# Distributed and Parallel Processing for the Embgrid Project

**Sanz, Cecilia[2], Tinetti Fernando[3], Russo Claudia[4], Denham Monica[5], De Giusti, Armando[1], Grau Oscar[6]**
**{degiusti, csanz, fernando, crusso, mdenham}@lidi.info.unlp.edu.ar**
**grau@biol.unlp.edu.ar**

*LIDI. Laboratorio de Investigación y Desarrollo en Informática [7].*
*Facultad de Informática. UNLP.*
*50 y 115. 1er Piso. La Plata*

## Summary

The goal of EMBgrid is to overcome relevant infrastructure needs in Bioinformatics development and services in order to address future needs for the provision of useful Bioinformatics solutions to the Biosciences community. The proposed approach will make use of an extended Grid architecture, built on top of existing EU-DATAGRID services and deployed over a wide number of European nodes, to enable delivery of advanced tools and solutions to the growing demands of Biosciences. We contemplate creation of a pan-european Grid resource, related to other EU Grid initiatives and devoted mainly to Bioinformatics. This initiative will raise Europe's competitiveness in this and related fields. Successful deployment and dissemination will be facilitated by EMBnet long track in delivering tools, services and training to the European scientific community.

There is a cooperative argentinian group involved in this project, and has the responsability of analyze the specification, transformation, optimization and verification of concurrent algorithms executable in distributed/parallel systems; they also have to study the optimization of solutions types in function of multiprocessor architecture models and complex and efficiency metrics related to the parallel processing.

In Biocomputing, the need of distributing advanced tools and solutions has arisen for the growing data generation which takes place in a hyper-exponential manner, and for the solution of complex problems.

The specific goal proposed for this project is to analyze some Emboss programs, which are used in Biocomputing area (in particular, those related with alignment, and pattern matching), in order to transform them into parallel and distributed algorithms.

---

[1] **Investigador Principal CONICET. Profesor Titular Ded. Exclusiva.**
[2] **Profesor Dedicación Exclusiva - Facultad de Informática. UNLP.**
[3] **Profesor Dedicación Semi Exclusiva - Facultad de Informática. UNLP.**
[4] **Profesor Dedicación Exclusiva - Facultad de Informática. UNLP.**
[5] **Ayudante Alumno - Becaria LIDI - Facultad de Informática. UNLP.**
[6] **Profesor Dedicación Exclusiva - Director IBBM - Facultad de Ciencias Exactas. UNLP.**
[7] **LIDI - Facultad de Informática. UNLP - Calle 50 y 115 1er Piso, (1900) La Plata, Argentina.**
**TE/Fax +(54)(221)422-7707.** http://lidi.info.unlp.edu.ar

**Research Topics**

**Processing Resources Use Optimization**

Obtaining the maximum efficiency of a multiprocessor architecture represents an essential objective in parallel processing. The research in techniques that lead to upgrade the performance of a given architecture, or to adjust dynamically the processing resource assignation, is the axis of this line of work. As examples, we can quote specific techniques of:

- Charge Balance: processing and/or communications, depending on the application and the parallelization. All deviation as regards the charge equilibrium leads to the loss of the resource use and, therefore, to the loss of the general output.
- Process and processors distribution: the processes distribution in the architecture is specifically related to the processing charge balance and, therefore, it must be analyzed and optimized **[1]**.
- Migration of data and/or processes: by analyzing the cost-benefit relation of keeping the data static distribution and processing of all the application, or adapt it dynamically on the basis of the parallel execution.

**Parallelization of Applications**

The parallelization of applications with large computing requirements are objects to be studied for their computational resolution with parallel processing **[2][3]**. The classical techniques of :

a) Algorithm transformations,
b) Adaptation of algorithms to architecture models, and
c) Specification and verification of real time parallel systems,

are under process of research, looking for their improvement (at least, by areas of massive data processing problems), trying to turn them into applicable systematic methodologies with the minor quantity of potential adaptations to different problems within the area of the given problem.

**Parallelism Metrics**

In general, they are dealt with from several points of view, such as **[4]**:

a) Complexity analysis of sequential and parallel algorithms. Considering this analysis as a basis, a type of solution can be adopted rationally.
b) Output analysis, with which the quality or the cost-benefit relation to the parallel solutions proposed for the solution of the specific problem can be assessed.
c) Characterization of parallel applications and architectures in order to clearly identify both the applications and architectures, and to attain a supported proposal of the proposed solution type.

**Parallel Processing Architectures [5]**

In data massive processing, the following topics are relevant:

a) The cost-benefit relation and the output in function of the processing potential capacity of parallel computing architectures (such as parallel supercomputers and workstations nets or clusters)
b) Dynamic reconfiguration of hardware and/or software in order to adapt to the application and/or to propose failure tolerance.

Parallel hardware design oriented to application classes.

**Experimental work [6][7]**

- Analysis of Emboss sequential programs

It is necessary to evaluate which programs need to be improve in order to obtain a better performance for them in cases of massive data treatment. The algorithm specification and the parallel programming paradigm should be studied in order to optimize time response.

- Parallel system evaluation

On of the analysis to make is evaluate the costs of the parallel solutions. There are different alternatives for the parallel systems implementation with homogeneous or heterogeneous hardware and with weekly or strongly coupled components. These alternatives give various possibilities that can be traduced in costs (including communication costs), which must be evaluated taken into account the relation cost/performance. The costs are not only related with hardware but with the development of algorithms and/or the algorithms adjustment for the processing architecture.

The performance of a parallel system is given by a complex relation among different factors (size of the problem, support architecture, process distribution among processors, balance charge algorithm existence, etc.). There are numerous metrics to evaluate parallel systems, such as, speed up, efficient, etc.

- Developed parallel solutions

Developed parallel/solutions for Emboss programs related with alignment and pattern matching on a cluster of homogeneous PCs, studying absolute performance and speed up.

Study of the best distribution (analyzing the local replication grade) for large data bases related with this project, in order to optimize their accessibility. To achieve this goal, a simulation environment for data base development (implemented by LIDI`s group) will be used.

The experimental architecture models proposed (and available for the project) are four: homogeneous multiprocessor architectures highly coupled, transputer cube type; distributed multiprocessor architectures (homogeneous or heterogeneous) weakly coupled NOW type; multiprocessors dedicated to DSPs-based image treatment; and multiprocessor architectures of shared memory SGI Origin 2000 type (Clementina).

**Conclusions and Future work**

An important research area is the specification, transformation, optimization and evaluation of distributed and parallel algorithms. The optimization of solutions in function of different architecture models and the study of complexity and efficiency metrics are of special interest. These encompass the development of parallel processes, the transformation of sequential algorithms in parallel (exploiting the implicit or explicit concurrency in the problem to be solved) and the performance assessment metrics on different supporting platforms (both of hardware and software).

There exist several areas in which parallel and distributed programming is required - as it can be observed in the project "Development of Grid Environment for Biological Applications", in which is involved this research line.

The group has experience in the area of parallel and distributed processing applied to complex problems where large volumes of data are involved. Actually, emboss programs are being analyzed and studied, in order to obtain parallel solutions.

**Bibliography**

**[1]** Bell, David; Grimson, Jane, "Distributed Database Systems",. Addison Wesley. 1992
**[2]** Brinch Hansen, P., "Studies in computational science: Parallel Programming Paradigms", Prentice-Hall, Inc., 1995.
**[3]** Gupta A., Kumar V., "Performance properties of large scale parallel systems", Journal of Parallel and Distributed Computing, November 1993.
**[4]** Hwang K., "Advanced Computer Architecture: Paralelism, Scalability, Programability", McGraw, 1993.
**[5]** IEEE Transactions on Parallel and Distributed Processing (colección de revistas 1990-2002)
**[6]** Zomaya A., "Parallel Computing. Paradigms and Applications", Int. Thomson Computer Press, 1996.
**[7]** http://www.embnet.org/