

Hi-per BLAST: High Performance BLAST on PC Cluster System

Akira Naruse¹ Naoki Nishinomiya²
 naruse@jp.fujitsu.com nishinomiya.nao@jp.fujitsu.com
 Kouichi Kumon¹ Masahito Yamaguchi²
 kumon@jp.fujitsu.com yamaguchi.ma-07@jp.fujitsu.com

¹ Grid Computing & Bioinformatics Laboratory, Fujitsu Laboratories Ltd.,
 4-1-1 Kamikodanaka, Nakahara, Kawasaki, Kanagawa 211-8588, Japan

² Life Science & Material Science Dept., Fujitsu Ltd.,
 1-9-3 Nakase, Mihama, Chiba, Chiba 261-8588, Japan

Keywords: bioinformatics, pc cluster system, homology search, blast

1 Introduction

BLAST [1] is a homology search tool to identify new DNA/protein sequences. It is considered as the fastest among several homology search tools, and improved and maintained by several organizations. Above all, NCBI BLAST [3] developed by NCBI is accepted as a defacto standard and widely used.

NCBI BLAST is fast, however, it is getting longer to complete a BLAST run. That is mainly because the size of database grows rapidly. As a matter of fact, the size of DNA database becomes twice a year recently [2]. The pace of database growth exceeds that of performance growth of a single CPU. Therefore, a parallel version of BLAST that can effectively use multiple CPUs is desired. Several organizations such as TurboWorx Inc. [5] and Paracel Inc. [4] are developing the parallel version of BLAST. Our Hi-per BLAST is such a parallel version of BLAST, based on NCBI BLAST. The performance of Hi-per BLAST shows almost linear speedup to the number of CPUs, and over linear speedup in some cases.

In this paper, we first explain the components and the execution process of Hi-per BLAST, then show the performance measurement results on a Fujitsu PRIMERGY server cluster system.

2 Components and Execution Process

There are two components in Hi-per BLAST: a control node and compute nodes. Figure 1 shows co-operation between the control node and the compute nodes. The main role of the control node is coordination. The actual sequence search is done by compute nodes. The speedup of the search depends on the number of compute nodes and property of submitted user request.

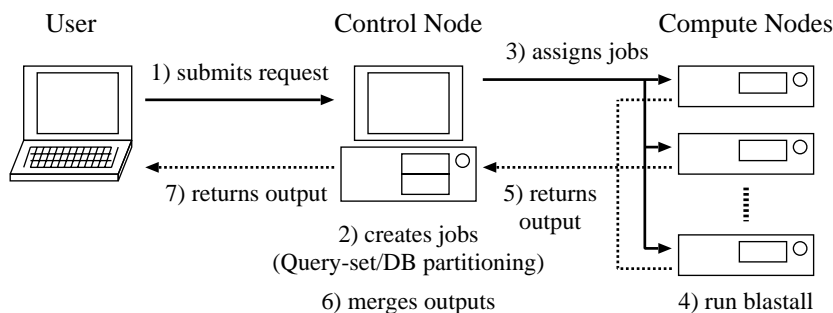


Figure 1: Overview of Hi-per BLAST.

3 Performance Measurement and Results

We prepared two DNA query-sets and two DNA databases for performance measurement, and used *blastn* program at all runs. Detail features of the cluster system, the query-sets and the databases are shown in Table 1.

Table 1: Cluster system, query-sets and databases used for performance measurement.

Cluster system		Query-sets (DNA)		
# of nodes	16	Name	# of sequences	average size of sequences
System	Fujitsu PRIMERGY TS220	Multiple	100	648
CPU	Intel PentiumIII 1GHz	Single	1	638
# of CPUs	2 /each node			
Memory	1GB /each node			
Network	100MbE			
OS	Red Hat Linux 7.1			
Databases (DNA)				
Name		# of sequences	# of letters	
Large (NCBI nt (Nov.2001))		1,000,461	4,365,428,743	
Small (Quarter NCBI nt)		250,556	1,021,297,621	

Table 2 and Figure 2 show the results of performance measurement. The performance of Hi-per BLAST shows almost linear speedup to the number of CPUs, and over linear speedup in case of the query-set consisting of multiple sequences and the large database.

Table 2: Performance measurement results.

Type of BLAST		Elapsed time (seconds)			
		NCBI	Hi-per BLAST		
# of CPUs		1	8	16	32
Query-set	Database				
Multiple	Large	4382.0	276.9	139.6	71.7
	Small	455.0	63.6	32.1	16.3
Single	Large	41.4	3.4	2.2	1.7
	Small	4.4	0.7	0.4	0.3

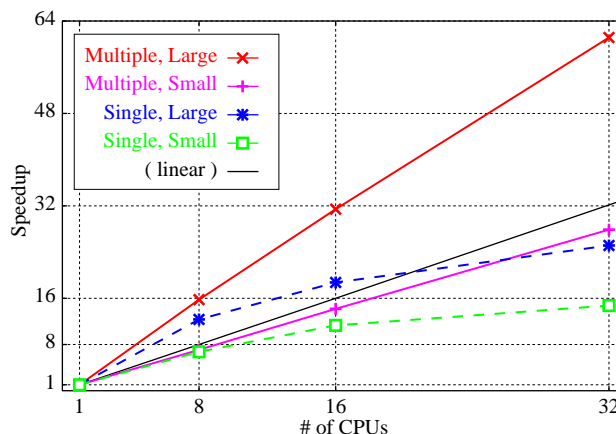


Figure 2: Performance measurement results.

4 Conclusions

Measured results indicate that the performance of Hi-per BLAST shows good scalability to the number of CPUs, especially when the number of sequences in a query-set is large and the size of database is large. Even in case of the query-set consisting of a single sequence, Hi-per BLAST shows sufficient scalability to the number of CPUs. That is because Hi-per BLAST can use a query-set/database partitioning technique and can create jobs for all compute nodes in any case.

Currently, we provide Hi-per BLAST based on NCBI BLAST 2.2.2~2.2.4, and will keep it up-to-date to the latest NCBI BLAST.

References

- [1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–410, 1990.
- [2] <http://www.dna.affrc.go.jp/htdocs/growth/index.html>
- [3] <http://www.ncbi.nlm.nih.gov/BLAST/>
- [4] <http://www.paracel.com/>
- [5] <http://www.turbogenomics.com/>