



# Parallel processing application for cluster computing on gridsim using Artificial Neural Network approach

Minal Dhoke<sup>1</sup>, Prof. Rajesh Dharmik<sup>2</sup>

<sup>1</sup>M.Tech IT Department,

Yeshwantrao Chavan College of Engineering, Nagpur,  
dhokeminal13@gmail.com

<sup>2</sup>H.O.D IT Department,

Yeshwantrao Chavan College of Engineering, Nagpur,  
raj\_dharmik@yahoo.com

## Abstract

Grid Computing is an emerging technology to provide high performance computing in a virtual organization composed of a large number of computers connected through web based technologies. We have implemented a parallel version of the Needleman-Wunsch algorithm for handling the DNA matching and alignment problem. We have presented the Needleman Wunsch algorithm for global alignments of sequences which obtains the best global alignments at the expense of very high computing power and huge memory requirements is integrated with the gridsim. Gridsim which supports the creation of *repeatable* and *controllable* Grid environments for quicker performance evaluation of scheduling strategies under different scenarios such as varying number of resources and users with different requirements.

**Keywords:** Needleman-Wunsch, Parallel Algorithm, Parallelization, Gridsim toolkit.

## 1. Introduction

Cluster computing is a fusion of the fields of parallel, high-performance, distributed, and high-availability computing, which also provides an excellent platform for solving a range of parallel and distributed applications which is integrated with the Artificial Neural Networks approach to provide a high degree of parallelism. Parallel computing is a form of computation in which many calculations are carried out simultaneously, operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently ("in parallel").

The primary objective of this project is to investigate effective resource allocation techniques based on computational economy through simulation. We like to simulate millions of resources and thousands of users with varied requirements and study scalability of systems, algorithms, efficiency of resource allocation policies and satisfaction of users. We are also interested to explore how significantly the local economy and the global positioning (e.g., the time zone) of a particular resource play role in securing jobs under various pricing and demand/supply situations. As such a large-scale simulation consumes large amount of computing power,

we would like to use parallel and cluster computing systems. In our simulation we would like to model applications in the areas of biotechnology, astrophysics, network design, and high-energy physics in order to study usefulness of our resource allocation techniques. The results of our work will have significant impact on the way resource allocation is performed for solving problems on cluster and grid computing systems.

Grid systems are classified into two categories: compute and data grids. In compute grids the main resource that is being managed by the resource management system is compute cycles while in data grids the focus is to manage data distributed over geographic allocations. The architecture and the services provided by the resource management system are affected by the type of grid system.

## 2. Parallel Needleman Wunsch's algorithm for global sequence alignment

The algorithm consists of two parts: the calculation of the total score indicating the similarity between the two given sequences, and the identification of the alignment(s) that lead to score. Thus to compare two sequences, we need to find the best alignment between

them, which is to place one sequence above the other making clear the correspondence between similar characters from the sequences [3]. In an alignment, gaps are inserted in arbitrary locations along the sequences so that they end up with the same size. Given an alignment between two sequences  $s$  and  $t$ , a score can be associated for it as follows. For each column, we associate +1 if the two characters are identical, 0 if the characters are different and -1 if one of them is a space. The score is the sum of the values computed for each column. The maximal score is the similarity between the two sequences, denoted by  $\text{sim}(s,t)$ . Figure 1 shows the alignment of sequences  $s$  and  $t$ , with the score for each column[2].

A	T	-	A	A	G	T
A	T	G	C	A	G	T
+1	+1	-1	0	+1	+1	+1

d=4

Figure 1. Alignment between  $s$  and  $t$ .

For long sequences, it is unusual to obtain a local alignment. Instead, the global alignment algorithm is executed to calculate the final score between sequences; local alignment algorithm is executed to detect regions inside both sequences that are similar. Global alignment algorithms are executed for similarity and final score. Figure 2 illustrates this.

Needleman and Wunsch proposed an algorithm (NW) based on dynamic programming to solve the global alignment problem. The time and space complexity of this algorithm is  $O(mn)$ , where  $m$  and  $n$  are the lengths of the two sequences, and, if both sequences have approximately the same length,  $n$ , we get  $O(n^2)$  [4]. The NW algorithm is divided into two parts: the calculation of the similarity array and the retrieval of the global alignments.

		A	T	G	C	A	G	T
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
T	-2	0	2	1	0	-1	-2	-3
A	-3	-1	1	2	1	1	0	-1
A	-4	-2	0	1	2	2	1	0
G	-5	-3	-1	1	1	2	3	2
T	-6	-4	-2	0	1	1	2	4

↑  
Final Score

Figure 2. Global alignment of two sequences with final score = 4

Using this algorithm in parallel, the reduction of time and space complexity is achieved. The results obtained to globally align real DNA sequences on two workstation of 12 cores each, present good speedup. We intend to implement this algorithm to compare very long DNA sequences (larger than IOMBP) in a computational grid.

### 3. Implementation of Needleman Wunsch algorithm on gridsim

Since the GridSim toolkit is an advanced and powerful simulation toolkit, its users will experience a high learning curve in order to utilize the toolkit functionalities for effective simulations. In addition, the users need to write Java code that use the toolkit packages to create the desired experimental scenarios. The GridSim toolkit provides a comprehensive facility for simulation different classes of heterogeneous resources, users, applications, and resource brokers. In Grid-like environment, resource brokers perform resource selection and aggregation depending on users requirements and hence they are *user-centric* in nature.

The gridsim has the facility so that the Users can be created with different requirements (application and quality of service requirements). These requirements include the baud rate of the network (connection speed), maximum time to run the simulation, time delay between each simulation, and scheduling strategy such as cost and/or time optimization for running the application jobs. The application jobs are modelled as Gridlets. The parameters of Gridlets that can be defined includes number of Gridlets, job length of Gridlets (in Million Instructions (MI)), and length of input and output data (in bytes). Each Grid user has its own economic requirements (deadline and budget) that constrains the running of application jobs. The Grid user can have the exact cost amount that it is willing to spend for the value-based option.

### 4. Building simulations with gridsim

The Java-based GridSim discrete event simulation toolkit provides Java classes that represent entities essential for application, resource modeling, scheduling of jobs to resources, and their execution along with management. In GridSim toolkit, we can create CPUs (also called *Processing Elements* (PEs)) with different MIPS (Million Instructions Per Second) or SPEC-like ratings. Then, one or more PEs can be put together to create a machine (a single CPU/SMP node). Such one or more machines can be put together to create a Grid resource. The resulting Grid resource can be a single processor, shared

memory multiprocessors (SMP), or a distributed memory cluster of computers.

A Gridlet is a tiny GridApp that contains all information related to jobs and job execution management details such as jobs processing requirements, expressed in MIPS, disk I/O operations, the size of input files, etc. that help in computing execution time of remote resource and the size of output files.

In the output, we have got the gridlet status, no. of gridlets, their resource id, and the execution time.

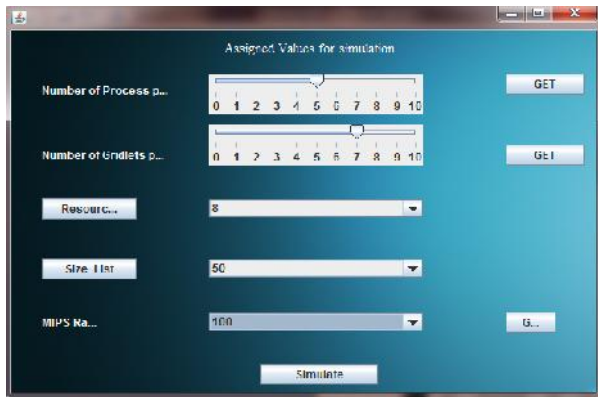


Figure 3. Assigning values to gridsim.

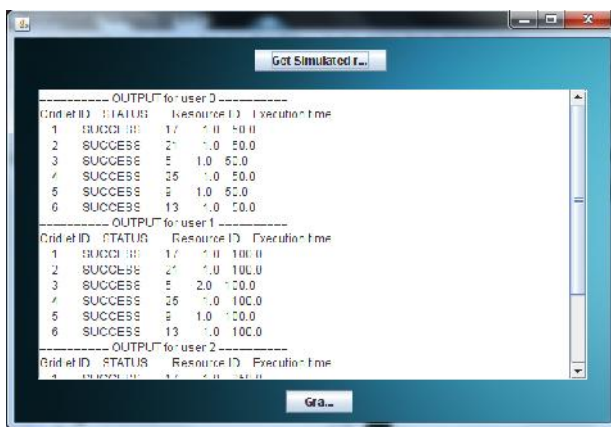


Figure 4. output of gridsim

The algorithm will be implemented parallelly using gridsim simulator like in cluster computing environment. We can select the architecture, operating system, no. of machines on which the algorithm will work or distributed.

The input given through the gridsim will be distributed over the machines that are used by it with all its specifications as shown in the figures above, and the output will be created on the gridsim only.

The design space for computers that can be used for ANN is very large, and even if it seems futile to talk about an optimal design there are a number of interesting trade-offs to be made. One basic trade-off is between *flexibility* and *speed*. That is, to increase speed we usually have to sacrifice some flexibility. In one end of the spectrum we have ordinary general purpose computers and in the other end we might have analog or optical computers. As in ANN the neurons are used, likewise the processing elements are used in parallel computing clusters. ANN approach is used to reduce more time to calculate the results in our system.

## 5. Experimental results

Sequences 20KBP, 50KBP, 100KBP, 250KBP and 500KBP align globally with 2, 4, 8, 12, 16, 20 and 24 processors with corresponding execution times on the gridsim are shown in figure 5 below and speedups with respect to processor are illustrated in figure 6. Speedups were calculated considering the total execution time and thus include times for initialization and collecting results.

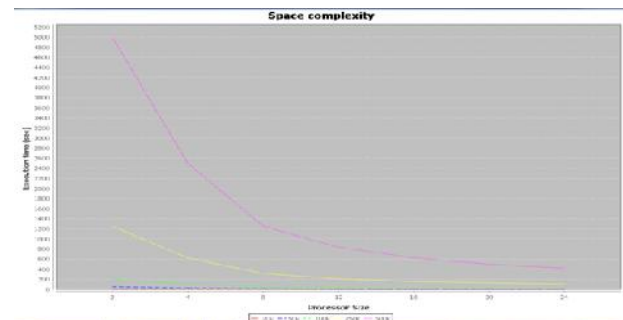


Figure 5. Execution time of 5 processors

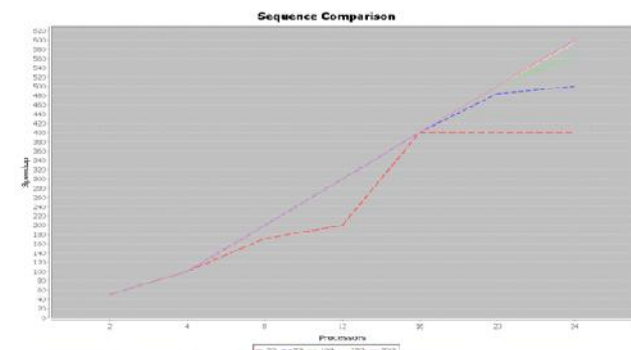


Figure 6. Speedups of 5 sequences

## 6. Related work

The two methods (heuristic and heuristic block) are used to reduce the time and space complexity in Needleman Wunsch algorithm. Heuristics to reduce the space complexity to  $O(n)$ , where  $n$  is the size of the sequences to be compared. For both the methods, the wavefront approach was used and work was assigned in a column basis. Synchronization was achieved with locks and condition variables. Barriers were only used at the beginning and at the end of computation.

## 7. Conclusion and Future work

In this paper, we are proposing an algorithm that will be used in parallel computing to reduce time and space complexity. We are using Gridsim toolkit that is the toolkit for creating clusters and it incorporates features such as easy-to-use wizards that enables users to create simulation models, and automatic source code generation facility that outputs ready-to-run simulation scenario code. ANN approach is used for reducing time for the calculations.

The very large DNA sequences for global alignment are intended to use in gridsim simulator by using large number of resources that can calculate the time more rapidly and reduce space as well.

## REFERENCES

- [1] Shrimankar D D, Sathé S R. "Implementation of Parallel Algorithms on Cluster of Workstations" 2nd IEEE International Conference on Parallel, Distributed and Grid Computing 2012
- [2] M. Baker, R. Buyya, D. Laforenza, "The Grid: International Efforts in Global Computing", International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, Italy, 2000.
- [3] Arie Keren and Amnon Barak "Opportunity Cost Algorithms for Reduction of I/O and Interprocess Communication Overhead in a Computing Cluster" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 14, NO. 1, JANUARY 2003.
- [4] H. Casanova, "Simgrid: A Toolkit for the Simulation of Application Scheduling", Proceedings of the First IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2001), May 15-18, 2001, Brisbane, Australia.
- [5] R. Buyya, D. Abramson, and J. Giddy, "A Case for Economy Grid Architecture for Service-Oriented Grid Computing", 10th IEEE International Heterogeneous Computing Workshop (HCW 2001), with IPDPS 2001, SF, California, USA, April 2001.
- [6] R. Melo, M. E. T. Walter, A. C. M. A. Melo, and R. B. Batista. "Comparing two long dna sequences using a dsm system". In Euro-Par 2003, pages 517-524. Springer-Verlag, 2003.
- [7] Leon Bottou and Yann LeCun. Large-scale on-line learning. In "In Advances in Neural Information Processing Systems" 15. MIT Press, 2004.
- [8] Naveed, Tahir, Imtiaz Saeed Siddiqui, and Shaftab Ahmed. "Parallel Needleman-Wunsch Algorithm for Grid." Proceedings of the PAK-US International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET 2005), Islamabad, Pakistan. 2005.
- [9] Ewing Lusk. Programming with MPI on clusters. In 3rd IEEE International Conference on Cluster Computing (CLUSTER'01), October 2001.
- [10] Buyya, R., Stockinger, H., Giddy, J., Abramson, D.: "Economic Models for Management of Resources in Peer-to-peer and Grid Computing". SPIE International Conference on Commercial Applications for High-Performance Computing, Denver, USA (August 20-24, 2001).
- [11] Buyya, R., Murshed, M.: "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing". The Journal of Concurrency and Computation: Practice and Experience, Vol. 14, Issue 13-15. Wiley Press (2002)
- [12] Buyya, R., Abramson, D., Giddy, J.: "Nimrod-G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid". Proceedings of 4th International Conference and Exhibition on High Performance Computing in Asia-Pacific Region (HPC Asia 2000). IEEE Computer Society Press, Beijing, China (May 14-17, 2000).
- [13] Alba Cristina M. A. Melo, Maria Emilia M. T. Walter, Renata Cristina F. Melo, Marcelo N. P. Santana, Rodolfo B. Batista, "Local DNA Sequence Alignment in a Cluster of Workstations: Algorithms and Tools", J. Braz. Comp. Soc. vol.10 no.2 Campinas Nov. 2004.