Original Message --------
Subject: another inquiry from Technology Research News
Date: Mon, 10 Dec 2001 22:37:39 -0500
From: Ted Bowen <tbowen@trnmag.com>
To: rajkumar@csse.monash.edu.au

Greetings,

I've just had a look at your paper, "The Virtual Laboratory: Enabling On-Demand Drug Design with the World Wide Grid".

I'd like to follow-up for a brief story.

A few questions:

1) How would you compare the performance of your grid-based database query and molecular docking applications to a theoretical single supercomputer performing docking tasks against a local database of molecular structures?

2) How does your speed-up method of performing network-based database queries compare to the 'current' state of database performance via grids? I understand database operations are sluggish in grid configurations. Does your method apply only to queries, or other operations?

3) What does your work do toward reducing latency in grid applications?

4) To clarify how the docking application was handled, were you re-writing the UCSF application to run better via grids, or did you write a separate application?

5) Can you clarify how the docking operations (comparison and fitting' of molecules) were handled remotely? What parts of the process ran where?

6) Have you completed the more extensive molecular screening tests, and if so, what did that show?

7) As per our usual TRN format, what was (were) the funding sources for the work, and how long would it be before you might see commercial/widespread use of the techniques and applications you developed?

Thanks very much,

Ted Bowen, contributing editor

Rajkumar Buyya with inputs from one of his collaborators, Kim Branson, answers to a press inquiry from the Technology Research News (www.trnmag.com). The work:

The Virtual Laboratory: Enabling Molecular Modelling for Drug Design on the World Wide Grid

is co-authored by Rajkumar Buyya, Kim Branson, Jon Giddy, and David Abramson. For a copy of the paper and software tools, please visit the Virtual Laboratory project website: http://www.buyya.com/vlab/

Answers to the Technology Research News Magazine Questions

1) How would you compare the performance of your grid-based database query and molecular docking applications to a theoretical single supercomputer performing docking tasks against a local database of molecular structures?

Fundamentally, molecular modeling for drug design involves screening millions of compounds in chemical databases to identify potential ones that can serve as drug candidates. This is both a computational and data challenge problem. Screening each compound, depending on structural complexity, can take from a few minutes to hours on a standard PC, which means screening all compounds in a single database can take years! Such a large time to discovery of drugs can be drastically reduced using parallel and distributed systems. Researchers have been using supercomputers such as clusters to solve such problems, but their ability to perform a large-scale exploration is still limited by the availability of processing power in a single supercomputer. With Grids, they can easily perform such large-scale exploration by using multiple computational resources (PCs, Workstations, SMPs, clusters, supercomputers) simultaneously and speed up the time to discovery.

Harnessing the Grid for drug design requires an extra effort to extract compounds from the remote databases, as they are not available locally on all resources for practical reasons. That means, performance improvement with Grid is not always proportional to the number of resources available, but scales well. For example, we are looking into a drug design problem involving screening 180,000 compounds and screening of each compound is expected to take 3 hours on a standard PC. That means, if we aim to screen all these compounds on a single PC, it can take up to 540000 hours, which is roughly equivalent to 61 years! If we use a typical cluster-based supercomputer with 64 nodes, we can solve this problem in one year. The problem can be solved with a large scale Grid of hundreds of supercomputers in a day. If we use a massive network of peer-to-peer style Grid computing infrastructure such as SETI@Home, the drug discovery problem could be solved within a few hours.

2) How does your speed-up method of performing network-based database queries compare to the 'current' state of database performance via grids? I understand database operations are sluggish in grid configurations. Does your method apply only to queries, or other operations?

Our CDB data management and access architecture supports several methods to speed up remote access to the database. They include indexing CDBs, multithreaded implementation of CDB servers, replication and federation of databases to release scalable distributed databases.

The chemical databases (CDBs) we use contain molecules whose structure is represented in the MOL2 text format. We have developed tools to index databases to enable a direct and faster access to a record in the database. When a remote request for compound structural information in the CDB is made, the parallel CDB server managing the database directly reads the record of interest and sends back to the requester. Multiple users can access the CDB server simultaneously over parallel communication streams.

Our CDB data management architecture supports a selective replication and federation of replicated CDBs to avoid a single CDB becoming a bottleneck. The data broker assists in the discovery and selection of a suitable CDB source within the database federation at runtime depending on their availability, network proximity, load, and the access price. The implications of such architecture and methodologies are enormous. They allow organizations or companies to share or provide access to their chemical databases as network service with customers or communities involved in the drug design for public good or profit.

3) What does your work do toward reducing latency in grid applications?

The Virtual Laboratory tools transform drug design application into a parameter sweep application for processing in parallel on distributed resources. The parameterized application contains multiple independent jobs, each screening different compounds to identify their drug potential. These jobs are computational intensive in nature and only a small proportion of the execution time is spent on data communication. Applications with this computational model have high *computation to communication* ratio and they tolerate low network latency when executed in parallel on Internet-wide distributed resources.

As I mentioned in our answer to the question 2, our data management system architecture supports faster, parallel, scalable, and remote access to databases through federation, distribution discovery, and selection of suitable databases. These techniques provide transparent, scalable, and faster access to datasets event though they are inter-connected through a network of different bandwidth, latency, load, and availability.

4) To clarify how the docking application was handled, were you re-writing the UCSF application to run better via grids, or did you write a separate application?

We have used our Nimrod-G parameter specification language to transform the UCSF '"DOCK" software as a parameter sweep application without any modification to the original code. This is good from a scientific standpoint since it means the calculations we perform are no different from calculations others can perform and are in the literature. Each job of parameterized "dock" application executes on the same code on different input data scenarios and compounds. We used the Nirmod-G resource broker to schedule docking jobs for execution on distributed Grid resources depending on their availability, performance delivery, and cost and user demands. It transparently handles parameterized input files into real input files at runtime. That means, Nimrod-G's capability of docking compounds in the databases in parallel without any modification to the original application code saves a significant amount of software engineering cost and the time.

5) Can you clarify how the docking operations (comparison and fitting' of molecules) were handled remotely? What parts of the process ran where?

As far as the execution of docking program for drug design is concerned, there is no difference between executing it on a single local computer or remote computer—the same code is executed in both the cases. What is different is, while docking on a remote Grid node, the execution environment and molecule record to be screened need to be staged before starting docking operation. The parts that are common for each docking operation are generation of the sphere centre to file the site and dock grid scores. These two programs are first executed on the user machine. The program *dock*, which matches spheres with ligand atoms and uses scoring grids to evaluate ligand orientations, is executed in parallel on distributed resources for all or selected compounds in the database. To make use of heterogeneous resources in the Grid, binary format input data files need to be generated for each type of machine. For example, scoring grids must be generated for each type of machine used. The broker selects suitable data files depending on the architecture of a machine on which the program dock is executed.

6) Have you completed the more extensive molecular screening tests, and if so, what did that show?

Our efforts so far have been dedicated to the development of Virtual Laboratory tools that enable the execution of molecular modeling for drug design application on the World Wide Grid testbed resources. The tools have been tested extensively with sample data sets in the CDB for a protein target involved in hypertension. Efforts are currently underway to perform extensive molecular screening, which is expected to produce interesting results.

7) As per our usual TRN format, what was (were) the funding sources for the work, and how long would it be before you might see commercial/widespread use of the techniques and applications you developed?

Our work is mainly supported by grants and scholarships from the Australian Government agencies, Monash University, CRC for the Enterprise Distributed Systems Technology (DSTC), Monash School of Computer Science and Software Engineering, and the IEEE Computer Society. Our collaborator Kim Branson's work is supported by the Walter and Eliza Hall Institute of Medical Research, CSIRO Health Sciences and Nutrition, the CRC for Cellular Growth Factors, and the Advanced Micro Device (AMD), which donated hardware components for building a cluster.

Given the current interest, revolution and investment in Life Sciences and Bioinformatics areas, we envision the widespread/commercial use of the Virtual Laboratory techniques within one to two years.