

Aplicaciones Grid en una organización virtual de química computacional

Grid Applications in a Computational Chemistry Virtual Organization

◆ S. Reyes, C. Muñoz-Caro y A. Niño

Resumen

En este trabajo se propone una nueva metodología para obtener, de forma automática, la hipersuperficie de energía potencial molecular. Para ello, se ha desarrollado una organización virtual de química computacional que agrupa recursos computacionales de diferentes equipos de investigación, localizados geográficamente en diferentes países. Sobre estos recursos se ha desarrollado una aplicación Grid, basada en el *middleware* resolución automática del problema.

Palabras clave: Programación Grid, organización virtual, hipersuperficie de energía potencial.

Summary

This work presents a new methodology to automatically obtain, the molecular potential energy hypersurface. Thus, a computational chemistry virtual organization has been developed in order to share computational resources from different research groups, geographically located in different countries. On these resources, a Grid application based in the *middleware* has been developed for the automatic resolution of the problem.

Keywords: Grid Programming, Virtual Organization, Potential Energy Hypersurfaces.

1. Introducción

El desarrollo de un hamiltoniano rotovibracional anarmónico para moléculas de tamaño arbitrario, que considere todos los grados internos de libertad, describa acoplamiento entre modos de vibración, e interacción rotación-vibración presenta un gran interés desde el punto de vista físico-molecular. El elemento clave es la determinación de una hipersuperficie de energía potencial para el movimiento nuclear, así como la variación de la estructura molecular sobre dicha hipersuperficie. A tal efecto, sería interesante poder realizar de forma automática una exploración masiva de la hipersuperficie de energía potencial como función de la estructura. Este proceso de *mapping* de la hipersuperficie se abordaría, en el marco de la aproximación de Born-Oppenheimer, usando los resultados de energía total de cálculos de estructura electrónica molecular [1]. Esta metodología podría aplicarse a la identificación y caracterización de moléculas de interés astrofísico y astrobiológico.

En términos computacionales, se trataría de trabajar a un nivel de abstracción mayor que el actual, donde los cálculos de estructura electrónica no serían el fin, sino el medio para construir nuevos modelos de movimiento nuclear. Dada la independencia de datos entre los cálculos de estructura electrónica podríamos automatizarlos usando técnicas de programación paralela sobre Grid de computadores. Así, utilizaríamos los recursos de los grupos de investigación involucrados en el trabajo, integrándolos en un solo sistema virtual. De esta forma, aprovecharíamos de manera más eficiente los recursos disponibles.

2. La organización virtual de química computacional

La tecnología Grid surgió ante la posibilidad de aprovechar los recursos disponibles en los sistemas informáticos conectados a Internet. Como los recursos de grupos de investigación, geográficamente

◆
En este artículo se propone una nueva metodología para obtener, de forma automática, la hipersuperficie de energía potencial molecular

◆
Se ha desarrollado una organización virtual de química computacional que agrupa recursos de diferentes equipos de investigación sobre los que se ha desarrollado una aplicación Grid, basada en el *middleware*



Actualmente los Grid computacionales se pueden estructurar como organizaciones virtuales de usuarios

Hasta el momento se ha utilizado una política de certificación privada para la organización virtual C2VO

distribuidos o no, o bien están conectados a Internet o podrían conectarse sin problemas. ¿Por qué no aprovechar los recursos de grupos de investigación afines? Es decir, desarrollar un metasisistema agrupando recursos computacionales (procesamiento, *software*, datos) de distintos grupos de investigación. Estos metasisistemas definen el Grid computacional. Actualmente estos Grid computacionales se pueden estructurar como organizaciones virtuales de usuarios, término acuñado en el trabajo seminal de Foster y col. sobre tecnología Grid [2].

Para el presente trabajo, se ha desarrollado una Organización Virtual de Química Computacional (*Computational Chemistry Virtual Organization*, C2VO) cuya infraestructura, véase figura 1, está formada por tres nodos. Dos de ellos están geográficamente localizados en Ciudad Real y pertenecen al grupo de Química Computacional y Computación de Alto Rendimiento de la Universidad de Castilla La-Mancha (QCyCAR-UCLM); el tercero está localizado en Puebla (México) y pertenece al Laboratorio de Química Teórica de la Benemérita Universidad Autónoma de Puebla (LQT-BUAP). Un cuarto nodo, accesible bajo demanda, está localizado geográficamente en Barcelona, en el Centro Nacional de Supercomputación (*Barcelona Supercomputing Center*, BSC-UPC).

Cabe destacar la heterogeneidad del Grid. En cuanto a los nodos Grid, los de Ciudad Real, Tales y Hermes, son *clusters* de computadores con placas monoprocesadoras Pentium IV; el nodo de Puebla, Popocatepetl, es un *cluster* de computadores con placas biprocesadoras AMD 64. Ambos nodos implementan el sistema de gestión y configuración Rocks [3]. A su vez, el nodo de Barcelona, Kadesh, es un *cluster* de computadores POWER PC gestionado por AIX [4]. Por otro lado, los gestores de colas son también heterogéneos. Los nodos de Ciudad Real utilizan OpenPBS [5], el nodo de Puebla SGE [6] y el de Barcelona LoadLeveler [7]. Para disponer de los servicios Grid básicos todos los nodos incorporan el middleware Globus Toolkit 2.X [8].

Con vistas al futuro, se tiene prevista la incorporación de nuevos nodos, permanentes, y bajo demanda, al sistema (ver figura 1). Los nodos permanentes corresponderían al Grupo de Astrofísica Molecular e Infrarrojo (DAMIR-CSIC) en Madrid, la Universidad Autónoma Metropolitana, UAM, en México DF, y la Thompson Rivers University, Canadá. Bajo demanda se incorporará, para futuras pruebas, el Centro de Supercomputación de Galicia, CESGA, en Santiago de Compostela.

Hasta el momento se ha utilizado una política de certificación privada para la organización virtual C2VO aunque en su momento se acordó comenzar las pruebas con la nueva PKI de IRISGrid [9].

FIGURA 1: ESTRUCTURA DE LA ORGANIZACIÓN VIRTUAL DE QUÍMICA COMPUTACIONAL (C2VO)



3. Aplicaciones

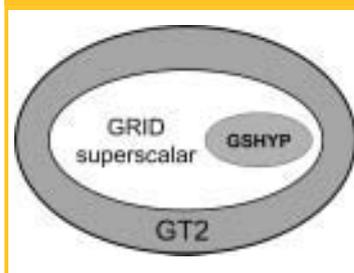
Para la resolución del problema de la generación automática de la hipersuperficie de energía potencial molecular ha sido necesario desarrollar dos aplicaciones: GSHYP, que es la encargada de generar la hipersuperficie, y GSSTAT, que se utiliza para monitorizar el proceso. A continuación, se detalla la estructura funcional de cada una de ellas.

3.1. GSHYP

GSHYP es una aplicación Grid (véase figura 2), desarrollada sobre el *middleware* [10] usando Globus Toolkit 2. Esta aplicación utiliza una metodología general para la determinación de hipersuperficies de energía potencial en Grids computacionales. El proceso se organiza en tres pasos:

1. Optimización de la geometría inicial de la molécula hasta alcanzar su estado de equilibrio. En la estructura general de la aplicación este módulo se ha denominado OPTHYP (véase figura 3).
2. Generación de las estructuras moleculares. Este paso consiste en la obtención de un conjunto de estructuras moleculares que definan la hipersuperficie de energía potencial. Este módulo ha sido denominado GENHYP (figura 3).
3. Realización en Grid de los cálculos de estructura electrónica, uno por cada estructura generada en el paso 2, e integración de datos. Cuando finaliza cada cálculo se filtra la información de salida necesaria y se integra en un único archivo denominado "molecula.hyp". Este módulo se ha denominado GSGMS y ha sido implementado (véase figura 3).

FIGURA 2: LA APLICACIÓN GSHYP EN CONTEXTO

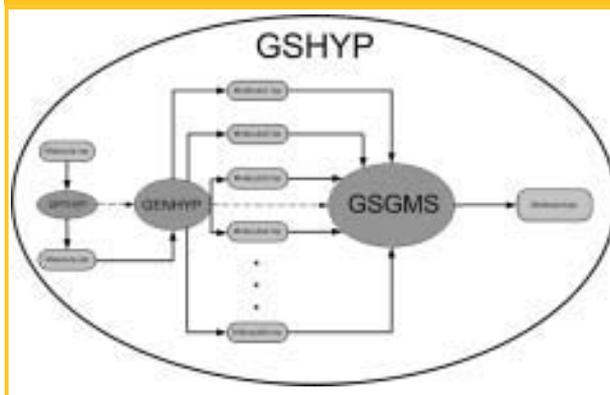


◆
GSHYP es una aplicación Grid que utiliza una metodología general para la determinación de hipersuperficies de energía potencial en Grids computacionales

Los cálculos de estructura electrónica molecular, llevados a cabo en los pasos 1 y 2 del algoritmo anterior, han sido realizados con el paquete de libre distribución GAMESS [11].

Las pruebas de la metodología anterior se han llevado a cabo con la molécula de acetona (véase figura 4). Para la obtención de la hipersuperficie de energía potencial se han considerado cuatro grados de libertad: los ángulos planos α y β y los diedros \varnothing_1 y \varnothing_2 , ver figura 4. El barrido sobre estos cuatro ángulos ha supuesto 1.120 cálculos individuales de estructura electrónica generados y resueltos de forma automática por la herramienta GSHYP [12]. Se llevaron a cabo dos pruebas con supuestos diferentes en cuanto al número de procesadores utilizados:

FIGURA 3: ESTRUCTURA FUNCIONAL DE LA APLICACIÓN GSHYP



◆
Las pruebas de la metodología anterior se han llevado a cabo con la molécula de acetona

1. En el primer caso, se utilizaron 32 procesadores distribuidos uniformemente en cada nodo del Grid C2VO, a razón de 8 procesadores por nodo Grid.
2. En el segundo caso, se usaron 64 procesadores distribuidos de la siguiente forma: 11 en Tales, 11 en Hermes, 14 en Popocatepeti y 28 en Kadesh.

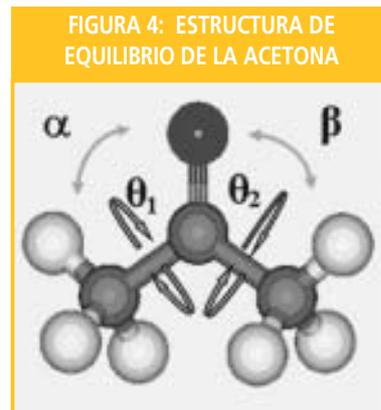
Los resultados obtenidos y el tiempo empleado en cada supuesto se recogen en la Tabla I. El tiempo total para el supuesto primero fue de 31 horas y 15 minutos, mientras que en el segundo se redujo a



17 horas. En el límite, un procesador por cálculo, el tiempo sería algo superior al necesario para obtener el resultado de un trabajo, debido a la sobrecarga asociada al tráfico en la red [12].

| | Caso 1 | Caso 2 |
|------------------|---------|--------|
| Tales | 310 | 223 |
| Hermes | 290 | 218 |
| Popocatepetl | 286 | 262 |
| Kadesh | 234 | 417 |
| Tiempo Total (s) | 112.500 | 61.500 |

Tabla I. Número de cálculos de estructura molecular finalizados en cada nodo del Grid en las pruebas descritas en el texto. Incluye el tiempo total (en segundos) para la resolución de los 1120.



GSSTAT es un monitor de trabajos para aplicaciones Grid

Consideremos una aplicación ejecutando n trabajos sobre m nodos de un Grid. Si no se dispone de un monitor de trabajos, no podríamos saber cómo evoluciona la ejecución de la aplicación. Por esta razón hemos construido GSSTAT, un monitor Grid (figura 5), desarrollado con el Perl Commodity Grid [13]. GSSTAT es un monitor de trabajos para aplicaciones Grid desarrolladas con el *middleware*. Se ha diseñado para que se ejecute junto con la aplicación Grid y para que muestre la información con un formato similar a los monitores que implementan los gestores de colas (figura 6). En nuestro caso, GSSTAT usa información generada por GSHYP.

La información mostrada por el monitor GSSTAT, está estructurada en dos secciones: la primera muestra información general, y la segunda la relativa al estado de los trabajos, en cada uno de los Workers

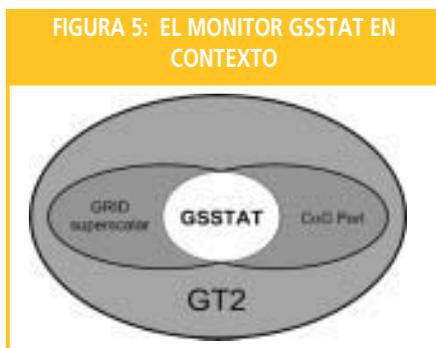


FIGURA 6: VISUALIZACIÓN DEL RESULTADO DE GSSTAT

```

=====
GRID Supervisor STATUS Job:
MASTER: tales.inf-cr.acm.es
=====
TIME: 12 h 14 m 36 s
TOTAL Jobs: 1120
DONE: 809 ACTIVE: 94 PENDING: 0 SUSPENDED: 0 FAILED: 0
REMAINING Jobs: 247
=====
WORKER: hermes.inf-cr.acm.es
DONE: 189 ACTIVE: 11 PENDING: 0 SUSPENDED: 0 FAILED: 0
WORKER: tales.inf-cr.acm.es
DONE: 189 ACTIVE: 11 PENDING: 0 SUSPENDED: 0 FAILED: 0
WORKER: kadesh.ccpba.upc.es
DONE: 295 ACTIVE: 25 PENDING: 0 FAILED: 0
WORKER: popocatepetl.fiqim.buap.mx
DONE: 189 ACTIVE: 14 PENDING: 0 SUSPENDED: 0 FAILED: 0
=====

```

La información mostrada por el monitor GSSTAT, está estructurada en dos secciones (figura 6). La primera muestra información general de la aplicación GSHYP: en primer lugar el nodo Máster desde donde se ha lanzado la ejecución de la aplicación; en segundo lugar el tiempo de ejecución de la aplicación; en tercer lugar el estado de los trabajos: finalizados, activos, pendientes, suspendidos, restantes, etc. La segunda sección está dedicada a mostrar la información, relativa al estado de los trabajos, en cada uno de los Workers. En las pruebas realizadas, GSSTAT ha permitido la monitorización de los 1.120 cálculos sin problemas de actualización de información.

4. Conclusiones

En este trabajo se muestra la utilidad de la programación Grid para el desarrollo de aplicaciones científicas de cálculo intensivo. En particular, para el campo físico-molecular, se demuestra la

factibilidad de la realización en Grid de los miles de cálculos de estructura electrónica necesarios para determinar una hipersuperficie de energía potencial molecular. Se observa que la existencia de conexiones intercontinentales en el Grid no supone una sobrecarga temporal, en tanto en cuanto se realice un procesamiento local y un filtrado de información en cada nodo del Grid, a fin de reducir el tamaño de los ficheros generados, en cada cálculo puntual y transferidos por la red.

Sebastián Reyes, Camelia Muñoz-Caro
 (Sebastian.Reyes@uclm.es), (Camelia.Munoz@uclm.es)
Alfonso Niño
 (Alfonso.Nino@uclm.es)
 Grupo de Química Computacional y
 Comp. de Alto Rendimiento (QCyCAR)
 E. S. de Informática - Universidad de Castilla-La Mancha



En este trabajo se muestra la utilidad de la programación Grid para el desarrollo de aplicaciones científicas de cálculo intensivo

Referencias

- [1] J Jensen, F. "Introduction to Computational Chemistry", John Wiley & Sons, 1999. ISBN: 0471980554
- [2] Foster, I.; Kesselman, C.; Tuecke, S.. "The Anatomy of the Grid: Enabling Scalable Virtual Organizations". "International Journal of High Performance Computing Applications". Volumen 15. 2001. Páginas (200-222).
- [3] Papadopoulos, P; Mason, J.; Bruno, G.. "NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters". Presentada en: Cluster: IEEE International Conference on Cluster Computing. 2001.
- [4] IBM. AIX cluster software documentation. Consultado en: www-03.ibm.com/servers/eserver/pseries/library/clusters/aix.html, 17-11-2005.
- [5] Altair Grid Technologies. OpenPBS. Consultado en: www.openpbs.org 17-11-2005
- [6] SUN microsystems. Sun Grid Engine. Consultado en: <http://gridengine.sunsource.net> 17-11-2005
- [7] Kannan, S.; Roberts, M.; Mayes, P.; Brelford, D.; Skovira, J.F.. "Workload Management with LoadLeveler". IBM. "IBM RedBooks". 2001. ISBN 0738422096.
- [8] The Globus Alliance. Globus Toolkit. Consultado en: www.globus.org 17-11-2005.
- [9] PKIRISGrid. Consultado en: www.irisgrid.es/pki/policy/ 23-11-2005
- [10] Badia, R. M.; Labarta, J.; Sirvent, R.; Pérez, J.M.; Cela, J.M.; Grima, R.. "Programming Grid Applications with Grid SuperScalar". "Journal of Grid Computing". Vol. 1. 2003. Páginas (151-170).
- [11] Gamess. Consultado en: www.msg.ameslab.gov/GAMESS/GAMESS.html 17-11-2005
- [12] Reyes, S.; Muñoz-Caro, C.; Niño, A.; Badia, R. M.; Cela, J. M. "Performance of Computationally Intensive Parameter Sweep Applications on Internet-based Grids of Computers: the Mapping of Molecular Potential Energy Hypersurfaces. Concurrency and Computation: Practice and Experience. Publicado Online: 28 sep 2006
- [13] Mock, S.; Thomas, M.; Von Lazewski, G.. "The Perl Commodity Grid Toolkit", Presentada en: Grid Computing Environments: Special Issue of Concurrency and Computation: Practice and Experience. 2001.