The Virtual Laboratory: A Toolset for Utilising the World-Wide Grid to Design Drugs

Rajkumar Buyya[†], Kim Branson[*], Jon Giddy[†], and David Abramson[†]

[†] School of Computer Science and Software Engineering
Monash University, Caulfield Campus
Melbourne, Australia
{rajkumar, davida, jon}@csse.monash.edu.au

[*] Structural Biology
Walter and Eliza Hall Institute
Royal Parade, Parkville, Melbourne
kbranson@wehi.edu.au

1 Introduction

Computational Grids [1] enable the sharing, *selection* and *aggregation* of distributed resources across multiple organizations for solving large-scale computational and data intensive problems. Molecular modeling for drug design is one of the scientific applications that can benefit from the availability of a large computational capability. Drug discovery is an extended process that can take as many as 15 years from the first compound synthesis in the laboratory until the therapeutic agent, or drug, is brought to market [6]. Reducing the research timeline in the discovery stage is a key priority for pharmaceutical companies worldwide. Many such companies are trying to achieve this goal through the application and integration of advanced technologies such as computational biology, chemistry, computer graphics, and high performance computing.

Molecular modeling for drug design involves screening millions of ligand records or molecules of compounds in a chemical database (CDB) to identify those that are potential drugs. This process is called molecular *docking*. It helps scientists explore how two molecules, such as a drug and an enzyme or protein receptor, fit together. Docking each molecule in the target chemical database is both a compute and data intensive task. It is our goal to use Grid technologies to provide cheap and efficient solutions for the execution of molecular docking tasks on large-scale, wide-area parallel and distributed systems.

While performing docking, information about the molecule must be extracted from one of a number of large chemical databases. Because the databases require storage space in the order of hundreds of megabytes to terabytes, it is not feasible to transfer the chemical database to all nodes in the Grid while processing. Therefore, access to a chemical database must be provided as *network service*. Also, the chemical database needs to be selectively replicated on a few nodes within the Grid to avoid any bottleneck due to providing access to the database from a single source. Intelligent mechanisms (e.g., CDB broker) need to be supported for selecting optimal sources for CDB services depending on the location of resources selected for processing docking jobs.

In this paper, we discuss a layered architecture of technologies and tools for creating the virtual laboratory environment for drug design application. We present the results of scheduling molecular docking jobs for processing on the WWG (World Wide Grid) resources [9].

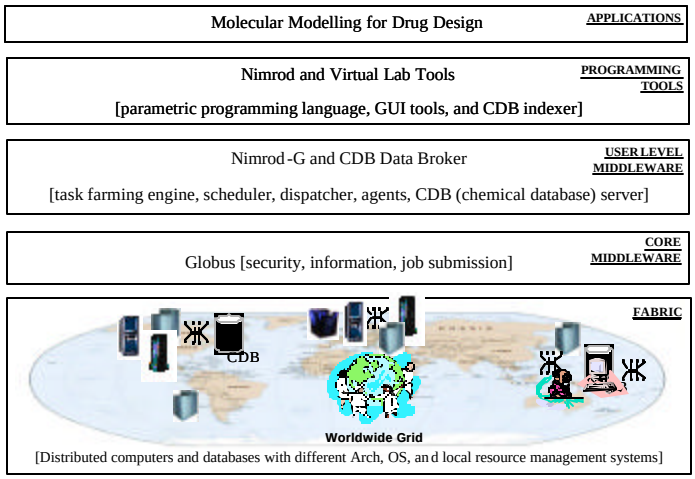

Figure 1: Layered architecture of Virtual Laboratory.

2 Architecture – The Software Stack

The Virtual Laboratory builds on the existing Grid technologies and tools for distributed processing of molecular docking application jobs. It uses the Nimrod-G parameter specification language [2] for composing a molecular docking application as a task-farming or parameter sweep application. The Nimrod-G grid resource broker [3] is used for scheduling and processing docking jobs on distributed resources. We developed new tools for proving access to ligand records in the Chemical Database (CDB), represented in MOL2 format [7]. A layered architecture and the software stack for creating Virtual Laboratory environment is shown in Figure 1. The components of the software stack are [10]:

- The DOCK software for Molecular Modeling [8].
- The Nimrod parameter-modeling language [2] for parameterising docking application.
- The Nimrod-G Grid resource broker [3] for scheduling DOCK jobs on the Grid.
- Chemical Database (CDB) Management and Intelligent Access Tools for:
  - Database lookup/index table generation.
  - Replica Catalogue for CDB resource discovery.
  - Servers for providing CDB services

- Brokering for selecting suitable database
- Fetching molecule record from the database.

- The Globus middleware for secure and uniform access to distributed resources [5].

3 Scheduling Experimentations

We have conducted scheduling experiments for docking 200 molecules (from the *aldrich_300* CDB) on a target receptor called endothelin converting enzyme (ECE), which is involved in hypotension. The WWG testbed resources used are show in Table 1. The scheduling experiments explored two different strategies, time optimization (TimeOpt) and cost optimization (CostOpt) algorithms [4], with 60 minutes as the deadline and 50,000 G$ as the budget limits. The *time* optimization experiment started on November 3, 2001 at 23:23:00, Australian Eastern Standard Time (AEST) and finished on November 3, 2001 by 23:57:00. It took 34 minutes to finish the processing of all jobs using resources available at that time with an expense of 17,702 G$. The cost optimization scheduling experiment was performed on November 4, 2001 at 00:08:00, AEST, and finished on November 4, 2001 by 01:07:30. It took almost 59.30 minutes to finish with an expense of 14,277 G$. The number of jobs processed at different times during these experiments is shown in Figure 2 and Figure 3.

Organization & Location, Resource-#CPUs	Price (G$/CPU sec.)	Number of Jobs Executed	
		TimeOpt	CostOpt
Monash, Australia: Sun Ultra-1	-- (Master node)	--	--
AIST, Japan: Sun Ultra-4	1	44	102
AIST, Japan: Sun Ultra-4	2	41	41
AIST, Japan: Sun Ultra-4	1	42	39
AIST, Japan: Sun Ultra-2	3	11	4
ANL, USA: Sun Ultra-8	1	62	14
Total Experiment Cost (G$)		17702	14277
Time to Finish Experiment (Min.)		34	59.30

Table 1: World-Wide Grid resources.

4 Conclusion

We have discussed the creation of a Virtual Laboratory environment for distributed processing of molecular docking application jobs on the World-Wide Grid. The results of scheduling experiments show the potential of Grids for distributed data-intensive computing.

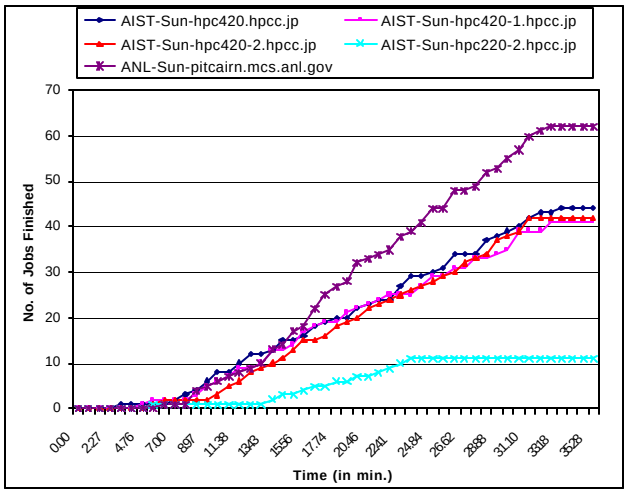

Figure 2: No. of jobs finished - time optimization.

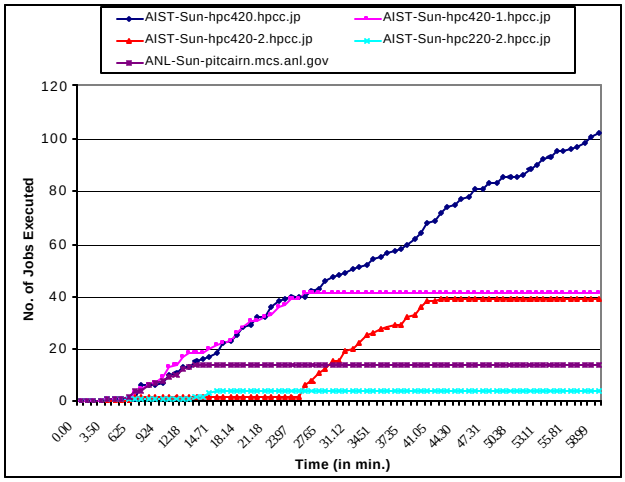

Figure 3: No. of jobs finished - cost optimization.

References

[1] I. Foster and C. Kesselman (eds.), *The Grid: Blueprint for a Future Computing Infrastructure*, Morgan Kaufmann Press, USA, 1999.

[2] D. Abramson et. al., *Nimrod: A Tool for Performing Parameterized Simulations using Distributed Workstations*, The 4th Intl. Symp. on High Performance Distributed Computing, August 1995.

[3] R. Buyya, D. Abramson, and J. Giddy, *Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid*, 4th Intl. Conf. on High Performance Computing in Asia-Pacific Region (HPC Asia 2000), China.

[4] R. Buyya, J. Giddy, and D. Abramson, *An Evaluation of Economy-based Resource Trading and Scheduling on Computational Power Grids for Parameter Sweep Applications*, The Second Workshop on Active Middleware Services (AMS 2000), Pittsburgh, USA

[5] Globus Project, *The Globus Toolkit*, http://www.globus.org

[6] E. Lunney, *Computing in Drug Discovery: The Design Phase*, http://computer.org/cise/homepage/2001/05Ind/05ind.htm

[7] *SYBYL Mol2 Format*, http://www.tripos.com/services/mol2/

[8] B. Shoichet, D. Bodian, and I. Kuntz, *Molecular docking using shape descriptors*, Journal of Compational Chemistry, 13(3), 1992.

[9] *World Wide Grid (WWG)*, http://www.buyya.com /ecogrid/wwg/

[10] R. Buyya, K. Branson, J. Giddy, and D. Abramson, *The Virtual Laboratory: Enabling Molecular Modeling for Drug Design on the World Wide Grid*, Tech. Report, Monash University, Dec. 2001.