Scalable BLAST Service in OBIGrid Environment

Fumikazu Konishi[1] **Yukimasa Shiroto**[2]
fumikazu@gsc.riken.jp shiroto@ats.nis.nec.co.jp
Ryo Umetsu[1] **Akihiko Konagaya**[1]
u-ryo@gsc.riken.jp konagaya@gsc.riken.jp

[1] Bioinformatics Group, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
[2] Scientific Platform Department Platform Software Division NEC Informatec Systems Ltd., 3-2-1 Sakado, Takatsu-ku, Kawasaki 213-0012, Japan

Keywords: Grid computing, OBIGrid, homology search

1 Introduction

OBIGrid [1, 2, 3, 4] is a grid environment composed of 27 organizations aimed for bioinformatics research, established by the Initiative for Parallel Bioinformatics (IPAB) and high performance biocomputing committee. The available computational resource at present (Oct. 2003) is as large as 492 CPU with 293 nodes over the Internet connected by virtual private net (VPN). Those resources are increasing year by year.

As increasing the amount of genome sequence information, more and more computational power is required especially for sequence matching such as BLAST search and DP matching. Parallel processing techniques are definitely helpful for this purpose. However, maximum computation power one can supply in his(her) institute is limited by physical conditions, such as lack of electronic power supply and machine room space, research budget and maintenance cost including manpower. In order to overcome this limitation and to avoid overinvestment by an institute, grid is one of the most attractive and promising approaches currently available.

In this study, we demonstrate how to make use of current grid technologies for bioinformatics applications by choosing BLAST ("OBIBlast") as the first example.

2 Method

Chief characteristics of our approach are in its scalability and expandability in terms of control parallelism and data transfer throughput. We can make use of three level of parallelism; query level, data level and thread level. Query level parallelism is obtained by dividing series of queries into several subtasks and by dispatching them to remote computers on grid. Data level parallelism can be obtained by decomposing target database into several subdatabases or restricting search range of the database on each node in PC clusters or shared memory parallel machine. Thread level parallelism can be obtained by executing plural commands in time sharing mode (user thread) or shared memory processor (kernel thread).

We developed OBIBlast service which is defined as the client-server application. Only application service interface is visible, however, implementation details such as distributed processing and parallel processing are invisible from users. OBIBlast automatically divides Blast queries into subtasks according to the ability of worker nodes. Authentication and user permission of the worker nodes are carried out through the Grid Security Infrastructure (GSI) in Globus Toolkits 2.0. All jobs are provided by the Globus Resource Allocation Manager (GRAM) and during the search against target databases,

the Blast-Server monitors job status (INITIALIZE, PENDING, ACTIVE, DONE, FAULT). When the all search is "DONE", the Blast-Server collects the results from the Blast-Workers and returns it to the users. The Blast-Server retries to submit the job to the others on the status of FAULT. The framework is general enough to apply to other bioinformatics applications such as Hmmer and FASTA.

3 Results and Discussion

The first prototype system consists of a 64-CPU PC cluster, a 12-CPU PC cluster and twenty worker PC nodes as well as an OBIBlast server. The PC clusters provide parallel Blast with data parallel search facilities while the worker PC nodes provide conventional NCBI Blast with multi-thread facilities. As shown in Table 1, OBIBlast achieved highly scalability in terms of the number of queries. It can compensate the overhead caused by grid operations, such as authentication and remote process invocation, when the number of queries is fairly large. Natural load balancing can be achieved by first-come first-served basis task assignment.

Table 1: A performance result of OBIBlast service*1.

Query Entry*2	Time (sec)*3	Time/Entry	Entry/Time(hour)*4
500	3,879	7.75	464
1,000	7,126	7.12	505
2,000	13,886	6.94	518
3,000	20,613	6.87	524

*1 14 sites 23 nodes 34 CPUs for each request

*2 Mycoplasma pneumoniae peptide sequence (278 letters)

*3 blastpgp -d nr -v 500 -b 500 -j 3 -T F -e 10 -F L
Database has an All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF

*4 Entry/Time*3600

References

[1] Konagaya, A., OBIGrid: towards a new distributed platform for bioinformatics, *21st IEEE Symposium on Reliable Distributed Systems (SRDS'02)*, 380–381, 2002.

[2] Konishi, F., Fukuzaki, A., Satou, K., Yamamoto, T., Defago, X., and Konagaya, A., OBIGrid: a new computing platfrom for bioinformatics, *Genome Informatics*, 13:484–485, 2002.

[3] Konishi, F., Umeda, H., Satou, K., and Konagaya, A., A network design for Open Bioinformatics Grid (OBIGrid), *Proc. The 3rd Annual Meeting '02, Chem-Bio Infomatics Society*, 192–193, 2002.

[4] `http://www.obigrid.org/`