecnología Grid en cada una de las instituciones se muestra en este experimento en la figura 5, donde observamos la ganancia en velocidad que tendrían cada una de las instituciones al realizar este experimento en el TestBed de IRISGrid en comparación a ejecutarlo localmente. La ganancia mostrada en la tabla, se calcula en base a la siguiente fórmula:

Tiempo Ganado=Tiempo Ejecucion Local Site/Tiempo en ejecución en el Grid

La ganancia mostrada en cada una de estas instalaciones podría ser del orden de 100 veces mayor en aquellos centros de investigación e instituciones donde su capacidad de cómputo sea limitada.

La ganancia mostrada en cada una de estas instalaciones podría ser del orden de 100 veces mayor en aquellos centros de investigación donde su capacidad de computo sea limitada

FIG. 5: GANANCIA EN VELOCIDAD USANDO LA TECNOLOGÍA GRID



7.- Conclusiones

A pesar de que la tecnología Grid no es madura y se está trabajando mucho en ella, vemos como ésta ya es una realidad, y diferentes centros utilizando aplicaciones adaptadas a este modelo distribuido pueden obtener muchos beneficios.

El experimento, realizado con diferentes centros integrados en IRISGrid, ha supuesto una punta de lanza en éste sentido, y su rendimiento será aumentado paulatinamente conforme se vayan integrando a la iniciativa más recursos computaciones, aportados desde las diferentes instituciones. Si

bien, el uso y capacitación en esta tecnología no es amplio, dada las ventajas que proporciona en diferentes ámbitos de la ciencia, este trabajo debe ser cada vez mayor en este sentido.

Como aspecto importante, señalar la heterogeneidad conseguida con el uso de la herramienta GridWay, que nos permitió la ejecución en diferentes centros aun cuando estos están en otros proyectos Grids con una arquitectura diferente a la de IRISGrid.

Además, como nota final, la apuesta por IRISGrid es una realidad a la que se deben sumar más instituciones, con el objeto de hacer esta tecnología alcanzable a todos los potenciales usuarios, además de integradora de los diferentes proyectos nacionales Grid en España como Iniciativa Nacional que es.

La apuesta por IRISGrid es una realidad a la que se deben sumar más instituciones, con el objeto de hacer esta tecnología alcanzable a todos los potenciales usuarios

Agradecimientos

El experimento que ha dado lugar a este artículo no hubiera sido posible sin la inestimable colaboración de los siguientes centros de investigación y universidades a quien agradecemos:

1.- RedIRIS, Red académica y de investigación española
2.- Dpto de Arquitectura de Computadores y Automática. UCM
3.- CAB, Centro de Astrobiología. CSIC/INTA
4.- CEPBA, Centro de Paralelismo de Barcelona
5.- CESGA, Centro de Supercomputación de Galicia
6.- IFCA, Instituto de Física de Cantabria
7.- IFIC, Instituto de Física Corpuscular
8.- CNB, Centro Nacional de Biotecnología
9. IMEDEA, Instituto Mediterráneo de Estudios Avanzados
10. Facultad de Informática. UM
11. CIEMAT, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
12. Dpto. de Física Teórica. UAM
13.- PIC. Puerto de Información Científica
14.- Grupo de Redes y Computación de Altas Prestaciones (GRyCAP). UPV
15.- BIFI. Instituto de Biocomputacion y Fisica de Sistemas Complejos. UNIZAR
16.- Dpto. de Informática y Matemática. UNAV
17. Servicio Central de Informática. UMA
18. ATC, Departamento de Electrónica y Computadores. UNICAN

Antonio Fuentes
(antonio.fuentes@rediris.es)
RedIRIS
José L. Vázquez, Rubén S. Montero
(jlvazquez@fdi.ucm.es),(rubensm@dacya.ucm.es)
Lab. de Computación Avanzada, Simulación y Aplicaciones Telemáticas,
Centro de Astrobiología (CSIC-INTA)
Eduardo Huedo
(huedoce@inta.es)
Ignacio M. Llorente
(llorente@dacya.ucm.es)
Centro de Astrobiología (CSIC-INTA) y
Dpto. de Arquitectura de Computadores y Automática - UCM