Título: GROMACS y los sistemas distribuidos

Autor: Indira Rodriguez Fernández
Centro Nacional de Sanidad Agropecuaria, Cuba
indira@censa.edu.cu

Introducción

Las ciencias de la vida han alcanzado un rápido desarrollo en los últimos años. Este acelerado desarrollo no hubiera sido posible sin la vinculación a las ciencias de la computación, naciendo así nuevas ramas de corte multidisciplinario como la Quimioinformática y Neuroinformática, entre otras.

Con el surgimiento del proyecto genoma humano, la Biología y la Informática se unieron dando nacimiento a una nueva rama del conocimiento denominada Bioinformática (1).

Los proyectos genoma han generado una desmesurada cantidad de información que se hace imposible procesar de forma manual, por lo que el uso de sistemas de bases de datos, algoritmos de minería de datos, inteligencia artificial, entre otros, han sido determinantes para el procesamiento y entendimiento de los sistemas biológicos.

En todas estas áreas del saber se encuentran problemas de costos computacionales, ya sea por lo complejo de los algoritmos a utilizar, por el volumen de datos a procesar o por ambos.

Para enfrentar estos problemas de gran demanda de cómputo existen dos tendencias a nivel mundial. Existen mercados en el área de hardware en el mundo que desarrollan microprocesadores cada vez más rápidos y potentes pero a precios muy altos, los cuales hacen del paralelismo transparente; un ejemplo de lo anteriormente expuesto es el microprocesador Xeon que funciona como si constara de dos microprocesadores virtuales, por lo que puede dividir el trabajo en procesos que se pueden planificar y ejecutar de forma independiente y el microprocesador Itanium que puede ejecutar simultáneamente diferentes instrucciones de un programa. La otra dirección es la de conexión de varios computadores en red, no necesariamente ubicados en el mismo lugar y que cooperan para resolver un problema común.

En esta dirección se destacan los Sistemas de Computación Paralelos (SP) y Distribuidos (SD), una vía de solución más asequible al brindar alta potencia de cálculo sin necesidad de invertir grandes sumas de dinero. Esto ha llevado a que instituciones u organizaciones creen su propia infraestructura de computación distribuida.

En la rama de la Bioinformática se realizan muchas simulaciones de procesos químicos como acoplamientos moleculares o reacciones químicas. Uno de los paquetes informáticos que permite realizar estas tareas es el GROMACS. La complejidad de los algoritmos contenidos dentro del paquete de software hace que las corridas del mismo sean costosas desde el punto de vista computacional. Basta señalar que para simular 10 nano segundos de un proceso de acoplamiento entre dos moléculas se pueden requerir días de cálculo. Si a esto le sumamos que en ocasiones es necesario simular estos procesos de acoplamiento para diferentes complejos moleculares podrá entenderse que el problema se hace más costoso aún.

Sistemas distribuidos

Existen varios conceptos en el mundo de sistemas distribuidos, entre los que se destacan:

- Un Sistema Distribuido no es más que una colección de computadoras independientes que aparentan ser para el usuario del sistema una única computadora. (2)
- Sistemas cuyos componentes hardware y software, que están en ordenadores conectados en red, se comunican y coordinan sus acciones mediante el paso de mensajes, para el logro de un objetivo. Se establece la comunicación mediante un protocolo prefijado por un esquema cliente-servidor. (3)

Hay dos aspectos importantes que un sistema debe cumplir para que sea distribuido, uno está relacionado con el hardware, pues las máquinas deben ser autónomas, y el otro, está relacionado con el software: para los usuarios, el sistema deberá ser una única computadora.
Aunque en la actualidad no se ha establecido un acuerdo, muchos autores plantean que entre los Sistemas Distribuidos se encuentran los Sistemas de Cómputo en Paralelo (SP) y los Sistemas de Cómputo Distribuido (SD).
Los SD se caracterizan por presentar un grupo de elementos: unidad computacional que puede ser un proceso, un procesador, un *switch*, entre otros. Estos elementos se interconectan a través de una red de comunicación, cooperan entre sí con el propósito de resolver una tarea común y se comunican mediante la recepción y el envío de mensajes actuando de forma espontánea y "autónoma" (4). En estos sistemas es común encontrarse un servidor que monitorea los servicios para llevar a cabo todo el proceso, por lo que el modelo del sistema responde a un Modelo Cliente-Servidor.
Los SP interconectan un grupo de elementos que se comunican para tarea; a diferencia a diferencia de los SD, tienen como principio dividir las aplicaciones en subtareas que son resueltas concurrentemente. Para aplicar el principio de la Computación Paralela se debe pensar en cómputos muy largos, con el requerimiento de ser posible dividirlos en subtareas que puedan procesarse de forma independiente.
Aunque ambos están concebidos para fortalecer la capacidad de cálculo en una red de computadoras, si se tiene en cuenta que el objetivo de los SP es alcanzar el máximo *speedup* en la solución de un problema, seguramente, muchas organizaciones se inclinarían por esta alternativa, sin embargo, puede ser compleja su aplicación. Entre otros aspectos, para aplicar los SP, se necesita comprender completamente el algoritmo secuencial que da solución al problema para poderlo dividir en subtareas totalmente independientes, y se requieren formas de coordinar los accesos a recursos compartidos, por ejemplo, MPI (Message Passing Information), que coordina los accesos mediante el paso de mensajes.
Otro de los aspectos a tener en cuenta, es que en estos sistemas no es frecuente que el número de procesadores o la forma de interconexión cambie en el transcurso del tiempo, por lo menos durante la ejecución, además, los elementos en la red de interconexión deben estar separados por pequeñas distancias.
Por otro, si se piensa en un SD, el programador no deberá conocer los códigos del programa secuencial que se quiere compartir, a diferencia de los SP, para reducir los tiempos de cálculos. Su principio es ejecutar en los clientes las mismas instrucciones con diferentes datos aprovechando los recursos existentes.

Aunque no suelen ser los SD completamente rápidos, sus elementos están diseñados para actuar de forma independiente, por lo que si falla uno, no implica que falle el sistema completo, estos pueden estar separados por centenares de metros o kilómetros.
Los SD permiten además, la aparición de recursos o su desaparición, incluso en tiempo de ejecución. Las instrucciones se pueden ejecutar en arquitecturas heterogéneas, y se admiten varias aplicaciones a la vez. Dentro de estos sistemas los Sistemas de *Computación Grid,* los cuales tienen como principio no solo la compartición de datos sino de otros recursos como: *Computadoras, Redes e Instrumentos.* En estos sistemas la seguridad es requerimiento básico para su funcionamiento. (5)

Ventajas de los Sistemas Distribuidos

- Buena relación costo beneficio: en la actualidad es común encontrarse procesadores que ejecuten un gran número de instrucciones, en consecuencia el precio de estos es elevado. Los SD ejecutan de igual modo un grupo de instrucciones y su precio lo hace más asequible.
- Velocidad: con la aplicación de los sistemas distribuidos se logra mayor velocidad que con un único computador.
- Independencia de fallo de recursos: el fracaso de algún recurso lógico o físico no implica que se destruya el sistema.
- Fiabilidad o confiabilidad: después de la ocurrencia de algún fallo, el sistema debe proporcionar los medios para reconfigurarlo o debe reasignar la tarea.
- Distribución inherente: un grupo de personas ubicadas en lugares diferentes trabajan juntas para realizar una tarea común.
- Compartición de datos: permiten el acceso de varios usuarios a datos comunes, como servidores de bases de datos, servidores de cómputo, virtualizaciones, entre otros.
- Compartición de dispositivos: permite a muchos usuarios compartir costosos periféricos como impresoras, dispositivos de almacenamiento, entre otros.
- Flexibilidad y extensibilidad: los SD son capaces de crecimiento incremental y proporcionan la extensión o modificación. Se adaptan a un ambiente cambiante sin desestabilizar sus operaciones.

Desventajas de los sistemas distribuidos

Uno de los problemas está relacionado con el software: aunque los SD están tomando gran fuerza, no se tiene mucha experiencia en su diseño, implementación y su uso. Es difícil aún dar respuestas a preguntas como: ¿qué sistema operativo es el más adecuado?, ¿qué lenguaje de programación usar?, ¿cuáles aplicaciones son apropiadas para el sistema?, ¿cuánto deberían conocer los usuarios sobre la distribución?
El segundo problema está relacionado con la red de comunicación, se pueden perder mensajes. Con el objetivo de recuperarlos el sistema deberá incluir software especiales y la red puede llegar a cargarse, cuando esto sucede debe ser reemplazada o una segunda red debe ser agregada. Una vez que el sistema dependa de la red, su saturación puede negar muchas ventajas a lograr para lo cual el sistema distribuido fue concebido.
Finalmente el intercambio de datos puede convertirse en un problema sino se trabaja bien con la seguridad ya que las personas pueden acceder a información sensible.

Aplicación de los Sistemas Distribuidos en la Bioinformática

La Bioinformática utiliza las tecnologías de la información para captar, organizar, analizar y distribuir informaciones biológicas con el objetivo de responder preguntas complejas en Biología. Es el resultado de la convergencia de la informática con la Bioiquímica, la Genética, la Biología molecular, entre otras, posibilitándoles valorar de manera integrada los datos que aceleran cada ves más los procesos de investigación.(6)

El uso de las computadoras para resolver cuestiones biológicas comenzó con el desarrollo de algoritmos y su aplicación al estudio de las interacciones de los procesos biológicos y las relaciones filogenéticas entre diferentes organismos. En los últimos años, el incremento de la cantidad de secuencias disponibles de proteínas, de ADN, entre otras, la alta demanda de cómputo, y la complejidad de las técnicas que emplean dichas computadoras para la adquisición y el análisis de los datos conlleva al uso de los SP y SD con el objetivo de procesar de forma más eficiente las informaciones biológicas.

GROMACS

El paquete de software GROMACS permite la simulación de procesos químicos como acoplamientos moleculares o reacciones químicas con la finalidad de determinar sus energías de asociación estadísticas y la obtención de la o las estructuras más favorecidas energéticamente en todo el espacio considerado(7).

GROMACS contiene todos los algoritmos que puede tener una aplicación moderna de dinámica molecular, conteniendo características que lo distinguen de la competencia:

- Proporciona un rendimiento muy alto en comparación con todos los demás programas. En su código presenta una gran cantidad de optimizaciones algorítmicas, que en tiempo de compilación son adoptadas para su arquitectura. Esto da como resultado un rendimiento excepcional en estaciones de trabajo de PC de bajo costo.
- Contiene topologías y archivos de parámetros escritos en forma de texto sin cifrar.
- A medida que la simulación se está llevando a cabo, GROMACS continuamente dirá que tan lejos se ha llegad, y a que fecha y hora debe de terminarse.
- Tanto los archivos de entrada para la ejecución y las trayectorias son independientes del hardware y pueden ser leídos por cualquier versión de GROMACS.
- Puede escribir las coordenadas utilizando la compresión con pérdida, lo que proporciona una forma muy compacta de almacenamiento de datos de la trayectoria. La exactitud puede ser seleccionada por el usuario.
- Contiene una amplia selección de herramientas flexibles para el análisis de la trayectoria, no se tendrá que escribir ningún código para realizarlos análisis de rutina.
- Contiene varios algoritmos que permiten reducir el tiempo de los intervalos en las simulaciones y así mejorar aún más el rendimiento.
- Es un paquete de software libre, disponible bajo la licencia Pública general (General Public License, GNU).

Conclusiones:

Los sistemas distribuidos constituyen actualmente la solución a los problemas de los costos computacionales, pues permite distribuir tareas y utilizar, entre

otras cosas, paquetes de software para diferentes tareas como las simulaciones y optimizar su uso. GROMACS requiere de soluciones informáticas optimizadas que faciliten el trabajo con grandes volúmenes de información mientras se realizan simulaciones como acoplamientos moleculares o reacciones químicas

Bibliografía

1. **Cañedo,R.;Arencibia,R. *Bioinformática: en busca de los secretos moleculares de la vida.* La Habana: s.n., 2004.**
2. **S.Tanenum, A. *Distributed Operating Systems***
3. **Colulouris, G. *Sistemas Distribuidos.* Tercera Edición. Addison Wesley. Madrid. 2001.**
4. **Colulouris, G. *Sistemas Distribuidos.* Madrid: s.n., 2001**
5. **Fernández, A. Arquitectura GRID orientadas a la gestión de recursos. 2004**
6. **Bioinformática. Concepto de Bioinformática con** [consultado el 1de febrero del 20013][En línea]http://www.solociencia.com/biologia/bioinformatica-concepto.htm]
7. **GROMACS. About GROMACS [consultado el 4 de febrero del 2013][En línea] http://www.gromacs.org/About_Gromacs**

SOCIEDAD DE LA INFORMACION

www.sociedadelainformacion.com

Edita: CEF

Director: José Ángel Ruiz Felipe
Jefe de publicaciones: Antero Soria Luján
D.L.: AB 293-2001
ISSN: 1578-326x